SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 1–30

STRUCTURED PERTURBATIONS PART I: NORMWISE DISTANCES*

SIEGFRIED M. RUMP†

Abstract. In this paper we study the condition number of linear systems, the condition number of matrix inversion, and the distance to the nearest singular matrix, all problems with respect to normwise structured perturbations. The structures under investigation are symmetric, persymmetric, skewsymmetric, symmetric Toeplitz, general Toeplitz, circulant, Hankel, and persymmetric Hankel matrices (some results on other structures such as tridiagonal and tridiagonal Toeplitz matrices, both symmetric and general, are presented as well). We show that for a given matrix the worst case structured condition number for all right-hand sides is equal to the unstructured condition number. For a specific right-hand side we give various explicit formulas and estimations for the condition numbers for linear systems, especially for the ratio of the condition numbers with respect to structured and unstructured perturbations. Moreover, the condition number of matrix inversion is shown to be the same for structured and unstructured perturbations, and the same is proved for the distance to the nearest singular matrix. It follows a generalization of the classical Eckart–Young theorem, namely, that the reciprocal of the condition number is equal to the distance to the nearest singular matrix for all structured perturbations mentioned above.

Key words. normwise structured perturbations, condition number, distance to singularity

AMS subject classifications. 15A12, 65F35

PII. S0895479802405732

1. Motivation. Consider a numerical problem in m input parameters producing k output parameters, that is, a function $f : \mathbb{R}^m \to \mathbb{R}^k$. An algorithm to solve the problem, i.e., to compute f, in finite precision may be considered as a function $\tilde{f}$. A finite precision arithmetic for general *real* numbers may be defined to produce the best finite precision approximation to the (exact) real result (with some tie-breaking strategy). This includes the definition of the arithmetic for finite precision numbers. Then, for given input data $p \in \mathbb{R}^m$, the numerical result $\tilde{f}(p)$ will in general be the same for all $\tilde{p}$ in a small neighborhood of p. So we cannot expect more from a numerical algorithm than its producing the exact function value $f(\tilde{p})$ for some $\tilde{p}$ near p. An algorithm with this property is commonly called *backward stable*. For example, the standard method for solving an $n \times n$ dense system of linear equations, namely, Gaussian elimination with partial pivoting, is backward stable.

But is it always possible that $\tilde{f}(p) = f(\tilde{p})$ for some $\tilde{p}$ near p? Consider the computation in double precision floating point arithmetic according to IEEE standard 754 [30] of the square of a matrix, for example, of $A = \begin{pmatrix} 1+u & 4 \\ 4 & -1 \end{pmatrix}$, where $u = 2^{-52}$ such that 1 and $1+u$ are adjacent floating point numbers. The result is $\tilde{B} = fl(A^2) = \begin{pmatrix} 17 & 4u \\ 4u & 17 \end{pmatrix}$. For a perturbation $\Delta A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ we obtain

$$(A+\Delta A)^2 = \begin{pmatrix} (1+u+\alpha)^2 + (4+\beta)(4+\gamma) & (4+\beta)(u+\alpha+\delta) \\ (4+\gamma)(u+\alpha+\delta) & (4+\beta)(4+\gamma) + (1-\delta)^2 \end{pmatrix}.$$

But $(A+\Delta A)^2 = \tilde{B}$ is impossible for a small perturbation ΔA because this implies, by comparing with $\tilde{B}_{11}$ and $\tilde{B}_{22}$, that $(1+u+\alpha)^2 = (1-\delta)^2$, so that $u+\alpha = -\delta$ for

*Received by the editors April 17, 2002; accepted for publication (in revised form) by N. J. Higham January 13, 2003; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/40573.html
†Technical University of Hamburg-Harburg, Schwarzenbergstr. 95, 21071 Hamburg, Germany (rump@tu-harburg.de).

a small perturbation ΔA. But then $(A+\Delta A)_{12} = 0$. In other words, ordinary matrix multiplication yields the best double precision floating point approximation $\tilde{B}$ to the exact result A^2 but is not backward stable. A similar behavior is not uncommon for other structured problems.

Consider, for example, a linear system $Cx = b$ with a circulant matrix C. Many algorithms take advantage of such information, in terms of computing time and storage. In this case only the first row of the matrix and the right-hand side are input to a structured solver, so $m = 2n$ input data are mapped to $k = n$ output data. By nature, a perturbation of the matrix must be a circulant perturbation.

It is easy to find examples of $Cx = b$ such that for a computed solution $\tilde{x}$ it is *likely* that $(C + \Delta C)\tilde{x} \neq b + \Delta b$ for *all* small perturbations ΔC and Δb such that $C + \Delta C$ is a circulant. This happens although, as above, $\tilde{x}$ may be very close to the *exact* solution of the original problem $Cx = b$. The reason is that, in contrast to general linear systems, the space of input data is not rich enough to produce perturbed input data with the desired property. Or, in other words, there is some hidden structure in the result in contradiction to a computed approximation $\tilde{x}$.

In such a case, about all an algorithm can do in finite precision is to produce some $\tilde{x}$ such that $(C + \Delta C)(\tilde{x} + \Delta x) = b + \Delta b$. In our previous setting this means that for given input data p we require an algorithm $\tilde{f}$ to produce $q = \tilde{f}(p)$ with $q+\Delta q = f(p+\Delta p)$. An algorithm $\tilde{f}$ with this property is called *stable* (more precisely, mixed forward-backward stable) with respect to the distance measure in use [27, section 1.5]. Indeed, there are (normwise) stable algorithms to solve a linear system with circulant matrix [40]. This leads to structured perturbations and structured condition numbers.

There has been substantial interest in algorithms for structured problems in recent years (see, for example, [1, 22, 15, 19, 33, 10, 40, 5] and the literature cited therein). Accordingly, there is growing interest in structured perturbation analysis; cf. [36, 8, 24, 25, 2, 16, 4, 15, 7, 39, 37, 38, 14]. Moreover, different kinds of structured perturbations are investigated in robust and optimal control, for example, the analysis of the μ-number or structured distances [11, 13, 34, 41, 35, 29].

Particularly, many very fast structured solvers have been developed. Frequently, however, perturbation and error analysis for structured solvers are performed with respect to *general* perturbations. This is obviously improvable because usually for a structured solver nothing else but structured perturbations are *possible*.

However, structured perturbations are not as easy to handle, and a perturbation analysis of an algorithm concerning structured perturbations is generally difficult. Before investing too much into solving a problem, it seems wise to estimate its worth. In our case that means estimating the ratio between the structured and the unstructured sensitivities of a problem. For example, it is known that for a symmetric linear system and for normwise distances it makes no difference at all whether matrix perturbations are restricted to symmetric ones or not. In such a case the "usual" (unstructured) perturbation analysis is perfectly sufficient.

Explicit formulas for other structured condition numbers are known, but not too much is known about the ratio between the structured and the unstructured condition numbers. The aim of this two-part paper is to investigate this problem for a number of common (linear) perturbations for linear systems and for matrix inversion. Part I deals with normwise distances and Part II with componentwise distances.

One result of this first part is that for normwise distances, and for structures that are symmetric Toeplitz or circulant, the general (unstructured) condition number of

a linear system may be up to about the square of the structured condition number, much as it is when solving a least squares problem using normal equations rather than some numerically stable method. Although for many structures there seems currently no stable algorithm in sight, that is, stable with respect to structured perturbations, this creates a certain challenge (see also the last section of Part II of this paper).

2. Introduction and notation. Let nonsingular $A \in M_n(\mathbb{R})$ and $x, b \in \mathbb{R}^n, x \neq 0$ be given with $Ax = b$. The (normwise) condition number of this linear system with respect to a weight matrix $E \in M_n(\mathbb{R})$ and a weight vector $f \in \mathbb{R}^n$ is defined by

$$\kappa_{E,f}(A,x) := \lim_{\varepsilon\to 0} \sup \left\{ \frac{\|\Delta x\|}{\varepsilon\|x\|} : \ (A+\Delta A)(x+\Delta x) = b + \Delta b,\ \Delta A \in M_n(\mathbb{R}), \right. \\ \left. \Delta b \in \mathbb{R}^n, \|\Delta A\| \leq \varepsilon\|E\|,\ \|\Delta b\| \leq \varepsilon\|f\| \right\}. \tag{2.1}$$

In definition (2.1) the parameters E and f are only used as scaling factors and may be replaced by $\|E\|$ and $\|f\|$, respectively. However, in Part II of this paper we treat componentwise perturbations, and there we need the matrix and vector information in E and f. So we use the indices E, f in (2.2) to display certain similarities between normwise and componentwise perturbations.

Throughout this paper we always use the spectral norm $\|\cdot\|_2$, where we denote the matrix norm and the vector norm by the same symbol $\|\cdot\|$. It is well known [27, Theorem 7.2] that

$$\kappa_{E,f}(A,x) = \|A^{-1}\|\,\|E\| + \frac{\|A^{-1}\|\,\|f\|}{\|x\|}. \tag{2.2}$$

Note that the (unstructured) condition number does not depend on x but only on $\|x\|$. For no perturbations in the right-hand side is the condition number even independent of x. That means ill-conditioning is a matrix intrinsic property. This will change for structured perturbations.

By definition (2.1), a perturbation of size eps in the input data A and b creates a distortion of size $\kappa \cdot \text{eps}$ in the solution. Therefore, we cannot expect a numerical algorithm to produce an approximation $\tilde{x}$ better than that; that is, $\|\tilde{x} - x\|/\|x\|$ will not be much less than $\kappa \cdot \text{eps}$. On the other hand, we may regard an algorithm to be stable if it produces an approximation $\tilde{x}$ of this quality, i.e., $\|\tilde{x} - x\|/\|x\| \sim \kappa \cdot \text{eps}$.

In case the matrix A has an additional structure such as symmetry or Toeplitz, the structure may be utilized to improve performance of a linear system solver. For example, we have the remarkable fact that the inverse of a (symmetric) Toeplitz matrix can be calculated in $\mathcal{O}(n^2)$ operations, the time it takes to print the entries of the inverse [18, Algorithm 4.7.3].

Usually, such a specialized solver utilizes only part of the input matrix, for example, only the first row in the symmetric Toeplitz case—the other entries are assumed to be defined according to the given structure. This implies that only structured perturbations of the input matrix *are possible*. Perturbations of the input matrix are structured by nature as, for example, symmetric Toeplitz. Accordingly, perturbation theory may use a *structured condition number* defined similarly to (2.1):

$$\kappa^{\text{struct}}_{E,f}(A,x) := \lim_{\varepsilon\to 0} \sup \left\{ \frac{\|\Delta x\|}{\varepsilon\|x\|} : \ (A+\Delta A)(x+\Delta x) = b + \Delta b,\ \Delta A \in M_n^{\text{struct}}(\mathbb{R}), \right. \\ \left. \Delta b \in \mathbb{R}^n, \|\Delta A\| \leq \varepsilon\|E\|,\ \|\Delta b\| \leq \varepsilon\|f\| \right\}. \tag{2.3}$$

For other definitions of structured condition numbers see [16] and [17]. The set $M_n^{\text{struct}}(\mathbb{R})$ depicts the set of $n \times n$ real matrices with a certain structure *struct*. In this paper we will investigate the linear structures

$$\text{struct} \in \{\text{sym}, \text{persym}, \text{skewsym}, \text{symToep}, \text{Toep}, \text{circ}, \text{Hankel}, \text{persymHankel}\} \tag{2.4}$$

depicting the set of symmetric, persymmetric, skewsymmetric, symmetric Toeplitz, general Toeplitz, circulant, Hankel, and persymmetric Hankel matrices. In view of (2.3) note that for $A \in M_n^{\text{struct}}(\mathbb{R})$ for any of the structures in (2.4) it is $\Delta A \in M_n^{\text{struct}}(\mathbb{R})$ equivalent to $A + \Delta A \in M_n^{\text{struct}}(\mathbb{R})$. We will derive explicit formulas or estimations for κ^{struct}. Particularly, we will investigate the ratio $\kappa^{\text{struct}}/\kappa$.

Consider, for example, the tridiagonal matrix

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \ddots \\ & & \ddots & \ddots \end{pmatrix}. \tag{2.5}$$

The traditional (unstructured) condition number (2.1), (2.2) for the natural weights $E = A$ and $f = b$ satisfies

$$\kappa_{A,Ax}(A, x) > 4 \cdot 10^{11}$$

for A as in (2.5) of size 10^6 rows and columns and for arbitrary solution x, and hence arbitrary right-hand side. For the specific solution $x = (x_i)$, $x_i = \sin(y_i)$ with y_i equally spaced in the interval $[a,\ k\pi - a]$ for $a = 13/6000$, $k = 690$, we have

$$\kappa_{A,Ax}^{\text{symtridiagToep}}(A, x) < 9.6 \cdot 10^5, \tag{2.6}$$

where perturbations are symmetric Toeplitz and tridiagonal. Note that in this case the matrix depends only on two parameters. For $x = (1, -1, 1, -1, \ldots)^T$ and no perturbations in the right-hand side we get

$$\kappa_{A}^{\text{symtridiagToep}}(A, x) < 0.6. \tag{2.7}$$

We will derive methods to estimate and compute structured condition numbers. We will especially focus on the ratio $\kappa^{\text{struct}}/\kappa$. We will prove (see Theorem 5.3)

$$\kappa_{E,f}^{\text{struct}}(A, x) = \kappa_{E,f}(A, x) \quad \text{for} \quad \text{struct} \in \{\text{sym}, \text{persym}, \text{skewsym}\}$$

and all $0 \neq x$ in $\mathbb{R}^n$. This extends a result in [24]. By estimations and examples we show that the ratio can be significantly less than 1 for perturbations subject to the other structures in (2.4). Among others, we will prove (see Theorems 8.4, 9.2, and 10.2)

$$1 \geq \frac{\kappa_{A,Ax}^{\text{struct}}(A, x)}{\kappa_{A,Ax}(A, x)} \geq \frac{1}{2\sqrt{2}\sqrt{\|A^{-1}\|\,\|A\|}}$$

for struct $\in \{\text{symToep}, \text{Toep}, \text{circ}, \text{Hankel}, \text{persymHankel}\}$. On the other hand, we will show that to every structure an easy-to-calculate matrix Ψ_x is assigned, depending only on the structure and the solution x, with the surprising result that the ratio $\kappa^{\text{struct}}/\kappa$ can only become small when the smallest singular value $\sigma_{\min}(\Psi_x)$ is small. So

the ratio can only become small for certain *solutions*, independent of the (structured) matrix.

Furthermore, we will investigate the structured condition number for matrix inversion

$$\kappa_E^{\text{struct}}(A) := \lim_{\varepsilon \to 0} \sup \left\{ \frac{\|(A+\Delta A)^{-1} - A^{-1}\|}{\varepsilon \|A^{-1}\|} : \ \Delta A \in M_n^{\text{struct}}(\mathbb{R}),\ \|\Delta A\| \le \varepsilon \|E\| \right\}.$$

The definition includes the traditional (unstructured) condition number $\kappa_E(A)$ for matrix inversion by setting $M_n^{\text{struct}}(\mathbb{R}) := M_n(\mathbb{R})$. It is well known that $\kappa_E(A) = \|A^{-1}\|\,\|E\|$ [27, Theorem 6.4]. Here we will show that

$$\kappa_E^{\text{struct}}(A) = \|A^{-1}\|\,\|E\| \quad \text{for all structures as in (2.4).}$$

In most cases this is not difficult to prove. However, for Hankel and general Toeplitz perturbations we have to show that

$$\text{for all } x \in \mathbb{R}^n \ \exists\, H \in M_n^{\text{Hankel}}(\mathbb{R}) : \ Hx = x \quad \text{and} \quad \|H\| \le 1.$$

It seems natural to consider an ill-conditioned matrix to be "almost singular." Indeed, for normwise and unstructured perturbations the distance to singularity

$$\delta_E(A) := \min \left\{ \frac{\|\Delta A\|}{\|E\|} : \ A + \Delta A \text{ singular} \right\} \tag{2.8}$$

is well known to be equal to the reciprocal of the condition number (with no perturbation in the right-hand side) [27, Theorem 6.5]:

$$\delta_E(A) = \kappa_E(A)^{-1} = (\|A^{-1}\|\,\|E\|)^{-1}.$$

We may ask whether this carries over to structured perturbations. The structured (normwise) distance to singularity is defined accordingly by

$$\delta_E^{\text{struct}}(A) := \min \left\{ \frac{\|\Delta A\|}{\|E\|} : \ A + \Delta A \text{ singular},\ \Delta A \in M_n^{\text{struct}}(\mathbb{R}) \right\}. \tag{2.9}$$

Indeed we will show that for all structures (2.4) under consideration δ_E^{struct} is equal to $\kappa_E^{\text{struct}}(A)^{-1}$.

We will use the following notation:

$M_n(\mathbb{R})$	set of real $n \times n$ matrices
$M_n^{\text{struct}}(\mathbb{R})$	set of structured real $n \times n$ matrices, struct as in (2.4)
$\|\cdot\|$	spectral norm
$\|A\|_F$	Frobenius norm $(\sum A_{ij}^2)^{1/2}$
E	some (weight) matrix, $E \in M_n(\mathbb{R})$
f	some (weight) vector, $f \in \mathbb{R}^n$
I, I_n	identity matrix (with n rows and columns)
e	vector of all 1's, $e \in \mathbb{R}^n$
$(\mathbf{1})$	matrix of all 1's, $(\mathbf{1}) = ee^T \in M_n(\mathbb{R})$
J, J_n	permutation matrix mapping $(1,\dots,n)^T$ into $(n,\dots,1)^T$
$\sigma_{\min}(A)$	smallest singular value of A
$\lambda_{\min}(A)$	smallest eigenvalue of symmetric A

3. Normwise perturbations. Throughout this paper we let nonsingular $A \in M_n(\mathbb{R})$ be given together with $0 \neq x \in \mathbb{R}^n$. Denote $b := Ax$ and let $E \in M_n(\mathbb{R}), f \in \mathbb{R}^n$.

We first prove (2.2) in a way which is suitable for general as well as structured perturbations. The standard proof [27, Theorem 7.2] for (2.2) uses the fact that $Ax = b$ and $(A + \Delta A)(x + \Delta x) = b + \Delta b$ imply

$$\Delta x = A^{-1}(-\Delta A x + \Delta b) + \mathcal{O}(\varepsilon^2). \tag{3.1}$$

For given ΔA with $\|\Delta A\| \leq \varepsilon \|E\|$ define $\Delta b := -\frac{\|f\|}{\|E\|\,\|x\|} \Delta A x$. Then $\|\Delta b\| \leq \varepsilon \|f\|$, and (3.1) implies

$$\Delta x = -A^{-1} \Delta A x \left(1 + \frac{\|f\|}{\|E\|\,\|x\|}\right) + \mathcal{O}(\varepsilon^2). \tag{3.2}$$

This is satisfied for arbitrary ΔA with $\|\Delta A\| \leq \varepsilon \|E\|$, the perturbations ΔA being structured or unstructured. This gives a reason for the following definition.

DEFINITION 3.1. *For nonsingular $A \in M_n(\mathbb{R}), 0 \neq x \in \mathbb{R}^n$, and $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ we define*

$$\varphi^{\text{struct}}(A, x) := \sup\{\|A^{-1} \Delta A x\| : \ \Delta A \in M^{\text{struct}}, \|\Delta A\| \leq 1\}.$$

For $M_n^{\text{struct}}(\mathbb{R}) = M_n(\mathbb{R})$ we omit the superindex struct*: $\varphi(A, x)$.*

Now the special choice of Δb that led to (3.2) and the definition (2.3) imply

$$\frac{\varphi^{\text{struct}}(A, x)}{\|x\|} \left(\|E\| + \frac{\|f\|}{\|x\|}\right) \leq \kappa_{E,f}^{\text{struct}}(A, x) \tag{3.3}$$

for all $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$. Furthermore, an obvious norm estimation using (2.3) and (3.1) yields

$$\kappa_{E,f}^{\text{struct}}(A, x) \leq \|A^{-1}\|\,\|E\| + \|A^{-1}\| \frac{\|f\|}{\|x\|}, \tag{3.4}$$

again for all $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$. Therefore, we have equality in (3.4) if $\varphi^{\text{struct}}(A, x) = \|A^{-1}\|\,\|x\|$. This is true (and well known) for unstructured perturbations

$$\varphi(A, x) = \|A^{-1}\|\,\|x\| \tag{3.5}$$

by choosing orthogonal ΔA with $\Delta A x = \|x\| y$ for $\|A^{-1}\| = \|A^{-1} y\|$ and $\|y\| = 1$.

THEOREM 3.2. *For nonsingular $A \in M_n(\mathbb{R})$, $0 \neq x \in \mathbb{R}^n$, and $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ we have*

$$\frac{\varphi^{\text{struct}}(A, x)}{\|x\|} \left(\|E\| + \frac{\|f\|}{\|x\|}\right) \leq \kappa_{E,f}^{\text{struct}}(A, x) \leq \|A^{-1}\|\,\|E\| + \|A^{-1}\| \frac{\|f\|}{\|x\|}. \tag{3.6}$$

Particularly, $\varphi^{\text{struct}}(A, x) = \|A^{-1}\|\,\|x\|$ implies

$$\kappa_{E,f}^{\text{struct}}(A, x) = \kappa_{E,f}(A, x) = \|A^{-1}\|\,\|E\| + \|A^{-1}\| \frac{\|f\|}{\|x\|}.$$

As we will see, the latter equality is true for symmetric, skewsymmetric, and persymmetric perturbations. For other perturbations the lower bound in (3.6) is usually too weak because $\varphi^{\text{struct}}(A,x)$ can be much less than $\|A^{-1}\|\,\|x\|$. An immediate upper bound by (2.3) and (3.1) is

$$\kappa_{E,f}^{\text{struct}}(A,x) \le \varphi^{\text{struct}}(A,x)\frac{\|E\|}{\|x\|} + \|A^{-1}\|\frac{\|f\|}{\|x\|}. \tag{3.7}$$

Although we are free in the perturbations Δb, the structure in ΔA may not allow equality in (3.7). However, for $u,v \in \mathbb{R}^n$ it is $\max(\|u+v\|,\|u-v\|) \ge \sqrt{\|u\|^2+\|v\|^2} \ge 2^{-1/2}(\|u\|+\|v\|)$ such that

$$u,v \in \mathbb{R}^n \quad \text{implies} \quad \max(\|u+v\|,\|u-v\|) = c(\|u\|+\|v\|),$$

where $2^{-1/2} \le c \le 1$. We are free in choosing the sign of Δb, so (3.7), $u = -A^{-1}\Delta A x$, $v = A^{-1}\Delta b$ together with (3.1) imply the following result.

THEOREM 3.3. *Let $A \in M_n(\mathbb{R}), 0 \ne x \in \mathbb{R}^n$, and $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ be given. Then the structured (normwise) condition number as defined in* (2.3) *satisfies*

$$\kappa_{E,f}^{\text{struct}}(A,x) = c \cdot \left[\varphi^{\text{struct}}(A,x)\frac{\|E\|}{\|x\|} + \|A^{-1}\|\frac{\|f\|}{\|x\|}\right], \tag{3.8}$$

where $2^{-1/2} \le c \le 1$. For no perturbations in the right-hand side we have

$$\kappa_{E}^{\text{struct}}(A,x) = \varphi^{\text{struct}}(A,x)\frac{\|E\|}{\|x\|}.$$

This moves our focus from analysis of structured condition numbers to the analysis of $\varphi^{\text{struct}}(A,x)$. In the following we will use Definition 3.1 of φ^{struct} together with Theorems 3.2 and 3.3 to establish formulas and bounds for structured condition numbers.

4. Condition number for general x. For general perturbations and for the natural choice $E = A$, $f = b$, we have $\frac{\|A^{-1}\|\,\|b\|}{\|x\|} \le \|A^{-1}\|\,\|A\|$ such that (2.2) yields

$$\kappa_A(A,x) = \|A^{-1}\|\,\|A\| \le \kappa_{A,Ax}(A,x) \le 2\|A^{-1}\|\,\|A\|. \tag{4.1}$$

In other words, in case of general perturbations it does not make a big difference whether we allow perturbations in the right-hand side or leave it unchanged. Moreover, the general condition number $\kappa_A(A,x)$ is independent of x. So the condition is an inherent property of the matrix.

This may change in case of structured condition numbers. A first result in this respect is that for all structures (2.4) the *worst case* structured (normwise) condition number, i.e., the supremum over all x, is equal to the worst case unstructured condition number.

THEOREM 4.1. *Let nonsingular $A \in M_n(\mathbb{R})$ be given and $M^{\text{struct}} \subseteq M_n(\mathbb{R})$ such that one of the following conditions is satisfied:*

(i) $I \in M_n^{\text{struct}}(\mathbb{R})$.
(ii) $J \in M_n^{\text{struct}}(\mathbb{R})$.
(iii) $\tilde{J} := \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right) \in M_n^{\text{struct}}(\mathbb{R})$ *in case n even.*

Let fixed $0 < \gamma \in \mathbb{R}$ *be given. Then for all* $\|x\| = \gamma$,

$$\sup_{\|y\|=\gamma} \kappa^{\rm struct}_{E,f}(A,y) = \kappa_{E,f}(A,x) = \|A^{-1}\|\,\|E\| + \|A^{-1}\|\frac{\|f\|}{\|x\|}, \tag{4.2}$$

so the worst case structured condition number is equal to the general condition number. Equation (4.2) *is especially true for all structures in* (2.4)*. It is of course also true for* $M_n^{\rm struct}(\mathbb{R}) = M_n(\mathbb{R})$.

Remark 4.2. Note that a nonsingular skewsmmetric matrix must be of even order.

Proof. Let $\|A^{-1}\| = \|A^{-1}y\|$ with $\|y\| = 1$. Choosing $\Delta A = I$, $\Delta A = J$, or $\Delta A = \tilde{J}$ in case (i), (ii), or (iii), respectively, observing $\Delta A^2 = \pm I$, and setting $x := \gamma \Delta A y$ imply $A^{-1}\Delta A x = \pm\gamma A^{-1}y$ and $\|x\| = \gamma$. Hence Definition 3.1 yields $\|A^{-1}\|\,\|x\| \geq \varphi^{\rm struct}(A,x) \geq \|A^{-1}\|\,\|x\|$ for that choice of x, and Theorem 3.2 finishes the proof. ∎

For specific x things may change significantly, at least if the structure imposes severe restrictions on ΔA. For symmetric, persymmetric, and skewsymmetric structures this is not yet the case.

5. Symmetric, persymmetric, and skewsymmetric perturbations. In the following we will show that those perturbations do not change the condition number at all. For symmetric perturbations this was already observed in [24]; see also [8]. In other words, "worst" perturbations may be chosen in the set $M_n^{\rm sym}(\mathbb{R})$, $M_n^{\rm persym}(\mathbb{R})$, or $M_n^{\rm skewsym}(\mathbb{R})$. We prove this by investigating our key to structured perturbations, the function $\varphi^{\rm struct}$. We first prove a lemma which will be of later use. For the symmetric case this was observed in [8].

LEMMA 5.1. *Let* $x, y \in \mathbb{R}^n$ *be given with* $\|x\| = \|y\| = 1$ *and let* struct $\in$ {sym, persym}. *Then there exists* $A \in M_n^{\rm struct}(\mathbb{R})$ *with*

$$y = Ax \quad \textit{and} \quad \|A\| = 1. \tag{5.1}$$

If, in addition, $y^T x = 0$, *then there exists* $A \in M_n^{\rm skewsym}(\mathbb{R})$ *with* (5.1).

Proof. For symmetric structure the Householder reflection H along $x + y$ satisfies $H = H^T$, $\|H\| = 1$, and $Hx = y$. A matrix B is persymmetric iff $B = JB^TJ$. Let H be the Householder reflection along $x + Jy$ and set $A := JH$. Then $A = JA^TJ$ is persymmetric, $\|A\| = 1$, and $Ax = JHx = J \cdot Jy = y$.

For skewsymmetric structure and x, y orthonormal there is orthogonal $Q \in M_n(\mathbb{R})$ with $[x|y] = Q[e_1| - e_2]$, e_i denoting the ith column of the identity matrix. Define $D := \operatorname{diag}\left(\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right), 0, \ldots, 0\right)$ and $A := QDQ^T$. Then $A = -A^T$, $\|A\| = 1$, and $Ax = QDe_1 = -Qe_2 = y$. ∎

LEMMA 5.2. *Let nonsingular* $A \in M_n(\mathbb{R})$ *and* $0 \neq x \in \mathbb{R}^n$ *be given. Then*

$$\varphi^{\rm struct}(A,x) = \varphi(A,x) = \|A^{-1}\|\,\|x\| \tag{5.2}$$

for struct $\in$ {sym, persym}. *Relation* (5.2) *is also true for* struct = skewsym *and* $A \in M_n^{\rm skewsym}$.

Proof. By Definition 3.1 and (3.5), $\varphi^{\rm struct}(A,x) \leq \varphi(A,x) = \|A^{-1}\|\,\|x\|$, so it remains to show $\varphi^{\rm struct}(A,x) \geq \|A^{-1}\|\,\|x\|$. Without loss of generality, assume $\|x\| = 1$ and let $\|A^{-1}\| = \|A^{-1}y\|$ for $\|y\| = 1$. It suffices to find $\Delta A \in M^{\rm struct}$ with $\|\Delta A\| \leq 1$ and $\Delta A x = y$. This is exactly the content of Lemma 5.1 for struct $\in$ {sym, persym}.

For skewsymmetric structure suppose $A \in M^{\mathrm{skewsym}}$. Eigenvalues of A are conjugate purely imaginary, and nonsingularity of A implies that n is even, and also implies that all singular values are of even multiplicity. That means there are orthogonal $y_1, y_2 \in \mathbb{R}^n$ with $\|y_1\| = \|y_2\| = 1$ and $\|A^{-1}y_1\| = \|A^{-1}y_2\| = \|A^{-1}\|$. Choose $y \in \mathrm{span}\{y_1, y_2\}$ with $x^T y = 0$ and $\|y\| = 1$. By construction, $\|A^{-1}y\| = \|A^{-1}\|$, and Lemma 5.1 finishes the proof. ∎

Together with Theorem 3.2 this proves the following.

THEOREM 5.3. *Let nonsingular $A \in M_n(\mathbb{R})$ and $0 \neq x \in \mathbb{R}^n$ be given. For* struct $\in \{\mathrm{sym}, \mathrm{persym}, \mathrm{skewsym}\}$ *we have*

$$\kappa^{\mathrm{struct}}_{E,f}(A, x) = \kappa_{E,f}(A, x),$$

where in case struct = skewsym *we suppose additionally $A \in M_n^{\mathrm{skewsym}}(\mathbb{R})$.*

The result was observed for symmetric structures in [24, 23]. As we will see, this nice fact is no longer true for the other structures. In fact, there may be quite a factor between κ^{struct} and κ.

6. Exploring the structure. Before we proceed we collect some general observations on structured condition numbers. To establish bounds for the ratio $\kappa^{\mathrm{struct}}/\kappa$ we need a relation between $\|E\|$ and $\|f\|$. Therefore we especially investigate the natural choice $E = A$ and $f = b$. The first statement is a useful lower bound.

LEMMA 6.1. *Let nonsingular $A \in M_n(\mathbb{R})$, $0 \neq x \in \mathbb{R}^n$, and some $M_n^{\mathrm{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ be given. Suppose*

$$\varphi^{\mathrm{struct}}(A, x) \geq \omega \|A^{-T}x\| \tag{6.1}$$

for $0 \leq \omega \in \mathbb{R}$. Then

$$\kappa^{\mathrm{struct}}_{A,Ax}(A, x) \geq \sqrt{\frac{\omega}{2}\|A^{-1}\|\,\|A\|}.$$

Proof. Without loss of generality assume $\|x\| = 1$. Then

$$1 = x^T A^{-1} A x \leq \|x^T A^{-1}\|\,\|Ax\| = \|A^{-T}x\|\,\|Ax\|. \tag{6.2}$$

In view of (3.8) for $E = A,\ f = b,\ \|x\| = 1$, and $Ax = b$, we are finished if we can show

$$\varphi^{\mathrm{struct}}(A, x)\|A\| + \|A^{-1}\|\,\|Ax\| \geq \sqrt{\omega\|A^{-1}\|\,\|A\|}.$$

This is true if $\|Ax\| \geq \sqrt{\omega\|A\|/\|A^{-1}\|}$. On the contrary, (6.2) yields $\|A^{-T}x\| \geq \|Ax\|^{-1} > \sqrt{\omega^{-1}\|A^{-1}\|/\|A\|}$, and combining this with (6.1) finishes the proof. ∎

The symmetric Toeplitz matrices are related to persymmetric Hankel matrices by

$$T \in M_n^{\mathrm{symToep}} \quad \Leftrightarrow \quad JT \in M_n^{\mathrm{persymHankel}} \quad \Leftrightarrow \quad TJ \in M_n^{\mathrm{persymHankel}}. \tag{6.3}$$

Similarly, (general) Toeplitz matrices are related to general Hankel matrices by

$$T \in M_n^{\mathrm{Toep}} \quad \Leftrightarrow \quad JT \in M_n^{\mathrm{Hankel}} \quad \Leftrightarrow \quad TJ \in M_n^{\mathrm{Hankel}}. \tag{6.4}$$

By rewriting (3.1) into

$$\Delta x = (JA)^{-1}(-J\Delta A x + J\Delta b) + \mathcal{O}(\varepsilon^2)$$

and

$$J\Delta x = (AJ)^{-1}(-\Delta AJ \cdot Jx + \Delta b) + \mathcal{O}(\varepsilon^2)$$

and observing $\|J\Delta A\| = \|\Delta AJ\| = \|\Delta A\|$ and $\|J\Delta b\| = \|\Delta b\|$, definition (2.3) yields the following.

THEOREM 6.2. *For nonsingular $A \in M_n(\mathbb{R})$ and $0 \neq x \in \mathbb{R}^n$ we have*

$$\kappa_{E,f}^{\text{symToep}}(A,x) = \kappa_{E,f}^{\text{persymHankel}}(JA,x) = \kappa_{E,f}^{\text{persymHankel}}(AJ,Jx)$$

and

$$\kappa_{E,f}^{\text{Toep}}(A,x) = \kappa_{E,f}^{\text{Hankel}}(JA,x) = \kappa_{E,f}^{\text{Hankel}}(AJ,Jx).$$

Therefore we will concentrate in the following on symmetric Toeplitz and Hankel structures. Every result for those is valid mutatis mutandis for persymmetric Hankel and general Toeplitz structures, respectively.

To further explore the structure we derive two-sided explicit bounds for $\varphi^{\text{struct}}(A,x)$. For linear structures in the matrix entries of $A \in M_n(\mathbb{R})$, every A_{ij} depends linearly on some k parameters. Denote by $\text{vec}(A) = (A_{11},\ldots,A_{1n},\ldots,A_{n1},\ldots,A_{nn})^T \in \mathbb{R}^{n^2}$ the vector of stacked columns of A. Then for every dimension there is some fixed structure matrix $\Phi^{\text{struct}} \in M_{n^2,k}(\mathbb{R})$ such that

$$A \in M_n^{\text{struct}}(\mathbb{R}) \quad \Leftrightarrow \quad \exists\, p \in \mathbb{R}^k :\ \text{vec}(A) = \Phi^{\text{struct}} \cdot p. \tag{6.5}$$

This idea was developed in [24]. For our structures (2.4) the number of independent parameters k is as shown in Table 6.1.

TABLE 6.1
Number of independent parameters.

Structure	sym	persym	skewsym	circ	symToep	Toep	Hankel	persymHankel
k	$(n^2+n)/2$	$(n^2+n)/2$	$(n^2-n)/2$	n	n	$2n-1$	$2n-1$	n

For the structures in (2.4) the structure matrix Φ^{struct} is sparse with entries $0/1$ except for skewsymmetric matrices with entries $0/+1/-1$. We can make Φ^{struct} unique by defining the parameter vector "columnwise"; i.e., $p \in \mathbb{R}^k$ is the unique vector of the first k independent components in $\text{vec}(A)$.

It is important to note that Φ^{struct} defines for every dimension n a one-to-one mapping between $\mathbb{R}^k$ and $M_n^{\text{struct}}(\mathbb{R})$. To compute bounds on φ^{struct} we relate the matrix norm $\|A\|_2$ to the vector norm $\|p\|_2$.

LEMMA 6.3. *Let $A \in M_n^{\text{struct}}(\mathbb{R})$ and $p \in \mathbb{R}^k$ be given such that* $\text{vec}(A) = \Phi^{\text{struct}} p$. *Then*

$$\alpha\|A\| \leq \|p\| \leq \beta\|A\| \tag{6.6}$$

with constants α, β according to the following table:

Structure	α	β
circ	$1/\sqrt{n}$	1
symToep	$1/\sqrt{2n-2}$	1
Toep	$1/\sqrt{n}$	$\sqrt{2}$
Hankel	$1/\sqrt{n}$	$\sqrt{2}$
persymHankel	$1/\sqrt{2n-2}$	1

All upper bounds and the lower bound for circulants are sharp, and the other lower bounds are sharp up to a factor $\sqrt{2}$.

Proof. For $A \in M_n^{\mathrm{circ}}$ we have

$$\|p\| = \|Ae_1\| \le \|A\| \le \|A\|_F = \Big(n \sum p_i^2\Big)^{1/2} = \sqrt{n}\|p\|.$$

The left and right estimations are sharp for $A = I$ and $A = (\mathbf{1})$, respectively. For $A \in M_n^{\mathrm{symToep}}$,

$$\|p\| = \|Ae_1\| \le \|A\| \le \|A\|_F \le \Big((2n-2) \sum p_i^2\Big)^{1/2} = \sqrt{2n-2}\|p\|.$$

For $A = I$ it is $\|A\| = \|p\| = 1$, and for $A = (\mathbf{1})$ it is $\|A\| = \sqrt{n}\|p\| = n$. For $A \in M_n^{\mathrm{Hankel}}$ we have

$$\|p\|^2 \le 2\max(\|Ae_1\|^2, \|e_1^T A\|^2) \le 2\|A\|^2$$

and

$$\|A\| \le \|A\|_F \le \Big(n \sum p_i^2\Big)^{1/2} = \sqrt{n}\|p\|.$$

For $A = (\mathbf{1})$ it is $\|A\| = n = \frac{n}{\sqrt{2n-1}}\|p\|$, and for the Hankel matrix with $A_{11} = A_{nn} = 1$ and zero entries elsewhere it is $\|p\| = \sqrt{2} = \sqrt{2}\|A\|$. The other estimations follow by (6.3) and (6.4). ∎

The bounds for circulants are noted for completeness; we will derive better methods to estimate $\kappa_{E,f}^{\mathrm{circ}}$ in the next section. The difficulty in estimating $\varphi^{\mathrm{struct}} = \sup\{\|A^{-1}\Delta A x\| : \ \Delta A \in M^{\mathrm{struct}}, \ \|\Delta A\| \le 1\}$ is that the supremum is taken only over structured matrices ΔA. With Lemma 6.3 this can be rewritten to the supremum over *all* parameter vectors $\Delta p \in \mathbb{R}^k$, $\|\Delta p\| \le \mathrm{const}$, where k is the number of independent parameters according to Table 6.1 and const follows by Lemma 6.3. We have

$$\begin{aligned} &\{\Delta A \in M_n(\mathbb{R}) : \ \mathrm{vec}(\Delta A) = \Phi^{\mathrm{struct}} \Delta p, \quad \Delta p \in \mathbb{R}^k, \|\Delta p\| \le \alpha\} \\ \subseteq\ &\{\Delta A \in M_n^{\mathrm{struct}}(\mathbb{R}) : \ \|\Delta A\| \le 1\} \\ \subseteq\ &\{\Delta A \in M_n(\mathbb{R}) : \ \mathrm{vec}(\Delta A) = \Phi^{\mathrm{struct}} \Delta p, \quad \Delta p \in \mathbb{R}^k, \|\Delta p\| \le \beta\}, \end{aligned} \tag{6.7}$$

where Δp varies freely in a norm ball of the $\mathbb{R}^k$. So (6.7) is the key to obtaining computable lower and upper bounds for the structured condition number, the bounds not being far apart.

To estimate $\varphi^{\mathrm{struct}}(A, x)$ we use the following ansatz as in [24]. Note that $\Delta A \cdot x = (x^T \otimes I)\ \mathrm{vec}(\Delta A)$, $\otimes$ denoting the Kronecker product. For $\mathrm{vec}(\Delta A) = \Phi^{\mathrm{struct}} \Delta p$ this implies

$$\Delta A \cdot x = (x^T \otimes I)\Phi^{\mathrm{struct}} \cdot \Delta p. \tag{6.8}$$

The matrix $(x^T \otimes I)\Phi^{\mathrm{struct}} \in M_{n,k}(\mathbb{R})$ depends only on x for every dimension. This leads us to the definition

$$\Psi_x^{\mathrm{struct}} := (x^T \otimes I)\Phi^{\mathrm{struct}} \in M_{n,k}(\mathbb{R}), \tag{6.9}$$

the dimension k as in Table 6.1. This definition holds for every linear structure. For the structures in (2.4), the matrices $\Psi_x^{\rm struct}$ can be calculated explicitly. For example, for the Hankel matrix

$$H = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{pmatrix}$$

we have $\Phi^{\rm Hankel} \in M_{n^2,k} = M_{9,5}$, a column block matrix with n blocks $\Phi_i \in M_{n,k}, 1 \le i \le n$, and

$$\Phi_i = \begin{pmatrix} \underbrace{\begin{matrix} 0 & \dots & 0 \\ & \dots & \\ 0 & \dots & 0 \end{matrix}}_{i-1} & & & & \underbrace{\begin{matrix} 0 & \dots & 0 \\ & \dots & \\ 0 & \dots & 0 \end{matrix}}_{n-i} \end{pmatrix} \begin{matrix} \\ \\ 1 \\ \ddots \\ \\ 1 \\ \\ \end{matrix} \in M_{3,5},$$

so that $\Psi_x^{\rm struct} = (x^T \otimes I)\Phi^{\rm struct}$ implies

$$\Psi_x^{\rm Hankel} = \sum x_i \Phi_i = \begin{pmatrix} x_1 & x_2 & x_3 & & \\ & x_1 & x_2 & x_3 & \\ & & x_1 & x_2 & x_3 \end{pmatrix} \in M_{n,k}. \tag{6.10}$$

We mention

$$\begin{aligned} &\Psi_x^{\rm circ} = {\rm circ}(x^T)^T, \\ &\Psi_x^{\rm symToep} = T(x) + JT(Jx) - xe_1^T, \\ &\Psi_x^{\rm Hankel} = {\rm Toeplitz}([x_1, z], [x^T, z]), \end{aligned} \tag{6.11}$$

where $z = {\rm zeros}(1, n-1)$ and $T(x) := {\rm Toeplitz}(x, [x_1, z])$ in Matlab [32] notation; that is, ${\rm Toeplitz}(c, r)$ denotes the Toeplitz matrix with first column c and first row r. With this we have explicit bounds for $\varphi^{\rm struct}$.

LEMMA 6.4. *Let nonsingular $A \in M_n(\mathbb{R})$ and $0 \ne x \in \mathbb{R}^n$ be given. Let* struct *be one of the structures mentioned in Lemma* 6.3*. Then*

$$\varphi^{\rm struct}(A, x) = \gamma \|A^{-1}\Psi_x^{\rm struct}\|,$$

where $\alpha \le \gamma \le \beta$ and α, β as in Lemma 6.3*.*

Proof. Combining (6.9), (6.8), and (6.7) with Definition 3.1 yields

$$\begin{aligned} \alpha\|A^{-1}\Psi_x^{\rm struct}\| &= \sup\{\|A^{-1}\Psi_x^{\rm struct}\Delta p\| : \ \Delta p \in \mathbb{R}^k, \|\Delta p\| \le \alpha\} \\ &\le \sup\{\|A^{-1}\Delta A x\| : \ \Delta A \in M_n^{\rm struct}(\mathbb{R}), \|\Delta A\| \le 1\} \\ &= \varphi^{\rm struct}(A, x) \\ &\le \sup\{\|A^{-1}\Psi_x^{\rm struct}\Delta p\| : \ \Delta p \in \mathbb{R}^k, \|\Delta p\| \le \beta\} \\ &= \beta\|A^{-1}\Psi_x^{\rm struct}\|. \qquad \square \end{aligned}$$

Combining this with Theorem 3.3 yields computable bounds for the structured condition number.

THEOREM 6.5. *Let nonsingular $A \in M_n(\mathbb{R})$ and $0 \ne x \in \mathbb{R}^n$ be given. Let* struct *be one of the structures mentioned in Lemma* 6.3*. Then*

$$\kappa_{E,f}^{\rm struct}(A, x) = c\frac{\gamma\|A^{-1}\Psi_x^{\rm struct}\| \, \|E\| + \|A^{-1}\| \, \|f\|}{\|x\|}, \tag{6.12}$$

where $2^{-1/2} \le c \le 1$ *and* $\alpha \le \gamma \le \beta$ *for* α, β *as in Lemma* 6.3. *In case of no perturbations in the right-hand side,*

$$\kappa_E^{\text{struct}}(A,x) = \gamma \frac{\|A^{-1}\Psi_x^{\text{struct}}\|}{\|x\|}\,\|E\|. \tag{6.13}$$

This implies the following remarkable property of the ratio between the structured and unstructured condition numbers.

COROLLARY 6.6. *Let nonsingular* $A \in M_n(\mathbb{R})$ *and* $0 \neq x \in \mathbb{R}^n$ *be given. Let* struct *be one of the structures mentioned in Lemma* 6.3. *Then*

$$\frac{\kappa_{E,f}^{\text{struct}}(A,x)}{\kappa_{E,f}(A,x)} \ge 2^{-1/2} \frac{\alpha\|A^{-1}\|\frac{\sigma_{\min}(\Psi_x^{\text{struct}})}{\|x\|}\,\|E\| + \|A^{-1}\|\,\frac{\|f\|}{\|x\|}}{\|A^{-1}\|\,\|E\| + \|A^{-1}\|\,\frac{\|f\|}{\|x\|}}, \tag{6.14}$$

for α *as in Lemma* 6.3. *Moreover, for no perturbations in the right-hand side,*

$$2^{1/2}\frac{\kappa_{E,f}^{\text{struct}}(A,x)}{\kappa_{E,f}(A,x)} \ge \frac{\kappa_E^{\text{struct}}(A,x)}{\kappa_E(A,x)} \ge \alpha\frac{\sigma_{\min}(\Psi_x^{\text{struct}})}{\|x\|}. \tag{6.15}$$

Proof. We have $\Psi_x^{\text{struct}} \in M_{n,k}(\mathbb{R})$ with $k \ge n$; therefore $\|A^{-1}\Psi_x^{\text{struct}}\| \ge \|A^{-1}\|\sigma_{\min}(\Psi_x^{\text{struct}})$. Now (2.2) and Theorem 6.5 finish the proof. ∎

This result allows us to estimate the minimum ratio of $\kappa^{\text{struct}}/\kappa$ *independent* of the matrix A only by examining the smallest singular value of Ψ_x^{struct}, where the latter can be computed, for example, by (6.11). So we have the surprising result that a small ratio $\kappa^{\text{struct}}/\kappa$ is only possible for certain *solutions* x, independent of the (structured) matrix. It also shows that for fixed x an arbitrarily small ratio of $\kappa^{\text{struct}}/\kappa$ is only possible if $\text{rank}(\Psi_x^{\text{struct}}) < n$. From a practical point of view this means that standard unstructured perturbation analysis suffices at least for all cases where $\sigma_{\min}(\Psi_x^{\text{struct}})$ is not too small.

The statistics in Table 6.2 show how often a small ratio $\kappa^{\text{struct}}/\kappa$ *can* occur. Note that this is a lower estimate of the ratio for all matrices A; it need not be attained for a specific matrix A. Table 6.2 shows the minimum and median of $\tau(x) := \sigma_{\min}(\Psi_x^{\text{struct}})/\|x\|$ for some 10^4 samples of x with entries uniformly distributed within $[-1,1]$. Also note that, in order to obtain the lower estimate for $\kappa^{\text{struct}}/\kappa$, by (6.15) the displayed numbers have to be multiplied by α according to the table in Lemma 6.3.

TABLE 6.2
Minimum value and median of $\tau(x) = \sigma_{\min}(\Psi_x^{\text{struct}})/\|x\|$.

	Symmetric Toeplitz		Circulant	
n	$\min(\tau(x))$	$\text{median}(\tau(x))$	$\min(\tau(x))$	$\text{median}(\tau(x))$
10	$2.4 \cdot 10^{-7}$	$4.9 \cdot 10^{-2}$	$1.3 \cdot 10^{-5}$	$2.6 \cdot 10^{-1}$
20	$5.0 \cdot 10^{-7}$	$2.2 \cdot 10^{-2}$	$3.9 \cdot 10^{-5}$	$2.0 \cdot 10^{-1}$
50	$1.1 \cdot 10^{-6}$	$8.8 \cdot 10^{-3}$	$3.1 \cdot 10^{-5}$	$1.4 \cdot 10^{-1}$
100	$1.5 \cdot 10^{-6}$	$4.3 \cdot 10^{-3}$	$7.5 \cdot 10^{-5}$	$1.0 \cdot 10^{-1}$

Table 6.2 shows that small ratios are possible but seem to be rare. We mention that rank-deficient Ψ_x^{struct} is possible, for example, for $x = (1,\ldots,1)^T$ and $n \ge 2$. That means that for this solution vector x the ratio $\kappa^{\text{struct}}/\kappa$ may become arbitrarily small. This is indeed the case, as we will see in the following sections. However, it changes for Hankel structures, as we will show in section 10.

Explicit computation of (6.12) is possible in $\mathcal{O}(n^3)$ flops. However, the computationally intensive part $\|A^{-1}\Psi_x^{\rm struct}\|$ can be estimated in some $\mathcal{O}(n^2)$ flops using well-known procedures for condition estimation as by [20]; see also [26].

The concept of $\Phi^{\rm struct}$ and $\Psi_x^{\rm struct}$ applies to all linear structures. Before we proceed, we give in the next section some examples of structures other than those in (2.4).

7. Some special structures. The concept of $\Phi^{\rm struct}$ and $\Psi_x^{\rm struct}$ especially can be used to calculate the structured condition number in case some elements of A remain unchanged, although we treat normwise distances to the matrix A. Typical examples are symmetric tridiagonal or general lower triangular matrices. In either case it is straightforward to calculate the corresponding $\Phi^{\rm struct}$, which is fixed for every dimension. Based on that, $\Psi_x^{\rm struct}$ is computed by (6.9) and, with constants α and β relating $\|A\|$ and $\|p\|$ as in Lemma 6.3, $\kappa_{E,f}^{\rm struct}(A,x)$ can be estimated by Theorem 6.5. Using this we calculated the condition numbers in (2.6) and (2.7). In the following we give some examples of tridiagonal structures.

Let a symmetric tridiagonal Toeplitz matrix A with diagonal element d and super- and subdiagonal element c be given. Then the eigenvalues of A are explicitly known [27, section 28.5] to be $\lambda_k(A) = d+2c\cos\frac{k\pi}{n+1}$ for $1 \le k \le n$, so $\|A\| = |d|+2|c|\cos\frac{\pi}{n+1}$. Furthermore, according to (6.5), $\mathrm{vec}(A) = \Phi^{\rm symtridiagToep}p$ for $p = (d,c)^T \in \mathbb{R}^2$, and a computation according to (6.9) yields

$$\Psi_x^{\rm symtridiagToep} = \begin{pmatrix} x_1 & x_2 \\ x_2 & x_1+x_3 \\ x_3 & x_2+x_4 \\ & \cdots \\ x_{n-1} & x_{n-2}+x_n \\ x_n & x_{n-1} \end{pmatrix}. \tag{7.1}$$

For $n \ge 2$ it follows that

$$\|A\| \ge |d| + 2|c|\cos\frac{\pi}{3} = |d| + |c| \ge \sqrt{d^2+c^2} = \|p\|$$

and

$$\frac{\|A\|}{\|p\|} \le \frac{|d|+2|c|}{\sqrt{d^2+c^2}} \le \max_{0\le x,y\le 1} \frac{x+2y}{\sqrt{x^2+y^2}} =: \beta.$$

A computation yields $\beta = \sqrt{5}$, so

$$\|p\| \le \|A\| \le \sqrt{5}\|p\| \quad \text{for} \quad A \in M_n^{\rm symtridiagToep}(\mathbb{R}).$$

Both estimations are sharp for $A = I$ and $c = 2$, $d = 1$, respectively. In the latter case $\|A\| \to 5$ as $n \to \infty$, whereas $\|p\| = \sqrt{5}$.

The explicit representation (7.1) for Ψ_x also shows that for specific solution vector x there is a big difference between the structured and unstructured condition numbers. Suppose n is divisible by 3 and let $x = (z, -z, z, -z, \ldots, \pm z)^T$ for $z = (\alpha, \alpha, 0)^T$, $\alpha \in \mathbb{R}$. A computation shows $Ax = (c+d)x$. Moreover, the second column of Ψ_x is equal to the first, so

$$A^{-1}\Psi_x = (A^{-1}x, A^{-1}x) = (c+d)^{-1}(x,x).$$

Therefore Theorem 6.5 implies

$$|(c+d)^{-1}| \le \kappa_A^{\text{symtridiagToep}}(A,x) \le \sqrt{10}|(c+d)^{-1}|.$$

Note that this is true for every x of the structure as defined above. For the matrix as in (2.5) this means

$$1 \le \kappa_A^{\text{symtridiagToep}}(A,x) \le \sqrt{10}$$

for every x as above, whereas, for $d+2c=0$,

$$\kappa_A(A,x) = \|A^{-1}\|\,\|A\| \sim n^2.$$

For a general tridiagonal Toeplitz matrix A with diagonal element d and off-diagonal elements c, e we have $\|p\| = \sqrt{c^2+d^2+e^2}$. Furthermore, $\|Ax\| \le (|c|+|d|+|e|)\|x\|$ for every $x \in \mathbb{R}^n$ and therefore

$$\frac{\|A\|}{\|p\|} \le \frac{|c|+|d|+|e|}{\sqrt{c^2+d^2+e^2}} \le \sqrt{3}.$$

The estimation is asymptotically sharp for $c=d=e=1$.

For x being the second column of the identity matrix and $n \ge 3$ it follows that

$$\frac{\|A\|}{\|p\|} \ge \frac{\sqrt{c^2+d^2+e^2}}{\|p\|} = 1.$$

This estimation is sharp for $A=I$. For $n=2$ it is $A = \left(\begin{smallmatrix} d & e \\ c & d \end{smallmatrix}\right)$ and $\|A\|/\|p\| \ge 1/\sqrt{2}$. This estimation is sharp for $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.

For symmetric tridiagonal A with diagonal elements d_ν and off-diagonal elements c_ν, we have $\|p\| = \sqrt{\|c\|^2+\|d\|^2}$ and $\|A\|_F = \sqrt{2\|c\|^2+\|d\|^2}$. This implies

$$\frac{\|A\|}{\|p\|} \ge \frac{1}{\sqrt{n}}\frac{\|A\|_F}{\|p\|} \ge \frac{1}{\sqrt{n}}$$

and

$$\frac{\|A\|}{\|p\|} \le \max_{0\le x,y\le 1} \frac{\sqrt{2x^2+y^2}}{\sqrt{x^2+y^2}} \le \sqrt{2}.$$

The first estimation is sharp for $A=I$, the second up to a small factor. Finally, for general tridiagonal A we have $\|A\|_F = \|p\|$, so

$$\frac{1}{\sqrt{n}}\|p\| \le \|A\| \le \|p\|.$$

The estimations are sharp for $A=I$ and the matrix with $A_{11}=1$ and $A_{ij}=0$ elsewhere, respectively.

Summarizing, we have the following result.

THEOREM 7.1. *Let $A \in M_n^{\text{struct}}(\mathbb{R})$ and $p \in \mathbb{R}^k$ be given such that $\text{vec}(A) = \Phi^{\text{struct}} p$. Then*

$$\alpha\|p\| \le \|A\| \le \beta\|p\|$$

with constants α, β according to the following table:

Structure	α			β
symtridiagToep	1			$\sqrt{5}$
tridiagToep	1 $1/\sqrt{2}$	*for* *for*	$n \neq 2$ $n = 2$	$\sqrt{3}$
symtridiag	$1/\sqrt{n}$			1
tridiag	$1/\sqrt{n}$			1

All lower bounds are sharp; all upper bounds are sharp up to a small constant factor.

Using the constants α, β and Theorem 6.5 the structured condition numbers are easily calculated.

Also, linear structures in the right-hand side can be treated by an augmented linear system of dimension $n+1$. Such structures appear, for example, in the Yule–Walker problem [18, section 4.7.2].

But more can be said, especially about $\kappa^{\text{struct}}/\kappa$. Things are particularly elegant for circulant matrices.

8. Circulant matrices. Circulant matrices are of the form

$$C = \begin{pmatrix} c_0 & c_1 & c_2 & c_3 \\ c_3 & c_0 & c_1 & c_2 \\ c_2 & c_3 & c_0 & c_1 \\ c_1 & c_2 & c_3 & c_0 \end{pmatrix}$$

and do have a number of remarkable properties [9]. Denote by P the permutation matrix mapping $(1, \ldots, n)^T$ into $(2, \ldots, n, 1)^T$. Then a circulant can be written as

$$C = \text{circ}(c_0, \ldots, c_{n-1}) = \sum_{\nu=0}^{n-1} c_\nu P^\nu \in M_n^{\text{circ}}.$$

From this polynomial representation it follows that circulants commute. Therefore, for $A \in M_n^{\text{circ}}$, Definition 3.1 implies

$$\varphi^{\text{circ}}(A, x) = \sup\{\|\Delta A \cdot A^{-1}x\| : \ \Delta A \in M_n^{\text{circ}}, \|\Delta A\| \leq 1\} \leq \|A^{-1}x\|,$$

and observing $\Delta A := I \in M_n^{\text{circ}}$ it follows that

$$\varphi^{\text{circ}}(A, x) = \|A^{-1}x\|. \tag{8.1}$$

THEOREM 8.1. *Let a nonsingular circulant $A \in M_n^{\text{circ}}(\mathbb{R})$ and $0 \neq x \in \mathbb{R}^n$ be given. Then*

$$\kappa_{E,f}^{\text{circ}}(A, x) = c\frac{\|A^{-1}x\| \, \|E\| + \|A^{-1}\| \, \|f\|}{\|x\|} \tag{8.2}$$

with $2^{-1/2} \leq c \leq 1$. In particular, for no perturbations in the right-hand side we have

$$\kappa_E^{\text{circ}}(A, x) = \frac{\|A^{-1}x\| \, \|E\|}{\|x\|} \tag{8.3}$$

and

$$\frac{\kappa_E^{\text{circ}}(A, x)}{\kappa_E(A, x)} = \frac{\|A^{-1}x\|}{\|A^{-1}\| \, \|x\|} \geq \frac{1}{\|A^{-1}\| \, \|A\|}. \tag{8.4}$$

The inequality is sharp.

Proof. The assertions follow by Theorem 3.3 and (8.1), where the last inequality stems from $\|x\| \leq \|A\| \, \|A^{-1}x\|$. Choosing x such that $\|A^{-1}x\| = \sigma_{\min}(A^{-1})\|x\| = \|A\|^{-1}\|x\|$ finishes the proof. ∎

So for no perturbations in the right-hand side we have $\kappa_A^{\text{circ}}(A, x) = 1$ for every circulant A and any x chosen such that $\|A^{-1}x\| = \sigma_{\min}(A^{-1})\|x\| = \|A\|^{-1}\|x\|$. Note that $\kappa_A(A, x) = \|A^{-1}\| \, \|A\|$ for every x. Also note that the ratio in (8.4) applies to general weight matrices E.

These are, however, extreme cases. Formula (8.2) also shows that, in general, $\kappa_{E,f}^{\text{circ}}$ and $\kappa_{E,f}$ are not too far apart because, in general, the same is true for $\|A^{-1}x\|$ and $\|A^{-1}\| \, \|x\|$.

To analyze the ratio $\kappa^{\text{circ}}/\kappa$ including perturbations in the right-hand side we need again a relation between $\|E\|$ and $\|f\|$. Therefore we switch to the natural choice $E = A$ and $f = b$. Furthermore, we need more details on circulants.

Every circulant is diagonalized by the scaled Fourier matrix $F \in M_n(\mathbb{C})$, $F_{ij} = \omega^{(i-1)(j-1)}/\sqrt{n}$, for ω denoting the nth root of unity [9]. Note that F is unitary and symmetric. So every circulant C is represented by $C = F^H D F$ for some diagonal $D \in M_n(\mathbb{C})$. We need some auxiliary results which will also be useful for Hankel matrices.

LEMMA 8.2. *Let $A \in M_n(\mathbb{C})$, $z \in \mathbb{C}^n$, and a circulant $C \in M_n^{\text{circ}}(\mathbb{C})$ be given. Then*

$$\|AC\| = \|AC^H\| \quad \textit{and} \quad \|Cz\| = \|C^H z\|.$$

Proof. Let $C = F^H D F$ for diagonal $D \in M_n(\mathbb{C})$. There is diagonal $S \in M_n(\mathbb{C})$ with $|S_{ii}| = 1$ for all i and $D = SD^H = D^H S$. Since F and S are unitary we obtain

$$\begin{aligned} \|AC^H\| &= \|AF^H D^H F\| = \|AF^H D^H\| = \|AF^H D^H S\| = \|AF^H D\| \\ &= \|AF^H DF\| = \|AC\| \end{aligned}$$

and

$$\|C^H z\| = \|F^H D^H F z\| = \|D^H F z\| = \|S^H DF z\| = \|F^H DF z\| = \|Cz\|. \qquad ∎$$

The next lemma characterizes real circulants. This result is definitely known; however, the only reference we found contains typos and is without proof. So we repeat the short proof.

LEMMA 8.3. *Every circulant C is equal to $F^H DF$ for (complex) diagonal D. Let P denote the permutation matrix mapping $(1, \ldots, n)^T$ into $(1, n, \ldots, 2)^T$. Then C is real iff $D = PD^H P$.*

Proof. The matrix C is real iff it is equal to its conjugate $\overline{C}$. Now the definitions of F and $F = F^T$ imply $\overline{F} = F^H$ and $F^H = PF = FP$; we get the latter equality because F, F^H, and P, are symmetric. Hence

$$\overline{C} = FD^H F^H = F^H \cdot PD^H P \cdot F$$

proves the assertion. ∎

For $A = F^H DF \in M_n^{\text{circ}}(\mathbb{R})$ being a circulant, $A^{-1} = F^H D^{-1} F$ is a circulant as well, so (8.1) and Lemma 8.2 show

$$\varphi^{\text{circ}}(A, x) = \|A^{-1}x\| = \|A^{-T}x\|.$$

Combining this with Lemma 6.1 yields

$$\kappa^{\rm circ}_{A,Ax}(A,x) \geq 2^{-1/2}\sqrt{\|A^{-1}\|\,\|A\|},$$

and (4.1) implies

$$\frac{\kappa^{\rm circ}_{A,Ax}(A,x)}{\kappa_{A,Ax}(A,x)} \geq \frac{1}{2\sqrt{2}\cdot\sqrt{\|A^{-1}\|\,\|A\|}}.$$

We give an explicit $n\times n$ example, $n\geq 5$, showing that this inequality is sharp up to a small constant factor. For $m\geq 0$ and $0<\varepsilon<1$ define

$$A = F^H\,{\rm diag}(1,v,\varepsilon,\varepsilon^{-1},[1,]\varepsilon^{-1},\varepsilon,v)F =: F^H D F, \tag{8.5}$$

where v denotes a row vector of m ones and $[1,]$ indicates that this diagonal element 1 may be left out. Accordingly, A is a circulant of dimension $n=2m+5$ or $n=2m+6$, depending on whether the diagonal element 1 is left out or not. In either case A is real by Lemma 8.3. The eigenvalues of A are the D_{ii} with corresponding columns of F^H as eigenvectors. Particularly, e is an eigenvector to $D_{11}=1$, so in our case $Ae = A^{-1}e = e$. Furthermore, $\|A\| = \varepsilon^{-1} = \|A^{-1}\|$. For $x = e/\sqrt{n}$ we have $\|b\| = \|Ax\| = \|A^{-1}x\| = \|x\| = 1$. So (2.2) gives

$$\kappa_{A,Ax}(A,x) = \varepsilon^{-2} + \varepsilon^{-1},$$

and Theorem 8.1 implies

$$\kappa^{\rm circ}_{A,Ax}(A,x) = c\frac{\varepsilon^{-1}+\varepsilon^{-1}}{1} \leq 2\varepsilon^{-1}$$

for $2^{-1/2}\leq c\leq 1$. Summarizing, we have the following result for circulants.

THEOREM 8.4. *For a nonsingular circulant $A\in M_n^{\rm circ}(\mathbb{R})$ and $0\neq x\in\mathbb{R}^n$ we have*

$$1 \geq \frac{\kappa^{\rm circ}_{A,Ax}(A,x)}{\kappa_{A,Ax}(A,x)} \geq \frac{1}{2\sqrt{2}\sqrt{\|A^{-1}\|\,\|A\|}}.$$

As by the matrix (8.5) *the second estimation is sharp up to a factor $4\sqrt{2}$ for all $n\geq 5$.*

Finally, we remark that in case of unstructured perturbations, allowing or not allowing perturbations in the right-hand side may alter the condition number by at most a factor of 2; see (4.1). This changes dramatically for circulant structured perturbations (and also for other structures). Following along the lines of example (8.5), define

$$A = F^H\,{\rm diag}(1,v,\varepsilon,[1,]\varepsilon,v)F \tag{8.6}$$

with v denoting a row vector of $m\geq 0$ ones. Thus, A is of dimension $n=2m+3$ or $n=2m+4$, depending on whether the diagonal element is left out or not. The same arguments as before apply to $x=e/\sqrt{n}$, and $\|A\| = \|A^{-1}x\| = \|Ax\| = 1$, $\|A^{-1}\| = \varepsilon^{-1}$, (2.2), and Theorem 8.1 yield

$$\kappa_A(A,x) = \varepsilon^{-1}, \quad \kappa^{\rm circ}_{A,Ax}(A,x) \geq 2^{-1/2}(1+\varepsilon^{-1}), \quad \text{but} \quad \kappa^{\rm circ}_A(A,x) = 1.$$

For a discussion of stability of a numerical algorithm for solving a linear system it seems inappropriate to ignore perturbations in the right-hand side. So (8.3) may

be of more theoretical interest. However, Theorem 8.4 shows that a linear system may be beyond the scope of a numerical algorithm which is only stable with respect to general perturbations, whereas it may be solved to some precision by a special circulant solver.

Notice that the ratio $\kappa^{\text{circ}}/\kappa$ may only become small for ill-conditioned matrices. This is also true for componentwise perturbations, as we will see in Part II of this paper (Theorem 7.2). In fact, this is the only structure out of (2.4) for which this statement is true.

9. Symmetric Toeplitz and persymmetric Hankel matrices. With Theorem 6.5 and (6.11) we already have computable bounds for κ^{symToep} and, therefore, in view of (6.3), for $\kappa^{\text{persymHankel}}$. More can be said about κ^{symToep} and also about how small the ratio $\kappa^{\text{symToep}}/\kappa$ can be.

Let $\tilde{J} \in \{+J, -J\}, \tilde{J} \in M_n(\mathbb{R})$, and $x \in \mathbb{R}^n$ be given such that $x = \tilde{J}x$. Then $A \in M_n^{\text{symToep}}(\mathbb{R})$ implies $A = \tilde{J}A\tilde{J}$ and $\tilde{J}Ax = \tilde{J}A\tilde{J}x = Ax$. That means every $A \in M_n^{\text{symToep}}(\mathbb{R})$ maps $X := \{x \in \mathbb{R}^n : \ x = \tilde{J}x\}$ into itself. For nonsingular A, the mapping $A : \ X \to X$ is bijective. Assume for the moment that n is even, set $m = n/2$, and split A into

$$A = \begin{pmatrix} T & U \\ U^T & T \end{pmatrix} \quad \text{with} \quad T, U \in M_m(\mathbb{R}). \tag{9.1}$$

Accordingly, split $\tilde{J}$ into $\tilde{J} = \left(\begin{smallmatrix} 0 & \overline{J} \\ J & 0 \end{smallmatrix}\right)$ such that $|\overline{J}| = J_m$ is the "flip"-matrix of dimension m. Then $A = \tilde{J}A\tilde{J}$ implies $U^T = \overline{J}U\overline{J}$. For $x \in X$ this means $x = \left(\begin{smallmatrix} \overline{x} \\ \overline{Jx} \end{smallmatrix}\right)$ with $\overline{x} \in \mathbb{R}^m$ and therefore

$$Ax = A\begin{pmatrix} \overline{x} \\ \overline{J}\overline{x} \end{pmatrix} = \begin{pmatrix} (T + U\overline{J})\overline{x} \\ \overline{J}(T + U\overline{J})\overline{x} \end{pmatrix}.$$

Thus nonsingularity of A implies nonsingularity of $T + U\overline{J}$.

To estimate φ^{symToep} let nonsingular $A \in M^{\text{symToep}}$, $\Delta A \in M^{\text{symToep}}$, and $x = \tilde{J}x = \left(\begin{smallmatrix} \overline{x} \\ \overline{Jx} \end{smallmatrix}\right) \in \mathbb{R}^n$ be given. Then

$$\Delta Ax = \begin{pmatrix} \overline{y} \\ \overline{J}\overline{y} \end{pmatrix} \quad \text{and} \quad A^{-1}\Delta Ax = A^{-1}\begin{pmatrix} \overline{y} \\ \overline{J}\overline{y} \end{pmatrix} = \begin{pmatrix} \overline{z} \\ \overline{J}\overline{z} \end{pmatrix}$$

for some $\overline{y}, \overline{z} \in \mathbb{R}^m$, where $\overline{y} = (T + U\overline{J})\overline{z}$. Therefore

$$\|A^{-1}\Delta Ax\| = \left\| \begin{pmatrix} (T + U\overline{J})^{-1}\overline{y} \\ \overline{J}(T + U\overline{J})^{-1}\overline{y} \end{pmatrix} \right\| \le \|(T + U\overline{J})^{-1}\| \ \left\| \begin{pmatrix} \overline{y} \\ \overline{y} \end{pmatrix} \right\|.$$

Moreover,

$$\left\| \begin{pmatrix} \overline{y} \\ \overline{y} \end{pmatrix} \right\| = \left\| \begin{pmatrix} \overline{y} \\ \overline{J}\overline{y} \end{pmatrix} \right\| = \|\Delta Ax\| \le \|\Delta A\| \cdot \|x\|$$

and therefore

$$\varphi^{\text{symToep}}(A, x) \le \|(T + U\overline{J})^{-1}\| \, \|x\|. \tag{9.2}$$

The same analysis, only more technical, is possible for odd n. In this case $m := (n+1)/2$ and

$$\pm\overline{J} = \begin{pmatrix} & & 1 & 0 \\ & \iddots & & \vdots \\ 1 & & & 0 \end{pmatrix} \in M_{m-1,m}(\mathbb{R}).$$

Note that $x = \tilde{J}x$ implies $x_m = 0$ in case $\tilde{J} = -J$. For the splitting

$$A = \begin{pmatrix} T_1 & U \\ U^T & T_2 \end{pmatrix},\ T_1 \in M_m^{\mathrm{symToep}},\ T_2 \in M_{m-1}^{\mathrm{symToep}},\ U \in M_{m,m-1}(\mathbb{R}), \tag{9.3}$$

we obtain $\overline{J}T_1\overline{J}^T = T_2$. In a similar way as before one can show

$$\varphi^{\mathrm{symToep}}(A,x) \le \|(T_1 + U\overline{J})^{-1}\|\,\|x\|. \tag{9.4}$$

The steps are technical and omitted. Combining (9.2) and (9.4) with Theorem 3.3 we obtain the following result.

THEOREM 9.1. *Let nonsingular $A \in M_n^{\mathrm{symToep}}$, and for $\tilde{J} = sJ, s \in \{-1,1\}$, let $0 \ne x \in \mathbb{R}^n$ be given with $x = \tilde{J}x$. Set $m := \lceil n/2 \rceil$ and define*

$$\overline{J} = \begin{pmatrix} & & s \\ & \iddots & \\ s & & \end{pmatrix} \in M_m(\mathbb{R}) \quad \textit{for } n \textit{ even}$$

and

$$\overline{J} = \begin{pmatrix} & & s & 0 \\ & \iddots & & \vdots \\ s & & & 0 \end{pmatrix} \in M_{m-1}(\mathbb{R}) \quad \textit{for } n \textit{ odd.}$$

Then for $T := A[1:m, 1:m]$ and $U := A[1:m, m+1:n]$ we have

$$\kappa_{E,f}^{\mathrm{symToep}}(A,x) \le \|(T + U\overline{J})^{-1}\|\,\|E\| + \|A^{-1}\|\frac{\|f\|}{\|x\|}. \tag{9.5}$$

Particularly for no perturbations in the right-hand side, we obtain

$$\frac{\kappa_E^{\mathrm{symToep}}(A,x)}{\kappa_E(A,x)} \le \frac{\|(T + U\overline{J})^{-1}\|}{\|A^{-1}\|}. \tag{9.6}$$

Note that the upper bound for $\kappa_E^{\mathrm{symToep}}(A,x)$ is only true for x with $x = \tilde{J}x$. The ratio in the right-hand side of (9.6) may become arbitrarily small as for

$$A = \mathrm{Toeplitz}(1, 0, \ldots, 0, 1+\varepsilon) \quad \text{and} \quad x = e/\sqrt{n}.$$

Again we use Matlab notation; that is, Toeplitz(c) denotes the symmetric Toeplitz matrix with first column c. In this case $T + UJ = \mathrm{diag}(2+\varepsilon, 1, \ldots, 1)$ and $\kappa_E^{\mathrm{symToep}}(A,x) \le \|E\|$. On the other hand, $y = (-1, 0 \ldots, 0, 1)^T$ is an eigenvector of A to the eigenvalue ε, so $\kappa_E(A,x) = \|A^{-1}\|\,\|E\| \ge \varepsilon^{-1}\|E\|$. However, allowing perturbations in the right-hand side, we obtain for the natural choice $E = A$, $f = b$

$$\kappa_{A,Ax}^{\mathrm{symToep}}(A,x) \ge (2\sqrt{2}\varepsilon)^{-1}\|A\|,$$

which is almost the same as $\kappa_A(A,x)$. Indeed, allowing perturbations in the right-hand side, the ratio $\kappa^{\mathrm{symToep}}_{A,Ax}/\kappa_{A,Ax}$ depends on the condition number $\kappa_{A,Ax}$. It can only become small for ill-conditioned matrices. The ratio can be estimated as before using Lemma 6.1.

THEOREM 9.2. *Let nonsingular $A \in M_n^{\mathrm{symToep}}(\mathbb{R})$ and $0 \neq x \in \mathbb{R}^n$ be given. Then*

$$1 \geq \frac{\kappa^{\mathrm{symToep}}_{A,Ax}(A,x)}{\kappa_{A,Ax}(A,x)} \geq \frac{1}{2\sqrt{2}\cdot\sqrt{\|A^{-1}\|\,\|A\|}}.$$

Proof. We have $I \in M_n^{\mathrm{symToep}}$, so $\varphi^{\mathrm{symToep}}(A,x) \geq \|A^{-1}x\| = \|A^{-T}x\|$, and Lemma 6.1 and (4.1) finish the proof. ∎

The lower bound in Theorem 9.2 seems not far from being sharp. Consider

$$A = \mathrm{Toeplitz}(1, -1-\varepsilon, 1-\varepsilon, -1+\varepsilon) \quad \text{and} \quad x = e,$$

the symmetric Toeplitz matrix with first row $[1, -1-\varepsilon, 1-\varepsilon, -1+\varepsilon]$. Then

$$\kappa_{A,Ax}(A,x) \sim 16\varepsilon^{-2} \quad \text{and} \quad \kappa^{\mathrm{symToep}}_{A,Ax}(A,x) < 11\varepsilon^{-1}.$$

Unfortunately, we do not have a generic $n \times n$ example. However, it is numerically easy to find examples of larger dimension. Therefore, we expect the second inequality in Theorem 9.2 to be sharp up to a small constant for all n.

Additional algebraic properties such as positive definiteness of the matrix do not improve the situation. An example is the symmetric positive definite Toeplitz matrix A with first row $(1+\varepsilon^2, -1+\varepsilon, 1-\varepsilon, -1+3\varepsilon)$ and $x := e$. One computes $\lambda_{\min}(A) = 0.75\varepsilon^2 + \mathcal{O}(\varepsilon^3)$, and (2.2) and Theorem 9.1 yield

$$\kappa_{A,Ax}(A,x) > 5.33\varepsilon^{-2} + \mathcal{O}(\varepsilon^{-1}) \quad \text{and} \quad \kappa^{\mathrm{symToep}}_{A,Ax}(A,x) < 7\varepsilon^{-1} + \mathcal{O}(1).$$

Note that the estimation in Theorem 9.1 is only valid for $x = \tilde{J}x$, $\tilde{J} = sJ$, $s \in \{+1,-1\}$. Let general $x \in \mathbb{R}^n$ be given and split $x = \binom{u}{v}$ into $u \in \mathbb{R}^m, v \in \mathbb{R}^{n-m}$. Define $\overline{J}$ as in Theorem 9.1 with $s = 1$, and set $\overline{y} := \frac{1}{2}(u + \overline{J}v)$ and $\overline{z} := \frac{1}{2}(u - \overline{J}v)$. Then for $y := \binom{\overline{y}}{\overline{J}\overline{y}} \in \mathbb{R}^n$ and $z := \binom{\overline{z}}{-\overline{J}\overline{z}} \in \mathbb{R}^n$ we have

$$Jy = y,\ -Jz = z, \text{ and } x = y + z.$$

For $\Delta A \in M_n^{\mathrm{symToep}}$ and $\|\Delta A\| \leq 1$ we can apply (9.2) and (9.4) to conclude that

$$\|A^{-1}\Delta A x\| = \|A^{-1}\Delta A(y+z)\| \leq \|(T+U\overline{J})^{-1}\|\,\|y\| + \|(T-U\overline{J})^{-1}\|\,\|z\|.$$

COROLLARY 9.3. *For nonsingular $A \in M_n^{\mathrm{symToep}}$, $0 \neq x \in \mathbb{R}^n$, and $T, U, \overline{J}$, y and z as defined above we have*

$$\kappa^{\mathrm{symToep}}_{E,f}(A,x) \leq \left(\|(T+U\overline{J})^{-1}\|\,\|y\| + \|(T-U\overline{J})^{-1}\|\,\|z\|\right)\frac{\|E\|}{\|x\|} + \|A^{-1}\|\frac{\|f\|}{\|x\|}.$$

Obviously $\|y\| \leq \|x\|$ and $\|z\| \leq \|x\|$, so one may replace the expression in the parentheses by $\mu\|x\|$ with $\mu := \max(\|(T+U\overline{J})^{-1}\|, \|(T-U\overline{J})^{-1}\|)$. However, such an approach does not give additional information. Let $A^{-1}w = \lambda w$ for $0 \neq w \in \mathbb{R}^n$. Then $A^{-1}\cdot Jw = JA^{-1}J\cdot Jw = \lambda Jw$, such that $A^{-1}(w \pm Jw) = \lambda(w \pm Jw)$. At

least one of $w \pm Jw$ is nonzero, so we conclude that to every eigenvalue of A^{-1} there is an eigenvector w such that $w = sJw$ for $s \in \{-1, +1\}$. For $w, \|w\| = 1$ being an eigenvector to $|\lambda| = \|A^{-1}\|$ and for the splitting $w = \begin{pmatrix} \overline{w} \\ \overline{Jw} \end{pmatrix}$ it follows that

$$\|A^{-1}\| = \|A^{-1}w\| = \left\| \begin{pmatrix} (T + sU\overline{J})^{-1}\overline{w} \\ \overline{J}(T + sU\overline{J})^{-1}\overline{w} \end{pmatrix} \right\| \leq \mu \cdot \left\| \begin{pmatrix} \overline{w} \\ \overline{w} \end{pmatrix} \right\| = \mu,$$

so that the above approach only verifies $\kappa_{E,f}^{\text{symToep}} \leq \kappa_{E,f}$.

We also see from this how to construct examples with small ratio $\kappa^{\text{symToep}}/\kappa$. If A is ill conditioned, at least one of the matrices $T + sU\overline{J}$ must be equally ill conditioned. Small ratios may occur if one of them, say for $s = 1$, is well conditioned and x is chosen with big part $y = Jy$ but small $z = -Jz$ in the splitting $x = y + z$.

Finally, note that Theorem 9.1 and Corollary 9.3 give upper bounds for κ^{symToep}. We do not know how sharp estimation (9.5) is. Numerical experience suggests that the overestimation is small. Can that be proved? Again, all statements in this section are valid mutatis mutandis for $A \in M_n^{\text{persymHankel}}$.

10. Hankel and general Toeplitz matrices. With Theorem 6.5 and (6.11) we already have computable bounds for the (normwise) Hankel condition number and therefore, in view of (6.4), for κ^{Toep}. In the following we investigate how small the ratio $\kappa^{\text{Hankel}}/\kappa$ can be. We first show a lower bound in the spirit of Theorems 8.4 and 9.2.

Suppose $A^T = A \in M_n(\mathbb{R})$, not necessarily $A \in M_n^{\text{Hankel}}(\mathbb{R})$. By definition,

$$\varphi^{\text{Hankel}}(A, x) = \sup\{\|A^{-1}\Delta A x\| : \ \Delta A \in M_n^{\text{Hankel}}(\mathbb{R}), \ \|\Delta A\| \leq 1\}.$$

Hankel matrices are symmetric. So if we can show that for every $0 \neq x \in \mathbb{R}^n$ there is a Hankel matrix ΔA with $\|\Delta A\| \leq 1$ and $\Delta A x = x$, then

$$\varphi^{\text{Hankel}}(A, x) \geq \|A^{-1}x\| = \|A^{-T}x\|, \tag{10.1}$$

and Lemma 6.1 delivers the desired bound. This is indeed true, as shown by the following lemma. We will prove it for the real and complex cases, the latter being needed in sections 11 and 12.

LEMMA 10.1. *Let $x \in \mathbb{C}^n$ be given. Then there exists $H \in M_n^{\text{Hankel}}(\mathbb{C})$ with $Hx = \bar{x}$ and $\|H\| \leq 1$, where $\bar{x}$ denotes the complex conjugate of x. In case $x \in \mathbb{R}^n$, H can be chosen real so that $Hx = x$.*

Proof. The expression (6.5) is of course also true for complex Hankel matrices because Φ^{Hankel} is a 0/1-matrix. So we are looking for a parameter vector $p \in \mathbb{C}^{2n-1}$ such that the Hankel matrix H with $\text{vec}(H) = \Phi^{\text{Hankel}}p$ satisfies the assertions of the lemma. Then

$$Hx = \Psi_x^{\text{Hankel}} p \tag{10.2}$$

for Ψ_x^{Hankel} as in (6.9), (6.10), and (6.11). We discuss the following for $n = 3$, which will give enough information for the general case. We first embed $\Psi_x := \Psi_x^{\text{Hankel}}$ into the circulant C_x with the first row identical to that of Ψ_x, i.e.,

$$C_x := \begin{pmatrix} x_1 & x_2 & x_3 & 0 & 0 \\ 0 & x_1 & x_2 & x_3 & 0 \\ 0 & 0 & x_1 & x_2 & x_3 \\ x_3 & 0 & 0 & x_1 & x_2 \\ x_2 & x_3 & 0 & 0 & x_1 \end{pmatrix}.$$

Then the matrix of the first n rows of C_x is equal to Ψ_x. Define

$$C := C_x^+ C_x^H, \tag{10.3}$$

with C_x^+ denoting the pseudoinverse of C_x. For $C_x = F^H D F$, the pseudoinverse $C_x^+ = F^H D^+ F$ is also a circulant, and we have

$$C_x C = F^H D F \cdot F^H D^+ F \cdot F^H D^H F = F^H D D^+ D^H F = F^H D^H F = C_x^H.$$

But Ψ_x comprises the first n rows of C_x, so $\Psi_x C$ is equal to the matrix of the first n rows of C_x^H. Define

$$p := Ce_1, \tag{10.4}$$

with e_1 denoting the first column of I_{2n-1}. The first n rows of $C_x^H e_1$ form the vector $\bar{x}$, so by (10.2),

$$Hx = \Psi_x p = \Psi_x C e_1 = \bar{x}$$

for the Hankel matrix H defined by the parameter vector $p = Ce_1$. Note that by (10.3) and (10.4) C, and therefore H, is real for real x so that $Hx = x$ in that case.

It remains to estimate the matrix norm of H. Denote the first column of the circulant C by $(c_1, \ldots, c_{2n-1})^T$. For $n = 3$, the definitions (10.4) and (10.2) imply

$$H = \begin{pmatrix} c_1 & c_2 & c_3 \\ c_2 & c_3 & c_4 \\ c_3 & c_4 & c_5 \end{pmatrix} \quad \text{and} \quad HJ = \begin{pmatrix} c_3 & c_2 & c_1 \\ c_4 & c_3 & c_2 \\ c_5 & c_4 & c_3 \end{pmatrix}.$$

The matrix HJ is the lower left $n \times n$ submatrix of C. So by Lemma 8.2 it follows that

$$\|H\| = \|HJ\| \le \|C\| = \|C_x^+ C_x^H\| = \|C_x^+ C_x\| = \|D^+ D\| = 1. \qquad \square$$

Combining Lemma 10.1 with (10.1), Lemma 6.1, and (4.1) proves the following lower bounds. Note that only symmetry of A was used in (10.1).

THEOREM 10.2. *Let nonsingular symmetric $A \in M_n(\mathbb{R})$, and let $0 \neq x \in \mathbb{R}^n$ be given. Then*

$$\kappa_{A,Ax}^{\text{Hankel}}(A, x) \ge 2^{-1/2} \sqrt{\|A^{-1}\| \, \|A\|} \tag{10.5}$$

and therefore

$$1 \ge \frac{\kappa_{A,Ax}^{\text{Hankel}}(A, x)}{\kappa_{A,Ax}(A, x)} \ge \frac{1}{2\sqrt{2} \cdot \sqrt{\|A^{-1}\| \, \|A\|}}. \tag{10.6}$$

The lower bound (10.6) is a severe underestimation, in fact, it is independent of A. By Corollary 6.6 we know that

$$2^{1/2} \frac{\kappa_{E,f}^{\text{Hankel}}(A, x)}{\kappa_{E,f}(A, x)} \ge \frac{\kappa_E^{\text{Hankel}}(A, x)}{\kappa_E(A, x)} \ge n^{-1/2} \frac{\sigma_{\min}(\Psi_x^{\text{struct}})}{\|x\|}.$$

If Ψ_x^{struct} were rank-deficient, this would imply that every minor of size n is zero. Then the minor of first n columns of Ψ_x^{Hankel} as in (6.10) implies $x_1 = 0$, and continuing

TABLE 10.1
Mean value and standard deviation of $\tau(x) = \sigma_{\min}(\Psi_x^{\text{Hankel}})/\|x\|$.

	Uniform x_i		Normal x_i	
n	mean($\tau(x)$)	std($\tau(x)$)	mean($\tau(x)$)	std($\tau(x)$)
10	0.49	0.135	0.50	0.136
20	0.42	0.110	0.42	0.111
50	0.35	0.084	0.35	0.084
100	0.31	0.069	0.31	0.069

with the minors of columns i to $i+n-1$ we conclude that $x = 0$. This implies $\sigma_{\min}(\Psi_x^{\text{struct}}) > 0$ for all $x \neq 0$ such that for fixed x there is a minimum ratio of the structured Hankel and the unstructured condition number.

Extensive numerical statistics on $\tau(x) := \sigma_{\min}(\Psi_x^{\text{Hankel}})/\|x\|$ suggest that this minimum is in general not too far from 1. In Table 10.1 we list the mean value and standard deviation of $\tau(x)$ for some 10^6 samples of x with entries uniformly distributed in $[-1, 1]$ and for entries of x with normal distribution with mean 0 and standard deviation 1.

We mention that the numbers in the two rightmost columns in Table 10.1 are almost the same for solution vectors x such that $x_i = s \cdot y_i$ with random sign $s \in \{-1, 1\}$ and uniform y_i with mean 1 and standard deviation 1.

Note again that this is a statistic on *solution vectors* x showing a lower bound for the ratio in Corollary 6.6 between the Hankel and the traditional (unstructured) condition numbers. This ratio applies to *every* matrix A regardless of its condition number.

Small values of $\tau(x) = \sigma_{\min}(\Psi_x^{\text{Hankel}})/\|x\|$ seem rare, but they are possible. Particularly, small values seem to occur for positive x and $x = Jx$. Statistically the means in Table 10.1 drop by about a factor of 2 to 3 for such randomly chosen x. A specific choice of x proposed by Heinig [21] is comprised of the coefficients of $(t+1)^{n-1}$. For this x we obtain

$$\tau(x) \sim 2.5^{-n}.$$

This generates a lower bound for $\kappa_{A,Ax}^{\text{Hankel}}(A, x)$. We indeed managed to find Hankel matrices with $\|A^{-1}\Psi_x^{\text{Hankel}}\|/\|x\| < 2^{-n}\|A^{-1}\|$ for that x and dimensions up to 15. That means for the unperturbed right-hand side it is $\kappa_A^{\text{Hankel}}(A, x)/\kappa_A(A, x) < 2^{-n}$. We could neither construct generic $n \times n$ matrices A with this property nor find examples with the ratio of condition numbers $\kappa_{A,Ax}^{\text{Hankel}}(A, x)/\kappa_{A,Ax}(A, x)$ (allowing perturbations in the right-hand side) getting significantly less than one. This includes in particular positive definite Hankel matrices which are known to be generally ill-conditioned [3].

An open problem is how small $\tau(x)$ can be; that is, what is the smallest possible value of $\sigma_{\min}(\Psi_x^{\text{Hankel}})$ for Ψ_x^{Hankel} as in (6.10) and $\|x\| = 1$? Based on that, how small may $\kappa_A^{\text{Hankel}}(A, x)/\kappa_A(A, x)$ become?

For general (normwise) perturbations in the matrix *and* the right-hand side we conjecture that Hankel structured and unstructured stabilities differ only by a small factor, supposedly only mildly or not at all, depending on n. In other words, $\kappa_{A,Ax}^{\text{Hankel}}(A, x)/\kappa_{A,Ax}(A, x) \geq \gamma$ for γ not much less than one.

Meanwhile Böttcher and Grudsky give a partial answer to that [6]. They show, based on a deep result by Konyagin and Schlag [31], that there exist universal constants $n_0 \in \mathbb{N}$ and $\varepsilon > 0$ such that the following is true. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$n \geq n_0$, comprise independent standard normal or independent Rademacher variables (recall that Rademacher variables are random with value 1 or -1 each with probability 1/2). Then, for all $A \in M_n^{\mathrm{Hankel}}(\mathbb{R})$,

$$\text{probability}\left(\frac{\kappa_A^{\mathrm{Hankel}}(A,x)}{\kappa_A(A,x)} \geq \frac{\varepsilon}{n^{3/2}} \right) > \frac{99}{100}.$$

11. Inversion of structured matrices. Similarly to the structured condition number for linear systems, the structured condition number for matrix inversion is defined by

(11.1)
$$\kappa_E^{\mathrm{struct}}(A) := \lim_{\varepsilon \to 0} \sup \left\{ \frac{\|(A+\Delta A)^{-1} - A^{-1}\|}{\varepsilon \|A^{-1}\|} : \ \Delta A \in M_n^{\mathrm{struct}}(\mathbb{R}), \|\Delta A\| \leq \varepsilon \|E\| \right\}.$$

For $M_n^{\mathrm{struct}}(\mathbb{R}) = M_n(\mathbb{R})$ this is the usual (unstructured) condition number which is well known [27, Theorem 6.4] to be

$$\kappa_E(A) = \|A^{-1}\| \, \|E\|.$$

Surprisingly, the same is true for all of the linear structures in (2.4). A reasoning is that by Theorem 4.1 the worst case condition number of a linear system maximized over all right-hand sides is equal to the unstructured condition number. So in some way the set of columns of the identity matrix is general enough to achieve the worst case.

THEOREM 11.1. *Let nonsingular $A \in M_n^{\mathrm{struct}}(\mathbb{R})$ be given for* struct $\in$ {sym, persym, skewsym, symToep, Toep, circ, Hankel, persymHankel}. *Then*

$$\kappa_E^{\mathrm{struct}}(A) = \|A^{-1}\| \, \|E\|.$$

Proof. As in the unstructured case we use the expansion

$$(A+\Delta A)^{-1} - A^{-1} = -A^{-1}\Delta A A^{-1} + \mathcal{O}(\|\Delta A\|^2).$$

Therefore, the result is proved if we can show that

$$\omega^{\mathrm{struct}}(A) := \sup\{\|A^{-1}\Delta A A^{-1}\| : \ \Delta A \in M^{\mathrm{struct}}, \|\Delta A\| \leq 1\} \geq \|A^{-1}\|^2 \tag{11.2}$$

because this obviously implies $\omega^{\mathrm{struct}}(A) = \|A^{-1}\|^2$. Let $x, y \in \mathbb{R}^n$, $\|x\| = \|y\| = 1$ be given with $A^{-1}x = \|A^{-1}\|y$. Then Definition 3.1 implies

$$\omega^{\mathrm{struct}}(A) \geq \sup\{\|A^{-1}\Delta A A^{-1}x\| : \ \Delta A \in M^{\mathrm{struct}}, \|\Delta A\| \leq 1\} = \|A^{-1}\| \varphi^{\mathrm{struct}}(A,y).$$

Therefore Lemma 5.2 proves (11.2) for struct $\in$ {sym, persym, skewsym}. For normal $A \in M_n^{\mathrm{struct}}(\mathbb{R})$, it is $A^{-1}x = \lambda x$ with $\|x\| = 1$ and $|\lambda| = \|A^{-1}\|$. Hence (11.2) is also proved for symmetric Toeplitz and circulant structures by using $\Delta A := I$. For $A \in M_n^{\mathrm{persymHankel}}(\mathbb{R})$, $AJ \in M_n^{\mathrm{symToep}}(\mathbb{R})$ and $JA^{-1}x = \lambda x$ with $\|x\| = 1$, and $|\lambda| = \|JA^{-1}\| = \|A^{-1}\|$ proves (11.2) by using $\Delta A := J \in M_n^{\mathrm{persymHankel}}(\mathbb{R})$. For Hankel matrices again $A^{-1}x = \lambda x$ for $\|x\| = 1$ and $|\lambda| = \|A^{-1}\|$, and Lemma 10.1 yields existence of $\Delta A \in M_n^{\mathrm{Hankel}}(\mathbb{R})$ with $\|\Delta A\| \leq 1$ and $\Delta A x = x$, and for $A \in M_n^{\mathrm{Toep}}(\mathbb{R})$ we have $AJ \in M_n^{\mathrm{Hankel}}(\mathbb{R})$. ∎

The theorem shows that among the worst case perturbations for the inverse of a structured matrix there are always perturbations of the same structure, the same result (cf. Theorem 5.3) as for linear systems with fixed right-hand side and struct $\in \{\mathrm{sym}, \mathrm{persym}, \mathrm{skewsym}\}$.

The proof basically uses the fact that A or JA is normal. It also can be extended to the complex case. Here the structure is still strong enough, although the singular values need not coincide with the absolute values of the eigenvalues. We have the following result.

THEOREM 11.2. *Let nonsingular $A \in M_n^{\mathrm{struct}}(\mathbb{C})$ be given for* struct *being Hermitian, skew-Hermitian, Toeplitz, circulant, or Hankel. Then*

$$\kappa_E^{\mathrm{struct}}(A) = \|A^{-1}\| \, \|E\|.$$

Proof. We proceed as in the proof of Theorem 11.1 and have to show $\omega^{\mathrm{struct}}(A) \geq \|A^{-1}\|^2$ for the $\omega^{\mathrm{struct}}(A)$ defined in (11.2). For normal A, there is $A^{-1}x = \lambda x$ with $\|x\| = 1$ and $|\lambda| = \|A^{-1}\|$. So the theorem is proved for the Hermitian and circulant cases by using $\Delta A = I$, and for the skew-Hermitian case by using $\Delta A = \sqrt{-1}I$.

For A being Hankel, A is especially (complex) symmetric. So a result by Takagi [28, Corollary 4.4.4] implies $A = U\Sigma U^T$ for nonnegative diagonal Σ and unitary U. For x denoting the nth column of U we have $A\bar{x} = \sigma_{\min}(A)x$, and therefore $A^{-1}x = \|A^{-1}\|\bar{x}$. By Lemma 10.1 there exists $\Delta A \in M_n^{\mathrm{Hankel}}(\mathbb{C})$ with $\|\Delta A\| \leq 1$ and $\Delta A\bar{x} = x$, so that $A^{-1}\Delta A A^{-1}x = \|A^{-1}\|^2 x$ and

$$\omega^{\mathrm{struct}}(A) \geq \|A^{-1}\Delta A A^{-1}x\| = \|A^{-1}\|^2.$$

Finally, for complex Toeplitz A, $H := JA$ is Hankel and, as above, we conclude that there is x and ΔH with $H^{-1}\Delta H H^{-1}x = \|H^{-1}\|^2 x$. Then $\Delta A := J\Delta H$ is Toeplitz with $\|\Delta A\| \leq 1$, and $y := Jx$ with $\|y\| = 1$ yields

$$\omega^{\mathrm{struct}}(A) \geq \|A^{-1}\Delta A A^{-1}y\| = \|H^{-1}\Delta H H^{-1}x\| = \|H^{-1}\|^2 = \|A^{-1}\|^2. \qquad \square$$

One might conjecture that the result in Theorems 11.1 and 11.2 is true for all linear structures. This is, however, not the case, for example, for (general) tridiagonal Toeplitz matrices or, more generally, for (general) tridiagonal matrices. Consider

$$A = \begin{pmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{pmatrix} \tag{11.3}$$

for small $\alpha > 0$. Then $\|A\| \sim 1$ and $\|A^{-1}\| \sim \alpha^{-3}$. For general $\Delta A \in M_3^{\mathrm{tridiag}}(\mathbb{R})$ with $\|\Delta A\| \leq 1$ one computes $\|A^{-1}\Delta A A^{-1}\| = \mathcal{O}(\alpha^{-5})$, so that $\omega^{\mathrm{struct}}(A)$ defined in (11.2) is of the order $\alpha\|A^{-1}\|^2$. This implies that $\kappa_E^{\mathrm{tridiag}}(A)$ is of the order $\alpha\|A^{-1}\| \, \|E\|$ instead of $\|A^{-1}\| \, \|E\|$. The same applies for general tridiagonal Toeplitz perturbations. Nevertheless one may ask: Is Theorem 11.1 true for other structures?

Usually linear systems are not solved by multiplying the right-hand side by a computed inverse. For structured matrices with small ratio $\kappa_{A,Ax}^{\mathrm{struct}}/\kappa_{A,Ax}$, lack of stability is yet another reason for that.

12. Distance to singularity. The condition number $\kappa(A) = \|A^{-1}\| \, \|A\|$ of a matrix is infinite iff the matrix is singular. Therefore it seems plausible that the distance to singularity of a matrix is inversely proportional to its condition number. Define

$$\delta_E^{\mathrm{struct}}(A) := \min\{\alpha : \ \Delta A \in M_n^{\mathrm{struct}}(\mathbb{R}), \|\Delta A\| \leq \alpha\|E\|, A + \Delta A \text{ singular}\},$$

where $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$. For $M_n^{\text{struct}}(\mathbb{R}) = M_n(\mathbb{R})$ this number $\delta_E(A)$ is the traditional (normwise) distance to the nearest singular matrix with respect to unstructured perturbations. A classical result [27, Theorem 6.5] by Eckart and Young [12] for the 2-norm with generalizations by Gastinel and Kahan to other norms is

$$\delta_E(A) = \{\|A^{-1}\| \, \|E\|\}^{-1} = \kappa_E(A)^{-1} \tag{12.1}$$

for general perturbations $M_n^{\text{struct}}(\mathbb{R}) = M_n(\mathbb{R})$. Thus the distance to singularity for general perturbations is not only inversely proportional to but *equal* to the reciprocal of the condition number. Note that the distance to singularity as well as the condition number may change with diagonal scaling, the former being contrary to componentwise perturbations (cf. Part II, section 9).

There are a number of results on some blockwise structured distance to singularity and on the so-called μ-number (cf. [11, 13, 34, 41, 35]). There also are results on distance to singularity with respect to certain symmetric structures [29]. The question remains of whether a result similar to (12.1) can be obtained for the structured condition number and distance to singularity. It was indeed shown by D. Higham [23] that (12.1) is also true for symmetric perturbations.

In the previous section we have seen that the structured condition number $\kappa_E^{\text{struct}}(A)$ is equal to the unstructured condition number $\|A^{-1}\| \, \|E\|$ for any E and for *all* structures in (2.4).

We conclude with the remarkable fact that the reciprocal of the condition number is equal to the structured distance to the nearest singular matrix for *all* structures in (2.4).

THEOREM 12.1. *Let nonsingular* $A \in M_n^{\text{struct}}(\mathbb{R})$ *for* struct $\in$ {sym, persym, skewsym, symToep, Toep, circ, Hankel, persymHankel} *be given. Then*

$$\delta_E(A) = \delta_E^{\text{struct}}(A) = \kappa_E^{\text{struct}}(A)^{-1} = \kappa_E(A)^{-1} = \{\|A^{-1}\| \, \|E\|\}^{-1}. \tag{12.2}$$

Proof. Without loss of generality we may assume $\|E\| = 1$. Then obviously $\delta_E^{\text{struct}}(A) \geq \delta_E(A) = \sigma_{\min}(A)$, and it remains to show $(A + \Delta A)x = 0$ for some $0 \neq x \in \mathbb{R}^n$ and $\Delta A \in M_n^{\text{struct}}(\mathbb{R})$ with $\|\Delta A\| = \sigma_{\min}(A)$.

For symmetric matrices there is real λ and $0 \neq x \in \mathbb{R}^n$ with $Ax = \lambda x$ and $|\lambda| = \sigma_{\min}(A)$. If $I \in M_n^{\text{struct}}(\mathbb{R})$, then $\Delta A = -\lambda I$ does the job. This proves (12.2) for struct $\in$ {sym, symToep}. For struct $\in$ {persym, persymHankel} and $A \in M_n^{\text{struct}}(\mathbb{R})$, JA is symmetric and $JAx = \lambda x$ for $0 \neq x \in \mathbb{R}^n$ and $|\lambda| = \sigma_{\min}(JA) = \sigma_{\min}(A)$. Therefore $\det(J(A+\Delta A)) = 0 = \det(A+\Delta A)$ for $\Delta A := -\lambda J \in M_n^{\text{struct}}(\mathbb{R})$.

For nonsingular skewsymmetric A we conclude as in the proof of Lemma 5.2 that all singular values have even multiplicity and that there are $u, v \in \mathbb{R}^n$ with $\|u\| = \|v\| = 1$, $u^T v = 0$, and $Av = \sigma_{\min}(A)u$. By Lemma 5.1 we find $\Delta A \in M_n^{\text{skewsym}}(\mathbb{R})$ with $\Delta A v = u$ and $\|\Delta A\| = 1$, so that $A - \sigma_{\min}(A)\Delta A$ is singular.

For a given real circulant $A = F^H D F$ there is $Ax = \lambda x$ with $0 \neq x \in \mathbb{C}^n$ and $|\lambda| = \sigma_{\min}(A)$. If λ is real, $\Delta A = -\lambda I \in M_n^{\text{circ}}(\mathbb{R})$ yields $\det(A + \Delta A) = 0$. For complex λ, define diagonal $\tilde{D} \in M_n(\mathbb{C})$ with all entries zero except the two entries λ and $\overline{\lambda}$ in the same position as in D. Define $\Delta A := F^H \tilde{D} F$. Then Lemma 8.3 implies that ΔA is a real circulant. Moreover, we have $\|\Delta A\| = \max |\tilde{D}_{\nu\nu}| = |\lambda| = \sigma_{\min}(A)$, and $A - \Delta A = F^H(D - \tilde{D})F$ is singular.

For Hankel matrices there is $Ax = \lambda x$, $\|x\| = 1$, and $|\lambda| = \sigma_{\min}(A)$, and Lemma 10.1 proves this part. Finally, $A \in M_n^{\text{Toep}}(\mathbb{R})$ implies $AJ \in M_n^{\text{Hankel}}$ and we proceed as before. ∎

As in the previous section we can formulate this theorem also for complex structures. For nonnormal matrices such as complex Hankel and Toeplitz matrices the key is again the complex part of Lemma 10.1.

THEOREM 12.2. *Let nonsingular* $A \in M_n^{\text{struct}}(\mathbb{C})$ *be given for* struct *being Hermitian, skew-Hermitian, Toeplitz, circulant, or Hankel. Then*

$$\delta_E(A) = \delta_E^{\text{struct}}(A) = \kappa_E^{\text{struct}}(A)^{-1} = \kappa_E(A)^{-1} = \{\|A^{-1}\|\,\|E\|\}^{-1}.$$

Proof. The proof of Theorem 12.1 obviously carries over to the normal case, that is, to complex Hermitian, skew-Hermitian, and circulant matrices. For a Hankel matrix A we use [28, Corollary 4.4.4] the factorization $A = U\Sigma U^T$ with nonnegative diagonal Σ and unitary U as in the previous section. For x denoting the nth column of U we have $A\bar{x} = \sigma_{\min}(A)x$. By Lemma 10.1 there exists $\Delta H \in M_n^{\text{Hankel}}(\mathbb{C})$ with $\|\Delta H\| \le 1$ and $\Delta H\bar{x} = x$. Obviously, $-\Delta H \in M_n^{\text{Hankel}}(\mathbb{C})$ as well, so that $\Delta A := \sigma_{\min}(A)\Delta H$, $\|\Delta A\| = \sigma_{\min}(A)$, and $(A + \Delta A)x = 0$ finish this part of the proof. For A being Toeplitz, JA is Hankel and we proceed as in the proof of Theorem 12.1. ∎

So our results are a structured version of the Eckart–Young theorem, valid for all of our structures in (2.4) including the complex case. Does the result extend to other structures?

13. Conclusion. We proved that for some problems and structures it makes no, or not much, difference whether perturbations are structured or not; for other problems and structures we showed that the sensitivity with respect to structured (normwise) perturbations may be much less than with respect to unstructured perturbations. This was especially true for the important cases of linear systems with a symmetric Toeplitz or circulant matrix. Surprisingly, it turned out that the ratio $\kappa^{\text{struct}}/\kappa$ can only become small for certain *solutions*, independent of the matrix.

The results show that a small ratio $\kappa^{\text{struct}}/\kappa$ seems not typical. So our results may be used to rely on the fact that unstructured and structured sensitivities are, in general, not too far apart. However, it may also define the challenge to design numerical algorithms to solve problems with structured data being stable not only with respect to unstructured perturbations but *being stable with respect to the corresponding structured perturbations*. There exists a result in that direction for normwise perturbations and circulant matrices [40], [27, Theorem 24.3]. However, structured analysis for circulants is assisted by the fact that circulants commute. Beyond that, there are similar results for nonlinear structures such as Cauchy or Vandermonde-like matrices (see the last section in Part II of this paper). We hope our results stimulate further research in that direction for other structures.

Acknowledgments. I wish to thank many colleagues for inspiring discussions, especially Chris Beattie, Gene Golub, Ilse Ipsen, and Nick Trefethen. Also my thanks to the students of a winter course 2001/02 on this subject, especially to Frank Blömeling and Christian Keil, for many helpful discussions, remarks, and for their indulgence during the development of the results. My special thanks go to Albrecht Böttcher from TU Chemnitz for his thorough reading and most constructive and valuable comments. He also granted me the privilege of having a part of this paper appear in his forthcoming book [6]. Last but not least, many thanks to the two anonymous referees for their thorough reading and most constructive and useful comments.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[2] S. G. Bartels and D. J. Higham, *The structured sensitivity of Vandermonde-like systems*, Numer. Math., 62 (1992), pp. 17–33.

[3] B. Beckermann, *The condition number of real Vandermode, Krylov and positive definite Hankel matrices*, Numer. Math., 85 (2000), pp. 553–577.

[4] A. W. Bojanczyk, R. P. Brent, and F. R. de Hoog, *Stability analysis of a general Toeplitz systems solver*, Numer. Algorithms, 10 (1995), pp. 225–244.

[5] T. Boros, T. Kailath, and V. Olshevsky, *Pivoting and backward stability of fast algorithms for solving Cauchy linear equations*, Linear Algebra Appl., 343/344 (2001), pp. 63–99.

[6] A. Böttcher and S. Grudsky, *Special Characteristics of Toeplitz Band Matrices*, to appear.

[7] R. P. Brent, *Stability of Fast Algorithms for Structured Linear Systems*, Technical Report TR-CS-97-18, Department of Computer Science, Australian National University, Canberra, 1997.

[8] 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.

[9] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[10] J. Demmel, *Accurate singular value decompositions of structured matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 562–580.

[11] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEE-D, 129 (1982), pp. 242–250.

[12] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[13] M. K. H. Fan, A. L. Tits, and J. C. Doyle, *Robustness in presence of mixed parametric uncertainty and unmodeled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.

[14] V. Frayssé, S. Gratton, and V. Toumazou, *Structured backward error and condition number for linear systems of the type* $A * Ax = b$, BIT, 40 (2000), pp. 74–83.

[15] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comput., 64 (1995), pp. 1557–1576.

[16] I. Gohberg and I. Koltracht, *Mixed, componentwise, and structured condition numbers*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 688–704.

[17] I. Gohberg and I. Koltracht, *Structured condition numbers for linear matrix structures*, in Linear Algebra for Signal Processing, A. Bojanczyk and G. Cybenko, eds., IMA Vol. Math. Appl. 69, Springer-Verlag, New York, 1995, pp. 17–26.

[18] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[19] M. Gu, *Stable and Efficient Algorithms for Structured Systems of Linear Equations*, Technical Report LBL-37690, Lawrence Berkeley National Laboratory, Berkeley, CA, 1995.

[20] W. Hager, *Condition estimates*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 311–316.

[21] G. Heinig, *private communication*, Kuwait University, 2000.

[22] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra for Signal Processing, A. Bojanczyk and G. Cybenko, eds., IMA Vol. Math. Appl. 69, Springer-Verlag, New York, 1994, pp. 63–81.

[23] D. J. Higham, *Condition numbers and their condition numbers*, Linear Algebra Appl., 214 (1995), pp. 193–213.

[24] 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.

[25] D. J. Higham and N. J. Higham, *Componentwise perturbation theory for linear systems with multiple right-hand sides*, Linear Algebra Appl., 174 (1992), pp. 111–129.

[26] N. J. Higham, *Experience with a matrix norm estimator*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 804–809.

[27] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[28] R. A. Horn and Ch. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[29] T. Hu and L. Qiu, *On structured perturbation of Hermitian matrices*, Linear Algebra Appl., 275/276 (1998), pp. 287–314.

[30] *ANSI/IEEE* 754-1985*, Standard for Binary Floating-Point Arithmetic*, IEEE, Piscataway, NJ, 1985.

[31] S. V. Konyagin and W. Schlag, *Lower bounds for the absolute value of random polynomials on a neighborhood of the unit circle*, Trans. Amer. Math. Soc., 351 (1999), pp. 4963–4980.
[32] D. J. Higham and N. J. Higham, *MATLAB Guide*, SIAM, Philadelphia, 2000.
[33] V. Olshevsky, *Pivoting for structured matrices with applications*, Linear Algebra Appl., to appear.
[34] A. Packard and J. Doyle, *The complex structured singular value*, Automatica, 29 (1993), pp. 71–109.
[35] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, and J. C. Doyle, *A formula for computation of the real stability radius*, Automatica, 31 (1995), pp. 879–890.
[36] J. Rohn, *A new condition number for matrices and linear systems*, Computing, 41 (1989), pp. 167–169.
[37] S. M. Rump, *Structured perturbations and symmetric matrices*, Linear Algebra Appl., 278 (1998), pp. 121–132.
[38] S. M. Rump, *Ill-conditionedness need not be componentwise near to ill-posedness for least squares problems*, BIT, 39 (1999), pp. 143–151.
[39] J.-G. Sun, *Bounds for the structured backward errors of Vandermonde systems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 45–59.
[40] P. Y. Yalamov, *On the almost strong stability of the circular deconvolution algorithm*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 358–363.
[41] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 31–56

STRUCTURED PERTURBATIONS PART II: COMPONENTWISE DISTANCES*

SIEGFRIED M. RUMP†

Abstract. In the second part of this paper we study condition numbers with respect to componentwise perturbations in the input data for linear systems and for matrix inversion, and the distance to the nearest singular matrix. The structures under investigation are linear structures, namely symmetric, persymmetric, skewsymmetric, symmetric Toeplitz, general Toeplitz, circulant, Hankel, and persymmetric Hankel structures. We give various formulas and estimations for the condition numbers. For all structures mentioned except circulant structures we give explicit examples of linear systems $A_\varepsilon x = b$ with parameterized matrix A_ε such that the unstructured componentwise condition number is $\mathcal{O}(\varepsilon^{-1})$ and the structured componentwise condition number is $\mathcal{O}(1)$. This is true for the important case of componentwise relative perturbations in the matrix and in the right-hand side. We also prove corresponding estimations for circulant structures. Moreover, bounds for the condition number of matrix inversion are given. Finally, we give for all structures mentioned above explicit examples of parameterized (structured) matrices A_ε such that the (componentwise) condition number of matrix inversion is $\mathcal{O}(\varepsilon^{-1})$, but the componentwise distance to the nearest singular matrix is $\mathcal{O}(1)$. This is true for componentwise relative perturbations. It shows that, unlike the normwise case, there is no reciprocal proportionality between the componentwise condition number and the distance to the nearest singular matrix.

Key words. componentwise structured perturbations, condition number, distance to singularity

AMS subject classifications. 15A12, 65F35

PII. S0895479802405744

1. Motivation.

In the first part of this paper we investigated structured perturbations with respect to normwise distances. There is some drawback to that. For example, many matrices arising from some discretization are sparse. When using a normwise distance, system zeros may be altered by a perturbation into nonzero elements, which usually does not correspond to the underlying model.

System zeros can be modeled in the context of normwise distances such as, for example, symmetric tridiagonal or tridiagonal Toeplitz matrices (see section 7 of Part I of this paper). If components differ much in size, there is the problem that normwise distances alter small components relatively more often than larger components.

To overcome this difficulty a common approach is to use componentwise distances. Consider a linear system $Ax = b$. For some (structured) weight matrix E, structured perturbations $A + \Delta A$ with $|\Delta A| \le \varepsilon |E|$ are considered, where absolute value and comparison are to be understood componentwise. This offers much freedom. For example, for E being the matrix of all 1's, the inequality above is equivalent to $||\Delta A||_m \le \varepsilon$, where $||A||_m := \max |A_{ij}|$, so that there is a finite ratio between this (structured) componentwise condition number and the (structured) normwise condition number (as considered in Part I). That means, in a way, the componentwise approach includes the normwise.

But componentwise perturbations offer much more freedom. For example, for Toeplitz perturbations one need not change the structure when dealing with banded or triangular Toeplitz matrices; in fact, for the common case of componentwise relative

*Received by the editors April 17, 2002; accepted for publication (in revised form) by N. J. Higham January 13, 2003; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/40574.html
†Technical University of Hamburg-Harburg, Schwarzenbergstr. 95, 21071 Hamburg, Germany (rump@tu-harburg.de).

perturbations such a structure is preserved per se. Also, the sensitivity with respect to one, or a couple of, components fits easily into the concept. Therefore, there has been quite some interest in componentwise perturbations in recent years; cf. [13, 14, 2, 10, 20, 22].

However, this much freedom implies drastic consequences for the ratio between the structured and the unstructured condition numbers. It has been mentioned as an advantage in the application of componentwise perturbations that certain components may be excluded from perturbations by setting the corresponding weights to zero. But zero weights *change the structure* of the perturbations and they lower the degree of freedom of perturbations.

One of the most common weights is $E = |A|$, corresponding to componentwise relative perturbations in the matrix A. In this case, zero components of the matrix shrink the space of admissible perturbations. Consider, for example, symmetric Toeplitz perturbations. Then for a specific $n \times n$ symmetric Toeplitz matrix with two nonzero components in the first row, normwise distances allow n degrees of freedom, whereas for componentwise distances *the specific matrix* reduces the degrees of freedom to 2. On the other hand, if this is the given data and if the zeros in the matrix are intrinsic to the model, then there is no more freedom for the perturbation of the input data.

As a consequence there are examples where the structured condition number is near 1, whereas the unstructured condition number can be arbitrarily large. Surprisingly, this is even the case for symmetric linear systems and the case of componentwise relative perturbations of the matrix *and* the right-hand side. This fact does not create much hope that algorithms can be found at all that are stable with respect to componentwise (relative) perturbations. We add more comments about that in the last section.

2. Introduction and notation. Let nonsingular $A \in M_n(\mathbb{R})$ and $x, b \in \mathbb{R}^n$, $x \neq 0$, be given with $Ax = b$. The componentwise condition number of this linear system with respect to a weight matrix $E \in M_n(\mathbb{R})$ and a weight vector $f \in \mathbb{R}^n$ is defined by

$$\begin{aligned} \operatorname{cond}_{E,f}(A,x) := \lim_{\varepsilon \to 0} \sup \Big\{ \frac{\|\Delta x\|_\infty}{\varepsilon \|x\|_\infty} : \; & (A + \Delta A)(x + \Delta x) = b + \Delta b, \Delta A \in M_n(\mathbb{R}), \\ & \Delta b \in \mathbb{R}^n, |\Delta A| \leq \varepsilon |E|, \; |\Delta b| \leq \varepsilon |f| \Big\}. \end{aligned} \tag{2.1}$$

Note that the weights E, f may have negative entries, but only $|E|, |f|$ is used. Other definitions assume nonnegative weights beforehand. Usually this does not cause problems. In the following, however, we will use skewsymmetric E as well; therefore we choose the definition as in (2.1).

We use the same symbol $\| \cdot \|_\infty$ for the vector maximum norm and the matrix row sum norm. Here and throughout the paper *we use absolute value and comparison of vectors and matrices always componentwise.* For example, $|\Delta A| \leq \varepsilon |E|$ is equivalent to $|\Delta A_{ij}| \leq \varepsilon |E_{ij}|$ for all i, j. It is well known [16, Theorem 7.4] that

$$\operatorname{cond}_{E,f}(A,x) = \frac{\| \, |A^{-1}| \, |E| \, |x| + |A^{-1}| \, |f| \, \|_\infty}{\|x\|_\infty}. \tag{2.2}$$

This generalizes the Skeel condition number [24]

$$\operatorname{cond}_A(A,x) = \frac{\| \, |A^{-1}| \, |A| \, |x| \, \|_\infty}{\|x\|_\infty}, \tag{2.3}$$

where $E = A$ depicts componentwise relative perturbations in A, and the omitted f indicates that the right-hand side is unchanged. As usual we use $\|\cdot\|_\infty$ in case of componentwise perturbations, whereas the spectral norm $\|\cdot\|_2$ is used in case of normwise perturbations (see Part I of this paper).

For specific right-hand sides the normwise and componentwise condition number can be arbitrarily far apart. For instance, for the well-known example by Kahan [18]

$$A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & \varepsilon & \varepsilon \\ 1 & \varepsilon & \varepsilon \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} \varepsilon \\ -1 \\ 1 \end{pmatrix} \tag{2.4}$$

one computes for normwise and componentwise relative perturbations in the matrix and the right-hand side

$$\kappa_{|A|,|Ax|}(A,x) = 1.4\varepsilon^{-1} \quad \text{but} \quad \operatorname{cond}_{|A|,|Ax|}(A,x) = 2.5.$$

In case of linear systems with special matrices such as Toeplitz or band matrices, algorithms are known that are faster than a general linear system solver. For such a special solver only structured perturbations are possible, for example, Toeplitz or band. Therefore one may ask whether the sensitivity of the solution changes when restricting perturbations to structured perturbations.

Let $M_n^{\text{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ denote a set of matrices of a certain structure. In this paper we will focus on linear structures, namely

$$\text{struct} \in \{\text{sym}, \text{persym}, \text{skewsym}, \text{symToep}, \text{Toep}, \text{circ}, \text{Hankel}, \text{persymHankel}\}, \tag{2.5}$$

that is, symmetric, persymmetric, skewsymmetric, symmetric Toeplitz, general Toeplitz, circulant, Hankel, and persymmetric Hankel matrices. We define, similarly to (2.1), the *structured componentwise condition number* by

$$\begin{aligned} \operatorname{cond}_{E,f}^{\text{struct}}(A,x) := \lim_{\varepsilon\to 0} \sup \Big\{ \frac{\|\Delta x\|_\infty}{\varepsilon \|x\|_\infty} :\ & (A+\Delta A)(x+\Delta x) = b+\Delta b,\ \Delta A \in M_n^{\text{struct}}(\mathbb{R}), \\ & \Delta b \in \mathbb{R}^n, |\Delta A| \le \varepsilon |E|, |\Delta b| \le \varepsilon |f| \Big\}. \end{aligned} \tag{2.6}$$

We mention that for $A \in M_n^{\text{struct}}(\mathbb{R})$ and all structures under investigation $\Delta A \in M_n^{\text{struct}}(\mathbb{R})$ is equivalent to $A + \Delta A \in M_n^{\text{struct}}(\mathbb{R})$. Therefore it suffices to assume $\Delta A \in M_n^{\text{struct}}(\mathbb{R})$ in (2.6).

For a specialized solver, for example, for symmetric Toeplitz A, only symmetric Toeplitz perturbations of A are possible because only the first row of A is input to the algorithm. A considerable factor between the structured condition number (2.6) and the general, unstructured condition number (2.1) may shed light on the stability of an algorithm.

Indeed, there may be huge factors between (2.1) and (2.6). Let, for example,

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \ddots \\ & & \ddots & \ddots \end{pmatrix} \in M_n(\mathbb{R}) \quad \text{and} \quad x = (1,-1,1,-1,\ldots)^T \in \mathbb{R}^n.$$

Then for $n = 200$ we have

$$\mathrm{cond}_A(A,x) = 2.02\cdot 10^4, \quad \mathrm{cond}_A^{\mathrm{sym}}(A,x) = 1.02\cdot 10^4, \quad \text{and} \quad \mathrm{cond}_A^{\mathrm{symToep}}(A,x) = 1.$$

In this example no perturbations in the right-hand side are allowed. This does not conform with Wilkinson's classical ansatz of error analysis, where he relates the computational error to a perturbation in the input data. It may seem artificial to allow variations of some input data, namely the input matrix A, and not of others such as the right-hand side b. However, this depends on the point of view. The right-hand side may be given exactly where perturbations do not make sense, for example, when solving $Ax = e_1$ for computing the first column of A^{-1}, where e_1 denotes the first column of the identity matrix. For this problem, and also in case the problem is solved through $Ax = e_1$, the input of the problem is only A.

A numerical algorithm solves a nearby problem, and for the judgment of stability of an algorithm disregarding perturbations in some input data seems inadequate. However, in case of banded A, for example, the zeros outside the band are *not input* to a band solver, and therefore perturbations in those should *not* be taken into account. Among others, these are motivations for looking at componentwise perturbations.

However, we will see that things change significantly in the rugged world of componentwise perturbations compared to the smooth world of normwise perturbations. We are especially interested in the estimation of $\mathrm{cond}^{\mathrm{struct}}/\mathrm{cond}$, a question also posed in [13]. Even for handsome structures such as symmetric matrices there are examples where $\mathrm{cond}_{A,b}(A,x)$ is arbitrarily large compared to $\mathrm{cond}_{A,b}^{\mathrm{sym}}(A,x) \sim 1$. Note that this is true for the important case of componentwise relative perturbations in the matrix *and* in the right-hand side.

We will give similar examples for all structures in (2.5) except circulant matrices. In the latter case we give almost sharp estimations for $\mathrm{cond}^{\mathrm{struct}}/\mathrm{cond}$. As we will see, for circulant structures the ratio may only become small for ill-conditioned matrices. The worst case is about $\mathrm{cond}^{\mathrm{struct}} \sim \sqrt{\mathrm{cond}}$.

The examples mentioned above are valid for a specific solution x and corresponding right-hand side. The worst case structured condition number for componentwise relative perturbations, the supremum over all x, however, will be shown to be not far away from the corresponding unstructured condition number for all structures in (2.5).

Moreover, bounds for the condition number of matrix inversion are given. Finally, we give for all structures in (2.5) explicit examples of parameterized (structured) matrices A_ε such that the condition number of matrix inversion is $\mathcal{O}(\varepsilon^{-1})$, but the componentwise distance to the nearest singular matrix is $\mathcal{O}(1)$. This is again true for the important case of componentwise relative perturbations. It shows that, unlike in the normwise case, there is no reciprocal proportionality between the componentwise condition number and distance to the nearest singular matrix. Recall that for normwise perturbations the structured condition number is equal to the reciprocal of the (structured) distance to the nearest singular matrix (Part I, Theorem 12.1).

We will use the following notation:

$M_n(\mathbb{R})$	set of real $n \times n$ matrices
$M_n^{\mathrm{struct}}(\mathbb{R})$	set of structured real $n \times n$ matrices
$\|\cdot\|_\infty$	infinity or row sum norm
E	some (weight) matrix, $E \in M_n(\mathbb{R})$
f	some (weight) vector, $f \in \mathbb{R}^n$
I, I_n	identity matrix (with n rows and columns)
e	vector of all 1's, $e \in \mathbb{R}^n$

$(\mathbf{1})$	matrix of all 1's, $(\mathbf{1}) = ee^T \in M_n(\mathbb{R})$
S	signature matrix, i.e., $\lvert S\rvert = I$ or $S = \mathrm{diag}(\pm 1, \ldots, \pm 1)$
J, J_n	permutation matrix mapping $(1, \ldots, n)^T$ into $(n, \ldots, 1)^T$
$\sigma_{\min}(A)$	smallest singular value of A
$\lambda_{\min}(A)$	smallest eigenvalue of symmetric A

In this paper we treat explicitly the important (linear) structures in (2.5). However, we also derive formulas for general linear structures similar to those derived in [13]. We mention that this includes structures in the right-hand side by treating an augmented linear system of dimension $n+1$. Such structures appear, for example, in the Yule–Walker problem [11, section 4.7.2].

3. Componentwise perturbations. Throughout this paper let nonsingular $A \in M_n(\mathbb{R})$ be given together with $0 \neq x \in \mathbb{R}^n$ and weights $E \in M_n(\mathbb{R})$, $f \in \mathbb{R}^n$. Denote $b := Ax$.

The standard proof [16, Theorem 7.4] of (2.2) uses that $(A+\Delta A)(x+\Delta x) = b+\Delta b$ and $Ax = b$ imply

$$\Delta x = A^{-1}(-\Delta A x + \Delta b) + \mathcal{O}(\varepsilon^2). \tag{3.1}$$

This is true independent of ΔA, structured or not. It follows that

$$\mathrm{cond}_{E,f}^{\mathrm{struct}}(A,x) = \sup\left\{ \frac{\|A^{-1}\Delta A x + A^{-1}\Delta b\|_\infty}{\|x\|_\infty} : \Delta A \in M_n^{\mathrm{struct}}(\mathbb{R}), \Delta b \in \mathbb{R}^n, \ \lvert\Delta A\rvert \leq \lvert E\rvert, \lvert\Delta b\rvert \leq \lvert f\rvert \right\}. \tag{3.2}$$

This is again true for all structures including the unstructured case $M_n^{\mathrm{struct}}(\mathbb{R}) = M_n(\mathbb{R})$. For the estimation of $\|\Delta x\|_\infty$, in case of structured perturbations of ΔA we use the ansatz as in [13] (see also Part I of this paper). All structures in (2.5) are linear structures. That means for given "struct" every matrix ΔA in $M_n^{\mathrm{struct}}(\mathbb{R})$ depends linearly on some k parameters $\Delta p \in \mathbb{R}^k$. The number of parameters k depends on the structure; see Table 6.1 in Part I of this paper. Denote the vector of stacked columns of ΔA by $\mathrm{vec}(\Delta A) \in \mathbb{R}^{n^2}$. Then there is a bijective correspondence

$$\mathrm{vec}(\Delta A) = \Phi^{\mathrm{struct}} \cdot \Delta p \tag{3.3}$$

between $\mathrm{vec}(\Delta A)$ and the parameters Δp by some matrix $\Phi^{\mathrm{struct}} \in M_{n^2,k}(\mathbb{R})$. Note that Φ^{struct} is fixed for every structure and given size $n \in \mathbb{N}$. Also note that Φ^{struct} contains for all structures in (2.5) exactly one nonzero entry in each row.

In case of structured componentwise perturbations it seems natural to assume $E \in M_n^{\mathrm{struct}}(\mathbb{R})$. This implies existence of $p_E \in \mathbb{R}^k$ with $\mathrm{vec}(E) = \Phi^{\mathrm{struct}} \cdot p_E$, and because Φ^{struct} contains exactly one nonzero entry per row we have the nice equivalence

$$\Delta A \in M_n^{\mathrm{struct}}(\mathbb{R}) \quad \text{and} \quad \lvert\Delta A\rvert \leq \lvert E\rvert \Leftrightarrow \mathrm{vec}(\Delta A) = \Phi^{\mathrm{struct}} \cdot \Delta p \quad \text{and} \quad \lvert\Delta p\rvert \leq \lvert p_E\rvert. \tag{3.4}$$

That means the set of all $\lvert\Delta p\rvert \leq \lvert p_E\rvert$ maps one-to-one to the set of ΔA allowed in (3.2). In that respect structured componentwise perturbations are easier to handle than structured normwise perturbations. Finally, observe $\Delta A x = (x^T \otimes I)\Phi^{\mathrm{struct}}\Delta p$ for $\otimes$ denoting the Kronecker product, and with the abbreviation

$$\Psi_x^{\mathrm{struct}} := (x^T \otimes I)\Phi^{\mathrm{struct}} \tag{3.5}$$

we obtain the following formula for $\mathrm{cond}^{\mathrm{struct}}_{E,f}(A,x)$, which was also observed in [13].

THEOREM 3.1. *For nonsingular* $A \in M_n(\mathbb{R})$, $0 \neq x \in \mathbb{R}^n$, $E \in M_n^{\mathrm{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$, *and* $f \in \mathbb{R}^n$ *such that* $\mathrm{vec}(E) = \Phi^{\mathrm{struct}} p_E$ *for* $p_E \in \mathbb{R}^k$ *we have*

$$\mathrm{cond}^{\mathrm{struct}}_{E,f}(A,x) = \frac{\| \, |A^{-1}\Psi_x| \, |p_E| + |A^{-1}| \, |f| \, \|_\infty}{\|x\|_\infty}. \tag{3.6}$$

We note that Theorem 3.1 contains (2.2) for unstructured perturbations. In that case it is just $\Phi = I_{n^2}$ and $\Psi_x = x^T \otimes I_n$. Then $\mathrm{vec}(E) = p_E$ implies $|A^{-1}(x^T \otimes I)| \, |p_E| = |x^T \otimes A^{-1}| \, |p_E| = (|x^T| \otimes |A^{-1}|)|p_E| = |A^{-1}| \, |E| \, |x|$ using [17, Lemmas 4.2.10 and 4.3.1].

In Part I of this paper, we concluded in Corollary 6.6 of the corresponding Theorem 6.5 that a lower bound on the ratio of the structured and unstructured normwise condition number only depends on the solution vector x and not on the matrix A. This is not possible in the componentwise case. In the normwise case the factor $\|E\|$ cancelled, whereas in the componentwise case p_E may consist of components large in absolute value corresponding to columns of $|A^{-1}\Psi_x|$ being small in absolute value. This does indeed happen, as we will see in the explicit examples in the next sections.

For the structures in (2.5) the matrix Ψ_x is large but sparse. A mere count of operations shows that $A^{-1}\Psi_x$ requires not more than n^3 multiplications and additions for all structures in (2.5). Moreover, frequently it is not the exact value but rather an approximation of (3.6) that is sufficient. For that purpose efficient methods requiring some $\mathcal{O}(n^2)$ flops are available; see, for example, [12, 15].

To simplify and focus the discussion we observe that

$$\begin{gathered} A \in M_n^{\mathrm{sym}}(\mathbb{R}) \Leftrightarrow JA \in M_n^{\mathrm{persym}}(\mathbb{R}) \Leftrightarrow AJ \in M_n^{\mathrm{persym}}(\mathbb{R}), \\ A \in M_n^{\mathrm{symToep}}(\mathbb{R}) \Leftrightarrow JA \in M_n^{\mathrm{persymHankel}}(\mathbb{R}) \Leftrightarrow AJ \in M_n^{\mathrm{persymHankel}}(\mathbb{R}), \\ A \in M_n^{\mathrm{Toep}}(\mathbb{R}) \Leftrightarrow JA \in M_n^{\mathrm{Hankel}}(\mathbb{R}) \Leftrightarrow AJ \in M_n^{\mathrm{Hankel}}(\mathbb{R}). \end{gathered} \tag{3.7}$$

By rewriting (3.1) into

$$\Delta x = (JA)^{-1}(-J\Delta A x + J\Delta b) + \mathcal{O}(\varepsilon^2) \quad \text{and} \quad J\Delta x = (AJ)^{-1}(-\Delta AJ \cdot Jx + \Delta b) + \mathcal{O}(\varepsilon^2)$$

and observing $|\Delta A| \leq |E| \;\Leftrightarrow\; |J\Delta A| \leq |JE| \;\Leftrightarrow\; |\Delta AJ| \leq |EJ|$ and $|\Delta b| \leq |f| \;\Leftrightarrow\; |J\Delta b| \leq |Jf|$ we obtain the following.

THEOREM 3.2. *For nonsingular* $A \in M_n(\mathbb{R})$ *and* $0 \neq x \in \mathbb{R}^n$ *there holds*

$$\begin{gathered} \mathrm{cond}^{\mathrm{sym}}_{E,f}(A,x) = \mathrm{cond}^{\mathrm{persym}}_{JE,Jf}(JA,x) = \mathrm{cond}^{\mathrm{persym}}_{EJ,f}(AJ,Jx), \\ \mathrm{cond}^{\mathrm{symToep}}_{E,f}(A,x) = \mathrm{cond}^{\mathrm{persymHankel}}_{JE,Jf}(JA,x) = \mathrm{cond}^{\mathrm{persymHankel}}_{EJ,f}(AJ,Jx), \\ \mathrm{cond}^{\mathrm{Toep}}_{E,f}(A,x) = \mathrm{cond}^{\mathrm{Hankel}}_{JE,Jf}(JA,x) = \mathrm{cond}^{\mathrm{Hankel}}_{EJ,f}(AJ,Jx). \end{gathered}$$

Therefore we will focus our discussion on symmetric, symmetric Toeplitz, and Hankel matrices, and the results will, mutatis mutandis, be valid for persymmetric, persymmetric Hankel, and general Toeplitz matrices, respectively.

4. Condition number for general x. For the case of unstructured componentwise relative perturbations $E = A$ and $f = b$ it does not make much difference

whether perturbations in the right-hand side are allowed or not. Indeed, (2.2) and $|b| = |Ax| \le |A|\,|x|$ imply

$$\mathrm{cond}_A(A,x) \le \mathrm{cond}_{A,b}(A,x) \le 2\mathrm{cond}_A(A,x). \tag{4.1}$$

A similar estimation is valid for the condition number under normwise perturbations (see Part I, equation (4.1)). Since the normwise condition number $\kappa_E(A,x) = \|A^{-1}\|_2\|E\|_2$ without perturbations in the right-hand side does not depend on x, condition is an inherent property of the matrix—at least for unstructured *normwise* perturbations. This is no longer the case for *componentwise* perturbations. Consider

$$A := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1+\varepsilon \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{4.2}$$

Then (2.2) implies

$$\begin{aligned} \mathrm{cond}_A(A,x) &= 4\varepsilon^{-1} + \mathcal{O}(1) \quad \text{but} \quad \mathrm{cond}_A(A,y) = 1, \quad \text{and} \\ \mathrm{cond}_{A,b}(A,x) &= 8\varepsilon^{-1} + \mathcal{O}(1) \quad \text{but} \quad \mathrm{cond}_{A,b}(A,y) = 2. \end{aligned} \tag{4.3}$$

So condition subject to componentwise perturbations is no longer an intrinsic matrix property but depends on the solution x (and therefore on the right-hand side). Note that the norms of rows and columns of A in (4.2) are of similar size.

There are similar examples for structured perturbations. For instance, the same data (4.2) yield for symmetric perturbations

$$\mathrm{cond}^{\mathrm{sym}}_{A,b}(A,x) = 6\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \mathrm{cond}^{\mathrm{sym}}_{A,b}(A,y) = 2,$$

and there are similar examples for the other structures in (2.5).

We may ask what is the worst case condition number for all x. We first observe the following.

LEMMA 4.1. *For nonsingular $A \in M_n(\mathbb{R})$ and $M_n^{\mathrm{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ we have*

$$\sup_{x \ne 0} \mathrm{cond}^{\mathrm{struct}}_{E,f}(A,x) = \sup_{|x|=e} \mathrm{cond}^{\mathrm{struct}}_{E,f}(A,x). \tag{4.4}$$

Proof. In view of (3.2) the supremum over all $0 \ne x \in \mathbb{R}^n$ in (4.4) can obviously be replaced by the supremum over all $\|x\|_\infty = 1$, the same as $|x| \le e$ with at least one $|x_i| = 1$. The assertion follows easily. ∎

For unstructured perturbations, formula (2.2) and Lemma 4.1 imply

$$\sup_{x \ne 0} \mathrm{cond}_{E,f}(A,x) = \mathrm{cond}_{E,f}(A,e) = \| \, |A^{-1}|\,|E|e + |A^{-1}|\,|f| \, \|_\infty, \tag{4.5}$$

and for no perturbations in the right-hand side

$$\sup_{x \ne 0} \mathrm{cond}_E(A,x) = \mathrm{cond}_E(A,e) = \| \, |A^{-1}|\,|E| \, \|_\infty.$$

For some structures the supremum for structured perturbations (4.4) is equal to the worst case (4.5) for unstructured perturbations.

THEOREM 4.2. *Let $M_n^{\mathrm{struct}}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ and nonsingular $A \in M_n^{\mathrm{struct}}(\mathbb{R})$, $E \in M_n^{\mathrm{struct}}(\mathbb{R})$, and $f \in \mathbb{R}^n$ be given. If, for every signature matrix S, $B \in M_n^{\mathrm{struct}}(\mathbb{R})$ implies $SBS \in M_n^{\mathrm{struct}}(\mathbb{R})$, then*

$$\sup_{x \ne 0} \mathrm{cond}^{\mathrm{struct}}_{E,f}(A,x) = \sup_{x \ne 0} \mathrm{cond}_{E,f}(A,x) = \| \, |A^{-1}|\,|E|\,e + |A^{-1}|\,|f| \, \|_\infty. \tag{4.6}$$

Proof. Let i denote the row of $|A^{-1}|\,|E|\,e + |A^{-1}|\,|f|$ for which the maximum is achieved in the ∞-norm, and denote by S the signature matrix with $S_{\nu\nu} := sign(A^{-1})_{i\nu}$. Then

$$(A^{-1} \cdot S|E|S \cdot Se)_i = (|A^{-1}|\,|E|\,e)_i,$$

and the result follows by (3.2), choosing $\Delta A := S|E|S \in M_n^{\rm struct}(\mathbb{R})$ with $|\Delta A| = |E|$ and the obvious choice of Δb. ∎

COROLLARY 4.3. *For* struct $\in \{\mathrm{sym}, \mathrm{persym}, \mathrm{skewsym}\}$ *and nonsingular* $A \in M_n^{\rm struct}(\mathbb{R})$, $E \in M_n^{\rm struct}(\mathbb{R})$, $f \in \mathbb{R}^n$, *it follows that*

$$\begin{aligned}\sup_{x\neq 0} \mathrm{cond}_{E,f}^{\rm struct}(A,x) &= \sup_{x\neq 0} \mathrm{cond}_{E,f}(A,x) = \mathrm{cond}_{E,f}(A,e) \\ &= \| \, |A^{-1}|\,|E|\,e + |A^{-1}|\,|f| \, \|_\infty ,\end{aligned}$$

and therefore

$$\sup_{x\neq 0} \mathrm{cond}_{E}^{\rm struct}(A,x) = \| \, |A^{-1}|\,|E| \, \|_\infty$$

for no perturbations in the right-hand side.

Proof. For struct $\in \{\mathrm{sym}, \mathrm{skewsym}\}$ the result follows by Theorem 4.2, and for persymmetric matrices by Theorem 3.2. ∎

For other structures things change if the structure imposes too many restrictions on the choice of the elements. If not, the following theorem gives at least two-sided bounds for the worst case condition number for all x.

THEOREM 4.4. *Let* $M_n^{\rm struct}(\mathbb{R}) \subseteq M_n(\mathbb{R})$ *be given such that for every individual column there is no dependency between the elements; in other words, for every* $c \in \mathbb{R}^n$ *and every index* $i \in \{1, \ldots, n\}$ *there exists* $B \in M_n^{\rm struct}(\mathbb{R})$ *with the* i*th column* B_i *of* B *equal to* c. *For such* $M_n^{\rm struct}(\mathbb{R})$ *and given nonsingular* $A \in M_n^{\rm struct}(\mathbb{R})$, $E \in M_n^{\rm struct}(\mathbb{R})$, $f \in \mathbb{R}^n$, *it follows that*

$$n^{-1}\alpha \leq \sup_{x\neq 0} \mathrm{cond}_{E,f}^{\rm struct}(A,x) \leq \alpha, \tag{4.7}$$

where

$$\alpha := \mathrm{cond}_{E,f}(A,e) = \sup_{x\neq 0} \mathrm{cond}_{E,f}(A,x) = \| \, |A^{-1}|\,|E|\,e + |A^{-1}|\,|f| \, \|_\infty .$$

Estimation (4.7) *is especially true for* struct $\in \{\mathrm{circ}, \mathrm{Toep}, \mathrm{Hankel}\}$.

Proof. Denote by $i \in \{1, \ldots, n\}$ an index with

$$\alpha = \sup_{\|x\|_\infty = 1} \mathrm{cond}_{E,f}(A,x) = \| \, |A^{-1}|\,|E|\,e + |A^{-1}|\,|f| \, \|_\infty = (|A^{-1}|\,|E|\,e + |A^{-1}|\,|f| \,)_i .$$

There is $j \in \{1, \ldots, n\}$ with $(|A^{-1}|\,|E|\,e)_i \leq n(|A^{-1}|\,|E|)_{ij}$. Choose $\Delta A \in M_n^{\rm struct}(\mathbb{R})$ with $|\Delta A| \leq |E|$ such that $\Delta A_{\nu i} = sign((A^{-1})_{i\nu}) \cdot |E_{\nu j}|$. Then $(A^{-1}\Delta A)_{ij} = (|A^{-1}|\,|E|)_{ij}$, and for suitable x with $|x| = e$ and for suitable Δb we obtain

$$\begin{aligned}|(A^{-1}\Delta A x + A^{-1}\Delta b)_i| &\geq (|A^{-1}|\,|E|)_{ij} + (|A^{-1}|\,|f|)_i \\ &\geq n^{-1}(|A^{-1}|\,|E|\,e + |A^{-1}|\,|f|)_i = n^{-1}\alpha.\end{aligned}$$

Now the left inequality in (4.7) follows by (3.2), and the right inequality is obvious. ∎

The assumptions on $M_n^{\mathrm{struct}}(\mathbb{R})$ are also satisfied for struct $\in \{\mathrm{sym}, \mathrm{persym}, \mathrm{skewsym}\}$, where we already obtained the sharp result in Corollary 4.3. However, the assumptions are not satisfied for symmetric Toeplitz structures of dimension $n \geq 3$, and therefore also are not for persymmetric Hankel structures. Indeed, we will give general examples of symmetric Toeplitz matrices of dimension $n = 3$ and $n \geq 5$ such that

$$\mathrm{cond}_A(A, e) = \varepsilon^{-1} + \mathcal{O}(1) \quad \text{but} \quad \sup_{x \neq 0} \mathrm{cond}_A^{\mathrm{symToep}}(A, x) = 1 + \mathcal{O}(\varepsilon). \tag{4.8}$$

Note that we use the weight matrix $E = A$ but do not allow perturbations in the right-hand side. Of course, after introducing some small weight f for a perturbation in the right-hand side formula, (4.8) is still valid in weaker form.

Let us explore this example for $n = 3$ in more detail. Consider (for small $\alpha \in \mathbb{R}$)

$$A = \begin{pmatrix} 0 & 1 & \alpha \\ 1 & 0 & 1 \\ \alpha & 1 & 0 \end{pmatrix} \quad \text{such that} \quad A^{-1} = (2\alpha)^{-1} \begin{pmatrix} -1 & \alpha & 1 \\ \alpha & -\alpha^2 & \alpha \\ 1 & \alpha & -1 \end{pmatrix}. \tag{4.9}$$

General $\Delta A \in M_3^{\mathrm{symToep}}(\mathbb{R})$ with $|\Delta A| \leq |A|$ is of the form

$$\Delta A = \begin{pmatrix} 0 & a & \alpha b \\ a & 0 & a \\ \alpha b & a & 0 \end{pmatrix} \quad \text{with} \quad |a| \leq 1, |b| \leq 1.$$

Then

$$A^{-1}\Delta A = \frac{1}{2} \begin{pmatrix} a+b & 0 & a-b \\ \alpha(b-a) & 2a & \alpha(b-a) \\ a-b & 0 & a+b \end{pmatrix} \quad \text{such that} \quad \|A^{-1}\Delta A\|_\infty \leq 1 + \mathcal{O}(\alpha).$$

But

$$|A^{-1}|\,|A| = \begin{pmatrix} 1 & \alpha^{-1} & 1 \\ \alpha & 1 & \alpha \\ 1 & \alpha^{-1} & 1 \end{pmatrix} \quad \text{implies} \quad \| \, |A^{-1}|\,|A| \, \|_\infty = \alpha^{-1} + \mathcal{O}(1)$$

such that (3.2) and (2.2) imply

$$\sup_{x \neq 0} \mathrm{cond}_A^{\mathrm{symToep}}(A, x) \leq 1 + \mathcal{O}(\alpha) \quad \text{but} \quad \mathrm{cond}_A(A, e) = \alpha^{-1} + \mathcal{O}(1). \tag{4.10}$$

Note that $\mathrm{cond}_A(A, x) \sim \alpha^{-1}$ is true for all $x \in \mathbb{R}^n$ with $|x_2|$ not too small.

The situation as in (4.10) cannot happen if perturbations in the right-hand side are allowed, at least not for the important case of relative perturbations $E = A, f = b$. In that case the worst case condition number is of the order of $\| \, |A^{-1}|\,|A| \, \|_\infty$, as shown by the following theorem.

THEOREM 4.5. *Let arbitrary $M_n^{\mathrm{struct}}(\mathbb{R})$ be given and nonsingular $A \in M_n^{\mathrm{struct}}(\mathbb{R})$. Then for componentwise relative perturbations in the matrix and in the right-hand side, i.e., for $E = A$ and $f = Ax$, we have*

$$n^{-1} \| \, |A^{-1}|\,|A| \, \|_\infty \leq \sup_{x \neq 0} \mathrm{cond}_{A,Ax}^{\mathrm{struct}}(A, x) \leq 2 \| \, |A^{-1}|\,|A| \, \|_\infty.$$

Remark 4.6. Note that the weight $f = Ax$ for the right-hand side depends on x. This problem does not occur in the normwise case because in that case the worst case structured condition number (for all x) is equal to the unstructured condition number for *all* weights E, f (see Part I, Theorem 4.1).

Proof. On the one hand, (3.2) implies

$$\sup_{x\neq 0} \mathrm{cond}^{\mathrm{struct}}_{A,Ax}(A,x) \geq \sup_{|x|=e} \| \, |A^{-1}| \, |Ax| \, \|_\infty \geq n^{-1} \| \, |A^{-1}| \, |A| \, \|_\infty .$$

On the other hand, (3.2) and (2.2) yield

$$\begin{array}{rcl} \sup\limits_{x\neq 0} \mathrm{cond}^{\mathrm{struct}}_{A,Ax}(A,x) & \leq & \sup\limits_{x\neq 0} \mathrm{cond}_{A,Ax}(A,x) = \mathrm{cond}_{A,Ae}(A,e) \\ & = & \| \, |A^{-1}| \, |A|e + |A^{-1}| \, |Ae| \, \|_\infty \leq 2 \| \, |A^{-1}| \, |A| \, \|_\infty . \end{array} \qquad \square$$

We mention that

$$\| \, |A^{-1}| \, |A| \, \|_\infty = \inf_{D_1, D_2} \kappa_\infty(D_1 A D_2) = \varrho(|A^{-1}| \, |A|), \tag{4.11}$$

where ϱ denotes the spectral radius and the infimum is taken over nonsingular diagonal D_i. So this quantity is the infimum ∞-norm condition number with respect to unstructured and normwise perturbations in the matrix. The right equality in (4.11) was proved by Bauer [3] for the case where $|A|$ and $|A^{-1}|$ have positive entries. The proof gives D_1 and D_2 explicitly by using the right Perron vector of $|A^{-1}| \, |A|$. The argument is also valid for general A, as shown by [23].

The question remains of whether at least some of the previous results for the worst case structured condition number (for all x) can be shown for specific x, i.e., specific right-hand side. A worst case scenario in that respect would be if for the natural weights $E = A$ and $f = b$, i.e., componentwise relative perturbations in the matrix and the right-hand side, there exist A, b, and x with $\mathrm{cond}_{A,b}(A,x)$ arbitrarily large, whereas $\mathrm{cond}^{\mathrm{struct}}_{A,b}(A,x) = \mathcal{O}(1)$. Indeed, we will show that for all structures mentioned in (2.5) except circulants there are such general examples.

5. Symmetric, persymmetric, and skewsymmetric matrices. For the case of no perturbations in the right-hand side it is fairly easy to find parameterized $A = A_\varepsilon$ and x such that $\mathrm{cond}_A(A,x) = \mathcal{O}(\varepsilon^{-1})$ and $\mathrm{cond}^{\mathrm{struct}}_A(A,x) = \mathcal{O}(1)$ for $\mathrm{struct} \in \{\mathrm{sym}, \mathrm{persym}, \mathrm{skewsym}\}$. We found it more difficult to find such examples with perturbations in the right-hand side; in fact, we did not expect there to be any; however, they do exist. We illustrate the first example in more detail. Consider

$$A = A_\varepsilon = \begin{pmatrix} 0 & 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & -1 & 1 & \varepsilon & 1 \\ -1 & 0 & 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Then

$$A^{-1} = \varepsilon^{-1} \begin{pmatrix} 1 & -1+\frac{\varepsilon}{2} & 0 & 1 & -1-\frac{\varepsilon}{2} \\ -1+\frac{\varepsilon}{2} & 1-\frac{\varepsilon}{2} & \frac{\varepsilon}{2} & -1 & 1 \\ 0 & \frac{\varepsilon}{2} & 0 & 0 & \frac{\varepsilon}{2} \\ 1 & -1 & 0 & 1 & -1 \\ -1-\frac{\varepsilon}{2} & 1 & \frac{\varepsilon}{2} & -1 & 1+\frac{\varepsilon}{2} \end{pmatrix}.$$

Furthermore,

$$|A^{-1}|\,|A|\,|x| = \varepsilon^{-1}\begin{pmatrix} 8+3\varepsilon \\ 8+5\varepsilon \\ 2\varepsilon \\ 8+\varepsilon \\ 8+5\varepsilon \end{pmatrix} \quad \text{and} \quad |A^{-1}|\,|Ax| = |A^{-1}|\begin{pmatrix} 0 \\ 0 \\ 4 \\ \varepsilon \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 0 \\ 1 \\ 3 \end{pmatrix}. \tag{5.1}$$

Note that $|A^{-1}|\,|Ax|$ is of size $\mathcal{O}(1)$ because the third column of A^{-1}, which meets the component 4 in $|Ax|$, is of size $\mathcal{O}(\varepsilon)$. This is important because this term $|A^{-1}|\,|Ax|$ occurs in both the unstructured condition number (2.2) *and* the structured condition number (cf. Theorem 3.1). Now (2.2) implies

$$\operatorname{cond}_{A,Ax}(A,x) = 8\varepsilon^{-1} + \mathcal{O}(1). \tag{5.2}$$

On the other hand, according to (3.5),

$$\Psi_x^{\mathrm{sym}} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

such that $A^{-1}\Psi_x^{\mathrm{sym}}$ has large elements of size $\mathcal{O}(\varepsilon^{-1})$ in columns 1, 4, 6, 9, 13, and 15, whereas all other columns are comprised of elements of magnitude $\mathcal{O}(1)$. However, the parameter vector p_A such that $\operatorname{vec}(A) = \Phi^{\mathrm{sym}} \cdot p_A$ has zero elements in components 1, 4, 6, 9, 10, 15, a value ε in component 13, and ± 1's otherwise. Therefore

$$|A^{-1}\Psi_x^{\mathrm{sym}}| \cdot |p_A| = \begin{pmatrix} 3 \\ 4 \\ 2 \\ 1 \\ 4 \end{pmatrix} + \mathcal{O}(\varepsilon)$$

such that (5.1) and Theorem 3.1 imply

$$\operatorname{cond}^{\mathrm{sym}}_{A,Ax}(A,x) = 7 + \mathcal{O}(\varepsilon). \tag{5.3}$$

The numbers in (5.2) and (5.3) do not change when replacing A by $A \oplus B$ and prolonging x by k zeros, where $B \in M_k(\mathbb{R})$ denotes any symmetric matrix. Furthermore, Theorem 3.2 implies that the same example applies for persymmetric structures. We proved the following result.

THEOREM 5.1. *For $n \geq 5$, there exist parameterized symmetric $A := A_\varepsilon \in M_n^{\mathrm{sym}}(\mathbb{R})$ and $x \in \mathbb{R}^n$ such that*

$$\operatorname{cond}_{A,Ax}(A,x) = 8\varepsilon^{-1} + \mathcal{O}(1) \quad \textit{and} \quad \operatorname{cond}^{\mathrm{sym}}_{A,Ax}(A,x) = 7 + \mathcal{O}(\varepsilon).$$

For the persymmetric matrix $JA \in M_n^{\mathrm{persym}}(\mathbb{R})$ similar assertions are true.

For the skewsymmetric matrix $A = A_\varepsilon \in M_n^{\mathrm{skewsym}}(\mathbb{R})$ with

$$A := \begin{pmatrix} 0 & 1 & 0 & -1 \\ -1 & 0 & \varepsilon & 1 \\ 0 & -\varepsilon & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

one computes using (2.2) and Theorem 3.1

$$\text{(5.4)}\qquad \operatorname{cond}_{A,Ax}(A,x) = 6\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \operatorname{cond}^{\text{skewsym}}_{A,Ax}(A,x) = 4 + \mathcal{O}(\varepsilon).$$

Skewsymmetric nonsingular matrices are of even dimension. So replacing A by $A \oplus B$ and prolonging x by $2k$ zeros, where $B \in M_k^{\text{skewsym}}(\mathbb{R})$ denotes any skewsymmetric matrix, does not change the numbers in (5.4). We have the following result.

THEOREM 5.2. *For even $n \geq 4$, there exist parameterized skewsymmetric $A := A_\varepsilon \in M_n^{\text{skewsym}}(\mathbb{R})$ and $x \in \mathbb{R}^n$ such that*

$$\operatorname{cond}_{A,Ax}(A,x) = 6\varepsilon^{-1} + \mathcal{O}(1) \quad \textit{and} \quad \operatorname{cond}^{\text{skewsym}}_{A,Ax}(A,x) = 4 + \mathcal{O}(\varepsilon).$$

This result shows a major difference between normwise and componentwise perturbations. In Part I, Theorem 5.3 we proved that for struct $\in \{\text{sym}, \text{persym}, \text{skewsym}\}$ and for *all* x the structured normwise condition number is *equal* to the unstructured normwise condition number. For componentwise perturbations and *specific* x the condition numbers can be arbitrarily far apart, although Corollary 4.3 shows that in the worst case they are identical. We had similar results for normwise perturbations for the other structures in (2.5) (Part I, Theorems 8.4, 9.2, and 10.2). However, the worst case was essentially $\kappa^{\text{struct}} \sim \kappa^{1/2}$; i.e., a big ratio $\kappa/\kappa^{\text{struct}}$ was only possible for ill-conditioned matrices. For componentwise perturbations, $\operatorname{cond}^{\text{struct}}_{A,Ax} \sim 1$ is possible compared to arbitrarily large $\operatorname{cond}_{A,Ax}$—always for the important case of componentwise relative perturbations in the matrix and the right-hand side.

6. Toeplitz and Hankel matrices. Symmetric Toeplitz matrices depend only on n parameters. That makes it less difficult to find examples in the spirit of the previous section. Consider

$$A = A_\varepsilon := \text{Toeplitz}(0, 0, 1+\varepsilon, -1, 1) \quad \text{and} \quad x = (1,1,0,1,1)^T$$

such that A is the symmetric Toeplitz matrix with first row $[0, 0, 1+\varepsilon, -1, 1]$. Then

$$\operatorname{cond}_{A,Ax}(A,x) = 4\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \operatorname{cond}^{\text{symToep}}_{A,Ax}(A,x) = 5 + \mathcal{O}(\varepsilon).$$

In case of Toeplitz structures it is a little more subtle to find general $n \times n$ examples. The structure does not permit us to use a simple direct sum as in the previous section. We found the following examples. For even order greater than or equal to 6 consider

$$\begin{aligned} &A = \text{Toeplitz}(0, 0, z, -1, 1, z, 1, \varepsilon) \in M_{6+2k}^{\text{symToep}}(\mathbb{R}) \quad \text{and} \\ &x = (1, 0, z, -1, -1, z, 0, 1) \in \mathbb{R}^{6+2k}, \end{aligned}$$

where $z \in \mathbb{R}^k$ denotes a vector of $k \geq 0$ zeros. Then (2.2) and Theorem 3.1 yield

$$\operatorname{cond}_{A,Ax}(A,x) = 4\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \operatorname{cond}^{\text{symToep}}_{A,Ax}(A,x) = 6 + \mathcal{O}(\varepsilon)$$

for all $k \geq 0$. For odd order greater than or equal to 7 consider

$$\begin{aligned} A &= \text{Toeplitz}(\varepsilon, -\varepsilon, 0, 0, 0, 0, z, 1, z) \in M_{7+2k}^{\text{symToep}}(\mathbb{R}) \quad \text{and} \\ x &= (z, 0, 1, 1, 0, 1, 1, 0, z) \in \mathbb{R}^{7+2k}, \end{aligned}$$

where $z \in \mathbb{R}^k$ denotes again a vector of k zeros. Then

$$\begin{aligned}
\operatorname{cond}_{A,Ax}(A,x) &= 4\varepsilon^{-1} + \mathcal{O}(1) && \text{for} \quad n = 7,\\
\operatorname{cond}_{A,Ax}(A,x) &= 4\varepsilon^{-2} + \mathcal{O}(\varepsilon^{-1}) && \text{for} \quad \text{odd } n \geq 9 \quad \text{but}\\
\operatorname{cond}^{\text{symToep}}_{A,Ax}(A,x) &= 5 + \mathcal{O}(\varepsilon) && \text{for} \quad \text{odd } n \geq 7.
\end{aligned}$$

For persymmetric Hankel structures we use Theorem 3.2 and, summarizing, we have the following result.

THEOREM 6.1. *For $n \geq 5$, there exist parameterized symmetric Toeplitz matrices $A := A_\varepsilon \in M_n^{\text{symToep}}(\mathbb{R})$ and $x \in \mathbb{R}^n$ such that*

$$\operatorname{cond}_{A,Ax}(A,x) \geq 4\varepsilon^{-1} + \mathcal{O}(1) \quad \textit{and} \quad \operatorname{cond}^{\text{symToep}}_{A,Ax}(A,x) \leq 6 + \mathcal{O}(\varepsilon).$$

For the persymmetric Hankel matrix $JA \in M_n^{\text{persymHankel}}(\mathbb{R})$ similar assertions are true.

For Hankel structures consider

$$\begin{aligned}
A &= \text{Hankel}([0,\varepsilon,-1+\varepsilon,-1,0],[0,1,1,0,0]) \in M_5^{\text{Hankel}}(\mathbb{R}) \quad \text{and}\\
x &= (1,1,0,1,1)^T \in \mathbb{R}^5,
\end{aligned}$$

where $\text{Hankel}(c,r)$ denotes the Hankel matrix with first column c and last row r. Then (2.2) and Theorem 3.1 give

$$\operatorname{cond}_{A,Ax}(A,x) = 8\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \operatorname{cond}^{\text{Hankel}}_{A,Ax}(A,x) = 8 + \mathcal{O}(\varepsilon). \tag{6.1}$$

For general even $n \geq 6$ consider

$$\begin{aligned}
A &= \text{Hankel}([\varepsilon,1,z,1,-1,z,0,0],[0,0,z,-1,1,z,1,0]) \in M_{6+2k}^{\text{Hankel}}(\mathbb{R}),\\
x &= (1,0,z,-1,-1,z,0,1)^T \in \mathbb{R}^{6+2k},
\end{aligned}$$

where z denotes a vector of $k \geq 0$ zeros. Then

$$\operatorname{cond}_{A,Ax}(A,x) = 8\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \operatorname{cond}^{\text{Hankel}}_{A,Ax}(A,x) = 7 + \mathcal{O}(\varepsilon)$$

for all even $n \geq 6$. For general odd $n \geq 7$ define

$$\begin{aligned}
A &= \text{Hankel}([\varepsilon,z,0,-1,-1,0,z],[z,0,1-\varepsilon,1,0,0,z]) \in M_{5+2k}^{\text{Hankel}}(\mathbb{R}) \quad \text{and}\\
x &= (1,1,z,0,z,1,1)^T \in \mathbb{R}^{5+2k},
\end{aligned}$$

where z denotes a vector of $k \geq 1$ zeros. Then (6.1) is valid as well. Using Theorem 3.2 for general Toeplitz structures we have the following result.

THEOREM 6.2. *For $n \geq 5$, there exist parameterized Hankel matrices $A := A_\varepsilon \in M_n^{\text{Hankel}}(\mathbb{R})$ and $x \in \mathbb{R}^n$ such that*

$$\operatorname{cond}_{A,Ax}(A,x) = 8\varepsilon^{-1} + \mathcal{O}(1) \quad \textit{and} \quad \operatorname{cond}^{\text{Hankel}}_{A,Ax}(A,x) \leq 8 + \mathcal{O}(\varepsilon).$$

For the general Toeplitz matrix $JA \in M^{\text{Toep}}(\mathbb{R})$ similar assertions are true.

In summary, for all of the structures $\text{struct} \in \{\text{sym}, \text{persym}, \text{skewsym}, \text{symToep}, \text{Toep}, \text{Hankel}, \text{persymHankel}\}$ there are general $n \times n$ examples, $n \geq 5$, such that the unstructured condition number is arbitrarily large, whereas the structured condition

number is $\mathcal{O}(1)$. Note that this includes perturbations in the right-hand side. The only exception of the structures in (2.5) to this statement are circulant structures, as we will see in the next section.

For no perturbations in the right-hand side things are even worse. Consider

$$\begin{array}{rcl} A := A_\varepsilon = \mathrm{Toeplitz}(\varepsilon, v, 1, v, 0) & \text{for odd } n \geq 3 & \text{and} \\ A := A_\varepsilon = \mathrm{Toeplitz}(\varepsilon, w, 1, w, 0, 0) & \text{for even } n \geq 6, & \end{array} \tag{6.2}$$

where $v \in \mathbb{R}^{k-1}$ and $w \in \mathbb{R}^{k-2}$ denote zero vectors for $k := \lfloor n/2 \rfloor$. We will show that for these matrices a linear system $Ax = b$ is *always well conditioned* with respect to componentwise symmetric Toeplitz perturbations, that is, *for all* x. On the other hand, the linear system is *ill conditioned* for generic x with respect to componentwise general perturbations. We illustrate the proof for $n = 5$. Let $x \in \mathbb{R}^5$ with $\|x\|_\infty = 1$ be given. According to (3.5) and (3.6) we calculate

$$A^{-1} = \frac{1}{2} \begin{pmatrix} \varepsilon^{-1} & 0 & 1 & 0 & -\varepsilon^{-1} \\ 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 \\ -\varepsilon^{-1} & 0 & 1 & 0 & \varepsilon^{-1} \end{pmatrix} + \mathcal{O}(\varepsilon) \quad \text{and}$$

$$\Psi_x^{\mathrm{symToep}} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ x_2 & x_1 + x_3 & x_4 & x_5 & 0 \\ x_3 & x_2 + x_4 & x_1 + x_5 & 0 & 0 \\ x_4 & x_3 + x_5 & x_2 & x_1 & 0 \\ x_5 & x_4 & x_3 & x_2 & x_1 \end{pmatrix},$$

and from this

$$A^{-1}\Psi_x = \frac{1}{2}\varepsilon^{-1} \begin{pmatrix} x_1 - x_5 & x_2 - x_4 & 0 & x_4 - x_2 & x_5 - x_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ x_5 - x_1 & x_4 - x_2 & 0 & x_2 - x_4 & x_1 - x_5 \end{pmatrix} + \mathcal{O}(1).$$

By construction (3.3) we have $\mathrm{vec}(A) = \Phi^{\mathrm{symToep}} \cdot p_A$ with $p_A = (\varepsilon, 0, 1, 0, 0)^T$. The only element of p_A of size 1 meets the zero column in $A^{-1}\Psi_x$, so $|A^{-1}\Psi_x|\,|p_A| = \mathcal{O}(1)$, and by (3.6)

$$\mathrm{cond}_A^{\mathrm{symToep}}(A, x) = \mathcal{O}(1) \quad \text{for all } 0 \neq x \in \mathbb{R}^n.$$

By (3.6) this remains true without the assumption $\|x\|_\infty = 1$. On the other hand,

$$|A^{-1}|\,|A| = \begin{pmatrix} 1 & 0 & \varepsilon^{-1} & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & \varepsilon^{-1} & 0 & 1 \end{pmatrix} + \mathcal{O}(\mathrm{eps}).$$

So (2.2) implies

$$\mathrm{cond}_A(A, x) \geq \varepsilon^{-1}|x_3|.$$

The computation above extends to all matrices in (6.2) and we have the following result.

THEOREM 6.3. *Let $\varepsilon > 0$ and a matrix $A := A_\varepsilon$ according to* (6.2) *be given. Then the following are true:*

(i) *For all $0 \neq x \in \mathbb{R}^n$ we have $\mathrm{cond}_A^{\mathrm{symToep}}(A, x) = \mathcal{O}(1)$.*

(ii) *Let $x \in \mathbb{R}^n$ be given and denote $\alpha := |x_{k+1}|$ for odd n and $\alpha := \max(|x_k|, |x_{k+1}|)$ for even n, where $k = \lfloor n/2 \rfloor$. Then*

$$\mathrm{cond}_A(A, x) \geq \varepsilon^{-1}\alpha.$$

The example is possible because the symmetric Toeplitz structure imposes severe restrictions on the possible perturbations so that the assumptions of Theorem 4.4 are not satisfied.

7. Circulant matrices. Better estimations of the ratio $\mathrm{cond}^{\mathrm{circ}}/\mathrm{cond}$ of the componentwise condition numbers are possible because circulant matrices commute (because they are diagonalized by the Fourier matrix; cf. [7, 11]). This implies for $A, \Delta A \in M_n^{\mathrm{circ}}(\mathbb{R})$ that $A^{-1}\Delta A = \Delta A \cdot A^{-1}$ and therefore

$$\Delta x = -\Delta A \cdot A^{-1}x + A^{-1}\Delta b + \mathcal{O}(\varepsilon^2).$$

This implies the following nice characterization.

THEOREM 7.1. *Let nonsingular $A \in M_n^{\mathrm{circ}}(\mathbb{R})$, $x \in \mathbb{R}^n$, $E \in M_n^{\mathrm{circ}}(\mathbb{R})$, $f \in \mathbb{R}^n$ be given. Then*

$$\mathrm{cond}_{E,f}^{\mathrm{circ}}(A, x) = \frac{\| \, |E| \, |A^{-1}x| + |A^{-1}| \, |f| \, \|_\infty}{\|x\|_\infty}.$$

To estimate the ratio $\mathrm{cond}^{\mathrm{circ}}/\mathrm{cond}$ we first show that

$$\| \, |A| \, |x| \, \|_\infty \geq n^{-1}\|A\|_\infty \|x\|_2 \quad \text{for} \quad A \in M_n^{\mathrm{circ}}(\mathbb{R}). \tag{7.1}$$

For $A \in M_n^{\mathrm{circ}}(\mathbb{R})$ we have $\sum_i |A_{ij}| = \|A\|_1$ for any j and $\|A\|_1 = \|A\|_\infty$, and therefore

$$\begin{aligned} \| \, |A| \, |x| \, \|_\infty &\geq n^{-1} \sum_i (|A| \, |x|)_i = n^{-1} \sum_i \sum_j |A_{ij}| \, |x_j| = n^{-1} \sum_j \|A\|_1 |x_j| \\ &= n^{-1}\|A\|_1 \|x\|_1 \geq n^{-1}\|A\|_\infty \|x\|_2. \end{aligned}$$

This implies

$$\begin{aligned} \| \, |A| \, |A^{-1}x| \, \|_\infty &\geq n^{-1}\|A\|_\infty \|A^{-1}x\|_2 \quad \text{and} \\ \| \, |A^{-1}| \, |Ax| \, \|_\infty &\geq n^{-1}\|A^{-1}\|_\infty \|Ax\|_2. \end{aligned}$$

For $x \in \mathbb{R}^n$ we have

$$\|x\|_2^2 = x^T A^{-1} A x \leq \|x^T A^{-1}\|_2 \|Ax\|_2 = \|A^{-1}x\|_2 \|Ax\|_2$$

using $\|C^T x\|_2 = \|Cx\|_2$ for $C \in M_n^{\mathrm{circ}}(\mathbb{R})$; see Part I, Lemma 8.2. Putting things together we obtain for relative perturbations $E = A$ and $f = Ax$

$$\begin{aligned} \mathrm{cond}_{A,Ax}^{\mathrm{circ}}(A, x) &= \| \, |A| \, |A^{-1}x| + |A^{-1}| \, |Ax| \, \|_\infty / \|x\|_\infty \\ &\geq n^{-1} \max(\|A\|_\infty \|A^{-1}x\|_2, \|A^{-1}\|_\infty \|Ax\|_2) / \|x\|_\infty \\ &\geq n^{-1} \sqrt{\|A\|_\infty \|A^{-1}\|_\infty \|A^{-1}x\|_2 \|Ax\|_2} / \|x\|_\infty \\ &\geq n^{-1} \sqrt{\|A\|_\infty \|A^{-1}\|_\infty} \, \frac{\|x\|_2}{\|x\|_\infty}. \end{aligned} \tag{7.2}$$

With this and (4.1) and (2.3) we also obtain a lower bound on the ratio $\mathrm{cond}^{\mathrm{circ}}_{A,Ax}/\mathrm{cond}_{A,Ax}$ by

$$\begin{aligned}\mathrm{cond}^{\mathrm{circ}}_{A,Ax}(A,x) &\geq n^{-1}\sqrt{\| \, |A^{-1}| \, |A| \, \|_\infty}\,\frac{\|x\|_2}{\|x\|_\infty} \\ &\geq n^{-1}\sqrt{\frac{\| \, |A^{-1}| \, |A|x\|_\infty}{\|x\|_\infty}}\cdot\frac{\|x\|_2}{\|x\|_\infty} \\ &\geq n^{-1}\sqrt{\frac{1}{2}\mathrm{cond}_{A,Ax}(A,x)}\cdot\frac{\|x\|_2}{\|x\|_\infty} \\ &\geq 2^{-1/2}n^{-1}\sqrt{\mathrm{cond}_{A,Ax}(A,x)}.\end{aligned} \tag{7.3}$$

We have the following result.

THEOREM 7.2. *Let a nonsingular circulant $A \in M_n^{\mathrm{circ}}(\mathbb{R})$ and $0 \neq x \in \mathbb{R}^n$ be given. Then*

$$\begin{aligned}\mathrm{cond}^{\mathrm{circ}}_{A,Ax}(A,x) &\geq n^{-1}\sqrt{\|A\|_\infty\|A^{-1}\|_\infty}\cdot\frac{\|x\|_2}{\|x\|_\infty} \\ &\geq 2^{-1/2}n^{-1}\sqrt{\mathrm{cond}_{A,Ax}(A,x)}\cdot\frac{\|x\|_2}{\|x\|_\infty}.\end{aligned}$$

We think that the factor n^{-1} in both lower bounds of Theorem 7.2 can be replaced by the factor $n^{-1/2}$. If this is true, it is easy to find examples verifying that the overestimation in either case is bounded by a small constant factor.

The ratio $\mathrm{cond}^{\mathrm{circ}}_{A,Ax}/\mathrm{cond}_{A,Ax}$ can only become large for ill-conditioned linear systems. The question remains of whether this changes if we forbid perturbations in the right-hand side. This is indeed the case, and it is very simple to find examples. Just take a vector $r \in \mathbb{R}^n$ with uniformly distributed random first $n-1$ components in $[-1,1]$ and set $r_n := -\sum_{i=1}^{n-1} r_i + \varepsilon$. Then define $A = [\mathrm{circ}(r)]^{-1}$ and $x = e$. Obviously $\mathrm{circ}(r)e = \varepsilon e$, so A has an eigenvalue ε^{-1} to the eigenvector e; one can see that most likely $A > 0$, so Theorem 7.1 implies $\mathrm{cond}^{\mathrm{circ}}_A(A,e) = \| \, |A| \, |A^{-1}e| \, \|_\infty = 1$. On the other hand, (2.3) implies $\mathrm{cond}_A(A,e) = \| \, |A^{-1}| \, |A|e\|_\infty = \| \, |A^{-1}| \, |A| \, \|_\infty$, and extensive numerical experience shows that it is likely that $\mathrm{cond}_A(A,x) \sim \varepsilon^{-1}$. An explicit example is a matrix A constructed as above with $r_1 = \cdots = r_{n-1} = 1$. Then

$$\mathrm{cond}_A(A,e) = (2n-2)\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad \mathrm{cond}^{\mathrm{circ}}_A(A,e) = 1$$

for $n \geq 2$.

THEOREM 7.3. *Given $n \geq 2$, there exists $A \in M_n^{\mathrm{circ}}(\mathbb{R})$ with*

$$\mathrm{cond}_A(A,e) \geq \mathcal{O}(\varepsilon^{-1}) \quad \textit{and} \quad \mathrm{cond}^{\mathrm{circ}}_A(A,e) = 1.$$

The results show that with respect to normwise and componentwise perturbations circulants behave similarly (Part I, Theorems 8.1, 8.4, and equation (8.1)). Besides normality, a reason for that is that circulants commute.

8. Inversion of structured matrices. Similar to the structured componentwise condition number for linear systems, the structured componentwise condition number for matrix inversion is defined for $A \in M_n^{\mathrm{struct}}(\mathbb{R})$ and given weight matrix $E \in M_n^{\mathrm{struct}}(\mathbb{R})$ by

$$\mu^{\mathrm{struct}}_E(A) := \lim_{\varepsilon\to 0}\sup\left\{\frac{\|(A+\Delta A)^{-1} - A^{-1}\|_\infty}{\varepsilon\|A^{-1}\|_\infty};\quad \Delta A \in M_n^{\mathrm{struct}}(E),\ |\Delta A| \leq \varepsilon|E|\right\}. \tag{8.1}$$

The unstructured condition number $\mu_E(A)$, that is, for $M_n^{\rm struct}(\mathbb{R}) = M_n(\mathbb{R})$, satisfies the following bounds:

$$n^{-1}\alpha \le \mu_E(A) \le \alpha \quad \text{for} \quad \alpha := \frac{\| \, |A^{-1}| \, |E| \, |A^{-1}| \, \|_\infty}{\|A^{-1}\|_\infty}. \tag{8.2}$$

This follows by the well-known ansatz (see, for example, [16, proof of Theorem 6.4])

$$(A+\Delta A)^{-1} - A^{-1} = -A^{-1}\Delta A A^{-1} + \mathcal{O}(\|\Delta A\|^2). \tag{8.3}$$

From this the right inequality in (8.2) is obvious. Denoting the ith row and jth column of A^{-1} by $A^{-1}_{i,:}$ and $A^{-1}_{:,j}$, respectively, we have

$$(A^{-1}\Delta A A^{-1})_{ij} = (|A^{-1}| \, |E| \, |A^{-1}|)_{ij} \quad \text{for} \quad \Delta A := \operatorname{diag}(sign(A^{-1}_{i,:}))|E| \operatorname{diag}(sign(A^{-1}_{:,j})), \tag{8.4}$$

which implies the left inequality of (8.2).

In case of normwise perturbations the condition numbers for matrix inversion and for an arbitrary linear system with the same matrix (for no perturbations in the right-hand side) are both equal to $\|A^{-1}E\|_2$. In case of componentwise perturbations the condition number depends on the solution (see (4.2) and (4.3)). We may ask whether there is a relation between $\mu_E(A)$ and the supremum of $\operatorname{cond}_E(A,x)$ over all x.

DEFINITION 8.1. *Let nonsingular $A \in M_n^{\rm struct}(\mathbb{R})$ and $E \in M_n^{\rm struct}(\mathbb{R})$ be given. Then*

$$\operatorname{cond}_E^{\rm struct}(A) := \sup_{x \ne 0} \operatorname{cond}_E^{\rm struct}(A,x).$$

In Corollary 4.3 we saw

$$\operatorname{cond}_E(A) = \operatorname{cond}_E^{\rm struct}(A) = \| \, |A^{-1}|E| \, \|_\infty \quad \text{for struct} \in \{\text{sym}, \text{persym}, \text{skewsym}\}. \tag{8.5}$$

Obviously (8.2) implies

$$\mu_E(A) \le \operatorname{cond}_E(A).$$

However, this inequality may be arbitrarily weak. Consider

$$A = A_\varepsilon = \begin{pmatrix} \varepsilon & 1 & 0 \\ 1 & \varepsilon & 1 \\ 0 & 1 & \varepsilon \end{pmatrix} \quad \text{with} \quad \frac{\| \, |A^{-1}| \, |A| \, |A^{-1}| \, \|_\infty}{\|A^{-1}\|_\infty} = 3, \tag{8.6}$$

$$\text{but} \quad \operatorname{cond}_A(A,e) = \| \, |A^{-1}| \, |A| \, \|_\infty = \varepsilon^{-1}.$$

Note that this is for componentwise relative perturbations, i.e., $E = A$. Denote $b := Ae$. Then (8.6) implies that the linear system $Ax = b$ is ill conditioned for small ε, but matrix inversion of A is well conditioned for every $\varepsilon > 0$. This might lead to the apparent contradiction that solving the linear system by $x = A^{-1}b$ removes the ill-conditionedness. This is of course not the case. In our example we have

$$A^{-1} = (2\varepsilon)^{-1} \begin{pmatrix} 1 & \varepsilon & -1 \\ \varepsilon & 0 & \varepsilon \\ -1 & \varepsilon & 1 \end{pmatrix} + \mathcal{O}(1),$$

so an $\mathcal{O}(1)$ change in A^{-1} is a small perturbation. However, $b = Ae = (1+\varepsilon, 2+\varepsilon, 1+\varepsilon)^T$, so an $\mathcal{O}(1)$ change in A^{-1} causes an $\mathcal{O}(1)$ perturbation in $x = A^{-1}b = e$, which is of the order of 100% change.

The condition number $\mu_E(A)$ depends on diagonal scaling of A (and E). We may ask for the optimal condition number with respect to two-sided diagonal scaling. For this we obtain the following result.

THEOREM 8.2. *Let nonsingular $A \in M_n(\mathbb{R})$ and $E \in M_n(\mathbb{R})$ be given. Denote*

$$\mu_E^{\rm opt}(A) := \inf_{D_1, D_2} \mu_{D_1 E D_2}(D_1 A D_2), \tag{8.7}$$

where the infimum is taken over nonsingular diagonal matrices. Define

$$r := \min_{i,j} \frac{(|A^{-1}|\,|E|\,|A^{-1}|)_{ij}}{|A^{-1}|_{ij}}, \tag{8.8}$$

where $\alpha/0 := \infty$ for $\alpha \geq 0$. Then

$$n^{-1} r \leq \mu_E^{\rm opt}(A) \leq r. \tag{8.9}$$

Proof. Let i, j be indices realizing the minimum in the definition (8.8) of r and let $D^{(\nu)} := {\rm diag}(\varepsilon, \ldots, \varepsilon, 1, \varepsilon, \ldots, \varepsilon)$ with the 1 at the νth position. Defining $D_1^{-1} := D^{(j)}$ and $D_2^{-1} := D^{(i)}$ we obtain by (8.2)

$$\begin{aligned} \mu_E^{\rm opt}(A) \leq \mu_{D_1 E D_2}(D_1 A D_2) &\leq \frac{\|D^{(i)}|A^{-1}|\,|E|\,|A^{-1}|D^{(j)}\|_\infty}{\|D^{(i)} A^{-1} D^{(j)}\|_\infty} \\ &= \frac{(|A^{-1}|\,|E|\,|A^{-1}|)_{ij}}{|A^{-1}|_{ij}} + \beta\varepsilon = r + \beta\varepsilon \end{aligned}$$

for a constant β not depending on ε. This proves the right inequality in (8.9). Denote $C := |A^{-1}|\,|E|\,|A^{-1}|$ and let $\|A^{-1}\|_\infty = \sum_\nu |A^{-1}|_{i\nu}$ and $\|C\|_\infty = \sum_\nu C_{j\nu}$. Then by (8.2) and the definition (8.8) of r

$$n\mu_E(A) \geq \frac{\|C\|_\infty}{\|A^{-1}\|_\infty} = \frac{\sum\limits_\nu C_{j\nu}}{\sum\limits_\nu |A^{-1}|_{i\nu}} \geq \frac{\sum\limits_\nu C_{i\nu}}{\sum\limits_\nu |A^{-1}|_{i\nu}} \geq \frac{r\sum\limits_\nu |A^{-1}|_{i\nu}}{\sum\limits_\nu |A^{-1}|_{i\nu}} = r. \qquad \square$$

We note that one may measure the componentwise relative perturbation of $(A+\Delta A)^{-1}$ versus A^{-1} subject to componentwise perturbations of A. Then (cf. [4, 16])

$$\begin{aligned} \tilde{\mu}_E(A) &:= \lim_{\varepsilon \to 0} \sup \left\{ \frac{|(A+\Delta A)^{-1} - A^{-1}|_{ij}}{\varepsilon |A^{-1}|_{ij}} : \quad \Delta A \in M_n(\mathbb{R}),\ |\Delta A| \leq \varepsilon |E| \right\} \\ &= \max_{ij} \frac{(|A^{-1}|\,|E|\,|A^{-1}|)_{ij}}{|A^{-1}|_{ij}}. \end{aligned}$$

The structured componentwise condition number for the inverse can be bounded by adapting the approach for linear systems. Let $A, \Delta A \in M_n^{\rm struct}(\mathbb{R})$, ${\rm vec}(\Delta A) = \Phi^{\rm struct} p_{\Delta A}$, and ${\rm vec}(E) = \Phi^{\rm struct} p_E$. Then $|\Delta A| \leq |E|$ is equivalent to $|p_{\Delta A}| \leq |p_E|$. In view of (8.5) we note that (see [17, Lemma 4.3.1])

$${\rm vec}(A^{-1}\Delta A A^{-1}) = (A^{-T} \otimes A^{-1}){\rm vec}(\Delta A) = (A^{-T} \otimes A^{-1})\Phi^{\rm struct} p_{\Delta A}. \tag{8.10}$$

This implies the following.

THEOREM 8.3. *Let nonsingular $A \in M_n^{\text{struct}}(\mathbb{R})$ and $E \in M_n^{\text{struct}}(\mathbb{R})$ be given. Let $B \in M_n(\mathbb{R})$ with*

$$\text{vec}(B) = |(A^{-T} \otimes A^{-1})\Phi^{\text{struct}}|\,|p_E| \tag{8.11}$$

and denote

$$\alpha := \frac{\|B\|_\infty}{\|A^{-1}\|_\infty}.$$

Then

$$n^{-1}\alpha \le \mu_E^{\text{struct}}(A) \le \alpha. \tag{8.12}$$

Remark 8.4. The result includes (8.2) because, in the unstructured case, $\Phi = I_{n^2}$ and $\text{vec}(B) = |A^{-T} \otimes A^{-1}|\,|p_E| = (|A^{-T}| \otimes |A^{-1}|)|p_E| = \text{vec}(|A^{-1}|\,|E|\,|A^{-1}|)$ by [17, Lemma 4.3.1].

Proof. Let $dA \in M_n^{\text{struct}}(\mathbb{R})$ such that $\|A^{-1}dAA^{-1}\|_\infty = \sup\{\|A^{-1}\Delta AA^{-1}\| : |\Delta A| \le |E|\}$, and denote $\text{vec}(dA) = \Phi^{\text{struct}}p_{dA}$. Then $|p_{dA}| \le |p_E|$ implies $|\text{vec}(A^{-1}dAA^{-1})| = |(A^{-T} \otimes A^{-1})\Phi^{\text{struct}} \cdot p_{dA}| \le \text{vec}(B)$, and the right inequality in (8.12) follows by (8.10), (8.3), and the definition (8.1). On the other hand, let the index k, $1 \le k \le n^2$, be such that $\max_{\mu,\nu}|B_{\mu\nu}| = (\text{vec}(B))_k$. Denote $C := (A^{-T} \otimes A^{-1})\Phi^{\text{struct}}$ and set diagonal $D \in M_{n^2}(\mathbb{R})$ with $D_{\nu\nu} := sign(C_{k\nu})$. Furthermore, define $p_{\Delta A} := D|p_E|$ and let $\Delta A \in M_n^{\text{struct}}(\mathbb{R})$ with $\text{vec}(\Delta A) = \Phi^{\text{struct}}p_{\Delta A}$. Then

$$\beta := (\text{vec}(A^{-1}\Delta AA^{-1}))_k = ((A^{-T} \otimes A^{-1})\Phi^{\text{struct}}p_{\Delta A})_k = (|C|\,|p_E|)_k = (\text{vec}(B))_k$$

and

$$\mu_E^{\text{struct}}(A) \ge \frac{\|A^{-1}\Delta AA^{-1}\|_\infty}{\|A^{-1}\|} \ge \frac{\beta}{\|A^{-1}\|} \ge n^{-1}\frac{\|B\|_\infty}{\|A^{-1}\|}. \qquad \square$$

The question remains of whether, as for normwise perturbations, there is a relation between the reciprocal of the matrix condition number and the componentwise distance to the nearest singular matrix. This question will be treated in the next section.

9. Distance to singularity. For normwise and unstructured perturbations the condition number is equal to the reciprocal of the distance to the nearest singular matrix. Moreover, we showed in Part I, Theorem 12.1 that this is also true for structured (normwise) perturbations, that is,

$$\delta_E^{\text{struct}}(A) = \kappa_E(A)^{-1},$$

which is true for all our structures (2.5) under investigation. In the limit, a matrix has condition number ∞ iff it is singular, that is, the distance to singularity is 0.

The question arises of whether a similar result can be proved for componentwise perturbations, unstructured or structured. The componentwise (structured) distance to the nearest singular matrix is defined by

$$\text{d}_E^{\text{struct}}(A) := \min\{\alpha :\ \Delta A \in M_n^{\text{struct}}(\mathbb{R}),\ |\Delta A| \le \alpha|E|,\ A + \Delta A \text{ singular}\}. \tag{9.1}$$

For normwise perturbations, the distance to singularity $\delta_E^{\text{struct}}(A)$ as well as the condition number $\kappa_E(A)$ depend on row and column diagonal scaling of the matrix. This

is no longer true for componentwise perturbations. The unstructured distance to singularity $\mathrm{d}_E(A)$ as well as the structured distance is independent of row and column diagonal scaling (as long as, of course, the scaled matrix remains in the structure). That is, for positive diagonal D_1, D_2,

$$\mathrm{d}_{D_1ED_2}(D_1AD_2) = \mathrm{d}_E(A),$$

and for $A, E, D_1AD_2, D_1ED_2 \in M_n^{\mathrm{struct}}(\mathbb{R})$,

$$\mathrm{d}^{\mathrm{struct}}_{D_1ED_2}(D_1ED_2) = \mathrm{d}^{\mathrm{struct}}_E(A).$$

This is simply because $|\Delta A| \le \alpha|E| \;\Leftrightarrow\; |D_1\Delta AD_2| \le \alpha D_1|E|D_2$ and $\det(A+\Delta A) = 0 \;\Leftrightarrow\; \det(D_1AD_2 + D_1\Delta AD_2) = 0$. Furthermore, $A + \tilde{E} = A(I + A^{-1}\tilde{E})$ is singular iff -1 is an eigenvalue of $A^{-1}\tilde{E}$ so that definition (9.1) implies

(9.2)
$$\mathrm{d}^{\mathrm{struct}}_E(A) = \Big[\max\{|\lambda| :\; \tilde{E} \in M_n^{\mathrm{struct}}(\mathbb{R}),\, |\tilde{E}| \le |E|,\, \lambda \text{ real eigenvalue of } A^{-1}\tilde{E}\}\Big]^{-1}.$$

Note that the maximum is taken only over real eigenvalues of $A^{-1}\tilde{E}$. For unstructured perturbations the linearity of the determinant in each matrix element implies that the matrices $\tilde{E}$ can be restricted to the boundary $|\tilde{E}| = |E|$, i.e., finitely many matrices:

$$\mathrm{d}_E(A) = \Big[\max\{|\lambda| :\; \tilde{E} \in M_n(\mathbb{R}),\, |\tilde{E}| = |E|,\, \lambda \text{ real eigenvalue of } A^{-1}\tilde{E}\}\Big]^{-1}. \tag{9.3}$$

This is not true for structured perturbations; that is, the maximum may only be achieved for some $|\tilde{E}| \ne |E|$. An example for symmetric structures was given in [20].

With respect to the condition number things are even more involved. In the normwise case, we have $\kappa^{\mathrm{struct}}_E(A) = \|A^{-1}\|\,\|E\|$ (see Part I, Theorem 11.1) for all structures in (2.5), and it is the same condition number for matrix inversion as for linear systems when taking the supremum over all x. This is no longer true in the componentwise case. Here the condition numbers for matrix inversion and a linear system with the same matrix may be arbitrarily far apart (cf. the example in (8.6)). So if there is a relation at all between distance to singularity and the reciprocal of a condition number we first have to discuss which is the "right" condition number to choose.

Let us first consider the condition number $\mu_E(A)$ of the matrix inverse as defined in the previous section. Consider

$$A = A_\varepsilon = \begin{pmatrix} -\varepsilon & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad E = A.$$

By (9.3) we calculate

$$\mathrm{d}_A(A) = \frac{1}{4}\varepsilon^{1/2} + \mathcal{O}(1).$$

On the other hand (8.2) yields

$$\mu_A(A) \le \frac{\|\,|A^{-1}|\,|A|\,|A^{-1}|\,\|_\infty}{\|A^{-1}\|_\infty} \le 8. \tag{9.4}$$

Note that this is true for the most common case $E = A$ of componentwise relative perturbations. The same example applies to structured perturbations. The perturbation ΔA with $|\Delta A| = \mathrm{d}_A(A)|A|$ and $\det(A + \Delta A) = 0$ is a symmetric matrix. That means

$$\mathrm{d}_A^{\mathrm{sym}}(A) = \frac{1}{4}\varepsilon^{1/2} + \mathcal{O}(1) \quad \text{and} \quad \mu_A^{\mathrm{sym}}(A) \leq \mu_A(A) \leq 8.$$

So the condition number of the matrix inverse does not seem appropriate for our anticipated results.

To proceed let us first consider unstructured componentwise perturbations. Then, by Corollary 4.3, the worst case condition number for all x is $\mathrm{cond}_E(A) = \sup_{x \neq 0} \mathrm{cond}_E(A, x) = \mathrm{cond}_E(A, e) = \| \, |A^{-1}| \, |E| \, \|_\infty$. We choose no perturbations in the right-hand side because we are interested in the *matrix property* of distance to singularity. By column diagonal scaling, $\| \, |A^{-1}| \, |E| \, \|_\infty$ may become arbitrarily large. Therefore we choose optimal diagonal scaling for which [3, 9, 23]

$$\inf_{D_1, D_2} \mathrm{cond}_{D_1 E D_2}(D_1 A D_2) = \varrho(|A^{-1}| \, |E|), \tag{9.5}$$

the infimum taken over nonsingular diagonal D_ν, ϱ denoting the spectral radius. Note that $\varrho(|A^{-1}| \, |E|)$ is also equal to the minimum *normwise* condition number $\kappa_{E,\infty}(A)$ with respect to the ∞-norm achievable by diagonal scaling. For this minimum condition number we could indeed show an inverse proportionality to $\mathrm{d}_E(A)$ as by [21, Proposition 5.1]

$$\frac{1}{\varrho(|A^{-1}| \, |E|)} \leq \mathrm{d}_E(A) \leq \frac{(3 + 2\sqrt{2})n}{\varrho(|A^{-1}| \, |E|)}. \tag{9.6}$$

The left inequality is an equality for large classes of matrices, e.g., M-matrices. Moreover, explicit $n \times n$ examples, $n \geq 1$, were given [21] with $\mathrm{d}_A(A) = n\varrho(|A^{-1}| \, |A|)^{-1}$, so there is not much room for improvement in (9.6).

The question remains of whether a similar result is possible in case of structured componentwise perturbations. Unfortunately, for all structures in (2.5) the answer is no. Following we give a sequence of examples showing that. The first example for the symmetric case will be treated in more detail; the rest follow similarly. All examples will be given for the important case of componentwise relative perturbations of the matrix entries.

Let

$$A = A_\varepsilon = \begin{pmatrix} \varepsilon & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & \varepsilon \\ 1 & 1 & \varepsilon & 0 \end{pmatrix} \in M_4^{\mathrm{sym}}(\mathbb{R})$$

be given for $\varepsilon > 0$. A general symmetric perturbation of A subject to componentwise relative perturbations is

$$\tilde{A} = \begin{pmatrix} \varepsilon(1+\delta_1) & 0 & 1+\delta_2 & 1+\delta_3 \\ 0 & 0 & 1+\delta_4 & 1+\delta_5 \\ 1+\delta_2 & 1+\delta_4 & 0 & \varepsilon(1+\delta_6) \\ 1+\delta_3 & 1+\delta_5 & \varepsilon(1+\delta_6) & 0 \end{pmatrix} \in M_4^{\mathrm{sym}}(\mathbb{R}).$$

Then $\mathrm{d}_A^{\mathrm{sym}}(A)$ is the smallest α such that $|\delta_\nu| \le \alpha$ and $\det \tilde{A} = 0$. With Maple [25] we calculate

$$\begin{array}{rcl} \det \tilde{A} & = & c_0 + c_2\varepsilon^2 \quad \text{with} \\ c_0 & = & ((1+\delta_2)(1+\delta_5) - (1+\delta_3)(1+\delta_4))^2, \\ c_2 & = & 2(1+\delta_1)(1+\delta_4)(1+\delta_5)(1+\delta_6). \end{array}$$

In order to move $\det \tilde{A}$ into zero, the second summand $c_2\varepsilon^2$ must be zero or negative. This implies $\mathrm{d}_A^{\mathrm{sym}}(A) \ge 1$ and therefore, of course

$$\mathrm{d}_A^{\mathrm{sym}}(A) = 1$$

because $\tilde{A} \equiv 0$ for $\delta_\nu \equiv -1$. On the other hand,

$$\mathrm{cond}_A(A) = \mathrm{cond}_A^{\mathrm{sym}}(A) = \| \, |A^{-1}| \, |A| \, \|_\infty = 4\varepsilon^{-1} + \mathcal{O}(1).$$

Moreover, (9.5) implies

$$\inf_D \mathrm{cond}_{DAD}^{\mathrm{sym}}(DAD) \ge \varrho(|A^{-1}| \, |A|) = 2.8\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \quad d_A(A) = \varrho(|A^{-1}| \, |A|)^{-1}$$

so that there are arbitrarily ill-conditioned, though optimally scaled, symmetric matrices with $\mathrm{d}_A^{\mathrm{sym}}(A) = 1$. In other words, no relative perturbation less than 100% may move A into the manifold of (symmetric) singular matrices. The example above is extendable to higher dimensions by choosing $A \oplus I$. By Theorem 3.2 the example extends also to persymmetric structures.

For the skewsymmetric case consider

$$A = A_\varepsilon = \begin{pmatrix} 0 & 0 & -1 & 1-\varepsilon & 0 & 0 \\ 0 & 0 & -\varepsilon & 0 & 1+\varepsilon & 1 \\ 1 & \varepsilon & 0 & -1 & 0 & 0 \\ -1+\varepsilon & 0 & 1 & 0 & 0 & 0 \\ 0 & -1-\varepsilon & 0 & 0 & 0 & -\varepsilon \\ 0 & -1 & 0 & 0 & \varepsilon & 0 \end{pmatrix} \in M_6^{\mathrm{skewsym}}(\mathbb{R}).$$

Here and in the following examples we define $\tilde{A}$ (as for the symmetric case) by multiplying the components of A by $1 + \delta_\nu$ with rowwise numbering of the δ_ν. Then

$$\det \tilde{A} = \varepsilon^4(1-\varepsilon)^2(1+\delta_2)^2(1+\delta_3)^2(1+\delta_7)^2,$$

implying

$$\mathrm{d}_A^{\mathrm{skewsym}}(A) = 1.$$

On the other hand,

$$\mathrm{cond}_A(A) = 6\varepsilon^{-2} + \mathcal{O}(\varepsilon^{-1}) \quad \text{and} \quad \mathrm{cond}_A^{\mathrm{skewsym}}(A) = 2\varepsilon^{-2} + \mathcal{O}(\varepsilon^{-1}),$$

whereas

$$\inf_D \mathrm{cond}_{DAD}^{\mathrm{skewsym}}(DAD) \ge \varrho(|A^{-1}| \, |A|) = 6\varepsilon^{-3/2} + \mathcal{O}(\varepsilon^{1/2}) \quad \text{and}$$
$$\mathrm{d}_A(A) = \varrho(|A^{-1}| \, |A|)^{-1}$$

such that A is truly ill conditioned for arbitrary diagonal scaling with respect to relative componentwise skewsymmetric perturbations. Now (8.5) and (9.5) imply for $\text{struct} \in \{\text{sym}, \text{persym}, \text{skewsym}\}$

$$\inf_D \text{cond}_{DED}^{\text{struct}}(DAD) = \inf_D \text{cond}_{DED}(DAD) = \varrho(|A^{-1}|\,|E|),$$

the infimum taken over positive diagonal D. So we have the following result.

THEOREM 9.1. *Let* $\text{struct} \in \{\text{sym}, \text{persym}, \text{skewsym}\}$. *Then for every* $\varepsilon > 0$ *there exists* $A := A_\varepsilon \in M_n^{\text{struct}}(\mathbb{R})$ *with*

$$\inf_D \text{cond}_{DAD}^{\text{struct}}(DAD) > \varepsilon^{-1} \quad \textit{and} \quad \text{d}_A^{\text{struct}}(A) = 1.$$

For the symmetric Toeplitz case consider

$$A = A_\varepsilon = \text{Toeplitz}(0, 1, 1, -\varepsilon) \in M_4^{\text{symToep}}. \tag{9.7}$$

Then defining $\tilde{A}$ as before yields

$$\begin{aligned} \det \tilde{A} &= c_0 + c_1\varepsilon + c_2\varepsilon^2 \quad \text{with} \\ c_0 &= (2 + \delta_1 + \delta_2)^2(\delta_2 - \delta_1)^2, \\ c_1 &= 2(1 + \delta_1)(1 + \delta_3)((1 + \delta_1)^2 + (1 + \delta_2)^2), \quad \text{and} \\ c_2 &= (1 + \delta_1)^2(1 + \delta_3)^2. \end{aligned}$$

For $|\delta_\nu| < 1$, c_0 is nonnegative, whereas both c_1 and c_2 are positive. Therefore

$$\text{d}_A^{\text{symToep}}(A) = 1.$$

On the other hand, for $x = (1, -1, 1, -1)^T$,

$$\begin{aligned} &\text{cond}_A(A) = 2\varepsilon^{-1} + \mathcal{O}(1) \quad \text{and} \\ &\sup_{x \neq 0} \text{cond}_A^{\text{symToep}}(A) \geq \text{cond}_A^{\text{symToep}}(A, x) = 2\varepsilon^{-1} + \mathcal{O}(1) \end{aligned}$$

such that A is truly ill conditioned subject to relative componentwise symmetric Toeplitz perturbations. For Toeplitz structures, diagonal scaling is, in general, not possible. For completeness we note

$$\varrho(|A^{-1}|\,|A|) = 2\varepsilon^{-1/2} + \mathcal{O}(1) = [\text{d}_A(A)]^{-1}.$$

The same example applies, according to Theorem 3.2, to persymmetric Hankel structures.

For the general Toeplitz case consider

$$A = A_\varepsilon = \text{Toeplitz}([0, 1, -1, 0], [0, 1, -1, -\varepsilon]) \in M_4^{\text{Toep}}(\mathbb{R}),$$

where $\text{Toeplitz}(c, r)$ denotes the (general) Toeplitz matrix with first column c and first row r. Defining $\tilde{A}$ as before yields

$$\begin{aligned} \det \tilde{A} &= c_0 + c_1\varepsilon \quad \text{with} \\ c_0 &= ((1 + \delta_1)(1 + \delta_4) - (1 + \delta_1)(1 + \delta_5))^2 \quad \text{and} \\ c_1 &= (1 + \delta_3)(1 + \delta_4)^3 + (1 + \delta_1)(1 + \delta_3)(1 + \delta_5)^2. \end{aligned}$$

For $|\delta_\nu| < 1$ the determinant is positive, so

$$\mathrm{d}_A^{\mathrm{Toep}}(A) = 1.$$

On the other hand,

$$\mathrm{cond}_A(A, e) = 4\varepsilon^{-1} + \mathcal{O}(1) = \mathrm{cond}_A^{\mathrm{Toep}}(A, e),$$

so A is truly ill conditioned subject to relative componentwise general Toeplitz perturbations. By Theorem 3.2 this also covers the Hankel case. For completeness we note

$$\varrho(|A^{-1}|\,|A|) = 2\sqrt{2}\varepsilon^{-1/2} + \mathcal{O}(1) = [\mathrm{d}_A(A)]^{-1}.$$

Finally, define for the circulant case

$$A = A_\varepsilon = \mathrm{circ}(1, \varepsilon, 1, 0) \in M_4^{\mathrm{circ}}(\mathbb{R}).$$

For $\tilde{A}$ defined as before we get

$$\begin{array}{rcl} \det \tilde{A} & = & \alpha\beta \quad \text{with} \\ \alpha & = & (2 + \delta_1 + \delta_3)^2 - \varepsilon^2(1 + \delta_2)^2 \quad \text{and} \\ \beta & = & (\delta_1 - \delta_3)^2 + \varepsilon^2(1 + \delta_2)^2. \end{array}$$

For small ε both factors are nonzero for $|\delta_\nu| < 1$, so

$$\mathrm{d}_A^{\mathrm{circ}}(A) = 1.$$

On the other hand, for $x = (1, 1, 1, -1)^T$,

$$\begin{array}{l} \mathrm{cond}_A(A, x) = 2\varepsilon^{-1} + \mathcal{O}(1) = \mathrm{cond}_A^{\mathrm{circ}}(A, x) \quad \text{and} \\ \varrho(|A^{-1}|\,|A|) = 2\varepsilon^{-1} + \mathcal{O}(1) = [\mathrm{d}_A(A)]^{-1}. \end{array}$$

Summarizing, we have the following result.

THEOREM 9.2. *Let* struct $\in \{\mathrm{symToep}, \mathrm{Toep}, \mathrm{circ}, \mathrm{Hankel}, \mathrm{persymHankel}\}$. *Then for every* $\varepsilon > 0$ *there exists* $A := A_\varepsilon \in M_n^{\mathrm{struct}}(\mathbb{R})$ *and* $x \in \mathbb{R}^n$ *with* $|x| = e$ *such that*

$$\mathrm{cond}_A^{\mathrm{struct}}(A, x) > \varepsilon^{-1} \quad \textit{and} \quad \mathrm{d}_A^{\mathrm{struct}}(A) = 1.$$

10. Conclusion. Summarizing, depending on the perturbation in use, we face severe differences in the sensitivity of the solution of a linear system. An extreme example is symmetric Toeplitz perturbations. In that case, Theorem 6.3 implies that for the matrices defined in (6.2) the solution $A^{-1}b$ is well conditioned subject to structured componentwise perturbations in the matrix for *all* right-hand sides b. However, for unstructured componentwise perturbations it is ill conditioned for *generic* right-hand side b. This is true when perturbations are restricted to the matrix.

We saw similar examples with a perfectly well-conditioned linear system with respect to componentwise structured perturbations in the matrix *and* the right-hand side, but being arbitrarily ill conditioned with respect to componentwise general (unstructured) perturbations. We presented such examples for all perturbations under investigation except circulants, for which almost sharp estimations for the ratio between the structured and unstructured condition numbers were derived.

So far it seems that componentwise perturbations may produce some quite unexpected and unwanted effects. One reason, as mentioned in the first section, is that zero weights produce certain substructures of the given structure. In particular, the degrees of freedom may be significantly reduced. Then a problem may become well conditioned because not much room is left to produce "bad" perturbations.

This may lead to the conclusion that it is rather unlikely we will find algorithms for the problems and structures under investigation in this paper that are stable with respect to componentwise perturbations. One might even conclude that this seems to be an intrinsic property of componentwise perturbations.

Fortunately, this seems not to be the case. There are other structures for which very fast and accurate algorithms have been developed for the solution of linear systems or matrix inversion and also for other problems such as LU-decomposition and the computation of singular values. For example, those problems can be solved with small componentwise relative backward error for Vandermonde-like or Cauchy matrices [16, section 22], [5, 8, 6]. This is especially remarkable because Vandermonde and Cauchy matrices are reputed for being persistently ill conditioned (with respect to unstructured perturbations; see [3] in Part I).

This is of course a question of exploiting the data, or of developing the "right" algorithms, but also is sometimes facilitated by choosing a clever set of input data. Consider, for example, the problem of matrix inversion, LU-decomposition, or computation of singular values for weakly diagonally dominant M-matrices. Small perturbations in the diagonal elements can cause arbitrarily large perturbations in the result. However, another choice of input data changes the situation [19, 1]: The mentioned problems are well conditioned with respect to the off-diagonal elements and the row sums as input data.

The problem with stability with respect to componentwise (relative) perturbations, structured or not, is that in the course of a computation one single subtraction producing some cancellation may ruin the result in the componentwise backward sense. The backward error of the result of the subtraction is small with respect to uncorrelated perturbations of the operands. However, perturbations are correlated if the operands are the result of previous computations. A typical example can be seen when solving (2.4) with Gaussian elimination.

It seems more and more difficult to design structured solvers for linear systems over the structures in (2.5) being stable with respect to structured componentwise perturbations. Are there such algorithms?

A candidate might be circulant matrices because of their rich algebraic properties. In fact, a normwise stable algorithm already exists [26]. Moreover, in contrast to the other perturbations under investigation, the worst case unstructured componentwise condition number in this case is at most about the square of the structured condition number (for perturbations in the matrix and the right-hand side; see Theorem 7.2).

Finally, there does not seem to be much relation between the distance to singularity and the reciprocal of a condition number in case of componentwise structured perturbations. This is the case for the matrix inverse condition number $\mu_E(A)$ as well as for $\mathrm{cond}_E^{\mathrm{struct}}(A)$, the supremum of $\mathrm{cond}_E^{\mathrm{struct}}(A, x)$ for all x. But maybe an appropriate structured componentwise condition number for that purpose is still to be defined.

Acknowledgments. I wish to thank the students of the summer course on structured perturbations held during my sabbatical at Waseda University, Tokyo. Also my thanks to the two anonymous referees for their thorough reading and most constructive and useful comments.

REFERENCES

[1] A. S. Alfa, J. Xue, and Q. Ye, *Entrywise perturbation theory for diagonally dominant M-matrices with applications*, Numer. Math., 90 (2002), pp. 401–414.

[2] S. G. Bartels and D. J. Higham, *The structured sensitivity of Vandermonde-like systems*, Numer. Math., 62 (1992), pp. 17–33.

[3] F. L. Bauer, *Optimally scaled matrices*, Numer. Math., 5 (1963), pp. 73–87.

[4] F. L. Bauer, *Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme*, Z. Angew. Math. Mech., 46 (1966), pp. 409–421.

[5] A. Björck, *Component-wise perturbation analysis and error bounds for linear least squares solutions*, BIT, 31 (1991), pp. 238–244.

[6] T. Boros, T. Kailath, and V. Olshevsky, *The fast parallel Björck-Pereyra-type algorithm for parallel solution of Cauchy linear equations*, Linear Algebra Appl., 302/303 (1999), pp. 265–293.

[7] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[8] J. Demmel, *Accurate singular value decompositions of structured matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 562–580.

[9] J. W. Demmel, *The componentwise distance to the nearest singular matrix*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 10–19.

[10] I. Gohberg and I. Koltracht, *Mixed, componentwise, and structured condition numbers*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 688–704.

[11] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[12] W. Hager, *Condition estimates*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 311–316.

[13] 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.

[14] D. J. Higham and N. J. Higham, *Componentwise perturbation theory for linear systems with multiple right-hand sides*, Linear Algebra Appl., 174 (1992), pp. 111–129.

[15] N. J. Higham, *Experience with a matrix norm estimator*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 804–809.

[16] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[17] R. A. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[18] W. M. Kahan, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.

[19] C. O'Cinneide, *Relative-error bounds for the LU decomposition via the GTH algorithm*, Numer. Math., 73 (1996), pp. 507–519.

[20] S. M. Rump, *Structured perturbations and symmetric matrices*, Linear Algebra Appl., 278 (1998), pp. 121–132.

[21] S. M. Rump, *Ill-conditioned matrices are componentwise near to singularity*, SIAM Rev., 41 (1999), pp. 102–112.

[22] S. M. Rump, *Ill-conditionedness need not be componentwise near to ill-posedness for least squares problems*, BIT, 39 (1999), pp. 143–151.

[23] S. M. Rump, *Optimal scaling for p-norms and componentwise distance to singularity*, IMA J. Numer. Anal., 23 (2003), pp. 1–9.

[24] R. Skeel, *Scaling for numerical stability in Gaussian elimination*, J. ACM, 26 (1979), pp. 494–526.

[25] Maple V, *Release* 7.0*, Reference Manual*, Waterloo Maple, Inc., Waterloo, Ontario, Canada, 2001.

[26] P. Y. Yalamov, *On the almost strong stability of the circular deconvolution algorithm*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 358–363.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 57–79

OPTIMIZATION PROBLEMS OVER POSITIVE PSEUDOPOLYNOMIAL MATRICES*

Y. GENIN†, Y. HACHEZ†, YU. NESTEROV†, AND P. VAN DOOREN†

Abstract. The Nesterov characterizations of positive pseudopolynomials on the real line, the imaginary axis, and the unit circle are extended to the matrix case. With the help of these characterizations, a class of optimization problems over the space of positive pseudopolynomial matrices is considered. These problems can be solved in an efficient manner due to the inherent block Toeplitz or block Hankel structure induced by the characterization in question. The efficient implementation of the resulting algorithms is discussed in detail. In particular, the real line setting of the problem leads naturally to ill-conditioned numerical systems. However, adopting a Chebyshev basis instead of the natural basis for describing the polynomial matrix space yields a restatement of the problem and of its solution approach with much better numerical properties.

Key words. convex optimization, positive polynomials, Toeplitz matrices, Hankel matrices

AMS subject classifications. 47A68, 65F30, 65Y20, 90C22

PII. S0895479803374840

1. Introduction.

This paper is concerned with a convex optimization problem over the set of polynomial matrices, which are nonnegative definite on distinguished contours of the complex plane, namely, the real line, the imaginary axis, and the unit circle. The set of such polynomial matrices is convex. Moreover, it has been shown by Nesterov that scalar polynomials of this type [12] admit a compact parametrization in terms of constant nonnegative definite matrices satisfying simple linear algebraic constraints.

The aim of this paper is to extend this parametrization to the matrix case and, with the help of this result, to discuss and to solve an important class of related convex optimization problems. In fact, the dual formulation of these optimization problems appears to be considerably more attractive from a computational viewpoint. On the one hand, it is stated in an optimization space of reduced dimension. On the other hand, this dual space is characterized by nonnegative definite matrices that have block Hankel or block Toeplitz structure.

A well-established technique for solving such optimization problems involves the introduction of a barrier function [13] whose differential characteristics have to be repeatedly evaluated along the numerical optimization process. Due to the Hankel or Toeplitz structure of the optimization space, fast, and even superfast, algorithms, based on displacement rank techniques, can be proposed for that purpose. The computational aspects of their implementation are discussed in some detail. In addition, as the real line formulation of the problem is shown to be inherently ill-conditioned, a change of polynomial basis is considered and discussed. This problem reformulation

*Received by the editors July 4, 2000; accepted for publication (in revised form) by M. Hyman January 3, 2003; published electronically May 15, 2003. This research was supported by NSF contract CCR-97-96315 and by the Belgian Programme on Inter-university Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors.

http://www.siam.org/journals/simax/25-1/37484.html

†Université Catholique de Louvain, Department of Mathematical Engineering, CESAME, avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium (genin@csam.ucl.ac.be, hachez@csam.ucl.ac.be, nesterov@csam.ucl.ac.be, vdooren@csam.ucl.ac.be). The research of the second author was supported by a research fellowship from the Belgian National Fund for Scientific Research.

exhibits much more interesting numerical prospects from this viewpoint.

The theory of positive transfer functions is well known for playing a fundamental role in systems and control theory. Such functions represent, e.g., spectral density functions of stochastic processes, appear in spectral factorizations, and also are related to the Riccati equations. It has been known since the work of Youla [17] that, when such transfer functions are rational, they possess rational spectral factorizations. Later on, it was shown that, using state-space models of positive transfer functions, one could express the condition of positivity in terms of linear matrix inequalities (see, e.g., [16]). Positive transfer functions obviously form a convex set, and they were recently studied in the convex optimization literature [4, 12]. The parametrization of pseudopolynomial matrices proposed in this paper fits naturally into that context. In particular, this parametrization can also be obtained as a straightforward application of the celebrated positive real lemma to an appropriate subset of positive paraconjugate transfer functions.

In section 2, the definition of positive paraconjugate transfer functions is given; in particular, such functions are well known for enjoying a remarkable spectral factorization property.

In section 3, positive pseudopolynomial matrices on the real line, the imaginary axis, and the unit circle are considered. In each case, the positivity constraint is shown to induce some form of symmetry on the pseudopolynomial matrix coefficients and to impose some restrictions on their formal degree.

In section 4, parametrizations of nonnegative pseudopolynomial matrices are derived in terms of appropriate subsets of nonnegative constant Hermitian matrices. These appropriate subsets are defined by linear algebraic relations and can be parametrized in terms of an arbitrary Hermitian or skew-Hermitian constant matrix of reduced dimension, depending on the particular contour of the real plane considered.

In section 5, an alternative proof of this parametrization is derived from the theory of positive paraconjugate transfer functions. In particular, with the help of the positive real lemma, any state-space realization of such a function is proved to involve some degree of freedom, which can be expressed in terms of a linear matrix inequality (LMI). This is precisely the characterization obtained in the preceding section.

In section 6, a class of important optimization problems is defined over the set of nonnegative pseudopolynomial matrices satisfying linear constraints. These constraints are assumed to be expressible in terms of Frobenius scalar products. Next, the dual form of these optimization problems is shown to be computationally much more attractive. The dimension of the optimization space appears to be reduced to the number of linear constraints instead of the pseudopolynomial matrix dimension, as in the primal form. In addition, this optimization space is characterized by nonnegative definite block Hankel or block Toeplitz matrices, depending on the particular complex plane contour considered. Furthermore, modern techniques for the numerical solution of the optimization problem involve the introduction of a barrier function. Since the appropriate barrier functions inherit the Hankel or Toeplitz structure of the optimization space, this paves the way for fast evaluations of the differential characteristics of the barrier function. Such fast evaluations are of paramount importance because they have to be made repeatedly in such optimization schemes [4, 5, 12, 13].

In section 7, the computational aspects of the fast algorithms which can be used to solve these optimization problems are considered and analyzed in detail. The optimization scheme mainly involves recurrent computations of the differential

characteristics of the barrier function, namely, its gradient and its Hessian. These differential functions are evaluated by carrying out Frobenius scalar products of appropriate block Hermitian matrices with underlying Hankel or Toeplitz structure. Displacement rank techniques are especially suited for their fast evaluation. In particular, the required calculations can be broken down into fast, or even superfast, elementary numerical operations by exploiting the compact displacement rank representations resulting from the problem structure. It is also pointed out that the real line problem is inherently ill-conditioned. This fact is a well-known consequence of the Hankel structure.

In section 8, the real line optimization problem is reformulated to get around the above technical difficulty. As the Hankel structure is an obvious consequence of the expansion of polynomial matrices into the natural basis of their monomials $[I_m, x\, I_m, x^2\, I_m, \ldots]$, the remedy consists in a change of basis. In this light, it is proposed to substitute a basis of Chebyshev polynomials for the natural basis. Such a Chebyshev basis induces a Toeplitz-plus-Hankel structure to the problem with, in principle, a much better numerical conditioning. It is finally recalled how one can take advantage of the Toeplitz-plus-Hankel structure in fast algorithms based on appropriate displacement rank techniques [6].

2. Paraconjugate transfer functions. Paraconjugate transfer functions $\Phi(.)$ play an important role in systems theory. They are defined with respect to a curve in the complex plane, which is typically the imaginary axis (for continuous-time systems), the unit circle (for discrete-time systems), and the real axis $\mathbb{R}$ (for the moment problem).

Imaginary axis. This curve is the boundary of the stable region for continuous-time transfer functions in the complex variable s (which is also the variable of the Laplace transform of such dynamical systems): the imaginary axis is denoted $s \in j\mathbb{R}$.

Unit circle. This curve is the boundary of the stable region for discrete-time transfer functions in the complex variable z (which is also the variable of the so-called z-transform of such dynamical systems): the unit circle is denoted $z \in e^{j\mathbb{R}}$.

Real axis. This curve occurs in the standard treatment of the classical moment problem [1, 11]. In this case, the complex variable x will be used with the real axis and denoted $x \in \mathbb{R}$.

To stress that a result holds for a particular curve, the above particular variable notation will be adopted instead of the standard variable p. In this paper, only the case of square *rational* transfer matrices $\Phi(p)$ will be considered, i.e., $m \times m$ matrices $\Phi(p)$ whose entries are rational functions of the variable p.

DEFINITION 2.1. *The paraconjugate transfer function $\Phi_*(p)$ of a given transfer matrix $\Phi(p)$ is defined as follows:*

$$\Phi_*(s) = [\Phi(-\overline{s})]^* \text{ for the imaginary axis,}$$
$$\Phi_*(z) = [\Phi(1/\overline{z})]^* \text{ for the unit circle,}$$
$$\Phi_*(x) = [\Phi(\overline{x})]^* \text{ for the real axis,}$$

where M^* is the conjugate transposed matrix of a matrix M.

Let us point out that the paraconjugate $\Phi_*(p)$ is also a rational transfer function of the complex variable p. A para-Hermitian transfer function can then be defined as follows.

DEFINITION 2.2. *A square transfer function $\Phi(p)$ is para-Hermitian if it is equal to its paraconjugate: $\Phi_*(p) = \Phi(p)$.*

This definition depends on the choice of curve considered. However, a para-Hermitian transfer function *evaluated* on the corresponding curve is always a Hermitian matrix. Indeed, $\Phi_*(p) = \Phi(p)$ implies the following for each case:

$$\begin{aligned}\Phi_*(j\omega) &= [\Phi(j\omega)]^* \text{ for } s = j\omega \text{ on the imaginary axis,}\\ \Phi_*(e^{j\omega}) &= [\Phi(e^{j\omega})]^* \text{ for } z = e^{j\omega} \text{ on the unit circle,}\\ \Phi_*(\omega) &= [\Phi(\omega)]^* \text{ for } x = \omega \text{ on the real axis,}\end{aligned}$$

where $\omega \in \mathbb{R}$ is thus a real variable parametrizing the curve.

Since a paraconjugate transfer function is a Hermitian matrix when evaluated on the curve, all its eigenvalues are real. Therefore, a positivity constraint can be imposed on these eigenvalues. This leads to the following definition.

DEFINITION 2.3. *A paraconjugate transfer function is* positive *(*nonnegative*) if it is positive (*nonnegative*) when evaluated on the curve:* $\Phi(p) \succ 0$ $(\Phi(p) \succeq 0)$.

Note that nonnegative paraconjugate transfer functions always possess a so-called *spectral factorization*,

$$\Phi(p) = G_*(p)G(p), \tag{2.1}$$

where the spectral factor $G(p)$ is again a square rational transfer function in p. This result is proven in the systems theory literature [17, 14].

3. Positive pseudopolynomial matrices. Pseudopolynomial matrices are matrices with a finite expansion in positive and negative powers of the independent variable p:

$$\Phi(p) = \sum_{k=-r}^{t} \Phi_k p^k.$$

Depending on the type of curve one considers, the coefficient matrices of such pseudopolynomial matrices must possess a certain symmetry.

Real axis. For a para-Hermitian transfer function $\Phi(x)$ that is nonnegative on the real axis $x \in \mathbb{R}$, it follows from the para-Hermitian nature that the coefficient matrices of the expansion

$$\Phi(x) = \sum_{k=-r}^{t} \Phi_k x^k \tag{3.1}$$

must all be Hermitian: $\Phi_k = \Phi_k^*$. Moreover, since x^2 is nonnegative on the real axis $x \in \mathbb{R}$, such pseudopolynomial matrices can be reduced to polynomial matrices in x or in x^{-1}; in particular, they reduce to the form

$$\Phi(x) = \sum_{k=0}^{t} \Phi_k x^k.$$

From the nonnegativity of $\Phi(x)$, it turns out that the highest degree coefficient must be of even degree $t = 2n$. For polynomial matrices in x^{-1}, the highest degree coefficient is also of even degree. The standard form used here for nonnegative para-Hermitian matrices on the real axis is

$$\Phi(x) = \sum_{k=0}^{2n} \Phi_k x^k, \quad \Phi_k = \Phi_k^*. \tag{3.2}$$

Unit circle. For a para-Hermitian transfer function $\Phi(z)$ that is nonnegative on the unit circle $z \in e^{j\mathbb{R}}$, it follows from the para-Hermitian nature that the coefficient matrices of the expansion

$$\Phi(z) = \sum_{k=-r}^{t} \Phi_k z^k \tag{3.3}$$

must satisfy the condition $\Phi_{-k} = \Phi_k^*$; thus such a pseudopolynomial matrix must have a symmetric expansion. The standard form used here for nonnegative para-Hermitian matrices on the unit circle is

$$\Phi(z) = \sum_{k=-n}^{n} \Phi_k z^k, \quad \Phi_{-k} = \Phi_k^*. \tag{3.4}$$

Imaginary axis. For a para-Hermitian transfer function $\Phi(s)$ that is nonnegative on the imaginary axis $s \in j\mathbb{R}$, it follows from the para-Hermitian nature that the coefficient matrices of the expansion

$$\Phi(s) = \sum_{k=-r}^{t} \Phi_k s^k \tag{3.5}$$

are Hermitian if k is even and are skew-Hermitian if k is odd:

$$\Phi_{2k} = \Phi_{2k}^*, \quad \Phi_{2k+1} = -\Phi_{2k+1}^*.$$

This follows easily from the change of variables $s = jx$ converting the real axis into the imaginary axis. One can again multiply by a power of $-s^2$ (which is nonnegative on the imaginary axis) to obtain a polynomial matrix in s or s^{-1},

$$\Phi(s) = \sum_{k=0}^{t} \Phi_k s^k,$$

and it is easy to see from the nonnegativity that the highest degree coefficient must be of even degree $t = 2n$. For polynomial matrices in s^{-1} the highest degree coefficient is also of even degree. The standard form we use here for nonnegative para-Hermitian matrices on the imaginary axis is

$$\Phi(x) = \sum_{k=0}^{2n} \Phi_k x^k, \quad \Phi_{2k} = \Phi_{2k}^*, \quad \Phi_{2k+1} = -\Phi_{2k+1}^*. \tag{3.6}$$

To end this section, let us observe that the pseudopolynomial matrices of interest have, in the above cases, $(2n+1)m^2$ degrees of freedom.

4. Parametrization of nonnegative pseudopolynomial matrices. The main result of this section highlights a parametrization of nonnegative pseudopolynomial matrices in terms of constant Hermitian or skew-Hermitian matrices.

To begin and, for further use, let us introduce two particular $(n+1)m \times (n+1)m$ block matrices: the standard block shift operator

$$Z \doteq \begin{bmatrix} 0 & I_m & & \\ & 0 & \ddots & \\ & & \ddots & I_m \\ & & & 0 \end{bmatrix}$$

on the one hand, and the degenerate matrix

$$X \doteq \left[\begin{array}{c|c} X_0 & 0 \\ \hline 0 & 0 \end{array}\right], \tag{4.1}$$

with X_0 any $nm \times nm$ complex matrix on the other hand.

4.1. Real axis. Let

$$P(x) = \sum_{k=0}^{2n} P_k x^k \tag{4.2}$$

be an $m \times m$ para-Hermitian polynomial matrix with Hermitian coefficients, i.e., $P_k = P_k^*$, and consider the set of Hermitian matrices

$$Y = \begin{bmatrix} Y_{0,0} & Y_{0,1} & \cdots & Y_{0,n} \\ Y_{1,0} & Y_{1,1} & \cdots & Y_{2,n} \\ \vdots & \vdots & & \vdots \\ Y_{n,0} & Y_{n,1} & \cdots & Y_{n,n} \end{bmatrix},$$

with blocks of dimension $m \times m$. If $\Pi(x)$ stands for

$$\Pi(x) = \begin{bmatrix} I_m & xI_m & \cdots & x^n I_m \end{bmatrix}^T,$$

the relation

$$\Pi_*(x) Y \Pi(x) = P(x) \tag{4.3}$$

implies that

$$P_k = \sum_{i+j=k} Y_{i,j}, \quad k = 0, \ldots, 2n, \tag{4.4}$$

within the convention that $Y_{i,j} = 0$ for i and j outside their definition range. A simple choice for Y so as to obtain this identity is found to be

$$Y_0 = \begin{bmatrix} P_0 & \frac{1}{2}P_1 & & \\ \frac{1}{2}P_1 & P_2 & \ddots & \\ & \ddots & \ddots & \frac{1}{2}P_{2n-1} \\ & & \frac{1}{2}P_{2n-1} & P_{2n} \end{bmatrix}. \tag{4.5}$$

Then, the following characterization theorem can be stated.

THEOREM 4.1. *A Hermitian matrix Y satisfies* (4.3) *if and only if it can be expressed as*

$$Y = Y_0 + Z^T X - X Z, \tag{4.6}$$

where X has the form (4.1) *and is skew-Hermitian, i.e., $X = -X^*$.*

Proof. The "if" part is obvious since one has $\Pi_*(x)\left[Z^TX - XZ\right]\Pi(x) = 0$ for any matrix X of the form (4.1). Conversely, let Y be a solution of (4.3) and let us set X as

$$X = \sum_{k=0}^{n} (Z^{k+1})(Y - Y_0)(Z^k). \tag{4.7}$$

It turns out that X has the structure (4.1) with $X = -X^*$ and satisfies (4.6). To see this, observe first that X has the structure (4.1) as an immediate consequence of relations (4.4). Next, inserting (4.7) in (4.6), one obtains successively

$$\begin{aligned} Y_0 + Z^TX - XZ &= Y_0 + Z^TZ\sum_{k=0}^{n} Z^k(Y - Y_0)Z^k - \sum_{k=0}^{n} Z^{k+1}(Y - Y_0)Z^{k+1} \\ &= Y_0 + Z^TZ(Y - Y_0) + (Z^TZ - I_{(n+1)m})\sum_{k=0}^{n-1} Z^{k+1}(Y - Y_0)Z^{k+1} \\ &= Y_0 + (Y - Y_0) \\ &= Y \end{aligned}$$

again in view of relations (4.4). Finally, one establishes the skew-Hermitian property of X from the fact that $Z^TX - XZ = X^*Z - Z^TX^*$ necessarily implies $X = -X^*$ for any matrix X of algebraic structure (4.1). ∎

Imposing the condition that $P(x)$ is also a nonnegative transfer function leads to the following theorem.

THEOREM 4.2. *A pseudopolynomial matrix $P(x) = \sum_{k=0}^{2n} P_k x^k$ is nonnegative definite on the real axis if and only if there exists a nonnegative definite Hermitian matrix Y with blocks $Y_{i,j}, i, j = 0, \ldots, n$, such that ($Y_{i,j} = 0$ for i and j outside their definition range)*

$$P_k = \sum_{i+j=k} Y_{i,j} \quad \text{for } k = 0, \ldots, 2n. \tag{4.8}$$

Proof. Because of the previous theorem, the "only if" part only needs a proof. It is obtained from the existence of a spectral factorization

$$P(x) = G_*(x)G(x),$$

where $G(x)$ is polynomial in x: $G(x) = \sum_{k=0}^{n} G_n x^n$. Indeed, choose

$$Y = \begin{bmatrix} G_0 & G_1 & \cdots & G_n \end{bmatrix}^* \begin{bmatrix} G_0 & G_1 & \cdots & G_n \end{bmatrix}.$$

This matrix Y is nonnegative and satisfies the constraints of the theorem. ∎

Let us point out that if $\det(P(x))$ has zeros, then Y cannot be strictly positive definite. This characterization of matrix polynomials nonnegative on the real axis extends a result obtained earlier by Nesterov [12] for scalar polynomials.

4.2. Unit circle. Let us now consider the case of the nonnegative transfer functions on the unit circle. It follows from its finite expansion and from its para-Hermitian character that such a pseudopolynomial matrix

$$P(z) = \sum_{k=-n}^{n} P_k z^k \tag{4.9}$$

has $m \times m$ coefficient matrices that satisfy $P_{-k} = P_k^*$. The set of Hermitian matrices of interest here is defined by the equation

$$\Pi_*(z) Y \Pi(z) = P(z), \tag{4.10}$$

where the same notation as above is used for the matrix Y and $\Pi(.)$. This is algebraically equivalent to the relations

$$P_k = \sum_{i-j=k} Y_{i,j}, \tag{4.11}$$

assuming $Y_{i,j} = 0$ for i and j outside their definition range. Clearly, the choice

$$Y_0 = \begin{bmatrix} P_0 & P_1 & \cdots & P_n \\ P_1^* & 0 & \vdots & 0 \\ \vdots & \vdots & & \vdots \\ P_n^* & 0 & \cdots & 0 \end{bmatrix} \tag{4.12}$$

is an admissible matrix Y. The characterization theorem now takes the following form.

THEOREM 4.3. *A Hermitian matrix Y satisfies* (4.10) *if and only if it can be expressed as*

$$Y = Y_0 + X - Z^T X Z, \tag{4.13}$$

where X has the form (4.1) *and is Hermitian, i.e., $X = X^*$.*

Proof. By duplicating the argument used in the proof of Theorem 4.1, one shows that the solution X of (4.13) is given by

$$X = \sum_{k=0}^{n} (Z^k)^T (Y - Y_0)(Z^k)$$

and that the resulting matrix X has the stated form because of (4.11). ∎

The positive pseudopolynomial matrices on the unit circle can then be characterized as follows.

THEOREM 4.4. *A pseudopolynomial matrix $P(z) = \sum_{k=-n}^{n} P_k z^k$ is nonnegative definite on the unit circle if and only if there exists a nonnegative definite Hermitian matrix Y with blocks $Y_{i,j}, i, j = 0, \ldots, n$, such that (assuming $Y_{i,j} = 0$ for i and j outside their definition range)*

$$P_k = \sum_{i-j=k} Y_{i,j} \quad \text{for } k = -n, \ldots, 0, \ldots, n. \tag{4.14}$$

The proof of this theorem is again based on the same spectral factorization argument as in Theorem 4.2 and is therefore omitted. This characterization of pseudopolynomials nonnegative on the unit circle also extends a result previously obtained by Nesterov [12] for trigonometric polynomials.

4.3. Imaginary axis. The third kind of nonnegative pseudopolynomial matrices is that with respect to the imaginary axis. This formulation of the problem does not

require any specific treatment since it can be reduced to the case of the real axis in a straightforward manner. Indeed, consider the para-Hermitian polynomial matrix

$$P(s) = \sum_{k=0}^{2n} P_k s^k \tag{4.15}$$

with $s \in j\mathbb{R}$. If $s = jx$, one derives from $P(s)$ the para-Hermitian polynomial matrix

$$\hat{P}(x) = \sum_{k=0}^{2n} (j^k P_k) x^k = \sum_{k=0}^{2n} \hat{P}_k x^k$$

with respect to the real line. In particular, this implies $P_k^* = (-1)^k P_k$ for all k. Therefore, applying Theorem 4.2 to $\hat{P}(x)$, one obtains for $P(s)$ the following result.

THEOREM 4.5. *A pseudopolynomial matrix $P(s) = \sum_{k=0}^{2n} P_k s^k$ is nonnegative on the imaginary axis if and only if there exists a nonnegative definite Hermitian matrix Y with blocks $Y_{i,j}, i, j = 0, \ldots, n$, such that ($Y_{i,j} = 0$ for i and j outside their definition range)*

$$P_k = (-j)^k \sum_{i+j=k} Y_{i,j} \quad \text{for } k = 0, \ldots, 2n.$$

5. Positive paraconjugate transfer functions. The parametrization of positive pseudopolynomial matrices, derived in the preceding section, can alternatively be obtained from the theory of positive paraconjugate transfer functions. More precisely, it follows from a straightforward application of the celebrated positive real lemma to the subclass of positive paraconjugate transfer functions that has a pseudopolynomial form.

To see this, let us start from a well-known result of state-space theory [14] that states that any proper paraconjugate transfer function admits minimal realizations of the form

$$\Phi(s) = \begin{bmatrix} B^*(-sI_n - A^*)^{-1}, & I_m \end{bmatrix} Y_0 \begin{bmatrix} (sI_n - A)^{-1}B \\ I_m \end{bmatrix}, \tag{5.1}$$

where Y_0 is some appropriate Hermitian matrix. Note that the assumption $\Phi(s)$ proper (i.e., $\Phi(s)$ bounded at $s = \infty$) is made for the sake of simplicity and could be lifted with the help of generalized state-space representations or with an appropriate transformation of the variable s. Clearly, Y_0 is not uniquely defined from $\Phi(s)$. Indeed, replace the matrix Y_0 with the matrix $Y(\tilde{X})$ defined as follows:

$$Y(\tilde{X}) = Y_0 + \begin{bmatrix} \tilde{X}A + A^*\tilde{X} & \tilde{X}B \\ B^*\tilde{X} & 0 \end{bmatrix}, \tag{5.2}$$

where $\tilde{X}$ is any $n \times n$ block Hermitian matrix. The transfer function $\Phi(s)$ is easily verified by direct inspection not to be affected by this substitution, which clearly preserves the Hermitian property of the realization.

The well-known positive real lemma [8, 14, 18] states that the existence of a Hermitian matrix $\tilde{X}$ such that $Y(\tilde{X})$ is nonnegative definite is a necessary and sufficient condition for $\Phi(s)$ to be a para-Hermitian transfer function nonnegative on the whole of the imaginary axis. Let us apply this result to the transfer function

$$\Phi(s) = [-jE(-sI_n + jZ^T)^{-1}, I_m] Y_0 \begin{bmatrix} (sI_n - jZ)^{-1}jE^T \\ I_m \end{bmatrix}, \tag{5.3}$$

where $E = [0, \ldots, 0, I_m]$, and Y_0 is defined as in (4.5). Since $P_k = P_k^*$ for all k by assumption, $\Phi(s)$ is a well-defined paraconjugate transfer function. Moreover, one has by construction the relation

$$\Phi(jx) = x^{-2n} \sum_{k=0}^{2n} P_k x^k = x^{-2n} P(x).$$

Therefore, $\Phi(s)$ is a nonnegative paraconjugate transfer function if and only if $P(x)$ is a nonnegative polynomial matrix. In view of the positive real lemma, it finally appears that $P(x)$ is nonnegative if and only if there exists a Hermitian matrix $\tilde{X}$ such that the Hermitian matrix

$$Y(\tilde{X}) = Y_0 + \begin{bmatrix} j\tilde{X}Z - jZ^T\tilde{X} & j\tilde{X}E^T \\ -jE\tilde{X} & 0 \end{bmatrix}$$

is nonnegative definite. If one sets $X_0 \doteq -j\tilde{X}$, this is precisely the characterization provided by Theorems 4.1 and 4.2.

An alternative proof of Theorems 4.3 and 4.4 can be obtained on the basis of a similar argument. Consider a state-space realization of a paraconjugate transfer function of the form

$$\Phi(z) = \begin{bmatrix} zB^*(I_n - zA^*)^{-1}, & I_m \end{bmatrix} Y_0 \begin{bmatrix} (zI_n - A)^{-1}B \\ I_m \end{bmatrix} \tag{5.4}$$

with Y_0 some Hermitian matrix. Incidentally, this realization can also be deduced from (5.1) by means of the variable transformation $s = (z-1)/(z+1)$, which maps the unit circle onto the imaginary axis. The transfer function $\Phi(z)$ is nonnegative on the unit circle if the matrix $\Phi(e^{j\theta})$ is nonnegative definite for all θ in the interval $[0, 2\pi]$. In this setting, the positive real lemma states that $\Phi(z)$ will be a well-defined nonnegative paraconjugate transfer function if and only there exists a Hermitian matrix $\tilde{X}$ such that

$$Y(\tilde{X}) = Y_0 + \begin{bmatrix} A^*\tilde{X}A - \tilde{X} & A^*\tilde{X}B \\ B^*\tilde{X}A & B^*\tilde{X}B \end{bmatrix} \tag{5.5}$$

is nonnegative definite. With Y_0 as in (4.12), $A = Z$, and $B = E^T$, the following equality holds:

$$\Phi(z) = \sum_{k=-n}^{+n} P_k z^k.$$

Therefore, the pseudopolynomial matrix $P(z)$ is found to be nonnegative definite on the unit circle if and only if there exists a Hermitian matrix $\tilde{X}$ such that the matrix

$$Y(\tilde{X}) = Y_0 + \begin{bmatrix} Z^T\tilde{X}Z - \tilde{X} & Z^T\tilde{X}E^T \\ E\tilde{X}Z & E\tilde{X}E^T \end{bmatrix}$$

is nonnegative definite. Here again, this is exactly the characterization proposed in the previous section provided one substitutes $\tilde{X}$ for $-X_0$.

6. The optimization problem. The optimization problems considered in this paper are assumed to be stated in terms of appropriate scalar products defined over the space of complex matrices. For any couple of matrices X and Y let us set their scalar product as follows:

$$\langle X, Y\rangle \doteq \mathrm{Re}(\mathrm{Trace}\, XY^*) \equiv \mathrm{Re}\sum_i \sum_j x_{i,j}\overline{y}_{i,j}, \tag{6.1}$$

where $x_{i,j}$ and $y_{i,j}$ are the scalar entries of the matrices X and Y, respectively. It follows from this definition that

$$\langle X, Y\rangle = \langle \mathrm{Re}(X), \mathrm{Re}(Y)\rangle + \langle \mathrm{Im}(X), \mathrm{Im}(Y)\rangle.$$

Since this scalar product induces the Frobenius norm, i.e., $\|X\|_F^2 = \langle X, X\rangle$, it is called the Frobenius scalar product in what follows. If X and Y are partitioned conformably into blocks $X_{i,j}$ and $Y_{i,j}$, the above relation entails, in particular, the identity

$$\langle X, Y\rangle = \sum_i \sum_j \langle X_{i,j}, Y_{i,j}\rangle.$$

Let us now formulate several classes of optimization problems. Each class is defined on a particular curve of the complex plane and requires the definition of an inner product that is conformable with the above definition.

6.1. Real axis. For any couple of nonnegative polynomials $P(x) = \sum_{k=0}^{2n} P_k x^k$ and $Q(x) = \sum_{k=0}^{2n} Q_k x^k$, let us define their scalar product $\langle P, Q\rangle_{\mathbb{R}}$ as follows:

$$\langle P, Q\rangle_{\mathbb{R}} = \sum_{k=0}^{2n} \langle P_k, Q_k\rangle.$$

Several important optimization problems can be formulated in the following standard form:

$$\min_{P\in\mathcal{K}_{\mathbb{R}}} \{\langle C, P\rangle_{\mathbb{R}} : \langle A_\ell, P\rangle_{\mathbb{R}} = b_\ell, \ell = 1, \ldots, q\}, \tag{6.2}$$

for given C, A_ℓ, and b_ℓ, and where $\mathcal{K}_{\mathbb{R}}$ is the cone of matrix coefficients

$$P \doteq [P_0,\ P_1, \ldots,\ P_{2n}\,]$$

of the polynomial matrix $P(x)$ which is nonnegative on the real axis, i.e.,

$$P(x) \succeq 0, \quad x \in \mathbb{R}.$$

As $P \in \mathcal{K}_{\mathbb{R}}$ necessarily implies $P_k = P_k^*$ for all k, we are not restricted to assuming that all the $m \times m$ blocks C_k of C and blocks $A_{\ell,k}$ of A_ℓ are Hermitian as well, since the anti-Hermitian part of these matrices would disappear anyway in the scalar products. As shown in the preceding section, P belongs to the cone $\mathcal{K}_{\mathbb{R}}$ if and only if there exists a nonnegative block matrix Y with blocks $Y_{i,j}, i, j = 0, \ldots n$, of dimension $m \times m$ satisfying

$$P_k = \sum_{i+j=k} Y_{i,j}, \quad k = 0, 1, \ldots, 2n. \tag{6.3}$$

By definition, the dual cone $\mathcal{K}_{\mathbb{R}}^*$ is the set of the matrix coefficients $Q \doteq [Q_0,\ Q_1, \ldots,\ Q_{2n}\,]$ of the para-Hermitian matrix polynomials satisfying the constraint

$$\langle Q, P\rangle_{\mathbb{R}} \geq 0 \quad \forall P \in \mathcal{K}_{\mathbb{R}}.$$

If $H(Q)$ denotes the block Hankel matrix

$$H(Q) \doteq \begin{bmatrix} Q_0 & Q_1 & \cdots & Q_n \\ Q_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & Q_{2n-1} \\ Q_n & \cdots & Q_{2n-1} & Q_{2n} \end{bmatrix}, \tag{6.4}$$

the properties of the scalar product and (6.3) allow one to write the equalities

$$\begin{aligned} \langle Q, P\rangle_{\mathbb{R}} &= \sum_{k=0}^{2n} \langle Q_k, P_k\rangle = \sum_{k=0}^{2n} \sum_{i+j=k} \langle Q_k, Y_{i,j}\rangle \\ &= \langle H(Q), Y\rangle. \end{aligned}$$

Moreover, the following equivalence is well known ("Fejer's theorem"; see [7]):

$$\langle H(Q), Y\rangle \geq 0 \quad \forall Y \succeq 0 \iff H(Q) \succeq 0.$$

Therefore the dual cone $\mathcal{K}_{\mathbb{R}}^*$ is characterized by $H(Q) \succeq 0$.

As a consequence, the optimization problem (6.2) can be restated in its dual form,

$$\max_{u_1,\ldots,u_q} \left\{ \sum_{\ell=1}^{q} b_\ell u_\ell : H\left(C - \sum_{\ell=1}^{q} u_\ell A_\ell\right) \succeq 0 \right\}. \tag{6.5}$$

From a numerical point of view, dual formulation (6.5) has a considerable advantage over the primal form (6.2) since it involves an optimization scheme in a space of variables of dimension q rather than $(2n+1)m^2$. Any optimization problem of this type can be solved efficiently with the help of interior-point methods [13]. Their numerical implementation requires the calculation of the first and second derivatives of the barrier function

$$f(u) = -\ln\det H\left(C - \sum_{\ell=1}^{q} A_\ell u_\ell\right).$$

These derivatives can be expressed as follows:

$$\begin{aligned} \frac{\partial f(u)}{\partial u_\ell} &= \langle H(S)^{-1}, H(A_\ell)\rangle, \\ \frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} &= \langle H(S)^{-1} H(A_\ell) H(S)^{-1}, H(A_s)\rangle, \end{aligned} \tag{6.6}$$

where $S = C - \sum_{\ell=1}^{q} A_\ell u_\ell$.

6.2. Unit circle. The same property holds for optimization over the set of nonnegative pseudopolynomial matrices on the unit circle. The scalar product to be used

for pseudopolynomials $P(z) = \sum_{k=-n}^{n} P_k z^k$ and $Q(z) = \sum_{k=-n}^{n} Q_k z^k$ is defined as follows:

$$\langle P, Q \rangle_{\mathbb{C}} \doteq \sum_{k=-n}^{n} \langle P_k, Q_k \rangle.$$

The optimization problem now reads

$$\min_{P \in \mathcal{K}_{\mathbb{C}}} \{ \langle C, P \rangle_{\mathbb{C}} : \langle A_\ell, P \rangle_{\mathbb{C}} = b_\ell, \ell = 1, \ldots, q \}, \tag{6.7}$$

where $\mathcal{K}_{\mathbb{C}}$ is the cone of matrix coefficients

$$P \doteq [P_{-n}, \ldots, P_n\,]$$

of nonnegative pseudopolynomial matrices

$$P(z) \succeq 0, \quad z \in e^{j\mathbb{R}},$$

on the unit circle. Note that the coefficients of such matrices satisfy $P_{-k} = P_k^*$ and that $P \in \mathcal{K}_{\mathbb{C}}$ necessarily implies

$$P_k = \sum_{i-j=k} Y_{i,j}, \quad k = -n, \ldots, n, \tag{6.8}$$

where Y is a nonnegative block matrix with blocks $Y_{i,j}, i, j = 0, \ldots n$, of dimension $m \times m$.

As before, we are not restricted to assuming that the $m \times m$ blocks C_k of C and $m \times m$ blocks $A_{\ell,k}$ of A_ℓ have the same type of symmetry as the blocks of P, since this does not affect the scalar products.

The dual cone $\mathcal{K}_{\mathbb{C}}^*$ is made of the matrix coefficients

$$Q \doteq [Q_{-n}, \ldots, Q_n\,]$$

of the para-Hermitian pseudopolynomials satisfying the constraint

$$\langle Q, P \rangle_{\mathbb{C}} \geq 0 \quad \forall P \in \mathcal{K}_{\mathbb{C}}.$$

If $T(Q)$ denotes the block Toeplitz matrix

$$T(Q) \doteq \begin{bmatrix} Q_0 & Q_1 & \cdots & Q_n \\ Q_1^* & Q_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & Q_1 \\ Q_n^* & \cdots & Q_1^* & Q_0 \end{bmatrix}, \tag{6.9}$$

one has the relations

$$\begin{aligned} \langle Q, P \rangle_{\mathbb{C}} &= \sum_{k=-n}^{n} \langle Q_k, P_k \rangle = \sum_{k=-n}^{n} \sum_{i-j=k} \langle Q_k, Y_{i,j} \rangle \\ &= \langle T(Q), Y \rangle \end{aligned}$$

so that the dual cone $\mathcal{K}_{\mathbb{C}}^*$ is characterized by $T(Q) \succeq 0$.

Therefore the dual optimization problem (6.7) becomes

$$\max_{u_1,\ldots,u_\ell} \left\{ \sum_{\ell=1}^{q} b_\ell u_\ell : T\left(C - \sum_{\ell=1}^{q} u_\ell A_\ell\right) \succeq 0 \right\} \tag{6.10}$$

for which the appropriate barrier function is

$$f(u) = -\ln\det T\left(C - \sum_{\ell=1}^{q} A_\ell u_\ell\right).$$

As in the block Hankel case, its derivatives can be expressed as follows:

$$\begin{aligned} \frac{\partial f(u)}{\partial u_\ell} &= \langle T(S)^{-1}, T(A_\ell)\rangle, \\ \frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} &= \langle T(S)^{-1}T(A_\ell)T(S)^{-1}, T(A_s)\rangle, \end{aligned} \tag{6.11}$$

where $S = C - \sum_{\ell=1}^{q} A_\ell u_\ell$.

6.3. Imaginary axis. The imaginary case reformulation is left to the reader. As shown in the previous section, it is reducible to the real line situation in a trivial manner.

7. Computational aspects. Efficient numerical schemes to solve the optimization problems considered require repeated calculations of the differential characteristics of the barrier function, i.e., the gradient $\partial f(u)/\partial u_\ell$ and the Hessian $\partial^2 f(u)/\partial u_\ell \partial u_s$. The block Toeplitz or block Hankel structure underlying the optimization space allows one to carry out these computations in a fast, and even superfast, manner. The aim of this section is to explain this procedure in some detail.

7.1. Displacement structure. Let us first consider Hermitian $(n+1)\times(n+1)$ block Toeplitz matrices with arbitrary $m \times m$ matrix blocks T_i,

$$T \doteq \begin{bmatrix} T_0 & T_1 & \cdots & T_n \\ T_1^* & T_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_1 \\ T_n^* & \cdots & T_1^* & T_0 \end{bmatrix},$$

and $(n+1)\times(n+1)$ block Hankel matrices with Hermitian $m \times m$ matrix blocks H_i,

$$H \doteq \begin{bmatrix} H_0 & H_1 & \cdots & H_n \\ H_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & H_{2n-1} \\ H_n & \cdots & H_{2n-1} & H_{2n} \end{bmatrix}.$$

Note that T and H are defined by $(2n+1)m^2$ parameters.

Also, let us set the block permutation matrix J,

$$J \doteq \begin{bmatrix} 0 & \cdots & 0 & I_m \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ I_m & 0 & \cdots & 0 \end{bmatrix},$$

that will play a special role in the subsequent developments.

The displacement theory of Toeplitz and Hankel matrices is well established [9, 10] and is the basis underlying most fast algorithms for decomposing such matrices. Using the block shift matrix one defines a "Toeplitz displacement operator" ∇_t and a "Hankel displacement operator" ∇_h as follows:

$$\nabla_t T \doteq T - Z^T T Z, \quad \nabla_h H \doteq H - ZHZ. \tag{7.1}$$

The reader may easily check that the following equalities hold:

$$\nabla_t T = \begin{bmatrix} T_0 & T_1 & \cdots & T_n \\ T_1^* & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ T_n^* & 0 & \cdots & 0 \end{bmatrix}, \tag{7.2}$$

$$\nabla_h H = \begin{bmatrix} H_0 & 0 & \cdots & 0 \\ H_1 & \vdots & & \vdots \\ \vdots & 0 & \cdots & 0 \\ H_n & \cdots & H_{2n-1} & H_{2n} \end{bmatrix}. \tag{7.3}$$

From the above expressions, ones notices that the original matrices T and H can be recovered from their respective displacement. The inverse operators are obtained by merely applying the displacement operator again and again to both sides of (7.2) to produce

$$T = \nabla_t T + Z^T \cdot \nabla_t T \cdot Z + \cdots + Z^{nT} \cdot \nabla_t T \cdot Z^n \tag{7.4}$$

and

$$H = \nabla_h H + Z \cdot \nabla_h H \cdot Z + \cdots + Z^n \cdot \nabla_h H \cdot Z^n. \tag{7.5}$$

It is also useful to point out that both displacements are closely related to each other. Permuting the block rows of a block Hankel matrix H indeed yields a block Toeplitz matrix JH, which can be defined as T by setting $T_i = H_{i+n},\ i = -n, \ldots, n$. Since $Z^T = JZJ$, the displacement operators are related in a similar fashion as follows:

$$T = JH \iff \nabla_t T = J \nabla_h H.$$

From the sparsity structure of matrices (7.2) and (7.3) it is obvious that the ranks of $\nabla_t T$ and $\nabla_h H$ cannot be larger than $2m$. This rank is called the "displacement rank" of the corresponding matrix.

The theory of displacement ranks [9, 10] tells us that the inverse of T or H (when it exists) has the same displacement as that of the matrix itself as follows:

$$\operatorname{rank} \nabla_t^* T^{-1} = \operatorname{rank} \nabla_t T, \quad \operatorname{rank} \nabla_h H^{-1} = \operatorname{rank} \nabla_h H,$$

where ∇_t^* stands for the transposed Toeplitz displacement operator, i.e., $\nabla_t^* T^{-1} = T^{-1} - Z T^{-1} Z^T$. Since the displacement rank of a block Toeplitz or block Hankel matrix is typically much lower than the dimensions of the corresponding matrix, and since the displacement operator can be inverted, it is economical to represent such a

matrix by a rank factorization of its displacement. From the expressions (7.2), (7.3), it is simple to construct low rank factorizations of $\nabla_t T$ or $\nabla_h H$ as follows:

$$\nabla_t T = F_t^* \cdot G_t, \quad \nabla_h H = F_h^* \cdot G_h,$$

where the number of rows of F_t and G_t equals $r_t \doteq \operatorname{rank} \nabla_t T$, and the number of rows of F_h and G_h equals $r_h \doteq \operatorname{rank} \nabla_h H$.

Given such factorizations, fast generalized Schur-based algorithms can be used [9, 10] to derive from them the corresponding factorizations of the displacement of the inverses as follows:

$$\nabla_t^* T^{-1} = A_t^* \cdot B_t, \quad \nabla_h H^{-1} = A_h^* \cdot B_h,$$

and these precise decompositions are used in what follows. Moreover, as Schur algorithms can be implemented in a superfast manner by means of a divide-and-conquer strategy, the complexity of the above construction is found to be $\mathcal{O}(rm^2 n \log^2 n)$. Incidentally, let us note that these factorizations are not unique and that for positive definite matrices T and H there exist particular choices of factorizations that can benefit from these properties. For instance, one can choose in the Toeplitz case

$$G_t = \begin{bmatrix} T_0 & T_1 & T_2 & \cdots & T_n \\ 0 & -T_1 & -T_2 & \cdots & -T_n \end{bmatrix}, \tag{7.6}$$

$$F_t = \begin{bmatrix} T_0 & 0 \\ 0 & -T_0 \end{bmatrix}^{-1} G_t. \tag{7.7}$$

In what follows, these aspects will be disregarded since they only marginally affect the complexity results.

Let us focus first on the case of Toeplitz displacement of an $m(n+1) \times m(n+1)$ matrix X and suppose that a rank r_t factorization of its Toeplitz displacement $\nabla_t X$ has been computed,

$$\nabla_t X = F^* \cdot G,$$

where F and G have dimensions $r_t \times m(n+1)$. Let us also define an upper block triangular Toeplitz matrix $U(G)$ as a function of the partitioned matrix G, where each subblock has dimensions $r_t \times m$,

$$G \doteq \begin{bmatrix} G_0 & G_1 & \cdots & G_n \end{bmatrix},$$

$$U(G) \doteq \begin{bmatrix} G_0 & G_1 & \cdots & G_n \\ 0 & G_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_1 \\ 0 & \cdots & 0 & G_0 \end{bmatrix}.$$

Doing the same for the matrix F, one obtains

$$F \doteq \begin{bmatrix} F_0 & F_1 & \cdots & F_n \end{bmatrix},$$

$$U(F)^* \doteq \begin{bmatrix} F_0^* & 0 & \cdots & 0 \\ F_1^* & F_0^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ F_n^* & \cdots & F_1^* & F_0^* \end{bmatrix}.$$

It follows from the displacement equation $\nabla_t X = F^* \cdot G$ that

$$X = \sum_{j=0}^{n} (FZ^j)^*(GZ^j) = U(F)^* U(G)$$
$$= \begin{bmatrix} F_0^* & 0 & \cdots & 0 \\ F_1^* & F_0^* & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ F_n^* & \cdots & F_1^* & F_0^* \end{bmatrix} \cdot \begin{bmatrix} G_0 & G_1 & \cdots & G_n \\ 0 & G_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_1 \\ 0 & \cdots & 0 & G_0 \end{bmatrix}.$$

This formula, when applied to a particular choice of displacement factors F and G for the inverse of a Toeplitz matrix T, is also known as the Gohberg–Semencul formula for $X = T^{-1}$.

For the Hankel displacement $\nabla_h X$ of an $m(n+1) \times m(n+1)$ matrix X, there exists a similar representation starting based upon the rank r_h factorization of $\nabla_h X$,

$$\nabla_h X = F^* \cdot G,$$

where F and G have dimension $r_h \times m(n+1)$. If the matrix F is partitioned in reverse order,

$$F \doteq \begin{bmatrix} F_0 & \cdots & F_n \end{bmatrix} \iff FJ \doteq \begin{bmatrix} F_n & \cdots & F_0 \end{bmatrix},$$

then it follows from the relation $J\nabla_h X = \nabla_t (JX)$ that

$$X = J \sum_{j=0}^{n} (FJZ^j)^*(GZ^j) = JU(FJ)^* U(G) \tag{7.8}$$
$$= \begin{bmatrix} F_0^* & \cdots & F_{n-1}^* & F_n^* \\ \vdots & \iddots & \iddots & 0 \\ F_{n-1}^* & F_n^* & \iddots & \vdots \\ F_n^* & 0 & \cdots & 0 \end{bmatrix} \cdot \begin{bmatrix} G_0 & G_1 & \cdots & G_n \\ 0 & G_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_1 \\ 0 & \cdots & 0 & G_0 \end{bmatrix}. \tag{7.9}$$

When applied to a particular choice of displacement factors F and G for the inverse of a Hankel matrix, this formula is also known as the Christoffel–Darboux formula for $X = H^{-1}$.

7.2. Implementation. The numerical solution of the optimization problem considered in section 6 requires evaluations of the gradient $\partial f(u)/\partial u_\ell$ and the Hessian $\partial^2 f(u)/\partial u_\ell \partial u_s$ as given by (6.6) or (6.11). Let us now focus on the fast computation of these elements using the displacement techniques mentioned above.

Consider the inner product $\langle X, T(A_s) \rangle$ which appears in (6.11) with $X = T(S)^{-1}$ or $X = T(S)^{-1} T(A_l) T(S)^{-1}$, and let $\operatorname{diag}\{W\}$ be the block diagonal matrix with all blocks equal to $W \in \mathbb{C}^{m \times m}$. Since

$$T(A_s) = \operatorname{diag}\{A_{s,0}\} + \sum_{k=0}^{n} [Z^k \operatorname{diag}\{A_{s,k}\} + (Z^k)^T \operatorname{diag}\{A_{s,k}^*\}],$$

the computation can be broken down into a summation of scalar products of the type

$$\langle X, Z^i \operatorname{diag}\{W\} \rangle, \quad \langle X, (Z^i)^T \operatorname{diag}\{W^*\} \rangle. \tag{7.10}$$

For Hermitian matrices X, it turns out that $\langle X, (Z^i)^T \operatorname{diag}\{W^*\}\rangle = \langle X, Z^i \operatorname{diag}\{W\}\rangle$ so that only one expression has to be evaluated.

Similarly, the inner product $\langle X, H(A_s)\rangle$ which appears in (6.6) with $X = H(S)^{-1}$ or $X = H(S)^{-1}H(A_l)H(S)^{-1}$ requires the evaluation of scalar products of the type

$$\langle X, JZ^i \operatorname{diag}\{W_1\}\rangle, \quad \langle X, J(Z^i)^T \operatorname{diag}\{W_2\}\rangle, \tag{7.11}$$

where W_1 and W_2 are Hermitian matrices of order m.

In addition, since the matrices X can be described by their Hankel or Toeplitz displacement, one can speed up the computation of (7.10) and (7.11). Let us first consider matrices X given by their Toeplitz displacement $\nabla_t X = F^* \cdot G$. Since

$$U(F) = \sum_{k=0}^{n} Z^k \operatorname{diag}\{F_k\}, \quad U(G) = \sum_{k=0}^{n} Z^k \operatorname{diag}\{G_k\},$$

and as

$$\langle Z^j \operatorname{diag}\{X\}, Z^i \operatorname{diag}\{Y\}\rangle = \delta_{i,j}(n+1-i)\langle X, Y\rangle,$$

one obtains the expression

$$\begin{aligned}\langle U(F)^*U(G), Z^j \operatorname{diag}\{W\}\rangle &= \langle (n+1-j)F_j^*G_0 + \cdots + 2F_{n-1}^*G_{n-j-1} + F_n^*G_{n-j},\, W\rangle \\ &\doteq \langle M_j, W\rangle.\end{aligned}$$

Since the matrix $X = U(F)^*U(G)$ is Hermitian, the roles of F_i and G_i can be interchanged in the above formula. Moreover, the quantities $\{M_j\}_{j=0}^n$ can be evaluated as the convolution of the block vectors

$$[(n+1)F_0, nF_1, \ldots, 2F_{n-1}, F_n], \quad [G_0, G_1, \ldots, G_{n-1}, G_n],$$

which has a complexity of $\mathcal{O}(r_t m^2 n \log_2 n)$ flops [10]. As the computation of the inner product $\langle M_j, W\rangle$ requires $\mathcal{O}(m^2)$ operations, the overall complexity of computing $\langle X, T(A_s)\rangle$ is thus found to be $\mathcal{O}(r_t m^2 n \log_2 n + m^2 n)$ flops for a matrix of displacement rank r_t, *provided that* the matrices F and G are given. If the matrix X is given by its transposed displacement $\nabla_t^* X = A^* \cdot B$, one can easily adapt the above formula and check that the overall complexity is also $\mathcal{O}(r_t m^2 n \log_2 n + m^2 n)$ flops, *provided that* the matrices A and B are given.

The calculations involving the Hessian, i.e., when $X = T(S)^{-1}T(A_\ell)T(S)^{-1}$, require some elaboration. With the matrix $\hat{T}$ defined by

$$\hat{T} = \begin{bmatrix} -T(A_\ell) & T(S) \\ T(S) & 0 \end{bmatrix},$$

note first that the following relation holds:

$$\hat{T}^{-1} = \begin{bmatrix} 0 & T(S)^{-1} \\ T(S)^{-1} & X \end{bmatrix}.$$

Furthermore, as $T(S)$ and $T(A_\ell)$ are block Toeplitz matrices, the rank of the matrix factors F and G in the block displacement equation

$$\nabla_t \hat{T} = \hat{T} - \begin{bmatrix} Z^T & 0 \\ 0 & Z^T \end{bmatrix} \hat{T} \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} = F^*G$$

is equal to $4m$, as is easily verified. The corresponding factorization of the block displacement of the inverse can be achieved at low computational cost in the form

$$\nabla_t^* \hat{T}^{-1} = \hat{T}^{-1} - \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} \hat{T}^{-1} \begin{bmatrix} Z^T & 0 \\ 0 & Z^T \end{bmatrix} = [A_1, A_2]^* \cdot [B_1, B_2].$$

Therefore, the expression of the transposed Toeplitz displacement of X is given by

$$\nabla_t^* X = A_2^* \cdot B_2.$$

The formalism described above for the fast computation of the relevant inner products can therefore be applied to construct the entries of the Hessian (6.11). If the displacement factors are computed using a superfast algorithm, the overall complexity of constructing the Hessian is therefore $\mathcal{O}(qr_t m^2 n \log^2 n + q^2 m^2 n)$.

Let us now consider matrices X given by their Hankel displacement $\nabla_h X = F^* \cdot G$. The inner products of interest can be rewritten in terms of JX as follows:

$$\begin{aligned} \langle X, JZ^i \operatorname{diag}\{W_1\}\rangle &= \langle JX, Z^i \operatorname{diag}\{W_1\}\rangle, \\ \langle X, J(Z^i)^T \operatorname{diag}\{W_2\}\rangle &= \langle (JX)^*, Z^i \operatorname{diag}\{W_2\}\rangle, \end{aligned}$$

where W_1 and W_2 are Hermitian matrices of order m. Since JX is block Toeplitz, the above formulas could, in theory, be applied mutatis mutandis. From a practical viewpoint, however, this does not make much sense. As explained in the next section, the Hankel setting of the optimization problem considered is numerically ill-conditioned. Hence, the problem formulation itself needs to be redesigned so as to circumvent this inherent difficulty. This issue is addressed in the next section.

The actual solution of the optimization problem of section 6 is often achieved with the help of an iterative Newton scheme. In particular, this iterative process requires frequent evaluations of the so-called Newton directions, which involve the product of the inverse of the current Hessian by an appropriate given vector. From a practical viewpoint, this approach is efficient only if the Hessian dimension q is small. Otherwise, a conjugate gradient scheme could be more attractive since it does not require the inversion of the Hessian but rather its product with a vector. Such computations can be made at low cost with the help of the inner product formalism explained in the present section.

Let us briefly clarify this issue. Assume that the optimization problem is defined on the unit circle, and consider the product of the Hessian by a vector x to yield a vector y. By definition, one has in view of (6.11) that the sth component of y is given by

$$\begin{aligned} y_s &= \sum_\ell \frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} x_\ell, \\ &= \sum_\ell \langle T(S)^{-1} T(A_\ell) T(S)^{-1}, T(A_s)\rangle x_\ell, \\ &= \langle T(S)^{-1} T(D) T(S)^{-1}, T(A_s)\rangle, \\ &= \langle T(S)^{-1} T(A_s) T(S)^{-1}, T(D)\rangle, \end{aligned}$$

where $T(D)$ stands for the block Toeplitz matrix $T(D) = \sum_\ell T(A_\ell x_\ell)$. Expressions of this type can be computed efficiently using the results derived above in this section. Performing k conjugate gradient steps at each Newton iteration therefore requires $\mathcal{O}(qr_t m^2 n \log^2 n + kqm^2 n)$ operations.

7.3. Complexity of the optimization scheme. Since interior-points methods require $\mathcal{O}(\sqrt{nm}\log\frac{1}{\epsilon})$ Newton steps to solve the optimization problems (6.5) and (6.10) up to an accuracy ϵ [13], the overall complexity of solving these problems depends on the method used to compute the Newton directions and is found to be

- $\mathcal{O}(\sqrt{nm}\log\frac{1}{\epsilon}[qr_tm^2n\log^2 n + q^2m^2n + q^3])$ flops for the "inversion" of the Hessian;
- $\mathcal{O}(\sqrt{nm}\log\frac{1}{\epsilon}[qr_tm^2n\log^2 n + kqm^2n])$ flops for the conjugate gradient scheme.

By solving the dual problem and using the matrix structures, we get a remarkable result for solving an optimization problem in a $(2n+1)m^2$-dimensional vector space, subject to q linear constraints and m semi-infinite inequality constraints (see (6.2) and (6.7)).

In particular, for nonnegative *scalar* polynomials, i.e., $m = 1$, each Newton iteration requires $\mathcal{O}(qn(\log^2 n + q) + q^3)$ and $\mathcal{O}(qn(\log^2 n + k))$, respectively.

8. Chebyshev reformulation of the real line optimization problem. The formulation of the real line optimization problem exhibits a serious drawback: it involves positive definite Hankel matrices, which are numerically ill-conditioned [3, 15]. The celebrated Hilbert matrix is a good illustration of this fact. More generally, the Euclidean condition number $\kappa(H)$ of any positive definite Hankel matrix H of order $n+1$ was shown recently [3] to be bounded from below by

$$\kappa(H) \geq \frac{(1.792)^{2n}}{16(n+1)}, \quad n \geq 2.$$

Therefore, solving the real line optimization problem as considered in section 6 is inherently hazardous, and all the more so if the problem dimension is large. To get around this, let us first observe that the occurrence of the block Hankel structure originates from the choice of the natural powers $1, x, x^2, \ldots$ as a basis for describing the optimization space of the polynomial matrices $P(x) = \sum_{k=0}^{n} P_k x^k$, positive semidefinite on the real line. Obviously, other choices are possible. In this section, the alternative use of a basis of Chebyshev polynomials to describe the optimization is specifically investigated together with the consequences of this choice.

The first order Chebyshev polynomials $T_k(x)$ are well known to satisfy, for $k \geq 1$, the recurrence formula

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$$

initialized with $T_0(x) = 1$ and $T_1(x) = x$. In particular, one has the relation

$$T_i(x)T_j(x) = \frac{1}{2}[T_{i+j}(x) + T_{|i-j|}(x)] \quad \forall\, i, j \geq 0. \tag{8.1}$$

In order to emphasize our choice of the Chebyshev basis, let us denote by $\tilde{P}_k = \tilde{P}_k^*$ the matrix coordinates of any para-Hermitian polynomial matrix $P(x)$ in this basis, i.e.,

$$P(x) = \sum_{k=0}^{2n} \tilde{P}_k T_k(x).$$

Using the notation introduced in section 4, let us consider the set of Hermitian matrices Y such that one has the identity

$$\tilde{\Pi}^*(x) Y \tilde{\Pi}(x) = P(x) \tag{8.2}$$

with $\tilde{\Pi}(x) = [T_0(x)I_m,\ T_1(x)I_m, \ldots, T_n(x)I_m]^T$.

The algebraic constraints on Y implied by (8.2) can be expressed in terms of the Chebyshev basis in a simple manner. Using identities (8.1), one can easily check that the following relations hold:

$$\tilde{P}_k = \frac{1}{2}\left(\sum_{i+j=k} Y_{i,j} + \sum_{|i-j|=k} Y_{i,j}\right) \qquad \text{for } k = 0, \ldots, 2n. \tag{8.3}$$

If L stands for the block lower triangular matrix transforming $\Pi(x)$ into $\tilde{\Pi}(x)$, i.e., $\tilde{\Pi}(x) = L\,\Pi(x)$, a simple consequence of Theorem 4.1 is that the set of all solutions Y to (8.2) is parametrized by the relation

$$Y = Y_0 + L^{-T}(Z\,X - X\,Z^T)L^{-1},$$

where Y_0 is a particular solution of (8.2) and X is any skew-Hermitian matrix of form (4.1). Moreover, if Theorem 4.2 is applied to $L^T\,Y\,L$, the existence of a positive definite solution Y to (8.2) is also found to be the necessary and sufficient condition such that $P(x)$ is a well-defined positive polynomial matrix on the real line. Note incidentally that a particular matrix Y_0 satisfying (8.2) is provided by

$$Y_0 = \begin{bmatrix} \frac{1}{2}\tilde{P}_0 - \sum_{k=1}^{n} \tilde{P}_{2k} & \frac{1}{2}\tilde{P}_1 - \frac{1}{2}\sum_{k=2}^{n} -\tilde{P}_{2k-1} & & & \\ \frac{1}{2}\tilde{P}_1 - \frac{1}{2}\sum_{k=2}^{n} \tilde{P}_{2k-1} & 2\tilde{P}_2 & \tilde{P}_3 & & \\ & \tilde{P}_3 & 2\tilde{P}_4 & \ddots & \\ & & \ddots & \ddots & \tilde{P}_{2n-1} \\ & & & \tilde{P}_{2n-1} & 2\tilde{P}_{2n} \end{bmatrix}.$$

When such a Chebyshev basis is chosen, the optimization space is transformed into the convex cone $\tilde{\mathcal{K}}_{\mathbb{R}}$ of matrix coefficients

$$\tilde{P} \doteq [\tilde{P}_0, \tilde{P}_1, \ldots, \tilde{P}_{2n}]$$

of the polynomial matrices nonnegative definite on the real axis, i.e.,

$$P(x) \succeq 0, \quad x \in \mathbb{R}.$$

Furthermore and as shown above, $\tilde{P}$ belongs to the cone $\tilde{\mathcal{K}}_{\mathbb{R}}$ if and only if there exists a nonnegative block matrix Y with blocks $Y_{i,j}$ satisfying (8.3). By definition, the dual cone $\tilde{\mathcal{K}}^*_{\mathbb{R}}$ consists of the matrix coefficients $\tilde{Q} \doteq [\tilde{Q}_0, \tilde{Q}_1, \ldots, \tilde{Q}_{2n}]$ satisfying the constraints

$$\langle \tilde{Q}, \tilde{P} \rangle_{\mathbb{R}} \geq 0 \quad \forall \tilde{P} \in \tilde{\mathcal{K}}_{\mathbb{R}}.$$

Recall that we are not restricted to assuming that the matrix coefficients $\tilde{P}_k$ and $\tilde{Q}_k$ are Hermitian for all k. For any $\tilde{Q}$, let us set the block Toeplitz-plus-Hankel matrix

$$T_H(\tilde{Q}) \doteq \begin{bmatrix} \tilde{Q}_0 & \tilde{Q}_1 & \cdots & \tilde{Q}_n \\ \tilde{Q}_1 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \tilde{Q}_{2n-1} \\ \tilde{Q}_n & \cdots & \tilde{Q}_{2n-1} & \tilde{Q}_{2n} \end{bmatrix} + \begin{bmatrix} \tilde{Q}_0 & \tilde{Q}_1 & \cdots & \tilde{Q}_n \\ \tilde{Q}_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{Q}_1 \\ \tilde{Q}_n & \cdots & \tilde{Q}_1 & \tilde{Q}_0 \end{bmatrix}. \tag{8.4}$$

In view of (8.3) and the scalar product definition, one derives the relation

$$\begin{aligned}\langle \tilde{P}, \tilde{Q}\rangle_{\mathbb{R}} &= \sum_{k=0}^{2n} \langle \tilde{Q}_k, \tilde{P}_k\rangle \\ &= \frac{1}{2}\sum_{k=0}^{2n}\left[\sum_{i+j=k}\langle \tilde{Q}_k, Y_{i,j}\rangle + \sum_{|i-j|=k}\langle \tilde{Q}_k, Y_{i,j}\rangle\right] \\ &= \frac{1}{2}\langle T_H(\tilde{Q}), Y\rangle,\end{aligned}$$

which shows that the dual cone $\mathcal{K}_{\mathbb{R}}^*$ is characterized by $T_H(\tilde{Q}) \succeq 0$.

Therefore, the dual form of the optimization problem (6.2) can be expressed in the present case as

$$\max_{u_1,\ldots,u_\ell}\left\{\sum_{\ell=1}^{q} b_\ell u_\ell : T_H\left(C - \sum_{\ell=1}^{q} u_\ell A_\ell\right) \succeq 0\right\}. \tag{8.5}$$

The corresponding barrier function is $f(u) = -\ln\det T_H(C - \sum_{\ell=1}^{q} u_\ell A_\ell)$ and the differential characteristics of interest now read

$$\begin{aligned}\frac{\partial f(u)}{\partial u_\ell} &= \langle T_H(S)^{-1}, TH(A_\ell)\rangle, \\ \frac{\partial^2 f(u)}{\partial u_\ell \partial u_s} &= \langle T_H(S)^{-1} T_H(A_\ell)\, T_H(S)^{-1}, T_H(A_s)\rangle,\end{aligned} \tag{8.6}$$

where $S = C - \sum_{\ell=1}^{q} A_\ell u_\ell$.

From a numerical viewpoint, this reformulation of the optimization problem on the real line exhibits a considerable advantage over its initial formulation in the sense that is not intrinsically ill-conditioned. Indeed, for all degrees n there exist nonnegative matrices $T_H(\tilde{Q})$ with a condition number equal to 2, as illustrated by the trivial example $\tilde{Q} = [I_m, 0, \ldots, 0]$. As a result, the numerical behavior of the computational optimization scheme is expected to be substantially improved.

Finally, let us point out that the differential characteristics of the Chebyshev basis reformulated barrier function (8.6) can also be computed in a fast way with the help of displacement techniques. This problem is not a straightforward generalization of the results presented in this paper. Nevertheless one expects to apply, as above, a divide-and-conquer strategy to get low complexity algorithms.

9. Conclusion. Cones of positive pseudopolynomial matrices are often encountered in practice as well as the corresponding dual cones, which are related to moment spaces. In this paper semidefinite representation of these cones is shown to be interesting from a computational viewpoint. In particular the dual optimization problems can be solved very efficiently using displacement-based factorizations as well as an appropriate divide-and-conquer strategy. These results are direct consequences of the Hankel or Toeplitz structure in the dual constraints.

During the review process Alkire and Vandenberghe [2] obtained an algorithm to solve optimization problems involving autocorrelation sequences. The associated cone consists of nonnegative cosine polynomials, which are particular pseudopolynomials. In their case the barrier function $f(u)$ is thus equal to the logarithmic barrier of a Toeplitz matrix $T(u)$. As the Levinson–Durbin algorithm is applied to factor the

inverse Toeplitz matrix and DFT is then applied to assemble the gradient and the Hessian, the complexity of one iteration in their scheme is equal to $\mathcal{O}(n^3)$. Although this method is similar to the one proposed in this paper (if applied to this particular setting) the techniques presented above are more general. On the one hand, they can be applied to structured matrices with low displacement rank, in particular, block Hankel or Toeplitz. On the other hand, we consider the generic setting of conic optimization problems, for which the barrier function is more general.

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Oliver and Boyd, Edinburgh, 1965.

[2] B. Alkire and L. Vandenberghe, *Convex optimization problems involving finite autocorrelation sequences*, Math. Program., (2002), 93 (2002), pp. 331–359.

[3] B. Beckermann, *The condition number of real Vandermonde, Krylov and positive definite Hankel matrices*, Numer. Math., 85 (2000), pp. 553–577.

[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Stud. Appl. Math. 15, SIAM, Philadelphia, PA, 1994.

[5] Y. Genin, Y. Nesterov, and P. Van Dooren, *Positive transfer functions and convex optimization*, in Proceedings of the European Control Conference ECC-99, Karlsruhe, Germany, 1999. Paper F-143.

[6] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Math. Res. 19, Akademie-Verlag, Berlin, 1984.

[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[8] V. Ionescu, C. Oară, and M. Weiss, *Generalized Riccati Theory and Robust Control: A Popov Function Approach*, John Wiley, Chichester, UK, 1999.

[9] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[10] T. Kailath and A. H. Sayed, eds., *Fast Reliable Algorithms for Matrices with Structure*, SIAM, Philadelphia, PA, 1999.

[11] S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Pure Appl. Math. 15, Wiley Interscience, New York, 1966.

[12] Y. Nesterov, *Squared functional systems and optimization problems*, in High Performance Optimization, Appl. Optim. 33, Kluwer, Dordrecht, 2000, pp. 405–440.

[13] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Math. 13, SIAM, Philadelphia, PA, 1994.

[14] V.-M. Popov, *Hyperstability of Control Systems*, Grundlehren Math. Wiss. 204, Springer-Verlag, Berlin, 1973.

[15] E. E. Tyrtyshnikov, *How bad are Hankel matrices?* Numer. Math., 67 (1994), pp. 261–269.

[16] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[17] D. C. Youla, *On the factorization of rational matrices*, IRE Trans. Inform. Theory, 7 (1961), pp. 172–189.

[18] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 80–104

OPTIMIZATION AND PSEUDOSPECTRA, WITH APPLICATIONS TO ROBUST STABILITY*

J. V. BURKE†, A. S. LEWIS‡, AND M. L. OVERTON§

Abstract. The ϵ-pseudospectrum of a matrix A is the subset of the complex plane consisting of all eigenvalues of all complex matrices within a distance ϵ of A. We are interested in *two* aspects of "optimization and pseudospectra." The first concerns *maximizing* the function "real part" over an ϵ-pseudospectrum of a fixed matrix: this defines a function known as the ϵ-pseudospectral abscissa of a matrix. We present a bisection algorithm to compute this function. Our second interest is in *minimizing* the ϵ-pseudospectral abscissa over a set of feasible matrices. A prerequisite for local optimization of this function is an understanding of its variational properties, the study of which is the main focus of the paper. We show that, in a neighborhood of any nonderogatory matrix, the ϵ-pseudospectral abscissa is a nonsmooth but locally Lipschitz and subdifferentially regular function for sufficiently small ϵ; in fact, it can be expressed locally as the maximum of a finite number of smooth functions. Along the way we obtain an eigenvalue perturbation result: near a nonderogatory matrix, the eigenvalues satisfy a Hölder continuity property on matrix space—a property that is well known when only a single perturbation parameter is considered. The pseudospectral abscissa is a powerful modeling tool: not only is it a robust measure of stability, but it also reveals the transient (as opposed to asymptotic) behavior of associated dynamical systems.

Key words. pseudospectrum, eigenvalue optimization, spectral abscissa, nonsmooth analysis, subdifferential regularity, robust optimization, robust control, stability radius, distance to instability, $\mathbf{H}_\infty$ norm

AMS subject classifications. Primary, 15A18, 65K05; Secondary, 90C30, 93D09

PII. S0895479802402818

1. Introduction.

The ϵ-pseudospectrum of a matrix A, denoted $\Lambda_\epsilon(A)$, is the subset of the complex plane consisting of all eigenvalues of all complex matrices within a distance ϵ of A (see [20, 39, 40]). We are interested in *two* aspects of "optimization and pseudospectra." The first concerns *maximizing* a simple real-valued function over a fixed pseudospectrum $\Lambda_\epsilon(A)$. We focus specifically on the case where this function is simply "real part." Then the optimal value defines the ϵ-pseudospectral abscissa of A, denoted $\alpha_\epsilon(A)$. Just as the spectral abscissa of a matrix provides a measure of its stability, that is, the asymptotic decay of associated dynamical systems, so the ϵ-pseudospectral abscissa provides a measure of *robust stability*, where by robust we mean with respect to complex perturbations in the matrix. One of the contributions of this paper is a bisection algorithm that computes $\alpha_\epsilon(A)$ for any A; this algorithm also identifies all maximizing points in the pseudospectrum.

In many applications, matrices are not fixed but dependent on parameters that may be adjusted. Our second interest in optimization concerns *minimizing* the ϵ-pseudospectral abscissa α_ϵ over a feasible set of matrices. A prerequisite for local

*Received by the editors February 15, 2002; accepted for publication (in revised form) by L. Vandenberghe January 3, 2003; published electronically May 15, 2003.

http://www.siam.org/journals/simax/25-1/40281.html

†Department of Mathematics, University of Washington, Seattle, WA 98195 (burke@math.washington.edu). The research of this author was supported in part by National Science Foundation grant DMS-9971852.

‡Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (aslewis@sfu.ca, www.cecm.sfu.ca/~aslewis). The research of this author was supported in part by NSERC.

§Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (overton@cs.nyu.edu). The research of this author was supported in part by National Science Foundation grant CCR-0098145.

minimization of α_ϵ is an understanding of its variational properties as a function of the matrix A. This provides the focus for most of the paper. Our main result shows that, in a neighborhood of any nonderogatory matrix, the ϵ-pseudospectral abscissa is a nonsmooth but locally Lipschitz and subdifferentially regular function for sufficiently small ϵ; in fact, it can be expressed locally as the maximum of a finite number of smooth functions. Such a property is desirable from the point of view of numerical methods for local optimization, but we defer a computational study to future work.

The paper is organized as follows. After setting up some notation in section 2, we begin in section 3 by discussing related ideas in the robust control literature. We review the connections between the pseudospectral abscissa and the "distance to instability" [28, 16], or "complex stability radius" [21], and the $\mathbf{H}_\infty$ norm of a transfer function [8]. The outcome of minimization of α_ϵ over a set of matrices obviously depends on the crucial issue of the choice of ϵ. We show that as ϵ is increased from zero to an arbitrarily large quantity, the corresponding optimization problem evolves from minimization of the spectral abscissa (enhancing the *asymptotic* decay rate of the associated dynamical system) to the minimization of the largest eigenvalue of the symmetric part of the matrix (minimizing the *initial* growth rate of the associated system). Regarding the first of these extremes (optimization of the spectral abscissa), variational analysis of this non-Lipschitz function is well understood [15, 12], global optimization is known to be hard [5], and some progress has been made in local optimization methods [14]. Regarding the second extreme, optimization of this convex function over a polyhedral feasible set is a semidefinite programming problem, and the global minimum can be found by standard methods [4, 38]. Intermediate choices of ϵ control *transient* peaking in the dynamical system associated with the matrix, and one particular choice corresponds exactly to the complex stability radius (or $\mathbf{H}_\infty$ norm) optimization problem. Thus the pseudospectral approach gives a whole range of stabilizing optimization problems, each with a quantifiable interpretation in terms of the allowable perturbations. Furthermore, unlike maximization of the complex stability radius, which simply optimizes the "robustness" of the stability, minimizing the pseudospectral abscissa preserves some explicit emphasis on optimizing the asymptotic decay rate of the system.

In section 4, we analyze the topology of the pseudospectrum, observing that points on the boundary are accessible from the interior by analytic paths, and discussing conditions under which the boundary is differentiable at points that maximize the real part. This sets the stage for the description of a simple bisection algorithm to compute $\alpha_\epsilon(A)$, the pseudospectral abscissa of a fixed matrix, in section 5. In section 6, we show that the bisection algorithm locates *all* maximizers of the real part over the pseudospectrum. The bisection algorithm is very much analogous to Byers' algorithm for measuring the distance to instability [16], which has spawned more sophisticated variants for the calculation of stability radii (real as well as complex) and $\mathbf{H}_\infty$ norms, both globally and quadratically (or higher order) convergent; see [8, 7, 11, 18, 36]. Along similar lines, we have also developed a quadratically convergent variant algorithm for computing α_ϵ, described and analyzed in a companion paper [13].

In section 7, we continue our study of analytical properties of the pseudospectrum. As is well known, the pseudospectrum of a matrix A is defined by an inequality on $\sigma_{\min}(A - zI)$, the least singular value of $A - zI$. In Theorem 7.4 (growth near an eigenvalue) we give an interesting estimate relating $\sigma_{\min}(A - zI)$ to $|z - \lambda_0|^m$, where

λ_0 is a nonderogatory eigenvalue (one whose geometric multiplicity is one), and m is its algebraic multiplicity. The coefficient relating these quantities is a ratio of two products, of eigenvalue separations and of singular values, respectively. One corollary of this result is that an ϵ-pseudospectrum component around a nonderogatory eigenvalue is strictly convex for sufficiently small ϵ, an intuitively appealing but apparently nontrivial fact. Another corollary is that the nonderogatory eigenvalue λ_0 satisfies a Hölder continuity property in a neighborhood of A in matrix space, with Hölder exponent equal to $1/m$. While this result might not seem surprising, in light of well-known classical spectral perturbation theory [27, 23, 3, 30], we have not seen it in the literature. The classical analysis focuses almost exclusively on single perturbation parameters.

The analytical results of section 7 allow us to achieve our primary goal in section 8: a detailed variational analysis of the pseudospectral abscissa α_ϵ. The main result has already been mentioned above. Finally, in section 9, we examine the boundary properties of the pseudospectrum at points where the boundary is not smooth, using techniques from modern variational analysis [17, 34]. We show that, under a nondegeneracy condition, the complement of the pseudospectrum is Clarke regular at such a point, and give a formula for the normal cone.

2. Notation. We consider a matrix A in the space of $n \times n$ complex matrices $\mathbf{M}^n$. We denote the spectrum of A by $\Lambda = \Lambda(A)$, and we denote by $\alpha = \alpha(A)$ the *spectral abscissa* of A, which is the largest of the real parts of the eigenvalues.

For a real $\epsilon > 0$, the *ϵ-pseudospectrum* of A is the set

$$\Lambda_\epsilon = \{z \in \mathbf{C} : z \in \Lambda(X) \text{ where } \|X - A\| \leq \epsilon\}.$$

(Throughout, $\|\cdot\|$ denotes the operator 2-norm on $\mathbf{M}^n$.) For the most part, ϵ is fixed, so where it is understood we drop it from the terminology. Any element of the pseudospectrum is called a *pseudoeigenvalue*. Unless otherwise stated, we shall always assume $\epsilon > 0$, but it is occasionally helpful to extend our notation to allow $\epsilon = 0$, so $\Lambda_0 = \Lambda$. Analogously, the *strict pseudospectrum* is the set

$$\Lambda'_\epsilon = \{z \in \mathbf{C} : z \in \Lambda(X) \text{ where } \|X - A\| < \epsilon\}.$$

The *pseudospectral abscissa* α_ϵ is the maximum value of the real part over the pseudospectrum:

$$\alpha_\epsilon = \sup\{\mathrm{Re}\, z : z \in \Lambda_\epsilon\}. \tag{2.1}$$

We call this optimization problem *the pseudospectral abscissa problem*. Note $\alpha_0 = \alpha$.

The function $\sigma_{\min} : \mathbf{M}^n \to \mathbf{R}$ denotes the smallest singular value. We define a function $g : \mathbf{C} \to \mathbf{R}$ by

$$g(z) = \sigma_{\min}(A - zI) = \|(A - zI)^{-1}\|^{-1},$$

where we interpret the right-hand side as zero when $z \in \Lambda(A)$. Thus g is the reciprocal of the norm of the resolvent. Using this notation, a useful characterization of the pseudospectrum is

$$\Lambda_\epsilon = \{z \in \mathbf{C} : g(z) \leq \epsilon\},$$

and analogously

$$\Lambda'_\epsilon = \{z \in \mathbf{C} : g(z) < \epsilon\}$$

(see [39]). Clearly as ϵ increases, both families of sets are monotonic increasing.

We will sometimes want to allow the matrix A (and the parameter ϵ) to vary. We therefore define the pseudospectral abscissa *function* $\alpha_\epsilon : \mathbf{M}^n \to \mathbf{R}$ by

$$\alpha_\epsilon(Z) = \sup\{\mathrm{Re}\, z : \sigma_{\min}(Z - zI) \le \epsilon\}.$$

3. Related ideas. The pseudospectral abscissa is related to several other functions important for stability analysis. In this section we briefly sketch the connections with two such functions, in particular, the "distance to instability" and the $\mathbf{H}_\infty$ norm.

A matrix A is *stable* if all its eigenvalues have strictly negative real parts; in other words, the spectral abscissa of A satisfies $\alpha(A) < 0$. From any given matrix A, the distance to the set of matrices which are not stable [28, 19] (also known as the *complex stability radius* [21]) is

$$\beta(A) = \min\{\|X - A\| : X \in \mathbf{M}^n,\ \alpha(X) \ge 0\}.$$

Since the set of matrices which are not stable is closed, this minimum is attained. Notice in particular that $\beta(A) = 0$ if and only if A is not stable. It is now easy to check the relationship

$$\beta(A) \le \epsilon \quad \Leftrightarrow \quad \alpha_\epsilon(A) \ge 0, \tag{3.1}$$

and more generally, for any real x,

$$\alpha_\epsilon(A) \ge x \quad \Leftrightarrow \quad \alpha_\epsilon(A - xI) \ge 0 \quad \Leftrightarrow \quad \beta(A - xI) \le \epsilon.$$

Notice that we can write the pseudospectral abscissa in the form

$$\alpha_\epsilon(A) = \max\{\alpha(X) : \|X - A\| \le \epsilon\},$$

a special case of "robust regularization" [26] and "minimum stability degree" [2]. Since the spectral abscissa α is continuous, standard arguments [26] show that the function

$$(\epsilon, A) \in \mathbf{R}_+ \times \mathbf{M}^n \mapsto \alpha_\epsilon(A) \tag{3.2}$$

is continuous.

In this paper we consider almost exclusively a fixed choice of the parameter ϵ, but for the moment let us consider the effect of varying ϵ on the solution of a pseudospectral abscissa minimization problem. For any fixed set of feasible matrices $F \subset \mathbf{M}^n$, the continuity of the map (3.2) guarantees various useful continuity properties of the optimal value and solutions of the optimization problem $\inf_F \alpha_\epsilon$ (see [34, Chap. 7]). In particular, if F is nonempty and compact, then

$$\lim_{\epsilon \to \bar{\epsilon}} \inf_F \alpha_\epsilon = \inf_F \alpha_{\bar{\epsilon}},$$

and any cluster point of a sequence of matrices A_r minimizing α_{ϵ_r} over F, where $\epsilon_r \to \bar{\epsilon}$, must minimize $\alpha_{\bar{\epsilon}}$ over F.

Notice that any stable matrix A satisfies

$$\alpha_{\beta(A)}(A) = 0.$$

To see this, note that the implication (3.1) shows $\alpha_{\beta(A)} \ge 0$, while if $\alpha_{\beta(A)} > 0$, then by the continuity of α_ϵ with respect to ϵ, there would exist $\epsilon \in (0, \beta(A))$ such that $\alpha_\epsilon(A) \ge 0$, whence we get the contradiction $\beta(A) \le \epsilon < \beta(A)$.

We return to our pseudospectral abscissa minimization problem $\inf_F \alpha_\epsilon$. The following easy result shows that, under reasonable conditions, for a particular choice of ϵ, this problem is equivalent to maximizing the distance to instability over the same set of feasible matrices.

PROPOSITION 3.1 (maximizing the distance to instability). *If the optimal value* $\bar{\beta} = \max_F \beta$ *is attained by some stable matrix, then* $\min_F \alpha_{\bar{\beta}} = 0$ *and*

$$\operatorname{argmin}\{\alpha_{\bar{\beta}}(X) : X \in F\} = \operatorname{argmax}\{\beta(X) : X \in F\}.$$

Proof. Any matrix $A \in F$ satisfies $\beta(A) \leq \bar{\beta}$. If A is stable, then $\alpha_{\bar{\beta}}(A) \geq \alpha_{\beta(A)}(A) = 0$, while on the other hand, if A is not stable, then $\alpha_{\bar{\beta}}(A) \geq \alpha_0(A) \geq 0$. Hence $\inf_F \alpha_{\bar{\beta}} \geq 0$.

By assumption, $\bar{\beta}$ is finite and strictly positive, so clearly every matrix in the (nonempty) set of optimal solutions $\operatorname{argmax}_F \beta$ is stable. Any such matrix A satisfies $\alpha_{\bar{\beta}}(A) = \alpha_{\beta(A)}(A) = 0$, and hence $A \in \operatorname{argmin}_F \alpha_{\bar{\beta}}$. We deduce $\operatorname{argmax}_F \beta \subset \operatorname{argmin}_F \alpha_{\bar{\beta}}$ and $\min_F \alpha_{\bar{\beta}} = 0$.

Consider, conversely, a matrix $A \in F$ such that $A \notin \operatorname{argmax}_F \beta$. Suppose first that A is stable. Since $\beta(A) < \bar{\beta}$, we know $\alpha_{\bar{\beta}}(A) > \alpha_{\beta(A)}(A) = 0$, because as we shall see in the next section, $\alpha_\epsilon(A)$ is strictly increasing in ϵ. On the other hand, if A is not stable, then the same reasoning shows $\alpha_{\bar{\beta}}(A) > \alpha_0(A) \geq 0$. In either case, we have shown $A \notin \operatorname{argmin}_F \alpha_{\bar{\beta}}$, so $\operatorname{argmin}_F \alpha_{\bar{\beta}} \subset \operatorname{argmax}_F \beta$ as required. ∎

We thus see that, under reasonable conditions, as ϵ increases from zero, the set of optimal solutions $\operatorname{argmax}_F \alpha_\epsilon$ evolves from the set of minimizers of the spectral abscissa through the set of maximizers of the stability radius. This raises the question of what happens for *large* ϵ. The following result shows that the limiting version of $\inf_F \alpha_\epsilon$ as $\epsilon \to +\infty$ is the optimization problem

$$\inf_{X \in F} \lambda_{\max}\Big(\frac{X + X^*}{2}\Big),$$

where $\lambda_{\max}$ denotes the largest eigenvalue of a Hermitian matrix.

THEOREM 3.2 (large ϵ). *For any matrix* $A \in \mathbf{M}^n$,

$$\big[\alpha_\epsilon(X) - \epsilon\big] \to \lambda_{\max}\Big(\frac{A + A^*}{2}\Big) \quad \textit{as} \quad \epsilon \to +\infty \textit{ and } X \to A.$$

Proof. If we denote the right-hand side by λ, then there is a unit vector $u \in \mathbf{C}^n$ satisfying $u^*(A + A^*)u = 2\lambda$. Consider any sequence $\epsilon_r \to +\infty$ and $X_r \to A$. Since $\|uu^*\| = 1$, we know

$$\alpha_{\epsilon_r}(X_r) - \epsilon_r \geq \alpha(X_r + \epsilon_r uu^*) - \epsilon_r = \epsilon_r\Big(\alpha\Big(uu^* + \frac{1}{\epsilon_r}X_r\Big) - 1\Big).$$

Now standard perturbation theory [23] shows α is analytic around the matrix uu^* with gradient $\nabla\alpha(uu^*) = uu^*$, so as $r \to \infty$, the right-hand side in the above relationship converges to

$$\operatorname{Re}\left(\operatorname{tr}\left(uu^*A\right)\right) = \operatorname{Re} u^*Au = \lambda.$$

We have thus shown

$$\liminf_r (\alpha_{\epsilon_r}(X_r) - \epsilon_r) \geq \lambda.$$

Now suppose

$$\limsup_r(\alpha_{\epsilon_r}(X_r) - \epsilon_r) > \lambda.$$

We will derive a contradiction. Without loss of generality, there exists a real $\delta > 0$ such that

$$\alpha_{\epsilon_r}(X_r) - \epsilon_r > \lambda + \delta \quad \text{for all } r.$$

For each r we can choose a matrix D_r satisfying $\|D_r\| \le 1$ and

$$\alpha_{\epsilon_r}(X_r) = \alpha(X_r + \epsilon_r D_r),$$

and a unit vector $w_r \in \mathbf{C}^n$ satisfying

$$\alpha(X_r + \epsilon_r D_r) = \mathrm{Re}\,(w_r^*(X_r + \epsilon_r D_r)w_r).$$

Hence

$$\begin{aligned}\lambda + \delta &< \mathrm{Re}\,(w_r^* X_r w_r) + \epsilon_r(\mathrm{Re}\,(w_r^* D_r w_r) - 1)\\ &\le \mathrm{Re}\,(w_r^* X_r w_r) = w_r^*\Big(\frac{X_r + X_r^*}{2}\Big)w_r\\ &\le \lambda_{\max}\Big(\frac{X_r + X_r^*}{2}\Big).\end{aligned}$$

But as $r \to \infty$, the right-hand side above converges to λ, which is the desired contradiction. □

We see from this result that, for example, if the set F is a polyhedron, then the limiting version of the optimization problem $\inf_F \alpha_\epsilon$ as $\epsilon \to \infty$ is a computationally straightforward, convex minimization problem, whereas when $\epsilon = 0$ the problem may be hard [5].

The idea of the $\mathbf{H}_\infty$ norm of a transfer matrix is also closely related to the complex stability radius. Consider the linear time-invariant dynamical system

$$\dot{p} = Ap + u,$$

where p denotes the state vector (in this simple case coinciding with the output) and u denotes the input vector. The "transfer matrix" of this system is the function $H(s) = (sI - A)^{-1}$ (where s is a complex variable). Assuming the matrix A is stable, the corresponding $\mathbf{H}_\infty$ norm is defined by

$$\|H\|_\infty = \sup_{\omega \in \mathbf{R}} \sigma_{\max}(H(i\omega)),$$

where $\sigma_{\max}$ denotes the largest singular value. Clearly

$$\|H\|_\infty = \sup_{\omega \in \mathbf{R}} \frac{1}{\sigma_{\min}(A - i\omega I)},$$

so $\|H\|_\infty < \epsilon^{-1}$ if and only if we have

$$\sigma_{\min}(A - i\omega I) > \epsilon \quad \text{for all } \omega \in \mathbf{R}.$$

As a consequence of Theorem 5.4 below, for example, this is equivalent to $\alpha_\epsilon(A) < 0$. In summary, for a stable matrix A, we have

$$\alpha_\epsilon(A) < 0 \quad \Leftrightarrow \quad \beta(A) > \epsilon \quad \Leftrightarrow \quad \|H\|_\infty < \frac{1}{\epsilon}. \tag{3.3}$$

We can characterize the condition $\alpha_\epsilon(A) < x$ analogously in terms of a "shifted" $\mathbf{H}_\infty$ norm [9, p. 67].

An important topic in robust control has been the design of controllers which minimize the $\mathbf{H}_\infty$ norm [44, 43]. In the language above, this corresponds to choosing the parameters defining the stable matrix A in order to maximize the minimum value of $\sigma_{\min}(A - zI)$ as z varies along the imaginary axis. Our ultimate aim of optimizing the pseudospectral abscissa is related, but rather different, being motivated by the broad idea of robust optimization [4]. We first fix the "level of robustness" ϵ (precisely the quantity that we seek to maximize in an $\mathbf{H}_\infty$ norm problem) and then vary A to move the corresponding pseudospectrum as far as possible to the left in the complex plane. In other words, we try to maximize a real parameter x such that the $\mathbf{H}_\infty$ norm corresponding to the shifted matrix $A - xI$ is not more than ϵ^{-1}.

What are the relative merits of different choices of ϵ in a pseudospectral minimization problem $\inf_F \alpha_\epsilon$? Here we are motivated by Trefethen's well-known viewpoint [39, 40], but we add an optimization "twist." When $\epsilon = 0$, optimization amounts to minimizing the spectral abscissa of a matrix $A \in F$, in other words, optimizing the *asymptotic* rate of decay of trajectories of the dynamical system $\dot{p} = Ap$. On the other hand, for large ϵ, by Theorem 3.2 (large ϵ), optimization amounts to minimizing $\lambda_{\max}(A + A^*)/2$. This corresponds to optimizing the *initial* decay rate of the dynamical system, since at time $t = 0$,

$$\frac{d}{dt}\frac{\|p\|^2}{2} = p(0)^*\Big(\frac{A + A^*}{2}\Big)p(0) \le \lambda_{\max}\Big(\frac{A + A^*}{2}\Big)\|p(0)\|^2,$$

with equality if $p(0)$ is an eigenvector corresponding to the largest eigenvalue. For intermediate choices of ϵ, minimizing the pseudospectral abscissa balances the two objectives of improving asymptotic stability and restricting the size of *transient* peaks in the trajectories. In particular, Proposition 3.1 (maximizing the distance to instability) shows that, under reasonable conditions, for some choice of ϵ, minimizing α_ϵ is equivalent to minimizing the $\mathbf{H}_\infty$ norm, that is, maximizing the complex stability radius.

To summarize, minimizing the $\mathbf{H}_\infty$ norm of a matrix A optimizes the robustness of the stability of the dynamical system $\dot{p} = Ap$, but with no explicit reference to its asymptotic decay rate. By minimizing the pseudospectral abscissa α_ϵ instead, for different choices of the parameter ϵ we obtain a range of different balances between robustness and asymptotic decay, one choice giving exactly the $\mathbf{H}_\infty$ norm problem. One could achieve a similar range of balances by minimizing the $\mathbf{H}_\infty$ norm corresponding to the shifted matrix $A - xI$ as the real parameter x varies; however, working with ϵ-pseudospectra for fixed ϵ provides a natural interpretation in terms of allowable perturbations to A. Yet another range of balances is achieved by the "robust spectral abscissa" defined in [14].

Just as with the $\mathbf{H}_\infty$ norm, the pseudospectral abscissa can be characterized via semidefinite programming. Specifically, by [9, p. 67] or [4, Prop. 4.4.2], a real x satisfies

$$\alpha_\epsilon(A) < x$$

if and only if there exist reals $\mu < 0$ and λ, and an $n \times n$ positive definite Hermitian matrix P such that the matrix

$$\begin{bmatrix} (\mu - \lambda)I + 2xP - A^*P - PA & -\epsilon P \\ -\epsilon P & \lambda I \end{bmatrix}$$

is positive semidefinite. As discussed in [9, pp. 3–4], the power of such semidefinite characterizations derives from their amenability to efficient interior point methods for convex optimization, pioneered in [31]. The disadvantage is the appearance of subsidiary semidefinite matrix variables: if the underlying matrices A are large, and we need to calculate the pseudospectral abscissa for many different matrices (in an optimization routine, for example), involving these subsidiary variables may be prohibitive computationally; see, for example, [33, 14, 42]. For this reason, in this work we consider more direct approaches to the pseudospectral abscissa.

4. Boundary properties. We begin our direct, geometric approach to the pseudospectral abscissa by studying the boundary of the pseudospectrum.

PROPOSITION 4.1 (compactness). *The pseudospectrum Λ_ϵ is a compact set contained in the ball of radius $\|A\| + \epsilon$. It contains the strict pseudospectrum Λ'_ϵ, which is nonempty and open.*

Proof. The strict pseudospectrum is nonempty since it contains the spectrum. It is open since $\sigma_{\min}$, and hence g are continuous. This also shows that the pseudospectrum is closed. For any point $z \in \Lambda_\epsilon$ there is a unit vector $u \in \mathbf{C}^n$ satisfying $\|(A - zI)u\| \leq \epsilon$. On the other hand, $\|Au\| \leq \|A\|$, so we have the inequality

$$|z| = \|zu\| \leq \|(A - zI)u\| + \|Au\| \leq \|A\| + \epsilon, \tag{4.1}$$

which shows boundedness. ∎

The next result is slightly less immediate.

THEOREM 4.2 (local minima). *The only local minimizers of the function*

$$g(z) = \sigma_{\min}(A - zI)$$

are the eigenvalues of the matrix A.

Proof. Suppose the point z_0 is a local minimizer that is not an eigenvalue. Then z_0 is a local maximizer of the norm of the resolvent $\|(A - zI)^{-1}\|$. We can choose unit vectors $u, v \in \mathbf{C}^n$ satisfying

$$\|(A - z_0 I)^{-1}\| = |u^*(A - z_0 I)^{-1} v|,$$

and then we have, for all points z close to z_0, the inequalities

$$|u^*(A - zI)^{-1}v| \leq \|(A - zI)^{-1}\| \leq \|(A - z_0 I)^{-1}\| = |u^*(A - z_0 I)^{-1}v|.$$

Hence the modulus of the function $u^*(A - zI)^{-1}v$ has a local maximum at z_0. But this contradicts the maximum modulus principle, since this function is analytic and nonconstant near z_0. ∎

COROLLARY 4.3 (closure of strict pseudospectrum). *The closure of the strict pseudospectrum is the pseudospectrum, so for $\epsilon > 0$ the pseudospectral abscissa is*

$$\alpha_\epsilon = \sup\{\operatorname{Re} z : z \in \Lambda'_\epsilon\}.$$

Proof. A point in the pseudospectrum that is outside the closure of the strict pseudospectrum must be a local minimizer of the function g. ∎

An easy exercise now shows that the pseudospectral abscissa α_ϵ is a continuous, strictly increasing function of $\epsilon \in [0, +\infty)$. Note also that, by contrast with the above result, the function g may have local *maximizers* and, consequently, the strict pseudospectrum may not equal the interior of the pseudospectrum.

We can refine the above corollary with a more delicate argument, showing that we can "access" any point in the pseudospectrum via a smooth path through the strict pseudospectrum.

THEOREM 4.4 (accessibility). *Given any point z_0 in the pseudospectrum, there is a real-analytic path $p : [0, 1] \to \mathbf{C}$ such that $p(0) = z_0$ and $p(t)$ lies in the strict pseudospectrum for all $t \in (0, 1]$.*

Proof. We may as well assume $g(z_0) = \epsilon$. By Corollary 4.3, there exists a sequence $z_r \in \Lambda'_\epsilon$ approaching z_0. For each index r there exists a vector $u^r \in \mathbf{C}^n$ satisfying the inequalities

$$1 < \|u^r\| < 1 + \frac{1}{r} \quad \text{and} \quad \|(A - z_r I)u^r\| < \epsilon.$$

By taking a subsequence, we may as well assume that the sequence $\{u^r\}$ converges to a limit u^0, and then we have $(z_0, u^0) \in \operatorname{cl} S$, where

$$S = \Big\{(z, u) : \|u\|^2 > 1,\ \|(A - zI)u\|^2 < \epsilon^2\Big\}.$$

Since the set S is defined by a finite number of strict algebraic inequalities, we can apply the accessibility lemma [29]. Hence there is a real-analytic path $q : [0, 1] \to \mathbf{C} \times \mathbf{C}^n$ such that $q(0) = (z_0, u^0)$ and $q(t) \in S$ for all $t \in (0, 1]$. The result now follows by taking p to be the first component of q. ∎

In most cases the boundary of the pseudospectrum is straightforward to analyze without recourse to the above result. We make the following definition.

DEFINITION 4.5. *A point $z \in \mathbf{C}$ is* degenerate *if the smallest singular value of $A - zI$ is nonzero and simple (that is, has multiplicity one) and the corresponding right singular vector u satisfies $u^*(A - zI)u = 0$.*

We need the following elementary identity.

LEMMA 4.6. *Given any unit vector $u \in \mathbf{C}^n$, matrix $B \in \mathbf{M}^n$, and scalar $w \in \mathbf{C}$, we have*

$$\|(B + wI)u\|^2 - \|Bu\|^2 = |u^*(B + wI)u|^2 - |u^*Bu|^2.$$

The next result shows that, except possibly at degenerate points, the pseudospectrum can never be "pointed" outwards.

PROPOSITION 4.7 (pointedness). *Any nondegenerate point in the pseudospectrum lies on the boundary of an open disk contained in the strict pseudospectrum.*

Proof. Consider a nondegenerate point $z_0 \in \Lambda_\epsilon$. We may as well assume $g(z_0) = \epsilon$. Choose a unit right singular vector $u \in \mathbf{C}^n$ satisfying the condition $u^*(A - z_0 I)u \neq 0$. We now claim

$$|z - u^*Au| < |z_0 - u^*Au| \quad \Rightarrow \quad z \in \Lambda'_\epsilon.$$

To see this, observe that if z satisfies the left-hand side, then

$$\begin{aligned}
\sigma^2_{\min}(A - zI) - \epsilon^2 &\leq \|(A - zI)u\|^2 - \|(A - z_0 I)u\|^2 \\
&= |u^*(A - zI)u|^2 - |u^*(A - z_0 I)u|^2 \\
&= |z - u^*Au|^2 - |z_0 - u^*Au|^2 \\
&< 0,
\end{aligned}$$

using the preceding lemma. ∎

In particular, this result shows that Theorem 4.4 is elementary in the case when the point of interest z_0 is nondegenerate.

Thus the pseudospectrum is not pointed outward, except possibly at a degenerate point. In fact, a more detailed analysis due to Trefethen shows that the pseudospectrum is *never* pointed outward [41]. However, it can certainly be pointed inward, as the following example shows.

Example 1 (nonsmooth points). Consider the matrix

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

The pseudospectrum Λ_ϵ consists of the union of two disks of radius ϵ, centered at the two eigenvalues, $\pm i$. For example, if $\epsilon = \sqrt{2}$, then $z = 1$ is a nonsmooth point where the boundary of the pseudospectrum is pointed inward. In the case where $\epsilon = 1$, the pseudospectrum consists of two disks tangent to each other at the origin.

Nonetheless, even though the boundary of the pseudospectrum can be nonsmooth, this cannot occur at any nondegenerate optimal solution of the pseudospectral abscissa problem.

PROPOSITION 4.8 (optimal solutions). *Any locally optimal solution z_0 of the pseudospectral abscissa problem* (2.1) *must lie on the boundary of the pseudospectrum. Furthermore, unless z_0 is degenerate, the boundary is differentiable there.*

Proof. The fact that z_0 cannot lie in the interior of Λ_ϵ is immediate. Now assume z_0 is nondegenerate. Since z_0 is optimal, Λ_ϵ lies on or to the left of the vertical line through z_0. But since z_0 is nondegenerate, Λ_ϵ contains a closed disk whose boundary contains z_0, by Proposition 4.7 (pointedness). Thus the boundary of Λ_ϵ lies between the disk and the vertical line, which are tangent at z_0. This completes the proof. ∎

Again, the nondegeneracy hypothesis may be dropped using the more general result on pointedness mentioned above [41].

5. Components of the pseudospectrum. We recall some basic ideas from plane topology. A *domain* is a nonempty, open, arcwise connected subset of $\mathbf{C}$. Given a point z in an open set $\Omega \subset \mathbf{C}$, a particular example of a domain is the *component* of z, which consists of all points that can be joined to z by a continuous path in Ω [35].

The following result is in essence well known (see, for example, [10]).

THEOREM 5.1 (eigenvalues and components). *Every component of the strict pseudospectrum of the matrix A contains an eigenvalue of A.*

Proof. Suppose the set S is a component of the strict pseudospectrum Λ'_ϵ that contains no eigenvalues of A. The function g attains its minimum on the compact set $\operatorname{cl} S$ at some point z, and clearly $g(z) < \epsilon$, so $z \in \Lambda'_\epsilon$. Since S is open and contains no eigenvalues, Theorem 4.2 (local minima) implies $z \notin S$.

But since Λ'_ϵ is open, it contains an open disk D centered at z. Since $z \in \operatorname{cl} S$, we know $D \cap S \neq \emptyset$, and hence $D \cup S$ is an arcwise connected subset of Λ'_ϵ strictly larger than S. But this contradicts the definition of S. ∎

In Example 1 (nonsmooth points), when $\epsilon = 1$ the strict pseudospectrum consists of two components, namely the two open disks centered at the two eigenvalues, $\pm i$. By contrast, the pseudospectrum is arcwise connected.

The simplest case of the above result occurs when each eigenvalue has geometric multiplicity one and ϵ is small. In this case we show later (Corollary 7.5) that the pseudospectrum consists of disjoint compact convex neighborhoods of each eigenvalue (cf. [32]).

Our next aim is to try to bracket the pseudospectral abscissa. We first need a subsidiary result.

LEMMA 5.2 (moving to the boundary). *For any point z_1 in Λ_ϵ there exists a point z_2 satisfying $\operatorname{Re} z_1 = \operatorname{Re} z_2$ and $g(z_2) = \epsilon$.*

Proof. We simply take z_2 on the boundary of the intersection of the vertical line through z_1 and the pseudospectrum Λ_ϵ (which is compact). □

Byers' algorithm for calculating the distance to instability [16] and its subsequent variants (see the introduction) all depend on versions of the following easy piece of linear algebra, relating singular values to imaginary eigenvalues of a certain Hamiltonian matrix. We include a proof for completeness.

LEMMA 5.3 (imaginary eigenvalues). *For real numbers x and y, and $\epsilon \geq 0$, the matrix $A - (x + iy)I$ has a singular value ϵ if and only if the matrix*

$$\begin{bmatrix} xI - A^* & \epsilon I \\ -\epsilon I & A - xI \end{bmatrix}$$

has an eigenvalue iy.

Proof. Plus and minus the singular values of any matrix $B \in \mathbf{M}^n$ are exactly the eigenvalues of the matrix

$$\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}.$$

Thus the matrix $A - (x + iy)I$ has a singular value ϵ if and only if ϵ is an eigenvalue of the matrix

$$\begin{bmatrix} 0 & A - (x + iy)I \\ A^* - (x - iy)I & 0 \end{bmatrix}$$

or, in other words, if and only if the matrix

$$\begin{bmatrix} -\epsilon I & A - (x + iy)I \\ A^* - (x - iy)I & -\epsilon I \end{bmatrix}$$

is singular. Since

$$\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} -\epsilon I & A - (x + iy)I \\ A^* - (x - iy)I & -\epsilon I \end{bmatrix} = \begin{bmatrix} (A^* - xI) & -\epsilon I \\ \epsilon I & (xI - A) \end{bmatrix} + iy \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

this is equivalent to iy being an eigenvalue of the matrix

$$\begin{bmatrix} xI - A^* & \epsilon I \\ -\epsilon I & A - xI \end{bmatrix}. \qquad \square$$

The following result is our key test. Geometrically it states simply that a given real x (bigger than the spectral abscissa α) is less than the pseudospectral abscissa exactly when the vertical line through x intersects the boundary of the pseudospectrum. As we shall see, this is a straightforward computational test.

THEOREM 5.4 (bracketing the pseudospectral abscissa). *For any real $x \geq \alpha$, the following statements are equivalent:*

(i) $x \leq \alpha_\epsilon$;

(ii) *the equation*

$$g(x+iy) = \epsilon, \quad y \in \mathbf{R}, \tag{5.1}$$

is solvable;

(iii) *the system*

$$iy \in \Lambda \begin{bmatrix} xI - A^* & \epsilon I \\ -\epsilon I & A - xI \end{bmatrix}, \quad y \in \mathbf{R}, \tag{5.2}$$

is solvable.

Proof. We first show (i) ⇒ (ii). If $x = \alpha_\epsilon$, then choose any point z solving the pseudospectral abscissa problem (2.1). Clearly $z = x + iy$ for some real y, and $g(z) = \epsilon$, so we have shown that (5.1) has a solution.

We can therefore assume $x < \alpha_\epsilon$, in which case there exists a point z_1 such that $\operatorname{Re} z_1 > x$ and $g(z_1) < \epsilon$. The component of z_1 in the strict pseudospectral abscissa Λ'_ϵ contains an eigenvalue z_2 by Theorem 5.1 (eigenvalues and components). Hence there is an arc in this component connecting z_1 and z_2. But since $\operatorname{Re} z_1 > x \geq \operatorname{Re} z_2$, this arc must contain a point z_3 with $\operatorname{Re} z_3 = x$. Now applying Lemma 5.2 (moving to the boundary) gives a solution to (5.1).

The implication (ii) ⇒ (iii) is immediate from Lemma 5.3 (imaginary eigenvalues), so it remains to show (iii) ⇒ (i). But this is again an easy consequence of Lemma 5.3: if system (5.2) holds, then ϵ is a singular value of the matrix $A - (x + iy)I$, and hence the smallest singular value of this matrix is no greater than ϵ, whence we get the result. □

Using this result, the relationship (3.3) between the pseudospectral abscissa and the $\mathbf{H}_\infty$ norm is an easy exercise.

We can now approximate the pseudospectral abscissa α_ϵ by a bisection search as follows.

Algorithm 5.5 (bisection method). We begin with the initial interval

$$[\alpha, \|A\| + \epsilon].$$

We know α_ϵ lies in this interval by the argument of Proposition 4.1 (compactness). Now at each iteration we let x be the midpoint of the current interval and compute all the eigenvalues of the matrix

$$\begin{bmatrix} xI - A^* & \epsilon I \\ -\epsilon I & A - xI \end{bmatrix}. \tag{5.3}$$

If any of the eigenvalues are purely imaginary, then we deduce $x \leq \alpha_\epsilon$ and replace the current interval with its right half. Otherwise, by Theorem 5.4 (bracketing the pseudospectral abscissa), we know $x > \alpha_\epsilon$, so we replace the current interval with its left half. The intervals generated by this algorithm are guaranteed to converge to the pseudospectral abscissa α_ϵ.

The difference between this algorithm and Byers' bisection method for the distance to instability [16] is that the former searches for x by bisection, while the latter searches for ϵ by bisection.

Notice that at each iteration of the bisection method we can easily solve (5.1). We first list the purely imaginary eigenvalues of the matrix (5.5), namely $\{iy_1, iy_2, \ldots, iy_k\}$. We then form the index set

$$J = \{j : \sigma_{\min}(A - (x + iy_j)I) = \epsilon\}.$$

The set of solutions of (5.1) is then simply $\{y_j : j \in J\}$. As we shall see in the next section, the points $x+iy_j$ (for $j \in J$) provide good approximations to *all* the solutions of the pseudospectral abscissa problem (2.1).

A more sophisticated, quadratically convergent algorithm for the pseudospectral abscissa, based on similar ideas and analogous to $\mathbf{H}_\infty$ norm algorithms such as [7, 11, 24], is developed in [13].

6. Approximate solutions. The results generated by the bisection algorithm (or the algorithm in [13]) approximate all the global maximizers in the pseudospectral abscissa problem (2.1). To make this precise we use the following standard notion of set convergence [34]. We say that a sequence of sets $Y^1, Y^2, \ldots \subset \mathbf{R}$ *converges* to a set $Y \subset \mathbf{R}$ if the following properties hold:

(i) For any number $y \in Y$ there exists a sequence of numbers $y^r \in Y^r$ converging to y;

(ii) any cluster point of a sequence of numbers $y^r \in Y^r$ lies in Y.

(This notion is weaker than the idea of convergence with respect to the Pompeiu–Hausdorff distance [34, Ex 4.13], although it is equivalent in the case when the sets Y^r and Y are uniformly bounded, as will be the case in our application below.)

We now prove a rather general result.

THEOREM 6.1 (global maximizers). *The number of global maximizers of the pseudospectral abscissa problem* (2.1) *does not exceed* n. *Denote these*

$$\{\alpha_\epsilon + iy : y \in Y\},$$

where $Y \subset \mathbf{R}$. *Consider any real sequence* $\alpha \leq x^r \uparrow \alpha_\epsilon$. *Then the sets*

$$Y_r = \{y \in \mathbf{R} : g(x^r + iy) = \epsilon\}$$

converge to Y.

Proof. The pseudospectral abscissa problem (2.1) has at least one maximizer, by compactness. Furthermore, any solution $z = \alpha_\epsilon + iy$ must satisfy the equation $g(z) = \epsilon$. Just as in the proof of Theorem 5.4 (bracketing the pseudospectral abscissa), this implies that y must satisfy the equation

$$\det \begin{bmatrix} -\epsilon I & A - (\alpha_\epsilon + iy)I \\ A^* - (\alpha_\epsilon - iy)I & -\epsilon I \end{bmatrix} = 0.$$

But this polynomial equation has at most $2n$ solutions, so we can write

$$Y = \{y_1, y_2, \ldots, y_m\},$$

where $1 \leq m \leq 2n$.

Fix the index $j \in \{1, 2, \ldots, m\}$. Theorem 4.4 (accessibility) and Theorem 5.1 (eigenvalues and components) together imply the existence of a continuous function $p : [0, 1] \to \mathbf{C}$ such that $p(0) = \alpha_\epsilon + iy_j$, $p(1)$ is an eigenvalue, and $p(t) \in \Lambda'_\epsilon$ for all $t > 0$.

Using the continuity of p, we can now iteratively construct a nonincreasing sequence $\{t_r\} \subset [0, 1]$ such that $\operatorname{Re} p(t_r) = x^r$ for all r. Taking limits shows

$$\operatorname{Re} p(\lim_r t_r) = \lim_r x^r = \alpha_\epsilon.$$

But for $t > 0$ we have $g(p(t)) < \epsilon$, which implies $\operatorname{Re} p(t) < \alpha_\epsilon$, so we deduce $t_r \downarrow 0$. Hence if we define $v_j^r = \operatorname{Im} p(t_r)$, we have $v_j^r \to y_j$.

For each index r, consider the bounded open set

$$\{y \in \mathbf{R} : g(x^r + iy) < \epsilon\}.$$

If $t_r = 0$, this set is empty, and we define $l_j^r = u_j^r = y_j$. Otherwise, denote the component of v_j^r in this set by the open interval (l_j^r, u_j^r). By continuity, l_j^r and u_j^r are both zeros of the function

$$y \in \mathbf{R} \mapsto g(x^r + iy) - \epsilon. \tag{6.1}$$

We now claim both $l_j^r \to y_j$ and $u_j^r \to y_j$.

If this claim fails, then without loss of generality, after taking a subsequence, we can assume $l_j^r \to w < y_j$. By definition, we know

$$g(x^r + i(sv_j^r + (1-s)l_j^r)) < \epsilon \text{ for all } s \in (0,1], \qquad r = 1, 2, \ldots,$$

so taking limits shows

$$g(\alpha_\epsilon + i(sy_j + (1-s)w)) \leq \epsilon \text{ for all } s \in [0,1].$$

But in this case every point in the line segment $\alpha_\epsilon + i[w, y_j]$ solves the pseudospectral abscissa problem (2.1), contradicting the fact that there are only finitely many solutions. This proves the claim. We have thus shown property (i) in the definition of set convergence: the constructed sequence (l_j^r) converges to the desired point y_j. Property (ii) is immediate.

Finally, suppose $m > n$. Choose any nondecreasing sequence $\{x^r\} \subset [\alpha, \alpha_\epsilon)$ converging to α_ϵ, and for each index r construct the set

$$\{l_j^r, u_j^r : j = 1, 2, \ldots, m\}$$

as above. Then for r sufficiently large, this is a set of $2m$ distinct zeros of the function (6.1), and hence of the polynomial

$$\det \begin{bmatrix} -\epsilon I & A - (x^r + iy)I \\ A^* - (x^r - iy)I & -\epsilon I \end{bmatrix}.$$

But this polynomial is not identically zero, and has degree $2n$, which is a contradiction. ∎

The algorithmic significance of the above result is this: Consider any algorithm that generates a sequence of lower approximations to the pseudospectral abscissa, $x^r \uparrow \alpha_\epsilon$. In particular, we could consider the bisection algorithm of the previous section. For each step r, an eigenvalue computation generates the set $Y^r \subset \mathbf{R}$, as described after Algorithm 5.5. The above result now shows that this set is a good approximation to Y, and hence gives us a good approximation to the set of all optimal solutions to the pseudospectral abscissa problem.

7. Smoothness. To study the smoothness of the function g, and hence the boundary of the pseudospectrum, we rely on the following well-known result. We consider $\mathbf{M}^n$ as a Euclidean space with inner product

$$\langle X, Y \rangle = \operatorname{Re} \operatorname{tr} (X^* Y) \quad (X, Y \in \mathbf{M}^n).$$

A real-valued function on a real vector space is *real-analytic* at zero if in some neighborhood of zero is can be written as the sum of an absolutely convergent power series

in the coordinates relative to some basis, and we make an analogous definition at other points. In particular, such functions are smooth (C^∞) near the point in question.

We call vectors $u, v \in \mathbf{C}^n$ *minimal left and right singular vectors* for a matrix $Z \in \mathbf{M}^n$ if

$$Zv = \sigma_{\min}(Z)u \quad \text{and} \quad Z^*u = \sigma_{\min}(Z)v.$$

THEOREM 7.1 (analytic singular value). *If the matrix Z has a simple smallest singular value, then the function $\sigma_{\min}^2$ is real-analytic at Z. If, furthermore, $\sigma_{\min}(Z) > 0$, then $\sigma_{\min}$ is real-analytic at Z, with gradient*

$$\nabla\sigma_{\min}(Z) = uv^*$$

for any unit minimal left and right singular vectors $u, v \in \mathbf{C}^n$.

Proof. The matrix

$$(X^T - iY^T)(X + iY)$$

depends analytically on the matrices $X, Y \in \mathbf{M}^n$ and has a simple eigenvalue $\sigma_{\min}^2(Z)$ when $(X, Y) = (X_0, Y_0)$ for real matrices $X_0, Y_0 \in \mathbf{M}^n$ satisfying $Z = X_0 + iY_0$. Hence by standard perturbation theory [23], the above matrix has a unique eigenvalue near $\sigma_{\min}^2(Z)$ for all (X, Y) close to (X_0, Y_0), depending analytically on (X, Y). When X and Y are real, this eigenvalue is exactly $\sigma_{\min}^2(X + iY)$, so the first part follows. The second part follows by taking square roots. The gradient calculation is standard (see, for example, [37]). ∎

We next turn to smoothness properties of the function $g : \mathbf{C} \to \mathbf{R}$ defined by

$$g(z) = \sigma_{\min}(A - zI).$$

We will often find it more convenient to work with the squared function $g^2(z) = (g(z))^2$.

We can treat $\mathbf{C}$ as a Euclidean space, where we define the inner product by $\langle w, z\rangle = \mathrm{Re}\,(w^*z)$.

COROLLARY 7.2 (analytic boundary). *If the singular value $\sigma_{\min}(A - z_0I)$ is simple, then the function g^2 is real-analytic at z_0. If, furthermore, this singular value is strictly positive, then g is real-analytic at z_0, with gradient*

$$\nabla g(z_0) = -v^*u,$$

where the vectors $u, v \in \mathbf{C}^n$ are unit minimal left and right singular vectors for $A - z_0I$.

Proof. This follows from the previous result by the chain rule. ∎

Thus what we called "degenerate" points are simply smooth critical points of g, distinct from the eigenvalues. At a nondegenerate smooth point z_0 with $g(z_0) = \epsilon$, the gradient of g is nonzero, and hence the boundary of the pseudospectrum

$$\Lambda_\epsilon = \{z \in \mathbf{C} : g(z) \le \epsilon\}$$

is simply a smooth curve locally, with normal $u^*(A - z_0)u$ at z_0.

We call an eigenvalue of A *nonderogatory* if it has geometric multiplicity one. This is the most common type of multiple eigenvalue (from the perspective of the dimensions of the corresponding manifolds in $\mathbf{M}^n$ [1]). The following result is very well known.

PROPOSITION 7.3 (nonderogatory eigenvalues). *The point $\lambda_0 \in \mathbf{C}$ is a nonderogatory eigenvalue of the matrix A if and only if 0 is a simple singular value of $A - \lambda_0 I$.*

Proof. First, note that λ_0 is an eigenvalue of A if and only if $A - \lambda_0 I$ is singular, which is equivalent to 0 being a singular value of $A - \lambda_0 I$. Second, v is a corresponding eigenvector of A exactly when $(A - \lambda_0 I)v = 0$, which says that v is a right singular vector of $A - \lambda_0 I$ corresponding to the singular value 0. Thus the eigenspace of A corresponding to the eigenvalue λ_0 coincides with the subspace of right singular vectors of $A - \lambda_0 I$ corresponding to the singular value 0, so in particular these spaces have the same dimension. The result now follows. □

We can now show that the function g is well behaved near any nonderogatory eigenvalue of A.

THEOREM 7.4 (growth near an eigenvalue). *Let λ_0 be a nonderogatory eigenvalue of multiplicity m for the matrix A. Then*

$$\sigma_{\min}(A - zI) = g(z) = \frac{\prod_{j=1}^{n-m} |\lambda_j - \lambda_0|}{\prod_{k=1}^{n-1} \sigma_k} |z - \lambda_0|^m + O(|z - \lambda_0|^{m+1})$$

for complex z near λ_0, where $\lambda_1, \lambda_2, \ldots, \lambda_{n-m}$ are the eigenvalues of A distinct from λ_0 (listed by multiplicity) and $\sigma_1, \sigma_2, \ldots, \sigma_{n-1}$ are the nonzero singular values of $A - \lambda_0 I$ (listed by multiplicity). (In the case $n = 1$ or $m = n$, we interpret the empty products appearing in the above expression as 1.)

Furthermore, the function g^2 has positive definite Hessian at all points $z \neq \lambda_0$ near λ_0.

Proof. We prove the case $\lambda_0 = 0$: the general case follows by a simple transformation.

Since 0 is a nonderogatory eigenvalue of A, Proposition 7.3 (nonderogatory eigenvalues) shows 0 is a simple singular value of A. Hence by Corollary 7.2 (analytic boundary), the function $f : \mathbf{R}^2 \to \mathbf{R}$ defined by

$$f(x, y) = (g(x + iy))^2$$

is real-analytic at $(0, 0)$.

Consider any point $(x, y) \in \mathbf{R}^2$, and let $z = x + iy$. The matrix

$$(A - zI)^*(A - zI)$$

is Hermitian, so its characteristic polynomial

$$p_z(\mu) = \det((A - zI)^*(A - zI) - \mu I)$$

has all real coefficients. Hence we can write

$$p_z(\mu) = \sum_{r=0}^{n} q_r(x, y)\mu^r$$

for some real polynomials q_r. The smallest zero of p_z is $f(x, y)$.

We concentrate on the two lowest-order coefficients of the above polynomial. First, note

$$q_0(x, y) = p_z(0)$$

$$\begin{aligned} &= \det((A - zI)^*(A - zI)) \\ &= |\det(A - zI)|^2 \\ &= |z|^{2m} \prod_{j=1}^{n-m} |\lambda_j - z|^2. \end{aligned}$$

Hence for small (x, y) we have

$$q_0(x, y) = \left((x^2 + y^2)^m \prod_{j=1}^{n-m} |\lambda_j|^2 \right) + O(\|(x, y)\|^{2m+1}). \tag{7.1}$$

Turning to the coefficient of μ, notice

$$p_0(\mu) = \det(A^*A - \mu I) = -\mu \prod_{k=1}^{n-1} (\sigma_k^2 - \mu),$$

so

$$q_1(0, 0) = -\prod_{k=1}^{n-1} \sigma_k^2$$

(notice this is nonzero), and hence

$$q_1(x, y) = -\prod_{k=1}^{n-1} \sigma_k^2 + O(\|x, y\|). \tag{7.2}$$

Since the function f is real-analytic at $(0, 0)$, we know for some integer $t = 1, 2, \ldots,$

$$f(x, y) = s(x, y) + O(\|(x, y)\|^{t+1})$$

for some nonzero homogeneous polynomial s of degree t. Now substituting into the relationship

$$\sum_{r=0}^{n} q_r(x, y)(f(x, y))^r = 0$$

and using (7.1) and (7.2) shows $t = 2m$, and

$$f(x, y) = \frac{\prod_{j=1}^{n-m} |\lambda_j|^2}{\prod_{k=1}^{n-1} \sigma_k^2} (x^2 + y^2)^m + O(\|(x, y)\|^{2m+1})$$

as required.

It remains to show that the Hessian $\nabla^2 f(x, y)$ is positive definite for all small $(x, y) \neq (0, 0)$. We have shown that f is analytic at $(0, 0)$ and

$$f(x, y) = \tau(x^2 + y^2)^m + O(\|(x, y)\|^{2m+1})$$

for some nonzero constant τ. Since we can differentiate the power series for f term-by-term, a short calculation shows

$$\nabla^2 f(x, y) = 2m\tau(x^2 + y^2)^{m-2} H(x, y),$$

where

$$H(x,y) = \begin{bmatrix} (2m-1)x^2 + y^2 & 2(m-1)xy \\ 2(m-1)xy & x^2 + (2m-1)y^2 \end{bmatrix} + O(\|(x,y)\|^3).$$

We deduce

$$H_{11}(x,y) = (2m-1)x^2 + y^2 + O(\|(x,y)\|^3) > 0$$

and furthermore

$$\det H(x,y) = (2m-1)(x^2+y^2)^2 + O(\|(x,y)\|^5) > 0,$$

so the matrix $H(x,y)$ is positive definite for all small $(x,y) \neq (0,0)$. The result now follows. ∎

The following results are immediate consequences.

COROLLARY 7.5 (convexity). *If λ_0 is a nonderogatory eigenvalue of the matrix A, then for all small $\epsilon > 0$ the pseudospectrum Λ_ϵ near λ_0 consists of a compact, strictly convex neighborhood of λ_0.*

Proof. We can consider the pseudospectrum as a level set of the function g^2, which is strictly convex near λ_0. ∎

COROLLARY 7.6 (smoothness). *If λ_0 is a nonderogatory eigenvalue of the matrix A, then the function g is smooth with nonzero gradient at all nearby points distinct from λ_0.*

Proof. Since the real-analytic function g^2 is strictly convex near the eigenvalue λ_0, with a strict local minimizer there, it follows that λ_0 is an isolated critical point of g^2. It is then easy to see that g is smooth and noncritical near λ_0. ∎

If the matrix A has an eigenvalue λ_0 of multiplicity m, then, by continuity of the set of eigenvalues, any matrix close to A will have exactly m eigenvalues close to λ_0 (counted by multiplicity). Our last corollary bounds how far these eigenvalues can be from λ_0.

COROLLARY 7.7 (Hölder continuity). *With the assumptions and notation of Theorem* 7.4, *consider any constant*

$$\kappa > \left(\frac{\prod_{k=1}^{n-1} \sigma_k}{\prod_{j=1}^{n-m} |\lambda_j - \lambda_0|} \right)^{1/m}.$$

For any matrix Z close to A, any eigenvalue z of Z close to λ_0 satisfies

$$|z - \lambda_0| \leq \kappa \|Z - A\|^{1/m}.$$

Proof. This follows easily from Theorem 7.4 (growth near an eigenvalue), using the elementary property that

$$\sigma_{\min}(A - zI) \leq \|Z - A\|$$

for any eigenvalue z of Z. ∎

In the above result, if we specialize to the case of a perturbation $Z = A + tB$ (where t is a complex parameter), then the result shows that the eigenvalues of $A + tB$ near a nonderogatory eigenvalue of A of multiplicity m satisfy an m^{-1}-Hölder continuity condition in t. This is a well-known result; see [27, 23, 3, 30].

8. Smoothness and regularity of the pseudospectral abscissa. Our ultimate goal is an understanding of how the pseudospectral abscissa α_ϵ depends on the underlying matrix A. We therefore now allow A (and ϵ) to vary. Recall that the pseudospectral abscissa function $\alpha_\epsilon : \mathbf{M}^n \to \mathbf{R}$ is given by

$$\alpha_\epsilon(Z) = \max\{\operatorname{Re} z : \sigma_{\min}(Z - zI) \le \epsilon\},$$

and for any nonempty set $\Omega \subset \mathbf{C}$ we define the refinement

$$\alpha^\Omega(Z, \epsilon) = \sup\{\operatorname{Re} z : z \in \Omega,\ \sigma_{\min}(Z - zI) \le \epsilon\}. \tag{8.1}$$

Thus for $\Omega = \mathbf{C}$, we obtain exactly the pseudospectral abscissa. We now apply classical sensitivity analysis to differentiate this function.

THEOREM 8.1 (smoothness of pseudospectral abscissa). *Suppose that, for $\epsilon = \epsilon_0 > 0$ and $Z = A$, the supremum (8.1) is attained by a point $z_0 \in \operatorname{int}\Omega$, where the singular value $\sigma_{\min}(A - z_0 I)$ is simple. Then for any corresponding unit minimal left and right singular vectors $u, v \in \mathbf{C}^n$, the number v^*u is real and nonpositive.*

Now suppose furthermore that z_0 is the unique attaining point in (8.1), that it is nondegenerate (or, in other words, $v^*u \neq 0$), and that the Hessian $\nabla^2(g^2)(z_0)$ is nonsingular. Then the function α^Ω is smooth around the point (A, ϵ_0), with

$$\nabla_Z \alpha^\Omega(A, \epsilon_0) = \frac{uv^*}{v^*u} \quad \text{and} \quad \nabla_\epsilon \alpha^\Omega(A, \epsilon_0) = -\frac{1}{v^*u}.$$

Proof. Consider the optimization problem

$$\left\{ \begin{array}{lrcl} \sup & \operatorname{Re} z & & \\ \text{subject to} & \sigma_{\min}^2(Z - zI) & \le & \epsilon^2, \\ & z & \in & \Omega. \end{array} \right.$$

When $(Z, \epsilon) = (A, \epsilon_0)$ this problem becomes

$$\left\{ \begin{array}{lrcl} \sup & \operatorname{Re} z & & \\ \text{subject to} & g^2(z) & \le & \epsilon_0^2, \\ & z & \in & \Omega, \end{array} \right.$$

with optimal solution z_0. By Corollary 7.2 (analytic boundary), the function g^2 is smooth near z_0, with gradient

$$\nabla g^2(z_0) = 2g(z_0)\nabla g(z_0) = -2\epsilon_0 v^* u.$$

Either this gradient is zero or there is a Lagrange multiplier $\mu \in \mathbf{R}_+$ such that the gradient of the Lagrangian

$$z \mapsto \operatorname{Re} z - \mu(g^2(z) - \epsilon_0^2)$$

at $z = z_0$ is zero. In this case,

$$1 + 2\mu\epsilon_0 v^* u = 0, \tag{8.2}$$

so the first part follows.

Moving to the second part, (8.2) implies $\mu = -(2\epsilon_0 v^* u)^{-1}$. Under the additional assumptions we can apply a standard sensitivity result (for example, [6, Thm 5.5.3])

to deduce that the gradient of the optimal value of the original optimization problem at (A, ϵ_0) equals the gradient of the Lagrangian

$$(Z, \epsilon) \mapsto \operatorname{Re} z_0 + (2\epsilon_0 v^* u)^{-1}(\sigma_{\min}^2(Z - z_0 I) - \epsilon^2)$$

at (A, ϵ_0). The result now follows by Theorem 7.1 (analytic singular value). ∎

An eigenvalue of the matrix A with real part equal to the spectral abscissa α is called *active*.

THEOREM 8.2 (regular representation). *If the matrix A has s distinct active eigenvalues, all of which are nonderogatory, then there exist s functions*

$$\gamma_j : \mathbf{M}^n \times \mathbf{R}_{++} \to \mathbf{R} \quad (j = 1, 2, \ldots, s),$$

such that for small $\epsilon > 0$ and matrices Z close to A, each map

$$(Z, \epsilon) \mapsto \gamma_j(Z, \epsilon)$$

is smooth and satisfies $\gamma_j(A, 0) = \alpha(A)$, the pseudospectral abscissa can be expressed as

$$\alpha_\epsilon(Z) = \max\{\gamma_j(Z, \epsilon) : j = 1, 2, \ldots, s\},$$

and the set of gradients

$$\{\nabla_Z \gamma_j(A, \epsilon) : j = 1, 2, \ldots, s\}$$

is linearly independent.

Proof. Denote the distinct eigenvalues of A by $\lambda_1, \lambda_2, \ldots, \lambda_k$, where

$$\operatorname{Re} \lambda_j \begin{cases} = & \alpha \quad (j \le s), \\ < & \alpha \quad (j > s). \end{cases}$$

Let D denote the open unit disk in $\mathbf{C}$. Providing we choose a radius $\delta > 0$ sufficiently small, we have

$$2\delta < |\lambda_p - \lambda_q| \text{ for all } p \ne q,$$
$$\delta + \operatorname{Re} \lambda_j < \alpha \text{ for all } j > s,$$

and so the open disks $\lambda_j + \delta D$ are disjoint, and those with $j > m$ lie in the half-plane $\operatorname{Re} z < \alpha$. Furthermore, again by reducing δ if necessary, Theorem 7.4 (growth near eigenvalues) guarantees that each of the functions

$$g^2\big|_{\lambda_j + \delta D} \quad (j = 1, 2, \ldots, s)$$

is smooth, with everywhere positive definite Hessian except possibly at λ_j.

We claim that the small pseudospectra (by which we mean pseudospectra corresponding to small ϵ) of matrices close to A lie in small disks around the eigenvalues of A. More precisely, for small $\epsilon \ge 0$ and matrices Z close to A, we claim

$$\{z \in \mathbf{C} : \sigma_{\min}(Z - zI) \le \epsilon\} \subset \{\lambda_1, \lambda_2, \ldots, \lambda_k\} + \delta D. \tag{8.3}$$

Otherwise there would exist sequences $\epsilon_r \to 0$, $Z_r \to A$, and $z_r \in \mathbf{C}$ satisfying, for all $r = 1, 2, \ldots$,

$$\sigma_{\min}(Z_r - z_r I) \le \epsilon_r,$$
$$|z_r - \lambda_j| \ge \delta \quad (j = 1, 2, \ldots, k).$$

The first inequality above implies the sequence $\{z_r\}$ is bounded, so has a cluster point z_0, which must satisfy the inequalities

$$\begin{aligned}\sigma_{\min}(A - z_0 I) &\leq 0,\\ |z_0 - \lambda_j| &\geq \delta \quad (j = 1, 2, \ldots, k).\end{aligned}$$

The first inequality above can only hold if z_0 is an eigenvalue of A, which contradicts the second inequality. Hence inequality (8.3) holds, as we claimed.

Using the notation of (8.1), we can, for small $\epsilon > 0$, matrices Z close to A and, for each $j = 1, 2, \ldots, m$, define functions

$$\gamma_j(Z, \epsilon) = \alpha^{\lambda_j + \delta D}(Z, \epsilon) = \sup\{\operatorname{Re} z : |z - \lambda_j| < \delta,\ \sigma_{\min}(Z - zI) \leq \epsilon\},$$

and as a consequence of inclusion (8.3), we can then write

$$\alpha_\epsilon(Z) = \max\{\gamma_j(Z, \epsilon) : j = 1, 2, \ldots, m\}.$$

We claim each function γ_j is smooth around the point (A, ϵ_0) for any small $\epsilon_0 > 0$.

To prove this claim, we use Theorem 8.1 (smoothness of pseudospectral abscissa). For any $j = 1, 2, \ldots, s$, consider the supremum

$$\begin{aligned}\gamma_j(A, \epsilon_0) &= \sup\{\operatorname{Re} z : |z - \lambda_j| < \delta,\ \sigma_{\min}(A - zI) \leq \epsilon_0\}\\ &= \sup\left\{\operatorname{Re} z :\ g^2\big|_{\lambda_j + \delta D}(z) \leq \epsilon_0^2\right\}.\end{aligned}$$

By our choice of the radius δ, this supremum is attained at a unique point z_j (cf. Corollary 7.5 (convexity)), which is nondegenerate (cf. Corollary 7.6 (smoothness)), and at which the Hessian $\nabla^2(g^2)(z_j)$ is positive definite. Hence the function γ_j is smooth around (A, ϵ_0), with gradient

$$\nabla_Z \gamma_j(A, \epsilon_0) = \frac{u_j v_j^*}{v_j^* u_j},$$

where u_j, v_j are unit minimal left and right singular vectors for $A - z_j I$, and $v_j^* u_j$ is real and strictly negative.

To complete the proof, it suffices to show that the set of matrices

$$\{u_j v_j^* : j = 1, 2, \ldots, s\} \subset \mathbf{M}^n$$

is linearly independent providing our choice of radius $\delta > 0$ is sufficiently small. If this fails, then for each j there is a sequence of points $z_j^r \to \lambda_j$ and sequences of unit minimal left and right singular vectors u_j^r, v_j^r for $A - z_j^r I$ such that the set of matrices

$$\{u_j^r (v_j^r)^* : j = 1, 2, \ldots, s\} \subset \mathbf{M}^n$$

is linearly dependent. By taking subsequences, we can suppose $u_j^r \to u_j^0$ and $v_j^r \to v_j^0$ for each j, and then the set

$$S = \{u_j^0 (v_j^0)^* : j = 1, 2, \ldots, s\} \subset \mathbf{M}^n$$

must be linearly dependent. But it also follows that u_j^0, v_j^0 are unit left and right eigenvectors for A corresponding to the eigenvalue λ_j. Since the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$ are distinct, the sets of eigenvectors

$$\{u_j^0 : j = 1, 2, \ldots, m\} \quad \text{and} \quad \{v_j^0 : j = 1, 2, \ldots, s\}$$

are each linearly independent, and a standard exercise then shows the contradiction that the set S above is linearly independent. ∎

As a consequence of this result, the pseudospectral abscissa must be a reasonably well-behaved nonsmooth function near a matrix with all nonderogatory eigenvalues. Specifically, we have the following result. We refer the reader to [17, 34] for standard nonsmooth terminology.

COROLLARY 8.3 (regularity). *If all the active eigenvalues of a matrix $A \in \mathbf{M}^n$ are nonderogatory, then for all small $\epsilon > 0$ the pseudospectral abscissa α_ϵ is locally Lipschitz and subdifferentially regular around A.*

Proof. This follows immediately from the representation as a maximum of smooth functions in the previous result [17, Prop. 2.3.12]. ∎

This corollary presents an interesting parallel with a key result in [15]. This result states that the spectral abscissa, even though non-Lipschitz, is a subdifferentially regular function around the matrix A if and only if each active eigenvalue of A is nonderogatory.

By combining the representation of α_ϵ constructed in the proof of Theorem 8.2 with the growth estimate of Theorem 7.4, we can also see how the pseudospectral abscissa depends on the parameter ϵ.

COROLLARY 8.4 (dependence on ϵ). *If all the active eigenvalues of the matrix A are nonderogatory, with maximum algebraic multiplicity m, then as a function of $\epsilon \geq 0$ we have*

$$\alpha_\epsilon - \alpha \sim \gamma \epsilon^{1/m} \quad \text{as } \epsilon \downarrow 0,$$

for some constant $\gamma > 0$.

9. Nonsmooth geometry. What about points $z_0 \in \mathbf{C}$, where $\sigma_{\min}(A - z_0 I)$ is multiple? The function $\sigma_{\min}$ is nonsmooth at any matrix with a multiple smallest singular value, so the function g may be nonsmooth at z_0. An appropriate approach to studying the pseudospectrum near z_0 is therefore to use nonsmooth analysis. We again refer to [17, 34] for the standard concepts.

For any point $z \in \mathbf{C}$ we consider the subspace $U(z) \subset \mathbf{C}^n$ spanned by all right singular vectors corresponding to $\sigma_{\min}(A - zI)$, and we define a subset of $\mathbf{C}$ by

$$G(z) = \{u^*(A - zI)u : u \in U(z),\ \|u\| = 1\}.$$

PROPOSITION 9.1 (convexity). *The set $G(z)$ is nonempty, compact, and convex.*

Proof. Define a linear map $B : U(z) \to U(z)$ by

$$Bu = P_{U(z)}((A - zI)u),$$

where $P_{U(z)} : \mathbf{C}^n \to U(z)$ denotes the orthogonal projection. Now notice for all vectors $u \in U(z)$ we have

$$\begin{aligned}\langle u, Bu\rangle = \langle u, P_{U(z)}((A - zI)u)\rangle &= \langle P^*_{U(z)}u, (A - zI)u\rangle \\ &= \langle u, (A - zI)u\rangle = u^*(A - zI)u,\end{aligned}$$

since the map $P^*_{U(z)} : U(z) \to \mathbf{C}^n$ is just the embedding. We deduce

$$G(z) = \{\langle u, Bu\rangle : u \in U(z),\ \|u\| = 1\},$$

and this set is nonempty, compact, and convex, by the Toeplitz–Hausdorff theorem [22]. ∎

The next result gives another perspective on the pointedness of the pseudospectrum (recall Proposition 4.7 (pointedness)).

THEOREM 9.2 (nonsmooth boundary behavior). *For complex z_0 satisfying $g(z_0) = \epsilon$ and $0 \notin G(z_0)$, the complement of the strict pseudospectrum,*

$$\{z \in \mathbf{C} : g(z) \geq \epsilon\},$$

is Clarke regular at z_0, with normal cone $\mathrm{cone}(G(z_0))$.

Proof. The complement of the strict pseudospectrum is

$$\begin{aligned}
&\{z \in \mathbf{C} : \sigma_{\min}(A - zI) \geq \epsilon\} \\
&= \{z : \lambda_{\min}((A - zI)^*(A - zI)) \geq \epsilon^2\} \\
&= \{z : F(z) \in \mathbf{H}^n_+\} = F^{-1}(\mathbf{H}^n_+),
\end{aligned}$$

where $\mathbf{H}^n$ denotes the Euclidean space of $n \times n$ Hermitian matrices, with inner product $\langle X, Y \rangle = \mathrm{Re}\,(\mathrm{tr}\,(XY))$ and positive semidefinite cone $\mathbf{H}^n_+$, the function $\lambda_{\min} : \mathbf{H}^n \to \mathbf{R}$ is the smallest eigenvalue, and the function $F : \mathbf{C} \to \mathbf{H}^n$ is defined by

$$F(z) = (A - zI)^*(A - zI) - \epsilon^2 I.$$

The gradient map $\nabla F(z_0) : \mathbf{C} \to \mathbf{H}^n$ is given by

$$\nabla F(z_0)(w) = -w^* A - wA^* + 2\langle w, z_0 \rangle I,$$

and a short calculation shows that the adjoint map $\nabla F(z_0)^* : \mathbf{H}^n \to \mathbf{C}$ is given by

$$\nabla F(z_0)^* X = 2\mathrm{tr}\,((z_0 I - A)X).$$

It is well known (see, for example, [25]) that the positive semidefinite cone is Clarke regular at $F(z_0)$ (being convex), with normal cone

$$\begin{aligned}
N_{\mathbf{H}^n_+}(F(z_0)) &= -\mathrm{cone}\{uu^* : F(z_0)u = 0\} \\
&= -\mathrm{cone}\{uu^* : u \in U(z_0),\ \|u\| = 1\}.
\end{aligned}$$

Now consider any matrix

$$X \in N_{\mathbf{H}^n_+}(F(z_0)) \cap N(\nabla F(z_0)^*).$$

By the calculations above, we deduce

$$-X = \sum_{j=1}^{k} \mu_j u_j u_j^*$$

for some integer k, reals $\mu_j \geq 0$, and unit vectors $u_j \in U(z_0)$ $(j = 1, 2, \ldots, k)$, and

$$0 = \mathrm{tr}\,((A - z_0 I)X) = \sum_{j=1}^{k} \mu_j u_j^*(A - z_0 I)u_j.$$

But since $0 \notin G(z_0)$, by Proposition 9.1 (convexity) this implies that each μ_j is zero. We have therefore proved the condition

$$N_{\mathbf{H}^n_+}(F(z_0)) \cap N(\nabla F(z_0)^*) = \{0\}.$$

Under this condition we can apply a standard chain rule [34] to the set of interest, $F^{-1}(\mathbf{H}_+^n)$, to deduce that it is Clarke regular at the point z_0, with normal cone

$$N_{F^{-1}(\mathbf{H}_+^n)}(z_0) = \nabla F(z_0)^* N_{\mathbf{H}_+^n}(F(z_0)) = \text{cone}(G(z_0)),$$

as required. □

Consider, for instance, Example 1 (nonsmooth points). When $\epsilon = \sqrt{2}$, we saw that the point $z_0 = 1$ is a nonsmooth point on the boundary of the pseudospectrum, which consists of the union of two disks of radius $\sqrt{2}$, centered at $\pm i$. A calculation shows that the set $G(z_0)$ in this case is the line segment $[1 - i, 1 + i]$, so according to the above result, the normal cone to the complement of the strict pseudospectrum is the cone $\{x + iy : |y| \leq -x\}$, as we expect.

By contrast, when $\epsilon = 1$ we saw that the pseudospectrum consists of two unit disks, tangent at 0. A calculation shows $G(0)$ is the line segment $[-i, i]$, which contains 0, so the above theorem does not apply.

Acknowledgments. Many thanks to Peter Benner, Carsten Scherer, and two anonymous referees for many helpful suggestions and references.

REFERENCES

[1] V. I. Arnold, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.

[2] V. Balakrishnan, S. Boyd, and S. Balemi, *Branch and bound algorithm for computing the minimum stability degree of parameter-dependent linear systems*, Internat. J. Robust Nonlinear Control, 1 (1991), pp. 295–317.

[3] H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser-Verlag, Basel, 1985.

[4] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, SIAM, Philadelphia, 2001.

[5] V. Blondel and J. N. Tsitsiklis, *NP-hardness of some linear control design problems*, SIAM J. Control Optim., 35 (1997), pp. 2118–2127.

[6] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.

[7] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞-norm*, Systems Control Lett., 15 (1990), pp. 1–7.

[8] S. Boyd, V. Balakrishnan, and P. Katamba, *A bisection method for computing the H_∞ norm of a transfer matrix and related problems*, Math. Control Signals Systems, 2 (1989), pp. 207–219.

[9] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[10] M. Brühl, *A curve tracing algorithm for computing the pseudospectrum*, BIT, 36 (1996), pp. 441–454.

[11] N. A. Bruinsma and M. Steinbuch, *A fast algorithm to compute the $\mathbf{H}_\infty$-norm of a transfer function matrix*, Systems Control Lett., 14 (1990), pp. 287–293.

[12] J. V. Burke, A. S. Lewis, and M. L. Overton, *Optimal stability and eigenvalue multiplicity*, Foundations Comput. Math., 1 (2001), pp. 205–225.

[13] J. V. Burke, A. S. Lewis, and M. L. Overton, *Robust stability and a criss-cross algorithm for pseudospectra*, IMA J. Numer. Anal., to appear.

[14] J. V. Burke, A. S. Lewis, and M. L. Overton, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351/352 (2002), pp. 117–145.

[15] J. V. Burke and M. L. Overton, *Variational analysis of non-Lipschitz spectral functions*, Math. Programming, 90 (2001), pp. 317–352.

[16] 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.

[17] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. Republished as Classics in Applied Math. 5, SIAM, Philadelphia, 1990.

[18] Y. Genin, P. Van Dooren, and V. Vermaut, *Convergence of the calculation of* $\mathbf{H}_\infty$ *norms and related questions*, in Proceedings MTNS-98, G. Picci and D. S. Gilliam, eds., Birkhäuser-Verlag, Basel, 1998.
[19] C. He and G. A. Watson, *An algorithm for computing the distance to instability*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 101–116.
[20] *Pseudospectra Gateway*, http://web.comlab.ox.ac.uk/projects/pseudospectra.
[21] D. Hinrichsen and A. J. Pritchard, *Stability radii of linear systems*, Systems Control Lett., 7 (1986), pp. 1–10.
[22] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[23] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer, New York, 1982.
[24] C. Lawrence, A. L. Tits, and P. Van Dooren, *A fast algorithm for the computation of an upper bound on the* μ*-norm*, Automatica, 36 (2000), pp. 449–456.
[25] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.
[26] A. S. Lewis, *Robust regularization*, Math. Program., submitted.
[27] V. B. Lidskii, *Perturbation theory of non-conjugate operators*, USSR Comput. Math. and Math. Phys., 1 (1965), pp. 73–85.
[28] C. F. Van Loan, *How near is a stable matrix to an unstable matrix?* Contemp. Math., 47 (1985), pp. 465–477.
[29] J. W. Milnor, *Singular Points of Complex Hypersurfaces*, Princeton University Press, Princeton, NJ, 1968.
[30] J. Moro, J. V. Burke, and M. L. Overton, *On the Lidskii–Vishik–Lyusternik perturbation theory for the eigenvalues of matrices with arbitrary Jordan structure*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 793–817.
[31] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[32] B. T. Polyak, *Convexity of nonlinear image of a small ball with applications to optimization*, Set-Valued Anal., 9 (2001), pp. 159–168.
[33] X. Rao, K. Gallivan, and P. Van Dooren, *Stabilization of large scale dynamical systems*, in Proceedings MTNS 2000, S. L. Campbell, ed., Int. J. Appl. Math. Comput. Sci. II, University of Zielona Góra, Zielona, Góra, Poland, 2001.
[34] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[35] R. A. Silverman, *Introductory Complex Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[36] J. Sreedhar, P. Van Dooren, and A. L. Tits, *A fast algorithm to compute the real structured stability radius*, in Stability Theory: Proceedings of Hurwitz Centenary Conference, R. Jeltsch and M. Mansour, eds., Birkhäuser-Verlag, Basel, 1996, pp. 219–230.
[37] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[38] M. J. Todd, *Semidefinite optimization*, Acta Numer., 10 (2001), pp. 515–560.
[39] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.
[40] L. N. Trefethen, *Computation of pseudospectra*, Acta Numer., 8 (1999), pp. 247–295.
[41] L. N. Trefethen, *private communication*, 2002.
[42] A. Varga and P. Parrilo, *Fast algorithms for solving* $\mathbf{H}_\infty$ *norm minimization problems*, in Proceedings of the 40th IEEE Conference on Decision and Control (Orlando, FL), IEEE, Piscataway, NJ, 2001.
[43] M. Vidyasagar, *Control Systems Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[44] G. Zames and B. A. Francis, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, 28 (1983), pp. 585–601.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 105–127

CONVERGENCE RATES OF SPECTRAL DISTRIBUTIONS OF LARGE SAMPLE COVARIANCE MATRICES*

Z. D. BAI†, BAIQI MIAO‡, AND JIAN-FENG YAO§

Abstract. In this paper, we improve known results on the convergence rates of spectral distributions of large-dimensional sample covariance matrices of size $p \times n$. Using the Stieltjes transform, we first prove that the expected spectral distribution converges to the limiting Marčenko–Pastur distribution with the dimension sample size ratio $y = y_n = p/n$ at a rate of $O(n^{-1/2})$ if y keeps away from 0 and 1, under the assumption that the entries have a finite eighth moment. Furthermore, the rates for both the convergence in probability and the almost sure convergence are shown to be $O_p(n^{-2/5})$ and $o_{a.s.}(n^{-2/5+\eta})$, respectively, when y is away from 1. It is interesting that the rate in all senses is $O(n^{-1/8})$ when y is close to 1.

Key words. convergence rate, random matrix, spectral distribution, Marčenko–Pastur distribution

AMS subject classifications. Primary 60F15; Secondary 62H99

PII. S0895479801385116

1. Introduction.

The spectral analysis of large-dimensional random matrices has been actively developed in the last decades since the initial contributions of Wigner (1955, 1958); also see the recent review by Bai (1999) and the book by Mehta (1991). Various limiting distributions were discovered including the Wigner semicircular law (Wigner, 1955), the Marčenko–Pastur law (Marčenko and Pastur, 1967), the limiting law for multivariate F matrices (Bai, Yin, and Krishnaiah (1987) and Silverstein (1985)) and the circular law (Bai and Yin (1986), Bai (1997)). The spectrum separation problem for large-dimensional sample covariance matrices was investigated in Bai and Silverstein (1998, 1999).

Let A be an $n \times n$ symmetric matrix, and $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of A. The spectral distribution F^A of A is defined as

$$F^A(x) = \frac{1}{n} \times \text{number of elements in } \{k : \lambda_k \le x\}.$$

Let $\mathbf{X}_p = (x_{ij})_{p\times n}$ be a $p \times n$ observation matrix whose entries are mutually independent and have a common mean zero and variance 1. The entries of $\mathbf{X}_p$ may depend on n but we suppress the index n for simplicity. In this paper, we consider the sample covariance matrix $\mathbf{S} = n^{-1}\mathbf{X}_p\mathbf{X}_p^T$, where $\mathbf{X}^T$ denotes the transpose of the matrix $\mathbf{X}$. Assume that the ratio p/n of sizes tends to a positive limit y as $n \to \infty$. Under suitable moment conditions on the x_{ij} entries, it is known that the empirical spectral distribution (ESD) $F_p := F^{\mathbf{S}}$ converges to the Marčenko–Pastur distribution F_y with

*Received by the editors February 15, 2001; accepted for publication (in revised form) by A. Edelman November 16, 2002; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/38511.html

†Department of Mathematics, Northeast Normal University, Changchun, 130024 China, and Department of Statistics and Applied Probability, National University of Singapore, Singapore, 117543 (stabaizd@leonis.nus.edu.sg). The research of this author was partially supported by NSFC grant 201471000 as well as by the NUS Research grant R-155-030-112.

‡Department of Statistics and Finance, University Science and Technology of China, Hefei, Anhui, China (bqmiao@ustc.edu.cn). The research of this author was supported in part by the National Foundation of Natural Science of China.

§IRMAR, Université de Rennes 1, Campus de Beaulieu, F-35042 Rennes Cedex, France (jian-feng.yao@univ-rennes1.fr).

index y with density

$$F_y'(x) = \begin{cases} \frac{1}{2\pi xy}\sqrt{(x-a)(b-x)} & \text{if } a < x < b, \\ 0 & \text{otherwise,} \end{cases}$$

where $a = (1-\sqrt{y})^2$, $b = (1+\sqrt{y})^2$.

An important question here concerns the problem of the convergence rates. However, no significant progress was made before the introduction of a novel and powerful tool, namely, the Berry–Esseen inequalities in terms of Stieltjes transforms, by Bai (1993a, 1993b). Using this methodology, Bai (1993b) proved that the expected ESD $\mathbb{E}F_p$ converges to F_{y_n} at a rate of $O(n^{-1/4})$ or $O(n^{-5/48})$ depending on whether y_n is far away or close to 1, respectively, where $y_n = p/n$. In another work by Bai, Miao, and Tsay (1997), these rates are also established for the convergence in probability of the ESD F_p itself. In later works of Bai, Miao, and Tsay (1999, 2002), the convergence rates for large Wigner matrices are significantly improved.

In this work, we further investigate the convergence rates for empirical spectral distributions for large sample covariance matrices and improve those results in the theorems to follow.

The following conditions will be used:

(C.1) $\mathbb{E}x_{ij} = 0, \quad \mathbb{E}x_{ij}^2 = 1, \quad 1 \le i \le p,\ 1 \le j \le n.$

(C.2) $\sup_{i,j,n} \mathbb{E}|x_{ij}|^8 < \infty.$

(C.3) For any positive constant δ,

$$\sum_{ij} \mathbb{E}x_{ij}^8 I_{(|x_{ij}| \ge \delta\sqrt{n})} = o(n^2).$$

It is easy to see that condition (C.3) guarantees that there is a sequence $\{\delta = \delta_n \to 0\}$ such that

$$\sum_{ij} \mathbb{E}x_{ij}^8 I_{(|x_{ij}| \ge \delta\sqrt{n})} = o(n^2\delta^8). \tag{1.1}$$

(C.2′) $\sup_{i,j,n} \mathbb{E}|x_{ij}|^k < \infty$ for any integer $k \ge 1$.

Throughout the paper, we use the notation $Z_n = O_p(a_n)$ if the sequence $(a_n^{-1}Z_n)$ is tight and use $Z_n = o_p(a_n)$ when $a_n^{-1}Z_n$ tends to 0 in probability. We shall also set $\|f\| = \sup_x |f(x)|$.

For simplicity, from now on we drop the index n from y and use the notation $y = y_n = p/n$. Finally, let us define

$$\theta = \theta(n,y) = \begin{cases} \frac{-2\log_n(1-\sqrt{y})}{1+4\log_n(1-\sqrt{y})} & \text{if } \quad y \le (1-n^{-1/8})^2, \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{1.2}$$

We now introduce the main results of the paper.

THEOREM 1.1. *Assume that the conditions* (C.1)–(C.3) *are satisfied. Then,*

$$\|\mathbb{E}F_p - F_y\| = O\left(\frac{n^{-1/[4\theta+2]}}{[1-\sqrt{y}+n^{-1/[8\theta+4]}]}\right),$$

THEOREM 1.2. *Assume that the conditions* (C.1)–(C.3) *are satisfied. Then,*

$$\|F_p - F_y\| = O_p\left(\max\left\{\frac{n^{-(2/(5+\theta))}}{[1-\sqrt{y}+n^{-(1/(5+\theta))}]}, \frac{n^{-1/[4\theta+2]}}{[1-\sqrt{y}+n^{-1/[8\theta+4]}]}\right\}\right).$$

THEOREM 1.3. *Assume that the conditions* (C.1)–(C.3) *are satisfied. Then, with probability* 1*,*

$$\|F_p - F_y\| = o\left(\max\left\{\frac{n^{-(2/(5+\theta))+\eta}}{[1-\sqrt{y}+n^{-(1/(5+\theta))}]^2}, \frac{n^{-1/[4\theta+2]}}{[1-\sqrt{y}+n^{-1/[8\theta+4]}]^2}\right\}\right).$$

Remark on the convergence rates. If y is not close to 1, then $\theta \sim c/\log n$, and hence the convergence rates in the above three theorems are $O(n^{-1/2})$, $O_p(n^{-2/5})$, and $o_{a.s.}(n^{-2/5+\eta})$, respectively. When $y > 1 - O(n^{-1/8})$, $\theta = 1/2$, and hence the rates of the three theorems are $O(n^{-1/8})$, $O_p(n^{-1/8})$, and $O_{a.s.}(n^{-1/8})$, respectively. When y goes to 1 with intermediate rates, we may have intermediate convergence rates.

It is worth noticing that the convergence rates given above for the case $0 < y \le 1$ also apply to the case $y > 1$, since the last case can be reduced to the first case by interchanging the roles of row and column sizes p and n.

The proofs of these main results will be given in section 3. For convenience, we first introduce some necessary notation and preliminary consequences in section 2. Some necessary lemmas are postponed to section 4.

2. Definitions and easy consequences. Throughout the paper, the transpose of a possibly complex matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{\mathrm{T}}$ and its conjugate by $\overline{\mathbf{A}}$. For each fixed p, n, and $k = 1, \ldots, p$, let us denote by $\mathbf{x}_k = (x_{k1}, \ldots, x_{kn})^T$ the kth row of $\mathbf{X}_p$ arranged as a column vector, and let $\mathbf{X}_p(k)$ be the $(p-1) \times n$ submatrix obtained from $\mathbf{X}_p$ by deleting its kth row. Let us define

$$\begin{aligned}
&\alpha_k := \frac{1}{n}\mathbf{X}_p(k)\mathbf{x}_k, \qquad \mathbf{S}_k := \frac{1}{n}\mathbf{X}_p(k)\mathbf{X}_p^T(k), \qquad \mathbf{B}_k := \frac{1}{n}\mathbf{X}_p^T(k)\mathbf{D}_k\mathbf{X}_p(k),\\
&\mathbf{B} := \frac{1}{n}\mathbf{X}_p^T\mathbf{D}\mathbf{X}_p, \qquad \mathbf{D}_k := (\mathbf{S}_k - z\mathbf{I}_{p-1})^{-1}, \qquad \mathbf{D} := (\mathbf{S} - z\mathbf{I}_p)^{-1},\\
&\mathbf{\Gamma}_k := \mathbf{D}_k\overline{\mathbf{D}}_k, \qquad \mathbf{\Lambda}_k := \mathbf{D}_k\mathbf{S}_k\overline{\mathbf{D}}_k.
\end{aligned} \tag{2.1}$$

Here $\mathbf{I}_m$ is the m-dimensional identity matrix and z a complex number with a positive imaginary part.

Following Bai (1993b), the Stieltjes transform of the spectral distribution F_p of the sample covariance matrix $\mathbf{S}$ is defined for $z = u + iv$ with $v > 0$ by

$$m_p(z) = \int_{-\infty}^{\infty} \frac{1}{x - z} dF_p(x),$$

and it is well known that

$$m_p(z) = \frac{1}{p} tr(\mathbf{S} - z\mathbf{I}_p)^{-1}.$$

Similarly, the Stieltjes transform of the spectral distribution $F_p^{(k)}$ of the submatrix $\mathbf{S}_k$ satisfies

$$m_p^{(k)}(z) = \int_{-\infty}^{\infty} \frac{1}{x - z} dF_p^{(k)}(x) = \frac{1}{p-1} tr(\mathbf{S}_k - z\mathbf{I}_{p-1})^{-1}.$$

Finally, the Stieltjes transform of the "limiting" (by noting that $y = y_n$) Marčenko–Pastur distribution F_y is

$$m(z) = \int_{-\infty}^{\infty} \frac{1}{x - z} dF_y(x) = -\frac{y + z - 1 - \sqrt{(1-y-z)^2 - 4yz}}{2yz} \tag{2.2}$$

for $0 < y \le 1$. Here the square root $\sqrt{z}$ is the one with a positive imaginary part. Note that $m(z)$ is a root of the quadratic equation

$$yzm^2 + (y + z - 1)m + 1 = 0,$$

which implies that $m(z)m^*(z) = \frac{1}{yz}$, where

$$m^*(z) = -\frac{y + z - 1 + \sqrt{(1 + y - z)^2 - 4y}}{2yz}$$

is the other root of the equation. We claim that

$$|m(z)| < |m^*(z)| \quad \text{for all } \ z = u + iv,\ v > 0. \tag{2.3}$$

To see this, set

$$\alpha + i\beta = \sqrt{(1 + y - z)^2 - 4y} \quad \text{with} \quad \beta \ge 0.$$

We have

$$\alpha\beta = v(u - y - 1), \tag{2.4}$$

$$\beta^2 - \alpha^2 = (b - u)(u - a) + v^2. \tag{2.5}$$

First, note that $\beta = 0$ is impossible; otherwise we should have $u = 1 + y$ and (2.5) would be violated. Hence, $\beta > 0$ and $\alpha \ge 0$ if and only if $u \ge 1 + y$.

It is easy to see that

$$\begin{aligned} |m(z)| < |m^*(z)| &\Leftrightarrow |y + z - 1 - (\alpha + i\beta)| < |y + z - 1 + (\alpha + i\beta)| \\ &\Leftrightarrow \alpha(y - 1 + u) + \beta v > 0. \end{aligned}$$

The last inequality clearly holds if $u \ge (1 + y)$ or $u \le 1 - y$ (in this case the result was proved in Bai (1993b, p. 651)). Now assume for a $u \in (1 - y, 1 + y) \subset [a, b]$ that the inequality does not hold, i.e., $\alpha(y - 1 + u) + \beta v \le 0$. This implies that $\beta v \le |\alpha|[u - (1 - y)]$ (noting that $\alpha < 0$). Multiplying both sides by β and using (2.4), we get

$$\beta^2 \le [u - (1 - y)]\,[1 + y - u] \le (b - u)(u - a),$$

which contradicts (2.5). The claim (2.3) is then proved.

This claim implies that $|m(z)| \le 1/\sqrt{y|z|}$ for any z. On the other hand, when $u < a - v$, both real and imaginary parts of $m(z)$ are positive and increasing (a consequence of the integral formula (2.2) of $m(z)$). Thus, $|m(z)|$ can only reach its maximum when $u > a - v$. When $a < 2v$, we have $|m(z)| \le 1/\sqrt{yv} \le \frac{2\sqrt{2}}{\sqrt{y}(\sqrt{a} + \sqrt{v})} = \frac{2\sqrt{2}v_y}{\sqrt{y}v}$. When $a \ge 2v$, by noticing that $\sqrt{|z|} \ge \sqrt[4]{a^2/4 + v^2} \ge \frac{1}{2\sqrt{2}}(\sqrt{a} + \sqrt{v})$, we obtain the same bound as in the first case. Therefore, we obtain

$$|m(z)| \le \frac{2\sqrt{2}v_y}{\sqrt{y}v}, \tag{2.6}$$

where

$$v_y := v/[\sqrt{a} + \sqrt{v}] = v/[1 - \sqrt{y} + \sqrt{v}]. \tag{2.7}$$

LEMMA 2.1. *Let* $\mathbf{x} = (x_1, \dots, x_n)^T$ *and* $\mathbf{y} = (y_1, \dots, y_n)^T$ *be independent real random vectors with independent elements. Suppose that for all* $1 \le j \le n$, $\mathbb{E}x_j = \mathbb{E}y_j = 0$, $\mathbb{E}|x_j|^2 = \mathbb{E}|y_j|^2 = 1$, $\mathbb{E}|x_j|^4 \le L < \infty$ *and that* $\mathbf{A}$ *is an* $n \times n$ *complex symmetric matrix. Let* $\mu_k = \max_{j \le n}(\mathbb{E}|x_j|^k, \mathbb{E}|y_j|^k)$. *Then*

(i) $\mathbb{E}|\mathbf{x}^T\mathbf{A}\mathbf{y}|^2 = tr(\mathbf{A}\bar{\mathbf{A}})$;
(ii) $\mathbb{E}|\mathbf{x}^T\mathbf{A}\mathbf{x}|^2 \le Ltr(\mathbf{A}\bar{\mathbf{A}}) + |tr\mathbf{A}|^2$;
(iii) $\mathbb{E}|\mathbf{x}^T\mathbf{A}\mathbf{x} - tr\mathbf{A}|^2 \le L(tr\mathbf{A}\bar{\mathbf{A}})$;
(iv) $\mathbb{E}|\mathbf{x}^T\mathbf{A}\mathbf{x} - tr\mathbf{A}|^{2k} \le d_k\left[\mu_{4k}tr(\mathbf{A}\bar{\mathbf{A}})^k + (Ltr(\mathbf{A}\bar{\mathbf{A}}))^k\right]$ *for* $k \ge 2$ *and some positive constant* d_k *depending on* k *only.*

The proofs of (i)–(iii) are elementary and therefore omitted. The statement (iv) follows from Lemma 2.7 of Bai and Silverstein (1998).

LEMMA 2.2. *Let* G_1 *and* G_2 *be probability distribution functions and* $z = u + iv$, $v > 0$. *Then for each positive integer* m,

$$\left|\int_{-\infty}^{\infty} \frac{1}{|x-z|^m} d(G_1(x) - G_2(x))\right| \le \frac{2}{v^m}\|G_1 - G_2\|.$$

Proof. Let be $G^* := G_1 - G_2$. We have, by integration by parts,

$$\begin{aligned}
&\left|\int_{-\infty}^{\infty} \frac{1}{|x-z|^m} d(x)G^*\right| \\
&= \left|-\int_{-\infty}^{\infty} G^*(x) d\left[\frac{1}{|x-z|^m}\right]\right| \\
&= \left|-\int_{-\infty}^{Re(z)} G^*(x) d\left[\frac{1}{|x-z|^m}\right] + \int_{Re(z)}^{\infty} G^*(x) d\left[-\frac{1}{|x-z|^m}\right]\right| \\
&\le \|G^*\|\left\{\int_{-\infty}^{Re(z)} d\left[\frac{1}{|x-z|^m}\right] + \int_{Re(z)}^{\infty} d\left[-\frac{1}{|x-z|^m}\right]\right\} \\
&= \|G^*\|\left\{\frac{1}{|x-z|^m}\Big|_{-\infty}^{Re(z)} + \left(-\frac{1}{|x-z|^m}\Big|_{Re(z)}^{\infty}\right)\right\} = \|G^*\|\frac{2}{v^m}. \qquad \square
\end{aligned}$$

We will need the following auxiliary variables:

$$\begin{aligned}
\varepsilon_k &= -\frac{1}{n}\sum_{j=1}^{n}(x_{kj}^2 - 1) + \frac{1}{n}({\mathbf{x}_k}^T\mathbf{B}_k\mathbf{x}_k - \mathbb{E}tr\mathbf{B}), \\
\varepsilon_k^* &= -\frac{1}{n}\sum_{j=1}^{n}(x_{kj}^2 - 1) + \frac{1}{n}({\mathbf{x}_k}^T\mathbf{B}_k\mathbf{x}_k - tr\mathbf{B}_k), \\
\widetilde{\varepsilon}_k &= \frac{1}{n}(tr\mathbf{B}_k - \mathbb{E}tr\mathbf{B}_k) = \frac{z}{n}(tr\mathbf{D}_k - \mathbb{E}tr\mathbf{D}_k), \\
\pi_k &= \frac{1}{n}\mathbb{E}(tr\mathbf{B}_k - tr\mathbf{B}) = \frac{z}{n}\mathbb{E}(tr\mathbf{D}_k - tr\mathbf{D}), \\
\beta_k &= -\frac{1}{n}\sum_{j=1}^{n}(x_{kj}^2 - 1) + z - 1 + \frac{1}{n}{\mathbf{x}_k}^T\mathbf{B}_k\mathbf{x}_k, \\
\beta_k^* &= z - 1 + \frac{1}{n}tr\mathbf{B}_k, \\
\beta &= z - 1 + \frac{1}{n}tr\mathbf{B}.
\end{aligned}$$

We summarize below some inequalities which will be used in the derivations. Let $\Delta = \| \mathbb{E}F_p - F_y\|$ and $M := \sup_{i,j,n} \mathbb{E}|x_{ij}|^4$. For fixed (n,p) and $1 \le k \le p$, we define the σ-algebra

$$\mathcal{F}^{(k)} = \sigma(\mathbf{x}_i \ : \ 1 \le i \le p,\ i \ne k), \quad \mathcal{F}_k = \sigma(\mathbf{x}_i \ : \ k < i \le p).$$

1. (from Lemma 3.3 of Bai (1993a)):

$$|(p-1)F_p^{(k)}(x) - pF_p(x)| \le 1. \tag{2.8}$$

2. (from Lemma 2.2 and (2.8)):

$$|tr\mathbf{D} - tr\mathbf{D}_k| = \left| \int_{-\infty}^{\infty} \frac{d[pF_p(x) - (p-1)F_p^{(k)}(x)]}{x-z} \right| \le 2v^{-1}. \tag{2.9}$$

3. (from equation (3.14) of Bai (1993b)):

$$m_p(z) = \int_0^{\infty} \frac{1}{x-z}\, dF_p(x) = \frac{1}{p} tr\mathbf{D} = -\frac{1}{p} \sum_{k=1}^{p} \frac{1}{\beta_k}. \tag{2.10}$$

4. (from Lemma 2.2 of Bai, Miao, and Tsay (1997)):

$$\mathbb{E}|m_p(z) - \ \mathbb{E}(m_p(z))|^2 \le p^{-1}v^{-2}. \tag{2.11}$$

5. (from $|\beta_k^*| \ge Im(\beta_k^*) = v(1 + n^{-1}tr\mathbf{\Lambda}_k)$):

$$|\beta_k^*|^{-1}(1 + n^{-1}tr\mathbf{\Lambda}_k) \le v^{-1}. \tag{2.12}$$

6.

$$|\beta_k| \ge Im(\beta_k) = v\left(1 + \frac{1}{n}\alpha_k^T \mathbf{D}_k \overline{\mathbf{D}_k} \alpha_k\right). \tag{2.13}$$

7.

$$\left|1 + \frac{1}{n}\alpha_k^T \mathbf{D}_k{}^2 \alpha_k\right| \le 1 + \frac{1}{n}\alpha_k^T \mathbf{D}_k \overline{\mathbf{D}_k} \alpha_k. \tag{2.14}$$

Let λ_{kj}, $j = 1, 2, \ldots, p-1$, be the eigenvalues of $\mathbf{S}_k$ which can be decomposed into a diagonal form on the basis of orthonormal and real eigenvectors. Let $\mathbf{L}$ be a complex matrix having the product form $\mathbf{L} = \mathbf{M}^{\ell}\mathbf{N}^{\ell'}$ for some integers ℓ, ℓ' and factors $\mathbf{M}$, $\mathbf{N}$ equal to one of the matrices $\{\mathbf{D}_k, \overline{\mathbf{D}}_k, \mathbf{S}_k\}$. An important feature that we will frequently use in what follows is that such a matrix $\mathbf{L}$ can be decomposed into a diagonal form *on the same basis as the eigenvectors of* $\mathbf{S}_k$. Moreover, the eigenvalues of $\mathbf{L}$ can be straightforwardly expressed in terms of the λ_{kj}'s. In particular, we have the following.

LEMMA 2.3. *Assume that* $|z| \le T$, *where* $T \ge 1$. *Then for all integers* $\ell \ge 1$,

$$tr(\mathbf{\Gamma}_k)^{\ell} \le \left(\frac{1}{v^2}\right)^{\ell-1} tr\mathbf{\Gamma}_k, \tag{2.15}$$

$$tr(\mathbf{\Lambda}_k)^{\ell} \le \left(\frac{T}{v^2}\right)^{\ell-1} tr\mathbf{\Lambda}_k. \tag{2.16}$$

Proof. (i) The inequality (2.15) follows from

$$tr(\mathbf{\Gamma}_k)^\ell = \sum_{j=1}^{p-1} \frac{1}{|\lambda_{kj} - z|^{2\ell}} \le v^{-2(\ell-1)} \sum_{j=1}^{p-1} \frac{1}{|\lambda_{kj} - z|^2} = v^{-2(\ell-1)} tr\mathbf{\Gamma}_k.$$

(ii) For the inequality (2.16), we have

$$tr(\mathbf{\Lambda}_k)^\ell = \sum_{j=1}^{p-1} \frac{\lambda_{kj}^\ell}{|\lambda_{kj} - z|^{2\ell}}.$$

The conclusion follows from the fact that the function $\varphi(\lambda) := \lambda^{-1}|\lambda - z|^2$ defined on $(0, \infty)$ is convex and has a unique minimum value φ^* satisfying

$$\varphi^* = 2(\sqrt{u^2 + v^2} - u) = 2\frac{v^2}{|z| + u} \ge \frac{v^2}{T}. \qquad \square$$

LEMMA 2.4. *For the Marčenko–Pastur distribution F_y, we have*

$$\int_a^b \frac{1}{|x - z|^2} dF_y(x) \le \frac{2}{y^{3/4}} v_y v^{-2}. \tag{2.17}$$

Proof. Since for $x \in [a, b]$, $(x - a)(b - x) \le x(b - a) = 4x\sqrt{y}$, we have for any z,

$$\begin{aligned} \int_a^b \frac{1}{|x - z|^2} dF_y(x) &= \int_a^b \frac{1}{|x - z|^2} \frac{1}{2\pi xy} \sqrt{(x - a)(b - x)}\, dx \\ &\le \frac{1}{\pi y^{3/4}} \int_a^b \frac{1}{\sqrt{x}|x - z|^2}\, dx. \end{aligned}$$

The maximum of $\int_a^b 1/(\sqrt{x}|x - z|^2)\, dx$ can only be attained when $u \ge a$. We then have

$$\begin{aligned} \int_a^b \frac{1}{|x - z|^2} dF_y(x) &\le \frac{1}{\pi y^{3/4}} \int_0^\infty \frac{1}{\sqrt{x}|x - z|^2}\, dx \\ &= \frac{1}{y^{3/4}} \frac{1}{|z|\sqrt{2|z| - 2u}} \le \frac{1}{y^{3/4}} |z|^{-1/2} v^{-1}. \end{aligned}$$

From this and by noticing that $|z|^{1/2} \ge \frac{1}{2}(\sqrt{a} + \sqrt{v})$ when $u > a$, the lemma is proved. $\square$

3. Proofs. We first truncate and centralize the random variables so that all random variables could be further considered as bounded (up to some order of n). In the subsection 3.2, we introduce a Bai inequality for the proofs of the main theorems. These proofs are then given in subsequent sections.

3.1. Truncation and centralization. Define $\hat{x}_{ij} = x_{ij} I(|x_{ij}| \le \delta\sqrt{n})$ and $\tilde{x}_{ij} = (\hat{x}_{ij} - \mathbb{E}(\hat{x}_{ij}))/\sigma_{ij}$, where $\sigma_{ij}^2 = \mathbb{E}(\hat{x}_{ij} - \mathbb{E}(\hat{x}_{ij}))^2$. Here $\delta = \delta_n$ is chosen such that $\delta_n \to 0$ with a slow rate and such that (1.1) holds. We remind the reader that all the above variables depend on n, but the index is suppressed.

Define $p \times n$ matrices $\hat{X} = (\hat{x}_{ij})$ and $\tilde{X} = (\tilde{x}_{ij})$ and define $p \times p$ matrices $\hat{S} = \frac{1}{n}\hat{X}\hat{X}^T$ and $\tilde{S} = \frac{1}{n}\tilde{X}\tilde{X}^T$. Denote the ESDs of $\hat{S}$ and $\tilde{S}$ by $\hat{F}_p$ and $\tilde{F}_p$, respectively.

We first estimate the truncation error $\|F_p - \hat{F}_p\|$. By (1.1),

$$\begin{aligned}&\sum_{i,j} P\left(|x_{ij}| \ge \delta_n\sqrt{n}\right)\\ &= \sum_{i,j} \mathbb{E}I\left(|x_{ij}| \ge \delta_n\sqrt{n}\right) \le (\delta_n\sqrt{n})^{-8}\sum_{i,j}\mathbb{E}\left[x_{ij}^8 I\left(|x_{ij}| \ge \delta_n\sqrt{n}\right)\right]\\ &\le cn^{-2}. \qquad (3.1)\end{aligned}$$

Let $\alpha \in (0,1)$. By the Markov inequality,

$$P\left(\sum_{i,j} I(|x_{ij}| \ge \delta_n\sqrt{n}) \ge n^{-\alpha}\right) \le cn^{-2+\alpha},$$

which, together with the Borel–Cantelli lemma, implies that

$$\operatorname{rank}(X - \hat{X}) \le \sum_{i,j} I(|x_{ij}| \ge \delta_n\sqrt{n}) = O(n^{-\alpha}) \quad \text{a.s.} \qquad (3.2)$$

By Lemma 2.6 of Bai (1999), we have

$$\|F_p - \hat{F}_p\| = O(1/n^{1+\alpha}) \quad \text{a.s.} \qquad (3.3)$$

The estimation (3.3) reduces the proofs to show that the three theorems remain true when F_p is replaced with $\hat{F}_p$.

Furthermore, recalling the proof of Lemma 2.7 of Bai (1999), we find that

$$\begin{aligned}\int |\hat{F}_p(x) - \tilde{F}_p(x)|dx &= \frac{1}{p}\sum_{k=1}^{p}|\hat{\lambda}_k - \tilde{\lambda}_k|\\ &\le \left(\frac{1}{np}tr(\hat{X}-\tilde{X})(\hat{X}-\tilde{X})^T\frac{2}{np}tr(\hat{X}\hat{X}^T + \tilde{X}\tilde{X}^T)\right)^{1/2}, \qquad (3.4)\end{aligned}$$

where $\hat{\lambda}_k$ and $\tilde{\lambda}_k$, arranged in increasing order, are the eigenvalues of $\hat{S}$ and $\tilde{S}$, respectively.

Under the uniform boundedness of the fourth moments of the entries, it is easy to show that

$$\frac{1}{np}tr\hat{X}\hat{X}^T \le \frac{1}{np}\sum_{ij}|x_{ij}|^2 \to 1 \quad \text{a.s.} \qquad (3.5)$$

Also,

$$1 \ge \max_{ij}\sigma_{ij}^2 \ge \min_{ij}\sigma_{ij}^2 \to 1. \qquad (3.6)$$

Furthermore, by (3.5) and (3.6),

$$\begin{aligned}\left(\frac{1}{np}tr\tilde{X}\tilde{X}^T\right)^{1/2} &= \frac{1}{\sqrt{np}}\left(\sum_{ij}|\tilde{x}_{ij}|^2\right)^{1/2}\\ &\le \frac{1}{\sqrt{np}(\min_{ij}\sigma_{ij})}\left(\left(\sum_{ij}|x_{ij}|^2\right)^{1/2} + \left(\sum_{ij}\mathbb{E}|x_{ij}|^2 I(|x_{ij}| \ge \delta\sqrt{n})\right)^{1/2}\right)\\ &\to 1 \quad \text{a.s.} \qquad (3.7)\end{aligned}$$

Note that

$$\frac{1}{np}tr(\hat{X}-\tilde{X})(\hat{X}-\tilde{X})^T$$

$$\leq \frac{2}{np}\sum_{ij}\Bigg[|x_{ij}^2|\max_{ij}|1-1/\sigma_{ij}|^2$$

$$+\mathbb{E}^2\{|x_{ij}|I(|x_{ij}|\geq\delta\sqrt{n})\}\Bigg]=O(n^{-6}) \text{ a.s.} \tag{3.8}$$

Here, the convergence rates follow from the facts that

(i) $\dfrac{1}{np}\sum_{ij}|x_{ij}^2|\to 1$ a.s.

(ii) $\max_{ij}|1-1/\sigma_{ij}|^2\leq\max_{ij}\ \mathbb{E}^2\{|x_{ij}|^2I(|x_{ij}|\geq\delta\sqrt{n})\}=O(\delta^{-12}n^{-6})$.

(iii) $\dfrac{1}{np}\sum_{ij}\ \mathbb{E}^2\{|x_{ij}|I(|x_{ij}|\geq\delta\sqrt{n})\}=O(\delta^6n^{-7})$.

It follows from (3.4)–(3.8) that under conditions (C.1)–(C.3),

$$\int|\hat{F}_p(x)-\tilde{F}_p(x)|dx=O(n^{-3}) \text{ a.s.} \tag{3.9}$$

Using Lemma 2.5 of Bai (1993b), the proofs of the three theorems reduce to show that the main theorem remains true when $\hat{F}_p$ is replaced with $\tilde{F}_p$. Note that the random variables $\tilde{x}_{ij}$ still satisfy the conditions (C.1)–(C.3). They also satisfy the additional condition

$$|x_{ij}|\leq\delta\sqrt{n}$$

(here, the constant δ should be 3δ if δ is the one we previously selected. For brevity, we still use δ). Also, for simplicity, we shall drop the tilde sign from various variables.

3.2. The Bai inequality. Suppose that G is a function of bounded variation. The Stieltjes transform g of G is defined as

$$g(z)=\int_{-\infty}^{\infty}\frac{1}{x-z}\,dG(x),$$

where $z=u+iv$ and $v>0$. Our main tool is the following inequality (Bai (1993a)).

PROPOSITION 3.1. *Let G be a distribution function and H be a function of bounded variation satisfying $\int|G(x)-H(x)|\,dx<\infty$. Denote their Stieltjes transforms by $g(z)$ and $h(z)$, respectively. Then*

$$\|G-H\|\leq\frac{1}{\pi(1-\kappa)(2\gamma-1)}\Bigg[\int_{-A}^{A}|g(z)-h(z)|\,du+\frac{2\pi}{v}\int_{|x|>B}|G(x)-H(x)|\,dx$$

$$+\frac{1}{v}\sup_x\int_{|y|\leq 2va_*}|H(x+y)-H(x)|\,dy\Bigg],$$

where the constants $A > B$, γ, and a_* are restricted by

$$\gamma = \frac{1}{\pi}\int_{|u|\le a_*} \frac{1}{u^2+1}du > \frac{1}{2}, \quad \textit{and} \quad \kappa = \frac{4B}{\pi(A-B)(2\gamma-1)} \in (0,1).$$

Denote the Stieltjes transform of F_p and F_y (recall our convention that $y = y_n = p/n$) by $m_p(z)$ and $m(z)$, respectively. Application of the Bai inequality with $(G,H) = (F_p, F_y)$, $A = 25$, and $B = 5$ gives, for some constant $c > 0$,

$$\|F_p - F_y\| \le c\left[\int_{-A}^{A} |m_p(z) - m(z)|\, du + \frac{1}{v}\int_{|x|>5} |F_p(x) - e(x)|\, dx \right.$$

$$\left. + \frac{1}{v}\sup_x \int_{|u|\le 2va_*} |F_y(x+u) - F_y(x)|\, du\right], \tag{3.10}$$

where $e(x) = 1$ for $x > 0$ or $e(x) = 0$ otherwise. We shall estimate these three terms in the above bound successively and start with the last one.

(a) Estimate for $\sup_x \int_{|u|\le 2va_*} |F_y(x+u) - F_y(x)|\, du$.

LEMMA 3.1. *We have, for any $0 < v < 4\sqrt{y}$,*

$$\sup_x \int_{|u|\le v} |F_y(x+u) - F_y(x)|\, du \le \frac{11\sqrt{2(1+y)}}{3\pi y} v v_y,$$

where $v_y = v/(\sqrt{a} + \sqrt{v})$ is defined as in (2.7).

Proof. It is enough to consider the part $0 \le u \le v$ in the integral only since the remaining part for $-v \le u \le 0$ can be handled in a similar way. Set $\Phi(\lambda) := \int_0^v [F_y(x+u) - F_y(x)]du$ with $x = a + \lambda$; we are estimating the maximum of $\Phi(\lambda)$. Without loss of generality, we need only consider the case that $\lambda \ge 0$ because $\int_0^v [F_y(x+u) - F_y(x)]du$ increases when $x \le a$. Then

$$\begin{aligned}\Phi(\lambda) &= \int_0^v du \int_x^{x+u} F_y'(t)\, dt \\ &= \int_{a+\lambda}^{a+\lambda+v} \frac{a+\lambda+v-t}{2\pi y t}\sqrt{(t-a)(b-t)}\, I_{[a,b]}(t)dt \\ &= \int_{\lambda}^{\lambda+v} \frac{\lambda+v-u}{2\pi y(u+a)}\sqrt{u(4\sqrt{y}-u)}\, I_{[0,b-a]}(u)\, du. \end{aligned} \tag{3.11}$$

Let $\phi(u) := (u+a)^{-1}\sqrt{u(4\sqrt{y}-u)}$. The derivative of $\log(\phi(u))^2$ is

$$\frac{1}{u} - \frac{1}{4\sqrt{y}-u} - \frac{2}{u+a} = \frac{2(2\sqrt{y}a - (1+y)u)}{u(4\sqrt{y}-u)(u+a)}.$$

Note that the above equality holds also for $y = 1$ for which $a = 0$. Let $\rho := (1+y)^{-1}(2a\sqrt{y})$. Thus $\phi(u)$ is decreasing when $u > \rho$ and increasing when $u < \rho$. Since

$$\frac{d\Phi(\lambda)}{d\lambda} = \frac{1}{2\pi y}\left(\int_\lambda^{\lambda+v} [\phi(u) - \phi(\lambda)]\, du\right),$$

it follows that for $\lambda > \rho$, $\Phi(\lambda)$ is decreasing and then $\Phi(\lambda) \le \Phi(\rho)$; for $\lambda < \rho - v$, $\Phi(\lambda)$ is increasing and then $\Phi(\lambda) \le \Phi(\rho - v)$. Hence, $\Phi(\lambda)$ reaches its maximum only for some $\lambda \in (\max(\rho - v, 0), \rho)$. Considering such a λ yields by (3.11)

$$\begin{aligned}\Phi(\lambda) &\le \frac{2y^{1/4}}{2\pi y}\int_{\lambda}^{\lambda+v}\frac{\lambda+v-u}{u+a}\sqrt{u}du \\ &= 2(\pi y^{3/4})^{-1}\left\{(\lambda+v+a)\left[(\sqrt{\lambda+v}-\sqrt{\lambda})\right.\right. \\ &\qquad \left.\left.-\sqrt{a}\left(\arctan\sqrt{\frac{\lambda+v}{a}}-\arctan\sqrt{\frac{\lambda}{a}}\right)\right]-\frac{1}{3}\left[(\lambda+v)^{3/2}-\lambda^{3/2}\right]\right\}.\end{aligned}$$

Since

$$\sqrt{a}\left(\arctan\sqrt{\frac{\lambda+v}{a}}-\arctan\sqrt{\frac{\lambda}{a}}\right) \ge \frac{a}{\lambda+v+a}\left(\sqrt{\lambda+v}-\sqrt{\lambda}\right),$$

we get, by setting $\lambda^* = \sqrt{\lambda+v}-\sqrt{\lambda}$,

$$\begin{aligned}\Phi(\lambda) &\le \frac{2}{\pi y^{3/4}}\left\{(a+\lambda+v)\left(\lambda^*-\frac{a}{a+\lambda+v}\lambda^*\right)-\lambda^*\left(\lambda+\sqrt{\lambda}\lambda^*+\frac{1}{3}\lambda^{*2}\right)\right\} \\ &= \frac{2}{\pi y^{3/4}}\left[\sqrt{\lambda}\lambda^{*2}+\frac{2}{3}\lambda^{*3}\right]. \qquad (3.12)\end{aligned}$$

Let $c^2 = \frac{1+y}{2\sqrt{y}}$. Since $\lambda + v \ge c^{-2}a$ and

$$(\sqrt{\lambda+v}+\sqrt{\lambda})^2 \ge \lambda+v+2\sqrt{\lambda(\lambda+v)} \ge 2\sqrt{\lambda v}+2\sqrt{\lambda c^{-2}a},$$

we have

$$\begin{aligned}\frac{\sqrt{\lambda}}{(\sqrt{\lambda+v}+\sqrt{\lambda})^2} &\le \frac{c}{2\sqrt{a}+2c\sqrt{v}} \le \frac{c}{2\sqrt{a}+2\sqrt{v}}, \\ \frac{1}{(\sqrt{\lambda+v}+\sqrt{\lambda})^3} &\le \frac{2c}{(\sqrt{a}+\sqrt{v})v},\end{aligned}$$

where the last inequality follows from

$$\begin{aligned}(\sqrt{\lambda+v}+\sqrt{\lambda})^3 &\ge \sqrt{\lambda+v}v \\ &\ge \frac{1}{2}\left[\sqrt{c^{-2}a}+\sqrt{v}v\right] \\ &\ge \frac{1}{2c}\left[\sqrt{v}+\sqrt{a}\right]v.\end{aligned}$$

Hence

$$\Phi(\lambda) \le \frac{2}{\pi y^{3/4}}\cdot\frac{11c}{6(\sqrt{a}+\sqrt{v})}v^2 = \frac{11\sqrt{2(1+y)}}{6\pi y}\frac{1}{\sqrt{v}+(1-\sqrt{y})}v^2.$$

This completes the proof of the lemma. □

(b) Estimate for $\frac{1}{v}\int_{|x|>5}|F_p(x)-e(x)|\,dx$. Let λ_p denote the largest eigenvalue of $\mathbf{S}$. By Yin, Bai, and Krishnaiah (1988), for any positive constant (ε) and integer (ℓ_n) such that $\ell_n/\log n\to\infty$ and $\ell_n\delta_n^{1/4}/\log n\to 0$, we have

$$E(\lambda_p)^{\ell_n}\le c(b+\varepsilon)^{\ell_n}. \tag{3.13}$$

Therefore, for $x\ge 5$ and any fixed $t>0$,

$$P(\lambda_p>x)\le c\left(\frac{b+\varepsilon}{x}\right)^{\ell_n}\le c\left(\frac{b+\varepsilon}{x}\right)^{2}\left(\frac{b+\varepsilon}{5}\right)^{\ell_n-2}=o(x^{-2}n^{-t}). \tag{3.14}$$

Since $F_p(x)=e(x)=0$ for $x\le 0$, we have

$$\begin{aligned}\int_{|x|>5}|F_p(x)-e(x)|\,dx&=\int_5^\infty[1-F_p(x)]\,dx\\ &=\int_5^\infty\frac{1}{p}\sum_{k=1}^{p}P(\lambda_k>x)\,dx\le\int_5^\infty P(\lambda_p>x)\,dx\\ &\le\int_5^\infty o(x^{-2}n^{-t})\,dx=o(n^{-t})\ . \end{aligned}\tag{3.15}$$

By (3.15) we finally get

$$\int_{|x|>5}|F_p(x)-e(x)|\,dx=O(n^{-2})\ \text{a.s.}$$

Thus, for $v>cn^{-1}$, we have

$$v^{-1}\int_{|x|>5}|F_p(x)-e(x)|\,dx=O_{a.s.}(n^{-1})\ \text{a.s.}$$

(c) Conclusion. Summarizing previous steps gives

$$\|F_p-F_y\|\le c\left[\int_{-A}^{A}|m_p(z)-m(z)|\,du+O_{a.s.}(n^{-1})+v_y\right]. \tag{3.16}$$

To prove the main theorems, we need only estimate $|m_p(z)-m(z)|$. Bai (1993b) has proved that $\Delta=\Delta_{n,y}=\|\,\mathbb{E}F_p-F_y\|=O(n^{-5/48})$. In what follows, we shall treat Δ as at least of the order $O(n^{-5/48})$.

3.3. Proof of Theorem 1.1. We begin by estimating $|\,\mathbb{E}m_p(z)-m(z)|$, with various choices of v, subject to $cn^{-1/2}\le v\le 1$ for some $c>0$. With the formula of m_p given in (2.10), let us define δ_p such that

$$m_p(z)=-\frac{1}{z+y-1+yz\mathbb{E}m_p(z)}+\delta_p=-\frac{1}{\mathbb{E}\beta}+\delta_p. \tag{3.17}$$

Since

$$\frac{1}{\beta_k}=\frac{1}{\mathbb{E}\beta}\left(1-\frac{\varepsilon_k}{\beta_k}\right),$$

it is easy to see that

$$\delta_p = \frac{1}{p}\sum_{k=1}^{p}\frac{1}{\mathbb{E}\beta}\frac{\varepsilon_k}{\beta_k} = \frac{1}{(\mathbb{E}\beta)^2}\left(\frac{1}{p}\sum_{k=1}^{p}\varepsilon_k - \frac{1}{p}\sum_{k=1}^{p}\frac{\varepsilon_k^2}{\beta_k}\right).$$

Now

$$\begin{aligned}
&|\mathbb{E}\delta_p| \\
&\le \frac{1}{p|\mathbb{E}\beta|^2}\sum_{k=1}^{p}\left(|\mathbb{E}\varepsilon_k| + \left|\mathbb{E}\frac{\varepsilon_k^2}{\beta_k}\right|\right) \\
&= \frac{1}{p|\mathbb{E}\beta|^2}\sum_{k=1}^{p}\left[|\mathbb{E}(\varepsilon_k^* + \tilde{\varepsilon}_k) + \pi_k| + \left|\frac{1}{\mathbb{E}\beta}\,\mathbb{E}\varepsilon_k^2 - \frac{1}{(\mathbb{E}\beta)^2}\,\mathbb{E}\varepsilon_k^3 + \frac{1}{(\mathbb{E}\beta)^2}\,\mathbb{E}\left(\frac{\varepsilon_k^4}{\beta_k}\right)\right|\right] \\
&\le \frac{1}{p|\mathbb{E}\beta|^2}\left[\sum_{k=1}^{p}|\mathbb{E}(\varepsilon_k^* + \tilde{\varepsilon}_k) + \pi_k| + \sum_{k=1}^{p}\left|\frac{1}{\mathbb{E}\beta}\,\mathbb{E}\varepsilon_k^2\right| + \sum_{k=1}^{p}\left|\frac{1}{(\mathbb{E}\beta)^2}\,\mathbb{E}\varepsilon_k^3\right|\right. \\
&\quad \left. + \sum_{k=1}^{p}\left|\frac{1}{(\mathbb{E}\beta)^2}\,\mathbb{E}\left(\frac{\varepsilon_k^4}{\beta_k}\right)\right|\right] \\
&= |\mathbb{E}\beta|^{-2}\left[I_0 + I_1 + I_2 + I_3\right].
\end{aligned}$$

We will estimate each I_i to obtain a bound on $|\mathbb{E}\delta_p|$ (cf. (3.19) below). Since $\mathbb{E}(\varepsilon_k^* + \tilde{\varepsilon}_k) = 0$, by (2.9), we have

$$I_0 = \frac{|z|}{p}\sum_{k=1}^{p}|\pi_k| = \frac{|z|}{pn}\sum_{k=1}^{p}|\mathbb{E}tr\mathbf{D}_k - \mathbb{E}tr\mathbf{D}| \le |z|/(nv).$$

From Lemma 4.2, Remark 4.1, and noticing that $v \le v_y$, we have

$$\begin{aligned}
I_1 &\le \frac{1}{p|\mathbb{E}\beta|}\sum_{k=1}^{p}\mathbb{E}|\varepsilon_k|^2 = \frac{1}{p|\mathbb{E}\beta|}\sum_{k=1}^{p}(\mathbb{E}|\varepsilon_k^*|^2 + \mathbb{E}|\tilde{\varepsilon}_k|^2 + |\pi_k|^2) \\
&\le \frac{c}{|\mathbb{E}\beta|}\left(\left[\frac{1}{n} + \frac{\Delta + v_y}{nv^2}\right] + \frac{\Delta + v_y}{n^2v^4} + \frac{1}{n^2v^2}\right) \le \frac{c(\Delta + v_y)}{|\mathbb{E}\beta|nv^2}, \\
I_2 &= \frac{1}{p|\mathbb{E}\beta|^2}\sum_{k=1}^{p}|\mathbb{E}\varepsilon_k^3| \le \sum_{k=1}^{p}\left(\frac{1}{p|\mathbb{E}\beta|}\,\mathbb{E}|\varepsilon_k|^2 + \frac{1}{p|\mathbb{E}\beta|^3}\,\mathbb{E}|\varepsilon_k|^4\right).
\end{aligned}$$

Now

$$\frac{1}{p}\sum_{k=1}^{p}\mathbb{E}|\varepsilon_k|^4 \le \frac{27}{p}\sum_{k=1}^{p}(\mathbb{E}|\varepsilon_k^*|^4 + \mathbb{E}|\tilde{\varepsilon}_k|^4 + |\pi_k|^4) \triangleq c(I_{21} + I_{22} + I_{23}).$$

Since

$$tr\mathbf{B}_k\overline{\mathbf{B}}_k = tr(\mathbf{I}_{p-1} + z\mathbf{D}_k)(\mathbf{I}_{p-1} + \overline{z}\overline{\mathbf{D}}_k) \le 2(p + |z|^2 tr\mathbf{D}_k\overline{\mathbf{D}}_k),$$

We have from the proof of Lemma 4.1,

$$\begin{aligned}
\mathbb{E}|\varepsilon_k^*|^4 &\le cn^{-2}\left\{1 + n^{-2}\,\mathbb{E}(tr\mathbf{B}_k\overline{\mathbf{B}}_k)^2\right\} \\
&\le cn^{-2}\left\{1 + n^{-2}\,\mathbb{E}(tr\mathbf{D}_k\overline{\mathbf{D}}_k)^2\right\}.
\end{aligned}$$

Now

$$\begin{aligned}\mathbb{E}(tr(\mathbf{D}_k\overline{\mathbf{D}_k}))^2 &= v^{-2}\,\mathbb{E}(\mathrm{Im}(tr(\mathbf{D}_k)))^2\\ &\le 2v^{-2}[v^{-2} + \mathbb{E}(\mathrm{Im}(tr(\mathbf{D})))^2]\\ &= 2v^{-4} + 2p^2v^{-2}\,\mathbb{E}(\mathrm{Im}(m_p(z)))^2\\ &\le 2v^{-4} + 4p^2v^{-2}|\,\mathbb{E}m_p(z)|^2 + 4p^2v^{-2}\,\mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|^2\\ &\le cp^2v^{-4}(\Delta+v_y)^2 + cv^{-6}(\Delta+v_y) \le cp^2v^{-4}(\Delta+v_y)^2,\end{aligned}$$

where the second inequality follows from (2.9) and the last steps follow from Proposition 4.1 and

$$|\,\mathbb{E}m_p(z)| \le |\,\mathbb{E}m_p(z) - m(z)| + |m(z)| \le v^{-1}(2\Delta + \alpha_y v_y), \tag{3.18}$$

with $\alpha_y := 2\sqrt{2}/\sqrt{y}$ (see (2.6)). Thus

$$\begin{aligned}I_{21} &\le c\left\{n^{-2} + n^{-2}v^{-4}(\Delta+v_y)^2\right\}\\ &\le cn^{-2}v^{-4}(\Delta+v_y)^2.\end{aligned}$$

Also, considering $\mathbf{D}_k$ instead of $\mathbf{D}$ as in Proposition 4.1 and applying (2.8), one can show that for some L_0 such that for all $L_0n^{-1/2} \le v < 1$,

$$I_{22} \le c(\Delta+v_y)^2n^{-4}v^{-8}.$$

Since $|\pi_k| \le |z|(nv)^{-1}$, we have $I_{23} \le |z|^4(nv)^{-4}$, and hence,

$$\begin{aligned}p^{-1}\sum_{k=1}^{p}\mathbb{E}|\varepsilon_k|^4 &\le c(I_{21}+I_{22}+I_{23})\\ &\le c[n^{-2}v^{-4}(\Delta+v_y)^2 + (\Delta+v_y)^2n^{-4}v^{-8} + (nv)^{-4}]\\ &\le cn^{-2}v^{-4}(\Delta+v_y)^2.\end{aligned}$$

Consequently, for some constant $c > 0$,

$$I_2 \le \frac{c(\Delta+v_y)}{|\,\mathbb{E}\beta|nv^2} + \frac{c(\Delta+v_y)^2}{|\,\mathbb{E}\beta|^3n^2v^4}$$

and

$$I_3 \le \frac{1}{pv|\,\mathbb{E}\beta|^2}\sum_{k=1}^{p}\mathbb{E}|\varepsilon_k|^4 \le \frac{c}{n^2v^5|\,\mathbb{E}\beta|^2}(\Delta+v_y)^2.$$

Summing up the above results, we obtain

$$\begin{aligned}|\,\mathbb{E}\delta_p| &\le \frac{1}{|\,\mathbb{E}\beta|^2}[I_0+I_1+I_2+I_3]\\ &\le \frac{c}{|\,\mathbb{E}\beta|^2}\left[\frac{1}{nv} + \frac{\Delta+v_y}{nv^2|\,\mathbb{E}\beta|} + \frac{(\Delta+v_y)^2}{n^2v^5|\,\mathbb{E}\beta|^2}\right]\\ &\le \frac{c}{|\,\mathbb{E}\beta|^2}\left[\frac{\Delta+v_y}{nv^2|\,\mathbb{E}\beta|} + \frac{(\Delta+v_y)^2}{n^2v^5|\,\mathbb{E}\beta|^2}\right].\end{aligned} \tag{3.19}$$

For the positive constant L_0 (required by Proposition 4.1) and all $v \in [L_0 n^{-1/2}, 1)$, define

$$\varphi_n(v) = \sup_{|u|\le A} |\,\mathbb{E}\delta_p| - \gamma v \quad \text{with} \quad \gamma = 1/[10(A+1)^2].$$

Checking the proofs of (3.39)–(3.40) of Bai (1993b), we find that there is a constant c such that

$$\varphi_n(v) \le 0 \Longrightarrow \int_{-A}^{A} |\,\mathbb{E}m_p(z) - m(z)|\,du < cv.$$

In view of (3.16), we can then find a positive constant c_1 such that

$$\varphi_n(v) \le 0 \Longrightarrow \Delta < c_1 v_y. \tag{3.20}$$

The proof of the theorem will be complete once we have shown that for all large n and all $v \in [Ln^{-1/[4\theta+2]}, 1)$, we have $\varphi_n(v) \le 0$, where L is a constant such that $L \ge L_0$ and

$$cM_0^2\left[\frac{(c_1+1)M_0}{L^2} + \frac{(1+c_1)^2M_0^2}{L^4}\right] < \gamma$$

and $M_0 = \gamma + 2c_1 + \alpha_y$.

Assume the contrary; i.e., there exists a $v_1 \in [Ln^{-1/[4\theta+2]}, 1)$ for which $\varphi_n(v_1) > 0$. By continuity of φ_n, there exists a $v_0 \in [Ln^{-1/[4\theta+2]}, 1)$ for which $\varphi_n(v_0) = 0$. As $[-A, A]$ is compact, there exists a $u_0 \in [-A, A]$ such that $0 = \varphi_n(v_0) = |\,\mathbb{E}\delta_p(u_0, v_0)| - \gamma v_0$. Let $z_0 = u_0 + iv_0$. By (3.17), (3.18), and Lemma 2.2, with $z = z_0$,

(3.21)
$$\begin{aligned}\frac{1}{|\,\mathbb{E}\beta|} &= |-\,\mathbb{E}\delta_p(z_0) + \,\mathbb{E}[m_p(z_0) - m(z_0)] + m(z_0)| \le |\,\mathbb{E}\delta_p(z_0)| + \frac{2\Delta + \alpha_y v_{y,0}}{v_0}\\ &\le (\gamma + 2c_1 + \alpha_y)\frac{v_{y,0}}{v_0} = M_0\frac{v_{y,0}}{v_0}.\end{aligned}$$

On the other hand, by definition (1.2) of θ, we have

$$\frac{v}{v_y} = \sqrt{a} + \sqrt{v} \ge \sqrt{a} \vee \sqrt{v} \ge n^{-\frac{\theta}{4\theta+2}}\ .$$

Therefore, for any $v \in [Ln^{-1/[4\theta+2]}, 1)$ we have

$$\frac{1}{nv^2}\left(\frac{v_y}{v}\right)^4 \le 1/L^2, \qquad \frac{1}{n^2v^4}\left(\frac{v_y}{v}\right)^6 \le 1/L^4.$$

Thus, from (3.19) we have for $z = z_0$ (so $v = v_0$ and $v_y = v_{y,0}$)

$$\begin{aligned}|\,\mathbb{E}\delta_p(z_0)| &\le \frac{cM_0^2v_y^2}{v^2}\left[\frac{(c_1+1)M_0v_y^2}{nv^3} + \frac{(1+c_1)^2M_0^2v_y^4}{n^2v^7}\right]\\ &\le vcM_0^2\left[\frac{(c_1+1)M_0}{L^2} + \frac{(1+c_1)^2M_0^2}{L^4}\right] < \gamma v\end{aligned}$$

by noticing the selection of L.

This leads to a contradiction of $\varphi_n(z_0) = 0$. The proof of Theorem 1.1 is complete.

3.4. Proof of Theorem 1.2. As the proof of (3.16), one can show that

$$\begin{aligned}
&\mathbb{E}||F_p - F_y|| \\
&\le c\left[\int_{-A}^{A} \mathbb{E}|m_p(z) - m(z)|du + v_y\right] \\
&\le c\left[\int_{-A}^{A} \mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|du + \int_{-A}^{A} |\mathbb{E}m_p(z) - m(z)|du + v_y\right].
\end{aligned}$$

In the proof of Theorem 1.1, we have shown that $\int_{-A}^{A} |\mathbb{E}m_p(z) - m(z)|du = O(v)$ if $Ln^{-1/[2+4\theta]} < v < 1$.

Applying the Cauchy–Schwarz inequality, Remark 4.1, and the result $\Delta = O(n^{-1/[4\theta+2]})$ proved in Theorem 1.1, we conclude that

$$\begin{aligned}
\int_{-A}^{A} \mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|du &\le \int_{-A}^{A} (\mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|^2)^{1/2}du \\
&\le cn^{-1/2}v_y^{1/2}v^{-2} \le v
\end{aligned}$$

for some positive constant c and all $cn^{-2/[5+\theta]} \le v < 1$. Recall that we need a condition of $v > Ln^{-1/[4\theta+2]}$ to guarantee $\int_{-A}^{A} |\mathbb{E}m_p(z) - m(z)|du = O(v)$. The convergence rate we can guarantee is

$$O_p\left(\max\left\{\frac{n^{-(2/(5+\theta))}}{[1-\sqrt{y}+n^{-(1/(5+\theta))}]}, \frac{n^{-1/[4\theta+2]}}{[1-\sqrt{y}+n^{-1/[8\theta+4]}]}\right\}\right).$$

The proof of Theorem 1.2 in this case is complete. □

3.5. Proof of Theorem 1.3. Similarly, we have

$$\|F_p - F_y\| \le c\left[\int_{-A}^{A} |m_p(z) - \mathbb{E}m_p(z)|du + v_y\right]. \tag{3.22}$$

Thus, to complete the proof of Theorem 1.3, setting $v = n^{-2/[5+\theta]+\eta}$ it suffices to show that

$$v^{-1}\int_{-A}^{A} |m_p(z) - \mathbb{E}m_p(z)|\,du \to 0 \qquad \text{a.s.} \tag{3.23}$$

Now, applying Proposition 4.1, we obtain for each $\xi > 0$,

$$\begin{aligned}
&P\left(\int_{-A}^{A} |m_p(z) - \mathbb{E}m_p(z)|\,du \ge \xi v\right) \\
&\le (v\xi)^{-2k}(2A)^{2k-1}\int_{-A}^{A} \mathbb{E}\,|m_p(z) - \mathbb{E}m_p(z)|^{2k}\,du \\
&\le \xi^{-2k}(2A)^{2k}\left[c_k\left(n^{-2}v^{-6}v_y\right)^k\right] \\
&\le c_k'(\varepsilon\xi)^{-2k}n^{-(5+\theta)\eta k}.
\end{aligned}$$

The right-hand side of the above inequality is summable by choosing k such that $5\eta k > 1$. Recalling the condition used in the proof of Theorem 1.1, the convergence rate is

$$O_{a.s.}\left(\max\left\{\frac{n^{-(2/(5+\theta))+\eta}}{[1-\sqrt{y}+n^{-(1/(5+\theta))}]}, \frac{n^{-1/[4\theta+2]}}{[1-\sqrt{y}+n^{-1/[8\theta+4]}]}\right\}\right)$$

Thus, (3.23) is proved and the proof of Theorem 1.3 is complete. □

4. Intermediate lemmas. In this section, we establish a few more technical lemmas. Let $\nu_\ell = \sup_{i,j,n}\{E|x_{ij}|^\ell\}$.

LEMMA 4.1. *For each $\ell \geq 1$ with $\nu_{4\ell} < \infty$, there exist positive constants c_ℓ independent of n and v such that for all n, v satisfying $nv \geq T$, we have*

$$\mathbb{E}\left(|\varepsilon_k^*|^{2\ell}\big|\,\mathcal{F}^{(k)}\right) \leq c_\ell(1+\lambda_p)^{\ell/2}n^{-\ell}\left(1+\frac{1}{n}tr\mathbf{\Lambda}_k\right)^\ell \tag{4.1}$$

and

$$\mathbb{E}\left(\frac{(\varepsilon_k^*)^{2\ell}}{|\beta_k^*|^\ell}\,\middle|\,\mathcal{F}^{(k)}\right) \leq c_\ell(1+\lambda_p)^{\ell/2}n^{-\ell}v^{-\ell}. \tag{4.2}$$

Proof. We have

$$\begin{aligned}\mathbb{E}\left(|\varepsilon_k^*|^{2\ell}|\mathcal{F}^{(k)}\right) &= \mathbb{E}\left(\left|-\frac{1}{n}\sum_{j=1}^n(x_{kj}^2-1)+\frac{1}{n}(\mathbf{x}_k'\mathbf{B}_k\mathbf{x}_k-tr\mathbf{B}_k)\right|^{2\ell}\,\middle|\,\mathcal{F}^{(k)}\right)\\ &\leq 2^{2\ell-1}n^{-2\ell}\left\{\mathbb{E}\left|\sum_{j=1}^n(x_{kj}^2-1)\right|^{2\ell}+\mathbb{E}\left(|\mathbf{x}_k'\mathbf{B}_k\mathbf{x}_k-tr\mathbf{B}_k|^{2\ell}\big|\,\mathcal{F}^{(k)}\right)\right\}\\ &:= A+B.\end{aligned}$$

For the first term A, by the Burkholder inequality (Burkholder (1973, p. 22)), we get

$$\begin{aligned}&\mathbb{E}\left|\sum_{j=1}^n(x_{kj}^2-1)\right|^{2\ell}\\ &\leq c_\ell\,\mathbb{E}\left[\sum_{j=1}^n(x_{kj}^2-1)^2\right]^\ell \leq c_\ell n^{\ell-1}\,\mathbb{E}\left[\sum_{j=1}^n(x_{kj}^2-1)^{2\ell}\right] \leq c_\ell\nu_{4\ell}n^\ell.\end{aligned}$$

For the second term B, denoting the eigenvalues of $\mathbf{S}_k$ by λ_{kj} and noticing that their maximum is less than the largest eigenvalue λ_p of $\mathbf{S}$, we then have

$$\begin{aligned}&tr\left(\mathbf{B}_k\overline{\mathbf{B}}_k\right) = tr\mathbf{B}_k+\overline{z}tr\mathbf{\Lambda}_k,\\ &=\sum_{j=1}^{p-1}\frac{\lambda_{kj}}{\lambda_{kj}-z}+\overline{z}\sum_{j=1}^{p-1}\frac{\lambda_{kj}}{|\lambda_{kj}-z|^2} \leq \lambda_p^{1/2}\sum_{j=1}^{p-1}\frac{\lambda_{kj}^{1/2}}{|\lambda_{kj}-z|}+T\sum_{j=1}^{p-1}\frac{\lambda_{kj}}{|\lambda_{kj}-z|^2}\\ &\leq(\lambda_p^{1/2}+T)\,(p-1+tr\mathbf{\Lambda}_k)\,.\end{aligned}$$

Therefore by Lemma 2.1,

$$E\left(|\mathbf{x}_k'\mathbf{B}_k\mathbf{x}_k - tr\mathbf{B}_k|^{2\ell}\big|\,\mathcal{F}^{(k)}\right)$$
$$\le c_\ell(\nu_{4\ell}+M^\ell)(tr\mathbf{B}_k\overline{\mathbf{B}}_k)^\ell \le c_\ell(\lambda_p^{1/2}+T)^\ell n^\ell\left(1+\frac{1}{n}tr\mathbf{\Lambda}_k\right)^\ell .$$

Combining the bounds for A and B proves the first conclusion. The second conclusion immediately follows by taking into account inequality (2.12). □

LEMMA 4.2. *If $n^{-1/2}\le v<1$, then there are positive constants C_1, C_2 such that for large n and each $1\le k\le p$,*

$$\text{(i)}\quad |\mathbb{E}tr(\mathbf{D}_k\overline{\mathbf{D}}_k)| \le C_1 p\frac{\Delta+v_y}{v^2}.$$
$$\text{(ii)}\quad \mathbb{E}|\varepsilon_k^*|^2 \le C_2\frac{1}{n}\left(1+|z|^2\frac{\Delta+v_y}{v^2}\right).$$

Proof. (i) Recall that $\Delta = \|\,\mathbb{E}F_p - F_y\|$. By Lemma 2.2,

$$\left|\int_{-\infty}^{\infty}\frac{1}{|x-z|^2}d(\,\mathbb{E}F_p(x)-F_y(x))\right| \le \frac{2\Delta}{v^2}.$$

Application of Lemma 2.1 and inequality (2.8) yields that

$$\begin{aligned}|\,\mathbb{E}tr(\mathbf{D}_k\overline{\mathbf{D}}_k)| &= \left|(p-1)\int_{-\infty}^{\infty}\frac{1}{|x-z|^2}d[\,\mathbb{E}F_p^{(k)}(x)]\right| \\ &\le \left|\int_{-\infty}^{\infty}\frac{1}{|x-z|^2}d[(p-1)\,\mathbb{E}F_p^{(k)}(x)-p\,\mathbb{E}F_p(x)]\right| \\ &\quad + p\left|\int_{-\infty}^{\infty}\frac{1}{|x-z|^2}d[\,\mathbb{E}F_p(x)-F_y(x)]\right| + p\left|\int_{-\infty}^{\infty}\frac{1}{|x-z|^2}dF_y(x)\right| \\ &\le \frac{2}{v^2}+p\frac{2\Delta}{v^2}+p\left|\int_{-\infty}^{\infty}\frac{1}{|x-z|^2}dF_y(x)\right|.\end{aligned}$$

Here, the bound of the last term follows from Lemma 2.4. The proof of conclusion (i) is complete.

(ii) This conclusion follows from (i), (4.1), and the fact

$$tr\mathbf{B}_k\overline{\mathbf{B}}_k = tr(\mathbf{I}_{p-1}+z\mathbf{D}_k)(\mathbf{I}_{p-1}+\overline{z}\overline{\mathbf{D}}_k) \le 2(p+|z|^2 tr\mathbf{D}_k\overline{\mathbf{D}}_k).$$ □

LEMMA 4.3. *Assume $|z|\le T$ with $T\ge 2$. Then there are constants C_0, C_1 such that for all $v\ge C_0 n^{-1/2}$ and large n, we have*

$$\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}) \le C_1 n(\Delta+v_y)v^{-1}\,. \tag{4.3}$$

Proof. From the definition of ε_k^*, we notice that $(\beta_k^*)^{-1} = \beta_k^{-1}(1+\beta_k^{-1}\varepsilon_k^*)$. By (2.9),

$$|\beta_k^*-\beta| = \frac{1}{n}|-1+z(tr\mathbf{D}_k - tr\mathbf{D})| \le \frac{1}{n}\left(1+\frac{|z|}{v}\right) \le \frac{2T}{nv}.$$

By (3.13), it is easy to see that for any fixed $t > 0$ and all large n,

$$\mathbb{E}\left[(1+\lambda_p)^{\ell}|W|\right] \le 6^{\ell}\,\mathbb{E}|W| + o(\|W\|n^{-t}), \tag{4.4}$$

where W is a bounded random variable with a nonrandom bound $\|W\|$. By this and taking into account (2.10), (4.2), and (2.6), we obtain

$$\begin{aligned}
&\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1})\\
&\le \sum_{k=1}^{p}\mathbb{E}\left|\frac{1}{|\beta_k^*|}-\frac{1}{|\beta|}\right| + \mathbb{E}\left|\sum_{k=1}^{p}\left(\frac{1}{\beta}-\frac{1}{\beta_k^*}\right)\right| + \mathbb{E}\left|\sum_{k=1}^{p}\left(\frac{1}{\beta_k^*}-\frac{1}{\beta_k}\right)\right| + \mathbb{E}\left|\sum_{k=1}^{p}\beta_k^{-1}\right|\\
&\le 2\sum_{k=1}^{p}\mathbb{E}\frac{|\beta_k^*-\beta|}{|\beta||\beta_k^*|} + \sum_{k=1}^{p}\mathbb{E}\frac{|\varepsilon_k^*|}{|\beta_k^*|^2} + \sum_{k=1}^{p}\mathbb{E}\frac{|\varepsilon_k^*|^2}{|\beta_k||\beta_k^*|^2} + p\,\mathbb{E}|m_p(z)|\\
&\le \frac{c}{nv^2}\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}) + \sum_{k=1}^{p}\mathbb{E}\frac{(\,\mathbb{E}(|\varepsilon_k^*|^2|\mathcal{F}^{(k)}))^{1/2}}{|\beta_k^*|^2} + \sum_{k=1}^{p}\mathbb{E}\frac{\mathbb{E}(|\varepsilon_k^*|^2|\mathcal{F}^{(k)})}{v|\beta_k^*|^2} + p\,\mathbb{E}|m_p(z)|\\
&\le c_*\left(\frac{1}{nv^2}+n^{-1/2}v^{-1}+\frac{1}{nv^2}\right)\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}) + p\,\mathbb{E}|m_p(z)| + o(n^{-t})\\
&\le 3c_*(nv^2)^{-1/2}\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}) + p\,\mathbb{E}|m_p(z)| + o(n^{-t})
\end{aligned}$$

for all $v \ge 1/\sqrt{n}$. Let $C_0 = 1 \vee (6c_*)^2$. We have for all $v \ge \sqrt{C_0/n}$

$$\begin{aligned}
\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}) &\le \frac{1}{2}\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}) + p\,\mathbb{E}|m_p(z)| + o(n^{-t})\\
&\le 2p\,\mathbb{E}|m_p(z)| + o(n^{-t})\\
&\le 2p\,\mathbb{E}|m_p(z) - \mathbb{E}(m_p(z))| + 2p|\,\mathbb{E}(m_p(z)) - m(z)| + 2p|m(z)| + o(n^{-t})\\
&\le 2\sqrt{p}v^{-1} + 4p\Delta/v + 2pv_y/v.
\end{aligned}$$

The proof is now complete. □

LEMMA 4.4. *Let* $z_k = \mathbb{E}(tr\mathbf{D}|\mathcal{F}_{k-1}) - \mathbb{E}(tr\mathbf{D}|\mathcal{F}_k)$. *Then* $tr\mathbf{D} - \mathbb{E}tr\mathbf{D} = \sum_{k=1}^{p} z_k$ *and* (z_k) *is a martingale difference with respect to* $(\mathcal{F}_k)$, $k = p, p-1, \ldots, 0$. *Moreover, we have the following formula for* z_k*:*

$$z_k = \{\,\mathbb{E}\,(a_k|\mathcal{F}_{k-1}) - \,\mathbb{E}\,(a_k|\mathcal{F}_k)\} - \,\mathbb{E}\,(b_k|\mathcal{F}_{k-1})\,,$$

with

$$a_k = \frac{\varepsilon_k^*(1+\alpha_k^T\mathbf{D}_k{}^2\alpha_k)}{\beta_k^*\beta_k}, \quad b_k = \frac{\alpha_k^T\mathbf{D}_k{}^2\alpha_k - \frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]}{\beta_k^*}. \tag{4.5}$$

Proof. Since $\mathbb{E}(tr\mathbf{D}_k|\mathcal{F}_{k-1}) = \mathbb{E}(tr\mathbf{D}_k|\mathcal{F}_k)$, we have

$$z_k = \mathbb{E}[(tr\mathbf{D} - tr\mathbf{D}_k)|\mathcal{F}_{k-1}] - \mathbb{E}[(tr\mathbf{D} - tr\mathbf{D}_k)|\mathcal{F}_k].$$

On the other hand,

$$
\begin{aligned}
&tr\mathbf{D} - tr\mathbf{D}_k = -\frac{1+\frac{1}{n}\alpha_k^T\mathbf{D}_k{}^2\alpha_k}{\beta_k} \\
&= -\frac{1+\frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]}{\beta_k^*} + \frac{\varepsilon_k^*(1+\alpha_k^T\mathbf{D}_k{}^2\alpha_k)}{\beta_k^*\beta_k} - \frac{\alpha_k^T\mathbf{D}_k{}^2\alpha_k - \frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]}{\beta_k^*} \\
&= -\frac{1+\frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]}{\beta_k^*} + a_k - b_k.
\end{aligned}
$$

The conclusion follows from

$$
\mathbb{E}\left(\frac{1+\frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]}{\beta_k^*}\middle|\mathcal{F}_{k-1}\right) = \mathbb{E}\left(\frac{1+\frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]}{\beta_k^*}\middle|\mathcal{F}_k\right)
$$

and

$$
\mathbb{E}\left(\alpha_k^T\mathbf{D}_k{}^2\alpha_k\middle|\mathcal{F}^{(k)}\right) = \frac{1}{n}tr[(\mathbf{I}+z\mathbf{D}_k)\mathbf{D}_k]. \qquad \square
$$

PROPOSITION 4.1. *For each $\ell > 1/2$ with $\nu_{4\ell} < \infty$, there exist positive constants c_ℓ and L_0 independent of n and v such that for all n, v satisfying $L_0 n^{-1/2} \le v < 1$,*

$$
\mathbb{E}|m_p(z) - Em_p(z)|^{2\ell} \le c_\ell n^{-2\ell} v^{-4\ell}(\Delta + v_y)^\ell.
$$

Proof. In the proof of the proposition, c_ℓ and $c_{\ell,0}$ will be used to denote universal positive constants which may depend on the moments up to order ℓ of underlying variables and may represent different values at different appearances, even in one expression. Recall that we have

$$
m_p(z) - \mathbb{E}m_p(z) = \frac{1}{p}\left[tr\mathbf{D} - \mathbb{E}tr\mathbf{D}\right] = \sum_{k=1}^{p} z_k,
$$

where the $\{z_k\}$ are defined as in Lemma 4.4. We have

$$
\begin{aligned}
\mathbb{E}\left(|z_k|^{2\ell}\middle|\mathcal{F}_k\right) &= \mathbb{E}\left\{\left|\left[\mathbb{E}\left(a_k|\mathcal{F}_{k-1}\right) - \mathbb{E}\left(a_k|\mathcal{F}_k\right)\right] - \mathbb{E}\left(b_k|\mathcal{F}_{k-1}\right)\right|^{2\ell}\middle|\mathcal{F}_k\right\} \\
&\le 2^{2\ell-1}\,\mathbb{E}\left\{\left[\mathbb{E}\left(a_k|\mathcal{F}_{k-1}\right) - \mathbb{E}\left(a_k|\mathcal{F}_k\right)\right]^{2\ell} + \left[\mathbb{E}\left(b_k|\mathcal{F}_{k-1}\right)\right]^{2\ell}\middle|\mathcal{F}_k\right\} \\
&\le 2^{2\ell-1}\,\mathbb{E}\left\{\left[\mathbb{E}\left(a_k|\mathcal{F}_{k-1}\right)\right]^{2\ell} + \left[\mathbb{E}\left(b_k|\mathcal{F}_{k-1}\right)\right]^{2\ell}\middle|\mathcal{F}_k\right\} \\
&\le 2^{2\ell-1}\left\{\mathbb{E}\left((a_k)^{2\ell}\middle|\mathcal{F}_k\right) + \mathbb{E}\left((b_k)^{2\ell}\middle|\mathcal{F}_k\right)\right\}.
\end{aligned}
$$

Note that by (2.13) and (2.14), $|a_k| \le v^{-1}\left|\varepsilon_k^*/\beta_k^*\right|$. Hence by Lemma 4.1,

$$
\mathbb{E}\left(|a_k|^{2\ell}\middle|\mathcal{F}^{(k)}\right) \le \frac{1}{v^{2\ell}}\,\mathbb{E}\left(\left|\frac{\varepsilon_k^*}{\beta_k^*}\right|^{2\ell}\middle|\mathcal{F}^{(k)}\right) \le c_{\ell,0}(1+\lambda_p)^{\ell/2}n^{-\ell}v^{-3\ell}|\beta_k^*|^{-\ell}.
$$

On the other hand, by Lemma 2.1 and assuming $\ell \ge 1$,

$$
\mathbb{E}\left(|b_k|^{2\ell}\middle|\mathcal{F}^{(k)}\right) \le c_{\ell,0}(n\beta_k^*)^{-2\ell}(\nu_{4\ell} + M^\ell)\left[tr(\mathbf{I}+z\mathbf{D}_k)(\mathbf{I}+\overline{z}\overline{\mathbf{D}}_k)\mathbf{D}_k\overline{\mathbf{D}}_k\right]^\ell.
$$

By (2.12) and (2.16),

$$|\beta_k^*|^{-1} tr(\mathbf{I}+z\mathbf{D}_k)(\mathbf{I}+\overline{z}\overline{\mathbf{D}}_k)\mathbf{D}_k\overline{\mathbf{D}}_k \le |\beta_k^*|^{-1} tr\mathbf{\Lambda}_k^2 \le nTv^{-3},$$

which implies

$$\mathbb{E}\left\{ |b_k|^{2\ell} \middle| \mathcal{F}_k \right\} \le c_{\ell,0} n^{-\ell} v^{-3\ell}\, \mathbb{E}\left[|\beta_k^*|^{-\ell} \middle| \mathcal{F}_k \right].$$

Therefore, for all $\ell \ge 1$,

$$\begin{aligned} \mathbb{E}\left(|z_k|^{2\ell} \middle| \mathcal{F}_k \right) &\le c_{\ell,0}(1+\lambda_p)^{\ell/2} n^{-\ell} v^{-3\ell}\, \mathbb{E}\left[|\beta_k^*|^{-\ell} \middle| \mathcal{F}_k \right] \\ &\le c_{\ell,0}(1+\lambda_p)^{\ell/2} n^{-\ell} v^{-4\ell+1}\, \mathbb{E}\left[|\beta_k^*|^{-1} \middle| \mathcal{F}_k \right]. \end{aligned} \tag{4.6}$$

Applying Lemma 4.3 and (4.4), it follows that, for $\ell \ge 1$,

$$\sum_{k=1}^{p} \mathbb{E}|z_k|^{2\ell} \le c_{\ell,0} n^{-\ell+1}(\Delta + v_y) v^{-4\ell}. \tag{4.7}$$

Case $\ell = 1$. Since $\{z_k\}$ is a martingale difference sequence, the above inequality yields

$$\mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|^2 = n^{-2} \sum_{k=1}^{p} \mathbb{E}|z_k|^2 \le c_{1,0} n^{-2}(\Delta + v_y) v^{-4}. \tag{4.8}$$

The proposition is proved in this case.

Case $\frac{1}{2} < \ell < 1$. By applying the Burkholder inequality for the martingale and using the concavity of the function x^ℓ, we find

$$\begin{aligned} &\mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|^{2\ell} \\ &\le c_\ell p^{-2\ell}\ \mathbb{E}\left(\sum_{k=1}^{p} |z_k|^2 \right)^{\ell} \le c_\ell n^{-2\ell}\ \left[\mathbb{E}\left(\sum_{k=1}^{p} |z_k|^2 \right) \right]^{\ell} \le c_\ell n^{-2\ell}\ \left[(\Delta + v_y) v^{-4} \right]^{\ell}, \end{aligned}$$

where the last step follows from the previous case $\ell = 1$. The lemma is then proved in this case.

Case $\ell > 1$. We proceed by induction in this general case. First, by another Burkholder inequality for the martingale (Burkholder (1973), p. 39), we have

$$\begin{aligned} \mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|^{2\ell} &\le c_\ell p^{-2\ell} \left\{ \sum_{k=1}^{p} \mathbb{E}|z_k|^{2\ell} + \mathbb{E}\left(\sum_{k=1}^{p} \mathbb{E}(|z_k|^2 | \mathcal{F}_k) \right)^{\ell} \right\} \\ &\hat{=} I_1 + I_2. \end{aligned} \tag{4.9}$$

By (4.7)

$$I_1 \le c_{\ell,0}(\Delta + v_y) n^{-3\ell+1} v^{-4\ell} \le c_{\ell,0}(\Delta + v_y)^{\ell} n^{-2\ell} v^{-4\ell}. \tag{4.10}$$

The proposition already has been proved for the case $\frac{1}{2} < \ell \le 1$. Suppose that the lemma is true for $\ell \le 2^t$. Now, we consider the case where $2^t < \ell \le 2^{t+1}$. Application of (4.6) with $\ell = 1$ gives

$$\sum_{k=1}^{n} \mathbb{E}\left(|z_k|^2 \middle| \mathcal{F}_k \right) \le c_{1,0} n^{-1} v^{-3} (1+\lambda_p)^{1/2} \sum_{k=1}^{p} \mathbb{E}\left(|\beta_k^*|^{-1} \middle| \mathcal{F}_k \right).$$

Hence, by taking into account (4.4) we get

$$I_2 \le c_{\ell,0}(nv)^{-3\ell}\,\mathbb{E}(1+\lambda_p)^{\ell/2}\left(\sum_{k=1}^{p}\mathbb{E}(|\beta_k^*|^{-1}|\mathcal{F}_k)\right)^{\ell}$$

$$\le c_{\ell,0}n^{-2\ell-1}v^{-3\ell}\sum_{k=1}^{p}\mathbb{E}|\beta_k^*|^{-\ell}+o(n^{-4\ell-1}v^{-4\ell}), \tag{4.11}$$

since $P(\lambda_p>5)=o(n^{-4\ell})$. Notice that if $L_0>\sqrt{2}$, then $nv^2>2$ and that

$$\left||\beta|^{-1}-|\beta_k^*|^{-1}\right| \le |\beta^{-1}-(\beta_k^*)^{-1}| = \frac{|tr\mathbf{D}-tr\mathbf{D}_k|}{p|\beta||\beta_k^*|} \le \frac{1}{pv^2}\min(|\beta|^{-1},|\beta_k^*|^{-1})$$

(this comes from (2.9) and $|\beta\beta_k^*|^{-1}\le v^{-1}\min(|\beta|^{-1},|\beta_k^*|^{-1})$). This yields

$$|\beta_k^*|^{-1}\le|\beta|^{-1}+p^{-1}v^{-2}|\beta_k^*|^{-1}\le 2|\beta|^{-1}$$

and

$$|p\beta^{-1}| \le \left|\sum_{k=1}^{p}(\beta_k^*)^{-1}\right|+\sum_{k=1}^{p}|(\beta_k^*)^{-1}-\beta^{-1}| \le \left|\sum_{k=1}^{p}(\beta_k^*)^{-1}\right|+v^{-2}|\beta|^{-1}$$

$$\le 2\left|\sum_{k=1}^{p}(\beta_k^*)^{-1}\right| \le 2\left|\sum_{k=1}^{p}((\beta_k^*)^{-1}-\beta_k^{-1})\right|+2\left|\sum_{k=1}^{p}\beta_k^{-1}\right|$$

$$\le 2\sum_{k=1}^{p}\frac{|\varepsilon_k^*|^2}{|\beta_k||\beta_k^*|^2}+2p|m_p(z)|.$$

Therefore, by applying Lemma 4.1, and if we choose, $L_0>(2c_{\ell,0})^{1/\ell}$ so that $c_{\ell,0}n^{-\ell}v^{-2\ell}<1/2$, we have

$$\sum_{k=1}^{p}\mathbb{E}|\beta_k^*|^{-\ell} \le c_{\ell,0}\left(v^{-\ell}\sum_{k=1}^{p}\mathbb{E}\frac{|\varepsilon_k^*|^{2\ell}}{|\beta_k^*|^{2\ell}}+p\,\mathbb{E}|m_p(z)|^{\ell}\right)$$

$$\le c_{\ell,0}\left(n^{-\ell}v^{-2\ell}\sum_{k=1}^{p}\mathbb{E}|\beta_k^*|^{-\ell}+p\,\mathbb{E}|m_p(z)|^{\ell}\right)$$

$$\le 2c_{\ell,0}p\,\mathbb{E}|m_p(z)|^{\ell}.$$

From the above inequality and (4.11), we get by induction,

$$I_2 \le c_\ell n^{-2\ell}v^{-3\ell}\,\mathbb{E}|m_p(z)|^{\ell}+o(n^{-4\ell-1}v^{-4\ell})$$

$$\le c_\ell n^{-2\ell}v^{-3\ell}\left[\,\mathbb{E}|m_p(z)-\,\mathbb{E}m_p(z)|^{\ell}+|\,\mathbb{E}m_p(z)-m(z)|^{\ell}+|m(z)|^{\ell}\right]$$

$$+o(n^{-4\ell-1}v^{-4\ell})$$

$$\le c_\ell n^{-2\ell}v^{-3\ell}\left[\,\mathbb{E}|m_p(z)-\,\mathbb{E}m_p(z)|^{\ell}+(\Delta+v_y)^{\ell}v^{-\ell}\right]+o(n^{-4\ell-1}v^{-4\ell})$$

$$\le c_\ell n^{-2\ell}v^{-4\ell}(\Delta+v_y)^{\ell}\left[\left(n^2v^2(\Delta+v_y)\right)^{-\ell/2}+1\right]+o(n^{-4\ell-1}v^{-4\ell})$$

$$\le c_\ell n^{-2\ell}v^{-4\ell}(\Delta+v_y)^{\ell}. \tag{4.12}$$

Therefore by (4.9) and (4.12), it follows that

$$\mathbb{E}|m_p(z) - \mathbb{E}m_p(z)|^{2\ell} \le c_\ell n^{-2\ell}(\Delta + v_y)^\ell v^{-4\ell}. \tag{4.13}$$

The proof of Proposition 4.1 is complete. □

Remark 4.1. Application of Proposition 4.1 to the case $\ell = 1$ gives that there is some constant $c_1 > 0$ such that

$$\mathbb{E}|tr\mathbf{D}_k - \mathbb{E}tr\mathbf{D}_k|^2 \le c_1(\Delta + v_y)v^{-4}. \tag{4.14}$$

It is also worth noticing that if we substitute $\mathbf{D}$ for any $\mathbf{D}_k$ with $k \le n$, Proposition 4.1 as well as the above consequence (4.14) are still valid, with slightly different constants c_ℓ.

Acknowledgments. The authors warmly thank the referees for their comments that actually lead to several important improvements. J. F. Yao thanks the National University of Singapore for the support during his visit to its Department of Statistics and Applied Probability.

REFERENCES

Z. D. Bai (1993a), *Convergence rate of expected spectral distributions of large random matrices, Part* I: *Wigner matrices*, Ann. Probab., 21, pp. 625–648.
Z. D. Bai (1993b), *Convergence rate of expected spectral distributions of large random matrices, Part* II: *Sample covariance matrices*, Ann. Probab., 21, pp. 649–672.
Z. D. Bai (1997), *Circular law*, Ann. Probab., 25, pp. 494–529.
Z. D. Bai (1999), *Methodologies in spectral analysis of large dimensional random matrices. A review*, Statist. Sinica, 9, pp. 611–677.
Z. D. Bai, B. Miao, and J. Tsay (1997), *A note on the convergence rate of the spectral distribution of large random matrices*, Statist. Probab. Lett., 34, pp. 95–101.
Z. D. Bai, B. Q. Miao, and J. Tsay (1999), *Remarks on the convergence rate of the spectral distributions of Wigner matrices*, J. Theoret. Probab., 12, pp. 301–311.
Z. D. Bai, B. Miao, and J. Tsay (2002), *Convergence rates of the spectral distributions of large Wigner matrices*, Int. Math. J., 1, pp. 65–90.
Z. D. Bai and J. W. Silverstein (1998), *No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices*, Ann. Probab., 26, pp. 316–345.
Z. D. Bai and J. W. Silverstein (1999), *Exact separation of eigenvalues of large dimensional sample covariance matrices*, Ann. Probab., 27, pp. 1536–1555.
Z. D. Bai and Y. Q. Yin (1986), *Limiting behavior of the norm of products of random matrices and two problems of Geman-Hwang*, Probab. Theory Related Fields, 73, pp. 555–569.
Z. D. Bai, Y. Q. Yin, and P. K. Krishnaiah (1987), *On the limiting empirical distribution function of the eigenvalues of a multivariate F-matrix*, Probab. Theory Appl., 32, pp. 490–500.
D. L. Burkholder (1973), *Distribution function inequalities for martingales*, Ann. Probab., 1, pp. 19–42.
V. A. Marčenko and L. A. Pastur (1967), *Distribution of eigenvalues for some sets of random matrices*, Mat. Sb., 72, pp. 507–536.
M. L. Mehta (1991), *Random Matrices*, Academic Press, New York.
J. W. Silverstein (1985), *The limiting eigenvalue distribution of a multivariate F matrix*, SIAM J. Math. Anal., 16, pp. 641–646.
E. P. Wigner (1955), *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. of Math. (2), 62, pp. 548–564.
E. P. Wigner (1958), *On the distribution of the roots of certain symmetric matrices*, Ann. of Math. (2), 67, pp. 325–327.
Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah (1988), *On the limit of the largest eigenvalue of the large dimensional sample covariance matrix*, Probab. Theory Related Fields, 78, pp. 509–531.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 128–142

A MATRIX ANALYSIS APPROACH TO HIGHER-ORDER APPROXIMATIONS FOR DIVERGENCE AND GRADIENTS SATISFYING A GLOBAL CONSERVATION LAW*

J. E. CASTILLO† AND R. D. GRONE†

Abstract. One-dimensional, second-order finite-difference approximations of the derivative are constructed which satisfy a global conservation law. Creating a second-order approximation away from the boundary is simple, but obtaining appropriate behavior near the boundary is difficult, even in one dimension on a uniform grid. In this article we exhibit techniques that allow the construction of discrete versions of the divergence and gradient operator that have high-order approximations at the boundary. We construct such discretizations in the one-dimensional situation which have fourth-order approximation both on the boundary and in the interior. The precision of the high-order mimetic schemes in this article is as high as possible at the boundary points (with respect to the bandwidth parameter). This guarantees an overall high order of accuracy. Furthermore, the method described for the calculation of the approximations uses matrix analysis to streamline the various mimetic conditions. This contributes to a marked clarity with respect to earlier approaches.

This is a crucial preliminary step in creating higher-order approximations of the divergence and gradient for nonuniform grids in higher dimensions.

Key words. mimetic finite difference, high order, divergence, gradient

AMS subject classifications. 65D25, 65G99, 65M06

PII. S0895479801398025

1. Introduction.

In this article we show how to construct higher-order discrete approximations to derivatives in uniform staggered one-dimensional grids that exactly satisfy a global conservation law.

The underlying central problem is to find higher-order approximations of the divergence ($\nabla\cdot$) and gradient (grad) that satisfy a discrete analogue of the divergence theorem:

$$\int_\Omega \nabla\cdot\vec{v}\, f\, dV + \int_\Omega \vec{v}\ \mathrm{grad}\, f\, dV = \int_{\partial\Omega} f\, \vec{v}\cdot\vec{n}\, dS\,. \tag{1}$$

For these two operators, there are three closely related ideas: the *divergence theorem* (1), *local conservation*, and *global conservation*. Local conservation is a special case of the divergence theorem (1) with $f = 1$ and Ω taken to be a single cell, while global conservation is (1) with $f = 1$ applied to the full region under consideration. We refer to discretizations which possess properties analogous to these as *mimetic*. As either the divergence theorem or local conservation implies global conservation, we expect it to be easiest to find discretizations which satisfy analogues of global conservation.

In the one-dimensional setting and with the interval being $[0, 1]$, (1) is simply integration by parts:

$$\int_0^1 \frac{dv}{dx}\, f\, dx + \int_0^1 v\, \frac{df}{dx}\, dx = v(1)\, f(1) - v(0)\, f(0)\,. \tag{2}$$

*Received by the editors September 15, 2001; accepted for publication (in revised form) by G. H. Golub November 8, 2002; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/39802.html
†Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182-7720 (castillo@myth.sdsu.edu, grone@math.sdsu.edu).

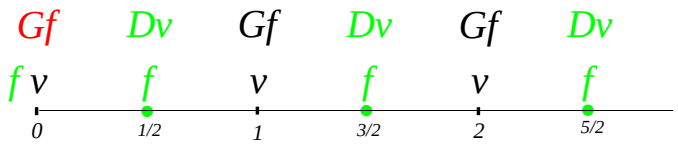


FIG. 1. *The grid.*

In some applications, such as hyperbolic conservation laws, a gradient is not needed, but a conservative divergence is, where conservative means that the integration-by-parts formula (1) holds with $f \equiv 1$ or that in one dimension (2) holds with $f \equiv 1$; that is,

$$\int_0^1 \frac{dv}{dx} dx = v(1) - v(0)\,. \tag{3}$$

In this paper we focus on the one-dimensional case and use techniques from matrix analysis to show how to construct higher-order approximations to d/dx that satisfy the global conservation law.

For our discretizations we will use the uniform finite-volume or support-operators grid shown in Figure 1, constructed by choosing $N > 0$, which is the number of cells, and then setting $h = 1/N$. The *nodes* or *points* in the grid are $x_i = i\,h$, $0 \le i \le N$. The *cells* are given by the interval $[i\,h, (i+1)\,h]$ with centers $x_{i+\frac{1}{2}} = (i+\frac{1}{2})\,h$, $0 \le i \le N-1$.

The discrete divergence will act on the v-values, while the discrete gradient will act on the f-values, as illustrated in Figure 1.

2. Second-order discrete operators. The simplest discrete divergence is defined by

$$(\mathbf{D}\,v)_{i+\frac{1}{2}} = \frac{v_{i+1} - v_i}{h}\,, \quad 0 \le i \le N-1\,, \tag{4}$$

and the discrete gradient is defined by

$$\begin{aligned} (\mathbf{G}\,f)_0 &= \frac{f_{\frac{1}{2}} - f_0}{h/2}\,, \\ (\mathbf{G}\,f)_i &= \frac{f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}}{h}\,, \quad 1 \le i \le N-1\,, \\ (\mathbf{G}\,f)_N &= \frac{f_N - f_{N-\frac{1}{2}}}{h/2}\,, \end{aligned} \tag{5}$$

where again the definition of $\mathbf{G}$ at the boundary points is standard for the support-operators approach. This divergence $\mathbf{D}$ is second-order accurate, while the gradient $\mathbf{G}$ is second-order accurate in the interior and first-order accurate at the boundary. Figure 1 illustrates the positions of the values of $\mathbf{D}\,v$ and $\mathbf{G}\,f$ in the grid.

To compare various works in this area it is important to realize that we are using a staggered grid, while [2, 12, 13, 14] use a nodal grid and approximate the derivative by

$$\left(\frac{df}{dx}\right)_i = \frac{f_{i+1} - f_{i-1}}{2\,h}\,, \quad 1 \le i \le N-1\,. \tag{6}$$

Also, the authors of [2, 12, 13, 14], along with those of [5, 6, 7], agree that, in either setting, the main difficulty in generating good higher-order schemes is in getting the appropriate behavior at the boundary (or endpoints, in the one-dimensional case).

The mimetic schemes constructed with this approach will preserve fundamental properties of the original continuous operators, allowing the discrete approximations of partial differential equations to mimic critical properties, including conservation laws and symmetries, in the solution of the underlying physical problems [4, 5, 6, 9, 10, 11]. In addition, mimetic schemes have been used with great success on Maxwell's first-order equations [8]. It is important to notice that the techniques presented here depart completely from the ones used in [4, 5, 6], where the construction of the discrete operators relies on heavy use of computer algebra. The method that we describe here for the calculation of the approximations uses matrix analysis to its advantage, streamlining the various conditions that they must satisfy and herewith contributing to a marked clarity with respect to earlier methods.

3. Mimetic discretization schemes. When using the uniform grid described in Figure 1, the function v becomes an $(N+1)$-tuple, and the divergence operator becomes an N-by-$(N+1)$ matrix, $\mathbf{D}$. Let $e = (1, 1, \ldots, 1)^t$ be the n-tuple of appropriate size. Since the divergence of $v = 1$ is zero, we wish the matrix $\mathbf{D}$ to satisfy the analogous property

$$\mathbf{D}\, e = 0\,, \tag{7}$$

where e is an $(N+1)$-tuple. This condition can also be expressed by saying that the row sums of $\mathbf{D}$ are $0, \ldots, 0$.

It is convenient to express the global conservation law in terms of inner products as

$$\langle \nabla \cdot\ v, 1\rangle = v(1) - v(0)\,, \tag{8}$$

which has the obvious discrete analogue

$$\langle \mathbf{D}\ v, e\rangle = v_N - v_0\,, \tag{9}$$

where e is an N-tuple. This condition can also be expressed by the statement that $\mathbf{D}$ has column sums equal to $-1, 0, \ldots, 0, 1$, which is equivalent to

$$e^t\mathbf{D}\ = (-1, 0, \ldots, 0, 1). \tag{10}$$

In addition to the mimetic conditions (7) and (9) or (10) there are some other natural conditions that $\mathbf{D}$ can be expected to possess due to the geometry of the situation. Also, the requirement that $\mathbf{D}$ be a higher-order approximation leads to further conditions on $\mathbf{D}$. It should be expected that our matrix $\mathbf{D}$ be sparse in order for these methods to be of interest. The values of the approximation at any node should be determined by its nearest neighbors, so $\mathbf{D}$ should be a banded matrix. The discretization scheme on the interior nodes can be expected to be similar. Hence the interior rows should exhibit a Toeplitz-type structure wherein the nonzero entries in row $i+1$ are just the nonzero entries of row i shifted one space to the right. The banded and Toeplitz-type properties play an important role in our discretization schemes. If the bandwidth of $\mathbf{D}$ is b, our techniques allow $\mathbf{D}$ to be described as independent of N, as long as $N \geq 3b - 1$. Another structural property of $\mathbf{D}$ is motivated as follows.

Suppose a function w is defined by $w(x) = v(1-x),\ 0 \le x \le 1$. Then the divergence of w is the negative of the divergence of v. In algebraic terms this imposes the following symmetry condition on $\mathbf{D}$. Let P_n denote the permutation matrix:

$$P_n = \begin{bmatrix} 0 & 0 & \cdots & & 0 & 0 & 1 \\ 0 & 0 & \cdots & & 0 & 1 & 0 \\ 0 & 0 & \cdots & & 1 & 0 & 0 \\ \vdots & & 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & \cdots & & 0 & 0 \\ 1 & 0 & 0 & \cdots & & 0 & 0 \end{bmatrix} \tag{11}$$

If $\mathbf{D}$ is N-by-$(N+1)$, then the mimetic version of $\nabla \cdot w = -\nabla \cdot v$ is

$$P_N \mathbf{D}\, P_{N+1} = -\mathbf{D}\,. \tag{12}$$

We will refer to a matrix which satisfies (12) as *centro-skew-symmetric* [1].

We can summarize the desired properties of our divergence matrix $\mathbf{D}$ as follows:

- $\mathbf{D}$ has zero row sums. Equivalently, $\mathbf{D}\, e = 0$, where $e = (1, 1, \ldots, 1)^t$.
- $\mathbf{D}$ has column sums $-1, 0, \ldots, 0, 1$. Equivalently, $e^t \mathbf{D} = (-1, 0, 0, \ldots, 0, 1)$.
- $\mathbf{D}$ is banded. (We let b denote the bandwidth of $\mathbf{D}$.)
- $\mathbf{D}$ has a "Toeplitz"-type structure on the interior rows and is defined independently of N, the number of grid points.
- $\mathbf{D}$ is centro-skew-symmetric.

The problem of finding a mimetic discrete version of the gradient is equivalent to that of finding a mimetic discrete version of the divergence, at least if the standard inner product is employed. Let $\mathbf{D}$ be an N-by-$(N+1)$ discrete version of the divergence; $\hat{\mathbf{D}}$ be the matrix $\mathbf{D}$ augmented with two rows, first and last, of zeroes; and $\mathbf{G}$ be an $(N+1)$-by-$(N+2)$ discrete version of the gradient. If v is an $(N+1)$-tuple, f is the N-tuple, $(f_{\frac{1}{2}}, f_{\frac{3}{2}}, \ldots, f_{N-\frac{1}{2}})$, and $\hat{f}$ is f with the endpoints f_0 and f_N added to its ends so that $\hat{f}$ is an $(N+2)$-tuple, the fundamental equation (2) is equivalent to

$$\langle \hat{\mathbf{D}}\, v, f \rangle + \langle v, \mathbf{G}\, \hat{f} \rangle = v_N f_N - v_0 f_0 \tag{13}$$

or

$$\langle \hat{\mathbf{D}}\, v, f \rangle + \langle \mathbf{G}^{\,t} v, \hat{f} \rangle = \langle \mathbf{B} v, \hat{f} \rangle\,, \tag{14}$$

where $\mathbf{B}$ is the $(N+2)$-by-$(N+1)$ matrix

$$\mathbf{B} = \begin{bmatrix} -1 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & & & & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}. \tag{15}$$

This, of course, implies that $\hat{\mathbf{D}} + \mathbf{G}^t = \mathbf{B}$, which yields that $\mathbf{G}$ must have the form

$$\mathbf{G} = \begin{bmatrix} -1 & & 0 \\ 0 & & 0 \\ \vdots & -\mathbf{D}^t & \vdots \\ 0 & & 0 \\ 0 & & 1 \end{bmatrix}. \tag{16}$$

Loosely speaking, except for some allowance for boundary behavior, $\mathbf{G}$ is the negative adjoint of $\mathbf{D}$, since (2) leads to (13) as a discrete analogue and hence to (14), which gives the relation between $\mathbf{G}$ and $\mathbf{D}$. In this article we will focus on the problem of constructing mimetic divergences, although this has obvious implications for the problem of constructing mimetic gradients.

4. Generalized inner products. So far we have described our problem in terms of inner product formulas with respect to the usual Euclidean inner product, $\langle u, v\rangle = v^t u$. We will find that it is not possible to satisfy all of the desired properties in addition to the "higher-order" conditions to be discussed in section 5. However, if we allow a generalized or weighted inner product as do the authors in [12, 13, 14], then solutions are possible. The most general form of an inner product on N-tuples is

$$\langle u, v\rangle = \langle Qu, v\rangle = v^t Q u, \tag{17}$$

where Q is N-by-N and positive definite. Alternatively, this can also be expressed as

$$\langle u, v\rangle = \langle Eu, Ev\rangle = v^t E^t E u, \tag{18}$$

where E is any rank N matrix which satisfies $E^t E = Q$. We refer to such an inner product defined by either E or Q as a *generalized inner product*. In case the matrix Q (or E) is diagonal, we refer to the corresponding inner product as a *weighted inner product*. If we are using a generalized inner product, then equations such as (9) or (10) become

$$\langle \mathbf{D}\, v, e\rangle_Q = v_N - v_0 \tag{19}$$

or

$$e^t Q \mathbf{D} = (-1, 0, \ldots, 0, 1). \tag{20}$$

We will show that our objectives of higher-order mimetic divergence cannot be attained with respect to the standard inner product but can be attained with a generalized inner product. In fact, the matrix Q will turn out to be diagonal, centrosymmetric, and independent of N.

5. High-order divergences. Let $\mathbf{D}$ be an N-by-$(N+1)$ matrix representing a discretized divergence as discussed in sections 1 and 2. We say that $\mathbf{D}$ is a *kth-order approximation* if $\mathbf{D}$ is exact for polynomials of order up to k but not $k+1$. There is a natural *stencil matrix*, $\mathbf{S}$, for pursuing kth-order approximations. For example, if $k = 4$, then the bandwidth of $\mathbf{S}$ equals 4 as well, and the interior rows of $\mathbf{S}$ have the form

$$\frac{1}{24}[0, \ldots, 0, 1, -27, 27, -1, 0, \ldots, 0]. \tag{21}$$

This canonical stencil is obtained by using Lagrange polynomials (see [6]).

Suppose $N = 7$ and $k = 4$. By using the stencil given by (21) and imposing the column sum conditions on $\mathbf{S}$, we immediately arrive at the stencil matrix

$$\mathbf{S} = \begin{bmatrix} -\frac{25}{24} & \frac{13}{12} & -\frac{1}{24} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{24} & -\frac{13}{12} & \frac{25}{24} \end{bmatrix}. \tag{22}$$

This matrix has most of our required properties. The row and column sums are as desired. The matrix is banded, with bandwidth $b = 4$, and it is centro-skew-symmetric. The matrix can be described independently of N for $N \geq 5$. The matrix satisfies fourth-order approximation conditions on the interior but not on the boundary. In the next section we employ matrix analysis techniques to modify the matrix $\mathbf{S}$ so that it also possesses fourth-order approximation on the boundary. The techniques are general in that they apply to any desired even order of approximation. Some of the matrices involved in the procedures are interesting in their own right.

For any k there is such a canonical N-by-$(N+1)$ matrix $\mathbf{S}$ which has bandwidth $b = k$, is a discrete version of our divergence operator, and gives kth order of approximation on the interior points of the interval but only low order at the endpoints. Our aim here will be to modify $\mathbf{S}$ in order to attain uniform kth-order approximation. We will generically denote such a modified version of $\mathbf{S}$ as $\mathbf{D}$ since $\mathbf{D}$ will be the desired divergence matrix. Since the behavior of $\mathbf{S}$ is optimal (for the bandwidth) on the interior nodes, we will assume that we will modify only the first and last t rows of $\mathbf{S}$ to obtain $\mathbf{D}$. For our techniques to be generally useful, we require that t be a function of b only and not of N. We will see that $t = b$ will be sufficient for our purposes. We will focus our attention on modifying the first t rows of $\mathbf{S}$, since the centro-skew-symmetric condition on $\mathbf{D}$ will determine the last t rows of $\mathbf{D}$ as well.

We will consider modifications to $\mathbf{S}$ of the following type. Suppose t and l are fixed and A is a t-by-l matrix. Let $\mathbf{D}(A)$ denote the matrix obtained by replacing the first t rows of $\mathbf{S}$ with t-by-$(N+1)$ matrix

$$[A \quad 0]. \tag{23}$$

As with t, we wish l to be "small" relative to b and independent of N. Our techniques require A to be k-by-$\frac{3}{2}k$. That is, $t = k, l = \frac{3}{2}k$, and k is both the order of approximation and the bandwidth of $\mathbf{S}$ on its interior rows. Throughout we assume $b = k$. We also assume that $N \geq 3b - 1$. This is to ensure that the upper left and lower right modified portions of $\mathbf{S}$ do not share common columns, which would lead to difficulties with the column sum condition. Perhaps the simplest way in which to effect such a modification of $\mathbf{S}$ is to multiply $\mathbf{S}$ on the left by a matrix of the form

$$Q = \Lambda \oplus I_{N-2t} \oplus \Lambda', \tag{24}$$

where Λ is t-by-t positive definite and $\Lambda' = P_t \Lambda P_t$ in order to preserve the centro-skew-symmetry of $\mathbf{S}$. This may be viewed as a particular case of using a generalized inner product while leaving $\mathbf{S}$ alone. A special case is when Λ is restricted to be a diagonal (positive definite) matrix. We note that this approach cannot succeed.

PROPOSITION 1. *There is no mimetic divergence matrix which is kth order for $k \geq 2$ when Q is diagonal with respect to the standard inner product.*

Proof. Note that the N-by-$(N+1)$ stencil matrix $\mathbf{S}$ satisfies the mimetic conditions

1. $\mathbf{S}e = 0$,
2. $\text{rank}(\mathbf{S}) = N$,
3. $\mathbf{S}$ is centro-skew-symmetric,
4. $e^t\mathbf{S} = (-1, 0, \ldots, 0, 1)$.

Let Q denote an arbitrary N-by-N diagonal matrix. We claim that it is impossible that both $\mathbf{S}$ and $Q\mathbf{S}$ satisfy the fourth condition unless $Q = I_N$. We have already noted that $\mathbf{S}$ does not afford kth order on the boundary for $k \geq 2$. Now, if $e^t\mathbf{S} = e^tQ\mathbf{S}$, then $e^t(I_N - Q)\mathbf{S} = 0$. Since $\text{rank}(\mathbf{S}) = N$, $e^t(I_N - Q) = 0$. Now because $I_N - Q$ is diagonal, this forces $I_N - Q = 0$. ∎

Our characterization of the order conditions on $\mathbf{D}$ and/or $\mathbf{S}$ will also make it obvious that $\mathbf{S}$ cannot have higher-order approximation on the endpoints with respect to the standard inner product. It should be noted that the weights are used only in the generalized inner product and are never part of the calculations. This proposition supports the findings, for nodal grids, of Kreiss and Scherer in [13]. The more general $\mathbf{D}(A)$ approach will nearly succeed. More precisely, when k is even, it will be possible to find k-by-$\frac{3}{2}k$ A such that $\mathbf{D}(A)$ satisfies the relevant mimetic and order conditions with respect to a weighted inner product defined by a matrix of the form Q in (24). We will see that Λ will be k-by-k diagonal in this case and that Λ will be independent of N. Our techniques will produce both A and Λ by computationally efficient techniques revolving around Gaussian elimination. Our method to obtain the above-described results, as well as to compute the desired matrices A and Λ, is to formulate our problem in matrix terms. In general, let $\mathbf{D}(A)$ be obtained by replacing the first t rows of S with the matrix (23), where A is t-by-l. Let a be the tl-tuple $a = [row_1(A), row_2(A), \ldots, row_t(A)]^t$. The conditions we wish $\mathbf{D}(A)$ to satisfy are the row sum conditions, the column sum conditions, and the order conditions (up to order k). These are linear conditions, and we need the following notation to express them in matrix terms. For a positive integer m and $x_1, \ldots, x_n \in \mathrm{R}$, let $V(m; x_1, \ldots, x_n)$ be the $(m+1)$-by-n *Vandermonde matrix*

$$V(m; x_1, , \ldots, x_n) = \begin{bmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ \vdots & & \vdots \\ x_1^m & \ldots & x_n^m \end{bmatrix}. \tag{25}$$

The mimetic and order conditions on $\mathbf{D}(A)$ may be expressed in the form

$$\mathbf{M}a = r, \tag{26}$$

where $\mathbf{M}$ is the $(tk + t + l)$-by-tl matrix

$$\mathbf{M} = \begin{bmatrix} V_1 & 0 & \ldots & 0 \\ 0 & V_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & V_t \\ I_l & I_l & \ldots & I_l \end{bmatrix} \tag{27}$$

and $V_1, \ldots, V_t$ are the following Vandermonde matrices:

$$\begin{aligned} V_1 &= V(k; 1, -1, -3, \ldots, -1-2k), \\ V_2 &= V(k; 3, 1, -1, \ldots, 1-2k), \\ V_3 &= V(k; 5, 3, 1, \ldots, 3-2k), \\ &\vdots\ \vdots \\ V_t &= V(k; 2t-1, \ldots, 2t-2k-3). \end{aligned} \tag{28}$$

Here the first row of V_i corresponds to the row sum condition on row i of $\mathbf{D}(A)$, and the last k rows of V_i correspond to the order conditions (up to k) on row i of A. The entries in the Vandermonde matrices are determined by the Taylor expansions of the functions used in our approximation. The last l rows of $\mathbf{M}$, or simply the last block row of $\mathbf{M}$, corresponds to the column sum conditions on $\mathbf{D}(A)$. The $(tk+t+l)$-tuple r in (26) is as follows. Let $c = (0, -2, 0, \ldots, 0) \in R^{k+1}$. Then let

$$r = \begin{bmatrix} c \\ c \\ \vdots \\ c \\ d \end{bmatrix}, \tag{29}$$

where $d \in R^l$ is chosen so that the column sums of $\mathbf{D}(A)$ are $-1, 0, \ldots, 0, 1$, as desired. Our problem now focuses on the matrix equation $\mathbf{M}a = r$. It is important to note that the condition $N \geq 2l - 1$ or $N \geq 3b - 1$ ensures that the nonzero portions of A and A' in $\mathbf{D}(A)$ do not overlap so that we may consider A independently of A'.

To illustrate our techniques we will describe the situation for $k = 4$ and $N \geq 11$. We will show that no hope of solution exists if $t \leq 3$ and that for $t = 4$ there exist fourth-order divergence matrices $\mathbf{D}(A)$ when $l \geq 6$. This $\mathbf{D} = \mathbf{D}(A)$ will satisfy the mimetic and kth-order conditions with respect to a weighted inner product defined by a positive definite diagonal matrix Q which is defined independently of N. These results can be generalized to any even integer k in the following way. We may use our technique to find a k-by-$\frac{3}{2}k$ matrix A for which $\mathbf{D}(A)$ satisfies the mimetic and kth-order conditions with respect to a positive definite diagonal matrix Q as defined above. As in the case $k = 4$, Q depends only on k and is independent of N. Our techniques provide a family of solutions for A as well as for the diagonal weighting matrix Q.

6. Fourth-order divergence. We are interested in approximations to the divergence which are of higher order than our previous example. To modify the matrix $\mathbf{S}$ in (22) to obtain higher-order approximation on the boundary we need only modify

the upper left and lower right corners of **S**. Let A be a k-by-l matrix and let

$$\mathbf{D} = \mathbf{D}(A) = \begin{bmatrix} A & & & & & & & 0 & 0 & 0 & 0 & 0 \\ & & & & & \ddots & & & & & & \\ 0 & \cdots & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & \cdots & 0 \\ \vdots & & & & & & & & & \ddots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & & & & & & & A' \end{bmatrix}, \tag{30}$$

where $A' = P_k A P_l$ and the intermediate rows of $\mathbf{D}(A)$ are just those of **S**.

To illustrate, we let A be 4-by-6, which is the largest part of **S** that can be modified and still allows the column sums to work with A and A' not overlapping. This justify that **D** is at least 11-by-12, which is an example of $N \geq 3b - 1$, so let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \end{bmatrix}. \tag{31}$$

Let $a_i = row_i(A)$ and

$$a = [a_1, a_2, a_3, a_4]^t \in R^{24}. \tag{32}$$

The conditions on A so that $\mathbf{D}(\mathrm{A})$ satisfies row sum, column sum, and order constraints are described by a matrix equation as follows. The conditions on a_1 corresponds to the matrix

$$V_1 = V(4; 1, -1, -3, -5, -7, -9) \tag{33}$$

with

$$V(4; 1, -1, -3, -5, -7, -9) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -3 & -5 & -7 & -9 \\ 1 & 1 & 9 & 25 & 49 & 81 \\ 1 & -1 & -27 & -125 & -343 & -729 \\ 1 & 1 & 81 & 625 & 2401 & 6561 \end{bmatrix}. \tag{34}$$

Similarly, let

$$V_2 = V(4; 3, 1, -1, -3, -5, -7), \tag{35}$$

$$V_3 = V(4; 5, 3, 1, -1, -3, -5), \tag{36}$$

$$V_4 = V(4; 7, 5, 3, 1, -1, -3). \tag{37}$$

Here V_i is a Vandermonde matrix which imposes the desired conditions on a_i. The equation

$$V_i a_i = \begin{bmatrix} 0 \\ -2 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{38}$$

establishes five constraints on a_i. The first is that the row sum of a_i is zero. The next four describe the conditions on a_i to obtain fourth-order approximation. This approach generalizes nicely to any order of approximation.

Now we use V_1, V_2, V_3, V_4 to form the 26-by-24 matrix:

$$\mathbf{M} = \begin{bmatrix} V_1 & 0 & 0 & 0 \\ 0 & V_2 & 0 & 0 \\ 0 & 0 & V_3 & 0 \\ 0 & 0 & 0 & V_4 \\ I_6 & I_6 & I_6 & I_6 \end{bmatrix}. \tag{39}$$

Let $r \in R^{26}$ be given as in (29), where $c^t = (0, -2, 0, 0, 0)$ and $d^t = (1, 0, 0, 1/24, 13/12, -1/24)$. The last 6 rows of $\mathbf{M}$ and the last 6 entries of r arise from the need to choose A so that the first 6 column sums of $\mathbf{D}(A)$ are $-1, 0, 0, 0, 0, 0$. The following matrix equation represents all the properties required:

$$\mathbf{M}a = r. \tag{40}$$

However this system is inconsistent for all $\mathbf{M}$ corresponding to orders (on the boundary) bigger than one. This leads us to use a weighted inner product (17) or, equivalently, to scale the first four rows of $\mathbf{D}(A)$. It is important to notice that our new divergence will satisfy the conditions stated in section 3 with the column condition being satisfied with respect to the weighted inner product. We have found two approaches that lead to solutions. We exhibit one of these.

Consider

$$\hat{V}(4; 1, -1, -3, -5, -7, -9) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 9 & 25 & 49 & 81 \\ 1 & -1 & -27 & -125 & -342 & -729 \\ 1 & 1 & 81 & 625 & 2401 & 6561 \end{bmatrix} \tag{41}$$

($\hat{V}$ and $\hat{b}$ are formed by deleting the second row of V and b, respectively), form

$$\hat{\mathbf{M}} = \begin{bmatrix} \hat{V_1} & 0 & 0 & 0 \\ 0 & \hat{V_2} & 0 & 0 \\ 0 & 0 & \hat{V_3} & 0 \\ 0 & 0 & 0 & \hat{V_4} \\ I_6 & I_6 & I_6 & I_6 \end{bmatrix}, \tag{42}$$

and solve

$$\hat{\mathbf{M}}a = \hat{r}. \tag{43}$$

This will give us our divergence. To get our weights we form a matrix $\mathbf{M_w}$ with the deleted rows from $\mathbf{M}$ and multiply $\mathbf{M_w}$ by a. This approach produces a three-parameter family of solutions. Here we present one divergence where there are changes only in the first row:

$$
\begin{array}{cccccccc}
-\frac{4751}{5192} & \frac{909}{1298} & \frac{6091}{15576} & -\frac{1165}{5192} & \frac{129}{2596} & -\frac{25}{15576} & 0 & 0 \\
\frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 & 0 & 0 \\
0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 & 0 \\
0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} & 0 \\
0 & 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} & -\frac{1}{24} \\
0 & 0 & 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} & \frac{9}{8} \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{24} & -\frac{9}{8} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{24}
\end{array}
\tag{44}
$$

where the weights are

$$
Q_0 = \frac{649}{576}\,, \quad Q_1 = \frac{143}{192}\,, \quad Q_2 = \frac{75}{64}\,, \quad Q_3 = \frac{551}{576}\,, \quad Q_4 = 1\,, \quad Q_5 = 1\,, \quad \ldots\,. \tag{45}
$$

Equivalently, $\Lambda = \mathrm{diag}(\frac{649}{576}, \frac{143}{192}, \frac{75}{64}, \frac{551}{576})$ in (24). Thus we have created a nice divergence.

This techniques work for higher order. For example, here is a sixth-order divergence. We exhibit only the first two rows of the matrix of a divergence, where the formula for this divergence is the usual sixth-order approximation of the derivative except for the boundary and first interior point [4]:

$$
\begin{array}{cccccccc}
-\frac{1077397}{1273920} & \frac{15668474643803}{32472850116480} & \frac{49955527}{39491520} & -\frac{25369793}{19745760} & \frac{12220145}{15796608} & -\frac{21334421}{78983040} & \frac{460217}{9872880} & -\frac{101017}{39491520} \quad \frac{3369}{26327680} \\
\frac{31}{960} & -\frac{687}{640} & \frac{129}{128} & \frac{19}{192} & -\frac{3}{32} & \frac{21}{640} & -\frac{3}{640} & 0 \qquad\quad 0
\end{array}
\tag{46}
$$

The corresponding Q weights are

$$
\frac{41137}{34560}\,, \quad \frac{15667}{34560}\,, \quad \frac{2933}{1728}\,, \quad \frac{2131}{4320}\,, \quad \frac{41411}{34560}\,, \quad \frac{33437}{34560}\,, \quad 1\,, \quad \ldots\,. \tag{47}
$$

The second approach is equivalent and yields the same results. In this approach, we regard the weights as unknowns. The 26-by-24 matrix M is augmented to a 26-by-28 matrix by adding four columns that have a single nonzero entry in each of the positions corresponding to the rows deleted from M in the first approach. The vector $\mathbf{r}$ is modified by replacing the vector $\mathbf{c}$ with 0. With either approach, some latitude is available for choosing the weights. The question of how to determine optimal weights merits further investigation. It is worth noting that we obtain inconsistent systems when A is smaller than 4-by-6.

7. High-order gradients. Using the exactness when $v \equiv 1$ of the discrete divergence $\mathbf{D}\ 1 = 0$, the summation-by-parts formula (1) reduces to

$$
\langle 1, P\,\mathbf{G}\ f\rangle = f_n - f_0. \tag{48}
$$

The operator $\mathbf{G}$ given in (49) satisfies this condition and is a second-order gradient

in the interior with first-order truncation error at the boundary. The gradient matrix for this case is

$$\mathbf{G} = \begin{pmatrix} -2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ & & \ddots & \ddots & \ddots & & & \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & -2 & 2 \end{pmatrix}, \tag{49}$$

and the matrix P is diagonal with entries $(1/2,1,1,\dots,1,1/2)$. Note that it is possible to construct gradients which are second order all the way to the boundary [3]. As for the divergence, we first try for gradients that are a standard fourth-order approximation away from the boundary and local special formulas near the boundaries. The standard fourth-order approximation is

$$(\mathbf{G}\ f)_i = \frac{1}{h}\left(\frac{1}{24}\, f_{i-\frac{3}{2}} - \frac{27}{24}\, f_{i-\frac{1}{2}} + \frac{27}{24}\, f_{i+\frac{1}{2}} - \frac{1}{24}\, f_{i+\frac{3}{2}}\right). \tag{50}$$

As for the divergence, the approach described in section 5 produces higher-order gradients on the boundary as well as in the interior.

The weights are again independent of the particulars of the gradient:

$$\bar{P}_0 = \frac{407}{1152}, \quad \bar{P}_1 = \frac{473}{384}, \quad \bar{P}_2 = \frac{343}{384}, \quad \bar{P}_3 = \frac{1177}{1152}, \quad \bar{P}_3 = 1, \quad \dots. \tag{51}$$

The resulting discrete inner product is a fifth-order approximation of the continuous inner product. A three-parameter family of uniformly fourth-order accurate gradients is

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} & g_{13} & g_{14} & g_{15} & g_{16} & 0 & 0\dots \\ \frac{16}{105} - \alpha\,\frac{128}{35} & -\frac{31}{24} + \alpha\, 9 & \frac{29}{24} - \alpha\, 12 & -\frac{3}{40} + \alpha\,\frac{54}{5} & \frac{1}{168} - \alpha\,\frac{36}{7} & \alpha & 0 & 0\dots \\ -\beta\,\frac{128}{35} & \frac{1}{24} + \beta\, 9 & -\frac{27}{24} - \beta\, 12 & \frac{27}{24} + \beta\,\frac{54}{5} & -\frac{1}{24} - \beta\,\frac{36}{7} & \beta & 0 & 0\dots \\ -\frac{16}{105} - \gamma\,\frac{128}{35} & \frac{3}{8} + \gamma\, 9 & -\frac{11}{24} - \gamma\, 12 & -\frac{27}{40} + \gamma\,\frac{54}{5} & \frac{51}{56} - \gamma\,\frac{36}{7} & \gamma & 0 & 0\dots \\ 0 & 0 & 0 & \frac{1}{24} & -\frac{27}{24} & \frac{27}{24} & -\frac{1}{24} & 0\dots \end{pmatrix} \tag{52}$$

where

$$\begin{aligned} g_{11} &= -\frac{124832}{42735} + \alpha\,\frac{16512}{1295} + \beta\,\frac{18816}{2035} + \gamma\,\frac{13696}{1295}, \\ g_{12} &= \frac{10789}{3256} - \alpha\,\frac{1161}{37} - \beta\,\frac{9261}{407} - \gamma\,\frac{963}{37}, \\ g_{13} &= -\frac{421}{9768} + \alpha\,\frac{1548}{37} + \beta\,\frac{12348}{407} + \gamma\,\frac{1284}{37}, \\ g_{14} &= -\frac{12189}{16280} - \alpha\,\frac{6966}{185} - \beta\,\frac{55566}{2035} - \gamma\,\frac{5778}{185}, \\ g_{15} &= \frac{11789}{22792} + \alpha\,\frac{4644}{259} + \beta\,\frac{5292}{407} + \gamma\,\frac{3852}{259}, \\ g_{16} &= -\frac{48}{407} - \alpha\,\frac{129}{37} - \beta\,\frac{1029}{407} - \gamma\,\frac{107}{37}. \end{aligned}$$

The choice $\alpha = 1/24$, $\beta = 0$, $\gamma = -1/24$ gives

$$\mathbf{G} = \begin{pmatrix} \frac{-1152}{407} & \frac{10063}{3256} & \frac{2483}{9768} & \frac{-3309}{3256} & \frac{2099}{3256} & \frac{-697}{4884} & 0 & \dots \\ 0 & \frac{-11}{12} & \frac{17}{24} & \frac{3}{8} & \frac{-5}{24} & \frac{1}{24} & 0 & \dots \\ 0 & \frac{1}{24} & \frac{-27}{24} & \frac{27}{24} & \frac{-1}{24} & 0 & 0 & \dots \end{pmatrix} \tag{53}$$

and results in a gradient that is modified only at the first interior point.

8. Nodal grid results. The known results on nodal grids go back to G. Scherer's doctoral thesis, some of which was presented in [13]. This work has been continued by Olsson [12] and Strand [14]. In [13], G. Scherer has a theorem that also indicates that it is not possible to obtain high order approximation at the boundary with the standard inner product. This corresponds with our proposition in section 3, where we also proved that with the standard inner product we could not obtain a mimetic divergence (or gradient) high-order approximation on the boundary. She also proves a theorem that restricts the order of approximation possible at the boundary, given the approximation in the interior. Her results state that if the interior approximation is order γ, then with use of a diagonal norm (weighted inner product), only order $\frac{\gamma}{2}$ is possible at the boundary. In order to get a higher approximation at the boundary (e.g., $\gamma - 1$), it is necessary to use a norm defined by a full matrix (generalized inner product) and it is not possible with a norm defined by a diagonal matrix.

Our results are that for a staggered grid we can construct a uniformly (interior and boundary) fourth- (and sixth-) order divergence (and gradient) with a norm defined by a diagonal matrix (weighted inner product).

Related to this theorem, we tried our techniques in a nodal grid using the standard central difference approximation in the interior. Here we have a uniformly fourth-order approximation to the derivative using central difference approximations. This seems to contradict the theorem of Scherer.

$$D = \begin{bmatrix} \frac{-33989}{13640} & \frac{49453}{8184} & \frac{-28993}{4092} & \frac{7391}{1364} & \frac{-18763}{8184} & \frac{16717}{40920} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \frac{-1}{4} & \frac{-5}{6} & \frac{3}{2} & \frac{-1}{2} & \frac{1}{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \frac{1}{12} & \frac{-287}{348} & \frac{55}{87} & \frac{-49}{174} & \frac{191}{348} & \frac{-55}{348} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & \frac{-1}{12} & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & \frac{-2}{3} & 0 & \frac{2}{3} & 0 & \dots \end{bmatrix} \tag{54}$$

We have a norm defined by a diagonal matrix, where the weights are

$$Q_0 = \frac{1705}{6156}, \quad Q_1 = \frac{5995}{4104}, \quad Q_2 = \frac{1363}{2052}, \quad Q_3 = \frac{13519}{12312}, \quad Q_4 = 1, \quad Q_5 = 1, \quad \dots . \tag{55}$$

9. Conclusions. Using matrix analysis we have characterized the problem of constructing high-order approximations to the derivative that satisfy a global conservation law. We constructed high-order approximations to both the divergence and the gradient operators which are high order (fourth) both in the interior and on the boundary. This approach generalizes to higher-order approximations naturally; in fact we reproduced the sixth-order approximations presented in [4]. The results presented here are for a one-dimensional uniform grid.

Some areas of further investigation are suggested by the results in this article. First, the question arises about how to choose the parameters, in our family of solutions, in an optimal fashion. Second, we have found that the weights have positive solutions for $k = 4, 6$, or 8 but have provided no explanation why this should hold for general k. This is of more theoretical than practical interest since higher than fourth- and sixth-order approximations are unlikely to be necessary in actual schemes. Third, this work may be a preliminary step in constructing high-order approximations to the divergence and gradient in two or more dimensions for both nodal and staggered grids. High-order mimetic operators have been constructed in [4] and [5]. It is hoped that an approach which involves this method in this article along with Kronecker products and graph techniques will lead to similar results and methods to the ones obtained here for dimension one in higher dimensional cases.

Acknowledgments. The work done here is part of an ongoing project of the first author with his colleagues in New Mexico (Hyman, Shashkov, and Steinberg; see [6, 5]). In fact, M. Shashkov originally posed the question discussed in this article and did some preliminary work on the problem. Additionally, M. Hyman, G. Scherer, and S. Steinberg have had extensive discussions with the authors and are thanked for the stimulating interactions.

REFERENCES

[1] A. L. Andrew, *Centrosymmetric matrices*, SIAM Rev., 40 (1998), pp. 697–699.
[2] M. H. Carpenter, D. Gottlieb, and S. Abarbanel, *Time-stable boundary conditions for finite difference solving hyperbolic systems: Methodology and applications to high-order compact schemes*, J. Comput. Phys., 111 (1994), pp. 220–236.
[3] J. E. Castillo, *A family of, including boundary, mimetic schemes from second to sixth order*, in preparation.
[4] J. E. Castillo, J. M. Hyman, M. J. Shashkov, and S. Steinberg, *Fourth and sixth-order conservative finite-difference approximations of the divergence and gradient*, Appl. Numer. Math., 37 (2001), pp. 171–187.
[5] J. E. Castillo, J. M. Hyman, M. J. Shashkov, and S. Steinberg, *High-order mimetic finite-difference schemes on non-uniform grids*, in ICOSAHOM 95, Houston, TX, 1996, pp. 347–361.
[6] J. E. Castillo, J. M. Hyman, M. J. Shashkov, and S. Steinberg, *The sensitivity and accuracy of fourth order finite-difference schemes on non-uniform grids in one dimension*, Comput. Math. Appl., 30-8 (1995), pp. 41–55.
[7] J. M. Hyman, R. J. Knapp, and J. C. Scovel, *High order finite volume approximations of differential operators on nonuniform grids*, Phys. D, 60 (1992), pp. 112–138.
[8] J. Hyman and M. Shashkov, *Mimetic discretization for Maxwell's equation*, J. Comput. Phys., 151 (1999), pp. 881–909.
[9] J. Hyman, M. Shashkov, and S. Steinberg, *The numerical solution of diffusion problems in strongly heterogeneous non-isotropic materials*, J. Comput. Phys., 132 (1997), pp. 130–148.
[10] M. Shashkov, *Conservative Finite-Difference Methods on General Grids*, CRC Press, Boca Raton, FL, 1995.
[11] M. Shashkov and S. Steinberg, *A new numerical method for solving diffusion equations with rough coefficients*, J. Comput. Phys., 129 (1996), pp. 383–405.

[12] P. Olsson, *Summation by parts, projections, and stability* I, Math. Comp., 64 (1995), pp. 1035–1065.

[13] H.-O. Kreiss and G. Scherer, *Finite element and finite difference methods for hyperbolic partial differential equations*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. De Boor, ed., Academic Press, New York, 1974, pp. 195–212.

[14] B. Strand, *Summation by parts for finite difference approximations for d/dx*, J. Comput. Phys., 110 (1994), pp. 47–67.

SIAM J. Matrix Anal. Appl.
Vol. 25, No. 1, pp. 143–151

A NOTE ON THE COLUMN ELIMINATION TREE*

JOHN R. GILBERT† AND LAURA GRIGORI‡

Abstract. This short communication considers the LU factorization with partial pivoting and shows that an *all-at-once* result is possible for the structure prediction of the column dependencies in L and U. Specifically, we prove that for every square strong Hall matrix A there exists a permutation P such that every edge of its column elimination tree corresponds to a symbolic nonzero in the upper triangular factor U. In the symbolic sense, this resolves a conjecture of Gilbert and Ng [*Graph Theory and Sparse Matrix Computation*, A. George, J. R. Gilbert, and J. W. H. Liu, eds., Springer-Verlag, New York, 1994, pp. 107–139].

Key words. column elimination tree, sparse partial pivoting, structure prediction, lower bounds

AMS subject classifications. 65F50, 65F05, 68R10

PII. S0895479801393770

1. Introduction.

Sparsity in matrix computations offers opportunities to save memory space by storing only nonzero elements, shorten execution times by eliminating computations on zeros, and exploit parallelism exposed by independent nonzero structures. Exploiting these opportunities often relies on a symbolic computation phase that predicts as accurately as possible which elements will have or can have nonzero values during the numerical computation itself, based only on the nonzero structure of the input matrix.

In LU factorization with partial pivoting, a square matrix A is factored as $PA = LU$, where P is a permutation matrix that depends on the values of the nonzeros of A and cannot be predicted only from the nonzero structure of A. Two structure prediction questions have been studied for this problem. The first is to predict bounds on the nonzero structure of the factors L and U. The second is to predict which columns of L and U depend directly or indirectly on which earlier columns. We restrict our attention to the class of matrices that satisfy an irreducibility condition called the *strong Hall* property.

George and Ng [4] developed upper bounds on the nonzero structure of L and U by employing a *row merge graph*. Gilbert and Ng [6] showed that this upper bound is as tight as possible in what they called "the exact sense." This means that, given the nonzero structure of a strong Hall matrix, for every edge in the row merge graph there is a choice of values for the nonzeros of A and a pivoting permutation P such that the corresponding element of L or U is nonzero. This is a *one-at-a-time* result [6]: any single position in the predicted structure can be made nonzero, but it may be the case that no single choice of nonzero values makes all the predicted elements nonzero at once.

The *column elimination tree* is a tree whose vertices are the columns of A and whose edges correspond to potential dependencies between columns (a complete defi-

*Received by the editors August 15, 2001; accepted for publication (in revised form) by E. Ng May 24, 2002; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/39377.html

†Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304 (gilbert@parc.xerox.com). The work of this author was performed during a visit to CERFACS, Toulouse, France.

‡INRIA, Domaine de Voluceau Rocquencourt - B.P. 105, 78153 Le Chesnay Cedex, France; currently visiting Lawrence Berkeley National Laboratory (LGrigori@lbl.gov). This author conducted her research as a Ph.D. student at INRIA Lorraine, Université Henri Poincaré, France.

nition is below.) Gilbert and Ng [6] showed that if k is the parent of j in the column elimination tree of a strong Hall matrix A, there exists a choice of nonzero values of A that will make column j update column k during factorization with partial pivoting—that is, a choice of nonzero values for A that will make $u_{jk} \neq 0$. This is again a one-at-a-time result.

A stronger statement would be an *all-at-once* result, showing that *all* the predicted positions can be made nonzero for the same input values. Unlike the case of sparse QR factorization, no tight all-at-once prediction is possible for the structure of L and U. The purpose of this short communication is to show that if we consider only the edges of the column elimination tree, an all-at-once result is possible in the symbolic sense. We prove that for every square strong Hall matrix A, there exists a permutation P such that every edge of the column elimination tree corresponds to a symbolic nonzero in the upper triangular factor U of A with partial pivoting. This resolves a variant of a conjecture of Gilbert and Ng [6].

Our result is *symbolic* in the sense that we assume that addition or subtraction of nonzeros always yields a nonzero result. Gilbert and Ng [6] also consider what they call *exact* results; we discuss this further in the conclusion.

A motivation for the current result is its impact on solvers that use the column elimination tree to model factorization in parallel. In solvers like the one described by Gilbert [5] and in the shared memory version of SuperLU [3], the tasks are scheduled dynamically on processors by using the precedence given by the column elimination tree. Our result shows that, in fact, for every strong Hall nonzero structure there is a matrix for which every dependency in the column elimination tree is a real constraint on the order of computation of the columns of the factor.

The next section presents background results and notation used in the paper. Section 3 introduces new results on the structure of the matrix during elimination. These results help to prove the all-at-once structure prediction of the column elimination tree. Section 4 concludes the paper.

2. Background. Let $A = (a_{rc})$ be a square, possibly unsymmetric, sparse $n \times n$ matrix which is to be factored as $PA = LU$ using partial pivoting.

In the following we introduce the commonly used tree and graph structures, the strong Hall property, a previously published theorem and lemma that will subsequently be used in our proofs. Most of our notation is similar or identical to that of Gilbert and Ng [6].

The *column intersection graph* $G_\cap(A)$ is undirected and has n vertices (one for each column) and an edge (i, j) if there is an r such that $a_{ri} \neq 0$ and $a_{rj} \neq 0$. This graph is equal to the graph of $A^T A$, unless there is numerical cancellation; in general $G(A^T A) \subseteq G_\cap(A)$.

The *directed graph* $G(A)$ has n vertices and an edge (i, j) for each nonzero element a_{ij}. The *bipartite graph* $H(A)$ has $2n$ vertices (one for each row and one for each column) and an edge (r', c) whenever a_{rc} is nonzero. In the bipartite graph, we use primes on the names of row vertices. For any graph G and vertex v, we write $Adj(v, G)$ to represent the set of vertices w such that (v, w) is an edge of G.

The *elimination tree* structure (*etree*) was first introduced for the Cholesky factorization of symmetric positive definite (SPD) matrices [8]. If L is the Cholesky factor of the SPD matrix A, then this tree has n vertices, and k is the parent of j if and only if $k = \min\{r > j : l_{rj} \neq 0\}$. Later the elimination tree was adapted to the LU factorization with partial pivoting [5]; the *column elimination tree* is the elimination tree of the column intersection graph $G_\cap(A)$ or, equivalently, the elimination tree of $A^T A$ if there is no numerical cancellation when computing or factoring $A^T A$.

A *strong Hall graph* is a bipartite graph with m rows and n columns that has the strong Hall property [2, 6]: every set of k column vertices is adjacent to at least $k+1$ row vertices for all $1 \leq k < n$. A square matrix has the strong Hall property if and only if it is a *fully indecomposable matrix*, that is, there are no two permutations P and Q such that PAQ is block triangular.

Before introducing the necessary theorem and lemma, let us elaborate on an additional definition, that of a sequence of bipartite graphs which model the structure of L and U during the elimination. Let $H_0 = H(A)$ be the bipartite graph of A. Suppose a_{rc} is nonzero and is chosen as pivot at step 1. The *deficiency* of the edge (r', c) of H_0 is defined as the set of edges

$$\{(i', j) : c \in Adj(i', H_0),\ j \in Adj(r', H_0),\ \text{and } j \notin Adj(i', H_0)\}.$$

It corresponds to the zero elements of A that become nonzero when a_{rc} is used as a pivot in Gaussian elimination.

Knowing the sequence of pivoting elements $(r'_1, c_1), (r'_2, c_2), \ldots, (r'_{n-1}, c_{n-1})$, we can construct a sequence of bipartite graphs $H_0, H_1, \ldots, H_n$, where H_i describes the structure of the $(n-i)\times(n-i)$ Schur complement remaining after step i. The bipartite graph H_i of the $(n-i) \times (n-i)$ submatrix that remains after eliminating (r'_i, c_i) is obtained as follows: delete from H_{i-1} vertices r'_i and c_i and all edges incident to them, then add the edges in the deficiency of (r'_i, c_i). The *bipartite filled graph* $H^+(A)$ is the bipartite graph containing all the edges of all H_i.

If the diagonal elements of A are nonzero, and the pivots are chosen in the order $(1', 1),\ (2', 2), \ldots, (n', n)$, then we write $G^+(A)$ for the *filled graph* of A, which is obtained from $H^+(A)$ by merging each row vertex v' with its corresponding column vertex v. The *filled column intersection graph* $G^+_\cap(A)$ is the filled graph of the column intersection graph of A, that is, $G^+(G_\cap(A))$. If H_0 is the bipartite graph of A, then $G_\cap(H_0)$ is equivalent to $G_\cap(A)$. ($G^+_\cap(H_0)$ is equivalent to $G^+_\cap(A)$.)

With these definitions at hand we now mention two results on which ours is based.

THEOREM 1 (Gilbert and Ng [6]). *Let H_0 be a bipartite graph and let (r', c) be an edge of H_0. Let H_1 be the bipartite graph resulting from the elimination of edge (r', c). If H_0 has the strong Hall property, then H_1 also has the strong Hall property.*

For the following lemma (called the *fill path* lemma), a *path* is a sequence of edges $P = [(v_0, v_1), (v_1, v_2), \ldots, (v_{p-1}, v_p)] = [v_0, v_1, \ldots, v_p]$ in which all the vertices are distinct. The length of this path P is p.

LEMMA 2 (Rose and Tarjan [7]). *Let G be a directed or undirected graph whose vertices are the integers 1 through n, and let G^+ be its filled graph. Then (x, y) is an edge of G^+ if and only if there is a path in G from x to y whose intermediate vertices are all smaller than* $\min(x, y)$.

We conclude this section by presenting several previous results, outlining the role of the different graphs, introduced here, in the structure prediction of L and U. If the matrix A can be factored without row or column interchanges, then $G(L + U)$ is equal to $G^+(A)$ unless numerical cancellation occurs.

If pivoting is necessary during the Gaussian elimination, then only upper bounds on the structures of L and U can be predicted. The filled column intersection graph of A represents such an upper bound: $G(U) \subseteq G^+_\cap(A)$, and a slightly different representation of L is also a subgraph of $G^+_\cap(A)$. Thus the graph $G^+_\cap(A)$ contains an edge for each element of L and U that can possibly be nonzero during the numerical computation. If the matrix A has the strong Hall property, then the filled column intersection graph is a tight exact bound for the nonzero structure of U [6].

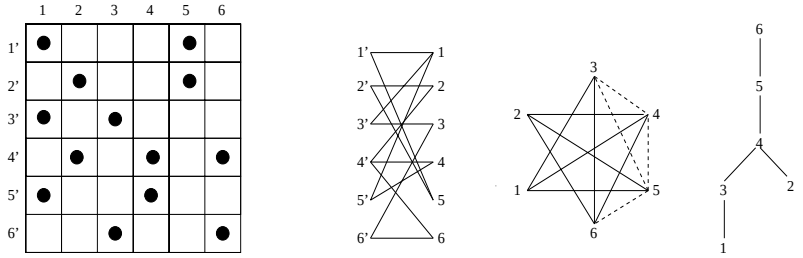


FIG. 1. *Matrix example A, the bipartite graph H_0, the filled column intersection graph $G^+_\cap(H_0)$, and its column elimination tree. The dotted lines in the filled column intersection graph $G^+_\cap(H_0)$ represent fill-in.*

3. Structure prediction and the column elimination tree.

3.1. An example. The elimination tree plays an important role in the parallel sparse Cholesky factorization of symmetric positive definite matrices. This tree describes all the dependencies between column computations, and it represents the task scheduling model of almost all parallel sparse Cholesky solvers.

In the LU factorization with partial pivoting, the column elimination tree predicts all potential dependencies between columns, and hence it can be used as a task scheduling model in the unsymmetric case. For example, in the shared memory version of SuperLU [3] this tree helps in identifying two levels of parallelism in the LU factorization with partial pivoting. As described in [3], a first level of parallelism exploits the property that computations in disjoint subtrees are independent, thus leading to assigning disjoint subtrees to different processors; a second level of parallelism sequences in a pipelining manner the computation of dependent columns in a subtree. This level is especially useful in the superior part of the tree, where there are more idle processors than disjoint subtrees.

The nonzero structure of U cannot be, in general, exactly determined prior to the numerical factorization, and thus the column elimination tree can overestimate the real column dependencies. Consider, for example, the strong Hall matrix A in Figure 1 with its bipartite graph H_0, the filled column intersection graph $G^+_\cap(H_0)$, and its column elimination tree.

Suppose that at the first elimination step the diagonal element is used as pivot. This means that the element u_{13} is zero and there is no dependency between the computations of columns 1 and 3. In other words, the dependency between the nodes 1 and 3 in the column elimination tree of A corresponds to an overestimation of the real dependencies.

Let us now analyze the later stages of the elimination. Consider the matrix P_1A in Figure 2 (P_1 describes the first elimination step), the bipartite graph H_1 resulting from the elimination of edge $(1', 1)$, followed by its filled column intersection graph $G^+_\cap(H_1)$ and the corresponding column elimination tree.

We note that while the edge $(3, 4)$ is present in the filled column intersection graph of H_0, it does not belong to the filled column intersection graph of H_1, and thus the structures of these two graphs are different. Hence, in general the graph $G^+_\cap(H_1)$ cannot be simply obtained by deleting the vertex 1 and its incident edges from the graph $G^+_\cap(H_0)$.

As a consequence, the structure of the column elimination trees related to the elimination graphs H_i can change from one elimination step to another. In our ex-

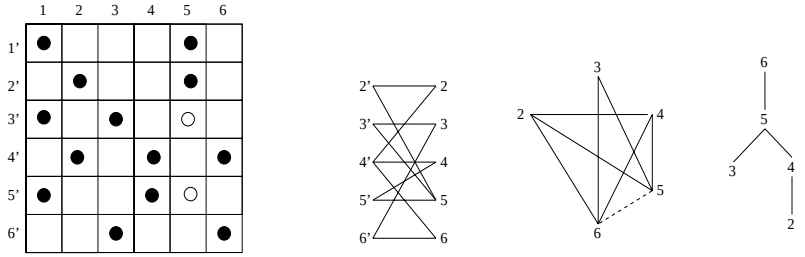


FIG. 2. *Matrix P_1A (including the deficiency of $(1', 1)$ represented by* o*), the bipartite graph H_1, the filled column intersection graph $G_\cap^+(H_1)$, and its column elimination tree.*

ample, after the first step of elimination there is no potential dependency between the computations of columns 3 and 4, and thus the edge $(3, 4)$ is not present in the column elimination tree of H_1 and is replaced by the edge $(3, 5)$.

This simple example shows that the column elimination tree may overestimate the dependencies between columns. However, it has been shown that for a strong Hall matrix, this is the tightest information we can obtain before the numerical factorization of A. In other words, for each edge of the column elimination tree, there exists a choice of numerical values of A such that this edge corresponds to a real column dependency.

3.2. Main result. In this section we prove the main result of the paper, which is that every strong Hall nonzero pattern admits a pivoting permutation for which every edge of the column elimination tree corresponds to a symbolic column dependency. We first prove a lemma saying that there is a choice of pivot element such that the first elimination step creates the correct dependency (corresponding to an edge in the column elimination tree) for the first column and also does not change the structure of the filled column intersection graph. The lemma essentially says that (with this pivoting order) we never learn anything more about the uncomputed rows of U than we knew from $G_\cap^+$ at the beginning. We then prove the main theorem by induction.

We make two observations about symbolic elimination. First, the fact that $i + 1$ is the least-valued vertex in H_i implies that there is no fill-in edge having $i + 1$ as an endpoint. This gives us

$$Adj(i + 1, G_\cap^+(H_i)) = Adj(i + 1, G_\cap(H_i)) \tag{1}$$

$$= \{v : i + 1 \in Adj(t', H_i) \text{ and } v \in Adj(t', H_i)\}. \tag{2}$$

Second, the fill path lemma implies that the vertices in the set $\{i + 1\} \cup Adj(i + 1, G_\cap^+(H_i))$ form a complete subgraph.

The next lemma shows that when we pivot on an element that is not the only element in its column, enough fill is added to preserve the structure of the filled column intersection graph. For each vertex i, we denote its parent in the column elimination tree by parent$[i]$; by definition this is $\min\{j > i : j \in Adj(i, G_\cap^+(H_0))\}$.

LEMMA 3. *Let H_0 be the structure of a square matrix A with at least two nonzero elements in column* 1*. Let P_1 be the permutation matrix that interchanges row r' with row* 1 *such that the edge* $(1, \text{parent}[1])$ *of the elimination tree of $G_\cap(A)$ corresponds to a nonzero in the upper triangular factor U. If H_1 is the bipartite graph resulting from the elimination of edge $(r', 1)$, then the filled column intersection graph of H_1 is*

obtained from the filled column intersection graph of H_0 just by deleting vertex 1 *and its incident edges. That is,*

$$G_{\cap}^{+}(H_1) = G_{\cap}^{+}(H_0) - \{1\}. \tag{3}$$

Proof. We will prove that the deficiency set of $(r', 1)$ introduces all the edges and preserves all the paths that can disappear by deleting row r' and column 1 while constructing the graph H_1. We will also show that this deficiency set does not introduce new edges or new fill paths in $G_{\cap}^{+}(H_1)$ compared to $G_{\cap}^{+}(H_0)$.

Let us analyze what happens when adding the deficiency of $(r', 1)$. Let $S = Adj(1, H_0)$ be the set of row indices of nonzeros in column 1. From the lemma statement, recall that column 1 is adjacent to at least two row vertices, so that r' is not the only element in set S.

By using the definition of the deficiency of $(r', 1)$, for each $t' \in S$ such that $t' \neq r'$ and for each edge $v \neq 1$ adjacent to r' in H_0, we see that v belongs to $Adj(t', H_1)$. For each two vertices $v_1, v_2 \in Adj(t', H_1)$, by using the definition of the column intersection graph, we see that (v_1, v_2) is an edge of $G_{\cap}(H_1)$.

Let us make an analysis depending on the origin of vertices v_1, v_2. First, if v_1, v_2 are adjacent to r' in H_0 (that is, $v_1, v_2 \in Adj(r', H_0)$), the fact that (v_1, v_2) is an edge of $G_{\cap}(H_1)$ proves that the deletion of the row r' does not change the structure of $G_{\cap}(H_1)$ compared to the structure of $G_{\cap}(H_0)$.

Second, if $v_1 \in Adj(r', H_0)$ and $v_2 \in Adj(t', H_0)$, then v_1 and v_2 are both adjacent to 1 in the column intersection graph of H_0, so $v_1, v_2 \in Adj(1, G_{\cap}(H_0))$. By using the observation at the beginning of this section, we see that (v_1, v_2) belongs to $G_{\cap}^{+}(H_0)$. This proves that the deficiency set does not introduce new edges in $G_{\cap}^{+}(H_1)$.

Using this analysis of edges introduced in H_1, we can easily check that

$$Adj(1, G_{\cap}(H_0)) - \{v\} \subseteq Adj(v, G_{\cap}(H_1)) \; \forall v \in Adj(r', H_0), v \neq 1. \tag{4}$$

Suppose that $[x_1, \ldots, x_r]$, $r > 2$, is a fill path in $G_{\cap}(H_0)$ and has 1 as an intermediate vertex. This means that $x_k < \min\{x_1, x_r\}$ for all $k = 2, \ldots, r-1$, and the edge (x_1, x_r) belongs to $G_{\cap}^{+}(H_0)$. Suppose that $x_k = 1, k > 1, k < r$. By using relation (4), we see that $x_{k-1}, x_{k+1} \in Adj(\text{parent}[1], G_{\cap}(H_1))$. If $x_{k-1} = \text{parent}[1]$ or $x_{k+1} = \text{parent}[1]$, it is evident that the fill path is preserved, since we can suppress 1 from the path while preserving adjacency in the path. Otherwise, vertex 1 can be replaced by parent[1] in the path $[x_1, \ldots, x_{k-1}, \text{parent}[1], x_{k+1}, \ldots, x_r]$. By using the definition of the column etree, we see that $x_{k-1}, x_{k+1} \geq \text{parent}[1]$, and this shows that $[x_1, \ldots, x_r]$ is a fill path in $G_{\cap}(H_1)$. This proves that all the fill paths are preserved in $G_{\cap}(H_1)$ and no new fill path is introduced. ∎

Figure 3 shows a matrix example A, its bipartite graph H_0, followed by the filled column intersection graph $G_{\cap}^{+}(H_0)$ with its column elimination tree. Figure 4 presents the permuted matrix $P_1 A$, the bipartite graph H_1 with its filled column intersection graph $G_{\cap}^{+}(H_1)$, and the corresponding column elimination tree.

Consider the elimination of edge $(4', 1)$ in matrix example A, Figure 3. The vertices 3 and 7 are adjacent to $4'$ in the bipartite graph H_0. Deleting the row $4'$ causes the edge $(3, 7)$ to disappear from $G_{\cap}^{+}(H_1)$. By adding the deficiency of $(4', 1)$, the edge $(3, 7)$ is introduced in $G_{\cap}^{+}(H_1)$ due to row vertex $2'$. Now consider the vertex 3 adjacent to vertex $4'$ and the vertex 5 adjacent to vertex $2'$ in the bipartite graph H_0. By the permutation of row $4'$ with row $1'$, the edge $(3, 5)$ is introduced in the filled column intersection graph $G_{\cap}^{+}(H_1)$. However, we remark that $(3, 5)$ was already

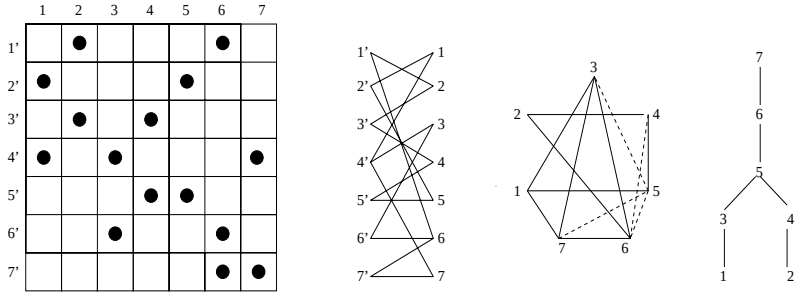


FIG. 3. *Matrix example A, the bipartite graph H_0, the filled column intersection graph $G_\cap^+(H_0)$, and its column elimination tree. The dotted lines in the filled column intersection graph $G_\cap^+(H_0)$ represent fill-in.*

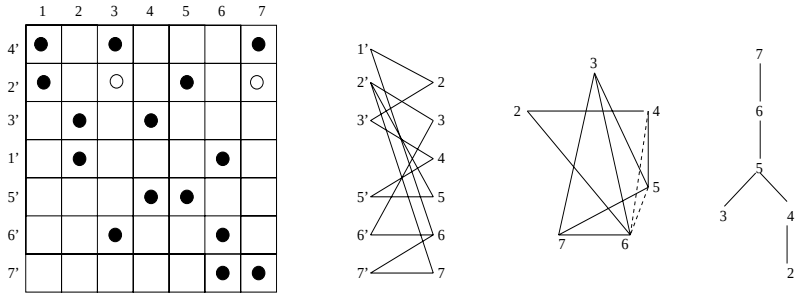


FIG. 4. *Matrix P_1A (including the deficiency of $(4', 1)$ represented by* o*), the bipartite graph H_1, the filled column intersection graph $G_\cap^+(H_1)$, and its column elimination tree.*

present in the filled column intersection graph $G_\cap^+(H_0)$. Finally, we consider the fill path [5 1 3 6] in $G_\cap(H_0)$, which is preserved in $G_\cap(H_1)$ in a compact form [5 3 6].

The next theorem is the main result of this paper. It proves the conjecture of Gilbert and Ng [6] in the symbolic sense, that is, if we assume that zeros are introduced only by explicit elimination and not by cancellation.

THEOREM 4. *Let A be an unsymmetric square sparse matrix having the strong Hall property. There is a permutation P such that every edge of the elimination tree of $G_\cap(A)$ corresponds to a nonzero in the upper triangular factor U in the symbolic sense, when the factorization $PA = LU$ is computed.*

Proof. We will prove this by induction. Let $H_0 = H(A)$ be the bipartite graph of A.

Initial phase. We show that there exists a permutation P_1 such that the element $u_{1,\text{parent}[1]}$ is nonzero.

Using relation (1), we see that $(1, \text{parent}[1])$ belongs to $G_\cap(H_0)$. There exists a row vertex r'_1 such that $(r'_1, 1)$ and $(r'_1, \text{parent}[1])$ are edges of H_0. We choose r'_1 as pivot, and P_1 describes this permutation. Row 1 is interchanged with row r'_1; therefore the element $u_{1,\text{parent}[1]}$ is nonzero.

Induction phase ($m-1 \to m$). We suppose that there is a sequence of permutations $P_{m-1}, \ldots, P_1$ such that for all $k = 1, \ldots, m-1$, $u_{k,parent[k]}$ is nonzero. We show that there is a permutation P_m such that the element $u_{m,parent[m]}$ is nonzero.

According to Theorem 1, at each elimination step k, the bipartite graph H_k is strong Hall because H_{k-1} is. In particular, this means that at each elimination step

we have at least two elements from which to choose to pivot on (column vertex k is adjacent to at least two row vertices in the graph H_{k-1}).

Therefore, Lemma 3 applies and says that at each elimination step k, the structure of the filled column intersection graph is preserved:

$$G_{\cap}^{+}(H_k) = G_{\cap}^{+}(H_{k-1}) - \{k\}, \quad 1 \leq k < m. \tag{5}$$

This relation shows that the induction hypothesis has as a direct consequence that the structure of the filled column intersection graph was preserved until this step m of elimination. We can deduce that $(m, \text{parent}[m])$ belongs to $G_{\cap}^{+}(H_{m-1})$. Even more, relation (1) says that this edge belongs to $G_{\cap}(H_{m-1})$.

Thus, there is some vertex r'_m such that (r'_m, m) and $(r'_m, \text{parent}[m])$ are edges of H_{m-1}. We choose r'_m as pivot and let P_m describe this permutation. The permutation of the row m with the row r'_m will make the element $u_{m,\text{parent}[m]}$ be nonzero.

Let $P = P_{n-1}, \ldots, P_1$ be the permutation matrix that includes the $n-1$ row interchanges. We have proved that every edge of the column elimination tree corresponds to a symbolic nonzero in the upper triangular factor U, when the factorization $PA = LU$ is computed with partial pivoting. □

4. Concluding remarks. The main result of this paper is Theorem 4, which gives an all-at-once structure prediction result, under the assumption that the matrix A is strong Hall. In the proof, we showed that if at each elimination step k the element $u_{k,\text{parent}[k]}$ is nonzero, then the structure of the filled column intersection graph is preserved during the elimination. One way to interpret this result is that (for a strong Hall matrix) there is a pivot sequence for which the only information about the structure of U exposed by each elimination step is the single newly computed row. In other words, the elimination does not give progressively more partial information about the uncomputed rows of U than was available from $G_{\cap}^{+}$ at the beginning.

We remark that, in the proof of Theorem 4, the strong Hall property was used in only one place for each elimination step. We used the strong Hall property to conclude that at each step (except the last), there is always a choice of at least two elements to pivot on. One could ask whether the strong Hall property is necessary as well as sufficient for this.

Our result is symbolic in the sense that we assume that during Gaussian elimination the result of adding or subtracting two nonzeros is never zero. A stronger result would be what Gilbert and Ng [6] called *exact*, which would assume only that the nonzero values in A were algebraically independent from each other; in other words, it would assume that any computed zeros were due to combinatorial properties of the nonzero structure rather than to coincidence in choice of values. We do not know whether or not the exact version of our main theorem holds, though we conjecture that it does. An exact version holds, for example, for the class of strong Hall matrices with exactly two nonzeros in every row and every column, because every elimination step creates exactly one new nonzero, and that nonzero is algebraically independent of the other remaining nonzeros.

We conclude by mentioning an open problem: What is the case for nonstrong Hall matrices, either for the elimination tree or for the structures of L and U? In this case, it is known that $G_{\cap}^{+}(A)$ may not be a tight bound for U. Is there a tight bound on U? If so, does it share the property that there is no new information revealed during the elimination except the structure of the current row of U?

Acknowledgments. The authors thank the anonymous reviewers for their helpful comments and suggestions to improve the presentation of the paper.

REFERENCES

[1] R. A. Brualdi and H. J. Ryser, *Combinatorial Matrix Theory*, Cambridge University Press, Cambridge, UK, 1991.

[2] R. A. Brualdi and B. L. Shader, *Strong Hall matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 359–365.

[3] J. W. Demmel, J. R. Gilbert, and X. S. Li, *An asynchronous parallel supernodal algorithm for sparse Gaussian elimination*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 915–952.

[4] A. George and E. Ng, *Symbolic factorization for sparse Gaussian elimination with partial pivoting*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 877–898.

[5] J. R. Gilbert, *An Efficient Parallel Sparse Partial Pivoting Algorithm*, Tech. Report 88/45052-1, Christian Michelsen Institute, Bergen, Norway, 1988.

[6] J. R. Gilbert and E. G. Ng, *Predicting structure in nonsymmetric sparse matrix factorizations*, in Graph Theory and Sparse Matrix Computation, A. George, J. R. Gilbert, and J. W. H. Liu, eds., Springer-Verlag, New York, 1994, pp. 107–139.

[7] D. J. Rose and R. E. Tarjan, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM J. Appl. Math., 34 (1978), pp. 176–197.

[8] R. Schreiber, *A new implementation of sparse Gaussian elimination*, ACM Trans. Math. Software, 8 (1982), pp. 256–276.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 152–164

SUPERLINEAR PRECONDITIONERS FOR FINITE DIFFERENCES LINEAR SYSTEMS*

STEFANO SERRA CAPIZZANO† AND CRISTINA TABLINO POSSIO‡

Abstract. We consider a preconditioning strategy for finite differences (FD) matrix sequences $\{A_n(a,\Omega)\}_n$ discretizing the elliptic problem

$$\begin{cases} A_a u \equiv (-)^k \nabla^k [a(x)\nabla^k u(x)] = f(x), & x \in \Omega, \\ \left(\dfrac{\partial^s}{\partial \nu^s} u(x)\right)_{|\partial\Omega} \equiv 0, & s = 0, \ldots, k-1, \end{cases}$$

with Ω being a plurirectangle of $\mathbf{R}^d$, with $a(x)$ being a uniformly positive (nonnegative) Riemann integrable function, and ν denoting the unit outward normal direction. More precisely, in connection with preconditioned conjugate gradient (PCG)–like methods, we consider the preconditioning sequence $\{P_n(a,\Omega)\}_n$, $P_n(a,\Omega) := \tilde{D}_n^{1/2}(a,\Omega) A_n(1,\Omega) \tilde{D}_n^{1/2}(a,\Omega)$, where $\tilde{D}_n(a,\Omega)$ is the suitable scaled main diagonal of $A_n(a,\Omega)$. Using embedding arguments and projection matrices, under the mild assumptions on $a(x)$, we show the weak clustering at the unity of the corresponding preconditioned sequence. If $a(x)$ is regular enough, then the preconditioned sequence shows a strong clustering at the unity so that the sequence $\{P_n(a,\Omega)\}_n$ turns out to be a superlinear preconditioning sequence for $\{A_n(a,\Omega)\}_n$. The computational interest is due to the fact that the solution of a linear system with coefficient matrix $A_n(a,\Omega)$ is reduced to computations involving diagonals and multilevel structures $\{A_n(1,\Omega)\}_n$ with banded pattern. In turn, the matrix $A_n(1,\Omega)$ can be reinterpreted as a projection of a multilevel banded Toeplitz matrix for which we use multigrid strategies. Some numerical experimentations confirm the efficiency of the discussed proposal and its strong superiority with respect to existing techniques in the case of semielliptic problems.

Key words. finite differences (FD), graph matrices, conditioning and preconditioning, multilevel (Toeplitz) structure, multigrid, preconditioned conjugate gradient (PCG)

AMS subject classifications. 15A12, 65F10, 65N22

PII. S0895479802416058

1. Introduction.

We consider finite differences (FD) discretizations of differential problems of the form

$$\begin{cases} A_a u \equiv (-)^k \nabla^k [a(x)\nabla^k u(x)] = f(x), & x \in \Omega, \\ \left(\dfrac{\partial^s}{\partial \nu^s} u(x)\right)_{|\partial\Omega} \equiv 0, & s = 0, \ldots, k-1, \end{cases} \tag{1.1}$$

with Ω being a plurirectangle of $\mathbf{R}^d$, with $a(x)$ being a uniformly positive (nonnegative) Riemann integrable function, and with ν denoting the unit outward normal direction. Using embedding arguments and projection matrices, it is easy to show that the powerful preconditioning techniques developed earlier (see [18, 22, 27]) in the simpler case of (1.1) acting on the hypercube $(0,1)^d$ can be translated in the new setting without any loss in the performances of the considered preconditioned conjugate gradient (PCG) procedures.

*Received by the editors October 12, 2002; accepted for publication (in revised form) by L. Reichel December 18, 2002; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/41605.html

†Dipartimento di Chimica, Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100 Como, Italy (Stefano.serrac@uninsubria.it, serra@mail.dm.unipi.it).

‡Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milano, Italy (Cristina.Tablinopossio@unimib.it).

The idea is quite general, at least for the following three reasons: Any set Ω that is simply Peano–Jordan measurable (i.e., its characteristic function is integrable in the Riemann sense) can be approximated in measure arbitrarily well by plurirectangles; in principle, the discretizations over nonequispaced grids can be included since this was partially done on the hypercube $(0,1)^d$ [28, 25]; and it is reasonable to expect the same result for finite elements approximations since, after using numerical formulae for evaluating the involved integrals, the resulting matrices are formally of the same type as FD matrices (see [19, 26] for preconditioning-oriented results and [2] for spectral analysis–oriented results in the presence of an L-shaped domain with finite element–style nonequispaced grids).

The paper is organized as follows. In section 2 we describe in full detail the case of the FD discretization of problem (1.1) with $k = 1$, $d = 2$, with FD formulae of minimal precision order 2 and with Ω being an L-shaped domain. We recall that a plurirectangle is an open set obtained by the union of a finite number of rectangles whose edges are parallel to the axes; therefore, the L-shaped domains represent the simplest plurirectangles which are not rectangles. The idea that we have in mind is to convince the reader of the structure of the analysis in a simple but significant case before presenting the results in their full generality.

For the basic case we show how the "embedding argument" both in the continuous (operator and domain) and in the discrete (matrix representation) settings is the key point in reducing the spectral and preconditioning analyses to the model case where the domain is a square (a hypercube in the d-level setting).

Section 3 is devoted to the general problem (1.1) with a special emphasis on the delicate points to be taken into account in extending the techniques from the basic example, and with a discussion on possible further applications.

Finally, section 4 is concerned with conclusions, open problems, and future work.

2. A model problem with $d = 2$, $k = 1$, and L-shaped Ω. The first step is to understand the relationships between the discretization of

$$\begin{cases} A_a u \equiv -\nabla[a(x)\nabla u(x)] = f(x), & x \in \Omega, \\ \\ \text{Dirichlet boundary conditions} \end{cases} \tag{2.1}$$

when $\Omega := L$ and $\Omega := Q$. Here $Q = (0,1)^2$, and L is a prototype of L-shaped domains and is defined as $Q \backslash Q'$ with $Q' = (0, 0.5]^2$. Q can be seen as the rectangle of minimal measure containing L.

We choose a given FD discretization process for the derivatives $\frac{\partial}{\partial x_i}$, $i = 1, 2$, with a uniform gridding of meshsize h and leaving the operator in divergence form: in this way, by following the same lexicographical order in the equations and in the unknowns, at least in the case of a rectangle, we preserve the symmetry and the positive definiteness of the resulting matrix (see [22, 24]).

We call $A_n(a, \Omega)$ the coefficient matrix, where Ω is either L or Q. We are now interested in the relationships between $A_n(a, Q)$ and $A_n(a, L)$.

Let us consider a given grid point $x^{(i)}$ of Q with $x^{(i)} = (i_1, i_2)h$, $i = (i_1, i_2)$, $i_1, i_2 \in \{1, \ldots, n\}$, and $h = 1/(n + 1)$. If the grid point belongs to L, then the resulting equation

$$(A_a u)(x^{(i)}) = f(x^{(i)})$$

will contribute in forming both the matrix $A_n(a, Q)$ and the matrix $A_n(a, L)$; in the matrix $A_n(a, Q)$ this equation will represent the row $(i_1 - 1)n + i_2$.

Conversely, if $x^{(i)} \notin L$, then the considered equation will not be present in the matrix $A_n(a, L)$, and the unknown $u(x^{(i)})$ will not be part of the unknowns of the associated linear system. We observe that the grid point $x^{(i)}$ defines the column of index $(i_1 - 1)n + i_2$ in $A_n(a, Q)$.

In conclusion, the matrix $A_n(a, L)$ can be obtained from the matrix $A_n(a, Q)$ by deleting the rows and columns belonging to the set of indices that are uniquely related to the grid points of Q not belonging to L. This fact is formally stated and resumed in the following proposition.

PROPOSITION 2.1. *Under the notation of section* 2*, there exists a matrix* Π *such that*

$$A_n(a, L) = \Pi A_n(a, Q)\Pi^T.$$

The matrix $A_n(a, Q)$ *has size* $d_n(Q) = n^2$*, and the matrix* $A_n(a, L)$ *has size* $d_n(L) = \frac{3}{4}n^2 + O(n)$*, and the matrix* $\Pi \in \mathbf{R}^{d_n(L)\times d_n(Q)}$ *and is obtained from the identity matrix of size* $d_n(Q)$ *by deleting the* j*th row if and only if* $j \equiv j(i) = (i_1 - 1)n + i_2$ *for some* $x^{(i)} \in Q\backslash L$*,* $i = (i_1, i_2) \in \{1, \ldots, n\}^2$*.*

The above reasoning is quite general and can be used whenever there is a PDE defined over two plurirectangles in which one is embedded in the other and is not affected by higher order operators or higher order formulas. A general analysis of this embedding argument can be performed in the context of graph theory and, more specifically, in the framework of (generalized) Laplacians of graphs (see [13]).

2.1. Spectral properties of $A_n(a, L)$. In this subsection we analyze some interesting consequences of Proposition 2.1. Despite the serendipity of the result, the spectral characterization that we can obtain for $A_n(a, L)$ is very rich. We present a useful definition.

DEFINITION 2.2. *Let* $\{A_n\}_n$ *be a sequence of matrices of increasing dimensions* d_n *and let* θ *be a measurable function defined over a set* K *of finite Lebesgue measure. We write that* $\{A_n\}_n$ *is distributed as the measurable function* θ *over* K *in the sense of the eigenvalues; i.e.,* $\{A_n\}_n \sim_\lambda (\theta, K)$ *if for every* F *continuous, real valued, and with bounded support, we have*

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F\left(\lambda_j(A_n)\right) = \frac{1}{m\{K\}} \int_K F(\theta(s))\, ds, \tag{2.2}$$

where $\lambda_j(A_n)$*,* $j = 1, \ldots, d_n$ *are the eigenvalues of* A_n*.*

The sequence $\{A_n\}_n$ *is clustered at* 1 *if it is distributed as the constant function* 1*. Finally, the sequence is properly (or strongly) clustered at* 1 *if for any* $\epsilon > 0$ *the number of the eigenvalues of* A_n *not belonging to* $(1 - \epsilon, 1 + \epsilon)$ *can be bounded by a pure constant possibly depending on* ϵ *but not on* n*.*

THEOREM 2.3. *Under the notation of section* 2 *and choosing the centered FD formulae of minimal precision order* 2*, the matrix sequence*

$$\{A_n(a, \Omega)\}_n, \quad \Omega \in \{L, Q\},$$

enjoys the following properties:

1. $A_n(a, \Omega)$ *is symmetric positive definite for any* n*;*
2. *the minimal eigenvalue of* $A_n(a, \Omega)$ *is asymptotic to* n^{-2}*;*
3. *the maximal eigenvalue of* $A_n(a, \Omega)$ *tends to* C*, where*

$$C = 8 \sup_\Omega a;$$

4. *the sequence* $\{A_n(a,\Omega)\}_n$ *is distributed as* $(f,\Omega\times(-\pi,\pi)^2)$ *in the sense of the eigenvalues, where*

$$f(x,y)=p(y)a(x),\quad x\in\Omega,\ y\in(-\pi,\pi)^2,$$

and with $p(s,t)=4-2\cos(s)-2\cos(t)$, $y=(s,t)$.

Proof. We first observe that the four statements are known in the case of $\Omega=Q$; in particular, statements 1 and 2 are really classical results [14], while statements 3 and 4 can be found in [21]. Concerning the case where $\Omega=L$ we will mainly use a four-step proof based on embedding arguments.

Proof of step 1. From the preceding discussion we know that $A_n(a,L)=\Pi A_n(a,Q)\Pi^T$ with symmetric $A_n(a,Q)$, and hence $A_n(a,L)$ is symmetric. Therefore, since $A_n(a,Q)$ is also positive definite and Π is full rank (by $\Pi\Pi^T=I$), we infer that $A_n(a,L)$ is positive definite.

Proof of step 2. Set $\hat{Q}=(0.5,1)^2$. It is evident that $\hat{Q}$ is a subdomain of L, and therefore $A_n(a,\hat{Q})$ can be seen as a principal submatrix of $A_n(a,L)$ in the sense that there exists a matrix $\hat{\Pi}$ (obtained by the identity of size $d_n(L)$ by deleting the rows related to grid points of L not belonging to $\hat{Q}$) such that

$$A_n(a,\hat{Q})=\hat{\Pi}A_n(a,L)\hat{\Pi}^T.$$

Setting $A_n(1,Q)=T_n(p)$, the Toeplitz matrix generated by $p(s,t)=4-2\cos(s)-2\cos(t)$, we have $A_n(1,\hat{Q})=T_{n/4}(p)$ and

$$\begin{array}{ccccc}\inf_Q aT_n(p) & \le & A_n(a,Q) & \le & \sup_Q aT_n(p),\\ \inf_{\hat{Q}} aT_{n/4}(p) & \le & A_n(a,\hat{Q}) & \le & \sup_{\hat{Q}} aT_{n/4}(p),\end{array}$$

where the relationship indicated by $\le$ denotes the usual partial ordering among Hermitian matrices. Under the assumptions of ellipticity and boundedness of a, we observe that $0<\inf_Q a\le\inf_{\hat{Q}} a\le\sup_{\hat{Q}} a\le\sup_Q a<\infty$ and, finally, the claimed thesis follows since

$$\lambda_{\min}(T_k(p))=8\sin^2\left(\frac{\pi}{2(k+1)}\right)\sim k^{-2}.$$

Proof of steps 3 *and* 4. Step 4 is known since $\{A_n(a,L)\}_n$ is a generalized locally Toeplitz sequence [21], while the desired limit relation in step 3 is a consequence of the following facts:

- $\{A_n(a,L)\}_n\sim\lambda(f,L\times(-\pi,\pi)^2)$ with $f(x,y)=p(y)a(x)$, $x\in L$, $y=(s,t)\in(-\pi,\pi)^2$, and $p(s,t)=4-2\cos(s)-2\cos(t)$;
- $8\sup_L a$ is the essential supremum of the function $f(x,y)=a(x)p(y)$ over the considered domain;
- we have, finally,

$$\begin{aligned}\lambda_{\max}(A_n(a,L))&\le\sup_L a\cdot\lambda_{\max}(A_n(1,L))\\&\le\sup_L a\cdot\lambda_{\max}(A_n(1,Q))\\&=\sup_L a\cdot\lambda_{\max}(T_n(p))\\&=\sup_L a\cdot 8\sin^2\left(\frac{n\pi}{2(n+1)}\right)\\&<8\sup_L a.\end{aligned}$$

□

2.2. Superlinear preconditioning. Using the results in [18] for the case of $A_n(a,Q)$, we can analyze the spectral features of $P_n^{-1}A_n$ in the case where $A_n = A_n(a,L)$ and $P_n = P_n(a,L)$. Here $P_n(a,\Omega)$, $\Omega \in \{L,Q\}$ is defined as

$$\tilde{D}_n^{1/2}(a,\Omega)A_n(1,\Omega)\tilde{D}_n^{1/2}(a,\Omega)$$

with

$$\tilde{D}_n(a,\Omega) = \frac{1}{4}\mathrm{diag}(A_n(a,\Omega)), \quad 4 = (A_n(1,\Omega))_{i,i},$$

for every $i = 1,\ldots,n^2$.

THEOREM 2.4. *Let $A_n := A_n(a,\Omega)$ and $P_n := P_n(a,\Omega)$ be the previously defined positive definite matrices with $\Omega \in \{L,Q\}$ and with the choice of the centered FD formulae of minimal precision order* 2.

1. *If the coefficient $a(x)$ is strictly positive and belongs to $\mathbf{C}^2(\overline{\Omega})$, then, for every $\epsilon > 0$, there exist an $\bar{n}$ and a constant q such that for every $n > \bar{n}$, $d_n(\Omega) - q$ eigenvalues of the preconditioned matrix $P_n^{-1}A_n$ belong to the open interval $(1-\epsilon, 1+\epsilon)$* [proper clustering]. *Moreover, all the eigenvalues belong to an interval $[c,C]$ well separated from zero* [spectral equivalence].
2. *If $a(x)$ is nonnegative, vanishes only on a finite set of curves, and is continuous over Ω, then, for every $\epsilon > 0$, there exists a function $q_n = o(d_n(\Omega))$ such that $d_n(\Omega) - q_n$ eigenvalues of the preconditioned matrix $P_n^{-1}A_n$ belong to the open interval $(1-\epsilon, 1+\epsilon)$* [weak clustering].

Proof. The statements contained in this theorem are known [18] when $\Omega = Q$.

Proof of step 1. Taking into account the relationship that links $A_n(a,Q)$ and $A_n(a,L)$, we infer the same properties from $[P_n(a,Q)]^{-1}A_n(a,Q)$ to

$$P_n^{-1}A_n := [P_n(a,L)]^{-1}A_n(a,L).$$

More specifically, by taking into account the fact that $\Pi^T\Pi$ is diagonal, we have

$$\begin{aligned}
P_n^{-1}A_n &= \left[\frac{1}{2}\mathrm{diag}^{1/2}(A_n(a,L))A_n(1,L)\frac{1}{2}\mathrm{diag}^{1/2}(A_n(a,L))\right]^{-1}\Pi A_n(a,Q)\Pi^T \\
&= \Bigg[\frac{1}{2}\Pi\mathrm{diag}^{1/2}(A_n(a,Q))\Pi^T\Pi A_n(1,Q)\Pi^T\frac{1}{2}\Pi \\
&\quad \mathrm{diag}^{1/2}(A_n(a,Q))\Pi^T\Bigg]^{-1}\Pi A_n(a,Q)\Pi^T \\
&= \Bigg[\frac{1}{2}\Pi\Pi^T\Pi\mathrm{diag}^{1/2}(A_n(a,Q))A_n(1,Q)\frac{1}{2} \\
&\quad \mathrm{diag}^{1/2}(A_n(a,Q))\Pi^T\Pi\Pi^T\Bigg]^{-1}\Pi A_n(a,Q)\Pi^T \\
&= \Bigg[\Pi\frac{1}{2}\mathrm{diag}^{1/2}(A_n(a,Q))A_n(1,Q)\frac{1}{2} \\
&\quad \mathrm{diag}^{1/2}(A_n(a,Q))\Pi^T\Bigg]^{-1}\Pi A_n(a,Q)\Pi^T \\
&= [\Pi P_n(a,Q)\Pi^T]^{-1}\Pi A_n(a,Q)\Pi^T.
\end{aligned}$$

Since Π is full rank, it is evident that the spectral behavior of

$$[\Pi P_n(a,Q)\Pi^T]^{-1}\Pi A_n(a,Q)\Pi^T$$

is, in principle, better than that of $[P_n(a,Q)]^{-1}A_n(a,Q)$. Indeed, from [18] we know that there exist positive constants $c(a,Q)$, $C(a,Q)$, and $q(a,Q,\epsilon)$ such that the spectrum of $[P_n(a,Q)]^{-1}A_n(a,Q)$ is contained in

$$[c(a,Q), C(a,Q)],$$

and the number of outliers with respect to the interval $(1-\epsilon, 1+\epsilon)$ is bounded by $q(a,Q,\epsilon)$. Finally, the use of the Cauchy interlace principle allows one to conclude the same statement for

$$[P_n(a,L)]^{-1}A_n(a,L) = [\Pi P_n(a,Q)\Pi^T]^{-1}\Pi A_n(a,Q)\Pi^T$$

with better constants $c(a,L)$, $C(a,L)$, and $q(a,L,\epsilon)$ since

$$c(a,L) \geq c(a,Q),\ C(a,L) \leq C(a,Q)$$

and

$$q(a,L,\epsilon) \leq q(a,Q,\epsilon).$$

Proof of step 2. The conclusion of step 2 follows in a completely similar manner. □

2.3. Numerical experiments. We give numerical evidences when basic L- and T-shaped domains in two dimensions with $k=1$ and $k=2$ are considered and when minimal precision formulae are employed for the discretization of $\frac{\partial}{\partial x_i}$, $i=1,2$. In the final part, we will present a comparison with other preconditioning/multigrid techniques available in the quoted literature.

2.3.1. PCG numerical results. We present the number of PCG iterations required to obtain $\|r_s\|_2/\|b\|_2 \leq 10^{-7}$ for increasing values n, and r_s denotes the residual at the sth step. The data vector either is made up of all ones or is a random vector. We consider three cases: a square domain $Q=(0,1)^2$, a basic L-shaped domain $L = Q\backslash Q'$, where $Q' = (0,1/2]^2$, and a basic T-shaped domain $T = Q\backslash(R_1 \cup R_2)$ with $R_1 = (0,1/2]\times(0,1/4]$ and $R_1 = (0,1/2]\times[3/4,1)$. The considered coefficient functions include elliptic ($a(x,y) = 1+x+y$), elliptic oscillating $a(x,y) = \sin^2(7(x+y))+1$, semielliptic ($a(x,y) = (1-x+y)^p$, $p=1,2$), and discontinuous ($a(x,y) = \exp(x+y)\mathrm{Ch}_{\{x+y\leq 2/3\}} + (2-(x+y))\mathrm{Ch}_{\{x+y>2/3\}}$) examples. The parameters are the basic ones $k=1$ and $k=2$, and the symbol Ch_X denotes the characteristic function of a set X. Looking at Tables 1 and 3 (with preconditioner $P_n(a,\Omega)$, $\Omega \in \{Q,L,T\}$), it is interesting to observe that there is no dependence of the iteration count on the domain and, strangely enough, in the semielliptic case with coefficient $a(x,y) = (1-x+y)^2$, we observe that our preconditioning technique leads to just one iteration of the PCG method for n large enough. In addition, in these examples, it is evident that our technique is really faster than the PCG with a usual incomplete Choleski (IC) factorization (refer to Table 2, where the differential problems are the same as in Table 1, but the preconditioner is the IC). In actuality, the related IC–PCG method is never optimal; in particular, we observe that the number of iterations is proportional to the square root of the size of the algebraic system. Conversely, for the preconditioner $P_n(a,\Omega)$, we have optimality, and the larger dimensions allow us to appreciate the superlinearity of the proposed preconditioning technique which leads, in many cases, to a decrease in the number of iterations as n increases (elliptic and semielliptic

TABLE 1
Number of PCG iterations in the case $k = 1$, Q, L- and T-shaped domains, with preconditioner $P_n(a, \Omega)$, exact solution xe of all ones, and random.

$k=1$ $d_n(Q)=n^2$, $d_n(L)=d_n(T)=\frac{3n^2}{4}$	$xe_i = 1$						xe_i random					
	$n+1$						$n+1$					
	16	32	64	128	256	512	16	32	64	128	256	512
$a(x,y) = 1+x+y$												
Q	3	3	3	3	3	3	3	3	2	2	2	2
L	3	3	3	3	3	3	3	3	2	2	2	2
T	3	3	3	3	3	3	3	3	2	2	2	2
$a(x,y) = \sin^2(7(x+y))+1$												
Q	10	10	10	9	9	8	9	9	9	8	7	6
L	9	9	9	8	8	8	8	8	8	7	7	6
T	9	9	9	9	8	8	9	9	8	7	7	6
$a(x,y) = 1-x+y$												
Q	4	4	4	4	4	3	4	4	3	3	3	2
L	4	4	4	4	4	3	4	4	3	3	3	2
T	4	4	4	4	4	3	4	4	3	3	3	2
$a(x,y) = (1-x+y)^2$												
Q	2	2	2	2	1	1	2	2	2	1	1	1
L	2	2	2	2	1	1	2	2	2	1	1	1
T	2	2	2	2	1	1	2	2	2	1	1	1
$a(x,y) = \exp(x+y)$ $\cdot\mathrm{Ch}_{\{x+y\le 2/3\}}$ $+(2-(x+y))$ $\cdot\mathrm{Ch}_{\{x+y>2/3\}}$												
Q	7	8	9	10	13	15	7	7	9	10	12	15
L	5	5	7	8	9	10	5	5	7	7	9	10
T	7	7	9	10	12	14	6	7	8	9	10	11

TABLE 2
Number of PCG iterations in the case $k = 1$, Q, L- and T-shaped domains, with IC preconditioner, exact solution xe of all ones, and random.

$k = 1$ $d_n(Q) = n^2$, $d_n(L) = d_n(T) = \frac{3n^2}{4}$	$xe_i = 1$						xe_i random					
	$n+1$						$n+1$					
	16	32	64	128	256	512	16	32	64	128	256	512
$a(x,y) = 1 + x + y$												
Q	16	28	53	100	196	371	16	27	48	85	155	294
L	13	23	42	80	154	300	14	24	43	74	133	251
T	15	27	50	95	186	362	14	26	47	87	158	283
$a(x,y) = \sin^2(7(x+y)) + 1$												
Q	16	29	54	104	202	391	17	30	51	95	165	300
L	13	23	44	83	159	303	14	26	43	74	139	265
T	15	27	52	99	192	367	15	26	47	89	167	290
$a(x,y) = 1 - x + y$												
Q	17	28	53	102	199	387	16	27	49	92	165	302
L	14	27	50	96	186	361	14	24	45	84	159	290
T	15	27	51	96	188	366	14	25	47	87	164	300
$a(x,y) = (1 - x + y)^2$												
Q	16	28	52	100	182	353	16	26	49	86	161	302
L	14	26	49	96	186	359	13	25	46	86	160	295
T	14	26	50	95	186	363	14	25	47	88	163	307
$a(x,y) = \exp(x+y)$ $\cdot \mathrm{Ch}_{\{x+y \le 2/3\}}$ $+(2-(x+y))$ $\cdot \mathrm{Ch}_{\{x+y > 2/3\}}$												
Q	16	28	53	102	199	385	16	29	49	93	171	324
L	13	23	44	84	162	309	14	24	44	76	143	267
T	15	27	51	98	189	370	14	26	47	89	165	298

TABLE 3
Number of PCG iterations in the case $k = 2$, Q, L- and T-shaped domains, with preconditioner $P_n(a, \Omega)$, exact solution xe of all ones, and random.

$k = 2$ $d_n(Q) = n^2$, $d_n(L) = d_n(T) = \frac{3n^2}{4}$	$xe_i = 1$						xe_i random					
	$n+1$						$n+1$					
	16	32	64	128	256	512	16	32	64	128	256	512
$a(x,y) = 1 + x + y$												
Q	3	3	2	2	2	2	3	3	2	2	2	2
L	3	2	2	2	2	2	3	2	2	2	2	2
T	3	2	2	2	2	2	3	2	2	2	2	2
$a(x,y) = \sin^2(7(x+y)) + 1$												
Q	10	12	10	9	8	8	9	9	9	8	7	6
L	8	9	8	7	7	7	8	8	8	7	7	6
T	9	10	9	8	7	7	9	9	8	7	7	6
$a(x,y) = 1 - x + y$												
Q	4	4	3	3	3	3	4	4	3	3	3	2
L	4	4	3	3	3	3	4	4	3	3	3	2
T	4	4	3	3	3	3	4	4	3	3	3	2
$a(x,y) = (1 - x + y)^2$												
Q	4	4	3	3	3	3	2	2	2	1	1	1
L	4	4	3	3	3	3	2	2	2	2	1	1
T	4	4	3	3	3	3	2	2	2	1	1	1
$a(x,y) = \exp(x+y)$ $\cdot \mathrm{Ch}_{\{x+y \le 2/3\}}$ $+(2-(x+y))$ $\cdot \mathrm{Ch}_{\{x+y>2/3\}}$												
Q	7	10	17	35	95	184	7	7	9	10	12	15
L	4	5	7	11	18	59	5	5	7	7	9	10
T	5	8	14	27	79	167	6	7	8	9	10	12

smooth examples). Finally, the presence of jumps of $a(x,y)$, or the case of a highly oscillating coefficient a, slightly deteriorates the performances of the preconditioner $P_n(a,\Omega)$. When $a(x,y)$ is discontinuous this is evident since Theorem 2.4 cannot hold (see [18, 27]). When $a(x,y)$ is smooth but highly oscillating, we observe a minor deterioration since the matrix $\tilde{D}_n(a,\Omega)$ (the diagonal part of the coefficient matrix) is given by equispaced samples of $a(x,y)$. Therefore, $\tilde{D}_n(a,\Omega)$ cannot, in general, be a faithful representation of a when a oscillates too much with regard to the grid parameter h.

2.3.2. The role of fast techniques and some comparisons with the literature. From a computational point of view it is worthwhile stressing that, in the case of plurirectangular domain Ω, the computation of the solution of the original linear system by the PCG method with preconditioner $P_n(a,\Omega)$ is reduced to the computation of the numerical solution of diagonal and two-level banded (projected) Toeplitz linear systems with nonnegative generating functions. We recall that the resolution of such linear systems can be performed within a linear arithmetic cost (linear time) by means of fast Poisson solvers, among which we count classical (direct) Poisson solvers based mainly on the cyclic reduction idea (see, e.g., [6, 7, 10, 11, 29]) and several specialized multigrid algorithms (see, e.g., [15, 5, 12, 8, 20]). Therefore, as remarked in subsection 2.3, the use of fast Poisson solvers ($a = 1$) is enough for numerically solving nonconstant coefficient PDEs: we stress that the clustering properties that hold in the elliptic case are observed in the semielliptic setting as well, even if there is a lack of adequate theoretical analysis. Finally, we mention that, in the past few years, semielliptic problems have received increasing attention from both numeric/modelistic and analytic points of view due to their occurrence in important applications, among which we mention electromagnetic field problems [16] and models in mathematical finance [30], where we encounter PDEs with a coefficient a either exploding or vanishing at the boundary of the domain.

3. The general case. In this section the goal is to demonstrate that the proposed techniques possess a natural flexibility: these techniques can be extended to higher dimensional domains, higher order FD formulae, more complicate shapes, higher order operators and, finally, different approximation methods (e.g., finite element methods). Due to a kind of "superposition of the effects," we can analyze these generalizations separately and will proceed in this way. We will stress the main points that can be extended and we will also stress the assumptions that we need. The nice discovery is that the assumptions are quite weak, while a critical point, which should be analyzed in future works, is the possibility of considering more complicated gridding strategies: this possibility will open wide the applicability of these ideas, especially in connection with variational approximations [1, 9] of the considered differential problems.

3.1. Higher dimensional problems: $d \geq 3$. The passage to higher dimensions does not pose specific problems. Indeed, both the spectral theory of Toeplitz/generalized locally Toeplitz structures and the asymptotic expansions of the form

$$\tilde{D}_n^{-1/2}(a,Q)A_n(a,Q)\tilde{D}_n^{-1/2}(a,Q) = A_n(1,Q) + h^2E_n + o(h^2), \quad a \in C^2(\bar{Q}),$$

are valid. Here E_n is a spectrally bounded, symmetric matrix having the same pattern as $A_n(a,Q)$ and $A_n(1,Q)=T_n(p)$ with

$$p(s_1,\dots,s_d)=\sum_{i=1}^{d}(2-2\cos(s_i)).$$

Therefore all the claims in Theorems 2.3 and 2.4 stand (with possibly different constants) without further assumptions.

3.2. Higher precision FD formulae. As in the case of the minimal precision order formulae, we leave the operator in divergence form and we use higher order formulae involving more discretization points [23, 27]. In this case, $A_n(1,Q)$ is still a Toeplitz matrix generated by a higher degree nonzero polynomial p: the involved polynomial is nonnegative and, by consistency [22], has a zero of order 2 at the origin. Under the additional assumption that p is strictly positive elsewhere, we easily deduce that Theorems 2.3 and 2.4 still are valid with possibly different constants.

3.3. More general domains. Let Ω be a generic plurirectangle. The assumption that we need for applying the embedding argument is the following: Calling l_i, $i=1,\dots,N_s$, $s=1,\dots,d$, the lengths of the edges of Ω parallel to the axis x_s, we assume that, for every $s=1,\dots,d$, the stepsize h_s is such that every l_i, $i=1,\dots,N_s$, is an integer multiple of h_s. In essence, the preceding one represents the hypothesis of "commensurability" (according to Pythagorean philosophy) among all the quantity l_i. In modern language, this means that we are just assuming that the numbers l_i, $i=1,\dots,N_s$, are relatively rational.

In conclusion, if this assumption is satisfied (we have a commensurable plurirectangle), then Theorems 2.3 and 2.4 are true, with L replaced by any commensurable plurirectangle Ω (with possibly different constants).

3.4. Higher order differential operators: $k\geq 2$. In the case of operators of order $2k$, $k\geq 2$, we have (see [23, 27])

$$\tilde{D}_n^{-1/2}(a,Q)A_n(a,Q)\tilde{D}_n^{-1/2}(a,Q)=A_n(1,Q)+h^2E_n+o(h^2),\quad a\in C^2(\bar{Q}).$$

The problem now is that, by consistency [22], the matrix $A_n(1,Q)$ is a Toeplitz matrix generated by a polynomial with a zero at the origin of order $2k$ with $k\geq 2$. Therefore, [4, 17] the minimal eigenvalues of $A_n(1,Q)$ are asymptotic to n^{-2k} and, consequently,

$$\{A_n^{-1}(1,Q)h^2E_n\}_n$$

is still (weakly) clustered at zero but it is not spectrally bounded. In conclusion we lose, at the same time, the guarantee of the proper clustering of the spectral equivalence. Hence, the following weakened versions of Theorems 2.3 and 2.4 hold.

THEOREM 3.1. *Under the notation of section* 2 *and choosing the centered FD formulae of minimal precision order* 2*, the matrix sequence*

$$\{A_n(a,\Omega)\}_n,\quad \Omega\in\{L,Q\},$$

enjoys the following properties:

1. $A_n(a,\Omega)$ *is symmetric positive definite for any* n*;*
2. *the minimal eigenvalue of* $A_n(a,\Omega)$ *is asymptotic to* n^{-2k}*;*

3. *the maximal eigenvalue of* $A_n(a,\Omega)$ *tends to* C*, where*

$$C = 2^{2k+1} \sup_{\Omega} a;$$

4. *the sequence* $\{A_n(a,\Omega)\}_n$ *is distributed as* $(f, \Omega \times (-\pi,\pi)^2)$ *in the sense of the eigenvalues, where*

$$f(x,y) = p(y)a(x), \quad x \in \Omega,\ y \in (-\pi,\pi)^2,$$

and with $p(s,t) = (2-2\cos(s))^k + (2-2\cos(t))^k$, $y = (s,t)$.

THEOREM 3.2. *Let* $A_n := A_n(a,\Omega)$ *and* $P_n := P_n(a,\Omega)$ *be the previously defined positive definite matrices with* $\Omega \in \{L, Q\}$ *and with the choice of the centered FD formulae of minimal precision order* 2*. If* $a(x)$ *is nonnegative, vanishes only on a finite set of curves, and is continuous over* Ω*, then for every* $\epsilon > 0$*, there exist a function* $q_n = o(d_n(\Omega))$ *such that* $n - q_n$ *eigenvalues of the preconditioned matrix* $P_n^{-1}A_n$ *belong to the open interval* $(1-\epsilon, 1+\epsilon)$ [weak clustering].

3.5. Further generalizations: The finite element method approximation. In a former paper [26], we analyzed the Toeplitz + diagonal preconditioning in connection with a uniform triangulation on a square domain and with specific types of finite elements (i.e., triangles or rectangles with linear or bilinear functions). If we restrict our attention to these basic discretization models, all the results in Theorems 2.3 and 2.4 hold.

Future work should be in the direction of nonuniform triangulation. We mention that a preliminary step in this sense has been done mainly with reference to the spectral distributional theory (see [2]).

3.6. Further generalizations: Nonsymmetric problems. Recently, the authors of [3] proposed a preconditioning technique for nonsymmetric positive definite problems arising, e.g., in the discretization of convection-diffusion differential equations. One of the main features of their method is that the whole convergence is essentially driven by the spectral properties of the preconditioned Hermitian part. Therefore, the preconditioning procedures proposed in this paper for Hermitian positive definite problems represent a key step for efficiently handling such nonsymmetric positive definite problems.

4. Conclusions. As mentioned in the introduction and section 3, our proposed ideas have a natural flexibility and embrace many different types of linear systems arising from coercive PDEs. However, there still exist important open questions, the main points of which concern the case of analogous preconditioning strategies when nonuniform meshings are involved, especially in connection with finite element method approximation techniques.

REFERENCES

[1] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press, New York, 1984.

[2] B. BECKERMANN AND S. SERRA CAPIZZANO, *Spectral Distributions of Finite Element Matrix Sequences*, manuscript, 2002.

[3] D. BERTACCINI, G. GOLUB, S. SERRA CAPIZZANO, AND C. TABLINO POSSIO, *Preconditioned HSS Method for the Solution of Non-Hermitian Positive Definite Linear Systems*, Tech. Report SCCM-02-10, Stanford University, Stanford, CA, 2002.

[4] A. BÖTTCHER AND S. GRUDSKY, *On the condition numbers of large semi-definite Toeplitz matrices*, Linear Algebra Appl., 279 (1998), pp. 285–301.

[5] J. Bramble, *Multigrid Methods*, Pitman Res. Notes in Math. Ser. 294, Longman Scientific, Harlow, UK, 1993.
[6] B. L. Buzbee, G. H. Golub, and C. W. Nielson, *On direct methods for solving Poisson's equations*, SIAM J. Numer. Anal., 7 (1970), pp. 627–656.
[7] B. L. Buzbee, F. W. Dorr, J. A. George, and G. H. Golub, *The direct solution of the discrete Poisson equation on irregular regions*, SIAM J. Numer. Anal., 8 (1971), pp. 722–736.
[8] R. H. Chan, Q.-S. Chang, and H.-W. Sun, *Multigrid method for ill-conditioned symmetric Toeplitz systems*, SIAM J. Sci. Comput., 19 (1998), pp. 516–529.
[9] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics in Appl. Math. 40, SIAM, Philadelphia, 2002.
[10] P. Concus and G. H. Golub, *Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 1103–1120.
[11] F. W. Dorr, *The direct solution of the discrete Poisson equation on a rectangle*, SIAM Rev., 12 (1970), pp. 248–263.
[12] G. Fiorentino and S. Serra, *Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions*, SIAM J. Sci. Comput., 17 (1996), pp. 1068–1081.
[13] A. Frangioni and S. Serra Capizzano, *Spectral Analysis and Preconditioning of Local Graph Matrices and Application to FD Linear Systems*, manuscript, 2002.
[14] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Elements Methods*, Cambridge University Press, Cambridge, UK, 1988.
[15] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[16] D. Marini and P. Pietra, *Mixing finite element approximation of a degenerate elliptic problem*, Numer. Math., 71 (1995), pp. 225–236.
[17] S. Serra Capizzano, *On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1998), pp. 109–129.
[18] S. Serra Capizzano, *The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems*, Numer. Math., 81 (1999), pp. 461–495.
[19] S. Serra Capizzano, *Locally X matrices, spectral distributions, preconditioning, and applications*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1354–1388.
[20] S. Serra Capizzano, *Convergence analysis of two grid methods for elliptic Toeplitz and PDEs matrix sequences*, Numer. Math., 92 (2002), pp. 433–465.
[21] S. Serra Capizzano, *Generalized locally Toeplitz sequences: Spectral analysis and applications to discretized partial differential equations*, Linear Algebra Appl., 366 (2003), pp. 371–402.
[22] S. Serra Capizzano and C. Tablino Possio, *Spectral and structural analysis of high order finite difference matrices for elliptic operators*, Linear Algebra Appl., 293 (1999), pp. 85–131.
[23] S. Serra Capizzano and C. Tablino Possio, *High-order finite difference schemes and Toeplitz based preconditioners for elliptic problems*, Electron. Trans. Numer. Anal., 11 (2000), pp. 55–84.
[24] S. Serra Capizzano and C. Tablino Possio, *Positive representation formulas for finite difference discretizations of (elliptic) second order PDEs*, Contemp. Math. 281 (2001), pp. 295–318.
[25] S. Serra Capizzano and C. Tablino Possio, *Structured preconditioning of optimal preconditioners for 2D collocation linear systems*, in Structured Matrices: Recent Developments in Theory and Computations, D. Bini, E. Tyrtyshnikov, and P. Yalamov, eds., Nova Science Publishers, Hauppauge, NY, 2001, pp. 191–204.
[26] S. Serra Capizzano and C. Tablino Possio, *Finite element matrix sequences: The case of rectangular domains*, Numer. Algorithms, 28 (2001), pp. 309–327.
[27] S. Serra Capizzano and C. Tablino Possio, *Preconditioning strategies for 2D finite difference matrix sequences*, Electron. Trans. Numer. Anal., 16 (2003), pp. 1–29.
[28] S. Serra Capizzano and C. Tablino Possio, *Analysis of preconditioning strategies for collocation linear systems*, Linear Algebra Appl., to appear.
[29] P. Swarztrauber, *The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle*, SIAM Rev., 19 (1977), pp. 490–501.
[30] P. Wilmott, S. Howison, and J. Dewynne, *The Mathematics of Financial Derivatives*, Cambridge University Press, Cambridge, UK, 1998.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 165–179

STRUCTURE PRESERVING DIMENSION REDUCTION FOR CLUSTERED TEXT DATA BASED ON THE GENERALIZED SINGULAR VALUE DECOMPOSITION*

PEG HOWLAND†, MOONGU JEON‡, AND HAESUN PARK†

Abstract. In today's vector space information retrieval systems, dimension reduction is imperative for efficiently manipulating the massive quantity of data. To be useful, this lower-dimensional representation must be a good approximation of the full document set. To that end, we adapt and extend the discriminant analysis projection used in pattern recognition. This projection preserves cluster structure by maximizing the scatter between clusters while minimizing the scatter within clusters. A common limitation of trace optimization in discriminant analysis is that one of the scatter matrices must be nonsingular, which restricts its application to document sets in which the number of terms does not exceed the number of documents. We show that by using the generalized singular value decomposition (GSVD), we can achieve the same goal regardless of the relative dimensions of the term-document matrix. In addition, applying the GSVD allows us to avoid the explicit formation of the scatter matrices in favor of working directly with the data matrix, thus improving the numerical properties of the approach. Finally, we present experimental results that confirm the effectiveness of our approach.

Key words. dimension reduction, discriminant analysis, pattern recognition, trace optimization, scatter matrix, generalized eigenvalue problem, generalized singular value decomposition, text classification

AMS subject classifications. 15A09, 68T10, 62H30, 65F15, 15A18

PII. S0895479801393666

1. Introduction.

The vector space–based information retrieval system, originated by Salton [13, 14], represents documents as vectors in a vector space. The document set comprises an $m \times n$ term-document matrix $A = (a_{ij})$, in which each column represents a document and each entry a_{ij} represents the weighted frequency of term i in document j. A major benefit of this representation is that the algebraic structure of the vector space can be exploited [1]. Modern document sets are huge [3], so we need to find a lower-dimensional representation of the data. To achieve higher efficiency in manipulating the data, it is often necessary to reduce the dimension severely. Since this may result in loss of information, we seek a representation in the lower-dimensional space that best approximates the document collection in the full space [8, 12].

The specific method we present in this paper is based on the discriminant analysis projection used in pattern recognition [4, 15]. Its goal is to find the mapping that transforms each column of A into a column in the lower-dimensional space, while preserving the cluster structure of the full data matrix. This is accomplished by

*Received by the editors August 13, 2001; accepted for publication (in revised form) by L. Eldén October 22, 2002; published electronically May 15, 2003. This research was supported in part by National Science Foundation (NSF) grant CCR-9901992. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

http://www.siam.org/journals/simax/25-1/39366.html

†Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 (howland@cs.umn.edu, hpark@cs.umn.edu). A part of this work was carried out while the third author was visiting the Korea Institute for Advanced Study, Seoul, Korea, for her sabbatical leave, from September 2001 to July 2002.

‡Department of Computer Science, University of California, Santa Barbara, CA 93106 (jeon@cs.ucsb.edu).

forming scatter matrices from A, the traces of which provide measures of the quality of the cluster relationship. After defining the optimization criterion in terms of these scatter matrices, the problem can be expressed as a generalized eigenvalue problem.

As we explain in the next section, the current discriminant analysis approach can be applied only in the case where $m \leq n$, i.e., when the number of terms does not exceed the number of documents. By recasting the generalized eigenvalue problem in terms of a related generalized singular value problem, we circumvent this restriction on the relative dimensions of A, thus extending the applicability to any data matrix. At the same time, we improve the numerical properties of the approach by working with the data matrix directly rather than forming the scatter matrices explicitly. Our algorithm follows the generalized singular value decomposition (GSVD) [2, 5, 16] as formulated by Paige and Saunders [11]. For a data matrix with k clusters, we can limit our computation to the *generalized right singular vectors* that correspond to the $k-1$ largest generalized singular values. In this way, our algorithm remains computationally simple while achieving its goal of preserving cluster structure. Experimental results demonstrating its effectiveness are described in section 5 of the paper.

2. Dimension reduction based on discriminant analysis. Given a term-document matrix $A \in \mathbb{R}^{m\times n}$, the general problem we consider is to find a linear transformation $G^T \in \mathbb{R}^{l\times m}$ that maps each column a_i, $1 \leq i \leq n$, of A in the m-dimensional space to a column y_i in the l-dimensional space:

$$G^T : a_i \in \mathbb{R}^{m\times 1} \rightarrow y_i \in \mathbb{R}^{l\times 1}. \tag{1}$$

Rather than looking for the mapping that achieves this explicitly, one may rephrase this as an approximation problem where the given matrix A is decomposed into two matrices B and Y as

$$A \approx BY, \tag{2}$$

where both $B \in \mathbb{R}^{m\times l}$ with $\text{rank}(B) = l$ and $Y \in \mathbb{R}^{l\times n}$ with $\text{rank}(Y) = l$ are to be found. Note that what we need ultimately is the lower-dimensional representation Y of the matrix A, where B and Y are both unknown. In [8, 12], methods that determine the matrix B have been presented. In those methods, after B is determined, the matrix Y is computed, for example, by solving the least squares problem [2]

$$\min_{B,Y} \|BY - A\|_F, \tag{3}$$

where B and A are given. The method we present here computes the matrix G^T directly from A without reformulating the problem as a matrix approximation problem as in (2).

Now our goal is to find a linear transformation such that the cluster structure existing in the full-dimensional space is preserved in the reduced-dimensional space, assuming that the given data are already clustered. For this purpose, first we need to formulate a measure of cluster quality. To have high cluster quality, a specific clustering result must have a tight within-cluster relationship while the between-cluster relationship has to be remote. To quantify this, in discriminant analysis [4, 15], within-cluster, between-cluster, and mixture scatter matrices are defined. For simplicity of discussion, we will assume that the given data matrix $A \in \mathbb{R}^{m\times n}$ is partitioned into k clusters as

$$A = [A_1 \quad A_2 \quad \cdots \quad A_k], \quad \text{where} \quad A_i \in \mathbb{R}^{m\times n_i}, \quad \text{and} \quad \sum_{i=1}^{k} n_i = n.$$

Let N_i denote the set of column indices that belong to the cluster i. The centroid $c^{(i)}$ of each cluster A_i is computed by taking the average of the columns in A_i, i.e.,

$$c^{(i)} = \frac{1}{n_i} A_i e^{(i)}, \quad \text{where} \quad e^{(i)} = (1, \ldots, 1)^T \in \mathbb{R}^{n_i \times 1},$$

and the global centroid is

$$c = \frac{1}{n} Ae, \quad \text{where} \quad e = (1, \ldots, 1)^T \in \mathbb{R}^{n \times 1}.$$

Then the within-cluster scatter matrix S_w is defined as

$$S_w = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T,$$

and the between-cluster scatter matrix S_b is defined as

$$\begin{aligned} S_b &= \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)(c^{(i)} - c)^T \\ &= \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T. \end{aligned}$$

Finally, the mixture scatter matrix is defined as

$$S_m = \sum_{j=1}^{n} (a_j - c)(a_j - c)^T.$$

It is easy to show [7] that the scatter matrices have the relationship

$$S_m = S_w + S_b. \tag{4}$$

Writing $a_j - c = a_j - c^{(i)} + c^{(i)} - c$ for $j \in N_i$, we have

$$S_m = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)} + c^{(i)} - c)(a_j - c^{(i)} + c^{(i)} - c)^T \tag{5}$$

$$= \sum_{i=1}^{k} \sum_{j \in N_i} [(a_j - c^{(i)})(a_j - c^{(i)})^T + (c^{(i)} - c)(c^{(i)} - c)^T] \tag{6}$$

$$+ \sum_{i=1}^{k} \sum_{j \in N_i} [(a_j - c^{(i)})(c^{(i)} - c)^T + (c^{(i)} - c)(a_j - c^{(i)})^T]. \tag{7}$$

This gives the relation (4), since each inner sum in (7) is zero.

Defining the matrices,

$$H_w = [A_1 - c^{(1)} e^{(1)^T}, A_2 - c^{(2)} e^{(2)^T}, \ldots, A_k - c^{(k)} e^{(k)^T}] \in \mathbb{R}^{m \times n}, \tag{8}$$

$$H_b = [\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c)] \in \mathbb{R}^{m \times k}, \tag{9}$$

and

$$H_m = [a_1 - c, \ldots, a_n - c] = A - ce^T \in \mathbb{R}^{m \times n}, \tag{10}$$

the scatter matrices can be expressed as

$$S_w = H_w H_w^T, \quad S_b = H_b H_b^T, \quad \text{and} \quad S_m = H_m H_m^T. \tag{11}$$

Note that another way to define H_b is

$$H_b = [(c^{(1)} - c)e^{(1)^T}, (c^{(2)} - c)e^{(2)^T}, \ldots, (c^{(k)} - c)e^{(k)^T}] \in \mathbb{R}^{m \times n},$$

but using the lower-dimensional form in (9) reduces the storage requirements and computational complexity of our algorithm.

Now, trace(S_w), which is

$$\text{trace}(S_w) = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})^T (a_j - c^{(i)}) = \sum_{i=1}^{k} \sum_{j \in N_i} \|a_j - c^{(i)}\|_2^2, \tag{12}$$

provides a measure of the closeness of the columns within the clusters over all k clusters, and trace(S_b), which is

$$\text{trace}(S_b) = \sum_{i=1}^{k} \sum_{j \in N_i} (c^{(i)} - c)^T (c^{(i)} - c) = \sum_{i=1}^{k} \sum_{j \in N_i} \|c^{(i)} - c\|_2^2, \tag{13}$$

provides a measure of the distance between clusters. When items within each cluster are located tightly around their own cluster centroid, then trace(S_w) will have a small value. On the other hand, when the between-cluster relationship is remote, and hence the centroids of the clusters are remote, trace(S_b) will have a large value. Using the values trace(S_w), trace(S_b), and relationship (4), the cluster quality can be measured. In general, when trace(S_b) is large while trace(S_w) is small, or trace(S_m) is large while trace(S_w) is small, we expect the clusters of different classes to be well separated and the items within each cluster to be tightly related, and therefore the cluster quality will be high. There are several measures of cluster quality which involve the three scatter matrices [4, 15], including

$$J_1 = \text{trace}(S_w^{-1} S_b) \tag{14}$$

and

$$J_2 = \text{trace}(S_w^{-1} S_m). \tag{15}$$

Note that both of the above criteria require S_w to be nonsingular or, equivalently, H_w to have full rank. For more measures of cluster quality, their relationships, and their extension to document data, see [6].

In the lower-dimensional space obtained from the linear transformation G^T, the within-cluster, between-cluster, and mixture scatter matrices become

$$S_w^Y = \sum_{i=1}^{k} \sum_{j \in N_i} (G^T a_j - G^T c^{(i)})(G^T a_j - G^T c^{(i)})^T = G^T S_w G,$$

$$S_b^Y = \sum_{i=1}^{k} \sum_{j \in N_i} (G^T c^{(i)} - G^T c)(G^T c^{(i)} - G^T c)^T = G^T S_b G,$$

$$S_m^Y = \sum_{j=1}^{n} (G^T a_j - G^T c)(G^T a_j - G^T c)^T = G^T S_m G,$$

where the superscript Y denotes values in the l-dimensional space. Given k clusters in the full dimension, the linear transformation G^T that best preserves this cluster structure in the reduced dimension would maximize $\text{trace}(S_b^Y)$ and minimize $\text{trace}(S_w^Y)$. We can approximate this simultaneous optimization using measure (14) or (15) by looking for the matrix G that maximizes

$$J_1(G) = \text{trace}((G^T S_w G)^{-1}(G^T S_b G))$$

or

$$J_2(G) = \text{trace}((G^T S_w G)^{-1}(G^T S_m G)).$$

For computational reasons, we will focus our discussion on the criterion of maximizing J_1. Although J_1 is a less obvious choice than the quotient

$$\text{trace}(G^T S_b G)/\text{trace}(G^T S_w G),$$

it is formulated to be invariant under nonsingular linear transformations, a property that will prove useful below.

When $S_w = H_w H_w^T$ is assumed to be nonsingular, it is symmetric positive definite. According to results from the symmetric-definite generalized eigenvalue problem [5], there exists a nonsingular matrix $X \in \mathbb{R}^{m\times m}$ such that

$$X^T S_b X = \Lambda = \text{diag}(\lambda_1 \dots \lambda_m) \quad \text{and} \quad X^T S_w X = I_m.$$

Letting x_i denote the ith column of X, we have

$$S_b x_i = \lambda_i S_w x_i, \tag{16}$$

which means that λ_i and x_i are an eigenvalue-eigenvector pair of $S_w^{-1} S_b$, and

$$\text{trace}(S_w^{-1} S_b) = \lambda_1 + \cdots + \lambda_m.$$

Expressing (16) in terms of H_b and H_w and premultiplying by x_i^T, we see that

$$\|H_b^T x_i\|_2^2 = \lambda_i \|H_w^T x_i\|_2^2. \tag{17}$$

Hence $\lambda_i \geq 0$ for $1 \leq i \leq m$.

The definition of H_b in (9) implies that $\text{rank}(H_b) \leq k-1$. Accordingly, $\text{rank}(S_b) \leq k-1$, and only the largest $k-1$ λ_i's can be nonzero. In addition, by using a permutation matrix to order Λ (and likewise X), we can assume that $\lambda_1 \geq \cdots \geq \lambda_{k-1} \geq \lambda_k = \cdots = \lambda_m = 0$.

We have

$$\begin{aligned} J_1(G) &= \text{trace}((S_w^Y)^{-1} S_b^Y) \\ &= \text{trace}((G^T X^{-T} X^{-1} G)^{-1} G^T X^{-T} \Lambda X^{-1} G) \\ &= \text{trace}((\tilde{G}^T \tilde{G})^{-1} \tilde{G}^T \Lambda \tilde{G}), \end{aligned}$$

where $\tilde{G} = X^{-1} G$. The matrix $\tilde{G}$ has full column rank provided G does, so it has the reduced QR factorization $\tilde{G} = QR$, where $Q \in \mathbb{R}^{m\times l}$ has orthonormal columns and R is nonsingular. Hence

$$\begin{aligned} J_1(G) &= \text{trace}((R^T R)^{-1} R^T Q^T \Lambda Q R) \\ &= \text{trace}(R^{-1} Q^T \Lambda Q R) \\ &= \text{trace}(Q^T \Lambda Q R R^{-1}) \\ &= \text{trace}(Q^T \Lambda Q). \end{aligned}$$

This shows that once we have diagonalized, the maximization of $J_1(G)$ depends only on an orthonormal basis for range$(X^{-1}G)$; i.e.,

$$\begin{aligned}\max_G J_1(G) &= \max_{Q^TQ=I} \operatorname{trace}(Q^T\Lambda Q)\\ &\le \lambda_1 + \cdots + \lambda_{k-1} = \operatorname{trace}(S_w^{-1}S_b).\end{aligned}$$

When $l \ge k-1$, this upper bound on $J_1(G)$ is achieved for

$$Q = \begin{pmatrix} I_l \\ 0 \end{pmatrix} \quad \text{or} \quad G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix} R.$$

Note that the transformation G is not unique in the sense that $J_1(G) = J_1(GW)$ for any nonsingular matrix $W \in \mathbb{R}^{l\times l}$ since

$$\begin{aligned}J_1(GW) &= \operatorname{trace}((W^TG^TS_wGW)^{-1}(W^TG^TS_bGW))\\ &= \operatorname{trace}(W^{-1}(G^TS_wG)^{-1}W^{-T}W^T(G^TS_bG)W)\\ &= \operatorname{trace}((G^TS_wG)^{-1}(G^TS_bG)WW^{-1}) = J_1(G).\end{aligned}$$

Hence, the maximum $J_1(G)$ is also achieved for

$$G = X \begin{pmatrix} I_l \\ 0 \end{pmatrix}.$$

This means that

$$\operatorname{trace}((S_w^Y)^{-1}S_b^Y) = \operatorname{trace}(S_w^{-1}S_b)$$

whenever $G \in \mathbb{R}^{m\times l}$ consists of l eigenvectors of $S_w^{-1}S_b$ corresponding to the l largest eigenvalues. Therefore, if we choose $l = k-1$, dimension reduction results in no loss of cluster quality as measured by J_1.

Now, a limitation of the criterion $J_1(G)$ in many applications, including text processing in information retrieval, is that the matrix S_w must be nonsingular. For S_w to be nonsingular, we can allow only the case $m \le n$, since S_w is the product of an $m \times n$ matrix, H_w, and an $n \times m$ matrix, H_w^T. In other words, the number of terms cannot exceed the number of documents, which is a severe restriction. We seek a solution which does not impose this restriction, and which can be found without explicitly forming S_b and S_w from H_b and H_w, respectively. Toward that end, we use (17) to express λ_i as α_i^2/β_i^2, and the problem (16) becomes

$$\beta_i^2 H_b H_b^T x_i = \alpha_i^2 H_w H_w^T x_i. \tag{18}$$

(λ_i will be infinite when $\beta_i = 0$, as we discuss later.) This has the form of a problem that can be solved using the GSVD [5, 11, 16], as described in the next section.

3. GSVD. The following theorem introduces the GSVD as was originally defined by Van Loan [16].

THEOREM 1. *Suppose two matrices $K_A \in \mathbb{R}^{m\times n}$ with $m \ge n$ and $K_B \in \mathbb{R}^{p\times n}$ are given. Then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{p\times p}$ and a nonsingular matrix $X \in \mathbb{R}^{n\times n}$ such that*

$$U^TK_AX = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \quad \text{and} \quad V^TK_BX = \operatorname{diag}(\beta_1, \ldots, \beta_q),$$

where $q = \min(p, n)$, $\alpha_i \ge 0$ for $1 \le i \le n$, and $\beta_i \ge 0$ for $1 \le i \le q$.

This formulation cannot be applied to the matrix pair K_A and K_B when the dimensions of K_A do not satisfy the assumed restrictions. Paige and Saunders [11] developed a more general formulation which can be defined for any two matrices with the same number of columns. We restate theirs as follows.

THEOREM 2. *Suppose two matrices $K_A \in \mathbb{R}^{m\times n}$ and $K_B \in \mathbb{R}^{p\times n}$ are given. Then for*

$$K = \begin{pmatrix} K_A \\ K_B \end{pmatrix} \quad \textit{and} \quad t = \operatorname{rank}(K),$$

there exist orthogonal matrices

$$U \in \mathbb{R}^{m\times m}, \quad V \in \mathbb{R}^{p\times p}, \quad W \in \mathbb{R}^{t\times t}, \quad \textit{and} \quad Q \in \mathbb{R}^{n\times n}$$

such that

$$U^T K_A Q = \Sigma_A(\underbrace{W^T R}_{t}, \underbrace{0}_{n-t}) \quad \textit{and} \quad V^T K_B Q = \Sigma_B(\underbrace{W^T R}_{t}, \underbrace{0}_{n-t}),$$

where

$$\underset{m\times t}{\Sigma_A} = \begin{pmatrix} I_A & & \\ & D_A & \\ & & 0_A \end{pmatrix}, \quad \underset{p\times t}{\Sigma_B} = \begin{pmatrix} O_B & & \\ & D_B & \\ & & I_B \end{pmatrix}, \tag{19}$$

and $R \in \mathbb{R}^{t\times t}$ is nonsingular with its singular values equal to the nonzero singular values of K. The matrices

$$I_A \in \mathbb{R}^{r\times r} \quad \textit{and} \quad I_B \in \mathbb{R}^{(t-r-s)\times(t-r-s)}$$

are identity matrices, where the values of r and s depend on the data,

$$0_A \in \mathbb{R}^{(m-r-s)\times(t-r-s)} \quad \textit{and} \quad 0_B \in \mathbb{R}^{(p-t+r)\times r}$$

are zero matrices with possibly no rows or no columns, and

$$D_A = \operatorname{diag}(\alpha_{r+1}, \ldots, \alpha_{r+s}) \quad \textit{and} \quad D_B = \operatorname{diag}(\beta_{r+1}, \ldots, \beta_{r+s})$$

satisfy

$$1 > \alpha_{r+1} \geq \cdots \geq \alpha_{r+s} > 0, \quad 0 < \beta_{r+1} \leq \cdots \leq \beta_{r+s} < 1, \tag{20}$$

and

$$\alpha_i^2 + \beta_i^2 = 1 \quad \textit{for } i = r+1, \ldots, r+s.$$

Paige and Saunders gave a constructive proof of Theorem 2, which starts with the complete orthogonal decomposition [5, 2, 10] of K, or

$$P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}, \tag{21}$$

where P and Q are orthogonal and R is nonsingular with the same rank as K. The construction proceeds by exploiting the SVDs of submatrices of P. Partitioning P as

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \quad \text{where} \quad P_{11} \in \mathbb{R}^{m\times t} \quad \text{and} \quad P_{21} \in \mathbb{R}^{p\times t},$$

implies $\|P_{11}\|_2 \leq 1$. This means that the singular values of P_{11} do not exceed one, so its SVD can be written as $U^T P_{11} W = \Sigma_A$, where $U \in \mathbb{R}^{m\times m}$ and $W \in \mathbb{R}^{t\times t}$ are orthogonal and Σ_A has the form in (19). Next $P_{21}W$ is decomposed as $P_{21}W = VL$, where $V \in \mathbb{R}^{p\times p}$ is orthogonal and $L = (l_{ij}) \in \mathbb{R}^{p\times t}$ is lower triangular with $l_{ij} = 0$ if $p - i > t - j$ and $l_{ij} \geq 0$ if $p - i = t - j$. This triangularization can be accomplished in the same way as QR decomposition except that columns are annihilated above the diagonal $p - i = t - j$, working from right to left. Then the matrix

$$\begin{pmatrix} \Sigma_A \\ L \end{pmatrix}$$

has orthonormal columns, which implies that $L = \Sigma_B$. These results can be combined with (21) to obtain

$$\begin{pmatrix} K_A \\ K_B \end{pmatrix} Q = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} P_{11}R & 0 \\ P_{21}R & 0 \end{pmatrix} = \begin{pmatrix} U\Sigma_A W^T R & 0 \\ V\Sigma_B W^T R & 0 \end{pmatrix},$$

which completes the proof. In [11], this form of GSVD is related to that of Van Loan by

$$U^T K_A X = (\Sigma_A, 0) \quad \text{and} \quad V^T K_B X = (\Sigma_B, 0), \tag{22}$$

where

$$\underset{n\times n}{X} = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}.$$

From the form in (22) we see that

$$K_A = U(\Sigma_A, 0)X^{-1} \quad \text{and} \quad K_B = V(\Sigma_B, 0)X^{-1},$$

which imply that

$$K_A^T K_A = X^{-T} \begin{pmatrix} \Sigma_A^T \Sigma_A & 0 \\ 0 & 0 \end{pmatrix} X^{-1} \quad \text{and} \quad K_B^T K_B = X^{-T} \begin{pmatrix} \Sigma_B^T \Sigma_B & 0 \\ 0 & 0 \end{pmatrix} X^{-1}.$$

Defining

$$\alpha_i = 1,\ \beta_i = 0 \quad \text{for } i = 1, \ldots, r$$

and

$$\alpha_i = 0,\ \beta_i = 1 \quad \text{for } i = r + s + 1, \ldots, t,$$

we have, for $1 \leq i \leq t$,

$$\beta_i^2 K_A^T K_A x_i = \alpha_i^2 K_B^T K_B x_i, \tag{23}$$

where x_i represents the ith column of X. For the remaining $n - t$ columns of X, both $K_A^T K_A x_i$ and $K_B^T K_B x_i$ are zero, so (23) is satisfied for arbitrary values of α_i and β_i when $t + 1 \leq i \leq n$. Therefore, the columns of X are the generalized right singular vectors for the matrix pair K_A and K_B.

In terms of the generalized singular values, or the α_i/β_i quotients, r of them are infinite, s are finite and nonzero, and $t - r - s$ are zero. To determine the number of

generalized singular values of each type, we write explicit expressions for the values of r and s. From (22) and (19), we see that

$$\operatorname{rank}(K_A) = r + s \quad \text{and} \quad \operatorname{rank}(K_B) = t - r.$$

Hence, the number of infinite generalized singular values is

$$r = \operatorname{rank}\begin{pmatrix} K_A \\ K_B \end{pmatrix} - \operatorname{rank}(K_B)$$

and the number of finite and nonzero generalized singular values is

$$s = \operatorname{rank}(K_A) + \operatorname{rank}(K_B) - \operatorname{rank}\begin{pmatrix} K_A \\ K_B \end{pmatrix}.$$

4. Application of the GSVD to dimension reduction. Recall that for the $m \times n$ term-document matrix A, when $m \leq n$ and the scatter matrix S_w is nonsingular, a criterion such as maximization of J_1 can be applied. However, one drawback of this criterion is that both $S_w = H_w H_w^T$ and $S_b = H_b H_b^T$ must be explicitly formed. Forming these cross-product matrices can cause a loss of information [5, p. 239, Example 5.3.2], but by using the GSVD, which works directly with H_w and H_b, we can avoid a potential numerical problem.

Applying the GSVD to the nonsingular case, we include in G those x_i's which correspond to the $k-1$ largest λ_i's, where $\lambda_i = \alpha_i^2/\beta_i^2$. When the GSVD construction orders the singular value pairs as in (20), the generalized singular values, or the α_i/β_i quotients, are in nonincreasing order. Therefore, the first $k-1$ columns of X are all we need. Our algorithm first computes the matrices H_b and H_w from the term-document matrix A. We then solve for a very limited portion of the GSVD of the matrix pair H_b^T and H_w^T. This solution is accomplished by following the construction in the proof of Theorem 2. The major steps are limited to the complete orthogonal decomposition of $K = (H_b, H_w)^T$, which produces orthogonal matrices P and Q and a nonsingular matrix R, followed by the SVD of a leading principal submatrix of P. The steps are summarized in Algorithm LDA/GSVD, where LDA stands for linear discriminant analysis.

When $m > n$, the scatter matrix S_w is singular. Hence, we cannot even define the J_1 criterion, and discriminant analysis fails. Consider a generalized right singular vector x_i that lies in the null space of S_w. From (18), we see that either x_i also lies in the null space of S_b or the corresponding β_i equals zero. We will discuss each of these cases in terms of the simultaneous optimization

$$\max_G \operatorname{trace}(G^T S_b G) \quad \text{and} \quad \min_G \operatorname{trace}(G^T S_w G) \tag{24}$$

that criterion J_1 is approximating.

When $x_i \in \operatorname{null}(S_w) \cap \operatorname{null}(S_b)$, (18) is satisfied for arbitrary values of α_i and β_i. As explained in section 3, this will be the case for the rightmost $m-t$ columns of X. To determine whether these columns should be included in G, consider

$$\operatorname{trace}(G^T S_b G) = \sum g_j^T S_b g_j \quad \text{and} \quad \operatorname{trace}(G^T S_w G) = \sum g_j^T S_w g_j,$$

where g_j represents a column of G. Adding the column x_i to G has no effect on these traces, since $x_i^T S_w x_i = 0$ and $x_i^T S_b x_i = 0$, and therefore does not contribute to

Algorithm 1 LDA/GSVD.

Given a data matrix $A \in \mathbb{R}^{m\times n}$ with k clusters, it computes the columns of the matrix $G \in \mathbb{R}^{m\times(k-1)}$, which preserves the cluster structure in the reduced-dimensional space, and it also computes the $(k-1)$-dimensional representation Y of A.

1. Compute $H_b \in \mathbb{R}^{m\times k}$ and $H_w \in \mathbb{R}^{m\times n}$ from A according to (9) and (8), respectively.
2. Compute the complete orthogonal decomposition of $K = (H_b, H_w)^T \in \mathbb{R}^{(k+n)\times m}$, which is
$$P^T K Q = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}.$$
3. Let $t = \text{rank}(K)$.
4. Compute W from the SVD of $P(1:k, 1:t)$, which is $U^T P(1:k, 1:t)W = \Sigma_A$.
5. Compute the first $k-1$ columns of
$$X = Q \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$$
and assign them to G.
6. $Y = G^T A$.

either maximization or minimization in (24). For this reason, we do not include these columns of X in our solution.

When $x_i \in \text{null}(S_w) - \text{null}(S_b)$, then $\beta_i = 0$. As discussed in section 3, this implies that $\alpha_i = 1$, and hence that the generalized singular value α_i/β_i is infinite. The leftmost columns of X will correspond to these. Including these columns in G increases $\text{trace}(G^T S_b G)$ while leaving $\text{trace}(G^T S_w G)$ unchanged. We conclude that, even when S_w is singular, the rule regarding which columns of X to include in G should remain the same as for the nonsingular case. Our experiments show that Algorithm LDA/GSVD works very well when S_w is singular, thus extending its applicability beyond that of the original discriminant analysis.

In terms of the matrix pair H_b^T and H_w^T, the columns of X correspond to the generalized singular values as follows. The first
$$r = \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} - \text{rank}(H_w^T)$$
columns correspond to infinite values and the next
$$s = \text{rank}(H_b^T) + \text{rank}(H_w^T) - \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix}$$
columns correspond to finite and nonzero values. The following
$$t - r - s = \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} - \text{rank}(H_b^T)$$
columns correspond to zero values and the last
$$m - t = m - \text{rank}\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix}$$

Algorithm 2 Centroid.

Given a data matrix $A \in \mathbb{R}^{m\times n}$ with k clusters, it computes a k-dimensional representation Y of A.

1. Compute the centroid $c^{(i)}$ of the ith cluster, $1 \le i \le k$.
2. Set $C = \begin{pmatrix} c^{(1)} & c^{(2)} & \cdots & c^{(k)} \end{pmatrix}$.
3. Solve $\min_Y \|CY - A\|_F$.

Algorithm 3 Orthogonal Centroid.

Given a data matrix $A \in \mathbb{R}^{m\times n}$ with k clusters, it computes a k-dimensional representation Y of A.

1. Compute the centroid $c^{(i)}$ of the ith cluster, $1 \le i \le k$.
2. Set $C = \begin{pmatrix} c^{(1)} & c^{(2)} & \cdots & c^{(k)} \end{pmatrix}$.
3. Compute the reduced QR decomposition of C, which is $C = Q_k R$.
4. Solve $\min_Y \|Q_k Y - A\|_F$ (in fact, $Y = Q_k^T A$).

columns correspond to the arbitrary values. If S_w is nonsingular, both $r = 0$ and $m-t = 0$, so $s = \text{rank}(H_b^T)$ generalized singular values are finite and nonzero, and the rest are zero. In either case, G should be comprised of the leftmost $r + s = \text{rank}(H_b^T)$ columns of X.

Assuming the centroids are linearly independent, we see from (9) that $\text{rank}(H_b)$ is $k - 1$, so Algorithm LDA/GSVD includes the minimum number of columns in G that are necessary to preserve the cluster structure after dimension reduction. If $\text{rank}(H_b) < k - 1$, then including extra columns in G (some which correspond to the $t - r - s$ zero generalized singular values and, possibly, some which correspond to the arbitrary generalized singular values) will have approximately no effect on cluster preservation.

5. Experimental results. We compare classification results in the full-dimensional space with those in the reduced-dimensional space using Algorithm LDA/GSVD and two other dimension reduction algorithms we have developed, namely, Algorithms Centroid and Orthogonal Centroid [8, 12]. The latter two algorithms assume that the centroids are linearly independent, an assumption for which we have encountered no counterexample in practice. As outlined in Algorithms 2 and 3, centroid and orthogonal centroid solve the same least squares problem (3) for different choices of B. The centroid method chooses the k cluster centroids as the columns of B, whereas orthogonal centroid chooses an orthonormal basis for the cluster centroids.

We employ both a centroid-based classification method and a nearest neighbor classification method [15], which are presented in Algorithms 4 and 5. For the full data matrix A, we apply the classification method with each column of A as the vector q and report the percentage that are misclassified. Likewise, for each dimension reduction method, we apply the classification method to the lower-dimensional representation Y of A. In addition, the quality of classification is assessed by examining traces of the within-class scatter matrix S_w and the between-class scatter matrix S_b.

Two different data types are used to verify the effectiveness of LDA/GSVD. In the first data type, the column dimension of the term-document matrix is higher than the row dimension. This can be dealt with by using the original J_1 criterion, assuming that S_w is nonsingular. In the second data type, the row dimension is higher than the column dimension, so S_w is singular. This means that neither criterion J_1 nor

Algorithm 4 Centroid-Based Classification.

Given a data matrix A with k clusters and k corresponding centroids, $c^{(i)}$, $1 \leq i \leq k$, it finds the index j of the cluster in which the vector q belongs.

- Find the index j such that $sim(q, c^{(i)})$, $1 \leq i \leq k$, is minimum (or maximum), where $sim(q, c^{(i)})$ is the similarity measure between q and $c^{(i)}$. (For example, $sim(q, c^{(i)}) = \|q - c^{(i)}\|_2$ using the L_2 norm, and we take the index with the minimum value. Using the cosine measure,

$$sim(q, c^{(i)}) = \cos(q, c^{(i)}) = \frac{q^T c^{(i)}}{\|q\|_2 \|c^{(i)}\|_2},$$

  and we take the index with the maximum value.)

Algorithm 5 k Nearest Neighbor (knn) Classification.

Given a data matrix $A = [a_1, \ldots, a_n]$ with k clusters, it finds the cluster in which the vector q belongs.

1. From the similarity measure $sim(q, a_j)$ for $1 \leq j \leq n$, find the k^* nearest neighbors of q. (We use k^* to distinguish the algorithm parameter from the number of clusters.)
2. Among these k^* vectors, count the number belonging to each cluster.
3. Assign q to the cluster with the greatest count in the previous step.

J_2 can be applied, but the dimension can be reduced very effectively using our new LDA/GSVD algorithm.

For the first data type, in Test I we use clustered data that are artificially generated by an algorithm adapted from [7, Appendix H]. Table 1 shows the dimensions of the term-document matrix and classification results using the L_2 norm similarity measure. The data consist of 2000 150-dimensional documents with seven clusters. Algorithm LDA/GSVD reduces the dimension from 150 to $k - 1 = 6$, where k is the number of classes. The other methods reduce it to $k = 7$. In Table 1, we also present the results obtained by using the LDA/GSVD algorithm to reduce the dimension to $k - 2 = 5$ and $k = 7$, which are one less than and one greater than the theoretical optimum of $k - 1$, respectively. The results confirm that the theoretical optimum does indeed maximize trace$((S_w^Y)^{-1} S_b^Y)$, and that its value is preserved exactly from the full dimension. In addition, using LDA/GSVD to reduce the dimension to $k - 1$ results in the lowest misclassification rates for both centroid-based and nearest neighbor methods. All three dimension reduction methods produce classification results that are, with one exception, at least as good as the results from the full space. This is remarkable in light of the fact that the row dimension was reduced from 150 to at most 7.

As mentioned in section 2, in a higher quality cluster structure, we will have a smaller value for trace(S_w) and a larger value for trace(S_b). With this in mind, the ratio trace(S_b)/trace(S_w) is another measure of how well trace$(G^T S_b G)$ is maximized while trace$(G^T S_w G)$ is minimized in the reduced space. We observe in Table 1 that the ratio produced by each of the three dimension reduction methods is greater than that of the full-dimensional data. This may explain why, in general, our dimension reduction methods give better classification results than those produced in the full-dimensional space.

TABLE 1
Test I: *Traces and misclassification rates (in %) with L_2 norm similarity.*

Method	Full	Orthogonal centroid	Centroid	LDA/GSVD		
Dim	150×2000	7×2000	7×2000	5×2000	6×2000	7×2000
$\text{trace}(S_w)$	299750	14238	942.3	1.6	2.0	3.0
$\text{trace}(S_b)$	23225	23225	1712	3.4	4.0	4.0
$\frac{\text{trace}(S_b)}{\text{trace}(S_w)}$	0.078	1.63	1.82	2.2	2.0	1.3
$\text{trace}(S_w^{-1} S_b)$	12.3	11.42	11.42	11.0	12.3	12.3
centroid	2.8	2.8	3.2	4.6	2.6	2.6
5nn	20.5	3.3	3.5	5.3	3.0	3.1
15nn	10.2	3.1	3.2	4.6	2.5	2.8
50nn	6.3	3.0	3.4	4.2	2.7	2.8

As proved in our previous work [8], the misclassification rates obtained using the centroid-based classification algorithm in the full space and in the orthogonal centroid-reduced space are identical. It is interesting to observe that the values of $\text{trace}(S_b)$ in these two spaces are also identical, although the motivation for the orthogonal centroid algorithm was not the preservation of $\text{trace}(S_b)$ after dimension reduction. We state this result in the following theorem.

THEOREM 3. *Let $Q_k \in \mathbb{R}^{m \times k}$ be the matrix with orthonormal columns in the reduced QR decomposition of the matrix $C \in \mathbb{R}^{m \times k}$ whose columns are the k centroids (see Algorithm Orthogonal Centroid). Then* $\text{trace}(S_b) = \text{trace}(Q_k^T S_b Q_k) = \text{trace}(S_b^Y)$, *where $Y = Q_k^T A$.*

Proof. There is an orthogonal matrix $\mathrm{Q} \in \mathbb{R}^{m \times m}$ such that

$$C = Q \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where $R \in \mathbb{R}^{k \times k}$ is upper triangular. Partitioning Q as $Q = (Q_k, \hat{Q})$, we have

$$C = (Q_k, \hat{Q}) \begin{pmatrix} R \\ 0 \end{pmatrix} = Q_k R. \tag{25}$$

Premultiplying (25) by $(Q_k, \hat{Q})^T$ gives $Q_k^T C = R$ and $\hat{Q}^T C = 0$. Therefore,

$$\begin{aligned} \text{trace}(S_b) &= \text{trace}(Q^T Q S_b) \\ &= \text{trace}(Q^T S_b Q) \\ &= \text{trace}((Q_k, \hat{Q})^T H_b H_b^T (Q_k, \hat{Q})) \\ &= \text{trace}(Q_k^T H_b H_b^T Q_k) \\ &= \text{trace}(Q_k^T S_b Q_k), \end{aligned}$$

where $H_b = [\sqrt{n_1}(c^{(1)} - c), \sqrt{n_2}(c^{(2)} - c), \ldots, \sqrt{n_k}(c^{(k)} - c)]$ and $\hat{Q}^T H_b = 0$, since $\hat{Q}^T c^{(i)} = 0$ and c is a linear combination of the $c^{(i)}$'s. ∎

In Test II, for the second data type, we use five categories of abstracts from the MEDLINE[1] database. Each category has 40 documents. The total number of terms is 7519 (see Table 2) after preprocessing with stopping and stemming algorithms [9]. For this 7519×200 term-document matrix, the original discriminant analysis breaks down, since S_w is singular. However, our improved LDA/GSVD method circumvents this singularity problem.

[1] http://www.ncbi.nlm.nih.gov/PubMed

TABLE 2
Medline data set for Test II.

Class	Data from MEDLINE	
	Category	No. of documents
1	heart attack	40
2	colon cancer	40
3	diabetes	40
4	oral cancer	40
5	tooth decay	40
	dimension	7519×200

TABLE 3
Test II: *Traces and misclassification rate with L_2 norm similarity.*

Method		Full	Orthogonal centroid	Centroid	LDA/GSVD
Dim		7519×200	5×200	5×200	4×200
Trace values	$\text{trace}(S_w)$	73048	4210	90	0.05
	$\text{trace}(S_b)$	6229	6229	160	3.95
	$\frac{\text{trace}(S_b)}{\text{trace}(S_w)}$	0.09	1.5	1.8	79
Misclassification rate in %	centroid	5	5	2	1
	1nn	40	3	2.5	1

By Algorithm LDA/GSVD the dimension 7519 is dramatically reduced to 4, which is one less than the number of classes. The other methods reduce the dimension to the number of classes, which is 5. Table 3 shows classification results using the L_2 norm similarity measure. As in the results of Test I, LDA/GSVD produces the lowest misclassification rate using both classification methods. Because the J_1 criterion is not defined in this case, we compute the ratio $\text{trace}(S_b)/\text{trace}(S_w)$ as an approximate optimality measure. We observe that the ratio is strikingly higher for the LDA/GSVD reduction than for the other methods, and that, once again, the ratio produced by each of the three dimension reduction methods is greater than that of the full-dimensional data.

6. Conclusion. Our experimental results verify that the J_1 criterion, when applicable, effectively optimizes classification in the reduced-dimensional space, while our LDA/GSVD extends the applicability to cases which the original discriminant analysis cannot handle. In addition, our LDA/GSVD algorithm avoids the numerical problems inherent in explicitly forming the scatter matrices.

In terms of computational complexity, the most expensive part of Algorithm LDA/GSVD is step 2, where a complete orthogonal decomposition is needed. Assuming $k \leq n$, $t \leq m$, and $t = \mathcal{O}(n)$, the complete orthogonal decomposition of K costs $\mathcal{O}(nmt)$ when $m \leq n$, and $\mathcal{O}(m^2 t)$ when $m > n$. Therefore, a fast algorithm needs to be developed for step 2.

Finally, we observe that dimension reduction is only a preprocessing stage. Even if this stage is a little expensive, it will be worthwhile if it effectively reduces the cost of the postprocessing involved in classification and document retrieval, which will be the dominating parts computationally.

Acknowledgments. The authors would like to thank Profs. Lars Eldén and Chris Paige for valuable discussions which improved the presentations in this paper. A part of this work was carried out during the summer of 2001, when H. Park was visiting the School of Mathematics, Seoul National University, Seoul, Korea, with the

support of the Brain Korea 21 program. She would like to thank Profs. Hyuck Kim and Dongwoo Sheen, as well as the School of Mathematics at SNU for their kind invitation.

REFERENCES

[1] M. W. Berry, S. T. Dumais, and G. W. O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Rev., 37 (1995), pp. 573–595.
[2] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[3] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice–Hall, Englewood Cliffs, NJ, 1992.
[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, 1990.
[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[6] P. Howland and H. Park, *Extension of discriminant analysis based on the generalized singular value decomposition*, in Proceedings of the Second SIAM International Conference on Data Mining/Text Mining Workshop, SIAM, Philadelphia, 2002.
[7] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice–Hall, Englewood Cliffs, NJ, 1988.
[8] M. Jeon, H. Park, and J. B. Rosen, *Dimension reduction based on centroids and least squares for efficient processing of text data*, in Proceedings of the First SIAM International Conference on Data Mining, CD-ROM, SIAM, Philadelphia, 2001.
[9] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, Kluwer Academic, Dordrecht, The Netherlands, 1997.
[10] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, SIAM, Philadelphia, 1995.
[11] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.
[12] H. Park, M. Jeon, and J. B. Rosen, *Lower dimensional representation of text data based on centroids and least squares*, BIT, to appear.
[13] G. Salton, *The SMART Retrieval System*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw–Hill, New York, 1983.
[15] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, 1999.
[16] C. F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 180–187

A NOTE ON THE SYMMETRIC RECURSIVE INVERSE EIGENVALUE PROBLEM*

RAPHAEL LOEWY† AND VOLKER MEHRMANN‡

Abstract. In [M. Arav et al., *SIAM J. Matrix Anal. Appl.*, 22 (2000), pp. 392–412] the recursive inverse eigenvalue problem for matrices was introduced. In this paper we examine an open problem on the existence of symmetric positive semidefinite solutions that was posed there. We first give several counterexamples for the general case and then characterize under which further assumptions the conjecture is valid.

Key words. inverse eigenvalue problem, recursive solution, symmetric matrices, positive semidefinite matrices

AMS subject classifications. 15A29, 15A18, 15A48, 15A57

PII. S0895479802408839

1. Introduction.

In [1] several classes of recursive inverse eigenvalue problems that construct matrices from eigenvalues and eigenvectors of leading principal submatrices were introduced. A simple application of such problems is the construction of Leontief models in economics (see, e.g., [2]) when a feasible model with $n-1$ inputs and $n-1$ outputs is extended (by adding an input and an output) to a larger feasible model with prescribed equilibrium point; see [1].

In this paper we discuss the particular case of the *real symmetric recursive inverse eigenvalue problem*, in the following denoted by **SRIEP**(n), which has the following form: For given scalars $s_1, \ldots, s_n \in \mathbb{R}$ and real vectors

$$r_1 = \begin{bmatrix} r_{1,1} \end{bmatrix}, \; r_2 = \begin{bmatrix} r_{1,2} \\ r_{2,2} \end{bmatrix}, \; \ldots, \; r_n = \begin{bmatrix} r_{1,n} \\ \vdots \\ r_{n,n} \end{bmatrix},$$

construct a symmetric matrix $A \in \mathbb{R}^{n,n}$ such that

$$A[i]r_i = s_i r_i, \qquad i = 1, \ldots, n,$$

where $A[i]$ denotes the ith leading principal submatrix of A.

We use the following notation; see [4]. By $\circ$ we denote the Hadamard (or elementwise) product of matrices. For an $n \times n$ matrix A and increasing sequences α, β of elements in $\{1, 2, \ldots, n\}$, $A[\alpha|\beta]$ denotes the submatrix of A given by the row indices α and the column indices β. Futhermore, A^T denotes the transpose of A, A^{-T} denotes the transpose of the inverse (if it exists), and e_i denotes the ith unit vector of appropriate dimension.

*Received by the editors June 4, 2002; accepted for publication (in revised form) by H. Woerdeman October 21, 2002; published electronically May 15, 2003.
http://www.siam.org/journals/simax/25-1/40883.html

†Department of Mathematics, Technion, Haifa 32000, Israel (loewy@techunix.technion.ac.il). The research of this author was supported by the Technische Universität Berlin, Deutscher Akademischer Austauschdienst, and the Fund for the Promotion of Research at the Technion.

‡Institut für Mathematik, TU Berlin, D-10623 Berlin, Germany (mehrmann@math.tu.berlin.de). The research of this author was supported by the DFG Research Center Mathematics for Key Technologies, TU Berlin.

The following matrices constructed from the data of the **SRIEP**(n) are used:

$$(1)\qquad R_n = \begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,n} \\ 0 & r_{2,2} & \dots & r_{2,n} \\ \vdots & & & \vdots \\ 0 & 0 & \dots & r_{n,n} \end{bmatrix}, \qquad S_n = \begin{bmatrix} s_1 & s_2 & s_3 & \cdots & s_n \\ s_2 & s_2 & s_3 & \cdots & s_n \\ s_3 & s_3 & s_3 & \cdots & s_n \\ \vdots & & & & \vdots \\ s_n & s_n & \cdots & \cdots & s_n \end{bmatrix}.$$

In [1] the existence and uniqueness of solutions to **SRIEP**(n) are characterized, and in particular it is shown that if R_n is invertible, i.e., all elements $r_{i,i}$ are nonzero, then the solution of **SRIEP**(n) exists, is unique, and is given by the formula

$$(2)\qquad A = R_n^{-T}(S_n \circ (R_n^T R_n))R_n^{-1}.$$

Thus the unique solution A is positive definite (positive semidefinite) if and only if $S_n \circ (R_n^T R_n)$ is positive definite (positive semidefinite). But if R_n is singular and if a solution exists, then it is not unique, so a natural question to ask is whether there exists a positive definite (positive semidefinite) solution. It was also shown in [1] that any solution of **SRIEP**(n) must satisfy the matrix equation

$$(3)\qquad R_n^T A R_n = S_n \circ (R_n^T R_n).$$

Hence it is clear that if there exists a positive definite (positive semidefinite) solution, then $S_n \circ (R_n^T R_n)$ has to be positive semidefinite. In [1] it was conjectured that the converse also holds; i.e., *let $n \geq 2$, and suppose that $S_n \circ (R_n^T R_n)$ is positive semidefinite (positive definite). Then there exists a positive semidefinite (positive definite) solution for* **SRIEP**(n).

In this article we show that this conjecture is generally false. We give an example which shows that **SRIEP**(n) does not have to possess a solution at all if the assumption of the conjecture holds. Furthermore, the conjecture fails to hold even if we add the assumption that the problem has a solution when $\operatorname{rank} S_n \circ (R_n^T R_n) \leq n-2$. We then prove that if a solution of **SRIEP**(n) exists and $\operatorname{rank} S_n \circ (R_n^T R_n) > n-2$, then there exists a positive semidefinite (positive definite) solution for **SRIEP**(n).

2. Counterexamples. In this section we present several counterexamples that show that the conjecture in [1] as well as several obvious modifications do not hold.

Example 1. Let $n = 2$, $r_1 = [1]$, $r_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $s_1 = 2$, $s_2 = 1$. Then

$$R_2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad R_2^T R_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix},$$

and clearly $S_2 \circ (R_2^T R_2) = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ is positive definite. Nevertheless, it is straightforward to check that there exists no matrix $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ such that $A[1]r_1 = 2r_1$ and $Ar_2 = r_2$. Hence the conjecture is false as stated. In Example 1 the problem has no solution at all, so an immediate modification of the conjecture would be to require that the problem is solvable.

The next example shows that even with this modification the conjecture is false.

Example 2. Let $n = 3$,

$$r_1 = [1], \quad r_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad r_3 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix},$$

$s_1 = s_2 = 3$, and $s_3 = 9$. Then

$$R_3 = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad R_3^T R_3 = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 4 & -2 \\ -1 & -2 & 3 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 3 & 3 & 9 \\ 3 & 3 & 9 \\ 9 & 9 & 9 \end{bmatrix},$$

and clearly

$$S_3 \circ (R_3^T R_3) = \begin{bmatrix} 3 & 6 & -9 \\ 6 & 12 & -18 \\ -9 & -18 & 27 \end{bmatrix}$$

is positive semidefinite and of rank 1. The system of 6 equations for the elements of A is

$$\begin{aligned} a_{1,1} &= 3, \\ 2a_{1,1} &= 6, \\ 2a_{1,2} &= 0, \\ -a_{1,1} + a_{1,2} + a_{1,3} &= -9, \\ -a_{1,2} + a_{2,2} + a_{2,3} &= 9, \\ -a_{1,3} + a_{2,3} + a_{3,3} &= 9, \end{aligned}$$

which has the general solution

$$A = \begin{bmatrix} 3 & 0 & -6 \\ 0 & 9 - a_{2,3} & a_{2,3} \\ -6 & a_{2,3} & 3 - a_{2,3} \end{bmatrix},$$

with $a_{2,3}$ to be chosen freely. But, since $\det A = 3[(9 - a_{2,3})(3 - a_{2,3}) - a_{2,3}^2] - 36(9 - a_{2,3}) = -245$ does not depend on $a_{2,3}$, clearly no positive semidefinite solution exists, although there exist symmetric solutions.

We can lift Example 2 to get counterexamples for all n, as long as $\operatorname{rank} S_n \circ (R_n^T R_n) \leq n - 2$.

Example 3. Let $n \geq 4$,

$$r_1 = [1], \quad r_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad r_3 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix},$$

$r_i = e_i$ for $i = 4, 5, \ldots, n$. Let, furthermore, $s_1 = s_2 = 3, s_3 = 9$, and let s_i, $i = 4, 5 \ldots, n$, be any positive numbers. Then

$$R_n = \left[\begin{array}{ccc|c} 1 & 2 & -1 & \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & \\ \hline & 0 & & I_{n-3} \end{array}\right],$$

and

$$S_n \circ (R_n^T R_n) = \left[\begin{array}{ccc|ccc} 3 & 6 & -9 & & & \\ 6 & 12 & -18 & & 0 & \\ -9 & -18 & 27 & & & \\ \hline & & & s_4 & & \\ & & & & \ddots & \\ & 0 & & & & s_n \end{array}\right]$$

is positive semidefinite of rank $n-2$. The direct sum of A from Example 2 and I_{n-3} is a solution. If B is any solution, then $B[3]$ necessarily is a solution for Example 2 and hence B cannot be positive semidefinite.

These examples demonstrate that to prove the conjecture we have to require that the rank of $S_n \circ (R_n^T R_n)$ is at least $n-1$. In the next section we show that in this case the conjecture is true.

3. Main result. In this section we present our main result and prove the conjecture for the case where $\operatorname{rank}(S_n \circ (R_n^T R_n)) \geq n-1$.

THEOREM 4. *Let matrices R_n and S_n be given such that $S_n \circ (R_n^T R_n)$ is positive semidefinite with* $\operatorname{rank}(S_n \circ (R_n^T R_n)) \geq n-1$. *If problem* **SRIEP**$(n)$ *has a solution, then it also has a positive semidefinite solution.*

Proof. Suppose first that $\operatorname{rank}(S_n \circ (R_n^T R_n)) = n$; i.e., $S_n \circ (R_n^T R_n)$ is positive definite. Let A be any solution of **SRIEP**(n). Then it has been shown in [1] that this solution must satisfy (3). This implies that R_n is invertible, and it has been shown in [1] that the solution is unique, given by (2), and hence positive definite.

It remains to study the case where $\operatorname{rank}(S_n \circ (R_n^T R_n)) = n-1$. If R_n is invertible, then again the solution A is unique and given by (2), which is a positive semidefinite matrix of rank $n-1$. Hence, we may assume in the following that R_n is singular.

Let A be any particular solution of **SRIEP**(n). Then it follows from (3) that $\operatorname{rank} R_n = n-1$.

Using a sequence of elementary row and column operations [3], i.e., adding scalar multiples of one row (or column) to another, it follows that there exist invertible matrices P, Q such that

$$PR_nQ = \operatorname{diag}(\Sigma_{n-1}, 0), \tag{4}$$

with Σ_{n-1} of size $n-1 \times n-1$, diagonal and invertible. Actually we could achieve $\Sigma_{n-1} = I_{n-1}$, but we will use a different factorization below.

It follows from (3) that

$$(Q^T R_n^T P^T)(P^{-T} A P^{-1})(PR_nQ) = Q^T (S_n \circ (R_n^T R_n))Q. \tag{5}$$

Partition $\tilde{A} = P^{-T} A P^{-1}$ conformally with (4) as

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{1,1} & \tilde{A}_{1,2} \\ \tilde{A}_{1,2}^T & \tilde{A}_{2,2} \end{bmatrix}.$$

Then it follows from (5) that

$$Q^T (S_n \circ (R_n^T R_n))Q = \begin{bmatrix} \Sigma_{n-1} \tilde{A}_{1,1} \Sigma_{n-1} & 0 \\ 0 & 0 \end{bmatrix},$$

and hence, since the left side has rank $n-1$, we have that $\tilde{A}_{11}$ is positive definite.

Note that $\tilde{A}$ does not depend on Q, so we may choose P and Q so that the factorization (4) holds, while P is as simple as possible. We now construct such a P, and since Q does not affect $\tilde{A}$, we do not record the column operations. Let

$$Z = \{i \in \{1, \ldots, n\} : r_{i,i} \neq 0\} = \{i_1, \ldots, i_m\},$$

where we assume that $1 \leq i_1 < i_2 < \cdots < i_m \leq n$. We call the entries $r_{i,i}$ with $i \in Z$ *pivot elements of the first type.*

For every $i \in Z$, using elementary column operations, we can eliminate all off-diagonal elements in row i, and since R_n is upper triangular, this will not alter any of the diagonal elements. Hence the only nonzero element in row $i \in Z$ of the transformed matrix $\tilde{R}_n$ is the original diagonal element $r_{i,i}$. Moreover, $\tilde{R}_n = [\tilde{r}_{i,j}]$ is still upper triangular and of rank $n-1$.

Partition the set of indices $\tilde{Z} = \{1, \ldots, n\} \backslash Z$ into maximal disjoint subsets $Z_1, \ldots, Z_k$ of consecutive integers, representing the row numbers with vanishing diagonal elements $r_{j,j}$. For example, if the zero diagonal elements of R_n are $r_{1,1}$, $r_{4,4}$, $r_{5,5}$, $r_{6,6}$, $r_{9,9}$, $r_{10,10}$, and $r_{14,14}$, then $Z_1 = \{1\}$, $Z_2 = \{4, 5, 6\}$, $Z_3 = \{9, 10\}$, and $Z_4 = \{14\}$.

Consider now an arbitrary Z_j, where $1 \le j \le k$, and assume for simplicity that $Z_j = \{p, p+1, \ldots, p+q\}$, where $q \ge 0$. Then, since $\operatorname{rank} \tilde{R}_n = n-1$, it follows that if $q \ge 1$, then all entries $\tilde{r}_{l,l+1}$, $l = p, \ldots, p+q-1$, are nonzero. We call these entries *pivot elements of the second type*. Furthermore, for all the blocks associated with index sets $Z_j = \{p_j, \ldots, p_j + q_j\}$, $j = 1, \ldots, k-1$, we have that there is at least the nonzero element $\tilde{r}_{p_j+q_j,s}$ in row $p_j + q_j$, where s is the smallest element in Z_{j+1}. If this were not the case, then we would have that $\operatorname{rank} \tilde{R}_n \le n-2$, a contradiction. We call the entries $\tilde{r}_{p_j+q_j,s}$ *pivot elements of the third type*.

Since there are no nonzero elements below the pivot elements of second type, we can perform further elementary column operations to eliminate more nonpivot elements. Consider first Z_1 and eliminate (in the natural order) all the nonpivot elements in the rows associated with the pivots of the second type. These operations do not affect any other rows associated with pivots of the second type or third type. Then we use the pivot element of the third type (if it exists) to annihilate the elements in its row, again without affecting any other rows. We proceed in the same way with the blocks associated with $Z_2, \ldots, Z_k$, again in the natural order.

Let w denote the largest element of Z_k, and let $\hat{R} = [\hat{r}_{p,q}]$ denote the matrix obtained via these column operations applied to $\tilde{R}_n$. The matrix $\hat{R}_n$ has as nonzero elements all the pivot elements of first, second, and third type, plus possibly some elements in row w. Since we have used only column operations, we have determined an invertible matrix $\hat{Q}$ such that $\hat{R}_n = R_n \hat{Q}$.

For the remainder of the proof we consider two cases.

Case 1. If $w = n$, then we have obtained (possibly after some additional permutation of columns) the desired form (4) with $P = I_n$, and hence $\tilde{A} = A$ and the submatrix $A[n-1]$ is positive definite. Since $r_{n,n} = 0$, it follows that the homogeneous linear system corresponding to **SRIEP**(n) has the matrix $E_{n,n} = e_n e_n^T$ as a solution. Thus all matrices of the form $\hat{A}(\alpha) = \alpha E_{n,n} + A$ with our particular solution A are solutions, and since $A[n-1]$ is positive definite, choosing $\alpha > 0$ sufficiently large, we obtain that $\hat{A}(\alpha)$ is positive definite.

Case 2. If $w < n$, then we need to perform elementary row operations using the pivots in rows $w+1$, $w+1$, $\ldots$, n of $\hat{R}_n$ to annihilate the entries in positions $(w, w+1)$, $(w, w+2)$, $\ldots$, (w, n) of $\hat{R}_n$. The corresponding pivot elements $r_{w+1,w+1}$, $r_{w+2,w+2}$, $\ldots, r_{n,n}$ are of the first type.

Using Cramer's rule we can exactly determine the elements of $\hat{R}_n$ that we still have to eliminate, i.e.,

$$\hat{r}_{w,w+1} = \det R_n[w|w+1],$$
$$\hat{r}_{w,w+2} = -\frac{\det R_n[w, w+1|w+1, w+2]}{r_{w+1,w+1}},$$

$$\hat{r}_{w,w+3} = \frac{\det R_n[w, w+1, w+2 | w+1, w+2, w+3]}{r_{w+1,w+1} r_{w+2,w+2}},$$

$$\vdots$$

$$\hat{r}_{w,n} = (-1)^{n-w-1} \frac{\det R_n[w, w+1, \ldots, n-1 | w+1, w+2, \ldots, n]}{r_{w+1,w+1} r_{w+2,w+2} \cdots r_{n-1,n-1}}.$$

Introducing the matrices of order $n - w + 1$,

$$C = \begin{bmatrix} 0 & 1 & 0 & \ldots & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & \ldots & 0 & 1 \\ 1 & 0 & \ldots & \ldots & 0 & 0 \end{bmatrix}, \qquad M = I_{n-w+1} - e_1 \begin{bmatrix} 0 \\ \frac{\hat{r}_{w,w+1}}{\hat{r}_{w+1,w+1}} \\ \cdots \\ \frac{\hat{r}_{w,n}}{\hat{r}_{n,n}} \end{bmatrix}^T,$$

and

$$P_1 = \begin{bmatrix} I_{w-1} & 0 \\ 0 & C \end{bmatrix}, \qquad P_2 = \begin{bmatrix} I_{w-1} & 0 \\ 0 & M \end{bmatrix},$$

then with $P = P_1 P_2$ we have obtained invertible matrices P, Q such that (4) holds. Recall that for $\tilde{A} = P^{-T} A P^{-1}$ we have $\tilde{A}[n-1]$ is positive definite.

As in Case 1, we show that there exists a rank 1 positive semidefinite solution A_0 of the homogeneous linear system corresponding to **SRIEP**(n) and a scalar $\alpha > 0$ such that $\tilde{A} + \alpha P^{-T} A_0 P^{-1} = P^{-T}(A + \alpha A_0) P^{-1}$ is a positive definite solution of **SRIEP**(n).

Let

$$\tilde{z}^T = \Big[\ 1, \quad \tfrac{-\det R_n[w|w+1]}{\hat{r}_{w+1,w+1}}, \quad \tfrac{\det R_n[w,w+1|w+1,w+2]}{\hat{r}_{w+1,w+1}\hat{r}_{w+2,w+2}}, \quad \ldots$$
$$\ldots, \quad (-1)^{n-w} \tfrac{\det R_n[w,w+1,\ldots,n-1|w+1,w+2,\ldots,n]}{\hat{r}_{w+1,w+1}\cdots\hat{r}_{n,n}} \ \Big],$$

and if $A_0 = zz^T$, where $z^T = \begin{bmatrix} 0 & \ldots & 0 & \tilde{z}^T \end{bmatrix}$, then A_0 satisfies the homogeneous system $A_0[i] r_i = 0$ for $i = 1, 2, \ldots, n$. To show this it suffices to prove that $z^T R_n e_i = 0$, for $i = 1, \ldots, n$. This is clear for $i = 1, \ldots, w-1$ because of the zeros in z and for $i = w$, since $r_{w,w} = 0$. To prove this for $i = w+1, \ldots, n$, we have to show that

$$\tilde{z}^T R_n[w, w+1, \ldots, n | w+1, w+2, \ldots, n] = 0,$$

but this is exactly how we have constructed $\tilde{z}$ and follows from Cramer's rule.

By construction we also have that $z^T P^{-1} = e_n^T$ and hence

$$P^{-T} A_0 P^{-1} = P^{-T} z z^T P^{-1} = e_n e_n^T = E_{n,n}.$$

The same proof as in Case 1 gives the existence of a positive definite solution. ∎

The interesting case in the proof of Theorem 4 is when $\operatorname{rank} R_n = n - 1$. In this case we needed to add a particular solution of the homogeneous system corresponding to **SRIEP**(n) in order to get a positive definite solution.

Thus it is interesting to study the homogeneous system in slightly more detail.

THEOREM 5. *Let $n \geq 2$; consider the homogeneous system*

$$A[i] r_i = 0, \qquad i = 1, 2, \ldots, n, \tag{6}$$

associated with **SRIEP**(n)*; and suppose that* $\operatorname{rank} R_n = n-1$. *Let* w *be the largest integer such that* $r_{i,i} = 0$. *Then the general solution of* (6) *has dimension* w *if* $r_w = 0$ *and dimension* $w-1$ *if* $r_w \neq 0$. *Moreover, for any solution* A *of* (6) *we have* $A[w-1] = 0$. *Furthermore, if* $r_w = 0$, *then the elements* $a_{1,w}, \ldots, a_{w,w}$ *can be chosen to be the free variables in the solution of* (6). *If* $r_w \neq 0$ *and* s *is the smallest integer such that* $r_{s,w} \neq 0$, *then* $a_{1,w}, a_{2,w}, \ldots, a_{s-1,w}, a_{s+1,w}, \ldots, a_{w,w}$ *can be chosen to be the free variables in the solution of* (6). *Here, if* $w = 1$, *we mean that* $a_{1,1}$ *is the only free variable.*

Proof. The proof is by induction on n. The case $n = 2$ is trivial. Suppose first that $w < n$. Consider the subsystem of (6) given by

$$A[i]r_i = 0, \qquad i = 1, 2, \ldots, w, \tag{7}$$

and apply the induction hypothesis. Since all diagonal entries $r_{w+1,w+1}, \ldots, r_{n,n}$ are nonzero, the system $A[w+1]r_{w+1} = 0$ will determine $a_{1,w+1}, \ldots, a_{w+1,w+1}$ uniquely in terms of the free variables of (7). Continuing in this way with the equations $A[w+j]r_{w+j} = 0$, $j = 2, \ldots, n-w$, we determine all the remaining entries of A in terms of the free variables of (7).

So we may assume that $w = n$, i.e., $r_{n,n} = 0$, and therefore the whole last row of R_n is zero. For $i = 1, 2, \ldots, n-1$ let

$$a_{(i)} = [a_{i,1}, a_{i,2}, \ldots, a_{i,n-1}],$$

and let $\hat{r}_j$ denote the vector obtained by deleting the last entry of r_j, $j = 1, 2, \ldots, n$. Since $\operatorname{rank} R_n = n-1$, the first $n-1$ rows of R_n are linearly independent, implying that $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$ span $\mathbb{R}^{n-1}$. Considering the row vector $a_{(1)}$, it follows from (6) that $a_{(1)}\hat{r}_j = 0$ for $j = 1, 2, \ldots, n$, and hence it follows that $a_{(1)} = 0$, in particular $a_{1,2} = a_{2,1} = 0$. Then for $a_{(2)} = [0, a_{2,2}, \ldots, a_{2,n-1}]$ we have $a_{(2)}\hat{r}_1 = 0$, since $a_{2,1} = 0$ and $a_{(2)}\hat{r}_j = 0$ for $j = 2, \ldots, n$ by (6), and hence $a_{(2)} = 0$. In particular we have $a_{1,3} = a_{3,1} = a_{2,3} = a_{3,2} = 0$. Proceeding inductively, we obtain in a similar way that $a_{(3)} = a_{(4)} = \ldots = a_{(n-1)} = 0$ and hence $A[n-1] = 0$. It remains to consider $Ar_n = 0$. If $r_n = 0$, this is automatically satisfied and hence $a_{1,n}, a_{2,n}, \ldots, a_{n,n}$ are the free variables. Otherwise, if $r_n \neq 0$, then there is a single linear equation

$$r_{1,n}a_{1,n} + r_{2,n}a_{2,n} + \cdots + r_{n,n-1}a_{n-1,n} = 0$$

for the free variables. This concludes the proof. ∎

We have given conditions so that there exists a positive semidefinite (positive definite) solution to **SRIEP**(n) that depends just on the fact that $S_n \circ (R_n^T R_n)$ is positive semidefinite (positive definite), but no use of the special structure of the matrix S_n is made. Some sufficient conditions that use just inequalities between the s_j are given in [1]. For example, it is shown there that if R_n is invertible and $s_1 > s_2 > \cdots > s_n \geq 0$, then the unique solution of **SRIEP**(n) is positive semidefinite, and if $s_n > 0$, then the unique solution is positive definite. However, these inequalities are not necessary to have a positive semidefinite solution.

4. Conclusion. We have presented counterexamples to a conjecture posed in [1] and conditions under which the conjecture holds.

REFERENCES

[1] M. Arav, D. Hershkowitz, V. Mehrmann, and H. Schneider, *The recursive inverse eigenvalue problem,* SIAM J. Matrix Anal. Appl., 22 (2000), pp. 392–412.
[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[3] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.
[4] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 188–212

A PROJECTION METHOD FOR LEAST SQUARES PROBLEMS WITH A QUADRATIC EQUALITY CONSTRAINT*

ZHENYUE ZHANG† AND YUSHAN HUANG†

Abstract. We consider the least squares problem with a quadratic equality constraint (LSQE), i.e., minimizing $\|Ax - b\|_2$ subject to $\|x\|_2 = \alpha$, without the assumption $\|A^\dagger b\|_2 > \alpha$ which is commonly imposed in the literature. Structure and perturbation analysis are given to demonstrate the sensitivity of the LSQE problem. We present a projection method combined with correction techniques (PMCT) for solving numerically the LSQE problem when the LSQE problem is ill-conditioned. We also give a detailed convergence analysis of our algorithms to illustrate the convergence behavior. Our algorithms have some obvious advantages over Newton's method and variants. Numerical experiments indicate that PMCT is much more efficient than Newton's methods when the LSQE problem is ill-conditioned; PMCT has a 90% success rate in terms of convergence, while commonly used Newton-type iterations almost always fail.

Key words. least squares, ill-conditioned problem, projection method, perturbation analysis, singular value

AMS subject classifications. 65F20, 65K10, 15A18, 15A12, 65F10

PII. S0895479801398712

1. Introduction.

Given an $m \times n$ matrix A and an m-dimensional column vector b, we consider numerical methods for solving the following least squares problem with a quadratic equality constraint (LSQE):

$$\min \|Ax - b\|_2 \quad \text{subject to} \quad \|x\|_2 = \alpha, \tag{1.1}$$

where α is a positive scalar. This problem is equivalent to solving the *constrained* normal equations

$$(A^T A + \lambda I)x = A^T b \quad \text{subject to} \quad x^T x = \alpha^2 \tag{1.2}$$

with respect to multiplier λ and vector x [5, 15]. The problem (1.1) (or (1.2)) arises in many important applications. For example, it is the Karush–Kuhn–Tucker condition for the quadratic inequality constrained least squares problem

$$\min \|Ax - b\|_2 \quad \text{subject to} \quad \|x\|_2 \le \alpha \tag{1.3}$$

when some regularization techniques are used to solve the ill-posed least squares problem [18, 22]

$$\min \|Ax - b\|_2.$$

Some interesting algorithms for solving the minimization problem (1.2) were recently discussed in [8, 14, 16]. Basically, those algorithms are designed for the case

*Received by the editors October 1, 2001; accepted for publication (in revised form) by G. H. Golub October 21, 2002; published electronically May 29, 2003.
http://www.siam.org/journals/simax/25-1/39871.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, People's Republic of China (zyzhang@zju.edu.cn, yhuang@arcsoft.com.cn). The work of these authors was supported in part by NSFC (project 19771073) and Zhejiang Provincial Natural Science Foundation of China. The first author was also supported by the Special Funds for Major State Basic Research Projects of China (project G1999032804) and Foundation for University Key Teacher by the Ministry of Education, China.

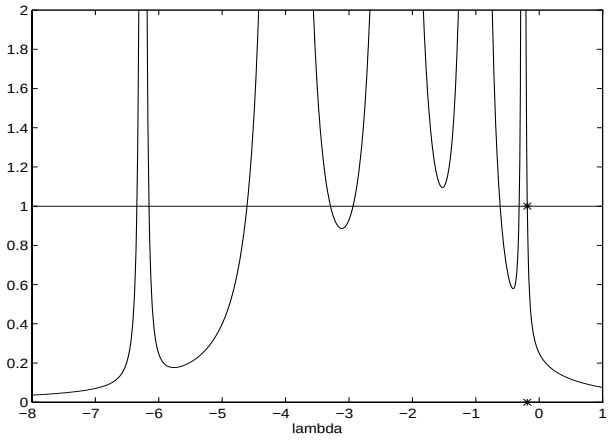


FIG. 1. *Function* $f(\lambda)$.

when the optimal (largest) Lagrange multiplier λ^* is positive. This condition is ensured by the assumption $\|A^\dagger b\|_2 > \alpha$, where $A^\dagger$ is the Moore–Penrose inverse of A [25]. In this case, if the matrix A is not very ill-conditioned, Newton's method and its variants are also particularly efficient because the optimal multiplier $\lambda^* > 0$ is positive and $\lambda_0 = 0$ is a good starting point for monotonical convergence of Newton-type iterations. In any case, Newton-type methods will converge if the initial guess satisfies

$$\lambda_0 \in (-\sigma_n^2(A),\ \lambda^*). \tag{1.4}$$

However, the optimal Lagrange multiplier λ^* may be negative and close to $-\sigma_n^2(A)$ if $\|A^\dagger b\|_2 < \alpha$. See Figure 1 for a small example with $\alpha = 1$, where function $f(\lambda)$ is defined in (1.5) and λ^* is the largest solution of the equation $f(\lambda) = 1$. It can be proven that for fixed A and $\alpha > 0$, the smaller in norm the vector b is, the closer to $-\sigma_n^2(A)$ the solution λ^* is. In general it is hard to verify the assumption (1.4) for λ_0 when $\lambda^* < 0$. In fact, Newton-type iterations with starting value $\lambda_0 = 0$ are often divergent if $\|A^\dagger b\|_2 \ll \alpha$. It is also possible that the equality $\lambda^* = -\sigma_n^2(A)$ holds in the case when $\|A^\dagger b\|_2 < \alpha$. In this case, the LSQE problem is equivalent to the problem of computing the right singular vectors of A corresponding to the smallest singular value under some assumptions. See Theorem 2.2 for details. In this paper we therefore always assume that $\lambda^* > -\sigma_n^2(A)$.

This paper will focus on the LSQE problem (1.1) when $\|A^\dagger b\|_2 < \alpha$. There are two reasons we propose a new algorithm for solving the problem without the restriction $\|A^\dagger b\|_2 > \alpha$. The first reason is that if $\|A^\dagger b\|_2 \ll \alpha$ and/or A has clustered smallest singular values, the LSQE problem (1.1) will be quite ill-conditioned—even the matrix A itself is well conditioned. It is related to the "hard case" in [27] and there are applications where this case occurs. Few results, however, have been reported on the algorithm development for the "hard case." The second reason is that the standard problem (1.1) with $\alpha = 1$ is the special case of the following Procrustes problem on the Stiefel manifold [10]:

$$\min \left\{ \ \|AQ - B\|_F \quad \middle| \quad Q^T Q = I_p \right\}$$

for given matrices $A \in R^{m\times n}$ and $B \in R^{m\times p}$. For the one-column Procrustes problem, $p = 1$, the inequality $\|A^\dagger b\|_2 < 1$ may hold in some cases. Our main idea consists in using a projection method combined with a correction technique. Differing from Newton's method, our projection method does not directly work on the so-called *secular function*

$$f(\lambda) = \|(A^TA + \lambda I)^{-1}A^Tb\|_2^2. \tag{1.5}$$

We project the parameterized vector $x(\lambda) = (A^TA+\lambda I)^{-1}A^Tb$ onto a one-dimensional subspace that also depends on the multiplier λ and a previous approximation of the optimal multiplier λ^*, which leads to a rational approximation of $f(\lambda)$. Our convergence analysis given in section 6 shows that this approach has a larger convergence range for the choice of a starting value λ_0, compared with Newton-type approaches. Furthermore, the combination generally guarantees the positive semidefiniteness of the matrix $A^TA + \lambda_k I$. Newton-type methods don't satisfy this property automatically and therefore need a lot of supplementary computations. In fact, in our method combined with the correction technique, the choice of a starting value λ_0 is not as crucial as in the case of Newton's methods, and the algorithm has a higher rate of success in terms of convergence, especially if $\|A^\dagger b\|_2$ is small or A has clustered smallest singular values.

The rest of the paper is organized as follows. We will first review some theoretical results about the characteristics of the solutions of the LSQE problem in section 2, and then discuss the effect of perturbations of A and b upon the solutions of the problem in section 3. We will review Newton's method and some variants in section 4 which will be compared to our projection method. In section 5, we will propose the projection method for solving the LSQE problem (1.1) iteratively. Some properties relative to the iterative scheme are discussed. A detailed convergence analysis of the projection method proposed will be given in section 6 to show the efficiency. To improve the convergence, we will further consider a correction technique in section 7. Numerical experiments will be given in section 8.

2. Least squares with quadratic equality constraints. In this section, we will review some theoretical results about the characteristics of the solutions of the LSQE problem (1.1). For simplicity, we assume in the rest of the paper that $\alpha = 1$. It is known that the *unconstrained* normal equations

$$(A^TA + \lambda I)x = A^Tb$$

follow directly from setting the gradient of Lagrange function

$$L(x, \lambda) = \|Ax - b\|_2^2 + \lambda(\|x\|_2^2 - 1)$$

to be zero. In general, there are many pairs (λ, x) satisfying the normal equations [11]. Moreover, for fixed multiplier $\lambda = -\sigma_j^2$ with the jth singular value σ_j of A, the unconstrained normal equations may have multiple solutions. If the multiplier $\lambda \neq -\sigma_j^2$ for all $j = 1, \ldots, n$, the corresponding vector x is uniquely determined by λ and

$$x = (A^TA + \lambda I)^{-1}A^Tb.$$

In this case, the multipliers required are the roots of the secular equation

$$f(\lambda) = \|(A^TA + \lambda I)^{-1}A^Tb\|_2^2 = 1.$$

What we are interested in are the solutions (λ, x) with respect to the optimal multiplier λ^*, the largest one denoted by

$$\lambda^* = \max\{\lambda \mid (\lambda, x) \text{ solves (1.2) for some } x\}. \tag{2.1}$$

In [11], it is proven that $\lambda^* \geq -\sigma_n^2$ and only the vectors x^*, the solutions of (1.2) corresponding to λ^*, solve the LSQE problem (1.1). We state the result as the following theorem.

THEOREM 2.1. *Let λ^* be defined by* (2.1). *Then a vector x^* is a minimizer of* (1.1) *if and only if (λ^*, x^*) is a solution of the normal equations* (1.2).

We will refer to the solution (λ^*, x^*) of the normal equations as an optimal solution. Note that the vector x^* may not be unique if $\lambda^* = -\sigma_n^2$. In that case, LSQE problem (1.1) is basically the problem of computing the smallest singular value and the relative right singular vectors and solving the singular linear system $(A^TA - \sigma_n^2 I)q = A^Tb$. The following theorem further characterizes the optimal solutions.

THEOREM 2.2. *Let σ_n be the smallest singular value of A. If the singular system $(A^TA - \sigma_n^2 I)q = A^Tb$ has no solutions, then $\lambda^* > -\sigma_n^2$ and $x^* = (A^TA + \lambda^* I)^{-1}A^Tb$.*

Assume, in reverse, that the singular system $(A^TA - \sigma_n^2 I)q = A^Tb$ has at least one solution and q^* is the unique minimum norm solution of the singular system. Then we have that

1. *if $\|q^*\|_2 > 1$, then $\lambda^* > -\sigma_n^2$ and $x^* = (A^TA + \lambda^* I)^{-1}A^Tb$;*
2. *if $\|q^*\|_2 = 1$, then $\lambda^* = -\sigma_n^2$ and $x^* = q^*$;*
3. *if $\|q^*\|_2 < 1$, then $\lambda^* = -\sigma_n^2$ and solutions of* (1.1) *are given in the form $x = q^* + \sqrt{1 - \|q^*\|_2^2}\, v_n$ with unit right singular vectors v_n corresponding to the smallest singular value σ_n of A.*

Proof. Let $A = U\Sigma V^T$ be the singular value decomposition of A with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$. Denote

$$J = \{j \mid \sigma_j = \sigma_n\}$$

and $c = x^Tb = (c_1, \ldots, c_n)^T$. If the singular system $(A^TA - \sigma_n^2 I)q = A^Tb$ has no solutions, then there exists an integer $j \in J$ such that $\sigma_j c_j \neq 0$. By the definition of $f(\lambda)$, if $\lambda > -\sigma_n^2$, then $f(\lambda)$ can be written in the form

$$f(\lambda) = \sum_{j \notin J} \left(\frac{\sigma_j c_j}{\sigma_j^2 + \lambda} \right)^2 + \frac{\beta}{(\sigma_n^2 + \lambda)^2} \equiv \hat{f}(\lambda) + \frac{\beta}{(\sigma_n^2 + \lambda)^2}$$

with a positive constant β. It is easy to verify that the secular equation $f(\lambda) = 1$ has the unique solution λ^* in the interval $(-\sigma_n^2, \infty)$ because $f'(\lambda) < 0$ for $\lambda > -\sigma_n^2$. Thus $x^* = (A^TA + \lambda^* I)^{-1}A^Tb$.

If $(A^TA - \sigma_n^2 I)q = A^Tb$ has one solution, then $\sigma_j c_j = 0$ for all $j \in J$ and $f(\lambda) = \hat{f}(\lambda)$ for $\lambda > -\sigma_n^2$. One can verify that the minimum norm solver q^* to the singular system $(A^TA - \sigma_n^2 I)q = A^Tb$ has the norm $\|q^*\|_2 = (\hat{f}(-\sigma_n^2))^{1/2}$. The condition $\|q^*\|_2 > 1$ thus implies there is a unique $\lambda^* > -\sigma_n^2$ that solves $f(\lambda) = 1$ because $f(\lambda)$ is strictly decreasing in $(-\sigma_n^2, \infty)$. For the case when $\|q^*\|_2 = 1$, we have that $\beta = 0$ and

$$f(\lambda) = \hat{f}(\lambda) < \hat{f}(-\sigma_n^2) = 1 \quad \text{for all} \quad \lambda > -\sigma_n^2.$$

Therefore (λ^*, q^*) solves (1.2) with the optimal multiplier $\lambda^* = -\sigma_n^2$. Recall that q^* is orthogonal to the right singular vectors of A corresponding to σ_n. For any

vector $x \neq q^*$ satisfying $(A^TA - \sigma_n^2 I)x = A^Tb$, $x - q^*$ is a right singular vector of A corresponding to σ_n. It follows that

$$\|x\|_2^2 = \|q^*\|_2^2 + \|x - q^*\|_2^2 > \|q^*\|_2^2 = 1.$$

By Theorem 2.1, $x = q^*$ is the unique solution of (1.1). Finally, if $\|q^*\|_2 < 1$, we can choose any right singular vector z of A with norm $\|z\|^2 = 1 - \|q^*\|_2^2$. Obviously, the pair $(-\sigma_n^2, q^* + z)$ solves (1.2). It is concluded that $\lambda^* = -\sigma_n^2$ because of the monotonicity of $f(\lambda)$ in $(-\sigma_n^2, \infty)$ and that $x = q^* + z$ is a solution of (1.1). ∎

In general, $\lambda^* > -\sigma_n^2$ and the corresponding optimal vector x^* has the explicit expression $x^* = (A^TA + \lambda^* I)^{-1}A^Tb$. Of course, λ^* depends continuously on the right vector b if A is fixed. It is not difficult to verify that the smaller the norm of b is, the closer to $-\sigma_n^2$ the largest multiplier λ^* is. The LSQE will be ill-conditioned if λ^* is tightly close to the pole $-\sigma_n^2$ of the function $f(\lambda) = \|(A^TA + \lambda I)^{-1}A^Tb\|_2$. In the next section, we will give a perturbation analysis of the LSQE problem to show the sensitivity of the LSQE problem. Roughly speaking, the condition number of problem (1.1) is $\|A\|_2/(\lambda^* + \sigma_n^2)$ provided $\lambda^* > -\sigma_n^2$.

3. Perturbation analysis. Throughout this section we assume that $\lambda^* > -\sigma_n^2$. We define by $x(\lambda)$ the unique solution of the linear system $(A^TA + \lambda I)x = A^Tb$ for $\lambda \neq \sigma_j$, i.e.,

$$x(\lambda) = (A^TA + \lambda I)^{-1}A^Tb. \tag{3.1}$$

Clearly, λ^* solves the secular equation

$$\|x(\lambda)\|_2 = 1$$

in the interval $(-\sigma_n^2, \infty)$, and $x^* = x(\lambda^*)$. In this section, we will consider the effect of the perturbations of A and b upon the optimal vector x^*.

Let $(\hat{\lambda}, \hat{x})$ be the optimal solution of (1.2) with A and b perturbed by δA and δb, respectively. To estimate $|\hat{\lambda} - \lambda^*|$ and $\|\hat{x} - x^*\|_2$, let's define $A(t)$ and $b(t)$ by

$$A(t) = A + t\,\delta A, \quad b(t) = b + t\,\delta b.$$

We also denote by $(\lambda^*(t), x^*(t))$ the optimal solution of (1.2) corresponding to $A(t)$ and $b(t)$; i.e., $(\lambda^*(t), x^*(t))$ satisfies that

$$(A(t)^TA(t) + \lambda^*(t)I)x^*(t) = A(t)^Tb(t) \quad \text{and} \quad x^*(t)^Tx^*(t) = 1. \tag{3.2}$$

Furthermore, we assume that $\lambda^*(t) > -\sigma_n^2(A(t))$ for all $t \in [0\ 1]$. By definition, $x^*(t) = x(\lambda^*(t))$, $x^* = x^*(0)$, $\lambda^* = \lambda^*(0)$, $\hat{x} = x^*(1)$, and $\hat{\lambda} = \lambda^*(1)$.

THEOREM 3.1. *Let (λ^*, x^*) and $(\hat{\lambda}, \hat{x})$ be the solutions of the maximum multiplier corresponding to (A, b) and $(A+\delta A,\ b+\delta b)$, respectively. There exist ξ_1 and $\xi_2 \in (0, 1)$ such that*

$$|\hat{\lambda} - \lambda^*| \leq \sqrt{\frac{\sigma_1^2(A + \xi_1\delta A) + \lambda^*(\xi_1)}{\sigma_n^2(A + \xi_1\delta A) + \lambda^*(\xi_1)}}\,\eta(\xi_1), \tag{3.3}$$

$$|\hat{x} - x^*| \leq \frac{1}{\sigma_n^2(A + \xi_2\delta A) + \lambda^*(\xi_2)}\,\eta(\xi_2), \tag{3.4}$$

where $\eta(t)$ is defined by

$$\eta(t) = \|(A + 2t\,\delta A)^T(\delta b - \delta A\,x^*(t)) + \delta A^T(b - Ax^*(t))\|_2.$$

Proof. For simplicity, we write $(\lambda^*(t), x^*(t))$ as $(\lambda(t), x(t))$. Differentiating the equations (3.2) with respect to t yields that

$$\begin{aligned}(3.5)\qquad &\big(A(t)^TA(t)+\lambda(t)I\big)x'(t)+\lambda'(t)x(t)+\big(A(t)^T\delta A+\delta A^TA(t)\big)x(t)\\ &\quad = A(t)^T\delta b+\delta A\, b(t)\end{aligned}$$

and $x^T(t)x'(t)=0$. Moving the last term on the left side of the equality to the right side and defining $R(t)$ by

$$R(t)=(A+2t\,\delta A)^T(\delta b-\delta A\,x(t))+\delta A^T\,(b-Ax(t)),$$

we obtain that

$$\big(A(t)^TA(t)+\lambda(t)I\big)x'(t)+\lambda'(t)x(t)=R(t).$$

It follows from $x^T(t)x'(t)=0$ that

$$(3.6)\qquad \lambda'(t)=\frac{x(t)^T\big(A(t)^TA(t)+\lambda(t)I\big)^{-1}R(t)}{x(t)^T\big(A(t)^TA(t)+\lambda(t)I\big)^{-1}x(t)},$$

and that

$$(3.7)\qquad |\lambda'(t)|\le\frac{\|\big(A(t)^TA(t)+\lambda(t)I\big)^{-1/2}R(t)\|_2}{\|\big(A(t)^TA(t)+\lambda(t)I\big)^{-1/2}x(t)\|_2}\le\sqrt{\frac{\sigma_1^2(A(t))+\lambda(t)}{\sigma_n^2(A(t))+\lambda(t)}}\|R(t)\|_2.$$

By the well-known differential mean value theorem, there exists $\xi_1\in(0,\,1)$ such that

$$|\hat\lambda-\lambda^*|=|\lambda(1)-\lambda(0)|=|\lambda'(\xi_1)|.$$

The first result follows.

To estimate $\|\hat x-x^*\|_2$, we define $\phi(t)=\|x(t)-x(0)\|_2$. Let $t_0=\max\{\,t\le 1\mid x(t)=x(0)\}$. Without loss of generality, we assume that $\hat x\ne x^*$, which implies $t_0<1$. It can be verified that $\phi(t)$ is differentiable in the interval $(t_0,\,1]$ and

$$\begin{aligned}\phi'(t)&=\frac{1}{\phi(t)}(x(t)-x(0))^Tx'(t)\\ &=\frac{1}{\phi(t)}(x(t)-x(0))^T\big(A(t)^TA(t)+\lambda(t)I\big)^{-1}\big(R(t)-\lambda'(t)x(t)\big).\end{aligned}$$

It follows that

$$(3.8)\qquad |\phi'(t)|\le\|\big(A(t)^TA(t)+\lambda(t)I\big)^{-1}\big(R(t)-\lambda'(t)x(t)\big)\|_2.$$

Substituting $\lambda'(t)$ of (3.6) into (3.8) gives that

$$\begin{aligned}(3.9)\qquad |\phi'(t)|&\le\|(A(t)^TA(t)+\lambda(t)I)^{-1/2}\|_2\,\|(A(t)^TA(t)+\lambda(t)I)^{-1/2}R(t)\|_2\\ &\le\frac{1}{\sigma_n^2(A(t))+\lambda(t)}\|R(t)\|_2.\end{aligned}$$

The result needed follows from

$$\|\hat x-x^*\|_2=\phi(1)-\phi(0)=\phi'(\xi_2)$$

TABLE 1
Sensitivities.

β	$\lvert\hat{\lambda}-\lambda\rvert$	$\mathrm{cond}_\lambda\eta$	$\lVert\hat{x}-x\rVert_2$	$\mathrm{cond}_x\eta$
1.0e−1	2.0118e−9	1.2942e−8	4.6724e−9	2.3245e−8
1.0e−3	2.5193e−8	1.2846e−7	3.4947e−7	2.3450e−6
1.0e−6	6.2968e−7	4.0620e−6	1.3120e−3	2.3452e−3

with a scalar $\xi_2 \in (t_0, 1)$. ∎

In general, for small δA and δb

$$\lambda'(\xi_1) \approx \lambda'(0), \quad \phi'(\xi_2) \approx \phi'(0).$$

By (3.7) and (3.9), we have that

$$|\hat{\lambda} - \lambda^*| \approx \sqrt{\frac{\sigma_1^2 + \lambda^*}{\sigma_n^2 + \lambda^*}}\,\eta, \quad \|\hat{x} - x^*\|_2 \approx \frac{1}{\sigma_n^2 + \lambda^*}\,\eta,$$

where

$$\eta = \|R(0)\|_2 = \|A^T(\delta b - \delta A\, x^*) + \delta A^T\,(b - Ax^*)\|_2.$$

It means that the condition numbers for computing λ^* and x^* are given by

$$\mathrm{cond}_\lambda = \sqrt{\frac{\sigma_1^2 + \lambda^*}{\sigma_n^2 + \lambda^*}} \quad \text{and} \quad \mathrm{cond}_x = \frac{1}{\sigma_n^2 + \lambda^*},$$

respectively.

Therefore the LSQE problem will be well conditioned if $\lambda^* > 0$ is not very small; even the matrix A itself is ill-conditioned. When λ^* is tightly close to $-\sigma_n^2$, LSQE will be very ill-conditioned. Below is a small example that confirms the conclusion.

Example. Let

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1.5 \\ \alpha \end{bmatrix}, \quad \delta A = \begin{bmatrix} 0 & 0 \\ \epsilon & 1 \end{bmatrix}, \quad \text{and} \quad \delta b = \begin{bmatrix} \epsilon \\ 0 \end{bmatrix},$$

where $\epsilon = 10^{-9}$. We write $\alpha = \frac{\sqrt{\beta(6+\beta)}\beta}{3+\beta}$ with $\beta > 0$. It is easy to verify that

$$\lambda^* = -1 + \beta, \quad x = \frac{1}{3+\beta}\begin{bmatrix} 3 \\ \sqrt{\beta(6+\beta)} \end{bmatrix};$$

i.e., the parameter β is just the gap between λ^* and $-\sigma_n^2$. The error term η reads

$$\eta = \|A^T(\delta b - \delta A x) + \delta A^T\,(b - Ax)\|_2 = \sqrt{4 + \left(\frac{3(3-\beta)}{2(3+\beta)}\right)^2}\,\epsilon \approx 2.3\epsilon.$$

In Table 1 we list the computed errors of λ and x produced by the perturbations of A and b and the first-order parts $\mathrm{cond}_\lambda * \eta$ and $\mathrm{cond}_x * \eta$.

As shown above, the LSQE problem may be ill-conditioned for well-conditioned matrix A. The following theorem indicates what will happen when A has clustered smallest singular values, which occurs in most applications for ill-posed problems.

THEOREM 3.2. *Assume that A has clustered smallest singular values*

$$\sigma_{k+1} \approx \cdots \approx \sigma_n \quad \textit{and} \quad \sigma_k \gg \sigma_{k+1}$$

which satisfy

$$(\sigma_k^2 - \sigma_n^2)^2 > \sigma_k^2 \sum_{i=1}^{k} \left(u_i^T b\right)^2,$$

where u_i is the left singular vector of A corresponding to σ_i. If $\lambda^* \le \sigma_k^2$, then

$$(\sigma_n^2 + \lambda^*)^2 \le \sigma_n^2 \frac{\sum_{i=k+1}^{n} \left(u_i^T b\right)^2}{1 - \frac{\sigma_k^2}{(\sigma_k^2 - \sigma_n^2)^2} \sum_{i=1}^{k} \left(u_i^T b\right)^2}.$$

Proof. Using the SVD of matrix A, we have

$$1 = \|x^*\|_2^2 = \sum_{i=1}^{n} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda^*)^2} \left(u_i^T b\right)^2.$$

By the assumption $\lambda^* \le \sigma_k^2$, one can easily verify that

$$\frac{\sigma_i^2}{(\sigma_i^2 + \lambda^*)^2} \le \frac{\sigma_k^2}{(\sigma_k^2 + \lambda^*)^2} \quad (i \le k), \qquad \frac{\sigma_i^2}{(\sigma_i^2 + \lambda^*)^2} \le \frac{\sigma_n^2}{(\sigma_n^2 + \lambda^*)^2} \quad (i > k).$$

It follows that

$$1 \le \frac{\sigma_k^2}{(\sigma_k^2 + \lambda^*)^2} \sum_{i=1}^{k} \left(u_i^T b\right)^2 + \frac{\sigma_n^2}{(\sigma_n^2 + \lambda^*)^2} \sum_{i=k+1}^{n} \left(u_i^T b\right)^2.$$

Substituting $(\sigma_k^2 + \lambda^*)^2 \ge (\sigma_k^2 - \sigma_n^2)^2$ gives

$$\frac{\sigma_n^2}{(\sigma_n^2 + \lambda^*)^2} \sum_{i=k+1}^{n} \left(u_i^T b\right)^2 \ge 1 - \frac{\sigma_k^2}{(\sigma_k^2 - \sigma_n^2)^2} \sum_{i=1}^{k} \left(u_i^T b\right)^2,$$

completing the proof. ∎

For many ill-posed problems, the matrix A is ill-conditioned with clustered smallest singular values and $\sigma_n \approx 0$. In that case,

$$(\sigma_n^2 + \lambda^*)^2 \le \sigma_n^2 \frac{\sum_{i=k+1}^{n} \left(u_i^T b\right)^2}{1 - \frac{\sigma_k^2}{(\sigma_k^2 - \sigma_n^2)^2} \sum_{i=1}^{k} \left(u_i^T b\right)^2} \approx \sigma_n^2 \frac{\sum_{i=k+1}^{n} \left(u_i^T b\right)^2}{1 - \sum_{i=1}^{k} \left(u_i^T b\right)^2 / \sigma_k^2},$$

which indicates that the LSQE problem is also ill-conditioned.

4. Newton methods. In [8, 14, 15, 16], algorithms for solving the minimization problem (1.2) are discussed. Basically, those algorithms were designed for the case when the largest Lagrange multiplier λ^* is positive, which is ensured by the assumption $\|A^\dagger b\|_2 > 1$. In this section, we review some Newton-type methods, one of which is discussed in [12, 26]. These Newton's methods will be compared with our projection scheme proposed in the next section.

Let $x(\lambda)$ be defined as in (3.1) and $f(\lambda)$ as in (1.5). We denote $y(\lambda)$ by

$$y(\lambda) = (A^T A + \lambda I)^{-1} x(\lambda). \tag{4.1}$$

It can be easily verified that $y(\lambda) = -x'(\lambda)$ and

$$f'(\lambda) = -2x^T(\lambda)y(\lambda), \quad f''(\lambda) = 6\|y(\lambda)\|_2^2.$$

Clearly the secular function $f(\lambda)$ is convex in each continuous interval and strictly monotone decreasing in the tightest continuous interval $(-\sigma_n^2, \infty)$.

There are many Newton-type iterations for solving $f(\lambda) = 1$ or other mathematically equivalent equations. Below we review three such schemes. For simplicity, we denote

$$x_k = x(\lambda_k), \quad y_k = y(\lambda_k).$$

[Newton I.] Applying Newton's method to $f(\lambda) = 1$ yields

$$\lambda_{k+1} = \lambda_k + \frac{\|x_k\|_2^2 - 1}{2x_k^T y_k}. \tag{4.2}$$

In our experience, the following Newton scheme (4.3), which was first used in [26], is much better than (4.2).

[Newton II.] Applying Newton's method to $f(\lambda)^{-1/2} = 1$ yields

$$\lambda_{k+1} = \lambda_k + \frac{(\|x_k\|_2 - 1)\|x_k\|_2^2}{x_k^T y_k}. \tag{4.3}$$

The basic idea of applying Newton's method to $\phi(\lambda) = 1$ is that we first construct the first-order approximation $\phi_k(\lambda) = \phi(\lambda_k) + \phi'(\lambda_k)(\lambda - \lambda_k)$ of $\phi(\lambda)$ at the previous guess λ_k and then solve $\phi_k(\lambda) = 1$ by replacing $\phi(\lambda) = 1$ to determine a new guess λ_{k+1}. This approach can work directly on the vector function $x(\lambda)$.

[Newton III.] By $z_k(\lambda)$ we define the first-order approximation of $x(\lambda)$ at $\lambda = \lambda_k$,

$$z_k(\lambda) = x_k - y_k \cdot (\lambda - \lambda_k).$$

Solving the minimization problem $\min | \|z_k(\lambda)\|_2^2 - 1 |$ yields the Newton scheme

$$\lambda_{k+1} = \lambda_k + \begin{cases} x_k^T y_k / y_k^T y_k, & \tilde{\Delta}_k \le 0, \\ (\|x_k\|_2^2 - 1)/(x_k^T y_k + \operatorname{sign}(x_k^T x_k)\sqrt{\tilde{\Delta}_k}), & \tilde{\Delta}_k > 0, \end{cases} \tag{4.4}$$

where

$$\tilde{\Delta}_k = (x_k^T y_k)^2 - (\|x_k\|_2^2 - 1)\|y_k\|_2^2.$$

The Newton iterations (4.2)–(4.4) converge monotonically if the starting value $\lambda_0 = 0$ when $\lambda^* > 0$ or $-\sigma_n^2 < \lambda_0 < \lambda^*$ when $\lambda^* < 0$. If we assume only that $\lambda_0 > -\sigma_n^2$, the new approximation λ_1 may be less than $-\sigma_n^2$. In that case the Newton sequence $\{\lambda_k\}$ may diverge or converge to a *pseudo*solution (a solution of $f(\lambda) = 1$ less than λ^*).

To guarantee convergence for general cases, it is necessary for the Newton-type methods (4.2)–(4.4) to check the inequality $\lambda_k > -\sigma_n^2$ or, equivalently, the positive

definiteness of the matrix $A^TA + \lambda I$. If $A^TA + \lambda I$ is not positive definite, one should increase λ_k and check again with the modified λ_k. This process should be repeated until the positive definite condition is satisfied. There are two issues that must be considered for this approach: cost of checking the positive definiteness and effectiveness of updating λ_k. For small and medium-sized problems, one can use bidiagonalization to compute σ_n and then check the inequality $\lambda_k > -\sigma_n^2$ [8], provided the smallest singular value of A is well separated from the others. For ill-conditioned A, the smallest singular value can't be computed reliably. On the other hand, assuming σ_n is approximately known, there are no easy ways to efficiently update λ_k so that the updated λ_k is a much better approximate to λ^* than λ_{k-1} and the condition $\lambda_k > -\sigma_n^2$ holds also. Bisection-like methods may result in slower convergence. In the next section, we will discuss a projection method that works well without checking the positive definiteness of $A^TA + \lambda I$.

5. Projection method. It is well known that Newton methods have only locally quadratic convergence. For the LSQE problem, the function $f(\lambda)$ has poles that may attract iterative points and then result in divergence. This motivates us to consider an approximation of $f(\lambda)$ that removes the poles and, at the same time, it is also a good approximation near λ^*. Differing from methods working on secular functions, our basic idea is to project the vector $x(\lambda)$ itself onto a one-dimensional subspace, say the subspace spanned by $w(\lambda)$, where the approximate vector $w(\lambda)$ should have no poles. As is known, the orthogonal projection $P_{w(\lambda)}x(\lambda)$ of $x(\lambda)$ is the (unique) optimal approximation to $x(\lambda)$ in the subspace $\mathrm{span}(w(\lambda))$,

$$\|P_{w(\lambda)}x(\lambda) - x(\lambda)\|_2 = \min_{y \in \mathrm{span}(w(\lambda))} \|y - x(\lambda)\|_2.$$

The following well-known lemma shows the gap between the two functions $f(\lambda) = \|x(\lambda)\|_2^2$ and $\|P_{w(\lambda)}x(\lambda)\|_2^2$.

LEMMA 5.1. *Assume $x \in R^n$ and $w \neq 0 \in R^n$. Let $P_w = ww^T/\|w\|_2^2$ be the orthogonal projector onto the subspace spanned by w. Then*

$$\left| \|x\|_2^2 - \|P_w x\|_2^2 \right| = \|(I - P_w)x\|_2^2.$$

When $w(\lambda)$ is available, we solve the secular equation $\|P_{w(\lambda)}x(\lambda)\|_2 = 1$ or, more generally, the minimization problem

$$\min_{\lambda} \left| \, \|P_{w(\lambda)}x(\lambda)\|_2^2 - 1 \, \right| \tag{5.1}$$

to get a good approximation of the optimal multiplier λ^*. From the viewpoint of numerical computation, besides the consideration of removing poles the vector function $w(\lambda)$ should be chosen such that

(1) it is quite cheap to evaluate $w(\lambda)$ for given λ,
(2) the residual $\|(I - P_{w(\lambda)})x(\lambda)\|_2$ is small, and
(3) it is not difficult to solve problem (5.1).

To this end, let

$$A = U\Sigma V^T = [u_1, \ldots, u_n] \, \mathrm{diag}(\sigma_1, \ldots, \sigma_n)[v_1, \ldots, v_n]^T$$

be the SVD of A. It is easy to verify that

$$x(\lambda) \approx \frac{\sigma_j c_j}{\sigma_j^2 + \lambda} v_j \quad \text{as} \quad \lambda \to -\sigma_j^2,$$

where $c_j = u_j^T b$. To remove the poles $-\sigma_j^2$, $w(\lambda)$ should be chosen such that $v_j^T w(\lambda) = O(\sigma_j^2 + \lambda)$, or at least $v_j^T w(\lambda) \to 0$ as $\lambda \to \lambda_j$. Assume that a previous guess λ_0 of λ^* is available. We have

$$x_0 = x(\lambda_0) = \sum \frac{\sigma_j c_j}{\sigma_j^2 + \lambda_0} v_j \quad \text{and} \quad y_0 = -x'(\lambda_0) = \sum \frac{\sigma_j c_j}{(\sigma_j^2 + \lambda_0)^2} v_j.$$

Taking inner products with v_j yields

$$v_j^T x_0 = \frac{\sigma_j c_j}{\sigma_j^2 + \lambda_0}, \quad v_j^T y_0 = \frac{\sigma_j c_j}{(\sigma_j^2 + \lambda_0)^2}.$$

Hence we have

$$v_j^T \big(x_0 + (\lambda - \lambda_0) y_0\big) = \frac{\sigma_j c_j}{\sigma_j^2 + \lambda_0} (\sigma_j^2 + \lambda),$$

which indicates that the required vector $w(\lambda)$ can be chosen as

$$w_0(\lambda) = x_0 + (\lambda - \lambda_0) y_0. \tag{5.2}$$

There are some interesting properties of the vector $w(\lambda)$. First we restate the orthogonality of $w_0(\lambda)$ to v_j as the following lemma.

LEMMA 5.2. *If $\lambda_0 \neq -\sigma_j$, then*

$$v_j^T w_0(\lambda) = O(\sigma_j^2 + \lambda) \quad \textit{as} \quad \lambda \to -\sigma_j^2.$$

Note that the vector $w_0(\lambda)$ in (5.2) differs from the "tangent line" $z_0(\lambda)$ of $x(\lambda)$ at $\lambda = \lambda_0$ used in Newton III. We refer to $w_0(\lambda)$ as the *skew-tangent* line at $\lambda = \lambda_0$. More interestingly, the inner product $w_0(\lambda)^T x(\lambda)$ is a constant with respect to $\lambda \neq -\sigma_j^2$, which can be easily verified by the expressions

$$x(\lambda) = B(\lambda) x_0, \quad w_0(\lambda) = B(\lambda)^{-1} x_0,$$

where $B(\lambda) = (A^T A + \lambda I)^{-1}(A^T A + \lambda_0 I)$. This property ensures that the problem (5.1) with $w(\lambda) = w_0(\lambda)$ can be easily solved because $\|P_{w_0(\lambda)} x(\lambda)\|_2^2$ is a rational function with respect to λ,

$$\phi_0(\lambda) \equiv \|P_{w_0(\lambda)} x(\lambda)\|_2^2 = \frac{\|x_0\|_2^4}{\|x_0\|_2^2 + 2(\lambda - \lambda_0) x_0^T y_0 + (\lambda - \lambda_0)^2 \|y_0\|_2^2}. \tag{5.3}$$

Obviously it is quite cheap to evaluate $\phi_0(\lambda)$ if x_0 and y_0 are known. The following lemma shows the quadratic approximation to $f(\lambda)$.

LEMMA 5.3. *Let $f(\lambda)$ be defined as in* (1.5) *and $\phi_0(\lambda)$ as in* (5.3). *Assume that $\lambda_0 \neq -\sigma_j^2$. Then*

(1) *$\phi_0(\lambda_0) = f(\lambda_0)$ and $\phi_0'(\lambda_0) = f'(\lambda_0)$;*

(2) *for all $\lambda \neq -\sigma_j^2$, $\phi_0(\lambda) \leq f(\lambda)$.*

The proof is simple and hence not given here. The function $\phi_0(\lambda)$ can be viewed as a rational Hermitian interpolation of $f(\lambda)$. Recalling that $f(\lambda)$ is also a rational function, $\phi_0(\lambda)$ is tightly close to $f(\lambda)$ for λ in the vicinity of λ_0. See Figure 2 for an example. It makes sense to choose λ_1, the largest solution of $\phi_0(\lambda) = 1$, as a better approximation to λ^* than λ_0 if $\phi_0(\lambda) = 1$ is solvable. Obviously, $\phi_0(\lambda) = 1$ has a real solution $\lambda_1 = \lambda_0 + (\sqrt{\Delta_0} - x_0^T y_0)/\|y_0\|_2^2$ if and only if $\Delta_0 \geq 0$, where $\Delta_0 = (x_0^T y_0)^2 + (\|x_0\|_2^2 - 1)\|x_0\|_2^2 \|y_0\|_2^2$. If $\Delta_0 < 0$, we use the unique maximum point

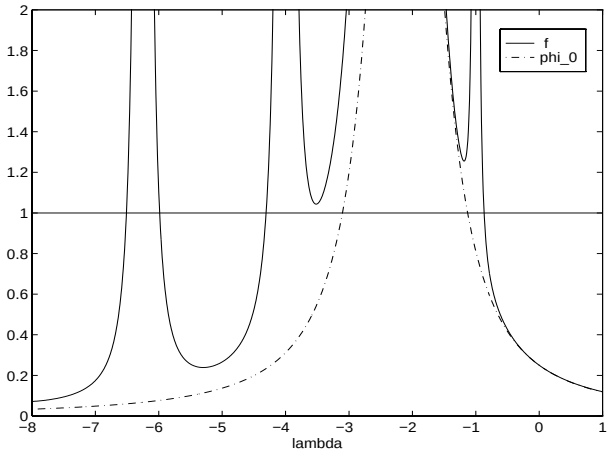


FIG. 2. *Functions $f(\lambda)$ (solid line) and $\phi_0(\lambda)$ (dashed line).*

of $\phi_0(\lambda)$ as λ_1, i.e., $\lambda_1 = \lambda_0 - x_0^T y_0/\|y_0\|_2^2$. This approach leads to our projection method for solving the LSQE problem (1.1) as follows:

$$\lambda_{k+1} = \lambda_k + \begin{cases} -x_k^T y_k / y_k^T y_k, & \Delta_k \le 0, \\ (\sqrt{\Delta_k} - x_k^T y_k)/y_k^T y_k, & \Delta_k > 0, \end{cases} \tag{5.4}$$

where

$$\Delta_k = (x_k^T y_k)^2 + (\|x_k\|_2^2 - 1)\|x_k\|_2^2 \|y_k\|_2^2. \tag{5.5}$$

In general, we define $\phi_k(\lambda)$ by

$$\phi_k(\lambda) \equiv \|P_{w_k(\lambda)} x(\lambda)\|_2^2 = \frac{\|x_k\|_2^4}{\|x_k\|_2^2 + 2x_k^T y_k(\lambda - \lambda_k) + \|y_k\|_2^2(\lambda - \lambda_k)^2} \tag{5.6}$$

corresponding to the skew-tangent line at λ_k,

$$w_k(\lambda) = x_k + y_k(\lambda - \lambda_k). \tag{5.7}$$

It can be verified that if $\Delta_k \ge 0$, then $\phi_k(\lambda_{k+1}) = 1$ and $\|x_{k+1}\|_2 \ge 1$ because $\|x_{k+1}\|^2 = f(\lambda_{k+1}) \ge \phi_k(\lambda_{k+1})$. In the case when $\Delta_k < 0$, which implies $\|x_k\|_2 < 1$, we have by Lemma 5.3 that

$$\|x_k\|_2^2 \le \frac{\|x_k\|_2^4 \|y_k\|_2^2}{\|x_k\|_2^2 \|y_k\|_2^2 - (x_k^T y_k)^2} = \phi_k(\lambda_{k+1}) \le f(\lambda_{k+1}) = \|x_{k+1}\|_2^2, \tag{5.8}$$

i.e., $\|x_{k+1}\|_2 \ge \|x_k\|_2$. In the next section, we will show that the sequence $\{\|x_k\|_2\}$ tends to 1 generally. A detailed convergence analysis for our projection method will be given in the next section.

6. Convergence analysis. First, we show the boundedness of the sequence $\{\lambda_k\}$ produced by (5.4) with any $\lambda_0 \ne -\sigma_j^2$. Note that there are no such properties for Newton sequences discussed in section 5.

LEMMA 6.1. *If $\lambda_0 \neq -\sigma_j^2$ for all j, then for all $k \geq 1$,*

$$-\sigma_1^2 \leq \lambda_k \leq \lambda^*. \tag{6.1}$$

Proof. Obviously, by definition we have that for $k \geq 0$

$$\lambda_{k+1} \geq \lambda_k - \frac{x_k^T y_k}{\|y_k\|_2^2} = \lambda_k - \frac{y_k^T (A^T A + \lambda_k I) y_k}{\|y_k\|_2^2} = -\frac{\|Ay_k\|_2^2}{\|y_k\|_2^2} \geq -\sigma_1^2.$$

On the other hand, if $\Delta_k \leq 0$,

$$\lambda_{k+1} = \lambda_k - \frac{x_k^T y_k}{\|y_k\|_2^2} = -\frac{\|Ay_k\|_2^2}{\|y_k\|_2^2} \leq -\sigma_n^2 \leq \lambda^*.$$

If $\Delta_k \geq 0$, $1 = \phi_k(\lambda_{k+1}) \leq f(\lambda_{k+1})$. It yields that the inequality $\lambda_{k+1} \leq \lambda^*$ is also true because $f(\lambda) < 1$ holds for all $\lambda > \lambda^*$, completing the proof. ∎

One can verify that the equality

$$\frac{y_k^T (A^T A + \lambda_k I) y_k}{y_k^T y_k} = \sigma_n^2 + \lambda_k$$

holds if and only if $A^T b$ is a right singular vector corresponding to the smallest singular value σ_n. In that case, λ^* can be immediately determined as

$$\lambda^* = \|A^T b\|_2 - \frac{\|A^T A A^T b\|_2}{\|A^T b\|_2}.$$

We therefore assume throughout the remaining part of the paper that $A^T b$ is not a right singular vector corresponding to the smallest singular value σ_n. The assumption ensures that all the λ_k produced by (5.4) are well defined and $\lambda_k \neq -\sigma_n^2$.

Lemma 6.1 shows that the optimal multiplier λ^* is bounded from below by λ_k. This property may motivate one to modify λ_{k+1}, when $\lambda_{k+1} < \lambda_k$, to keep $\{\lambda_k\}$ increasing. Theorem 6.2 below, however, guarantees automatically the monotonicity of $\{\lambda_k\}$ provided $\|x_0\|_2 > 1$.

THEOREM 6.2. *Let $\|x_0\|_2 \geq 1$. Then the sequence $\{\lambda_k\}$ of (5.4) converges increasingly to $\tilde{\lambda}$ which satisfies $f(\tilde{\lambda}) = 1$ and $f'(\tilde{\lambda}) \leq 0$.*

Proof. The condition $\|x_k\|_2 \geq 1$ implies that $\Delta_k \geq (x_k^T y_k)^2$. By definition,

$$\lambda_{k+1} - \lambda_k = \frac{\sqrt{\Delta_k} - x_k^T y_k}{\|y_k\|_2^2} \geq 0 \tag{6.2}$$

and $1 = \phi_k(\lambda_{k+1}) \leq f(\lambda_{k+1}) = \|x_{k+1}\|_2^2$. Hence we have that $\lambda_k \leq \lambda_{k+1} \leq \lambda^*$. By induction again, we can conclude from the given condition $\|x_0\|_2 \geq 1$ that the sequence $\{\lambda_k\}$ is monotonically increasing and bounded from above by λ^* and that $\|x_k\| \geq 1$ for all $k \geq k_0$. It guarantees that there exists a $\tilde{\lambda} \leq \lambda^*$ such that $\lambda_k \to \tilde{\lambda}$ as $k \to \infty$.

Now we show that $\tilde{\lambda}$ is not a pole of $f(\lambda)$. Suppose, to the contrary, $\{\lambda_k\}$ tends increasingly to a pole of $f(\lambda)$, say $-\sigma_{j_0}^2$, $f(\lambda_k) \to +\infty$, and $x_k^T y_k \to -\infty$. We obtain that for sufficiently large k, $x_k^T y_k \leq 0$, $f'(\lambda_k) = \|x_k\|_2^2 > 2$, and

$$\min_j |\sigma_j^2 + \lambda_k| = -(\sigma_{j_0}^2 + \lambda_k).$$

By definition,

$$\begin{aligned}\lambda_{k+1} - \lambda_k &= \frac{\sqrt{(x_k^T y_k)^2 + (\|x_k\|_2^2 - 1)\|x_k\|_2^2\|y_k\|_2^2} - x_k^T y_k}{\|y_k\|_2^2} \\ &\geq \frac{\sqrt{(\|x_k\|_2^2 - 1)\|x_k\|_2^2\|y_k\|_2^2}}{\|y_k\|_2^2} \\ &> \frac{\|x_k\|_2}{\|y_k\|_2} \quad = \quad \frac{\|(A^T A + \lambda_k I)y_k\|_2}{\|y_k\|_2} \\ &\geq \min_j |\sigma_j^2 + \lambda_k| = -(\sigma_{j_0}^2 + \lambda_k).\end{aligned}$$

It leads to a contradiction: $\lambda_{k+1} > -\sigma_{j_0}^2 = \tilde{\lambda}$. We therefore conclude that $x_k \to \tilde{x} \equiv x(\tilde{\lambda})$ and $y_k \to \tilde{v} \equiv y(\tilde{\lambda})$. It follows from (6.2) that

$$\sqrt{(\tilde{x}^T\tilde{v})^2 + (\|\tilde{x}\|_2^2 - 1)\|\tilde{x}\|_2^2\|\tilde{v}\|_2^2} - \tilde{x}^T\tilde{v} = 0.$$

Hence we have that $f'(\tilde{\lambda}) = -2\tilde{x}^T\tilde{v} \leq 0$ and $f(\tilde{\lambda}) = \|\tilde{x}\|_2^2 = 1$, completing the proof. □

The sequence $\{\lambda_k\}$ must be convergent and $\|x_k\|_2 \to 1$ if there is $\|x_{k_0}\|_2 \geq 1$, although the limit $\tilde{\lambda}$ may be less than λ^*. In general $\|x_k\|_2 \geq 1$ always holds for some k. Practically, we can conclude from the following theorem that there are a finite number of different x_k with norms less than 1 if $\{\|x_k\|_2\}$ does not converge to 1.

THEOREM 6.3. *If $\|x_k\|_2 < 1$ for all k, then $\{\|x_k\|_2\}$ is increasing and there exists k_0 such that for all $k \geq k_0$*

$$\lambda_k = \lambda_{k_0} < -\sigma_n^2 \quad \textit{and} \quad f'(\lambda_k) = 0.$$

Proof. The monotonicity of $\{\|x_k\|_2\}$ has been shown in (5.8) because the condition that $\|x_k\|_2 < 1$ holds for all k implies $\Delta_k < 0$. So $\{\|x_k\|_2\}$ is convergent and has the limit less than or equal to 1. Recalling that both sequences $\{\lambda_k\}$ and $\{x_k\}$ are bounded, we can pick up, respectively, convergent subsequences $\{\lambda_{n_k}\}$ and $\{x_{n_k}\}$. Let

$$\lambda_{n_k} \to \hat{\lambda} \leq \lambda^* \quad \text{and} \quad x_{n_k} \to \hat{x} \equiv x(\hat{\lambda}) \neq 0.$$

Obviously, $\hat{\lambda}$ is not a pole of $f(\lambda)$ because $\{\|x_{n_k}\|_2\}$ is bounded. We have that $y_{n_k} \to \hat{v} \equiv y(\hat{\lambda})$. By (5.8), we can verify that

$$0 \leq (x_{n_k}^T y_{n_k})^2 \leq \left(1 - \frac{\|x_{n_k}\|_2^2}{\|x_{n_k+1}\|_2^2}\right)\|x_{n_k}\|_2^2\|y_{n_k}\|_2^2 \longrightarrow 0$$

as $k \to \infty$. It follows that $f'(\hat{\lambda}) = -2\hat{x}^T\hat{v} = 0$; i.e., $f(\hat{\lambda})$ is a local minimum. Hence there is a constant $\delta > 0$ such that $f(\hat{\lambda}) \leq f(\lambda)$ holds for all $\lambda \in (\hat{\lambda} - \delta, \hat{\lambda} + \delta)$. Thus $\|\hat{x}\|_2 \leq \|x_{k_0}\|_2$ because $\lambda_{k_0} \in (\hat{\lambda} - \delta, \hat{\lambda} + \delta)$ for a certain index k_0. By the monotonicity

of $\{\|x_k\|_2\}$ we conclude that for all $k \geq k_0$

$$\|\hat{x}\|_2 \leq \|x_{k_0}\|_2 \leq \|x_k\|_2 \leq \|\hat{x}\|_2,$$

which implies $\|x_k\|_2 = \|\hat{x}\|_2$. It follows from (5.8) and (5.4) that $x_k^T y_k = 0$ and $\lambda_k = \lambda_{k_0} < -\sigma_n^2$ for all $k \geq k_0$ because $f'(\lambda) < 0$ for $\lambda > -\sigma_n^2$. ∎

Practically, we still have $\|x_k\|_2 \geq 1$ for some k because of finite arithmetic operations in numerical computation. Therefore for almost all the starting λ_0, the sequence $\{\lambda_k\}$ converges to a $\tilde{\lambda}$ that satisfies $f(\tilde{\lambda}) = 1$ and $f'(\tilde{\lambda}) < 0$. Note that both the equality and inequality hold for λ^*, too. Obviously, one can verify that the sequence $\{\lambda_k\}$ determined by (5.4) converges to λ^*, i.e., $\{x_k\}$ tends to the solution x^* of the problem (1.1), provided $\lambda_0 > -\sigma_n^2$ and $\|x_0\|_2 > 1$. It is worthwhile to point out that under the same assumption, the Newton iterations (4.2)–(4.4) are also convergent. However, we can further show that our projection scheme has stronger convergence. To this end, let $\underline{\lambda}$ be the largest value satisfying

$$f(\underline{\lambda}) \leq 1 \quad \text{and} \quad f'(\underline{\lambda}) = 0.$$

(If there exists no such $\underline{\lambda}$, we set $\underline{\lambda} = -\infty$.) It can be proven that $\underline{\lambda} < -\sigma_n^2$. The following theorem indicates a larger convergence interval for the choice of a starting value λ_0.

THEOREM 6.4. *For any $\lambda_0 \in (\underline{\lambda}, -\sigma_n^2) \cup (-\sigma_n^2, \lambda^*)$, the iteration sequence $\{\lambda_k\}$ of* (5.4) *converges monotonically to λ^*.*

Proof. We need only prove the theorem for $\lambda_0 \in (\underline{\lambda}, -\sigma_n^2)$. By Theorem 6.2, we need to show that there exists $\lambda_k > \underline{\lambda}$ such that $\|x_k\|_2 \geq 1$. Theorem 6.3 shows that it is true if $\underline{\lambda} = -\infty$. Therefore we can assume that $\underline{\lambda}$ is finite valued. By the definition of $\underline{\lambda}$, we have that for all $\lambda \in (\underline{\lambda}, -\sigma_n^2)$, $f'(\lambda) = -2x(\lambda)y(\lambda) > 0$. Hence we get $x_0^T y_0 < 0$ and

$$\lambda_1 > \lambda_0$$

whether the inequality Δ_0 holds or not. If $\lambda_1 < -\sigma_n^2$, we get again that $x_1^T y_1 < 0$ and $\lambda_2 > \lambda_1$, and so on. Therefore, by Theorem 6.3, one of the following cases must hold:

(1) There exists k such that $\lambda_k < -\sigma_n^2 < \lambda_{k+1}$.

(2) There exists k such that $\|x_k\|_2 > 1$.

By the proof of Theorem 6.1, the condition $\lambda_k < -\sigma_n^2 < \lambda_{k+1}$ implies that $\Delta_k > 0$ and $\|x_{k+1}\|_2 \geq 1$. Therefore we conclude that there is an integer k such that $\|x_k\|_2 \geq 1$ and $\lambda_k > \underline{\lambda}$, completing the proof. ∎

Remark. It is possible there is no $\underline{\lambda}$ such that $f'(\underline{\lambda}) = 0$ and $f(\underline{\lambda}) \leq 1$. In that case, the projection iterations $\{\lambda_k\}$ converge to λ^* for any starting λ_0 if every λ_k is well defined; i.e., λ_k is not a pole of $f(\lambda)$. Below is a small example demonstrating the phenomenon.

Example. The test matrix A is chosen in the form $A = Q\,\mathrm{diag}(1, 1.5, 2, 2.5, 3)Q^T$ with orthogonal matrix Q of order $n = 5$. We choose the vector b such that $c = Q^T b$ has components listed as

$$\begin{array}{r} -0.13377080593009 \\ 0.50677250315177 \\ -0.55034836933116 \\ -0.39905657003380 \\ -0.38842364243402. \end{array}$$

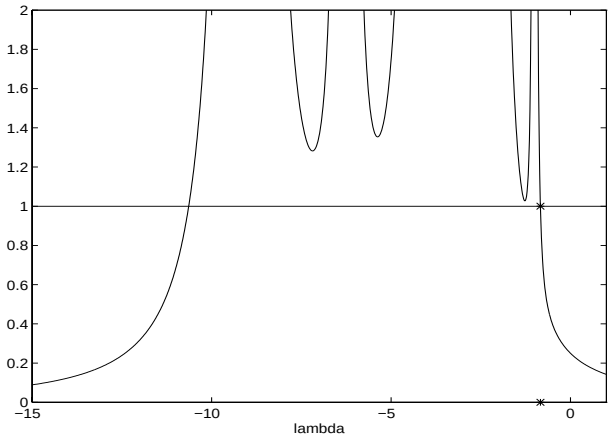


FIG. 3. *Function $f(\lambda)$ for which the projection algorithm converges started at any $\lambda_0 \neq -\sigma_j^2$.*

Figure 3 plots the function $f(\lambda)$. Our projection method converges quickly to $\lambda^* = -0.81866763239804$ for any starting λ_0. In the following table we list the starting values λ_0 and the numbers of the iterations needed for the projection method to achieve the accuracy $|\,\|x_k\|_2 - 1\,| < 1.e-14$:

λ_0	2	0	−2	−4	−6	−8	−10	−12	−14
Iteration number	9	9	9	10	13	15	14	15	15

Newton's methods I and III diverge for all λ_0 while Newton II, started at $\lambda_0 = 0$, achieves accuracy after 42 iterations.

Now we consider the convergence rate of the projection method. The results of Theorem 6.5 following can be guessed from Lemma 5.3.

THEOREM 6.5. *Let $\lambda_k \to \tilde{\lambda}$ as $k \to \infty$. If $f'(\tilde{\lambda}) \neq 0$, then*

$$\overline{\lim_{k\to\infty}} \frac{\tilde{\lambda} - \lambda_{k+1}}{(\tilde{\lambda} - \lambda_k)^2} \leq \frac{2}{3}\frac{f''(\tilde{\lambda})}{f'(\tilde{\lambda})} - f'(\tilde{\lambda}),$$

where $\overline{\lim}$ means a superior limit.

Proof. By the mean value theorem, there exists $\xi_1 \in (\lambda_{k+1}, \tilde{\lambda})$ depending on k such that

$$f(\tilde{\lambda}) - f(\lambda_{k+1}) = f'(\xi_1)(\tilde{\lambda} - \lambda_{k+1}). \tag{6.3}$$

On the other hand, for sufficiently large k, $\phi_k(\lambda_{k+1}) = 1 = f(\tilde{\lambda})$. Applying the mean value theorem and Lemma 5.3 again we obtain that

$$\begin{aligned} f(\tilde{\lambda}) - f(\lambda_{k+1}) &= \phi_k(\lambda_{k+1}) - f(\lambda_{k+1}) \\ &= \frac{1}{2}\left(f''(\xi_2) - \phi_k''(\xi_2)\right)(\lambda_{k+1} - \lambda_k)^2 \end{aligned} \tag{6.4}$$

for certain $\xi_2 \in (\lambda_k, \lambda_{k+1})$. Combining (6.3) and (6.4) and using the inequalities $\lambda_k \leq \lambda_{k+1} < \tilde{\lambda}$ we have

$$\frac{\tilde{\lambda} - \lambda_{k+1}}{(\tilde{\lambda} - \lambda_k)^2} \leq \frac{\tilde{\lambda} - \lambda_{k+1}}{(\lambda_{k+1} - \lambda_k)^2} = \frac{f''(\xi_2) - \phi_k''(\xi_2)}{2f'(\xi_1)}.$$

Let $k \to \infty$, ξ_1, $\xi_2 \to \tilde{\lambda}$ also. Note that $\phi_k(\xi_2) \to 2(f'(\tilde{\lambda}))^2 - \frac{1}{3}f'(\tilde{\lambda})$ as $k \to \infty$. We have that

$$\overline{\lim_{k\to\infty}} \frac{\tilde{\lambda} - \lambda_{k+1}}{(\tilde{\lambda} - \lambda_k)^2} \leq \frac{2}{3}\frac{f''(\tilde{\lambda})}{f'(\tilde{\lambda})} - f'(\tilde{\lambda}),$$

completing the proof. ∎

Theorem 6.5 shows that the projection method has at least quadratic convergence. A similar result holds for the vector sequence $\{x_k\}$.

THEOREM 6.6. *Let $\lambda_k \to \tilde{\lambda}$ as $k \to \infty$, and let $\tilde{x} = x(\tilde{\lambda})$. If $f'(\tilde{\lambda}) \neq 0$, then the vector sequence $\{x_k\}$ converges to $\tilde{x}$ quadratically.*

Proof. By definition, we have that

$$(A^TA + \tilde{\lambda})\tilde{x} = A^Tb \quad \text{and} \quad (A^TA + \lambda_k)x_k = A^Tb,$$

which yields that

$$(\lambda_k - \tilde{\lambda})x_k = (A^TA + \tilde{\lambda})(\tilde{x} - x_k).$$

Multiplying $(A^TA + \tilde{\lambda})^{-1}$ by the left, we obtain that for all k

$$\tilde{x} - x_k = (\lambda_k - \tilde{\lambda})(A^TA + \tilde{\lambda})^{-1}x_k.$$

Therefore, $x_k \to \tilde{x}$ as $k \to \infty$. Furthermore,

$$\frac{\|\tilde{x} - x_{k+1}\|_2}{\|\tilde{x} - x_k\|_2^2} = \frac{\|(A^TA + \tilde{\lambda})^{-1}x_{k+1}\|_2}{\|(A^TA + \tilde{\lambda})^{-1}x_k\|_2^2} \frac{\tilde{\lambda} - \lambda_{k+1}}{(\tilde{\lambda} - \lambda_k)^2}$$

$$\approx \|(A^TA + \tilde{\lambda})^{-1}\tilde{x}\|_2^{-1} \frac{\tilde{\lambda} - \lambda_{k+1}}{(\tilde{\lambda} - \lambda_k)^2}.$$

The result follows directly from Theorem 6.5. ∎

7. Corrections. As shown in the proof of Lemma 6.1, the inequality $\Delta_k < 0$ implies that $\lambda_{k+1} \leq -\sigma_n^2 < \lambda^*$. In that case, λ_{k+1} could not be a good approximation to λ^*. It is therefore required to modify the guess λ_{k+1}. In this section, we will consider two strategies to improve the projection method. First, the function $\phi_k(\lambda)$ will be modified by adding a positive term $\delta_k(\lambda)$ to it, and then we will choose a suitable solution of the minimization problem

$$\min \Big| \big(\phi_k(\lambda) + \delta_k(\lambda)\big) - M_k \Big|, \tag{7.1}$$

rather than that of (5.1), to be λ_{k+1}. Here the constant $M_k < 1$, depending on the previous approximation λ_k. We point out emphatically that λ_{k+1} will be modified only in the case when $\Delta_k < 0$. Otherwise we keep λ_{k+1} unchanged.

To this end, a natural step is to choose the corrector

$$\delta_k(\lambda) = \|(I - P_{w_k(\lambda)})x_k\|_2^2$$

because $f(\lambda) - \phi_k(\lambda) = \|(I - P_{w_k(\lambda)})x(\lambda)\|_2^2$. See (5.7) for the definition of $w_k(\lambda)$. A simple calculation yields that

$$\delta_k(\lambda) = \|x_k\|_2^2 - \frac{\big(x_k^Tw_k(\lambda)\big)^2}{\|w_k(\lambda)\|_2^2} = \frac{(\lambda - \lambda_k)^2\big(\|x_k\|_2^2\|y_k\|_2^2 - (x_k^Ty_k)^2\big)}{\|x_k\|_2^2 + 2(\lambda - \lambda_k)x_k^Ty_k + (\lambda - \lambda_k)^2\|y_k\|_2^2}.$$

Obviously, the modified approximate $\psi_k(\lambda) \equiv \phi_k(\lambda) + \delta(\lambda)$ to $x(\lambda)$ is also a rational function

$$\psi_k(\lambda) = \frac{\|x_k\|_2^4 + (\lambda - \lambda_k)^2\big(\|x_k\|_2^2\|y_k\|_2^2 - (x_k^T y_k)^2\big)}{\|x_k\|_2^2 + 2(\lambda - \lambda_k)x_k^T y_k + (\lambda - \lambda_k)^2\|y_k\|_2^2}.$$

However, for the case when $\|x_k\|_2$ is relatively small, adding $\delta_k(\lambda) = \|(I - P_{w_k(\lambda)})x_k\|_2^2$ to $\phi_k(\lambda)$ does not improve approximation. A better and more flexible step is to introduce a factor $\rho_k > 0$ to x_k in the expression of $\delta_k(\lambda)$ to get a new corrector

$$\delta_k(\lambda) = \|(I - P_{w_k(\lambda)})(x_k * \rho_k)\|_2^2. \tag{7.2}$$

This consideration leads to the following approximation to $f(\lambda)$:

$$\psi_k(\lambda,\, \rho_k) = \frac{\|x_k\|_2^4 + \rho_k^2(\lambda - \lambda_k)^2\big(\|x_k\|_2^2\|y_k\|_2^2 - (x_k^T y_k)^2\big)}{\|x_k\|_2^2 + 2(\lambda - \lambda_k)x_k^T y_k + (\lambda - \lambda_k)^2\|y_k\|_2^2}. \tag{7.3}$$

Since the correction process is adopted in the case when $\Delta_k < 0$ or, equivalently, $\phi_k(\lambda) = 1$ has no solution, we have that $\|x_k\|_2 < 1$. We therefore suggest writing ρ_k in the form

$$\rho_k = \|x_k\|_2^\alpha \tag{7.4}$$

with a certain parameter α. It is likely that the projection method converges to λ^* for a smaller parameter α, while the number of iterations will be larger. Although we have no idea how to choose the "best" sequence ρ_k or the parameter α such that the scheme has stronger convergence and less number of iterations, $\alpha = -0.5$ is often a quite good choice. See Test 3 of section 8.

Now let us consider the equation

$$\psi_k(\lambda,\, \rho_k) = M_k. \tag{7.5}$$

We will choose the constant M_k to be the maximal value of $\phi_k(\lambda)$, i.e.,

$$M_k = \max \phi_k(\lambda) = \frac{\|x_k\|_2^4\|y_k\|_2^2}{\|x_k\|_2^2\|y_k\|_2^2 - (x_k^T y_k)^2}.$$

It is easy to show that (7.5) has two solutions. We choose the largest one as a new approximate λ_{k+1}^ψ to λ^* when $\lim_{\lambda\to\infty} \psi(\lambda,\, \rho_k) \le M_k$. Otherwise we use the smallest solution as λ_{k+1}^ψ. It can be verified that the new approximation has the following expression:

$$\lambda_{k+1}^\psi = \lambda_k - \frac{x_k^T y_k}{\|y_k\|_2^2 + \rho_k\|x_k\|_2^{-2}\big(\|x_k\|_2^2\|y_k\|_2^2 - (x_k^T y_k)^2\big)}. \tag{7.6}$$

Setting $\rho_k = \|x_k\|_2^\alpha$ gives

$$\lambda_{k+1}^\psi = \lambda_k - \frac{x_k^T y_k}{\|y_k\|_2^2 + \|x_k\|_2^{\alpha-2}\big(\|x_k\|_2^2\|y_k\|_2^2 - (x_k^T y_k)^2\big)}. \tag{7.7}$$

Clearly, $\lambda_{k+1}^\psi \to \lambda_{k+1}$ as $\alpha \to +\infty$ because $\|x_k\|_2 < 1$.

Remark. The convergence results discussed in section 6 are also true for the iteration sequence produced by the projection method with corrections.

Remark. As we mentioned above, we just use the new approximation λ_{k+1}^{ψ} replacing λ_{k+1} defined in (5.4) only when $\Delta_k < 0$. One can always use the solution of the problem

$$\min \Big| \psi_k(\lambda, \rho_k) - 1 \Big|$$

to replace directly λ_{k+1} whether the condition $\Delta_k < 0$ holds or not. However, numerical experiments show that such a scheme is not competitive with the approaches proposed above. So we omit the discussion about it. More interestingly, the iteration scheme produced by the solution of $\min | \psi_k(\lambda, \rho_k) - 1 |$ with $\rho_k = \|x_k\|_2^{-1}$ is just the Newton scheme (4.3) for the case $x_k^T y_k > 0$.

8. Numerical experiments. In this section, we will present several numerical experiments to illustrate the effectiveness of the proposed projection method. For the first three tests, we construct, with MATLAB notation, the test matrices A using the following steps:

$$[\mathtt{U}, \mathtt{R}] = \mathtt{qr}(\mathtt{rand}(\mathtt{n}));$$

$$[\mathtt{V}, \mathtt{R}] = \mathtt{qr}(\mathtt{rand}(\mathtt{n}));$$

$$\mathtt{A} = \mathtt{U} * \mathtt{diag}(\mathtt{s}) * \mathtt{V}^{\mathtt{T}};$$

We also construct the vectors b as $\mathtt{b} = (\mathtt{1} - \mathtt{2} * \mathtt{rand}(\mathtt{n}, \mathtt{1})) * \mathtt{c}$ with certain chosen constant $c > 0$ such that

$$\|A^{\dagger} b\|_2 = r$$

for given $r < 1$. The set of singular values $s = (s(1), \ldots, s(n))$ will be chosen according to different test purposes. In all our numerical tests, the inequality $\lambda^* > -\sigma_n^2$ always holds. The starting value λ_0 is naturally chosen to be zero. The iteration process will terminate in the kth step if x_k corresponding to λ_k satisfies

$$\big| \, \|x_k\|_2 - 1 \, \big| < \epsilon.$$

By convergence we mean that the iteration sequence $\{\lambda_k\}$ converges to λ^* or, equivalently, $\{x_k\}$ converges to the solution x^* of the LSQE problem (1.1). Test 4 and Test 5 are two applications on a Procrustes problem and an ill-posed problem, respectively. We simply use the SVD of A to compute the normal equations for x_k and y_k at each iteration step, because the tested matrices A are dense and the matrix sizes are not very large.

[Test 1]. First we compare the efficiency of our proposed projection method (5.4) (without using the correction technique discussed in section 7) with Newton's methods (4.2)–(4.4). The test matrices are randomly chosen with singular values `s = 10*rand(1,n)` for $n = 20$ and `r = rand(1)`. 1000 pairs of such matrices A and vectors b are chosen. For each pair of A and b we implement the projection method and the Newton schemes, respectively. The convergence accuracy is chosen as $\epsilon = \mathtt{1.e-7}$. In the table below we list the number of successes for which the corresponding method is convergent and the average iteration number required to achieve the given accuracy.

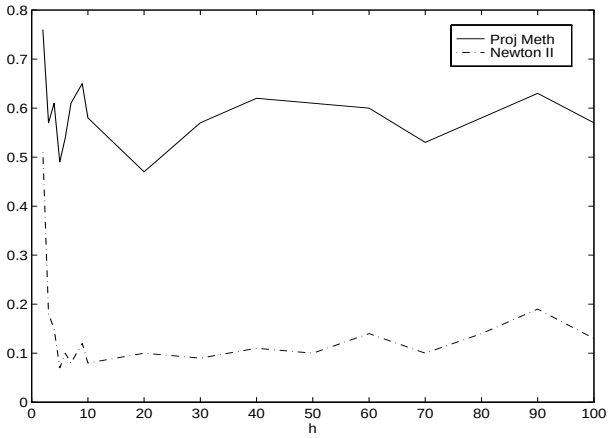


FIG. 4. *Percentage of successes.*

	PM	NT1	NT2	NT3
Number of successes	845	337	624	368
Average number of iterations	4.0	6.7	4.0	4.8

In this table, column PM corresponds to the projection method while columns NT1–NT3 correspond to Newton's methods. The numerical experiments show that even when the correction technique is not used, our projection method has a higher number of successes than Newton-type iterations. For these tests, Newton methods NT1–NT3 have moderate success because generally $r = \|A^\dagger b\|_2$ is not small for those cases. (The average value of r is 0.5078.) Newton's methods will fail if $\|A^\dagger b\|_2$ is small. See the next test for details.

[Test 2]. We now look at the effect of the size of $\|A^\dagger b\|_2$ on the convergence. The test data A and b are constructed so that

$$\|A^\dagger b\|_2 = 1/h \tag{8.1}$$

holds for h chosen from the set

$$[2:9,\ 10:10:100].$$

We construct matrix A with unit distributed singular values

$$\mathtt{s} = \mathtt{linspace}(1, 10, \mathtt{n})$$

and $n = 20$. For each h, we also implement the projection method with no corrections and Newton's methods 100 times with different $A = U * \mathrm{diag}(s)V^T$ and b which satisfy (8.1). In Figure 4 we plot, respectively, the curves of the percentage of convergent tests for the projection method with no corrections and Newton iteration II via h ($\epsilon = 1.e-7$). Newton I and III fail for almost all the tests and the results are deleted here. In general, the larger the value h is, the less the percentage of successes is.

[Test 3]. The purpose of this test is to compare the effectiveness of correction approaches discussed in section 7. The function $\phi_k(\lambda)$ will be modified by adding the corrector $\delta_k(\lambda)$ defined in (7.2) with the factor ρ_k of the form

$$\rho_k = \|x_k\|_2^\alpha$$

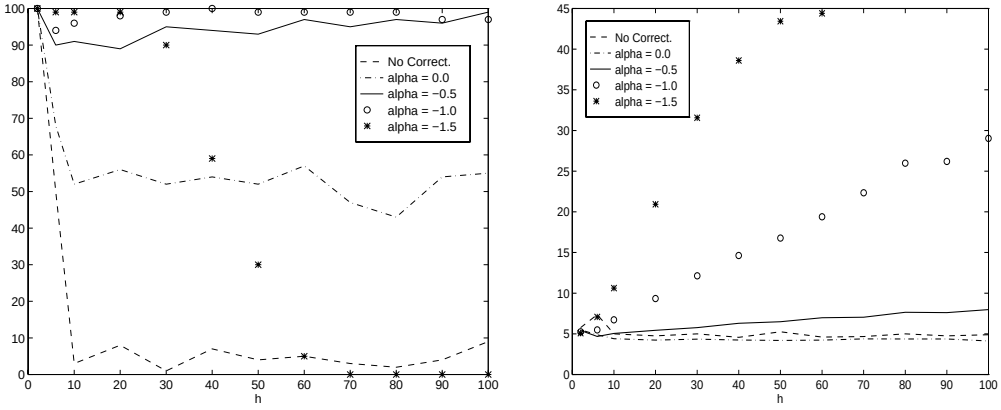


FIG. 5. *Percentage of successes (left) and average number of iterations (right).*

to $\phi_k(\lambda)$ only when $\max \phi_k(\lambda) < 1$. Note that $\|x_k\|_2 < 1$ if the correction step is implemented. As we mentioned before, it is possible to strengthen the convergence of the projection method by decreasing the value of α. Meanwhile, the iteration number required to achieve the given accuracy will increase, too. To illustrate this, we use four different values of α,

$$[0,\ -0.5,\ -1.0,\ -1.5].$$

Note that the projection scheme (7.7) with $\alpha = +\infty$ is just the projection scheme *without* correction. Test matrices we used have singular values with the following distributions:

Unit distributed singular values	$\mathtt{s} = [10 : -1 : 1]$;
Isolated smallest singular values	$\mathtt{s} = [10 : -0.1 : 9.2,\ 1]$;
Clustered smallest singular values	$\mathtt{s} = [10 : -0.5 : 7,\ 3.0 : 0.1 : 4,\ 1.5 : -0.1 : 1]$,

respectively. For each matrix A and each scale h chosen from the set

$$[2,\ 6,\ 10 : 10 : 100],$$

the right vector b is also randomly chosen and scaled to satisfy $\|A^{\dagger}b\|_2 = 1/h$. As in Test 2, for each h, 100 pairs of such $(A,\ b)$ are tested for the projection method (5.4) and those with the correction approaches. We also use $\epsilon = 1.e-7$.

Figure 5 plots the percentages of successes achieving the accuracy among the 100 tests (left) and the average number of iterations (right), corresponding to the matrices that have a cluster of smallest singular values

$$\mathtt{s} = [10 : -0.5 : 7,\ 3.0 : 0.1 : 4,\ 1.5 : -0.1 : 1]$$

with different singular vector matrices. These numerical examples indicate that for a certain choice of α the projection method with correction (7.7) can solve the LSQE problem within several iterations at a very high percentage. For matrices with the three different kinds of distributions of singular values, $\alpha = -0.5$ is still the best one among the four choices; for more than 90% of the implemented examples the

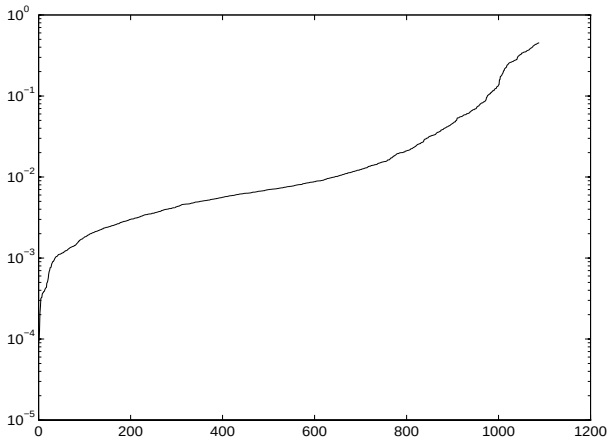


FIG. 6. *Gaps between* λ^* *and* $-\sigma_n^2(A)$.

projection method converges within eight iterations. Since the vector b is scaled such that $\|A^\dagger b\|_2$ is small for each test, all of the 1200 numerical examples have multipliers λ^* close to $-\sigma_n^2(A)$. We plot in Figure 6 the gaps between $-\sigma_n^2(A)$ and λ^* computed by the modified projection method with $\alpha = -0.5$ for a total of 1088 convergent tests. By the way, Newton II is convergent only for nine pairs A and b among the 1000 tests corresponding to $h \geq 10$ if A has a cluster of smallest singular values. For the matrices with unit distributed singular values or the isolated smallest singular value, our projection method succeeds at a higher percentage.

[Test 4]. Now we consider an application of the orthogonal Procrustes problem arising in factor analysis. The matrix A of order 60×11 is from [17]. It is well-conditioned and has clustered smallest singular values. We list its singular values below.

```
4.69983097105684
3.01248135577805
2.30884044477344
1.82861073858085
1.50301206135281
1.43270111380274
1.26420982332002
1.14341898893269
1.09792059765636
1.03306991250800
1.01934648149101
```

We randomly choose vector $\hat{b}$ as $\hat{b} = 1 - 2 * rand(m, 1)$. The vector $\hat{b}$ is then normalized as $b = \hat{b}/\|\hat{x}\|_2$, where vector $\hat{x}$ satisfies the normal equations $(A^T A + \lambda^* I)\hat{x} = \hat{b}$ for previously chosen $\lambda^* = \xi - \sigma_n^2$ with the parameter $\xi > 0$. The normalized vector $x^* = \hat{x}/\|\hat{x}\|_2$ is the optimal solution of the LSQE problem corresponding to the optimal multiplier λ^*. We use different values of ξ to control the ill-conditioning level of the LSQE problem. Table 2 gives the numerical results for the stopping criterion $\epsilon = 1.e - 7$.

[Test 5]. In this test, we show the numerical results of our projection algorithm

TABLE 2
Results of the projection method on the Procrustes problem.

ξ	$\|A^\dagger b\|_2$	$\big\|\|x\|_2 - 1\big\|$	$\|x - x^*\|_2$	k
1.00e−05	5.3769e−05	3.3307e−16	2.2204e−16	45
1.00e−04	5.3767e−04	4.4409e−16	4.4409e−16	18
1.00e−03	5.3565e−03	4.4409e−16	3.3337e−16	9
1.00e−02	4.4460e−02	8.4856e−11	9.3588e−11	5
1.00e−01	2.0570e−01	2.8721e−09	3.0860e−09	5

TABLE 3
Results of the projection method on the inverse heat equation.

ϵ_{noise}	$\|A^\dagger b\|_2$	$\big\|\|x\|_2 - 1\big\|$	$\|x - x^*\|_2$	k
1.00e−06	1.65e+11	9.61e−04	4.38e−02	10
1.00e−04	1.65e+13	5.37e−04	3.20e−02	11
1.00e−02	1.65e+15	9.50e−04	3.39e−02	17
1.00e−01	1.65e+16	8.58e−04	1.03e−01	13
3.00e−01	4.95e+16	1.83e−05	2.32e−01	12
5.00e−01	8.25e+16	4.66e−06	3.47e−01	11

applied on a regularization problem. The problem is to compute the unknown function $h(t)$ of the following inverse heat equation, a first kind Volterra integral equation with the integration interval [0,1]:

$$\int_0^1 K(s,t)h(t)dt = g(s),$$

for given kernel $K(s,t) = k(s-t)$ with

$$k(t) = \begin{cases} \frac{1}{2\kappa\sqrt{\pi t^3}} \exp(-\frac{1}{4\kappa^2 t}), & t > 0, \\ 0, & t \le 0, \end{cases} \tag{8.2}$$

and a right-hand side function $g(s)$ [6, 19]. The constant κ controls the ill-conditioning of the problem. We use the MATLAB routine `heat(n,kappa)` of regularization tools [19] with $n = 1000$ and $\kappa = 6$ to construct the matrix A, which is the discretization of kernel $K(s,t)$ at discrete points s_i in the integral interval [0, 1], and a discrete solution h_{DISC} with the right-hand side vector g_{DISC} produced as $g_{DISC} = A * h_{DISC}$. The matrix A has condition number $\text{cond}(A) = 1.3850e19$ and two smallest singular values

$$2.503232055747098\text{e}-19, \quad 6.140117608941735\text{e}-20.$$

We normalize h_{DISC} to get $x^* = h_{DISC}/\|h_{DISC}\|_2$ and denote $\hat{b} = g_{DISC}/\|h_{DISC}\|_2$. The right-hand side vector will be changed as $b = \hat{b} + \delta b$ with a noise vector δb that is randomly chosen and scaled to have a given relative noise level $\epsilon_{noise} = \|\delta b\|_2/\|\hat{b}\|_2$. In Table 3 we list the numerical results of the projection method applied to the problem for different noise levels.

9. Concluding remarks. In this paper, we have presented a projection method combined with a correction technique for computing numerically the least squares problem with quadratic equality constraints (1.1) when the LSQE problem is ill-conditioned, i.e., the optimal multiplier λ^* is negative and close to the largest pole

$-\sigma_n^2(A)$ of the secular function $f(\lambda) = \|(A^TA + \lambda I)^{-1}A^Tb\|_2^2$. We also gave detailed structure and perturbation analysis to demonstrate the sensitivity of the LSQE problem. Our algorithm has some obvious advantages over Newton's method and variants. It has a wider convergence range for a choice of initial approximation λ_0. In fact, in the new algorithm, the choice of a starting value λ_0 is not crucial; for any choice of starting value $\lambda_0 \neq -\sigma_j^2(A)$, the algorithm always produces a monotonic and bounded sequence $\{\lambda_k\}$ of multipliers. Numerical experiments indicate that the projection method with corrections is much more efficient than Newton's methods which almost always fail when $\|A^\dagger b\|_2$ is relatively small or A has clustered smallest singular values. We didn't touch the problem of computing the normal equations for large sparse problems.

Several issues deserve further investigation: 1) For the ill-conditioned problem LSQE, if an iterative approach such as the conjugate gradient or GMRES method [28] is used for the normal equations $(A^TA + \lambda I)x = A^Tb$ with given λ, there are some issues that need to be considered: preconditioning technique, initial guess, and suitable stopping criterion for the inner iteration (if we refer to the projection method as the outer iteration). If λ_0 is a good approximation to λ^* and $(A^TA+\lambda_0 I)$ is positive definite, the Cholesky decomposition of $(A^TA + \lambda_0 I)$ may be a good preconditioner. Clearly, one can use x_{k-1} as the initial guess of the inner iteration for the normal equations $(A^TA + \lambda_k I)x = A^Tb$. 2) It is still not clear whether $\alpha = -0.5$ or some other α is the optimal parameter for our correction approach presented in section 7. In our numerical experimentations, $\alpha = -0.5$ is always a good choice for the correction approach with which the projection method converges within a small number of iterations that is less than the iterations required in general. 3) It is also an interesting problem to investigate the relation with trust-region methods for ill-conditioned TLS problems [27].

Acknowledgments. We want to thank the anonymous referees for their careful reading of this paper. Their insightful suggestions and comments greatly improved the presentation.

REFERENCES

[1] A.A. Anda and H. Park, *Self-scaling fast rotations for stiff least squares problems*, Linear Algebra Appl., 234 (1996), pp. 137–162.
[2] J.L. Barlow, N.K. Nichols, and R.J. Plemmons, *Iterative methods for equality-constrained least squares problems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 892–906.
[3] Å. Björck, *A direct method for sparse least squares problems with lower and upper bounds*, Numer. Math., 54 (1988), pp. 19–32.
[4] Å. Björck, *Component-wise perturbation analysis and errors bounds for linear least square solutions*, BIT, 31 (1991), pp. 238–244.
[5] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[6] A. Carasso, *Determining surface temperatures from interior observations*, SIAM J. Appl. Math., 42 (1982), pp. 558–574.
[7] P.S. Dwyer, *A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solution*, Ann. Math. Statist., 15 (1944), pp. 82–89.
[8] L. Eldén, *Algorithms for the regularization of ill-conditioned least squares problems*, BIT, 17 (1977), pp. 134–145.
[9] L. Eldén, *Perturbation theory for the least squares problem with linear equality constraints*, SIAM J. Numer. Anal., 17 (1980), pp. 338–350.
[10] L. Eldén and H. Park, *A Procrustes Problem on the Stiefel Manifold*, Technical Report LiTH-MAT-R-97-6, Department of Mathematics, Linköping University, Sweden, 1995.
[11] W. Gander, *Least squares with a quadratic constraint*, Numer. Math., 36 (1981), pp. 291–307.

[12] W. Gander, G.H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114/115 (1989), pp. 815–839.
[13] G.H. Golub, *Numerical methods for solving least squares problems*, Numer. Math., 7 (1965), pp. 206–216.
[14] G.H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[15] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore and London, 1996.
[16] G.H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.
[17] M. Gulliksson, *Algorithms for the Partial Procrustes Problems*, Report UMINF-95, ISSN-0348-0542, Department of Computing Science, Umeå University, Sweden, 1995.
[18] M. Hanke and P.C. Hansen, *Regularization methods for large-scale problems*, Surveys Math. Indust., 3 (1993), pp. 253–315.
[19] P.C. Hansen, *Regularization Tools: A MATLAB package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[20] M.T. Heath, *Numerical methods for large sparse linear least squares problems*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 497–513.
[21] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Classics in Applied Mathematics 15, SIAM, Philadelphia, 1995.
[22] A. Neumaier, *Solving ill-conditioned and singular linear systems: A tutorial on regularization*, SIAM Rev., 40 (1998), pp. 636–666.
[23] C.C. Paige, *Fast numerically stable computations for generalized least squares problems*, SIAM J. Numer. Anal., 16 (1979), pp. 165–171.
[24] G. Peters and J.H. Wilkinson, *The least squares problem and pseudo-inverse*, J. Comput., 13 (1970), pp. 309–316.
[25] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.
[26] C.H. Reinsch, *Smoothing by spline functions* II, Numer. Math., 16 (1971), pp. 451–454.
[27] M. Rojas and Sorensen, *A Trust-Region Approach to the Regularization of Large-Scale Discrete Ill-Posed Problems*, Technical Report 99-26, Department of Computational and Applied Mathematics, Rice University, Houston, 1999.
[28] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[29] G.W. Stewart and J.G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 213–223

SOME RESULTS ON VANDERMONDE MATRICES WITH AN APPLICATION TO TIME SERIES ANALYSIS*

ANDRÉ KLEIN† AND PETER SPREIJ‡

Abstract. In this paper we study Stein equations in which the coefficient matrices are in companion form. Solutions to such equations are relatively easy to compute as soon as one knows how to invert a Vandermonde matrix (in the generic case where all eigenvalues have multiplicity one) or a confluent Vandermonde matrix (in the general case). As an application we present a way to compute the Fisher information matrix of an autoregressive moving average (ARMA) process. The computation is based on the fact that this matrix can be decomposed into blocks where each block satisfies a certain Stein equation.

Key words. ARMA process, Fisher information matrix, Stein's equation, Vandermonde matrix, confluent Vandermonde matrix

AMS subject classifications. 15A09, 15A24, 62F10, 62M10, 65F05, 65F10, 93B25

PII. S0895479802402892

1. Introduction.

In this paper we investigate some properties of (confluent) Vandermonde and related matrices aimed at and motivated by their application to a problem in time series analysis. Specifically, we show how to apply results on these matrices to obtain a simpler representation of the (asymptotic) Fisher information matrix of an autoregressive moving average (ARMA) process. The Fisher information matrix is prominently featured in the asymptotic analysis of estimators and in asymptotic testing theory, e.g., in the classical Cramér–Rao bound on the variance of unbiased estimators. See [10] for general results and see [2] for time series models. However, the Fisher information matrix has also attracted considerable attention in the signal processing literature, e.g., [6], [19], and [12]. We have previously shown (see [14]) that the Fisher information matrix of an ARMA process is the solution of a so-called Lyapunov equation. More precisely, although we don't go into detail about ARMA processes until section 5, the Fisher information matrix in this case can be decomposed into blocks that are solutions of equations such as

$$X + MXN^\top = R.$$

The coefficients M and N in this equation turn out to be in companion form in the given context of time series analysis, and the right-hand side R is another given matrix.

The plan of attack that we follow to solve such an equation is to break up the solution procedure into a number of steps that are each relatively easy to perform. First, we replace by a basis transformation the coefficient matrices with their Jordan forms, thereby also changing the variable matrix X and the right-hand side R. Since a basis of (generalized) eigenvectors of companion matrices can be represented as the columns of a (confluent) Vandermonde matrix, the basis transformation needed

*Received by the editors February 21, 2002; accepted for publication (in revised form) by S. Van Huffel November 15, 2002; published electronically May 29, 2003.
http://www.siam.org/journals/simax/25-1/40289.html

†Department of Actuarial Sciences and Econometrics, Universiteit van Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands (aklein@fee.uva.nl).

‡Korteweg-de Vries Institute for Mathematics, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (spreij@science.uva.nl).

for this can be expressed in terms of the above-mentioned Vandermonde matrices. Performing the basis transformation requires knowing how to compute inverses of confluent Vandermonde matrices. One of the aims of our paper is to derive rather simple, but explicit, representations for these inverses. Of course this whole procedure would be meaningless if the equation in the new coordinate system were more complex than the original one. In section 4 we will see that, fortunately, the resulting equation is much easier to solve than the original one, especially in a generic case, where the solution becomes almost trivial. By applying the developed procedure to the computation of the Fisher information matrix for an ARMA process, we reach our goal of giving an alternative way to represent this Fisher information matrix. This application also motivates, from a statistical perspective, the interest of analyzing (confluent) Vandermonde matrices.

The remainder of the paper is organized as follows. In section 2 we introduce the basic notation that we use throughout the paper. Section 3 is devoted to technical results on companion matrices and confluent Vandermonde matrices, the main results concerning inversion of confluent Vandermonde matrices. In section 4 we apply these results to describe solutions to Stein equations in which the coefficient matrices are in companion form. Finally, in section 5 we investigate the special case where the solutions to certain Stein equations are given by blocks of the Fisher information matrix of an ARMA process.

2. Notation and preliminaries. Consider the matrix $A \in \mathbb{R}^{n\times n}$ in the companion form

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & 0 & 1 \\ -a_n & & & -a_2 & -a_1 \end{pmatrix}. \tag{1}$$

Let $a^\top = (a_1, \dots, a_n)$, $u(z)^\top = (1, z, \dots, z^{n-1})$, and $u^*(z)^\top = (z^{n-1}, \dots, 1)$ (where $^\top$ denotes transposition). Define recursively the Hörner polynomials $a_k(\cdot)$ by $a_0(z) = 1$ and $a_k(z) = za_{k-1}(z) + a_k$. Notice that $a_n(z)$ is the characteristic polynomial of A. We will denote it by $\pi(z)$ and, occasionally, by $\pi_A(z)$ if we want to emphasize the role of the A-matrix.

Write $a(z)$ for the n-vector $(a_0(z), \dots, a_{n-1}(z))^\top$. Furthermore S will denote the shift matrix, so $S_{ij} = \delta_{i,j+1}$, and P will denote the *backward* or *antidiagonal* identity matrix, so $P_{ij} = \delta_{i+j,n+1}$ (assuming that $P \in \mathbb{R}^{n\times n}$). As an example we have $Pu(z) = u^*(z)$. The matrix P has the following property: If M is a Toeplitz matrix, then $PMP = M^\top$, in particular $P^2 = I$, the identity matrix.

We associate with the vector a the matrix $T_a \in \mathbb{R}^{n\times n}$ given by

$$T_a = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \\ a_{n-1} & \cdots & a_1 & 1 \end{pmatrix}.$$

Notice that the matrices T_a and S commute and that $a(z) = T_a u(z)$.

Denoting the kth basis vector in $\mathbb{R}^n$ by e_k, we can write

$$A = -e_n a^\top P + S^\top. \tag{2}$$

If $q(\cdot)$ is a polynomial and if for some natural number k the term $(z-\alpha)^k$ is a factor of $q(z)$ (which happens if α is a zero of $q(\cdot)$ with multiplicity greater than or equal to k), then we define the polynomial $q_k(\cdot;\alpha)$ by $q_k(z;\alpha) = \frac{q(z)}{(z-\alpha)^k}$. Notice the identity $q_k(\alpha;\alpha) = q_k^{(k)}(\alpha)/k!$. In what follows we will often use D for differentiation (w.r.t. z). For instance, instead of $\frac{d}{dz}q_k(z;\alpha)$ we then write $Dq_k(z;\alpha)$, and $Dq_k(z;\alpha)$ in $z=\alpha$ is denoted by $Dq_k(\alpha;\alpha)$. Notice also the formula

$$\pi(z) - \pi(\alpha) = (z-\alpha)u^*(z)^\top a(\alpha), \tag{3}$$

which follows from the definition of the Hörner polynomials by a direct computation.

We also need some results on Lagrange and Hermite interpolation problems. Assume we are given s pairwise different complex numbers $\alpha_1, \ldots, \alpha_s$ (so $\alpha_i \neq \alpha_j$ iff $i \neq j$) and we want to find n polynomials $p_1, \ldots, p_n$ of degree at most $n-1$ such that $p_j(\alpha_i)$ take on certain given values. Notice that we have n^2 unknown parameters to determine, but only ns conditions. Therefore we add constraints by prescribing certain values of the derivatives $p_j^{(k)}(\alpha_i)$ for $k = 1, \ldots, m_i - 1$, where the m_i are such that $\sum_{i=1}^s m_i = n$. In this way we obtain n^2 constraints. The total set of prescribed values of the polynomials p_j and their derivatives that we consider is given by the equations

$$\frac{p_j^{(k-1)}(\alpha_i)}{(k-1)!} = \delta_{\sum_{l=1}^{i-1} m_l + k, j},$$

where $j = 1, \ldots, n$, $i = 1, \ldots, s$, $k = 1, \ldots, m_i$, and δ denotes the Kronecker symbol. Notice that in the case where $s = n$, all m_i are equal to 1, and we only require $p_j(\alpha_i) = \delta_{ij}$.

In order to give the solution to this interpolation problem an elegant form we present the conditions as described below. We need some notation. First, we denote by $p(z)$ the column vector $(p_1(z), \ldots, p_n(z))^\top$. For each i we denote by $\Pi(i)$ the $n \times m_i$ matrix with columns $\Pi(i)_k = \frac{p^{(k-1)}(\alpha_i)}{(k-1)!}$, with $k = 1, \ldots, m_i$. The constraints are now given in compact form by the equality $(\Pi(1), \ldots, \Pi(s)) = I$, where I is the $n \times n$ identity matrix.

Write $\pi(z) = \prod_{i=1}^s (z-\alpha_i)^{m_i} = \sum_{j=0}^n a_j z^{n-j}$ and let A be the associated companion matrix of (1) so that π is its characteristic polynomial. Let $U_i(z)$ be the $n \times m_i$ matrix with kth column equal to $\frac{1}{(k-1)!}u^{(k-1)}(z)$ and write $U_i = U_i(\alpha_i)$. We define the $n \times n$ matrix V (often called the *confluent* Vandermonde matrix associated with the eigenvalues of A) by $V = (U_1, \ldots, U_s)$. Similar interpolation problems involving one polynomial only are known to have a unique solution; see e.g., [17, p. 306] or [5, p. 37]. Here the situation is similar and, as an almost straightforward result from the current setup, we have the following proposition.

PROPOSITION 2.1. *The unique solution to the interpolation problem is* $p(z) = V^{-1}u(z)$.

Write $p^*(z) = z^{n-1}p(\frac{1}{z})$ and notice that we use multiplication with the same power of z for all entries of $p(\frac{1}{z})$.

Let Π^* be defined by $\Pi^* = V^{-1}PV$. Then the matrix Π^* is involutive, i.e., $(\Pi^*)^2 = I$.

PROPOSITION 2.2. *The polynomials p and p^* are related by*

$$p^*(z) = V^{-1}PVp(z) = \Pi^* p(z). \tag{4}$$

In particular, $p^*(0) = V^{-1}e_n$.

Proof. This follows from

$$p^*(z) = z^{n-1}V^{-1}u\left(\frac{1}{z}\right) = V^{-1}Pu(z) = V^{-1}PVp(z). \qquad \square$$

3. Confluent Vandermonde matrices. The main point of this section is to give some formulas for the inverse of a confluent Vandermonde matrix. We need some auxiliary results. First we give an expression for $\mathrm{adj}(z - A)$, where A is a companion matrix of the form (2). The next proposition is an alternative to formula (31) in [7, p. 84].

PROPOSITION 3.1. *Let A be a companion matrix with π as its characteristic polynomial. The following equation holds true:*

$$\mathrm{adj}(z - A) = u(z)a(z)^\top P - \pi(z)\sum_{j=0}^{n-1} z^j S^{j+1}. \tag{5}$$

Proof. First we show that

$$a(z)^\top P(z - A) = \pi(z)e_1^\top. \tag{6}$$

Using (2), we have

$$\begin{aligned} a(z)^\top P(z - A) &= a(z)^\top P(z - S^\top + e_n a^\top P) \\ &= a(z)^\top (z - S + Pe_n a^\top)P \\ &= (\pi(z)e_n^\top - a^\top + a(z)^\top Pe_n a^\top)P \\ &= \pi(z)e_n^\top P, \end{aligned}$$

which gives (6). Multiply the right-hand side of (5) by $(z - A)$. First we consider $a(z)^\top P(z - A)$. In view of (6), this is just

$$\pi(z)e_1^\top. \tag{7}$$

Then we consider $\sum_{j=0}^{n-1} z^j S^{j+1}(z - A) = \sum_{j=0}^{n-1} z^{j+1}S^{j+1} + \sum_{j=0}^{n-1} z^j S^{j+1}(-S^\top + e_n a^\top P)$. Since $Se_n = 0$, this reduces to $\sum_{j=0}^{n-1} z^{j+1}S^{j+1} - \sum_{j=0}^{n-1} z^j S^{j+1}S^\top$. Now use the equality $SS^\top = I - e_1 e_1^\top$ to rewrite this as $\sum_{j=0}^{n-1} z^j S^j(zS - I + e_1 e_1^\top)$, which equals $\sum_{j=0}^{n-1} z^j S^j(zS - I) + \sum_{j=0}^{n-1} z^j e_{j+1}e_1^\top$. However, this is equal to $-I + u(z)e_1^\top$ because the first summation is just $-I$ and the latter one equals $u(z)e_1^\top$. Hence

$$\sum_{j=0}^{n-1} z^j S^{j+1}(z - A) = -I + u(z)e_1^\top. \tag{8}$$

So we obtain from (7) and (8) that the right-hand side of (5) multiplied by $z - A$ is equal to

$$u(z)\pi(z)e_1^\top + \pi(z)(I - u(z)e_1^\top),$$

which is $\pi(z)I$, precisely what we have to prove. □

For the application to time series that we have in mind, as explained in the introduction, we need the inverse of a (confluent) Vandermonde matrix. In the 1970s this was an especially popular topic and many papers appeared on the subject. Quite often attention has been paid to the finding of efficient procedures to carry out the inversion numerically. Recently, there has been a renewed interest in a related subject, the inversion of Cauchy–Vandermonde matrices. These matrices appear in rational interpolation problems and are beyond the scope of this paper.

Below we provide inversion formulas for confluent Vandermonde matrices. Some of these can be found in the older literature, but the derivation below is different. Of the many possible references we mention [11] and [4], which give results for the relatively simple case of a genuine Vandermonde matrix or, in the spirit of our Proposition 3.3 (but obtained by different methods), for a confluent Vandermonde matrix, and mention [20] which has elementwise expressions. Related results of a different nature include [9], [3], and [18].

We need the Jordan decomposition of A. We use the notation S_{m_i} to denote the shift matrix of size $m_i \times m_i$. Recall that the confluent Vandermonde matrix as we defined it is such that the columns are independent eigenvectors of A. The Jordan form of A is determined by the relation $V^{-1}AV = J_A$, and J_A is block diagonal with the ith block given by $\alpha_i I_{m_i} + S_{m_i}^\top$. As a first step toward expressions for the inverse of a Vandermonde matrix we will use the next proposition.

PROPOSITION 3.2. *Let J_A be the Jordan form of the companion matrix A. Then*

$$\operatorname{adj}(z - J_A) = p(z)a(z)^\top PV - \pi(z)V^{-1}\sum_{j=0}^{n-1} z^j S^{j+1}V. \tag{9}$$

In particular

$$\operatorname{adj}(\alpha_k - J_A) = \pi(\alpha_k)a(\alpha_k)^\top PV. \tag{10}$$

Proof. This follows from Propositions 3.1 and 2.1. □

Next we proceed with some results of a general nature. Let M be the block diagonal matrix with s blocks $M(i)$ of size $m_i \times m_i$ specified by

$$M(i)_{kl} = \begin{cases} \frac{1}{(k+l-m_i-1)!} D^{k+l-m_i-1}\pi_{m_i}(\alpha_i;\alpha_i) & \text{if } k+l-m_i-1 \geq 0, \\ 0 & \text{else.} \end{cases} \tag{11}$$

Notice that the $M(i)$ are symmetric Hankel matrices and that the $M(i)_{kl}$ are zero for $k + l \leq m_i$. We have for the matrices $M(i)$ the alternative expression

$$M(i) = \sum_{l=0}^{m_i-1} \delta_l S^l P, \text{ where } \delta_l = \frac{1}{l!} D^l \pi_{m_i}(\alpha_i;\alpha_i).$$

Here we denoted by S the $m_i \times m_i$ shift matrix and by P the $m_i \times m_i$ backward identity matrix.

The computation of the inverse of an $M(i)$ is simple because of its triangular structure and the fact that it is Hankel. Indeed, it is sufficient to know the first row of $M(i)^{-1}$, call it r_1, since all rows r_j are of the form $r_1 S^{j-1}$. As a matter of fact, the inverses of the matrices $M(i)$ have a particular simple structure. To clarify this

we introduce, for a given $m-1$ times continuously differentiable real function f, the matrix valued function $L^f(z)$ of size $m \times m$ defined by

$$L^f_{kl}(z) = \begin{cases} \frac{1}{(k-l)!} D^{k-l} f(z) & \text{if } k \geq l, \\ 0 & \text{else.} \end{cases}$$

Notice that the matrices $L^f(z)$ are lower triangular and Toeplitz. One readily verifies that $(L^f(z))^{-1} = L^{\frac{1}{f}}(z)$ in the points z where f doesn't vanish. In particular, the last row of $(L^f(z))^{-1}$ is given by

$$\left(\frac{1}{f(z)}, \dots, \frac{1}{(m-1)!} D^{m-1} \left(\frac{1}{f(z)} \right) \right) P,$$

where P is, as above, of size $m \times m$.

Now we apply this result to $f(z) = \pi_{m_i}(z; \alpha_i)$ and $m = m_i$ to get the inverse of $M(i)$. We then have for this choice of f that $M(i) = L^f(\alpha_i)P$. The first row of $M(i)^{-1}$ is then seen to be

$$\left(\frac{1}{\pi_{m_i}(\alpha_i; \alpha_i)}, \dots, \frac{1}{(m_i-1)!} D^{m_i-1} \left(\frac{1}{\pi_{m_i}(\alpha_i; \alpha_i)} \right) \right) P. \tag{12}$$

Next we define a matrix N consisting of blocks $N(ij)$ of size $m_i \times m_j$. To do so we need some additional notation. We write $\pi^*(z) = z^n \pi(\frac{1}{z})$ and $\pi^*_k(z; \alpha) = z^{n-1} \pi_k(\frac{1}{z}; \alpha)$. Then we define the entries of the $N(ij)$ by

$$N(ij)_{kl} = \frac{1}{(k-1)!} D^{k-1} \pi^*_l(\alpha_i; \alpha_j).$$

Unfortunately, the matrix N doesn't share the nice properties (block diagonal, block Hankel, block symmetric) with the matrix M above.

PROPOSITION 3.3. *The following equalities hold:*

$$u^*(z)^\top T_a = a(z)^\top P, \tag{13}$$

$$u^*(z)^\top T_a V = (\pi_1(z; \alpha_1), \dots, \pi_{m_1}(z; \alpha_1), \dots, \pi_1(z; \alpha_s), \dots, \pi_{m_s}(z; \alpha_s)), \tag{14}$$

$$V^\top P T_a V = M, \tag{15}$$

$$u(z)^\top T_a V = (\pi^*_1(z; \alpha_1), \dots, \pi^*_{m_1}(z; \alpha_1), \dots, \pi^*_1(z; \alpha_s), \dots, \pi^*_{m_s}(z; \alpha_s)), \tag{16}$$

$$V^\top T_a V = N, \tag{17}$$

$$V^{-1} = M^{-1} V^\top P T_a = M^{-1} (T_a V)^\top P. \tag{18}$$

Proof. The equality (13) is the result of the string $u^*(z)^\top T_a = u(z)^\top P T_a = u(z)^\top T_a^\top P = a(z)^\top P$.

We continue with showing (14). Consider (3) and differentiate k times w.r.t. α. We obtain $-D^k \pi(\alpha) = u^*(z)^\top ((z-\alpha) D^k a(\alpha) - k D^{k-1} a(\alpha))$.

If α is a zero with multiplicity m, then $D^k \pi(\alpha) = 0$ for $k \leq m-1$. So we get the system of equations $0 = u^*(z)^\top ((z-\alpha) D^k a(\alpha) - k D^{k-1} a(\alpha))$ for $1 \leq k \leq m-1$ and $\pi(z) = (z-\alpha) u^*(z)^\top a(\alpha)$. Now write $q_k(z) = u^*(z)^\top D^k a(\alpha)$, then $q_0(z) = \frac{\pi(z)}{z-\alpha}$, and we have the recursive system of equations $0 = (z-\alpha) q_k(z) - k q_{k-1}(z)$ for $k = 1, \dots, m-1$. Solving this system yields $q_k(z) = k! \frac{\pi(z)}{(z-\alpha)^{k+1}} = k! \pi_{k+1}(z; \alpha)$. In other words, we find

$$u^*(z)^\top D^k a(\alpha) = k! \pi_{k+1}(z; \alpha). \tag{19}$$

Consider now $a(w) = T_a u(w) = T_a V p(w)$, where p is the interpolation polynomial. Then we also have $u^*(z)^\top a(w) = u^*(z)^\top T_a V p(w)$. Take in this equation derivatives w.r.t. w, substitute α_i for w, and use the definition of the interpolation polynomial to get

$$u^*(z)^\top T_a V = \left(a(\alpha_1), \dots, \frac{D^{m_1-1} a(\alpha_1)}{(m_1-1)!}, \dots, a(\alpha_s), \dots, \frac{D^{m_s-1} a(\alpha_s)}{(m_s-1)!} \right). \tag{20}$$

Combining (19) and (20) yields (14).

To prove (15) we start from (14). Take the appropriate jth order derivatives, divide by $j!$, and substitute the α_i in the resulting expression. Doing so results in a block diagonal matrix, with the $M(i)$ on the diagonal.

Equation (16) immediately follows from (14) by definition of the polynomials $\pi_k^*(z;\alpha)$.

The proof of (17) completely parallels that of (15) and is therefore omitted. Now we turn to (18). First we observe that all the matrices $M(i)$ are invertible because of their triangular structure and the nonzero elements $\pi_{m_i}(\alpha_i;\alpha_i)$ (α_i had multiplicity m_i) on the antidiagonal. Therefore M also is invertible and, taking inverses in (15), yields the first equality of (18). The second then follows from $PT_a = T_a^\top P$. □

Remark 3.4. The most important formula of Proposition 3.3 is (18), which gives an expression for the inverse of the confluent Vandermonde matrix. We see that the only inversion that has to be carried out is that of M. For that we have (12) at our disposal.

COROLLARY 3.5. *The matrices M and N are related through the identities*

$$M = N^\top \Pi^*, \tag{21}$$

$$N = (\Pi^*)^\top M. \tag{22}$$

Moreover $NM^{-1} = MN^{-1}$, and thus NM^{-1} is involutive.

Proof. From (17) we get $V^{-\top} = T_a V N^{-1}$, and hence $V^\top P V^{-\top} N = V^\top P T_a V$ and, in view of (15), this equals M. Now Π^* was defined as $\Pi^* = V^{-1} P V$, so we get $(\Pi^*)^\top N = M$ and, since M is symmetric, we obtain (21). However, we also have $N = (\Pi^*)^{-\top} M = (\Pi^*)^\top M$ since Π^* is involutive. For the same reason the final assertion of the corollary follows. □

In the next proposition we present integral representations for the matrices M and M^{-1}. Below we use the notation $u_{m_i}(z)^\top = (1, z, \dots, z^{m_i-1})$ and $u_{m_i}^*(z)^\top = (z^{m_i-1}, \dots, z, 1)$, and the Γ_{α_i} are sufficiently small contours around α_i.

PROPOSITION 3.6. *The following integral representations for the matrices $M(i)$ and $M(i)^{-1}$ are valid:*

$$M(i) = \frac{1}{2\pi i} \oint_{\Gamma_{\alpha_i}} u_{m_i}^*(z-\alpha_i) u_{m_i}^*(z-\alpha_i)^\top \frac{\pi(z)}{(z-\alpha_i)^{2m_i}} dz, \tag{23}$$

$$M(i)^{-1} = \frac{1}{2\pi i} \oint_{\Gamma_{\alpha_i}} u_{m_i}(z-\alpha_i) u_{m_i}(z-\alpha_i)^\top \frac{1}{\pi(z)} dz. \tag{24}$$

As we have previously noticed, $M(i)^{-1}$ is completely determined by its first row (or column). From Proposition 3.6 we get, using Cauchy's theorem, that this first row is given by

$$\frac{1}{2\pi i} \oint_{\Gamma_{\alpha_i}} u_{m_i}(z-\alpha_i)^\top \frac{1}{\pi(z)} dz = \left(\frac{1}{\pi_{m_i}(\alpha_i;\alpha_i)}, \dots, \frac{1}{(m_i-1)!} D^{m_i-1} \frac{1}{\pi_{m_i}(\alpha_i;\alpha_i)} \right) P,$$

in agreement with what we already found in (12).

4. Application to Stein equations. The goal of this section is to obtain a way to compute the solution of Stein's equation, where the coefficients are matrices in companion form. Apart from its interest this is chiefly motivated by the computation of Fisher's information matrix of an ARMA process. As we stated in the introduction, the blocks of Fisher's information matrix are solutions to such a Stein equation; see [14]. We postpone the application to ARMA processes until section 5.

Let A be a complex matrix of size $n \times n$ (not necessarily in companion form). If f is a $\mathbb{C}^{n\times l}$ valued analytic function, then we define $f(A)$ as $\sum_{k=0}^{\infty} \frac{1}{k!} A^k f^{(k)}(0)$. We use the following known result (see, for instance, [17, section 9.9, Theorem 2]).

LEMMA 4.1. *Let A be a complex matrix ($n \times n$) whose eigenvalues lie strictly inside the unit disk Γ. Then for a $\mathbb{C}^{n\times l}$ valued analytic function f one has*

$$\frac{1}{2\pi i} \oint_{\Gamma} (z - A)^{-1} f(z) dz = f(A).$$

As an application of Lemma 4.1 we solve the Stein equation. Given matrices A, C, and H of appropriate dimensions (we also assume that the eigenvalues of both A and C lie inside the unit disk), we are looking for the solution for S of

$$S - ASC^{\top} = H. \tag{25}$$

This equation is of interest in matrix and operator theory (e.g., the operator that takes S to $S - ASC$ is called a displacement operator; see [8]). In [15] we study this equation further and relate solutions of various Stein equations to a certain Fisher information matrix.

The solution to (25) (see [16]) is given by $\frac{1}{2\pi i} \oint_{\Gamma} (z - A)^{-1} f(z) dz$, with $f(z) = H(I - zC)^{-\top}$, and hence is equal to $\sum_{k=0}^{\infty} A^k H (C^{\top})^k$.

We continue with presenting an alternative way to obtain a solution for the special case where both the matrices A and C are in companion form. Let V_A be the Vandermonde matrix associated with A and let V_C be associated with C. Let $\hat{S} = V_A^{-1} S V_C^{-\top}$ and $\hat{H} = V_A^{-1} H V_C^{-\top}$. The results of section 3 on inverses of confluent Vandermonde matrices enable us to compute $\hat{H}$.

Premultiplication of (25) with V_A^{-1}, together with postmultiplication with $V_C^{-\top}$, results in

$$\hat{S} - J_A \hat{S} J_C^{\top} = \hat{H}, \tag{26}$$

where J_A and J_C are the Jordan forms of A and B, respectively.

Let $v = \mathrm{vec}\,(\hat{S})$ and $h = \mathrm{vec}\,(\hat{H})$. Then it is known (see [16]) that v is given by $v = (I - J_C \otimes J_A)^{-1} h$ under the assumption that no product of an eigenvalue of A and an eigenvalue of C equals 1. This assumption is typically fulfilled in the context of stationary and invertible ARMA processes, where these eigenvalues are the zeros of both AR- and MA-polynomials and thus lie inside the unit circle; see section 5.

The computation of the inverse of the matrix $I - J_C \otimes J_A$ can now be done in an efficient way. Let $J_{A,i}$ be the Jordan block of J_A associated with the eigenvalue α_i and let $J_{C,j}$ be the Jordan block of J_C associated with the eigenvalue γ_j. Then $I - J_C \otimes J_A$ is block diagonal with diagonal blocks $I - J_{C,j} \otimes J_{A,i}$. Moreover, these blocks are upper triangular and even almost block diagonal. On the diagonal we find the blocks $I - \gamma_j J_{A,i}$ and on the subdiagonal just above it find the blocks $-J_{A,i}$. Therefore, $(I - J_{C,j} \otimes J_{A,i})^{-1}$ is again upper triangular with, on the diagonal, the blocks $(I - \gamma_j J_{A,i})^{-1}$ and, on the kth subdiagonal above it ($k \leq m_j - 1$ with m_j the

multiplicity of γ_j), one finds the blocks $(I - \gamma_j J_{A,i})^{-k-1} J_{A,i}^k$. Finally, the inverses of the $I - \gamma_j J_{A,i}$ are upper triangular Toeplitz matrices with kl-element given by $\gamma_j^{k-l}(1 - \alpha_i\gamma_j)^{-k+l-1}$ for $k \geq l$.

The generic case is that in which all the eigenvalues of A and all the eigenvalues of C have multiplicity 1. Consequently the matrices J_A and J_C are diagonal. In this case (26) has a very simple solution: $\hat{S}$ has elements $\hat{S}_{ij} = \frac{1}{1-\alpha_i\gamma_j}\hat{H}_{ij}$.

5. Application to ARMA processes. Consider an ARMA(p, q) process y, a stationary discrete time stochastic process that satisfies

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} = \varepsilon_t + c_1\varepsilon_{t-1} + \cdots + c_q\varepsilon_{t-q}, \tag{27}$$

where ε is a Gaussian white noise sequence with unit variance. The real constants $a_1, \ldots, a_p$ and $c_1, \ldots, c_q$ will be fixed throughout the rest of this section.

Introduce the monic polynomials $a(z) = \sum_{i=0}^p a_{p-i} z^i$ and $c(z) = \sum_{i=0}^q c_{q-i} z^i$ and let a^* and c^* be the corresponding reciprocal polynomials so that $a^*(z) = \sum_{i=0}^n a_i z^i$ and $c^*(z) = \sum_{i=0}^q c_i z^i$. We make the common assumption that the ARMA process is causal and invertible, meaning that a and c have their zeros strictly inside the unit circle [2, Chapter 3].

Write $\theta = (a_1, \ldots, a_p, c_1, \ldots, c_q)^\top$. Notice that the observations y (given random variables or their realized values) of course don't depend on the parameter θ, but then the noise sequence ε does. The Fisher information matrix $F_n(\theta)$ for n observations is defined (see [1]) as the covariance matrix of the score function and, because of the assumed Gaussian distribution of ε, it is asymptotically equal to n times the stationary Fisher information matrix

$$F(\theta) = \mathbb{E}_\theta \frac{\partial\varepsilon}{\partial\theta} \frac{\partial\varepsilon}{\partial\theta}^\top,$$

where $\mathbb{E}_\theta$ denotes expectation under the parameter θ. Knowledge of the Fisher information matrix is crucial for asymptotic statistical analysis. For instance, it is known (see, e.g., [2]) that maximum likelihood estimators of the parameters of an ARMA process are consistent and have (using n observations) an asymptotic covariance matrix that is n^{-1} times the inverse (provided that it exists) of the stationary Fisher information matrix. The inverse exists if the polynomials a and c have no common zeros; see [13].

The matrix $F(\theta)$ has a representation in the spectral domain given by the block decomposition

$$F(\theta) = \begin{pmatrix} F_{aa} & F_{ac} \\ F_{ac}^\top & F_{cc} \end{pmatrix}, \tag{28}$$

where the matrices appearing here have the elements

$$F_{aa}^{jk} = \frac{1}{2\pi i} \oint_{|z|=1} \frac{z^{j-k+p-1}}{a(z)a^*(z)} dz, (j, k = 1, \ldots, p),$$
$$F_{ac}^{jk} = \frac{1}{2\pi i} \oint_{|z|=1} \frac{z^{j-k+q-1}}{c(z)a^*(z)} dz, (j = 1, \ldots, p, k = 1, \ldots, q),$$
$$F_{cc}^{jk} = \frac{1}{2\pi i} \oint_{|z|=1} \frac{z^{j-k+q-1}}{c(z)c^*(z)} dz, (j, k = 1, \ldots, q).$$

With $k(z) = a(z)a^*(z)c(z)c^*(z)$, $u_p(z) = (1, \dots, z^{p-1})^\top$, $u_q(z)$ likewise, and u_p^* and u_q^* their reciprocal polynomials, we have the following compact expression for the whole Fisher information matrix:

$$F(\theta) = \frac{1}{2\pi i} \oint_{|z|=1} \frac{1}{k(z)} \begin{pmatrix} c^*(z)u_p(z) \\ -a^*(z)u_q(z) \end{pmatrix} \begin{pmatrix} c(z)u_p^*(z)^\top & -a(z)u_q^*(z)^\top \end{pmatrix} dz. \tag{29}$$

As in section 2 we let $A \in \mathbb{R}^{p\times p}$ be the companion matrix associated with the polynomial $a(\cdot)$ (its precise form is given by (1) for $n = p$). The matrix $C \in \mathbb{R}^{q\times q}$ associated with the polynomial $c(\cdot)$ has an analogous form.

Let the matrix $\tilde{A} \in \mathbb{R}^{(p+q)\times(p+q)}$ be given by

$$\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix}.$$

In [14] we showed that the Fisher information matrix $F(\theta)$ is the solution of the Stein equation

$$F(\theta) - \tilde{A}F(\theta)\tilde{A}^\top = ee^\top, \tag{30}$$

where $e^\top = (e_{pp}^\top, e_{qq}^\top)$ with e_{pp} the pth standard basis vector in $\mathbb{R}^p$ and e_{qq} the qth standard basis vector in $\mathbb{R}^q$. Using for $F(\theta)$ the block decomposition (28), we see that each of the blocks involved satisfies a Stein equation with appropriate coefficients. For instance, for $F_{ac} \in \mathbb{R}^{p\times q}$ we have

$$F_{ac} - AF_{ac}C^\top = H_{ac}, \tag{31}$$

with $H_{ac} = e_{pp}e_{qq}^\top$. As we already announced in the introduction, (31) as well as the analogous equation for the other blocks of Fisher's information matrix motivated the study of solutions to Stein's equation, in which the coefficient matrices are in companion form.

We apply the results of the previous sections as follows. Let V_A be a matrix whose columns are the generalized eigenvectors of A, and let V_C be the corresponding matrix for C. As we have seen, these matrices are confluent Vandermonde matrices. By J_A and J_C we denote the Jordan forms of A and C, respectively. Let also $\hat{F}_{ac} = V_A^{-1}F_{ac}V_C^{-\top}$ and $\hat{H}_{ac} = V_A^{-1}H_{ac}V_C^{-\top}$. Then we can replace (31) with the equivalent equation

$$\hat{F}_{ac} - J_A\hat{F}_{ac}J_C^\top = \hat{H}_{ac}. \tag{32}$$

A little more can be said. The matrix $\hat{H}_{ac}$ here becomes $V_A^{-1}e_{pp}(V_C^{-1}e_{qq})^\top$ and we observe that both $V_A^{-1}e_{pp}$ and $V_C^{-1}e_{qq}$ are the last columns of the inverse of a Vandermonde matrix. We have already seen in section 2 how these columns are related to interpolation polynomials. We have, for instance, that $V_A^{-1}e_{pp}$ is equal to $p_A^*(0)$, where $p_A^*(z) = z^{p-1}p_A(\frac{1}{z})$ and p_A is the interpolation polynomial related to the eigenvalues of A as described in Proposition 2.2. Likewise $V_C^{-1}e_{qq} = p_C^*(0)$.

Let us finish by considering the generic case of Fisher's information matrix; i.e., we assume that A and C only have eigenvalues of multiplicity 1. It then follows that $\hat{F}_{ac}$ has as its ijth element

$$\frac{p_A^*(0)_i p_C^*(0)_j}{1 - \alpha_i\gamma_j}.$$

Now it is easy to compute $F_{ac} = V_A\hat{F}_{ac}V_C^\top$. To the other blocks of the Fisher information matrix the same procedure applies.

Acknowledgment. We thank a referee for many detailed suggestions that helped us in writing the revision.

REFERENCES

[1] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis*, 3rd ed., Prentice Hall Inc., Englewood Cliffs, NJ, 1994.

[2] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., Springer-Verlag, Berlin, New York, 1991.

[3] F. C. Chang, *The inverse of the generalized Vandermonde matrix through the partial fraction expansion*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 151–152.

[4] F. G. Csáki, *Some notes on the inversion of confluent Vandermonde matrices*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 154–157.

[5] P. J. Davis, *Interpolation and Approximation*, Dover Publications Inc., New York, 1975.

[6] B. Friedlander, *On the computation of the Cramér-Rao bound for ARMA parameter estimation*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 721–727.

[7] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea Publishing, New York, 1959.

[8] I. Gohberg and V. Olshevsky, *The fast generalized Parker-Traub algorithm for inversion of Vandermonde and related matrices*, J. Complexity, 13 (1997), pp. 208–234.

[9] I. C. Göknar, *Obtaining the inverse of the generalized Vandermonde matrix of the most general type*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 530–532.

[10] I. A. Ibragimov and R. Z. Has'minskiĭ, *Statistical Estimation. Asymptotic Theory*, Springer-Verlag, New York, 1981.

[11] I. Kaufman, *The inversion of the Vandermonde matrix and the transformation to the Jordan canonical form*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 774–777.

[12] A. Klein and G. Mélard, *Computation of the Fisher information matrix for SISO models*, IEEE Trans. Signal Process., 42 (1994), pp. 684–688.

[13] A. Klein and P. J. C. Spreij, *On Fisher's information matrix of an ARMAX process and Sylvester's resultant matrices*, Linear Algebra Appl., 237/238 (1996), pp. 579–590.

[14] A. Klein and P. J. C. Spreij, *On Fisher's information matrix of an ARMA process*, in Stochastic Differential and Difference Equations, Birkhäuser Boston, Boston, MA, 1997, pp. 273–284.

[15] A. Klein and P. J. C. Spreij, *On Stein's equation, Vandermonde matrices and Fisher's information matrix of time series processes.* I. *The autoregressive moving average process*, Linear Algebra Appl., 329 (2001), pp. 9–47.

[16] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1995.

[17] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.

[18] L. Lupaş, *On the computation of the generalized Vandermonde matrix inverse*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 559–561.

[19] B. Porat and B. Friedlander, *Computation of the exact information matrix of Gaussian time series with stationary random components*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 118–130.

[20] R. H. Schappelle, *The inverse of the confluent Vandermonde matrix*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 724–725.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 224–236

SOME OBSERVATIONS ON GENERALIZED SADDLE-POINT PROBLEMS*

P. CIARLET, JR.†, JIANGUO HUANG‡, AND JUN ZOU§

Abstract. This paper studies the solvability and stability of a generalized saddle-point system in finite- and infinite-dimensional spaces. Sharp solvability conditions and stability estimates are derived.

Key words. generalized saddle-point problems, stability, solvability

AMS subject classifications. 35A15, 35A05, 65N12, 65L20

PII. S0895479802410827

1. Introduction.

We shall consider the solvability and stability of the following saddle-point system: Find $(u, p) \in V \times Q$ such that

$$a(u, v) + b_1(v, p) = f(v) \quad \forall v \in V, \tag{1.1}$$

$$b_2(u, q) - c(p, q) = g(q) \quad \forall q \in Q, \tag{1.2}$$

where a, b_1, b_2, and c are bounded bilinear forms and where f and g are bounded linear functionals on V and Q, respectively. The system (1.1)–(1.2) seems to be one of the most generalized saddle-point systems investigated in the literature. The case of bilinear forms $c = 0$ and $b_1 = b_2$ has been extensively studied [1, 3, 4, 7, 5, 9, 10]. Also, considerable research has been done on the system with $b_1 = b_2$ and $c \neq 0$ [4, 11, 14], while the well-posedness for the system with $c = 0$ but $b_1 \neq b_2$ was established in [13] and [2]. However, to our knowledge there have been no investigations into the solvability and stability for the most general form of system (1.1)–(1.2) with $b_1 \neq b_2$ and $c \neq 0$.

The aim of this paper is to establish the solvability and stability conditions for the generalized saddle-point system (1.1)–(1.2). The existence and uniqueness of the solutions to the system are shown under some standard conditions, and stability estimates of the solutions are derived in terms of the given data.

The system (1.1)–(1.2) arises in, for example, mixed variational formulations of some boundary value problems. The first of such examples is the following general non–self-adjoint elliptic problem:

$$-\nabla \cdot (\alpha(x)\nabla p + \mathbf{b}(x)\, p) + \gamma(x)\, p = \mu(x), \quad x \in \Omega, \tag{1.3}$$

where Ω is a bounded domain in R^d $(d = 2, 3)$ with boundary $\partial\Omega$, the solution p is assumed to take the boundary value $\omega(x)$ on $\partial\Omega$, and $\alpha(x)$, $\mathbf{b}(x)$, $\gamma(x)$, and $\mu(x)$

*Received by the editors July 10, 2002; accepted for publication (in revised form) by A. Wathen December 30, 2002; published electronically May 29, 2003.
http://www.siam.org/journals/simax/25-1/41082.html

†Ecole Nationale Supérieure de Techniques Avancées, 32, Boulevard Victor, 75739 Paris Cedex 15, France (ciarlet@ensta.fr).

‡Department of Mathematics, Shanghai Jiao Tong University, Shanghai, 200240, People's Republic of China (jghuang@online.sh.cn). The work of this author was partially supported by the National Natural Science Foundation of China under grant 19901018 and by Direct Grant of CUHK (2060226), Hong Kong.

§Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (zou@math.cuhk.edu.hk). The work of this author was supported by Hong Kong RGC grants CUHK4292/00P and CUHK4048/02P.

are given functions with appropriate smoothness [6]. By introducing the new variable $\mathbf{u} = -(\alpha \nabla p + \mathbf{b}\, p)$, and letting $\tilde{\alpha}(x) = \alpha(x)^{-1}$ and $\tilde{\mathbf{b}}(x) = \tilde{\alpha}(x)\mathbf{b}(x)$, we have that the weak form of (1.3) is then described by system (1.1)–(1.2) (see [6]), with two spaces $V = \{\mathbf{u} \in L^2(\Omega)^d; \mathrm{div}\mathbf{v} \in L^2(\Omega)\}$ and $Q = L^2(\Omega)$, and two linear functionals $f(\mathbf{v}) = -\langle \omega, \mathbf{v} \cdot \mathbf{n}\rangle$ and $g(q) = -(\mu, q)$, while the bilinear forms are given by

$$a(\mathbf{u}, \mathbf{v}) = (\tilde{\alpha}\, \mathbf{u}, \mathbf{v})\,, \quad b_1(\mathbf{v}, p) = -(\mathrm{div}\, \mathbf{v}, p) + (\tilde{\mathbf{b}}\, p, \mathbf{v})\,,$$
$$c(p, q) = (\gamma\, p, q)\,, \qquad b_2(\mathbf{u}, q) = -(\mathrm{div}\, \mathbf{u}, q)\,,$$

where $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ denote the scalar products in $L^2(\Omega)$ (or $L^2(\Omega)^d$) and $L^2(\partial\Omega)$, respectively.

A second example comes from some exterior electromagnetic interface problems [12, 13]. The weak formulations of such problems also take the form (1.1)–(1.2) if one introduces a Lagrange multiplier variable $\mathbf{u}$ for the current density $\nabla\phi$, where ϕ is the potential function [12, 13], and introduces another Lagrange multiplier variable ξ for the boundary value of ϕ on the boundary of the physical domain Ω.

2. Preliminaries. In this section, we introduce some existing saddle-point theory. Let V and Q be two finite- or infinite-dimensional Hilbert spaces equipped with the inner products $(\cdot, \cdot)_V$ and $(\cdot, \cdot)_Q$, and the induced norms $\|\cdot\|_V$ and $\|\cdot\|_Q$, respectively. Let $a(v, w)$, $b_1(v, q)$, and $b_2(v, q)$ be bilinear forms on $V \times V$, $V \times Q$, and $V \times Q$, respectively, which are bounded; i.e., there are positive constants $\|a\|$, $\|b_1\|$, and $\|b_2\|$ such that

$$|a(v, w)| \le \|a\|\|v\|_V\|w\|_V \quad \forall v,\ w \in V, \tag{2.1}$$
$$|b_1(v, q)| \le \|b_1\|\|v\|_V\|q\|_Q \quad \forall v \in V,\ q \in Q, \tag{2.2}$$
$$|b_2(v, q)| \le \|b_2\|\|v\|_V\|q\|_Q \quad \forall v \in V,\ q \in Q. \tag{2.3}$$

Associated with the three bilinear forms are three linear operators $A \in \mathcal{L}(V, V)$, B_1, $B_2 \in \mathcal{L}(V, Q)$ defined by

$$a(v, w) = (Av, w)_V \quad \forall\, v \in V,\ w \in V,$$
$$b_1(v, q) = (B_1 v, q)_Q = (v, B_1^t q)_V \quad \forall\, v \in V,\ q \in Q,$$
$$b_2(v, q) = (B_2 v, q)_Q = (v, B_2^t q)_V \quad \forall\, v \in V,\ q \in Q.$$

Clearly, the three constants in (2.1)–(2.3) can be taken as

$$\|a\| = \|A\|_{\mathcal{L}(V,V)}, \quad \|b_1\| = \|B_1\|_{\mathcal{L}(V,Q)}, \quad \|b_2\| = \|B_2\|_{\mathcal{L}(V,Q)}.$$

We first consider the saddle-point problem

$$\begin{aligned} a(u, v) + b_1(v, p) &= (f, v)_V \quad \forall\, v \in V, \\ b_2(u, q) &= (g, q)_Q \quad \forall\, q \in Q. \end{aligned} \tag{2.4}$$

This system is equivalent to the following operator equation (or matrix equation in finite dimensions):

$$Au + B_1^t p = f, \quad B_2 u = g. \tag{2.5}$$

Let $U_i = \mathrm{Ker}(B_i)$, $i = 1, 2$. Then we have the following results on system (2.4) (cf. [13, 2]).

THEOREM 2.1. *In addition to assumptions* (2.1)–(2.3), *we assume that*

$$\sup_{w\in U_1}\frac{a(v,w)}{\|w\|_V}\geq\alpha\|v\|_V\quad\forall\, v\in U_2, \tag{2.6}$$

$$\sup_{v\in U_2}a(v,w)>0\quad\forall\, w\in U_1,\ w\neq 0, \tag{2.7}$$

$$\sup_{v\in V}\frac{b_i(v,q)}{\|v\|_V}\geq\beta_i\,\|q\|_Q\quad\forall\, q\in Q\quad(i=1,2) \tag{2.8}$$

hold for some constants α, β_1, $\beta_2>0$. *Then for any* $f\in V$ *and* $g\in Q$, *there exists a unique solution* $(u,p)\in V\times Q$ *to system* (2.4), *and the following stability estimates hold:*

$$\|u\|_V\leq\beta_2^{-1}(1+\alpha^{-1}\|a\|)\|g\|_Q+\alpha^{-1}\|f\|_V, \tag{2.9}$$

$$\|p\|_Q\leq\beta_1^{-1}(\|f\|_V+\|a\|\|u\|_V). \tag{2.10}$$

Theorem 2.1 generalizes the standard saddle-point theory ($b_1=b_2$) [1, 3]. Equation (2.8) is the so-called *inf-sup* condition, which plays an important role in the entire saddle-point theory. We refer to [4, 7] and the references therein for more details.

To apply the saddle-point theory for the compressible Stokes equations, Kellogg and Liu [11] introduced another abstract framework; see also [4]. Let $c(p,q)$ be a bounded and weakly coercive bilinear form on $Q\times Q$; i.e., there exist a positive constant $\|c\|$ and a constant γ (possibly negative[1]) such that

$$|c(p,q)|\leq\|c\|\|p\|_Q\|q\|_Q\quad\forall p,\ q\in Q, \tag{2.11}$$

$$c(q,q)\geq-\gamma\|q\|_Q^2\quad\forall q\in Q\,. \tag{2.12}$$

Further, define the operator $C\in\mathcal{L}(Q,Q)$ by

$$c(p,q)=(Cp,q)_Q\quad\forall p,\ q\in Q\,. \tag{2.13}$$

Let $b(v,q)$ be a bilinear form on $V\times Q$ satisfying

$$\sup_{v\in V}\frac{b(v,q)}{\|v\|_V}\geq\beta\,\|q\|_Q\quad\forall\, q\in Q\,, \tag{2.14}$$

$$|b(v,q)|\leq\|b\|\|v\|_V\|q\|_Q\quad\forall v\in V,\ q\in Q \tag{2.15}$$

for some positive constants β and $\|b\|$. Then for the saddle-point problem

$$\begin{aligned}a(u,v)+b(v,p)&=(f,v)_V\quad\forall\, v\in V,\\ b(u,q)-c(p,q)&=(g,q)_Q\quad\forall\, q\in Q,\end{aligned} \tag{2.16}$$

which is equivalent to the operator equation (or matrix equation)

$$Au+B^tp=f,\quad Bu-Cp=g\,, \tag{2.17}$$

we have (cf. [11, 4]) the following.

[1]It is clear from (2.11) that the weak coerciveness (2.12) is always satisfied for any $\gamma\geq\|c\|$, but we are interested only in the case with $\gamma<\|c\|$.

THEOREM 2.2. *Assume that for some constant* $\alpha > 0$,

$$a(v,v) \geq \alpha \|v\|_V^2 \quad \forall\, v \in V, \tag{2.18}$$

and conditions (2.1), (2.11)–(2.15) *are satisfied. Then for any* $f \in V$ *and* $g \in Q$, *there exists a unique solution* $(u,p) \in V \times Q$ *to system* (2.16) *if* $\gamma < \alpha \|a\|^{-2}\beta^2$, *and the following stability estimates hold:*

$$\|p\|_Q \leq \frac{\alpha^{-1}\|b\|\|f\|_V + \|g\|_Q}{\alpha\|a\|^{-2}\beta^2 - \gamma}, \quad \|u\|_V \leq \alpha^{-1}(\|f\|_V + \|b\|\|p\|_Q). \tag{2.19}$$

Finite-dimensional case. Let us briefly discuss the equivalent forms of the inf-sup condition and other conditions used in Theorems 2.1 and 2.2 when V and Q are finite dimensional. Without loss of generality, we consider $V = R^n$ and $Q = R^m$ $(n \geq m)$, and both spaces are equipped with the standard Euclidean norms $\|\cdot\|_2$ and inner products $(\cdot,\cdot)$, with no distinction between the notation of the norms and inner products of R^n and R^m.

First, we claim that the inf-sup conditions (2.8) are equivalent to the conditions $\operatorname{rank}(B_1) = \operatorname{rank}(B_2) = m$. To see this, we write

$$\sup_{v \in V} \frac{b_1(v,q)}{\|v\|_V} = \sup_{v \in R^n} \frac{(v, B_1^t q)}{\|v\|_2} = \|B_1^t q\|_2,$$

so (2.8) with $i = 1$ is the same as the condition

$$\|B_1^t q\|_2 \geq \beta_1 \|q\|_2 \quad \forall\, q \in R^m,$$

or $\operatorname{rank}(B_1) = m$. Similar derivations lead to the fact that (2.8) with $i = 2$ is equivalent to the condition $\operatorname{rank}(B_2) = m$.

Second, one can directly check that conditions (2.18) and (2.12) amount to

$$\lambda_{\min}\left(\frac{A + A^t}{2}\right) \geq \alpha, \quad \lambda_{\min}\left(\frac{C + C^t}{2}\right) \geq -\gamma,$$

respectively (cf. [8]).

Finally, we analyze conditions (2.6)–(2.7). Let $\operatorname{rank}(B_i) = m_i \leq m$, $i = 1, 2$; then we know that $\dim(U_i) = \dim(\operatorname{Ker}(B_i)) = n - m_i$. Let N_i be the $n \times (n - m_i)$ matrix formed by an orthonormal basis of $\operatorname{Ker}(B_i)$. To rewrite condition (2.6), for any $w \in U_1$ and $v \in U_2$, let $w = N_1 x$ and $v = N_2 y$ with $x \in R^{n-m_1}$ and $y \in R^{n-m_2}$; then

$$\sup_{w \in U_1} \frac{a(v,w)}{\|w\|_V} = \sup_{x \in \mathbf{R}^{n-m_1}} \frac{(AN_2 y, N_1 x)}{\|N_1 x\|_2} = \sup_{x \in \mathbf{R}^{n-m_1}} \frac{(N_1^T A N_2 y, x)}{\|x\|_2} = \|N_1^T A N_2 y\|_2\,,$$

so (2.6) is equivalent to the condition

$$\|N_1^T A N_2 y\|_2 \geq \alpha \|y\|_2 \quad \forall\, y \in R^{n-m_2},$$

or $\operatorname{rank}(N_1^T A N_2) = n - m_2$. Similarly, we can rewrite condition (2.7) as

$$\sup_{v \in U_2} a(v,w) = \sup_{y \in R^{n-m_2}} (AN_2 y, N_1 x) = \sup_{y \in R^{n-m_2}} (y, N_2^T A^T N_1 x).$$

This indicates that (2.7) is equivalent to the condition

$$N_2^T A^T N_1 x \neq 0 \quad \forall\, x \neq 0,$$

or $\operatorname{rank}(N_2^T A^T N_1) = n - m_1$.

One can further conclude from the above that $\operatorname{rank}(B_1) = m_1 = m_2 = \operatorname{rank}(B_2)$ if both conditions (2.6) and (2.7) are satisfied.

3. Main results. This paper is concerned with the following generalized saddle-point problem: Find $(u,p) \in V \times Q$ such that

$$\begin{aligned} &a(u,v) + b_1(v,p) = (f,v)_V \quad \forall v \in V, \\ &b_2(u,q) - c(p,q) = (g,q)_Q \quad \forall q \in Q. \end{aligned} \tag{3.1}$$

The system can be written in the operator or matrix form

$$Au + B_1^t\, p = f, \quad B_2 u - Cp = g. \tag{3.2}$$

Obviously, problem (3.1) covers systems (2.4) and (2.16) as two special cases.

In this section, we present two results on the solvability and stability for system (3.1) under two sets of different conditions: the first result requires that only one of the bilinear forms $b_1(v,q)$ and $b_2(v,q)$ satisfy the inf-sup condition; the second does not assume the weak coerciveness (2.12) for the bilinear form $c(p,q)$ with $\gamma < \alpha\|a\|^{-2}\beta^2$.

3.1. Well-posedness with either $b_1(v,q)$ or $b_2(v,q)$ satisfying the inf-sup condition. The main results of this section are summarized in the following theorem.

THEOREM 3.1. *The same assumptions as in Theorem* 2.2 *are made but with* $b(v,q)$ *replaced by* $b_1(v,q)$ *here. Then for any* $f \in V$ *and* $g \in Q$, *there exists a unique solution* $(u,p) \in V \times Q$ *to the saddle-point problem* (3.1) (*or* (3.2)) *as long as*

$$\delta_1 = \frac{\alpha^{-1}\|b_1\|\|b_1 - b_2\|}{\alpha\|a\|^{-2}\beta_1^2 - \gamma} < 1\,, \tag{3.3}$$

where $\|b_1 - b_2\| = \|B_1 - B_2\|_{\mathcal{L}(V,Q)}$. *Further, the solution admits the stability estimates*

$$\|u\|_V \le \frac{1}{1-\delta_1}\|\tilde{u}\|_V, \quad \|p\|_Q \le \|\tilde{p}\|_Q + \frac{\|b_1 - b_2\|}{(\alpha\|a\|^{-2}\beta_1^2 - \gamma)(1-\delta_1)}\|\tilde{u}\|_V, \tag{3.4}$$

where $(\tilde{u}, \tilde{p})$ *solves* (2.16) *with* b *replaced by* b_1, *and thus has the bounds*

$$\|\tilde{p}\|_Q \le \frac{\alpha^{-1}\|b_1\|\|f\|_V + \|g\|_Q}{\alpha\|a\|^{-2}\beta_1^2 - \gamma}, \quad \|\tilde{u}\|_V \le \alpha^{-1}(\|f\|_V + \|b_1\|\|\tilde{p}\|_Q).$$

Proof. We choose $u^0 = 0 \in Q$ and determine a sequence $\{(u^n, p^n)\}$ by

$$Au^{n+1} + B_1^t p^{n+1} = f, \tag{3.5}$$

$$B_1 u^{n+1} - Cp^{n+1} = g + (B_1 - B_2)u^n, \tag{3.6}$$

for $n = 0, 1, 2, \ldots$. The sequence $\{(u^n, p^n)\}$ is well defined by Theorem 2.2. Subtracting (3.5)–(3.6) from (3.5)–(3.6) with n replaced by $n-1$, it follows that

$$\begin{aligned} &A(u^{n+1} - u^n) + B_1^t(p^{n+1} - p^n) = 0, \\ &B_1(u^{n+1} - u^n) - C(p^{n+1} - p^n) = (B_1 - B_2)(u^n - u^{n-1}). \end{aligned} \tag{3.7}$$

Now applying estimates (2.19) to (3.7), we have

$$\|u^{n+1} - u^n\|_V \le \alpha^{-1}\|b_1\|\|p^{n+1} - p^n\|_Q, \tag{3.8}$$

$$\|p^{n+1} - p^n\|_Q \le \frac{\|b_1 - b_2\|}{\alpha\|a\|^{-2}\beta_1^2 - \gamma}\|u^n - u^{n-1}\|_V, \tag{3.9}$$

which implies for $n \geq 1$ that

$$\|u^{n+1} - u^n\|_V \leq \delta_1 \|u^n - u^{n-1}\|_V \leq \delta_1^n \|u^1\|_V; \tag{3.10}$$

that is, for any nonnegative integers $m > n$,

$$\begin{aligned} \|u^m - u^n\|_V &\leq \sum_{i=n}^{m-1} \|u^{i+1} - u^i\|_V \leq \left(\sum_{i=n}^{m-1} \delta_1^i \right) \|u^1\|_V \\ &\leq \frac{\delta_1^n}{1-\delta_1} \|u^1\|_V. \end{aligned} \tag{3.11}$$

This means $\{u^n\}$ is a Cauchy sequence, and there exists a $u \in V$ such that

$$u^n \to u \quad \text{in } V. \tag{3.12}$$

On the other hand, it follows from (3.9) and (3.10) that

$$\|p^{n+1} - p^n\|_Q \leq \frac{\|b_1 - b_2\|}{\alpha \|a\|^{-2} \beta_1^2 - \gamma} \delta_1^{n-1} \|u^1\|_V,$$

which implies that

$$\|p^m - p^n\|_Q \leq \frac{\|b_1 - b_2\|}{\alpha \|a\|^{-2} \beta_1^2 - \gamma} \frac{\delta_1^{n-1}}{1-\delta_1} \|u^1\|_V. \tag{3.13}$$

Hence $\{p^n\}$ also is a Cauchy sequence, and there exists a $p \in Q$ such that

$$p^n \to p \quad \text{in } Q. \tag{3.14}$$

Letting n tend to infinity in (3.5)–(3.6), we know that $(u, p) \in V \times Q$ solves (3.2).

We next verify the uniqueness of problem (3.2). Assume that there are two solutions (u_1, p_1), $(u_2, p_2) \in V \times Q$ to the system. It is easy to see that the difference between the two solutions satisfies

$$\begin{aligned} &A(u_1 - u_2) + B_1^t (p_1 - p_2) = 0, \\ &B_1(u_1 - u_2) - C(p_1 - p_2) = (B_1 - B_2)(u_1 - u_2). \end{aligned} \tag{3.15}$$

Using the same technique for deriving estimate (3.10), we have

$$\|u_1 - u_2\|_V \leq \delta_1 \|u_1 - u_2\|_V,$$

which shows $u_1 = u_2$ since $\delta_1 < 1$. Equality $p_1 = p_2$ follows immediately by applying estimate (2.19) to (3.15).

Finally, we derive the stability estimates. As $u^0 = 0$, we see that (u^1, p^1) solves (2.16) with b replaced by b_1; thus (u^1, p^1) satisfies estimates (2.19). Taking $n = 1$ in (3.11) and letting m go to infinity, we obtain the first estimate in (3.4). Similarly, taking $n = 1$ in (3.13) leads to the second estimate in (3.4). ∎

Sharpness of the condition on δ_1 in (3.3). Below, we give a simple example to show that condition $\delta_1 < 1$ is a sharp condition guaranteeing the unique solvability of system (3.2). For this, consider $V = R^n$, $Q = R^m$, where $n \geq m$. We choose $A = I_n$, $C = I_m$, and $B_2 = -B_1$ with $B_1 \in R^{m \times n}$ such that $\operatorname{rank}(B_1) = m$. It is easy

to see that $\delta_1 = 2\sigma_{\max}^2(B_1)/(1+\sigma_{\min}^2(B_1))$, where $\sigma_{\min}(B_1)$ is the minimal singular value of B_1. Then $\delta_1 < 1$ means that

$$2\sigma_{\max}^2(B_1) < 1 + \sigma_{\min}^2(B_1), \tag{3.16}$$

which implies $\sigma_{\max}(B_1) < 1$. It is also easy to show that problem (3.2) is uniquely solvable if and only if the matrix $(I_m - B_1 B_1^t)$ is nonsingular. Hence, $\sigma_{\max}(B_1) < 1$ is indeed a sufficient condition for the unique solvability of (3.2).

On the other hand, for any $\delta_1 \geq 1$, choose the $m \times n$ matrices B_1 and B_2 as follows:

$$-B_2 = B_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \delta_1 I_{m-1} & 0 \end{pmatrix}.$$

Then the matrix $(I_m - B_1 B_1^t)$ is singular, and so (3.2) is not uniquely solvable.

Remark 3.1. In most applications, the constants γ in (2.12) are negative. Then the condition $\gamma \leq \alpha \|a\|^{-2} \beta_1^2$ required in Theorems 2.2 and 3.1 is automatically satisfied.

Remark 3.2. In Theorem 3.1, only $b_1(v,q)$, not $b_2(v,q)$, is required to satisfy the inf-sup condition. Similar results hold when $b_2(v,q)$ satisfies the inf-sup condition but $b_1(v,q)$ does not.

3.2. Well-posedness not assuming condition (2.12) for any $\gamma < \|c\|$. The main results of this section are summarized in the following theorem.

THEOREM 3.2. *If we make the same assumptions as in Theorem* 2.1*, then for any $f \in V$ and $g \in Q$, there exists a unique solution $(u,p) \in V \times Q$ to the saddle-point problem* (3.1) *as long as*

$$\delta_2 := \beta_1^{-1} \beta_2^{-1} \|a\| (1 + \alpha^{-1} \|a\|) \|c\| < 1. \tag{3.17}$$

Further, the following stability estimates hold:

$$\|p\|_Q \leq \frac{1}{1-\delta_2} \|\tilde{p}\|_Q, \quad \|u\|_V \leq \|\tilde{u}\|_V + \frac{\beta_2 (1 + \alpha^{-1}\|a\|)\|c\|}{1-\delta_2} \|\tilde{p}\|_Q, \tag{3.18}$$

where $(\tilde{u}, \tilde{p})$ is the solution to (2.4) *and thus has the bounds*

$$\|\tilde{u}\|_V \leq \beta_2^{-1}(1 + \alpha^{-1}\|a\|)\|g\|_Q + \alpha^{-1}\|f\|_V, \quad \|\tilde{p}\|_Q \leq \beta_1^{-1}(\|f\|_V + \|a\|\|\tilde{u}\|_V).$$

Proof. We first prove the existence of the solution to system (3.2), which is equivalent to (3.1). Choose $p^0 = 0 \in Q$, then determine a sequence $\{(u^n, p^n)\}$ by

$$Au^{n+1} + B_1^t p^{n+1} = f, \tag{3.19}$$

$$B_2 u^{n+1} = g + Cp^n \tag{3.20}$$

for $n = 0, 1, 2, \ldots$. By Theorem 2.1, $\{(u^n, p^n)\}$ is well defined. From (3.19)–(3.20) we have

$$A(u^{n+1} - u^n) + B_1^t (p^{n+1} - p^n) = 0, \tag{3.21}$$

$$B_2(u^{n+1} - u^n) = C(p^n - p^{n-1}). \tag{3.22}$$

Applying estimates (2.9)–(2.10) to this system, we obtain

$$\|u^{n+1} - u^n\|_V \leq \beta_2^{-1}(1 + \alpha^{-1}\|a\|)\|c\|\|p^n - p^{n-1}\|_Q, \tag{3.23}$$

$$\|p^{n+1} - p^n\|_Q \leq \beta_1^{-1}\|a\|\|u^{n+1} - u^n\|_V, \tag{3.24}$$

which implies for $n \geq 1$,

$$\|p^{n+1} - p^n\|_Q \leq \delta_2 \|p^n - p^{n-1}\|_Q \leq \delta_2^n \|p^1\|_Q. \tag{3.25}$$

Therefore, for any nonnegative integer $m > n$,

$$\begin{aligned} \|p^m - p^n\|_Q &\leq \sum_{i=n}^{m-1} \|p^{i+1} - p^i\|_Q \leq \left(\sum_{i=n}^{m-1} \delta_2^i \right) \|p^1\|_Q \\ &\leq \frac{\delta_2^n}{1-\delta_2} \|p^1\|_Q. \end{aligned} \tag{3.26}$$

That is, $\{p^n\}$ is a Cauchy sequence, and there exists a $p \in Q$ such that

$$p^n \to p \quad \text{in } Q. \tag{3.27}$$

On the other hand, it follows from (3.23) and (3.25) that

$$\|u^{n+1} - u^n\|_V \leq \beta_2^{-1}(1 + \alpha^{-1}\|a\|)\|c\|\delta_2^{n-1}\|p^1\|_Q,$$

which implies that for any integer $m > n$,

$$\|u^m - u^n\|_V \leq \beta_2^{-1}(1 + \alpha^{-1}\|a\|)\|c\| \frac{\delta_2^{n-1}}{1-\delta_2} \|p^1\|_Q. \tag{3.28}$$

Hence $\{u^n\}$ is also a Cauchy sequence, and there exists a $u \in V$ such that

$$u^n \to u \quad \text{in } V. \tag{3.29}$$

Letting n tend to infinity in (3.19)–(3.20), we see that $(u, p) \in V \times Q$ solves (3.2). The uniqueness of the solution can be shown using an argument similar to the one used in Theorem 3.1.

It remains to give stability estimates (3.18). As $p^0 = 0$, we see from (3.19)–(3.20) that (u^1, p^1) solves systems (2.4)–(2.5). Then taking $n = 1$ in the estimate (3.26) gives

$$\|p^m\|_Q \leq \frac{1}{1-\delta_2} \|p^1\|_Q,$$

which leads to the first estimate in (3.18) by letting m tend to infinity with the help of estimates (2.9)–(2.10) for (u^1, p^1). In the same manner, taking $n = 1$ in (3.28) leads to the second estimate in (3.18). ∎

Sharpness of the condition on δ_2 in (3.17). Next, we give some simple examples to show that condition (3.17) is a sharp condition guaranteeing the unique solvability of system (3.2). Clearly, in the finite-dimensional case, we have

$$\delta_2 = \beta_1^{-1}\beta_2^{-1}\|A\|_2(1 + \alpha^{-1}\|A\|_2)\|C\|_2. \tag{3.30}$$

Our first example shows that system (3.2) may not necessarily be uniquely solvable if $\delta_2 = 1$. For this, consider $V = R^3$ equipped with the Euclidean norm, $Q = R^1$ in (3.1) or (3.2). Then we choose $B_1 = (1, 0, 0)$, $B_2 = (k, 0, 0)$, $C = -1$, with k a nonzero

constant to be determined later. For the matrix A, we take the following symmetric form with $\varepsilon > 0$:

$$A = \begin{pmatrix} a_{11} & a_{12} & \varepsilon \\ a_{12} & \varepsilon & 0 \\ \varepsilon & 0 & \varepsilon \end{pmatrix}.$$

One can easily verify that conditions (2.6)–(2.7) hold with $\alpha = \varepsilon$, $\beta_1 = 1$, and $\beta_2 = |k|$.

The next steps are intended to construct the matrix A and constant k such that $\delta_2 = 1$, but system (3.2) is not uniquely solvable. That is,

$$|k| = a(1 + \varepsilon^{-1}a), \quad a = \|A\|_2, \tag{3.31}$$

and

$$\det \begin{pmatrix} A & B_1^t \\ B_2 & -C \end{pmatrix} = \det A - k\varepsilon^2 = 0. \tag{3.32}$$

To do this construction, we want to be able to choose the matrix A with three eigenvalues a, a_1, and $-a_1$, respectively, with $a_1 > 0$. In this case, we obtain from (3.31) and (3.32) that

$$aa_1^2 = |\det A| = |k|\varepsilon^2 = a(1 + \varepsilon^{-1}a)\varepsilon^2, \tag{3.33}$$

which gives

$$a_1 = \sqrt{\varepsilon^2 + a\varepsilon}. \tag{3.34}$$

As $a = \|A\|_2$, we must have $a_1 \le a$, i.e.,

$$\varepsilon^2 + a\varepsilon \le a^2. \tag{3.35}$$

On the other hand, the characteristic equation of A is

$$\lambda^3 - d_1\lambda^2 + d_2\lambda - \det A = 0,$$

with $d_1 = a_{11} + 2\varepsilon$ and $d_2 = 2a_{11}\varepsilon - a_{12}^2$. Then by the Vita theorem we know that

$$a + a_1 + (-a_1) = a_{11} + 2\varepsilon, \tag{3.36}$$

$$aa_1 + a(-a_1) + (-a_1^2) = 2a_{11}\varepsilon - a_{12}^2. \tag{3.37}$$

From (3.36) we get

$$a_{11} = a - 2\varepsilon. \tag{3.38}$$

Combining this with (3.34) and (3.37) leads to

$$a_{12}^2 = 2a_{11}\varepsilon + (\varepsilon^2 + a\varepsilon) = 3\varepsilon(a - \varepsilon). \tag{3.39}$$

Then if we take

$$\varepsilon \le a/2, \tag{3.40}$$

(3.35) is satisfied, and by (3.39) we may choose

$$a_{12} = \pm\sqrt{3\varepsilon(a - \varepsilon)}. \tag{3.41}$$

In summary, for any fixed constant $a > 0$, we may choose $\varepsilon \in (0, a/2]$, then compute a_{11} and a_{12} from (3.38) and (3.41), and k from (3.31). Clearly, with the matrix A constructed above, we have $\delta_2 = 1, \|A\|_2 = a$, and $|\det A| = |k|\,\varepsilon^2$ (see (3.33)).

Using $|\det A| = |k|\,\varepsilon^2$, we have either $\det A = k\varepsilon^2$ or $\det A = -k\varepsilon^2$. If the former is valid, then (3.32) holds, and system (3.2) is singular; otherwise we should choose $B_2 = -kB_1$. Then (3.32) holds with k replaced by $-k$, and (3.2) is again singular.

Our second example shows some very interesting results when $V = R^2$ and $Q = R^1$: system (3.2) is always uniquely solvable when $\delta_2 = 1$ but may not be when $\delta_2 > 1$.

To see this, we take $A = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$, $C = -1$, $B_1 = (1, 0)$, and $B_2 = (k, k)$. In this case, system (3.2) reads as follows:

$$\begin{pmatrix} a_{11} & a_{12} & 1 \\ a_{21} & a_{22} & 0 \\ k & k & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ p \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ g \end{pmatrix}. \tag{3.42}$$

Clearly, both B_1 and B_2 satisfy the inf-sup conditions (2.8) with $\beta_1 = 1$ and $\beta_2 = \sqrt{2}|k|$, respectively, and $U_1 = \text{span}\{(0, 1)^t\}$ and $U_2 = \text{span}\{(1, -1)^t\}$. For condition (2.6), a simple calculation gives

$$\inf_{v \in U_2} \sup_{w \in U_1} \frac{(Av, w)}{\|v\|_2 \|w\|_2} = \frac{|a_{21} - a_{22}|}{\sqrt{2}}; \tag{3.43}$$

thus (2.6) holds with $\alpha = |a_{21} - a_{22}|/\sqrt{2}$, and we should assume $a_{21} \neq a_{22}$. The condition $a_{21} \neq a_{22}$ also ensures condition (2.7). Furthermore, it follows from the definition of (3.30) that

$$\sqrt{2}|k| = \frac{1}{\delta_2} a(1 + \alpha^{-1} a), \tag{3.44}$$

where $a = \|A\|_2 = \sigma_{\max}(A)$ stands for the maximal singular value of A. On the other hand, problem (3.42) is uniquely solvable if and only if its coefficient matrix is nonsingular, namely,

$$\det \begin{pmatrix} a_{11} & a_{12} & 1 \\ a_{21} & a_{22} & 0 \\ k & k & 1 \end{pmatrix} = k(a_{21} - a_{22}) + \det A \neq 0. \tag{3.45}$$

Let a_1 be the other singular value of A. Then a^2 and a_1^2 are the two eigenvalues of $A^t A$, and we have

$$a^2 + a_1^2 = \text{tr}(A^t A), \quad a^2 a_1^2 = |\det A|^2. \tag{3.46}$$

Let us first consider any given $\delta_2 > 1$. We want to be able to find a constant $k > 0$ and a matrix A such that both (3.44) and

$$k(a_{21} - a_{22}) + \det A = 0 \tag{3.47}$$

hold. That is, it is possible to construct an example of the linear system (3.42), which is not uniquely solvable.

It follows from (3.43), (3.44), and (3.47) that

$$|\det A| = |k||a_{21} - a_{22}| = \sqrt{2}|k|\alpha = \frac{1}{\delta_2}\alpha a(1 + \alpha^{-1} a) = \frac{1}{\delta_2} a(\alpha + a). \tag{3.48}$$

Combining this with the second equation of (3.46), we see that

$$a^2 a_1^2 = \left(\frac{1}{\delta_2} a(\alpha + a)\right)^2,$$

or $a_1 = (a + \alpha)/\delta_2$. To ensure $a = \|A\|_2$ we need $a_1 \le a$, that is,

$$a \ge \frac{\alpha}{\delta_2 - 1}. \tag{3.49}$$

Now we have to check the first equation of (3.46), that is,

$$a^2 + a_1^2 = \operatorname{tr}(A^t A) = a_{11}^2 + a_{22}^2 + a_{12}^2 + a_{21}^2. \tag{3.50}$$

For simplicity, we take $a_{22} = 0$. Then from definition (3.43), we have $a_{21} = \sqrt{2}\alpha$ (or $-\sqrt{2}\alpha$), so the condition $a_{21} \ne a_{22}$ is fulfilled. Now we take $a_{12} = a_{21} = \sqrt{2}\alpha$, and (3.50) becomes

$$a^2 + a_1^2 = a_{11}^2 + 4\alpha^2,$$

which gives

$$a_{11} = \pm\sqrt{a^2 + a_1^2 - 4\alpha^2}$$

if $a \ge 2\alpha$. Therefore, given $\alpha > 0$, if a satisfies the condition

$$a \ge \alpha \max\left\{2, \frac{1}{\delta_2 - 1}\right\}, \tag{3.51}$$

we obtain a suitable matrix A by the above construction.

Hence, for any fixed $\alpha > 0$, we can choose $a > 0$ satisfying condition (3.51), and afterwards choose k from (3.44). Then we can compute A from (3.50)–(3.51). Clearly, with such a resulting matrix A, we have $\delta_2 = 1, \|A\|_2 = a$, and $|\det A| = |k\, a_{21}|$ (see (3.48)).

As $|\det A| = |k\, a_{21}|$, we have either $\det A = -k\, a_{21}$ or $\det A = k\, a_{21}$. If the former is valid, then (3.47) holds, and system (3.42) is singular. Otherwise we should choose $B_2 = -(k, k)$; then (3.47) is satisfied with k replaced by $-k$, and (3.42) is again singular.

Finally, we consider the case $\delta_2 = 1$. To our surprise, this condition guarantees the unique solvability of (3.42) when $V = R^2$ and $Q = R^1$. This is summarized in the next proposition.

PROPOSITION 3.3. *Let $V = R^2$, $Q = R^1$ and A, B_1, and B_2 satisfy conditions (2.6)–(2.8), $C \ne 0$. Then for any $f \in R^2$ and $g \in R^1$, problem (3.2) is uniquely solvable.*

Proof. Without loss of generality, assume $C = -1$. We proceed by contradiction. Assume that there exist $A = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$, $B_1 = (b_1, b_2)$, $B_2 = (c_1, c_2)$ such that $\delta_2 = 1$; however, system (3.2) is singular, that is,

$$\det \begin{pmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ c_1 & c_2 & 1 \end{pmatrix} = 0. \tag{3.52}$$

Clearly, $\mathrm{Ker}(B_1) = \mathrm{span}\{(-b_2, b_1)^t\}$, $\mathrm{Ker}(B_2) = \mathrm{span}\{(-c_2, c_1)^t\}, \beta_1 = \|B_1\|_2$, $\beta_2 = \|B_2\|_2$, and conditions (2.6)–(2.7) are equivalent to

$$\frac{\left|(-b_2, b_1)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} -c_2 \\ c_1 \end{pmatrix}\right|}{\|B_1\|_2\|B_2\|_2} = \frac{|\det A|}{\|B_1\|_2\|B_2\|_2} = \alpha > 0, \tag{3.53}$$

where we have used the fact, thanks to (3.52), that there holds

$$\det A = b_2c_2a_{11} - b_1c_2a_{21} - b_2c_1a_{12} + b_1c_1a_{22}.$$

On the other hand, it follows from (3.30) and $\delta_2 = 1$ that

$$\|B_1\|_2\|B_2\|_2 = \|A\|_2(1 + \alpha^{-1}\|A\|_2). \tag{3.54}$$

This with (3.53) implies

$$|\det A| = \|B_1\|_2\|B_2\|_2\alpha = \|A\|_2(\|A\|_2 + \alpha). \tag{3.55}$$

Let σ_1 be the smallest singular value of A in comparison with the singular value $a = \|A\|_2$; then from (3.46) and (3.55) it follows that

$$\|A\|_2^2\sigma_1^2 = |\det A|^2 = \|A\|_2^2(\|A\|_2 + \alpha)^2,$$

which gives $\sigma_1 = \|A\|_2 + \alpha$. This is a contradiction. □

Concluding remarks. We have studied the solvability and stability of a generalized saddle-point system in finite- or infinite-dimensional spaces. Sharp solvability conditions and stability estimates are derived. The results generalize some existing saddle-point theories in such a natural way that the results here reduce to the existing ones in the special cases. For example, Theorem 3.1 reduces to Theorem 2.2 when two bilinear forms b_1 and b_2 are equal, while Theorem 3.2 reduces to Theorem 2.1 when the bilinear form $c(p, q)$ vanishes.

Theorems 3.1 and 3.2 hold for both finite- and infinite-dimensional Hilbert spaces V and Q. In the case that V and Q are infinite dimensional, one may further consider their finite-dimensional approximations V_h and Q_h, e.g., by finite element methods, and establish the error estimates for the approximate solutions of problem (3.1) associated with the spaces V_h and Q_h. The detailed discussions on the error estimates are omitted here as they follow naturally from the standard error estimates for systems (2.4) and (2.16), as done in [13, 4]. For the solvability and stability of the resulting finite-dimensional system, one of the most important and difficult issues is to appropriately choose the pair (V_h, Q_h) such that the inf-sup conditions are held with the constants β_i in (2.8) and β in (2.14) independent of the mesh parameter h.

Acknowledgments. The authors wish to thank two anonymous referees for many constructive comments.

REFERENCES

[1] I. Babuska and A. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. Aziz, ed., Academic Press, New York, 1973, pp. 1–359.

[2] C. Bernardi, C. Canuto, and Y. Maday, *Generalized inf-sup conditions for Chebyshev spectral approximation of the Stokes problem,* SIAM J. Numer. Anal., 25 (1988), pp. 1237–1271.

[3] F. Brezzi, *On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers,* RAIRO Math. Model. Numer. Anal., 8 (1974), pp. 129–151.
[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Element Methods,* Springer-Verlag, New York, 1991.
[5] P. Ciarlet, Jr. and J. Zou, *Finite element convergence for the Darwin model to Maxwell's equations,* RAIRO Math. Model. Numer. Anal., 31 (1997), pp. 213–250.
[6] J. Douglas, Jr. and J. Roberts, *Global estimates for mixed methods for second order elliptic equations,* Math. Comput., 44 (1985), pp. 39–52.
[7] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations,* Springer-Verlag, New York, 1986.
[8] G. H. Golub and C. F. Van Loan, *Matrix Computation,* 2nd ed., Johns Hopkins University Press, Baltimore, 1989.
[9] Q. Hu and J. Zou, *An iterative method with variable relaxation parameters for saddle-point problems,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 317–338.
[10] Q. Hu and J. Zou, *Two new variants of nonlinear inexact Uzawa algorithms for saddle-point problems,* Numer. Math., 93 (2002), pp. 333–359.
[11] R. B. Kellogg and B. Liu, *A finite element method for compressible Stokes equations,* SIAM J. Numer. Anal., 33 (1996), pp. 780–788.
[12] R. MacCamy, *Variational procedures for a class of exterior interface problems,* J. Math. Anal. Appl., 78 (1980), pp. 248–266.
[13] R. A. Nicolaides, *Existence, uniqueness and approximation for generalized saddle point problems,* SIAM J. Numer. Anal., 19 (1982), pp. 349–357.
[14] D. Yang, *Iterative schemes for mixed finite element methods with applications to elasticity and compressible flow problems,* Numer. Math., 93 (2002), pp. 177–200.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 237–245

SOLVABILITY OF SYSTEMS OF LINEAR INTERVAL EQUATIONS*

JIRI ROHN†

Abstract. A system of linear interval equations is called solvable if each system of linear equations contained therein is solvable. In the main result of this paper it is proved that solvability of a general rectangular system of linear interval equations can be characterized in terms of nonnegative solvability of a finite number of systems of linear equations which, however, is exponential in matrix size; the problem is proved to be NP-hard. It is shown that three earlier published results are consequences of the main theorem, which is compared with its counterpart valid for linear interval inequalities that turn out to be much less difficult to solve.

Key words. linear interval equations, solvability, complexity, linear interval inequalities

AMS subject classifications. 15A06, 15A39, 65G10

PII. S0895479801398955

1. Introduction.

Let $\mathbf{A} = [\underline{A}, \overline{A}] = \{A;\ \underline{A} \le A \le \overline{A}\}$ be an $m \times n$ interval matrix and $\mathbf{b} = [\underline{b}, \overline{b}] = \{b;\ \underline{b} \le b \le \overline{b}\}$ an m-dimensional interval vector (inequalities are taken componentwise and it is assumed that $\underline{A} \le \overline{A}$ and $\underline{b} \le \overline{b}$, so that both sets are nonempty). A system of linear interval equations, formally written as

$$\mathbf{A}x = \mathbf{b}, \tag{1}$$

is defined to be the family of all systems of linear equations

$$Ax = b \tag{2}$$

with data satisfying

$$A \in \mathbf{A}, \qquad b \in \mathbf{b}. \tag{3}$$

During approximately the last 35 years, much attention has been paid to systems of linear interval equations (1) with square interval matrices (cf., e.g., the monographs by Alefeld and Herzberger [1], Neumaier [4], Kreinovich et al. [3]). On the contrary, the general rectangular case has been much less studied and remains much less understood.

In this paper we raise the question of solvability of general systems of linear interval equations with rectangular matrices. A system (1) is called solvable if each system in the family (2), (3) is solvable (i.e., has a solution). The reasons for introducing this property are obvious: assuming we are interested in solvability of a linear system $A_0x = b_0$, whose data A_0, b_0 are not known exactly but only known to belong to $\mathbf{A}$ and $\mathbf{b}$, respectively, we can be sure that the system $A_0x = b_0$ is solvable only if each system (2) with data satisfying (3) possesses this property.

Except for the trivial case of $\underline{A} = \overline{A}$ and $\underline{b} = \overline{b}$, the family (2), (3) consists of infinitely many linear systems. In the main result of this paper (Theorem 3) we prove that a system (1) is solvable if and only if a finite number of linear systems are

*Received by the editors November 29, 2001; accepted for publication (in revised form) by L. Vandenberghe January 22, 2003; published electronically May 29, 2003. This work was supported by the Czech Republic Grant Agency under grant 201/01/0343.
http://www.siam.org/journals/simax/25-1/39895.html
†Institute of Computer Science, Czech Academy of Sciences, Pod vodárenskou věží 2, 182 07 Prague, Czech Republic (rohn@cs.cas.cz).

nonnegatively solvable (i.e., have nonnegative solutions). These systems are formed in the following way: For each $i \in \{1, \dots, m\}$, the ith equation of such a system is either of the form

$$(\underline{A}x^1 - \overline{A}x^2)_i = \overline{b}_i \tag{4}$$

or of the form

$$(\overline{A}x^1 - \underline{A}x^2)_i = \underline{b}_i. \tag{5}$$

Since for each of the m equations we have two options to choose from, there are altogether 2^m linear systems of this form in general (notice that the matrix of each such system is of size $m \times 2n$). But if the ith rows of $\underline{A}$ and $\overline{A}$ are equal and if $\underline{b}_i = \overline{b}_i$, then (4) and (5) coincide. Hence the exact number of mutually different linear systems to be solved is 2^q, where q is the number of nonzero rows of the matrix $(\overline{A} - \underline{A}, \overline{b} - \underline{b})$. This shows that the characterization, although generally exponential, can be of practical use for problems with moderate values of q.

As shown in section 3, the proof of this result is nontrivial and relies on the Farkas lemma and on the Oettli–Prager theorem. In section 4 we show that the main result offers a unified view of three different, earlier results published independently: characterization of nonnegative solvability of (1) (Theorem 4), characterization of regularity of interval matrices (Theorem 5), and the convex-hull theorem (Theorem 6). Next it is shown that the problem of checking solvability of linear interval equations is NP-hard (Theorem 7); this explains the exponentiality inherent in formulation of the main result. Finally, we compare the characterization of solvability of linear interval equations in Theorem 3 with that of linear interval inequalities. Unlike the case of exact data, these two problems turn out to be of different complexity since solvability of a system of linear interval inequalities is characterized by solvability of *one* system of linear inequalities only (Theorem 8). A brief discussion of the reasons for this difference concludes the paper.

Throughout the paper we shall use the following notation. For an interval matrix $\mathbf{A} = [\underline{A}, \overline{A}]$ we define

$$A_c = \tfrac{1}{2}(\underline{A} + \overline{A})$$

(the center matrix) and

$$\Delta = \tfrac{1}{2}(\overline{A} - \underline{A})$$

(the radius matrix). Then $\underline{A} = A_c - \Delta$ and $\overline{A} = A_c + \Delta$, so that we also can write $\mathbf{A} = [A_c - \Delta, A_c + \Delta]$. Similarly, for the right-hand side $\mathbf{b} = [\underline{b}, \overline{b}]$, setting

$$b_c = \tfrac{1}{2}(\underline{b} + \overline{b})$$

and

$$\delta = \tfrac{1}{2}(\overline{b} - \underline{b}),$$

we have $\mathbf{b} = [b_c - \delta, b_c + \delta]$. This form of expressing the bounds turns out to be more useful, mainly due to the Oettli–Prager description of the solution set of (1) (Theorem 2 below). For a vector $x = (x_i)$, its absolute value is defined by $|x| = (|x_i|)$; Conv X denotes the convex hull of X. We define

$$Y_m = \{y \in \mathbb{R}^m;\ y_j \in \{-1, 1\} \text{ for each } j\};$$

i.e., Y_m is the set of all ± 1-vectors in $\mathbb{R}^m$; its cardinality is obviously 2^m. Finally, for each $y \in Y_m$ we denote

$$T_y = \operatorname{diag}(y_1, \ldots, y_m) = \begin{pmatrix} y_1 & 0 & \ldots & 0 \\ 0 & y_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & y_m \end{pmatrix}.$$

Notice that $A_c - T_y\Delta \in \mathbf{A}$, $A_c + T_y\Delta \in \mathbf{A}$, and $b_c + T_y\delta \in \mathbf{b}$ for each $y \in Y_m$ (these quantities appear in formulation of the main result, equation (8) below).

2. Preliminaries. In order to keep the paper self-contained, we give here explicit formulations of two well-known results that will be used in the proof of the main theorem. The first is the Farkas lemma.

LEMMA 1 (Farkas [2]). *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then the system*

$$\begin{aligned} Ax &= b, \\ x &\geq 0, \end{aligned}$$

has a solution if and only if each $p \in \mathbb{R}^m$ with $A^T p \geq 0$ satisfies $b^T p \geq 0$.

Our second auxiliary result is the Oettli–Prager theorem. If $\mathbf{A} = [A_c - \Delta, A_c + \Delta]$ is an $m \times n$ interval matrix and $\mathbf{b} = [b_c - \delta, b_c + \delta]$ is an m-dimensional interval vector, then the solution set of the system of linear interval equations

$$\mathbf{A}x = \mathbf{b}$$

is defined by

$$X = \{x;\ Ax = b \text{ for some } A \in \mathbf{A}, b \in \mathbf{b}\}. \tag{6}$$

The Oettli–Prager theorem gives a description of the solution set by means of a certain nonlinear inequality.

THEOREM 2 (Oettli and Prager [5]). *We have*

$$X = \{x;\ |A_c x - b_c| \leq \Delta|x| + \delta\}. \tag{7}$$

Hence, if x satisfies the inequality in (7), then $Ax = b$ for some $A \in \mathbf{A}$ and $b \in \mathbf{b}$. In fact, A and b can be explicitly expressed in terms of x (see the proof of Theorem 2.1 in [8]), but we shall not need it in this paper.

3. Solvability. In this section we present the main result of this paper, a characterization of solvability of linear interval equations defined in the following way. Let $\mathbf{A}$ be an $m \times n$ interval matrix and $\mathbf{b}$ an m-dimensional interval vector. The system of linear interval equations $\mathbf{A}x = \mathbf{b}$ is said to be *solvable* if each system $Ax = b$ with $A \in \mathbf{A}$, $b \in \mathbf{b}$ has a solution.

Except for the trivial case $\Delta = 0$ and $\delta = 0$, the family $\mathbf{A}x = \mathbf{b}$ consists of infinitely many systems. Yet the following theorem shows that solvability of $\mathbf{A}x = \mathbf{b}$ can be characterized in terms of nonnegative solvability of a finite number of linear systems, although this number is generally exponential in matrix size.

THEOREM 3. *A system of linear interval equations* $\mathbf{A}x = \mathbf{b}$ *is solvable if and only if for each* $y \in Y_m$ *the system*

$$(A_c - T_y\Delta)x^1 - (A_c + T_y\Delta)x^2 = b_c + T_y\delta, \tag{8}$$

$$x^1 \geq 0, \qquad x^2 \geq 0, \tag{9}$$

has a solution x^1_y, x^2_y. *Moreover, if this is the case, then for each* $A \in \mathbf{A}$, $b \in \mathbf{b}$ *the system* $Ax = b$ *has a solution in the set*

$$\mathrm{Conv}\{x^1_y - x^2_y;\ y \in Y_m\}.$$

Proof. "Only if": Let $\mathbf{A}x = \mathbf{b}$ be solvable. Assume to the contrary that (8), (9) does not have a solution for some $y \in Y_m$. Then the Farkas lemma implies existence of a $p \in \mathbb{R}^m$ satisfying

$$(A_c - T_y\Delta)^T p \geq 0, \tag{10}$$

$$(A_c + T_y\Delta)^T p \leq 0, \tag{11}$$

$$(b_c + T_y\delta)^T p < 0. \tag{12}$$

Now (10) and (11) together give

$$\Delta^T T_y p \leq A_c^T p \leq -\Delta^T T_y p,$$

hence

$$|A_c^T p| \leq -\Delta^T T_y p = |-\Delta^T T_y p| \leq \Delta^T |p|,$$

and the Oettli–Prager theorem as applied to the system $[A_c^T - \Delta^T, A_c^T + \Delta^T]z = [0, 0]$ shows that there exists a matrix $A \in \mathbf{A}$ such that

$$A^T p = 0. \tag{13}$$

In light of the Farkas lemma, (13) and (12) mean that the system

$$Ax = b_c + T_y\delta$$

has no solution, which contradicts our assumption since $A \in \mathbf{A}$ and $b_c + T_y\delta \in \mathbf{b}$.

"If": Conversely, let for each $y \in Y_m$ the system (8), (9) have a solution x^1_y, x^2_y. Let $A \in \mathbf{A}$, $b \in \mathbf{b}$. To prove that the system $Ax = b$ has a solution, we first show that $T_y A x_y \geq T_y b$ holds for each $y \in Y_m$, where $x_y = x^1_y - x^2_y$. Thus let $y \in Y_m$. Then we have

$$\begin{aligned} T_y(Ax_y - b) &= T_y(A_c x_y - b_c) + T_y(A - A_c)x_y + T_y(b_c - b) \\ &\geq T_y(A_c x_y - b_c) - \Delta|x_y| - \delta \end{aligned}$$

since $|T_y(A - A_c)x_y| \leq \Delta|x_y|$, which implies $T_y(A - A_c)x_y \geq -\Delta|x_y|$, and similarly $|T_y(b_c - b)| \leq \delta$ implies $T_y(b_c - b) \geq -\delta$; thus

$$\begin{aligned} T_y(Ax_y - b) &\geq T_y(A_c(x^1_y - x^2_y) - b_c) - \Delta|x^1_y - x^2_y| - \delta \\ &\geq T_y(A_c(x^1_y - x^2_y) - b_c) - \Delta(x^1_y + x^2_y) - \delta \\ &= T_y((A_c - T_y\Delta)x^1_y - (A_c + T_y\Delta)x^2_y - (b_c + T_y\delta)) \\ &= 0 \end{aligned}$$

since x_y^1, x_y^2 solve (8), (9). In this way we have proved that for each $y \in Y_m$, x_y satisfies

$$T_y A x_y \geq T_y b. \tag{14}$$

Using (14), we shall next prove that the system of linear equations

$$\sum_{y \in Y_m} \lambda_y A x_y = b, \tag{15}$$

$$\sum_{y \in Y_m} \lambda_y = 1, \tag{16}$$

has a solution $\lambda_y \geq 0$, $y \in Y_m$. In view of the Farkas lemma, it suffices to show that for each $p \in \mathbb{R}^m$ and each $p_0 \in \mathbb{R}^1$,

$$p^T A x_y + p_0 \geq 0 \text{ for each } y \in Y_m \tag{17}$$

implies

$$p^T b + p_0 \geq 0. \tag{18}$$

Thus let p and p_0 satisfy (17). Define $y \in Y_m$ by $y_i = -1$ if $p_i \geq 0$ and by $y_i = 1$ if $p_i < 0$ ($i = 1, \ldots, m$), then $p = -T_y|p|$, and from (14), (17) we have

$$p^T b + p_0 = -|p|^T T_y b + p_0 \geq -|p|^T T_y A x_y + p_0 = p^T A x_y + p_0 \geq 0,$$

which proves (18). Hence the system (15), (16) has a solution $\lambda_y \geq 0$, $y \in Y_m$. Put $x = \sum_{y \in Y_m} \lambda_y x_y$, then $Ax = b$ by (15), and x belongs to the set $\mathrm{Conv}\{x_y;\ y \in Y_m\} = \mathrm{Conv}\{x_y^1 - x_y^2;\ y \in Y_m\}$ by (16). This proves the "if" part and also the additional assertion. □

Let us have a closer look at the form of systems (8). If $y_i = 1$, then the ith rows of $A_c - T_y\Delta$ and $A_c + T_y\Delta$ are equal to the ith rows of $\underline{A}$ and $\overline{A}$, respectively, and $(b_c + T_y\delta)_i = \overline{b}_i$. This means that in this case the ith equation of (8) has the form

$$(\underline{A}x^1 - \overline{A}x^2)_i = \overline{b}_i, \tag{19}$$

and similarly, in case $y_i = -1$ it is of the form

$$(\overline{A}x^1 - \underline{A}x^2)_i = \underline{b}_i. \tag{20}$$

Hence we can see that the family of systems (8) for all $y \in Y_m$ is just the family of all systems whose ith equations are either of the form (19) or of the form (20) for $i = 1, \ldots, m$. The number of mutually different such systems is exactly 2^q, where q is the number of nonzero rows of the matrix $(\Delta,\ \delta)$. Hence, despite the inherent exponentiality, Theorem 3 can be of practical use if q is of moderate size.

In the "if" part of the proof we proved that for each $A \in \mathbf{A}$ and $b \in \mathbf{b}$ the equation $Ax = b$ has a solution in the set $\mathrm{Conv}\{x_y^1 - x_y^2;\ y \in Y_m\}$. The proof, relying on the Farkas lemma, was purely existential. It is worth noting, however, that such a solution can be found in a constructive way when using an algorithm described in [9]. For its description we need a special order of elements of Y_m defined inductively via the sets Y_j, $j = 1, \ldots, m-1$, in the following way:

(i) The order of Y_1 is $-1, 1$.

(ii) If $y_1, \dots, y_{2^j}$ is the order of Y_j, then $(y_1, -1), \dots, (y_{2^j}, -1)$, $(y_1, 1), \dots, (y_{2^j}, 1)$ is the order of Y_{j+1}.

Further, for a sequence $z_1, \dots, z_{2h}$ with an even number of elements, each pair z_j, z_{j+h} is called a conjugate pair, $j = 1, \dots, h$. As in Theorem 3, for each $y \in Y_m$, let x^1_y and x^2_y be a solution to (8), (9). Then the algorithm runs as follows:

1. Select $A \in \mathbf{A}$ and $b \in \mathbf{b}$.

2. Form a sequence of vectors $((x^1_{-y} - x^2_{-y})^T, (A(x^1_{-y} - x^2_{-y}) - b)^T)^T$ set in the order of the y's in Y_m.

3. For each conjugate pair x, x' in the current sequence compute

$$\lambda = \begin{cases} \frac{x'_k}{x'_k - x_k} & \text{if } x'_k \neq x_k, \\ 1 & \text{otherwise,} \end{cases}$$

where k is the index of the current last entry, and set

$$x := \lambda x + (1 - \lambda)x'.$$

4. Cancel the second part of the sequence and in the remaining part delete the last entry of each vector.

5. If there remains a single vector x, terminate: x solves $Ax = b$ and $x \in \mathrm{Conv}\{x^1_y - x^2_y;\ y \in Y_m\}$. Otherwise go to step 3.

The algorithm starts with 2^m vectors $((x^1_{-y} - x^2_{-y})^T, (A(x^1_{-y} - x^2_{-y}) - b)^T)^T \in \mathbb{R}^{n+m}$, $y \in Y_m$, and proceeds by halving the sequence and deleting the last entry; hence it is finite and at the end produces a single vector $x \in \mathbb{R}^n$. The assertion made in step 5 is a consequence of Theorem 2 in [9] because we have

$$T_y A x_y \geq T_y b$$

for each $y \in Y_m$; hence also

$$T_y A x_{-y} \leq T_y b$$

for each $y \in Y_m$, which is the form used in [9].

4. Remarks. In this section we show that Theorem 3 offers a unified view of three earlier published results whose original proofs were rather involved and that can be easily obtained, and perhaps also better understood, as consequences of the main result. Next we compare the results for linear interval equations with those for linear interval inequalities that, unlike the case of exact data, turn out to be of different complexity.

First we consider nonnegative solvability. A linear interval system $\mathbf{A}x = \mathbf{b}$ is called *nonnegatively solvable* if each system $Ax = b$ with $A \in \mathbf{A}$, $b \in \mathbf{b}$ is nonnegatively solvable. The following characterization (without the convex hull part) was proved in [7].

THEOREM 4. *A system of linear interval equations $\mathbf{A}x = \mathbf{b}$ is nonnegatively solvable if and only if for each $y \in Y_m$ the system*

$$(A_c - T_y \Delta)x = b_c + T_y \delta \tag{21}$$

has a nonnegative solution x_y. Moreover, if this is the case, then for each $A \in \mathbf{A}$, $b \in \mathbf{b}$ the system $Ax = b$ has a solution in the set

$$\mathrm{Conv}\{x_y;\ y \in Y_m\}.$$

Repeating the argument following the proof of Theorem 3, we can say that the ith row of (21) is of the form

$$(\underline{A}x)_i = \overline{b}_i$$

if $y_i = 1$ and of the form

$$(\overline{A}x)_i = \underline{b}_i$$

if $y_i = -1$ (hence, unlike (8), the system matrix always belongs to $\mathbf{A}$ in this case), and the number of mutually different systems (21) is again 2^q, where q is the number of nonzero rows of the matrix $(\Delta,\ \delta)$.

Next we turn to square matrices. A square interval matrix $\mathbf{A}$ is said to be *regular* if each $A \in \mathbf{A}$ is nonsingular. A number of necessary and sufficient regularity conditions was given in Theorem 5.1 in [8]. One of them is the following, which is again obtained as an easy consequence of Theorem 3.

THEOREM 5. *An interval matrix* $\mathbf{A}$ *is regular if and only if for each* $y \in Y_m$ *the system*

$$(A_c - T_y\Delta)x^1 - (A_c + T_y\Delta)x^2 = y,$$

$$x^1 \geq 0, \qquad x^2 \geq 0,$$

has a solution.

If $\mathbf{A}$ is regular, then for each right-hand side $\mathbf{b}$ the system of linear interval equations $\mathbf{A}x = \mathbf{b}$ is solvable, and hence the system (8), (9) has a solution for each $y \in Y_m$. But, as shown in Theorem 2.2 in [8], in this case we can do essentially better; namely, if we impose an additional complementarity constraint, then the solution turns out to be unique.

THEOREM 6. *Let* $\mathbf{A}$ *be regular. Then for each* $y \in Y_m$ *the system*

$$(A_c - T_y\Delta)x^1 - (A_c + T_y\Delta)x^2 = b_c + T_y\delta, \tag{22}$$

$$x^1 \geq 0, \qquad x^2 \geq 0, \tag{23}$$

$$(x^1)^T x^2 = 0, \tag{24}$$

has a unique solution x^1_y, x^2_y, *and for the solution set* X *of* $\mathbf{A}x = \mathbf{b}$ *defined by* (6) *we have*

$$\operatorname{Conv} X = \operatorname{Conv}\{x^1_y - x^2_y;\ y \in Y_m\}. \tag{25}$$

Because of (24), for each $y \in Y_m$ the system (22)–(24) can be equivalently written as

$$A_c x - T_y\Delta|x| = b_c + T_y\delta$$

and its unique solution x_y satisfies $x_y = x^1_y - x^2_y$, so that (25) takes the form

$$\operatorname{Conv} X = \operatorname{Conv}\{x_y;\ y \in Y_m\}. \tag{26}$$

This is the form used in [8]. Theorem 6 has important theoretical consequences. If $[\underline{x}, \overline{x}]$ is the interval hull (optimal enclosure) of the solution set X, then (26) gives

$$\underline{x}_i = \min_{y \in Y_m} (x_y)_i,$$
$$\overline{x}_i = \max_{y \in Y_m} (x_y)_i$$

for $i = 1, \ldots, n$. This result forms a basis for several enclosure algorithms; see [8] and [10].

The number of systems (8), (9) to be checked for solvability is exponential in the number of rows of $\mathbf{A}$ in general. This characterization is unlikely to be substantially improved because of the following complexity result.

THEOREM 7. *Checking solvability of linear interval equations is NP-hard.*

The proof follows easily from the fact that checking regularity of interval matrices, which is an NP-complete problem as proved in [6], can obviously be reduced in polynomial time to the problem of checking solvability of linear interval equations, which is thus NP-hard. NP-hardness of checking nonnegative solvability was established in part 2 of the proof of the main result in [11].

It is instructive to compare the main result of Theorem 3 with its counterpart valid for linear interval inequalities. Analogously to the terminology in section 3, we call a system of linear interval inequalities

$$\mathbf{A}x \leq \mathbf{b}$$

solvable if each system $Ax \leq b$ with $A \in \mathbf{A}$, $b \in \mathbf{b}$ has a solution. Yet the characterization in this case, as shown by Rohn and Kreslová [12], is qualitatively different: although the proof of the "only if" part follows rather similar lines as the respective part of the proof of Theorem 3, it turns out that only *one* system of linear inequalities is to be checked for solvability.

THEOREM 8. *A system of linear interval inequalities* $\mathbf{A}x \leq \mathbf{b}$ *is solvable if and only if the system*

$$\overline{A}x^1 - \underline{A}x^2 \leq \underline{b},$$

$$x^1 \geq 0, \qquad x^2 \geq 0,$$

has a solution.

As a byproduct of the proof we obtain a nontrivial fact which is worth mentioning explicitly [12].

THEOREM 9. *A system of linear interval inequalities* $\mathbf{A}x \leq \mathbf{b}$ *is solvable if and only if all the systems* $Ax \leq b$, $A \in \mathbf{A}$, $b \in \mathbf{b}$, *have a solution in common.*

Based on this comparison, we can conclude that, as regards solvability, linear interval equations and linear interval inequalities behave differently. In the case of exact data, a system of linear equations

$$Ax = b \tag{27}$$

can be equivalently written as

$$\begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}, \tag{28}$$

and hence any algorithm for checking solvability of (28) can be employed for checking solvability of (27). This is no more true in the case of inexact data: A system

$$\mathbf{A}x = \mathbf{b} \tag{29}$$

cannot be equivalently written as

$$\begin{pmatrix} \mathbf{A} \\ -\mathbf{A} \end{pmatrix} x \leq \begin{pmatrix} \mathbf{b} \\ -\mathbf{b} \end{pmatrix} \tag{30}$$

because of dependence of data in (28) which is not reflected in (30), where the same coefficient (say, a_{ij}) is allowed to take on different values within its two occurrences. Hence the solution set of (29) is always a part of that of (30), but the converse inclusion need not be true.

REFERENCES

[1] G. Alefeld and J. Herzberger, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] J. Farkas, *Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1902), pp. 1–27.

[3] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[4] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.

[5] 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.

[6] S. Poljak and J. Rohn, *Checking robust nonsingularity is NP-hard*, Math. Control Signals Systems, 6 (1993), pp. 1–9.

[7] J. Rohn, *Strong solvability of interval linear programming problems*, Computing, 26 (1981), pp. 79–82.

[8] J. Rohn, *Systems of linear interval equations*, Linear Algebra Appl., 126 (1989), pp. 39–78.

[9] J. Rohn, *An existence theorem for systems of linear equations*, Linear Multilinear Algebra, 29 (1991), pp. 141–144.

[10] J. Rohn, *Cheap and tight bounds: The recent result by E. Hansen can be made more efficient*, Interval Comput., 4 (1993), pp. 13–21.

[11] J. Rohn, *Linear programming with inexact data is NP-hard*, Z. Angew. Math. Mech., 78, Suppl. 3, (1998), pp. S1051–S1052.

[12] J. Rohn and J. Kreslová, *Linear interval inequalities*, Linear Multilinear Algebra, 38 (1994), pp. 79–82.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 246–265

AN IMPLICITLY RESTARTED REFINED BIDIAGONALIZATION LANCZOS METHOD FOR COMPUTING A PARTIAL SINGULAR VALUE DECOMPOSITION*

ZHONGXIAO JIA† AND DATIAN NIU‡

Abstract. The bidiagonalization Lanczos method can be used for computing a few of the largest or smallest singular values and corresponding singular vectors of a large matrix, but the method may encounter some convergence problems. In this paper the convergence of the method is analyzed, showing why it may converge erratically and perhaps fail to converge. To correct this possible nonconvergence and improve the method, a refined bidiagonalization Lanczos method is proposed. The implicitly restarting technique due to Sorensen is applied to the method, and an implicitly restarted refined bidiagonalization Lanczos algorithm (IRRBL) is developed. A new selection of shifts is proposed for use within IRRBL, called refined shifts, and a reliable and efficient algorithm is developed for computing the refined shifts. Numerical experiments show that IRRBL can perform better than the implicitly restarted bidiagonalization Lanczos algorithm (IRBL) proposed by Larsen, in particular when the smallest singular triplets are desired.

Key words. singular value, singular vector, the bidiagonalization Lanczos method, Ritz value, Ritz vector, refined Ritz vector, the refined bidiagonalization Lanczos method, implicit restart, exact shifts, refined shifts, convergence

AMS subject classifications. 65F15, 15A18

PII. S0895479802404192

1. Introduction.

We are concerned with the following problem.

PROBLEM 1. *Compute numerically the k largest or smallest singular values and corresponding left and right singular vectors of a large real $M \times N$ matrix $A \in \mathcal{R}^{M\times N}$, where k is much smaller than M and N.*

Such a problem arises from many applications, e.g., total least squares problems, determination of numerical rank of a matrix, regression analysis, and image processing and pattern recognitions.

Without loss of generality, we assume that $M \geq N$ (otherwise we work on A^{T}, the transpose of A). Let σ_i, $i = 1, 2, \ldots, N$, be the singular values of A, labeled in decreasing or increasing order, and u_i and v_i the corresponding left and right singular vectors. The triplets (σ_i, u_i, v_i) are called the singular triplets of A. We then have the singular value decomposition (SVD) of A:

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^{\mathrm{T}} = U_1 \Sigma V^{\mathrm{T}}, \tag{1.1}$$

where $U = (u_1, u_2, \ldots, u_M) = (U_1, U_2)$ is orthogonal with $U_1 = (u_1, u_2, \ldots, u_N)$, $V = (v_1, v_2, \ldots, v_N)$ orthogonal, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N)$.

Consider the augmented matrix

$$\tilde{A} = \begin{pmatrix} 0 & A \\ A^{\mathrm{T}} & 0 \end{pmatrix}. \tag{1.2}$$

*Received by the editors March 18, 2002; accepted for publication (in revised form) by H. van der Vorst December 17, 2002; published electronically May 29, 2003. This work was supported by Special Funds for State Major Basic Research Projects (G19990328).

http://www.siam.org/journals/simax/25-1/40419.html

†Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China (zjia@math.tsinghua.edu.cn).

‡Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China.

It is easily verified that $\tilde{A}$ has the $2N$ eigenvalues $\pm\sigma_1, \ldots, \pm\sigma_N$ and $M - N$ eigenvalues zero. The eigenvectors associated with σ_i and $-\sigma_i$ are $\frac{1}{\sqrt{2}} \left(u_i^{\mathrm{T}}, v_i^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\frac{1}{\sqrt{2}} \left(u_i^{\mathrm{T}}, -v_i^{\mathrm{T}}\right)^{\mathrm{T}}$, respectively, and the eigenvectors associated with the eigenvalues zero are $\left(u^{\mathrm{T}}, 0^{\mathrm{T}}\right)^{\mathrm{T}}$, where the u's are orthogonal to $u_1, \ldots, u_N$. Therefore, we get an eigenproblem equivalent to (1.1).

PROBLEM 2. *Compute numerically the k largest or smallest positive eigenvalues of $\tilde{A}$ and the associated eigenvectors.*

Since M and N are assumed to be large and the dimension of $\tilde{A}$ is $M + N$, only projection methods are reasonable to solve Problem 2. A typical method is the symmetric Lanczos method [26]. However, if the method is applied to solve Problem 2 directly and explicitly, then the computational complexity and the memory requirement will be greatly increased. So it is not preferable to work on $\tilde{A}$ directly. Another consequence of using $\tilde{A}$ explicitly is that the smallest positive eigenvalues of $\tilde{A}$ are now interior ones, while they are the leftmost (extreme) singular values of A. Note that the symmetric Lanczos method usually favors the extreme eigenvalues and the associated eigenvectors, and it is very difficult to compute interior eigenpairs [26]. Therefore, we should not work on $\tilde{A}$ directly for computing the smallest singular values of A.

Because of the mentioned drawbacks, we attempt to solve Problem 1 by working on $\tilde{A}$ implicitly. It will turn out that the bidiagonalization Lanczos method [4, 5, 9] and its refined version to be proposed in this paper can settle these problems elegantly.

Over the past decade, the implicit restarting technique due to Sorensen [27] has proven to be a powerful and efficient tool for restarting a Krylov subspace algorithm. It has been used in various contexts, e.g., [2, 3, 9, 14, 17, 28, 29, 30]. It may save computational cost considerably at each restart and maintain numerical stability. However, it should be kept in mind that for an overall performance one of the keys for the success of an implicitly restarted Krylov algorithm is reasonable selection of shifts involved [14, 17]. Other applications of the technique are possible. Björck, Grimme, and van Dooren [3] successfully applied the implicit restarting technique to the lower bidiagonalization Lanczos method for ill-posed least squares problems. Wang and Zha [30] proposed a variant of their algorithm for computing a few largest singular values of A. Both algorithms take zeros as shifts. Larsen [22] developed an implicitly restarted bidiagonalization Lanczos algorithm and discussed many issues, including selection of shifts and the maintenance of semiorthogonality of Lanczos vectors. A few packages are now available for computing a partial SVD of A, e.g., PROPACK and LANSO [21, 22] and ARPACK [23]. PROPACK works on A directly, and LANSO is a symmetric Lanczos algorithm with selective orthogonalization and solves the eigenproblem of $A^T A$ or $\tilde{A}$. Both packages work without restarting until the desired singular values and/or singular vectors have been found, while ARPACK solves the eigenproblems of $A^T A$ and $\tilde{A}$ whose MATLAB counterparts are eigs.m and svds.m, respectively.

The paper is organized as follows. In section 2, we describe the bidiagonalization Lanczos process, and we show how the process can be combined with the Rayleigh–Ritz procedure for computing a partial SVD of A. We then make a convergence analysis of approximate singular values (Ritz values) and approximate singular vectors (Ritz vectors). We show that, under the natural hypothesis that the deviations of a desired singular vector from a sequence of Krylov subspaces tend to zero, there is a Ritz value that converges to the desired singular value, while, on the other hand, the

associated Ritz vectors may converge erratically and even may fail to converge to the desired left and right singular vectors. In section 3, based on the refined projection methods for large matrix eigenproblems [28, 29] proposed by Jia [10, 12, 13, 15, 16], by exploiting the bidiagonalization Lanczos process we propose a refined bidiagonalization Lanczos method for Problem 1. The refined method has a different background from the standard method. The fundamental difference between the refined method and the standard method is that rather than using Ritz approximations, the former seeks new approximate singular vectors, called refined singular vector approximations or simply refined Ritz approximations, from certain Krylov subspaces that minimize the norms of certain residuals and use them to approximate the desired singular vectors. We analyze the convergence of refined Ritz approximations and show that they always converge, provided that the deviations tend to zero. In section 4, we review an implicitly restarted bidiagonalization Lanczos algorithm (IRBL) for Problem 1, in which the shifts are often selected as those unwanted approximate singular values (Ritz values) [21, 22], called exact shifts. In order to compute the large close singular values and improve performance, Larsen [22] proposed a simple adaptive shifting strategy that replaces bad shifts by zero. This strategy often appears to be quite effective. In section 5, motivated by Jia's work [14, 17], we discuss the selection of shifts involved in an implicitly restarted algorithm, and we propose a new shifts scheme, called refined shifts, for use within the implicitly restarted refined bidiagonalization Lanczos algorithm (IRRBL). Still, we exploit Larsen's adaptive shifting strategy to compute the large close singular values. We show qualitatively that the refined shifts are better than the exact shifts for use within IRBL. We discuss how to compute the refined shifts efficiently and reliably. However, Larsen's adaptive shifting strategy cannot work for computing the smallest close singular values. To this end, we give a heuristic analysis and propose to replace bad shifts by the largest Ritz value at the current cycle. In section 6 we make numerical experiments on several real-world problems, indicating that IRRBL can be more efficient than IRBL, in particular for computing the smallest singular triplets. To be complete, we also compare our algorithm with PROPACK, LANSO, and ARPACK and show the superiority of IRRBL. Finally, in section 7 we draw some conclusions.

Some notation to be used is introduced now. Throughout the paper, denote by $||\cdot||$ the Euclidean norm, by $\mathcal{K}_m(C, w_1) = span\{w_1, Cw_1, \ldots, C^{m-1}w_1\}$ the m-dimensional Krylov subspace generated by C and a unit length vector w_1, and by e_m the mth coordinate vector of dimension m.

2. The bidiagonalization Lanczos process and method.

2.1. The bidiagonalization Lanczos process.

We first describe the lower bidiagonalization Lanczos process due to Paige and Saunders [25], which is a variant of the upper bidiagonalization Lanczos process due to Golub and Kahan [7].

ALGORITHM 1. THE m-STEP BIDIAGONALIZATION LANCZOS PROCESS.

1. Start: Choose a unit length vector p_1 of dimension M, $\beta_1 = 1$ and let $q_0 = 0$.
2. For $i = 1, 2, \ldots, m$
   (a) $r_i = A^{\mathrm{T}} p_i - \beta_i q_{i-1}$
   $\alpha_i = ||r_i||, q_i = r_i/\alpha_i$
   (b) $z_i = Aq_i - \alpha_i p_i$
   $\beta_{i+1} = ||z_i||, p_{i+1} = z_i/\beta_{i+1}$
   Endfor

Define $Q_m = (q_1, q_2, \ldots, q_m)$ and $P_{m+1} = (p_1, p_2, \ldots, p_{m+1})$. Then Algorithm 1 can be written in matrix form

$$AQ_m = P_{m+1}B_m, \tag{2.1}$$

$$A^T P_{m+1} = Q_m B_m^{\mathrm{T}} + \alpha_{m+1} q_{m+1} e_{m+1}^{\mathrm{T}}. \tag{2.2}$$

Therefore, we have

$$P_{m+1}^T A Q_m = B_m, \tag{2.3}$$

where

$$B_m = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_m \\ & & & \beta_{m+1} \end{pmatrix} \in \mathcal{R}^{(m+1)\times m}$$

is called the projection matrix of A with the left subspace $span\{P_{m+1}\}$ and the right subspace $span\{Q_m\}$.

Note that the above three relations can also be written as

$$\begin{aligned} \tilde{A} \begin{pmatrix} P_{m+1} & 0 \\ 0 & Q_m \end{pmatrix} &= \begin{pmatrix} P_{m+1} & 0 \\ 0 & Q_m \end{pmatrix} \begin{pmatrix} 0 & B_m \\ B_m^{\mathrm{T}} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ r_{m+1} e_m^{\mathrm{T}} & 0 \end{pmatrix} \\ &= \begin{pmatrix} P_{m+1} & 0 & 0 \\ 0 & Q_m & q_{m+1} \end{pmatrix} \begin{pmatrix} 0 & B_m \\ B_m^{\mathrm{T}} & 0 \\ \alpha_{m+1} e_m^{\mathrm{T}} & 0 \end{pmatrix}. \end{aligned} \tag{2.4}$$

In finite precision arithmetic, it is well known [26] that the orthogonality of P_{m+1} and Q_m, Lanczos basis vectors, may lose soon. In order to maintain numerical (semi)orthogonality, an efficient approach is to use a partial reorthogonalization. For details, refer to Larsen [21, 22].

It is known that there is a close relationship between the above bidiagonalization process and the symmetric Lanczos process applied to $A^T A$ and AA^{T}, both of which have the same nonzero eigenvalues σ_i^2, $i = 1, 2, \ldots, N$, as well as $\tilde{A}$. For details, see [4, 8, 21].

2.2. The bidiagonalization Lanczos method. Let θ_i, $i = 1, 2, \ldots, m$, be the singular values of B_m, and let w_i and s_i be the corresponding left and right singular vectors. Define

$$\tilde{u}_i = P_{m+1} w_i, \qquad \tilde{v}_i = Q_m s_i.$$

It follows from (2.1) and (2.2) that

$$A\tilde{v}_i = \theta_i \tilde{u}_i, \tag{2.5}$$

$$A^{\mathrm{T}} \tilde{u}_i = \theta_i \tilde{v}_i + \alpha_{m+1} q_{m+1} e_{m+1}^{\mathrm{T}} w_i. \tag{2.6}$$

Therefore, if $\alpha_{m+1} = 0$, then $(\theta_i, \tilde{u}_i, \tilde{v}_i)$, $i = 1, 2, \ldots, m$, are exact singular triplets of A. The bidiagonalization Lanczos method uses the triplets $(\theta_i, \tilde{u}_i, \tilde{v}_i)$ as approximate singular triplets of A. This is the way of achieving the Ritz–Galerkin process on

the Krylov subspaces $\mathcal{K}_m(A^TA, A^Tq_1)$ and $\mathcal{K}_{m+1}(AA^{\mathrm{T}}, q_1)$. So, the triplets $(\theta_i, \tilde{u}_i, \tilde{v}_i)$ are simply called Ritz approximations of singular triplets. Similar to the symmetric Lanczos method, the largest and smallest singular values of B_m converge usually rapidly to the largest and smallest singular values of A [4, 8, 21].

We claim an approximate triplet $(\theta_i, \tilde{u}_i, \tilde{v}_i)$ to have converged if

$$\sqrt{\|A\tilde{v}_i - \theta_i\tilde{u}_i\|^2 + \|A^{\mathrm{T}}\tilde{u}_i - \theta_i\tilde{v}_i\|^2} = \alpha_{m+1} \mid e_{m+1}^T w_i \mid \leq tol, \tag{2.7}$$

where *tol* is a user-prescribed tolerance. Therefore, we do not need to form the Ritz approximations $\tilde{u}_i, \tilde{v}_i$ explicitly until the convergence occurs.

We next show that the method is an orthogonal projection method that projects $\tilde{A}$ onto a suitable subspace. Define the subspace

$$E = span\left\{\begin{pmatrix} P_{m+1} & 0 \\ 0 & Q_m \end{pmatrix}\right\}. \tag{2.8}$$

Then it follows from (2.4), (2.5), and (2.6) that the pairs

$$(\theta_i, \tilde{\varphi}_i) = \left(\theta_i, \frac{1}{\sqrt{2}}\begin{pmatrix}\tilde{u}_i \\ \tilde{v}_i\end{pmatrix}\right), \qquad i = 1, 2, \ldots, m,$$

satisfy the orthogonal projection (Rayleigh–Ritz approximation)

$$\begin{cases} \tilde{\varphi}_i \in E, \\ \tilde{A}\tilde{\varphi}_i - \theta_i\tilde{\varphi}_i \perp E, \end{cases} \tag{2.9}$$

and the projection matrix is $\tilde{B} = \left(\begin{smallmatrix} 0 & B_m \\ B_m^{\mathrm{T}} & 0 \end{smallmatrix}\right)$. The $(\theta_i, \tilde{\varphi}_i)$ are part of the Ritz pairs of $\tilde{A}$ with respect to E.

Jia [11, 15] and Jia and Stewart [18, 19] have proved that, for a general matrix and a general projection subspace, the Ritz vectors may fail to converge. In the context of this paper, note that the spectral condition number of $\tilde{B}$ is always one. Then from Theorem 2.1 of [19], we can get the following simplified result.

THEOREM 2.1. *Define* $\varepsilon = \sin\angle\left(\left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right), E\right)$ *and assume that* ε *is small enough. Then there is a matrix* F *satisfying*

$$||F|| \leq \frac{\varepsilon}{\sqrt{1-\varepsilon^2}}||A|| \tag{2.10}$$

such that σ *is an exact eigenvalue of*

$$\tilde{B}_m + F = \begin{pmatrix} 0 & B_m \\ B_m^{\mathrm{T}} & 0 \end{pmatrix} + F.$$

Furthermore, there exists a positive eigenvalue θ *of* $\tilde{B}_m$ *such that*

$$|\sigma - \theta| \leq ||F||. \tag{2.11}$$

This theorem shows that there is always a Ritz value θ that converges to a desired σ once the deviation ε of $(u^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}}$ from E tends to zero.

Theorem 3.2 in [19] reduces to the following result.

THEOREM 2.2. *Let* $(\theta, \tilde{w}, \tilde{s})$ *be a singular triplet of* B_m*, and let* $(\tilde{w}, \tilde{W}_\perp)$ *and* $(\tilde{s}, \tilde{S}_\perp)$ *be orthogonal matrices such that*

$$\begin{pmatrix} \tilde{w}^{\mathrm{T}} \\ \tilde{W}_\perp^{\mathrm{T}} \end{pmatrix} B_m \left(\tilde{s}, \tilde{S}_\perp\right) = \begin{pmatrix} \theta & 0 \\ 0 & C \end{pmatrix}. \tag{2.12}$$

Define the matrix $\tilde{C} = \left(\begin{smallmatrix} 0 & C \\ C^{\mathrm{T}} & 0 \end{smallmatrix}\right)$, *and assume that* $\sigma I - \tilde{C}$ *is nonsingular. Let the separation of* σ *and the spectra of* C *be defined by*

$$\mathrm{sep}(\sigma, \tilde{C}) = ||(\sigma I - \tilde{C})^{-1}||^{-1}. \tag{2.13}$$

Then if

$$\mathrm{sep}(\sigma, \tilde{C}) \geq \mathrm{sep}(\theta, \tilde{C}) - \mid \theta - \sigma \mid > 0, \tag{2.14}$$

we have

$$\begin{aligned} \sin \angle \left(\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right) &\leq \left(1 + \frac{||A||}{\sqrt{1-\varepsilon^2}\mathrm{sep}(\sigma, \tilde{C})} \right) \varepsilon \\ &\leq \left(1 + \frac{||A||}{\sqrt{1-\varepsilon^2}(\mathrm{sep}(\theta, \tilde{C}) - |\theta - \sigma|)} \right) \varepsilon. \end{aligned} \tag{2.15}$$

Suppose that Algorithm 1 does not break down, i.e., $\alpha_{m+1} \neq 0$. Then B_m only has simple singular values, i.e., θ is different from the singular values of C in (2.12). As a consequence, assumption (2.14) holds with $\varepsilon \to 0$ as $\theta \to \sigma$. However, we must point out that $\mathrm{sep}(\theta, \tilde{C}) - \mid \theta - \sigma \mid$ can be arbitrarily near zero because C may have a singular value that is arbitrarily close to σ, though it is different from σ. Thus, the right-hand side of (2.15) may converge to zero erratically and even may not approach zero although $\varepsilon \to 0$, which means that the Ritz vector $(\tilde{u}^{\mathrm{T}}, \tilde{v}^{\mathrm{T}})^{\mathrm{T}}$ may converge erratically and even may not converge to $(u^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}}$.

Next we establish an inequality on approximate left and right singular vectors $\tilde{u}$ and $\tilde{v}$.

THEOREM 2.3. *We have*

$$\sin^2 \angle(u, \tilde{u}) + \sin^2 \angle(v, \tilde{v}) \leq 2 \sin^2 \angle \left(\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right). \tag{2.16}$$

Proof. By definition, we obtain

$$\begin{aligned} \sin^2 \angle(u, \tilde{u}) + \sin^2 \angle(v, \tilde{v}) &= \min_{\alpha} ||u - \alpha\tilde{u}||^2 + \min_{\alpha} ||v - \alpha\tilde{v}||^2 \\ &\leq \min_{\alpha} (||u - \alpha\tilde{u}||^2 + ||v - \alpha\tilde{v}||^2) \\ &= \min_{\alpha} \left\| \begin{pmatrix} u \\ v \end{pmatrix} - \alpha \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right\|^2 \\ &= 2 \min_{\alpha} \left\| \frac{1}{\sqrt{2}} \begin{pmatrix} u \\ v \end{pmatrix} - \frac{1}{\sqrt{2}} \alpha \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right\|^2 \\ &= 2 \sin^2 \angle \left(\begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \right), \end{aligned}$$

which completes the proof. ∎

Combining Theorems 2.1–2.3, we conclude that under the natural hypothesis that $\varepsilon \to 0$ there is a Ritz value θ that converges to the desired singular value unconditionally, while the corresponding $\tilde{u}$ and $\tilde{v}$ may converge erratically and may even fail to converge to the desired left and right singular vectors u and v.

3. The refined bidiagonalization Lanczos method. As was seen previously, the bidiagonalization Lanczos method may have convergence problems for computing singular vectors. In order to correct this deficiency, we apply the principle of the refined eigenvector approximation advocated by Jia [10, 12] and popularized by Jia [13, 15, 16, 17] (also see [2, 28, 29]) to the bidiagonalization Lanczos method, and we propose a refined bidiagonalization Lanczos method. For $\tilde{A}$, a refined projection method seeks for each θ_i, $i = 1, 2, \ldots, k$, a unit length vector $\tilde{\psi}_i \in E$ satisfying the optimality property

$$||\tilde{A}\tilde{\psi}_i - \theta_i\tilde{\psi}_i|| = \min_{\psi\in E, ||\psi||=1} ||\tilde{A}\psi - \theta_i\psi|| \tag{3.1}$$

and uses them as new approximations to the desired eigenvectors $\frac{1}{\sqrt{2}}(u_i^{\mathrm{T}}, v_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, 2, \ldots, k$. We call $\tilde{\psi}_i$ a refined eigenvector approximation or simply a refined Ritz vector of $\tilde{A}$ with respect to θ_i and the spectral norm. Partition

$$\tilde{\psi}_i = (\tilde{\psi}_{i1}^{\mathrm{T}}, \tilde{\psi}_{i2}^{\mathrm{T}})^{\mathrm{T}}, \tag{3.2}$$

with $\tilde{\psi}_{i1}$ and $\tilde{\psi}_{i2}$ being $m+1$- and m-dimensional, respectively, and take

$$\hat{u}_i = \frac{\tilde{\psi}_{i1}}{\|\tilde{\psi}_{i1}\|}, \qquad \hat{v}_i = \frac{\tilde{\psi}_{i2}}{\|\tilde{\psi}_{i2}\|}. \tag{3.3}$$

Then accordingly, we call the triplet $(\theta, \hat{u}_i, \hat{v}_i)$ a refined Ritz triplet for approximating the singular triplet (σ_i, u_i, v_i) of A.

Based on Theorem 3.2 of Jia [12], we have the following result.

THEOREM 3.1. *Let $z_i = (x_i^{\mathrm{T}}, y_i^{\mathrm{T}})^{\mathrm{T}}$ be the right singular vector of the matrix*

$$\begin{pmatrix} 0 & B_m \\ B_m^{\mathrm{T}} & 0 \\ \alpha_{m+1}e_m^{\mathrm{T}} & 0 \end{pmatrix} - \theta_i \begin{pmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix}$$

associated with its smallest singular value $\sigma_{\min}$, where x_i and y_i are $m+1$- and m-dimensional, respectively. Then

$$\tilde{\psi}_i = \begin{pmatrix} P_{m+1} & 0 \\ 0 & Q_m \end{pmatrix} z_i, \tag{3.4}$$

$$\hat{u}_i = \frac{P_{m+1}x_i}{\|x_i\|}, \qquad \hat{v}_i = \frac{Q_m y_i}{\|y_i\|}, \tag{3.5}$$

$$||\tilde{A}\tilde{\psi}_i - \theta_i\tilde{\psi}_i|| = \sigma_{\min}. \tag{3.6}$$

The computational cost of each z_i is $O(m^3)$ flops. So if k is small, the extra cost of the refined bidiagonalization Lanczos method is very low, compared with the bidiagonalization Lanczos method. So, we can compute the refined approximate singular triplets efficiently and accurately.

Write

$$\hat{x}_i = \frac{x_i}{\|x_i\|}, \qquad \hat{y}_i = \frac{y_i}{\|y_i\|}.$$

Then it follows from (2.1) and (2.2) that

$$\begin{aligned} \|A\hat{v}_i - \theta_i\hat{u}_i\| &= \|AQ_m\hat{y}_i - \theta_i P_{m+1}\hat{x}_i\| \\ &= \|P_{m+1}B_m\hat{y}_i - \theta_i P_{m+1}\hat{x}_i\| \\ &= \|B_m\hat{y}_i - \theta_i\hat{x}_i\| \end{aligned} \tag{3.7}$$

and

$$\|A^{\mathrm{T}}\hat{u}_i - \theta_i\hat{v}_i\| = \sqrt{\|B_m^{\mathrm{T}}\hat{x}_i - \theta_i\hat{y}_i\|^2 + \alpha_{m+1}^2 \mid e_{m+1}^{\mathrm{T}}\hat{x}_i \mid^2}. \tag{3.8}$$

Therefore, we can claim a refined Ritz triplet $(\theta_i, \hat{u}_i, \hat{v}_i)$ to have converged if

$$\sqrt{\|B_m\hat{y}_i - \theta_i\hat{x}_i\|^2 + \|B_m^{\mathrm{T}}\hat{x}_i - \theta_i\hat{y}_i\|^2 + \alpha_{m+1}^2 \mid e_{m+1}^{\mathrm{T}}\hat{x}_i \mid^2} \le tol, \tag{3.9}$$

where *tol* is a user-prescribed tolerance. This important relation means that, similar to the bidiagonalization Lanczos method (cf. (2.7)), we do not need to form the refined Ritz approximations $\hat{u}_i$ and $\hat{v}_i$ explicitly before they converge.

Jia [20] proved that if $||\tilde{A}\tilde{\psi}_i - \theta_i\tilde{\psi}_i|| \neq 0$, i.e., the refined Ritz triplet $(\theta_i, \hat{u}_i, \hat{v}_i)$ is not an exact singular triplet of A, then $\tilde{\psi}_i \neq \tilde{\varphi}_i$, i.e., the refined approximations $\hat{u}_i$ and $\hat{v}_i$ are different from the Ritz approximations $\tilde{u}_i$ and $\tilde{v}_i$. Moreover, if $||\tilde{A}\tilde{\varphi}_i - \theta_i\tilde{\varphi}_i|| \neq 0$, then $||\tilde{A}\tilde{\psi}_i - \theta_i\tilde{\psi}_i|| < ||\tilde{A}\tilde{\varphi}_i - \theta_i\tilde{\varphi}_i||$. Furthermore, if θ_i is very close to one of the other distinct Ritz values θ_j, $j \neq i$, then it may happen that $||\tilde{A}\tilde{\psi}_i - \theta_i\tilde{\psi}_i|| \ll ||\tilde{A}\tilde{\varphi}_i - \theta_i\tilde{\varphi}_i||$. Therefore, $\hat{u}_i$ and $\hat{v}_i$ is more accurate and may be much more accurate than $\tilde{u}_i$ and $\tilde{v}_i$.

Jia and Stewart [18] derived a priori error bounds on the refined Ritz vector. The following result is a direct corollary of Theorem 4.1 of [18].

THEOREM 3.2. *Let (σ, u, v) be a singular triplet of A, and let $(u, U_\perp)$ and $(v, V_\perp)$ be orthogonal matrices such that*

$$\begin{pmatrix} u^{\mathrm{T}} \\ U_\perp^{\mathrm{T}} \end{pmatrix} A(v, V_\perp) = \begin{pmatrix} \sigma & 0 \\ 0 & L \end{pmatrix}, \tag{3.10}$$

where $L = U_\perp^T A V_\perp$. Define $\tilde{L} = \left(\begin{smallmatrix} 0 & L \\ L^{\mathrm{T}} & 0 \end{smallmatrix}\right)$. Assume that $(\theta, \tilde{\psi})$ is the refined Ritz pair approximating $(\sigma, \frac{1}{\sqrt{2}}(u^{\mathrm{T}}, v^{\mathrm{T}})^{\mathrm{T}})$. Then if

$$\mathrm{sep}(\theta, \tilde{L}) \ge \mathrm{sep}(\sigma, \tilde{L}) - |\theta - \sigma| > 0, \tag{3.11}$$

then

$$\sin\angle\left(\tilde{\psi}, \begin{pmatrix} u \\ v \end{pmatrix}\right) \le \frac{||\tilde{A} - \theta I||\varepsilon + |\theta - \sigma|}{\sqrt{1 - \varepsilon^2}(\mathrm{sep}(\sigma, \tilde{L}) - |\theta - \sigma|)}. \tag{3.12}$$

Recall that Theorem 2.1 shows $\theta \to \sigma$ as $\varepsilon \to 0$. Note that $\mathrm{sep}(\sigma, \tilde{L})$ is a positive constant independent of ε, assuming that A has only simple singular values. Therefore, Theorem 3.2 indicates that the refined Ritz approximations $\hat{u}$ and $\hat{v}$ converge to the left and right singular vectors u and v, respectively, as $\varepsilon \to 0$. Generally, they can be expected to be more accurate than the corresponding Ritz approximations $\tilde{u}$ and $\tilde{v}$. Hence the refined bidiagonalization Lanczos method corrects the possible nonconvergence of the standard bidiagonalization Lanczos method.

4. Implicit restart. In practice, due to the limitation of memory and computational complexity, m should not be large. However, for a small m, it is often the case that ε is not small enough, so that it cannot guarantee the convergence of the bidiagonalization Lanczos method and its refined counterpart. Therefore, we usually have to restart the methods in order to compute the desired singular triplets with prescribed accuracy. Over the past decade, the implicit restarting technique due to Sorensen [27] has proven to be a very successful and powerful restarting scheme and has been used either trivially or nontrivially in various contexts. In what follows, we review the technique and show how it is applied to the bidiagonalization Lanczos method and its refined counterpart.

For a general matrix C whose eigenpairs are (λ_i, φ_i), the m-step Arnoldi process [27] is

$$CV_m = V_m H_m + r_m e_m^{\mathrm{T}}. \tag{4.1}$$

Assume that the eigenpairs $(\lambda_i, \varphi_i), i = 1, 2, \ldots, k$, are desired. Given $m-k$ shifts μ_j, $j = 1, 2, \ldots, m-k$, for the $m \times m$ upper Hessenberg matrix H_m, we successively apply QR iterations to the shifted $H_m - \mu_j I$, deriving

$$(H_m - \mu_1 I)(H_m - \mu_2 I) \cdots (H_m - \mu_{m-k}) = QR, \tag{4.2}$$

where Q is orthogonal (unitary) and R is upper triangular. Define $H_m^+ = Q^* H_m Q$, $V_m^+ = V_m Q$, and H_k^+ to be the $k \times k$ leading principal matrix of H_m^+ and V_k^+ the first k columns of V_m^+. Then it holds by the k-step Arnoldi process that

$$CV_k^+ = V_k^+ H_k^+ + r_k^+ e_k^{\mathrm{T}}. \tag{4.3}$$

It has been shown [27] that the new initial vector

$$v_1^+ = p(C) v_1 \tag{4.4}$$

with $p(\lambda) = \alpha \prod_{j=1}^{m-k} (\lambda - \mu_j)$ and α a normalizing factor. Furthermore, it is shown [27] that

$$r_k^+ = 0 \quad \text{if and only if} \quad v_1^+ \in span\{\varphi_1, \varphi_2, \ldots, \varphi_k\}. \tag{4.5}$$

In this case the Arnoldi process breaks down at step k, V_k^+ spans an invariant subspace of C associated with $\lambda_1, \lambda_2, \ldots, \lambda_k$, and the eigenvalues of H_k^+ are just $\lambda_1, \lambda_2, \ldots, \lambda_k$. If r_k^+ is approximately zero, V_k^+ spans an approximate invariant subspace of C, and the eigenvalues of H_k^+ are accepted to have converged to $\lambda_1, \ldots, \lambda_k$.

The implicit restarting technique can be adapted to the bidiagonalization Lanczos process, as was done in [3, 22, 30]. They work in the following way: given the $m-k$ shifts $\mu_1, \ldots, \mu_{m-k}$, the implicit restarting technique leads to

$$\begin{cases} (B_m B_m^{\mathrm{T}} - \mu_1^2 I) \cdots (B_m B_m^{\mathrm{T}} - \mu_{m-k}^2 I) = \tilde{Q} R, \\ \tilde{P}^T B_m \tilde{Q} \quad \text{still (lower) bidiagonal}, \end{cases} \tag{4.6}$$

where $\tilde{P}$ and $\tilde{Q}$ are the accumulation matrices of Givens rotations applied to B_m from the left and right, respectively. Define $P_{m+1}^+ = P_{m+1}\tilde{P}$, $Q_m^+ = Q_m \tilde{Q}$, and $B_m^+ = \tilde{P}^T B_m \tilde{Q}$. The process is achieved implicitly from $B_m B_m^{\mathrm{T}}$ to $B_m^+ (B_m^+)^{\mathrm{T}}$ by working directly on B_m. This is a typical step of the Golub–Kahan SVD algorithm [7]

for a lower bidiagonal B_m. Then by exploiting the special structure of $\tilde{P}$ we obtain by manipulation

$$\tilde{A}\begin{pmatrix} P_{k+1}^+ & 0 \\ 0 & Q_k^+ \end{pmatrix} = \begin{pmatrix} P_{k+1}^+ & 0 \\ 0 & Q_k^+ \end{pmatrix}\begin{pmatrix} 0 & B_k^+ \\ B_k^{+\mathrm{T}} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ (\alpha_{k+1}\tilde{p}_{m+1,k+1}q_{m+1} + \alpha_{k+1}^+ q_{k+1}^+)e_{k+1}^{\mathrm{T}} & 0 \end{pmatrix}, \tag{4.7}$$

with $\tilde{p}_{m+1,k+1}$ being the $(m+1, k+1)$ entry of $\tilde{P}$. Since $\alpha_{k+1}\tilde{p}_{m+1,k+1}q_{m+1}+\alpha_{k+1}^+ q_{k+1}^+$ is orthogonal to Q_k^+, we get a k-step bidiagonalization Lanczos process (Algorithm 1). It is then extended to an m-step bidiagonalization Lanczos process in a standard way. So we avoid restarting Algorithm 1 from scratch and doing it from step $k+1$ upwards. This way saves computational cost of the first k steps of Algorithm 1 by performing a sequence of implicit shift SVD steps on the small B_m at low cost. As a result, we have formally sketched an implicitly restarted bidiagonalization Lanczos algorithm (IRBL) and an implicitly restarted refined bidiagonalization Lanczos algorithm (IRRBL) for computing a partial SVD of a large matrix, which will be detailed later.

5. Selection of shifts. As was seen previously, we can run IRBL and IRRBL once the shifts μ_j, $j = 1, 2, \ldots, m-k$, are given. However, in order to make them work as efficiently as possible, we must select the best possible shifts available for each of them. For an implicitly restarted Krylov subspace algorithm for the eigenproblem, it has been shown [14, 17] that if the shifts are more accurate approximations to some of the unwanted eigenvalues of the original matrix, then the resulting new Krylov subspace will contain more accurate eigenvectors to the desired eigenvectors, so that the algorithm may converge faster. For IRBL and IRRBL, the same conclusion still holds. In an ideal case, similar to Theorem 3 of [17], we can prove the following result.

THEOREM 5.1. *Assume that the sets $\{\sigma_1, \ldots, \sigma_k\}$ and $\{\sigma_{k+1}, \ldots, \sigma_N\}$ are disjoint and A has only simple singular values. Then if $m-k$ distinct ones among σ_j, $j = k+1, \ldots, N$, are selected as shifts at each restart, then IRBL and IRRBL converge after at most $\lceil \frac{N-k}{m-k} \rceil$ restarts.*

Note that p_1^+ can be expressed as

$$\gamma p_1^+ = \prod_{i=1}^{m-k} (AA^{\mathrm{T}} - \mu_i^2 I)p_1, \tag{5.1}$$

with γ being a normalizing factor. Then by a continuity argument of polynomials, it is seen from this theorem and the above relation that the better μ_j approximates an unwanted singular value σ_{j_i} with $j_i > k$, the smaller the component of p_1^+ is in the direction of u_{j_i}, so that $\mathcal{K}_m(A^TA, A^Tp_1)$ and $\mathcal{K}_{m+1}(AA^{\mathrm{T}}, p_1)$ contain more accurate approximations to $v_1, v_2, \ldots, v_k$ and $u_1, u_2, \ldots, u_k$. As a consequence, IRBL and IRRBL usually converges faster.

For the implicitly restarted Arnoldi algorithm (IRA), Sorensen [27] proposed to select those unwanted Ritz values as shifts, called exact shifts. In some sense, this selection scheme is best for the algorithm as the exact shifts are the best approximations available obtained by the algorithm to some unwanted eigenvalues. So, for IRBL, we still use the exact shifts θ_j, $j = k+1, \ldots, m$, as they are the best approximations to some unwanted singular values obtained by IRBL at the current cycle. However, these exact shifts are *not* best for IRRBL as we can find better possible shifts than them based on the refined approximations $\hat{u}_i, \hat{v}_i$, $i = 1, 2, \ldots, k$, as shown below.

Note that the refined Ritz approximations $\hat{u}_i, \hat{v}_i$ are more accurate than the corresponding Ritz approximations $\tilde{u}_i, \tilde{v}_i$, $i = 1, 2, \ldots, k$. This motivates us to seek better possible shifts than θ_j, $j = k+1, \ldots, m$ based on $\hat{u}_i, \hat{v}_i$, $i = 1, 2, \ldots, k$. Let us make the orthogonal direct sum decompositions

$$span\{P_{m+1}\} = span\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k\} \oplus span\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k\}^{\perp} \tag{5.2}$$

$$= span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\} \oplus span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\}^{\perp}, \tag{5.3}$$

$$span\{Q_m\} = span\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\} \oplus span\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}^{\perp} \tag{5.4}$$

$$= span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\} \oplus span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}^{\perp}, \tag{5.5}$$

where $\oplus$ denotes the direct sum. Then clearly

$$span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\}^{\perp} = span\{\tilde{u}_{k+1}, \ldots, \tilde{u}_{m+1}\},$$
$$span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}^{\perp} = span\{\tilde{v}_{k+1}, \ldots, \tilde{v}_m\}.$$

Define

$$\tilde{U}_k = (\tilde{u}_1, \tilde{u}_1, \ldots, \tilde{u}_k), \qquad \tilde{V}_k = (\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k)$$

and

$$\tilde{U}_{m-k} = (\tilde{u}_{k+1}, \ldots, \tilde{u}_{m+1}), \qquad \tilde{V}_{m-k} = (\tilde{v}_{k+1}, \ldots, \tilde{v}_m).$$

Then it is easily justified from the bidiagonalization Lanczos method that the *wanted* Ritz values $\theta_1, \theta_2, \ldots, \theta_k$ are the singular values of A with respect to the left and right subspaces $span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\}$ and $span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}$, that is, they are the singular values of the projection matrix

$$\tilde{U}_k^T A \tilde{V}_k,$$

while on the other hand the *unwanted* Ritz values $\theta_{k+1}, \ldots, \theta_m$ are the singular values of A with respect to the left and right subspaces $span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\}^{\perp}$ and $span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}^{\perp}$, that is, they are the singular values of the projection matrix

$$\tilde{U}_{m-k}^T A \tilde{V}_{m-k}.$$

Keep in mind that $\hat{u}_i, \hat{v}_i$ are generally more accurate than $\tilde{u}_i, \tilde{v}_i$, $i = 1, 2, \ldots, k$, respectively. Then it is clear that $span\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k\}^{\perp}$ and $span\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}^{\perp}$ contain more accurate approximations to the unwanted left and right singular vectors $u_{k+1}, \ldots, u_N$ and $v_{k+1}, \ldots, v_N$ than $span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\}^{\perp}$ and $span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}^{\perp}$, respectively. As a consequence, the Ritz values ξ_i, $i = 1, 2, \ldots, m-k$, of A with respect to the left and right subspaces $span\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k\}^{\perp}$ and $span\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}^{\perp}$ should be generally more accurate approximations to some $m-k$ unwanted singular values than the unwanted Ritz values $\theta_{k+1}, \ldots, \theta_m$ of A with respect to the left and right subspaces $span\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k\}^{\perp}$ and $span\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_k\}^{\perp}$. Therefore, this suggests that we take the ξ_i's as shifts for use within IRRBL. In terms of Jia's terminology [14, 17], they are called the refined shifts. Jia [14, 17] presented very efficient and reliable algorithms to compute the refined shifts for use within the implicitly restarted refined Arnoldi algorithm and the implicitly restarted refined harmonic Arnoldi algorithm, respectively. Adapted from Jia's trick [14], we can propose an efficient algorithm to compute the refined shifts ξ_i's for IRRBL as follows.

Note that $\hat{u}_i = P_{m+1}\hat{x}_i$, $\hat{v}_i = Q_m\hat{y}_i$, $i = 1, 2, \ldots, k$. Define

$$\hat{U}_k = (\hat{u}_1, \ldots, \hat{u}_k) = P_{m+1}(\hat{x}_1, \ldots, \hat{x}_k) = P_{m+1}\hat{X}_k$$

and

$$\hat{V}_k = (\hat{v}_1, \ldots, \hat{v}_k) = Q_m(\hat{y}_1, \ldots, \hat{y}_k) = Q_m\hat{Y}_k.$$

Then we compute the full QR decompositions

$$\hat{X}_k = WR_1, \qquad \hat{Y}_k = SR_2$$

and partition

$$W = (W_k, W_{m-k}), \qquad S = (S_k, S_{m-k}),$$

from which it can be proved, similarly to Jia [14], that

$$span\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k\}^{\perp} = span\{P_{m+1}W_{m-k}\}, \tag{5.6}$$

$$span\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}^{\perp} = span\{Q_mS_{m-k}\}. \tag{5.7}$$

Recall from (2.3) that $P_{m+1}^TAQ_m = B_m$. Then it is known that the projection matrix of A with respect to the left subspace $span\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k\}^{\perp}$ and the right subspace $span\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k\}^{\perp}$ is

$$G = (P_{m+1}W_{m-k})^TA(Q_mS_{m-k}) = W_{m-k}^{\mathrm{T}}(P_{m+1}^TAQ_m)S_{m-k} = W_{m-k}^TB_mS_{m-k},$$

which can be formed at cost of $(m-k)^2m$ flops. So we have exploited the relation $P_{m+1}^TAQ_m = B_m$ to form G, which avoids computing $G = (P_{m+1}W_{m-k})^TA(Q_mS_{m-k})$ directly and reduces the computational cost considerably.

Based on the above arguments and the algorithms for computing the refined shifts [14, 17], we are now able to present the following algorithm.

ALGORITHM 2. THE COMPUTATION OF REFINED SHIFTS ξ_i'S.

1. Form the projection matrix

$$G = W_{m-k}^TB_mS_{m-k}.$$

2. Compute the $m-k$ singular values ξ_j, $j = 1, 2, \ldots, m-k$, of G.
3. Take the ξ_j's as the refined shifts for use within IRRBL.

Thus, starting with the refined Ritz approximations $\hat{u}_i, \hat{v}_i$, $i = 1, 2, \ldots, k$, we can compute the refined shifts ξ_j's using $O(m^3)$ flops, which is negligible compared with one cycle of IRBL.

As Larsen [22] noted, when large close singular values are present, IRBL with exact shifts may have very poor performance and even stagnation. IRRBL inherits the same deficiency. This is explained as follows: By inspecting the relation (5.1), we see the component along the desired kth singular vector u_k is greatly damped if a shift μ_i is very close to σ_k, so that θ_k converges to σ_k very slowly. Since $\mu_1 = \theta_{k+1}$ in the exact shifts and it is an approximation to σ_{k+1}, it is a bad shift when σ_k and σ_{k+1} are close and θ_{k+1} is approximating σ_{k+1}. For this case, the refined shifts have the same deficiency as there is a refined shift that is approaching σ_{k+1}.

To correct this problem, Larsen [22], for IRBL with the exact shifts $\mu_i = \theta_{k+i}$, $i = 1, 2, \ldots, m-k$, proposed the adaptive shifting strategy that required that the relative gaps

$$\mathrm{relgap}_{ki} = \frac{(\theta_k - \epsilon_k) - \mu_i}{\theta_k} \tag{5.8}$$

between the smallest Ritz value θ_k (i.e., the desired kth largest singular value) and all the shifts μ_i, $i = 1, 2, \ldots, m-k$, be larger than some prescribed tolerance, where ϵ_i is the residual norm (2.7). Since $\theta_k - \epsilon_k$ is an approximation to σ_k, relgap_{ki} can be considered to be an approximation of the relative gap of σ_k and the shift μ_i.

However, there is an oversight in (5.8), as relgap_{ki} is only guaranteed to be positive when θ_k is approaching σ_k, i.e., the IRBL is starting to converge. Clearly, if θ_k is still not converging, then ϵ_k is not small. In this case, relgap_{ki} can be negative, so that the strategy cannot work. A simple correction we propose is to replace relgap_{ki} by its absolute value. As in Larsen [22], if

$$| \text{ relgap}_{ki} | \leq 10^{-3},$$

we claim μ_i to be a bad shift and set it to zero. Zero shifts will amplify the component along u_k in p_1^+ and thus overcome the drawback of the exact shifts.

So the combination of the exact shifts and zero shifts will amplify the components along u_i, $i = 1, 2, \ldots, k$, in p_1^+ and at the same time dampen those along the unwanted u_i, $i = k+1, \ldots, N$. It holds to the refined shifts. So we combine the refined shifts with zero shifts for use within IRRBL when computing the largest singular triplets. However, we must point out that the above adaptive strategy works only for computing the largest singular values σ_i, $i = 1, 2, \ldots, k$. It *cannot* be adapted to compute the smallest close singular values of A.

To see why, suppose that we are required to compute the k smallest singular values $\sigma_1 < \sigma_2 < \cdots < \sigma_k$, and we use the k smallest Ritz values θ_i, $i = 1, 2, \ldots, k$, to approximate them. Now the exact shifts are the remaining $m-k$ unwanted large Ritz values $\mu_i = \theta_{k+i}$, $i = 1, 2, \ldots, m-k$, as shifts. Expand q_1 in the left singular basis vectors $\{u_j\}_{j=1}^M$ as

$$p_1 = \sum_{j=1}^{N} \alpha_j u_j + \sum_{j=N+1}^{M} \alpha_j u_j.$$

Then

$$\gamma p_1^+ = \sum_{j=1}^{k} \alpha_j \prod_{i=k+1}^{m} (\sigma_j^2 - \theta_i^2) u_j + \sum_{j=k+1}^{N} \alpha_j \prod_{i=k+1}^{m} (\sigma_j^2 - \theta_i^2) u_j + \sum_{j=N+1}^{M} \alpha_j \prod_{i=k+1}^{m} (-\theta_i^2) u_j.$$

It is clear that if θ_{k+1} is close to σ_k, then the component of p_1^+ in u_k is very small relative to the others. A good cure for this is to replace such a θ_i by the largest Ritz value θ_{m-k}. This way will amplify the components of p_1^+ along u_i, $i = 1, 2, \ldots, k$, and meanwhile possibly dampen those along u_i, $i = k+1, \ldots, N$.

Obviously, the above adaptive shifting strategy can be combined with the refined shifts. The differences are now that ϵ_k is the residual norm (3.9) and bad shifts are replaced by the largest refined shift. Having done the above, we now come to the following practical algorithm.

ALGORITHM 3. IRRBL WITH THE REFINED SHIFTS.

1. Assume a unit length vector p_1 of dimension M and the steps m, the number k of the desired largest or smallest singular triplets (σ_i, u_i, v_i), $i = 1, 2, \ldots, k$, and a user-prescribed tolerance *tol*.
2. Run Algorithm 1 to construct B_m, P_{m+1}, and Q_{m+1}.
3. Compute the singular values θ_i, $j = 1, 2, \ldots, m$, and take the first k ones as approximations to the desired σ_i, $i = 1, \ldots, k$. For each θ_i, $i = 1, 2, \ldots, k$, compute z_i satisfying (3.4).

4. Check if (3.9) for $i = 1, 2, \ldots, k$ is below *tol*. If yes, stop and explicitly compute the refined Ritz approximations $\hat{u}_i = P_{m+1}x_i/\|x_i\|$ and $\hat{v}_i = Q_m y_i/\|y_i\|$, where x_i and y_i are the vectors consisting of the first $m+1$ components and the last m components of z_i, respectively (see (3.5)); otherwise, continue.
5. Use Algorithm 2 to compute the refined shifts ξ_i, $i = 1, 2, \ldots, m-k$.
6. Go to step 2 and implicitly restart combined with the adaptive shifting strategy.

We see it is not necessary to explicitly form the refined Ritz approximations $\hat{u}_i, \hat{v}_i$, $i = 1, 2, \ldots, k$, before the algorithm converges. This way saves some computational work.

6. Numerical experiments. We have tested IRBL, IRRBl, PROPACK, LANSO, and ARPACK, whose MATLAB counterparts are lansvd.m, laneig.m (downloaded from [22]), and eigs.m, respectively. We ran experiments on an Intel Celeron 1700 MHz with main memory 256MB using MATLAB 5.3 with machine precision $\mathbf{u} = 2.22 \times 10^{-16}$. Recall (2.7) and (3.9). The stopping criterion for IRBL and IRRBL is

$$stopcrit = \max_{1\le i\le k} \sqrt{\|A\hat{v}_i - \theta_i\hat{u}_i\|^2 + \|A^{\mathrm{T}}\hat{u}_i - \theta_i\hat{v}_i\|^2}.$$

If

$$stopcrit \le tol \times \max\{\|B_m\|, 1\},$$

then

$$stopcrit \Leftarrow \max_{1\le i\le k} \frac{stopcrit}{||A||_1}.$$

If $stopcrit < tol$, stop.

By taking $m = 2k$ we intend to make all the restarted algorithms as black-box solvers for computing the largest singular values. To make a fair comparison, we used the same starting vector generated randomly in a uniform distribution whenever possible for all the restarted algorithms. In experiments, we took $tol = 10^{-6}$. In all the tables, "iter" denotes the number of restarts, "CPU" the CPU timings in second, and $m > 1000$ denotes no convergence of LANSO or PROPACK when the steps m (i.e., the subspace dimension) exceeded 1000. We terminated LANSO and PROPACK and counted CPU timings when $m > 1000$.

Example 1. We took some test matrices from [1, 6] for our purpose. Keep $\tilde{A} = [0, A; A', 0]$ in mind. IRBL and IRRBL used the same initial vector p_1, eigs($\tilde{A}$) used $(p_1^{\mathrm{T}}, 0)^{\mathrm{T}}$, and eigs($A^TA$) used A^Tp_1 as initial vectors.

From Tables 6.1–6.3, we see that IRRBL works at least as efficiently as IRBL in terms of restarts. For $k = 50$ and well1850, illc1850, and tols4000, it consumed significantly more CPU time than IRBL for some of the test matrices. This is because we had to compute k small SVDs to obtain refined Ritz vectors. In all the other cases, IRRBL was as good as IRBL and could be significantly better than IRBL both in terms of restarts and CPU timings. In particular, for can1054, saylr4, and add32, IRRBL was much faster than IRBL. Both algorithms were significantly better than ARPACK applied to $\tilde{A}$. ARPACK applied to $A^{\mathrm{T}}A$ was faster than IRRBL for five of the eight test matrices but was considerably slower than IRRBL for af23560, saylr4, and add32. However, ARPACK applied to $A^{\mathrm{T}}A$ is not able to compute the left

TABLE 6.1
Computing 10 *largest singular triplets.*

Matrix	well1850		illc1850		tols4000		af23560	
Program	steps	time	steps	time	steps	time	steps	time
lansvd	70	0.47	70	0.53	21	0.16	47	9.00
laneig(A^TA)	75	0.27	75	0.27	155	2.80	51	7.30
laneig($\tilde{A}$)	139	1.45	115	1.14	225	11.3	155	67.7
	iter	time	iter	time	iter	time	iter	time
eigs(A^TA)	18	1.02	11	0.77	22	11.4	8	50.2
eigs($\tilde{A}$)	55	10.13	25	5.84	36	31.5	22	112.1
IRBL	7	2.16	7	2.23	19	16.5	5	23.5
IRRBL	7	2.41	7	2.59	17	15.5	5	23.6
Matrix	can1054		dwt1242		saylr4		add32	
Program	steps	time	steps	time	steps	time	steps	time
lansvd	45	0.27	70	0.41	369	14.1	349	28.4
laneig(A^TA)	67	0.30	115	0.66	401	18.8	371	29.8
laneig($\tilde{A}$)	129	2.23	183	4.84	807	349	531	208
	iter	time	iter	time	iter	time	iter	time
eigs(A^TA)	6	1.41	7	2.30	42	43.4	72	152
eigs($\tilde{A}$)	14	6.34	16	9.67	107	207.3	81	417
IRBL	43	11.3	27	7.98	n.c.	-	79	56.8
IRRBL	8	2.50	15	5.33	48	41.9	44	39.0

TABLE 6.2
Computing 20 *largest singular triplets.*

Matrix	well1850		illc1850		tols4000		af23560	
Program	steps	time	steps	time	steps	time	steps	time
lansvd	150	2.38	141	3.28	41	0.33	83	17.0
laneig(A^TA)	143	0.84	141	1.05	167	3.55	85	14.7
laneig($\tilde{A}$)	139	8.20	279	7.91	303	22.8	173	102.7
	iter	time	iter	time	iter	time	iter	time
eigs(A^TA)	14	2.89	20	4.05	9	21	6	82.9
eigs($\tilde{A}$)	32	27.2	53	38.7	15	59.8	17	299
IRBL	7	7.41	11	13.0	9	27.8	4	64.2
IRRBL	7	10.9	8	13.3	8	28.6	4	66.0
Matrix	can1054		dwt1242		saylr4		add32	
Program	steps	time	steps	time	steps	time	steps	time
lansvd	74	0.66	122	1.19	445	21.5	467	49
laneig(A^TA)	83	0.42	145	1.09	575	43.2	505	63
laneig($\tilde{A}$)	167	4.22	259	12.4	>1000	862	>1000	1844
	iter	time	iter	time	iter	time	iter	time
eigs(A^TA)	4	2.70	8	5.83	29	82.7	72	152
eigs($\tilde{A}$)	10	14.2	18	26.7	77	352	81	417
IRBL	4	3.23	8	8.17	33	93.5	38	116
IRRBL	4	5.48	8	12.5	31	103	12	43

singular vectors simultaneously and is less preferable, as it can lead to severe loss of accuracy of small singular values. LANSO failed in some cases when m exceeded 1000. It could be faster than IRBL and IRRBL in some cases but required (much) more memory to save Lanczos basis vectors for computing singular vectors. LANSO applied to $\tilde{A}$ could be much slower than IRRBL and meanwhile used much more memory. PROPACK was faster than IRRBL in most cases but used much more memory.

TABLE 6.3
Computing 50 *largest singular triplets.*

Matrix	well1850		illc1850		tols4000		af23560	
Program	steps	time	steps	time	steps	time	steps	time
lansvd	422	38.5	315	12.4	101	1.47	154	75.8
laneig(A^TA)	423	13.0	319	4.52	215	6.16	167	52.1
laneig($\tilde{A}$)	847	129	635	47.1	473	59.1	333	14340
	iter	time	iter	time	iter	time	iter	time
eigs(A^TA)	21	29	13	18	2	40	5	302
eigs($\tilde{A}$)	48	236	31	173	9	206	15	1560
IRBL	9	64	6	40	4	67	3	287
IRRBL	9	219	6	142	4	135	3	319
Matrix	can1054		dwt1242		saylr4		add32	
Program	steps	time	steps	time	steps	time	steps	time
lansvd	135	1.69	223	4.25	808	123	505	64.5
laneig(A^TA)	139	1.08	213	2.59	>1000	277	469	51.6
laneig($\tilde{A}$)	281	12.8	423	30.9	>1000	641	>1000	2459
	iter	time	iter	time	iter	time	iter	time
eigs(A^TA)	3	8.69	5	20.8	37	489	19	288
eigs($\tilde{A}$)	8	52.3	12	97.8	81	1680	27	1042
IRBL	2	7.83	4	22.8	1019	38774	13	339
IRRBL	2	43.6	4	92.9	139	6126	7	291

Example 2. We now report some test results for computing a few of the smallest singular triplets by IRBL and IRRBL. In contrast to Example 1, it appears that the computation of smallest singular triplets is much more difficult. It turns out that it is hard to use them as black-box solvers. So we test each case for several m. Since LANSO, PROPACK, and ARPACK exploit shift-and-invert to compute smallest singular triplets, we are not able to compare IRBL and IRRBL with them now and can only give a comparison between IRRBL and IRBL. The test matrices are from [1, 6]. In the tables, "n.c." denotes no convergence after 2000 restarts are used. Tables 6.4–6.13 list the results obtained.

We see that in contrast to Tables 6.1–6.3 it was much more difficult to compute the smallest singular triplets. We could use neither IRBL nor IRRBL as a black-box solver. Performance of IRBL and IRRBL depended heavily on m. However, it is clearly seen from Tables 6.4–6.13 that IRRBL was much more efficient than IRBL, and the latter often failed but the former solved a problem quite successfully.

TABLE 6.4
*well*1850*, computing the* k *smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
10	1077	204	697	138	1351	209	933	153
15	372	152	294	125	347	128	190	76
20	193	138	132	98	161	107	71	49
25	116	129	74	84	91	96	60	66

TABLE 6.5
*dw*2048, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
20	n.c.	-	n.c.	-	n.c.	-	1393	1067
30	n.c.	-	1716	3835	955	1566	667	1140
40	1516	6121	806	3350	493	1449	285	882
50	929	4470	481	2394	301	1433	209	1042

TABLE 6.6
*lshp*2233, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
20	1475	1574	734	808	n.c.	-	1587	1638
30	602	1440	311	756	949	2130	581	1348
40	328	1402	214	933	499	2084	284	1237
50	207	1380	165	1163	309	2197	207	1491

TABLE 6.7
*bcspwr*06, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
15	271	86.3	158	51.1	1179	338	829	250
20	137	72.3	68	36.7	521	251	419	221
25	85	69.8	48	40.7	293	224	192	156

TABLE 6.8
*bcspwr*07, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
10	n.c.	-	1231	199	n.c.	-	n.c.	-
15	709	246	417	149	n.c.	-	1685	548
20	361	211	265	160	1069	557	615	346
25	215	196	124	115	595	538	394	350

TABLE 6.9
*bcspwr*08, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
10	1385	237	691	113	n.c.	-	n.c.	-
15	485	182	287	102	n.c.	-	1691	561
20	245	144	153	93.3	1563	843	1067	619
25	149	137	116	110	885	786	582	547

TABLE 6.10
*bcspwr*09, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
10	1063	183	693	121	n.c.	-	n.c.	-
15	371	139	319	121	1041	347	497	169
20	189	117	165	102	463	262	274	159
25	113	110	83	82	263	236	186	171

TABLE 6.11
*pde*900, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
10	n.c.	-	1431	146	n.c.	-	1827	155
15	913	201	649	143	797	151	398	79.1
20	458	177	311	121	355	138	285	118
25	164	204	199	122	204	124	108	64.3

TABLE 6.12
*jpwh*991, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
15	1371	343	1039	255	n.c.	-	1627	380
20	665	279	421	176	968	397	746	317
25	381	252	284	194	527	322	349	221
30	265	246	193	181	325	280	234	214

TABLE 6.13
*plat*1919, *computing the k smallest singular triplets.*

	$k=3$				$k=5$			
	IRBL		IRRBL		IRBL		IRRBL	
m	iter	time	iter	time	iter	time	iter	time
15	1569	671	785	335	n.c.	-	n.c.	-
20	763	560	307	223	n.c.	-	n.c.	-
25	451	489	244	267	n.c.	-	1526	1642
30	145	179	117	183	n.c.	-	947	1450

7. Conclusion. Both IRRBL and IRBL can be used to compute a partial SVD of a large matrix. But IRRBL is much more efficient than IRBL for computing the smallest singular triplets; in some cases, it can be significantly better than IRBL for computing the largest singular triplets. In comparison with IRBL, it is safer to use IRRBL as a black-box solver for computing the largest singular triplets. For computing the smallest singular triplets, IRBL and IRRBL still cannot perform as black-box solvers, and their performance depends heavily on m. Numerical experiments have demonstrated that (1) the refined Ritz approximations can be much more accurate than the Ritz approximations and (2) the refined shifts can be much better than the exact shifts. For the effect of the refined approximations and the refined shifts on a refined restarted algorithm, see [14, 17] for more analysis.

Note the difficulty of computing the smallest singular triplets. It may be good

to combine IRBL and IRRBL with shift-and-invert. As is well known, however, each step may be very costly and even unacceptable since one has to solve a large linear system each step. Another possibly promising approach to settling the issue is to develop harmonic versions of IRBL and IRRBL, avoiding explicit shift-and-invert, as was done in [17, 24].

Acknowledgments. We thank two referees very much for their very helpful comments and suggestions, which enabled us to improve the presentation and numerical experiments of the paper considerably.

REFERENCES

[1] Z. Bai, R. Barret, D. Day, J. Demmel, and J. Dongarra, *Test Matrix Collection for Non-Hermitian Eigenvalue Problems*, Technical Report CS-97-355, University of Tennessee, Knoxville, 1997. LAPACK Note #123. Data available online from http://math.nist.gov/MarketMatrix.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. A. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, PA, 2000.

[3] A. Björck, E. Grimme, and P. van Dooren, *An implicit shift bidiagonalization algorithm for ill-posed systems*, BIT, 34 (1994), pp. 510–534.

[4] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.

[5] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[6] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[7] G. H. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 205–224.

[8] G. H. Golub, F. T. Luk, and M. L. Overton, *A block Lanczos method for computing the singular values and singular vectors of a matrix*, ACM Trans. Math. Software, 7 (1981), pp. 149–169.

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[10] Z. Jia, *Some Numerical Methods for Large Unsymmetric Eigenproblems*, Ph.D. thesis, Department of Mathematics, Bielefeld University, Bielefeld, Germany, 1994.

[11] Z. Jia, *The convergence of generalized Lanczos methods for large unsymmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 843–862.

[12] Z. Jia, *Refined iterative algorithms based on Arnoldi's process for large unsymmetric eigenproblems*, Linear Algebra Appl., 259 (1997), pp. 1–23.

[13] Z. Jia, *A refined iterative algorithm based on the block Arnoldi process for large unsymmetric eigenproblems*, Linear Algebra Appl., 270 (1998), pp. 170–189.

[14] Z. Jia, *Polynomial characterizations of the approximate eigenvectors by the refined Arnoldi method and an implicitly restarted refined Arnoldi algorithm*, Linear Algebra Appl., 287 (1999), pp. 191–214.

[15] Z. Jia, *Composite orthogonal projection methods for large matrix eigenproblems*, Sci. China Ser. A, 42 (1999), pp. 577–585.

[16] Z. Jia, *A refined subspace iteration algorithm for large sparse eigenproblems*, Appl. Numer. Math., 32 (2000), pp. 35–52.

[17] Z. Jia, *The refined harmonic Arnoldi method and an implicitly restarted refined algorithm for computing interior eigenpairs of large matrices*, Appl. Numer. Appl., 42 (2002), pp. 489–512.

[18] Z. Jia and G. W. Stewart, *An analysis of the Rayleigh-Ritz method for approximating eigenspaces*, Math. Comp., 70 (2001), pp. 637–647.

[19] Z. Jia and G. W. Stewart, *On the Convergence of Ritz Values, Ritz Vectors and Refined Ritz Vectors*, Technical Report TR-3986, Department of Computer Science, University of Maryland, College Park, MD, 1999. Also available online from http://www.cs.umd.edu/~stewart.

[20] Z. Jia, *A Theoretical Comparison of Two Classes of Projection Methods*, technical report, Department of Applied Mathematics, Dalian University of Technology, Dalian, People's Republic of China, 2001.

[21] R. M. Larsen, *Lanczos Bidiagonalization with Partial Reorthogonalization*, technical report, Department of Computer Science, University of Aarhus, Aarhus, Denmark, 1998. Also

available online from http://soi.stanford. edu/˜rmunk/PROPACK.
[22] R. M. Larsen, *Combining Implicit Restarts and Partial Reorthogonalization in Lanczos Bidiagonalization*, available online from http://soi.stanford.edu/˜rmunk/PROPACK.
[23] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, PA, 1998.
[24] R. B. Morgan, *Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1112–1135.
[25] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[26] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998.
[27] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[28] G. W. Stewart, *Matrix Algorithms. Volume* II: *Eigensystems*, SIAM, Philadelphia, PA, 2001.
[29] H. A. van der Vorst, *Computational methods for large eigenvalue problems*, in Handbook of Numerical Analysis, Vol. III, P. G. Ciarlet and J. L. Lions, eds., Elsevier, North-Holland, Amsterdam, 2002, pp. 3–179.
[30] X. Wang and H. Zha, *An Implicitly Restarted Bidiagonal Lanczos Method for Large-Scale Singular Value Problems*, Technical Report 42472, Scientific Computing Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 1998.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 266–284

THE QUADRATIC NUMERICAL RANGE AND THE LOCATION OF ZEROS OF POLYNOMIALS*

HANSJÖRG LINDEN†

Abstract. Containment regions for the zeros of a monic polynomial are given with the aid of the quadratic numerical range of different types of companion matrices of the polynomial.

Key words. quadratic numerical range, companion matrix, zeros of polynomials

AMS subject classifications. 15A60, 26C10, 30C15, 65H05

DOI. 10.1137/S0895479802411651

1. Introduction.

We consider monic polynomials of type

$$P_n(x) := x^n - \alpha_1 x^{n-1} - \cdots - \alpha_n, \qquad \alpha_j \in \mathbb{C},\ j = 1, 2, \ldots, n,\ \alpha_n \neq 0, \tag{1}$$

of degree $n \geq 3$. In this paper we give containment regions for the zeros of P_n with the aid of the quadratic numerical range of some companion matrices of P_n.

The localization of the zeros of a polynomial is a classical problem, which has been considered by many authors (see Marden [17], Milovanović, Mitrinović, and Rassias [22], and McNamee [18, 19, 20, 21]). In many cases the approach to this problem uses matrix analysis (cf. Abdurakhmanov [1], Deutsch [4, 5], Fujii and Kubo [6, 7], Linden [14, 15, 16], Marden [17, pp. 139–146], and Parodi [23, pp. 125–155]). In particular, there were used estimates of the numerical radius of companion matrices of the given polynomial to give bounds for the zeros of the polynomial (cf. Abdurakhmanov [1], Alpin, Chien, and Yeh [2], Fujii and Kubo [7], Linden [15], and Gustafson and Rao [8, p. 122]). Furthermore, Linden [16] has used containment regions for the numerical range for certain matrices to give regions for the zeros of polynomials.

In this paper we extend our earlier work and again use matrix analysis to give regions for the zeros of polynomials. Results on the quadratic numerical range of certain companion matrices of the given polynomial are applied to give containment regions for the zeros of that polynomial. The quadratic numerical range of a matrix (or, more generally, of a linear operator) has been introduced recently by Langer and Tretter [11] as a tool to localize the spectrum of a block operator matrix. In section 2 we give and prove a containment region for the quadratic numerical range in respect to a special decomposition of a matrix of order n, which we subsequently use in section 3 for the localization of the zeros of polynomials.

In section 3 we consider monic polynomials of the type given by (1) and consider two different types of generalized companion matrices of P_n, which we have already used in [14, 15, 16]. Each of these companion matrices can be decomposed in different ways in a block matrix. For some of these decompositions it is possible to derive containment regions for the quadratic numerical range by considering the numerical ranges of the (extended) elements of the block matrix. From this we get containment regions for the zeros of P_n, since these are contained in the quadratic numerical ranges

*Received by the editors July 18, 2002; accepted for publication (in revised form) by H. Woerdeman January 10, 2003; published electronically July 11, 2003.
http://www.siam.org/journals/simax/25-1/41165.html
†Fachbereich Mathematik, Fernuniversität-Gesamthochschule, Postfach 940, Lützowstr. 125, D-58084 Hagen, Germany (hansjoerg.linden@fernuni-hagen.de).

of the companion matrices. Here the result from section 2 is applied first, but also the extension of the method of proof of this result is applied to further decompositions of the companion matrices. These containment regions are new and are better than known ones, as examples show.

2. The quadratic numerical range. First, we recall the basic facts of the quadratic numerical range, which we subsequently use. For details and proofs see [11, 12, 13].

Let H_1, H_2 be Hilbert spaces with inner products $(,.,)_1$,$(.,.)_2$ and norms $\|.\|_1, \|.\|_2$, respectively. In the Hilbert space $H := H_1 \bigoplus H_2$ there is considered the linear operator

$$\mathcal{A} := \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{2}$$

where $A \in B(H_1), D \in B(H_2), B \in B(H_2, H_1)$, and $C \in B(H_1, H_2)$. Here $B(.), B(.,.)$ mean the Banach spaces of all bounded linear operators in the corresponding Hilbert spaces. Denote by Σ the set

$$\Sigma := \{(f,g)^T : f \in H_1, g \in H_2, \|f\|_1 = \|g\|_2 = 1\}.$$

The set

$$\begin{aligned} W_{\mathcal{A}}^2 := \{\ & \lambda \in \mathbb{C} : \lambda^2 - \lambda((Af,f)_1 + (Dg,g)_2) \\ & + (Af,f)_1(Dg,g)_2 - (Bg,f)_1(Cf,g)_2 = 0, (f,g)^T \in \Sigma\} \end{aligned}$$

is called the quadratic numerical range $W_{\mathcal{A}}^2$ of the operator $\mathcal{A}$ with respect to the block operator representation (2). The quadratic numerical range $W_{\mathcal{A}}^2$ is a bounded subset of $\mathbb{C}$; if $\dim H < \infty$, it is also closed. $W_{\mathcal{A}}^2$ is either connected or consists of two components. But in general it is not convex, and even its components do not need to be convex. The numerical range $W(\mathcal{A})$ of $\mathcal{A}$ is defined by

$$W(\mathcal{A}) := \{(\mathcal{A}f, f) : f \in H, \|f\| = 1\},$$

and the numerical radius $w(\mathcal{A})$ of $\mathcal{A}$ is defined by

$$w(\mathcal{A}) := \sup\{|z| : z \in W(\mathcal{A})\};$$

see [8, 9] for details on the numerical range and the numerical radius. We have $W_{\mathcal{A}}^2 \subset W(\mathcal{A})$. Furthermore, $\sigma(\mathcal{A}) \subset \overline{W_{\mathcal{A}}^2} \subset \overline{W(A)}$, where $\sigma(\mathcal{A})$ denotes the spectrum of $\mathcal{A}$.

In the following proposition we give a containment region for the quadratic numerical range of a matrix of order n with respect to a special decomposition of this matrix. We use the following notation: For a complex number $z \neq 0$ the expression $\pm(z)^{1/2}$ means the two square roots of z.

PROPOSITION 1. *Let $A := [\alpha_{i,j}]_{i,j=1}^n$ be a matrix of order n. Let*

$$A := \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix},$$

where A_1 is the matrix of order $n-1$ given by $A_1 := [\alpha_{i,j}]_{i,j=1}^{n-1}$, D_1 is the matrix of order 1 given by $D_1 := [\alpha_{nn}]$, B_1 is the $(n-1) \times 1$ matrix given by $B_1 := [\alpha_{i,n}]_{i=1}^{n-1}$,

and C_1 is the $1 \times (n-1)$ matrix given by $C_1 := [\alpha_{n,j}]_{j=1}^{n-1}$. Let ρ_1, ρ_2 denote the two square roots of

$$\alpha_{nn}^2 + 2\sum_{k=1}^{n-1} \alpha_{nk}\alpha_{kn}.$$

Then the quadratic numerical range W_A^2 of A with respect to this decomposition is contained in the union of the two circles centered at

$$\frac{1}{2}(\alpha_{nn} + \rho_1), \qquad \frac{1}{2}(\alpha_{nn} + \rho_2),$$

with the same radius

$$\frac{1}{2}\left(w(A_1) + \left(2w(A_1)|\alpha_{nn}| + (w(A_1))^2 + 2\left(\sum_{k=1}^{n-1} |\alpha_{nk}|^2 \sum_{k=1}^{n-1} |\alpha_{kn}|^2 \right)^{1/2} \right)^{1/2} \right), \tag{3}$$

respectively.

Proof. The quadratic numerical range W_A^2 of A with respect to this decomposition is the set

$$\begin{aligned} W_A^2 := \Bigg\{ & \lambda \in \mathbb{C} : \lambda^2 - \lambda\left((A_1 f, f) + \alpha_{nn}\right) + \alpha_{nn}(Af, f) \\ & - \sum_{k=1}^{n-1} \alpha_{kn}\overline{f_k} \sum_{k=1}^{n-1} \alpha_{nk} f_k = 0 : f = (f_1, \dots, f_{n-1})^T \in \mathbb{C}^{n-1}, \|f\| = 1 \Bigg\} \\ = \Bigg\{ & \frac{1}{2}\Bigg(\alpha_{nn} + (A_1 f, f) \\ & \pm \left((-\alpha_{nn} + (A_1 f, f)t)^2 + 4\sum_{k=1}^{n-1} \alpha_{kn}\overline{f_k} \sum_{k=1}^{n-1} \alpha_{nk} f_k \right)^{1/2} \Bigg) : \\ & f \in \mathbb{C}^{n-1}, \|f\| = 1 \Bigg\}. \end{aligned}$$

From this representation we derive a containment region for W_A^2.

The set $\left\{(A_1 f, f) : f \in \mathbb{C}^{n-1}, \|f\| = 1\right\}$ is equal to the numerical range of the matrix A_1, and therefore it is contained in the closed circular disk centered at the origin with radius less than or equal to $w(A_1)$. Therefore, the set

$$\left\{-\alpha_{nn} + (A_1 f, f) : f \in \mathbb{C}^{n-1}, \|f\| = 1\right\}$$

is contained in the closed circular disk centered at $-\alpha_{nn}$ and radius less than or equal to $w(A_1)$. It follows that the set

$$\left\{(-\alpha_{nn} + (A_1 f, f))^2 : f \in \mathbb{C}^{n-1}, \|f\| = 1\right\}$$

is contained in the closed circular disk centered at α_{nn}^2 with radius

$$2w(A_1)|\alpha_{nn}| + (w(A_1))^2.$$

Now, the set

$$\left\{ 4 \sum_{k=1}^{n-1} \alpha_{kn} \overline{f_k} \sum_{k=1}^{n-1} \alpha_{nk} f_k : f \in \mathbb{C}^{n-1}, \|f\| = 1 \right\}$$

is equal to the numerical range of the matrix of order $n-1$ given by $4B_1C_1$ of rank 1, and thus it is a closed elliptical disk with center at

$$2 \sum_{k=1}^{n-1} \alpha_{nk} \alpha_{kn}$$

and main axis

$$4 \left(\sum_{k=1}^{n-1} |\alpha_{nk}|^2 \sum_{k=1}^{n-1} |\alpha_{kn}|^2 \right)^{1/2}.$$

This elliptical disk is contained in the closed circular disk centered at

$$2 \sum_{k=1}^{n-1} \alpha_{nk} \alpha_{kn}$$

with radius

$$2 \left(\sum_{k=1}^{n-1} |\alpha_{nk}|^2 \sum_{k=1}^{n-1} |\alpha_{kn}|^2 \right)^{1/2}.$$

Therefore, the radicand in the expression above for W_A^2 is contained in the closed circular disk centered at

$$\alpha_{nn}^2 + 2 \sum_{k=1}^{n-1} \alpha_{nk} \alpha_{kn}$$

with radius

$$2w(A_1)|\alpha_{nn}| + (w(A_1))^2 + 2 \left(\sum_{k=1}^{n-1} |\alpha_{nk}|^2 \sum_{k=1}^{n-1} |\alpha_{kn}|^2 \right)^{1/2}.$$

Now we have to take the square root: We get two sets R_1 and R_2. R_1 is the closed circular disk centered at ρ_1 and radius

$$\left(2w(A_1)|\alpha_{nn}| + (w(A_1))^2 + 2 \left(\sum_{k=1}^{n-1} |\alpha_{nk}|^2 \sum_{k=1}^{n-1} |\alpha_{kn}|^2 \right)^{1/2} \right)^{1/2},$$

and R_2 is the closed circular disk centered at ρ_2 with the same radius. $R_1 \cup R_2$ contains the set of the two square roots of the radicand. Now the assertion follows from the representation of W_A^2. □

Remark 1. (a) An analogous result can be proved for a containment region of the quadratic numerical range W_A^2 of A with respect to the decomposition

$$A := \begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix},$$

where A_n is the square matrix of order 1 given by $A_n := [\alpha_{11}]$, D_n is the square matrix of order $n-1$ given by $D_n := [\alpha_{i,j}]_{i,j=2}^{n}$, B_n is the $1 \times (n-1)$ matrix given by $B_n := [\alpha_{1,j}]_{j=2}^{n}$, and C_n is the $(n-1) \times 1$ matrix given by $C_n := [\alpha_{i,1}]_{i=2}^{n}$.

(b) Of course, we can try to extend the technique of the proof of Proposition 1 to decompositions of A with the order of D_1 greater than 1, but the resulting formulas seem to be rather complicated. Thus, in section 3 we will apply such a procedure only to certain companion matrices, which is much easier.

3. Regions for the zeros of polynomials. In this section we apply the properties of the quadratic numerical range as described in section 2 (in particular Proposition 1) to different types of companion matrices of the monic polynomial P_n given by (1). From this containment, regions for the zeros of P_n are derived. In the following theorem we use companion matrices of P_n, which come from a diagonal similarity of the usual Frobenius companion matrix. Let P_n be as given by (1). We suppose that there exist complex numbers $\gamma_1, \dots, \gamma_n \in \mathbb{C}, 0 \neq \beta_1, \dots, \beta_{n-1} \in \mathbb{C}$ such that

$$\begin{aligned} &\alpha_1 := \gamma_1, \\ &\alpha_2 := \gamma_2\beta_1, \\ &\vdots \\ &\alpha_n := \gamma_n\beta_{n-1}\cdots\beta_1. \end{aligned} \tag{4}$$

Furthermore, let

$$\delta_j^{(n-1)} := \min\left\{ \cos\frac{\pi}{n-j+1} \max_{k=j+1,\dots,n-1} |\beta_k|, \right. \\ \left. \frac{1}{2} \max_{k=j+1,\dots,n-2} (|\beta_k| + |\beta_{k+1}|) \right\}, \qquad j = 1, \dots, n-3.$$

If $\beta_1 = \cdots = \beta_{n-1} =: \beta$, then

$$\delta_j^{(n-1)} = |\beta| \cos\frac{\pi}{n-j+1}, \qquad j = 1, \dots, n-3.$$

Decompositions of type (4) of the coefficients of P_n are always possible. Refer to [14, 16] for discussions of useful decompositions.

THEOREM 2. *Let P_n be as given by* (1)*, and let its coefficients satisfy* (4)*. Let ρ_1, ρ_2 denote the two square roots of $\alpha_1^2 + 2\alpha_2$. Then all zeros of P_n lie in the union of the two circles centered at*

$$\frac{1}{2}(\alpha_1 + \rho_1), \qquad \frac{1}{2}(\alpha_1 + \rho_2),$$

with the same radius

$$\frac{1}{2}\left(\delta_1^{(n-1)} + \left(2\delta_1^{(n-1)}|\alpha_1| + \left(\delta_1^{(n-1)}\right)^2 + 2|\beta_1| \left(\sum_{k=2}^{n} |\gamma_k|^2 \right)^{1/2} \right)^{1/2} \right), \tag{5}$$

respectively.

Proof. Let $A_c := (\tau_{i,j})_{i,j=1}^{n}$ be the matrix of order n given by

$$\tau_{i,j} := \begin{cases} \gamma_{n-j+1}, & i = n, j = 1, \dots, n, \\ \beta_{n-i}, & i = 1, \dots, n-1, j = i+1, \\ 0 & \text{otherwise;} \end{cases} \tag{6}$$

then $\det(xE_n - A_c) = P_n(x)$. That is, A_c is a companion matrix of P_n.

We decompose

$$A_c := \begin{bmatrix} A_{c1} & B_{c1} \\ C_{c1} & D_{c1} \end{bmatrix},$$

where A_{c1} is the matrix of order $n-1$ given by $A_{c1} := [\tau_{i,j}]_{i,j=1}^{n-1}$, D_{c1} is the matrix of order 1 given by $D_{c1} := [\gamma_1]$, B_{c1} is the $(n-1)\times 1$ matrix given by $B_{c1} := [\tau_{i,n}]_{i=1}^{n-1}$, and C_{c1} is the $1\times(n-1)$ matrix given by $C_{c1} := [\tau_{n,j}]_{j=1}^{n-1}$. We apply Proposition 1 to this decomposition of A_c. Then the assertion follows immediately since the numerical range of the matrix A_{c1} is a closed circular disk centered at the origin with radius less than or equal to $\delta_1^{(n-1)}$ (Theorem 3 in [3]). □

Remark 2. (a) For the case $n = 3$ the numerical range of A_{c1} is exactly the closed disk centered at the origin with radius $|\beta_2|/2$. Therefore, in this case the radius given by (5) is

$$\frac{1}{2}\left(\frac{1}{2}|\beta_2| + \left(|\beta_2\alpha_1| + \frac{1}{4}|\beta_2|^2 + 2|\beta_1|\left(|\gamma_2|^2 + |\gamma_3|^2\right)^{1/2}\right)^{1/2}\right).$$

(b) For the case $n = 4$ the numerical range of A_{c1} is exactly the closed disk centered at the origin with radius $(|\beta_2|^2 + |\beta_3|^2)^{1/2}/2$ (see [10]). Therefore, in this case the radius given by (5) is

$$\frac{1}{2}\left(\frac{1}{2}\left(|\beta_2|^2 + |\beta_3|^2\right)^{1/2} + \left(|\alpha_1|\left(|\beta_2|^2 + |\beta_3|^2\right)^{1/2} + \frac{1}{4}(|\beta_2|^2 + |\beta_3|^2) + 2|\beta_1|\left(|\gamma_2|^2 + |\gamma_3|^2 + |\gamma_4|^2\right)^{1/2}\right)^{1/2}\right).$$

(c) If $\beta_1 = \cdots = \beta_{n-1} = \beta$, then the radius given by (5) is

$$\frac{1}{2}\left(|\beta|\cos\frac{\pi}{n} + \left(2|\alpha_1||\beta|\cos\frac{\pi}{n} + |\beta|^2\cos^2\frac{\pi}{n} + 2|\beta|\left(\sum_{k=2}^{n}\frac{|\alpha_k|^2}{|\beta|^{2(k-1)}}\right)^{1/2}\right)^{1/2}\right).$$

(d) If $\beta_1 = \cdots = \beta_{n-1} = 1$, and $\alpha_1 = \alpha_2 = 0$, then the region given by Theorem 2 is properly contained in the regions given by Corollary 2(a),(b) in [16], respectively; in this case all regions are circular disks centered at the origin.

There is a further decomposition of the companion matrix used in the proof of Theorem 2 that can be handled in an analogous way and gives another region for the zeros of P_n.

THEOREM 3. *Let P_n be as given by* (1)*, and let its coefficients satisfy* (4)*. Then all zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2}\left(\widetilde{\delta}_2 + \left(\widetilde{\delta}_2|\alpha_1| + \widetilde{\delta}_2^2 + |\beta_2\gamma_3| + |\beta_2|\left(\sum_{k=3}^{n}|\gamma_k|^2\right)^{1/2}\right)^{1/2}\right),$$

respectively, where

$$\widetilde{\delta}_2 = \delta_2^{(n-1)} + \frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2 + 2\alpha_2\right| + \frac{1}{2}|\alpha_1|^2 + |\gamma_2|^2 + |\beta_1|^2\right)^{1/2}. \tag{7}$$

Proof. We decompose

$$A_c := \begin{bmatrix} A_{c2} & B_{c2} \\ C_{c2} & D_{c2} \end{bmatrix},$$

where A_{c2} is the matrix of order $n-2$ given by $A_{c2} := [\tau_{i,j}]_{i,j=1}^{n-2}$, D_{c2} is the matrix of order 2 given by $D_{c2} := [\tau_{i,j}]_{i,j=n-1}^{n}$, B_{c2} is the $(n-2)\times 2$ matrix given by $B_{c2} := [\tau_{i,j}]_{\substack{i=1,\dots,n-2\\ j=n-1,n}}$, and C_{c2} is the $2\times(n-2)$ matrix given by $C_{c2} := [\tau_{i,j}]_{\substack{i=n-1,n\\ j=1,\dots,n-2}}$. The quadratic numerical range $W_{A_c}^2$ of A_c with respect to this decomposition is then the set

$$\begin{aligned} W_{A_c}^2 := \Bigg\{ & \frac{1}{2}\Bigg(\sum_{k=1}^{n-3} \beta_{n-k} f_{k+1}\overline{f_k} + \beta_1 g_2\overline{g_1} + \gamma_2 g_1\overline{g_2} + \gamma_1|g_2|^2 \\ & \pm \Bigg(\Bigg(- \sum_{k=1}^{n-2} \beta_{n-k} f_{k+1}\overline{f_k} + \beta_1 g_2\overline{g_1} + \gamma_2 g_1\overline{g_2} + \gamma_1|g_2|^2 \Bigg)^2 \\ & + \ 4\beta_2 g_1\overline{g_2}\,\overline{f_{n-2}} \sum_{k=1}^{n-2} \gamma_{n-k+1} f_k \Bigg)^{1/2} \Bigg) : \\ & = (f_1,\dots,f_{n-2})^T \in \mathbb{C}^{n-2}, g = (g_1,g_2)^T \in \mathbb{C}^2, \|f\| = 1 = \|g\| \Bigg\}. \end{aligned}$$

From this representation we derive a containment region for $W_{A_c}^2$ and thus get the desired containment region for the zeros of P_n.

The set

$$\left\{ \sum_{k=1}^{n-3} \beta_{n-k} f_{k+1}\overline{f_k} : f \in \mathbb{C}^{n-2}, \|f\| = 1 \right\}$$

is equal to the numerical range of the matrix A_{c2} and therefore is a closed circular disk centered at the origin with radius less than or equal to $\delta_2^{(n-1)}$ (Theorem 3 in [3]). Furthermore, the set

$$\left\{ \beta_1 g_2\overline{g_1} + \gamma_2 g_1\overline{g_2} + \gamma_1|g_2|^2 : g \in \mathbb{C}^2, \|g\| = 1 \right\}$$

is the numerical range of the matrix of order 2,

$$\begin{bmatrix} 0 & \beta_1 \\ \gamma_2 & \gamma_1 \end{bmatrix},$$

and therefore is a closed elliptical disk with foci

$$\frac{1}{2}\alpha_1 \pm \left(\frac{1}{4}\alpha_1 + \alpha_2\right)^{1/2}$$

and center at $\alpha_1/2$, which is contained in the closed circular disk centered at $\alpha_1/2$ with radius

$$\frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2+2\alpha_2\right|+\frac{1}{2}|\alpha_1|^2+|\gamma_2|^2+|\beta_1|^2\right)^{1/2}$$

(this is half of the main axis of the ellipse). It follows that the set

$$\left\{-\sum_{k=1}^{n-2}\beta_{n-k}f_{k+1}\overline{f_k}+\beta_1 g_2\overline{g_1}+\gamma_2 g_1\overline{g_2}+\gamma_1|g_2|^2 : f\in\mathbb{C}^{n-2}, g\in\mathbb{C}^2, \|f\|=1=\|g\|\right\} \tag{8}$$

is contained in the closed circular disk centered at $\alpha_1/2$ with radius $\widetilde{\delta}_2$. Hence, the set

$$\left\{\left(-\sum_{k=1}^{n-2}\beta_{n-k}f_{k+1}\overline{f_k}+\beta_1 g_2\overline{g_1}+\gamma_2 g_1\overline{g_2}+\gamma_1|g_2|^2\right)^2 : f\in\mathbb{C}^{n-2}, g\in\mathbb{C}^2, \|f\|=1=\|g\|\right\}$$

is contained in the closed circular disk centered at $\alpha_1^2/4$ with radius $\widetilde{\delta}_2|\alpha_1|+\widetilde{\delta}_2^2$. Now, the set

$$\left\{4\beta_2 g_1\overline{g_2} : g\in\mathbb{C}^2, \|g\|=1\right\}$$

is equal to the numerical range of the matrix of order 2 given by

$$\begin{bmatrix} 0 & 0 \\ 4\beta_2 & 0 \end{bmatrix},$$

and thus it is a closed circular disk with center at 0 and radius $2|\beta_2|$. Further, the set

$$\left\{\overline{f_{n-2}}\sum_{k=1}^{n-2}\gamma_{n-k+1}f_k : f\in\mathbb{C}^{n-2}, \|f\|=1\right\}$$

is equal to the numerical range of the matrix of order $n-2$ given by

$$\begin{bmatrix} O \\ C_{c2} \end{bmatrix},$$

and thus (see [16]) it is a closed elliptical disk with center at $\gamma_3/2$ and half of the main axis is

$$\frac{1}{2}\left(\sum_{k=3}^{n}|\gamma_k|^2\right)^{1/2}.$$

This elliptical disk is contained in the closed circular disk centered at $\gamma_3/2$ with radius

$$\frac{1}{2}\left(\sum_{k=3}^{n}|\gamma_k|^2\right)^{1/2}.$$

Thus, the set

$$\left\{4\beta_2 g_1 \overline{g_2}\,\overline{f_{n-2}} \sum_{k=1}^{n-2} \gamma_{n-k+1} f_k : f \in \mathbb{C}^{n-2}, g \in \mathbb{C}^2, \|f\| = 1 = \|g\|\right\}$$

is contained in the closed disk centered at 0 with radius

$$|\beta_2\gamma_3| + |\beta_2| \left(\sum_{k=3}^{n} |\gamma_k|^2\right)^{1/2}.$$

Therefore, the radicand in the above expression for $W_{A_c}^2$ is contained in the closed circular disk centered at $\alpha_1^2/4$ with radius

$$\widetilde{\delta}_2|\alpha_1| + \widetilde{\delta}_2^2 + |\beta_2\gamma_3| + |\beta_2| \left(\sum_{k=3}^{n} |\gamma_k|^2\right)^{1/2}.$$

Now we have to take the square root: We get two sets S_1 and S_2. S_1 is the closed circular disk centered at $\alpha_1/2$ and radius

$$\left(\widetilde{\delta}_2|\alpha_1| + \widetilde{\delta}_2^2 + |\beta_2\gamma_3| + |\beta_2| \left(\sum_{k=3}^{n} |\gamma_k|^2\right)^{1/2}\right)^{1/2},$$

and S_2 is the closed circular disk centered at $-\alpha_1/2$ with the same radius. The set of both square roots of the radicand is contained in $S_1 \cup S_2$. Now the assertion follows from the representation of $W_{A_c}^2$ with (8). □

Remark 3. (a) For the case $n = 3$ the numerical range of A_{c2} is exactly the origin. Therefore, in this case $\delta_2^{(n-1)}$ in (7) can be replaced by 0.

(b) For the case $n = 4$ the numerical range of A_{c2} is exactly the closed disk centered at the origin with radius $|\beta_3|/2$. Therefore, in this case $\delta_2^{(n-1)}$ in (7) can be replaced by $|\beta_3|/2$.

(c) For the case $n = 5$ the numerical range of A_{c2} is exactly the closed disk centered at the origin with radius $(|\beta_3|^2 + |\beta_4|^2)^{1/2}/2$ (see [10]). Therefore, in this case $\delta_2^{(n-1)}$ in (7) can be replaced by this expression.

The procedure can be extended to some further decompositions of the companion matrix A_c under other special assumptions. If the order of the diagonal lower block in the decomposition of A_c is enlarged, then in general the numerical range of this matrix cannot be determined except in some special cases, which we will consider now.

THEOREM 4. *Let P_n be as given by* (1)*, and let its coefficients satisfy* (4)*. Furthermore, let $1 < m < n$ be such that $\alpha_1, \ldots, \alpha_m = 0$ and $\alpha_{m+1} \neq 0$. Then all zeros of P_n lie in the closed circle centered at* 0 *with radius*

$$\frac{1}{2}\left(\delta_m^{(n-1)} + \delta_1^{(m-1)} + \left(\left(\delta_m^{(n-1)} + \delta_1^{(m-1)}\right)^2 + |\beta_m\gamma_{m+1}| + |\beta_m| \left(\sum_{k=m+1}^{n} |\gamma_k|^2\right)^{1/2}\right)^{1/2}\right).$$

Proof. We decompose

$$A_c := \begin{bmatrix} A_{cm} & B_{cm} \\ C_{cm} & D_{cm} \end{bmatrix},$$

where A_{cm} is the matrix of order $n-m$ given by $A_{cm} := [\tau_{i,j}]_{i,j=1}^{n-m}$, D_{cm} is the matrix of order m given by $D_{cm} := [\tau_{i,j}]_{i,j=n-m+1}^{n}$, B_{cm} is the $(n-m)\times m$ matrix given by $B_{cm} := [\tau_{i,j}]_{\substack{i=1,\dots,n-m \\ j=n-m+1,\dots,n}}$, and C_{cm} is the $m\times(n-m)$ matrix given by $C_{cm} := [\tau_{i,j}]_{\substack{i=n-m+1,\dots,n \\ j=1,\dots,n-m}}$. The quadratic numerical range $W^2_{A_c}$ of A_c with respect to this decomposition then is the set

$$\begin{aligned} W^2_{A_c} := \Bigg\{ \frac{1}{2}\Bigg(& \sum_{k=1}^{n-m-1} \beta_{n-k} f_{k+1}\overline{f_k} + \sum_{k=1}^{m-1} \beta_{m-k} g_{k+1}\overline{g_k} \\ & \pm \Bigg(\Bigg(\sum_{k=1}^{n-m-1} \beta_{n-k} f_{k+1}\overline{f_k} - \sum_{k=1}^{m-1} \beta_{m-k} g_{k+1}\overline{g_k} \Bigg)^2 \\ & + \ 4\beta_m \overline{f_{n-m}} g_1 \overline{g_m} \sum_{k=1}^{n-m} \gamma_{n-k+1} f_k \Bigg) \Bigg)^{1/2} : \\ & f\in\mathbb{C}^{n-m}, g\in\mathbb{C}^m, \|f\| = 1 = \|g\| \Bigg\}. \end{aligned}$$

From this representation we derive a containment region for $W^2_{A_c}$, and thus we get the desired containment region for the zeros of P_n in a way analogous to the proofs of Theorems 2 and 3. ∎

Remark 4. For some special cases the radius in Theorem 4 can be described more precisely in a way analogous to Remark 1(a), (b) and Remark 3; we omit the details.

(a) For the case $n=4, m=2$, the radius in Theorem 4 can be replaced by

$$\frac{1}{2}\left(\frac{1}{2}\left(|\beta_1| + |\beta_3|\right) + \left(\frac{1}{4}\left(|\beta_1| + |\beta_3|\right)^2 + |\beta_2\gamma_3| + |\beta_2| \left(|\gamma_3|^2 + |\gamma_4|^2\right)^{1/2} \right)^{1/2} \right).$$

(b) For the case $n=5, m=2$, the radius in Theorem 4 can be replaced by

$$\begin{aligned} \frac{1}{2}\Bigg(& \frac{1}{2}\left(|\beta_1| + \left(|\beta_3|^2 + |\beta_4|^2\right)^{1/2} \right) + \Bigg(\frac{1}{4}\left(|\beta_1| + \left(|\beta_3|^2 + |\beta_4|^2\right)^{1/2} \right)^2 \\ & + \ |\beta_2\gamma_3| + |\beta_2| \left(|\gamma_3|^2 + |\gamma_4|^2 + |\gamma_5|^2\right)^{1/2} \Bigg)^{1/2} \Bigg). \end{aligned}$$

(c) For the case $n=5, m=3$, the radius in Theorem 4 can be replaced by

$$\begin{aligned} & \frac{1}{2}\Bigg(\frac{1}{2}\left(|\beta_4| + \left(|\beta_1|^2 + |\beta_2|^2\right)^{1/2} \right) \\ & + \ \left(\frac{1}{4}\left(|\beta_4| + \left(|\beta_1|^2 + |\beta_2|^2\right)^{1/2} \right)^2 + |\beta_3\gamma_4| + |\beta_3| \left(|\gamma_4|^2 + |\gamma_5|^2\right)^{1/2} \right)^{1/2} \Bigg). \end{aligned}$$

(d) For the case $n = 6, m = 3$, the radius in Theorem 4 can be replaced by

$$\frac{1}{2}\Bigg(\frac{1}{2}\left(\left(|\beta_4|^2+|\beta_5|^2\right)^{1/2}+\left(|\beta_1|^2+|\beta_2|^2\right)^{1/2}\right) + \left(\frac{1}{4}\left(\left(|\beta_4|^2+|\beta_5|^2\right)^{1/2}+\left(|\beta_1|^2+|\beta_2|^2\right)^{1/2}\right)^2 + |\beta_3\gamma_4|+|\beta_3|\left(|\gamma_4|^2+|\gamma_5|^2+|\gamma_6|^2\right)^{1/2}\right)^{1/2}\Bigg).$$

(e) For the case $n \geq 7, m = n-3$, in the expression for the radius in Theorem 4 the term $\delta_m^{(n-1)}$ can be replaced by $\left(|\beta_{n-1}|^2+|\beta_n|^2\right)^{1/2}$. For the case $n \geq 7, m = 3$, in the expression for the radius in Theorem 4 the term $\delta_1^{(m-1)}$ can be replaced by $\left(|\beta_1|^2+|\beta_2|^2\right)^{1/2}$.

Another type of generalized companion matrix, not coming from a diagonal similarity of the Frobenius companion matrix, gives more regions for the zeros of P_n with the aid of the quadratic numerical range. Let P_n be as given by (1). We suppose that there exist complex numbers $\alpha_1^{(1)}, \alpha_2^{(1)}, \alpha_2^{(2)}, \ldots, \alpha_n^{(1)}, \alpha_n^{(2)}, \ldots, \alpha_n^{(n)} \in \mathbb{C}$ such that

$$\begin{aligned}\alpha_1 &:= \alpha_1^{(1)},\\ \alpha_2 &:= \alpha_2^{(1)}\alpha_2^{(2)},\\ &\vdots\\ \alpha_n &:= \alpha_n^{(1)}\alpha_n^{(2)}\cdots\alpha_n^{(n)}.\end{aligned} \tag{9}$$

If any of the coefficients α_k is equal to 0, we choose $\alpha_k^{(1)} = \cdots = \alpha_k^{(k)} = 0$. Furthermore, let

$$\widehat{\beta}_1 := \max\left\{\frac{1}{2}\left|\alpha_3^{(2)}\right|, \max_{k=4,\ldots,n}\min\left\{\cos\frac{\pi}{k}\max_{j=2,\ldots,k-1}\left|\alpha_k^{(j)}\right|, \frac{1}{2}\max_{j=2,\ldots,k-2}\left(\left|\alpha_k^{(j)}\right|+\left|\alpha_k^{(j+1)}\right|\right)\right\}\right\},$$

$$\widehat{\beta}_2 := \max_{k=4,\ldots,n}\min\left\{\cos\frac{\pi}{k}\max_{j=2,\ldots,k-1}\left|\alpha_k^{(j)}\right|, \frac{1}{2}\max_{j=2,\ldots,k-2}\left(\left|\alpha_k^{(j)}\right|+\left|\alpha_k^{(j+1)}\right|\right)\right\}.$$

Decompositions of type (9) of the coefficients of P_n are always possible. Refer to [14, 16] for a discussion of useful decompositions.

THEOREM 5. *Let P_n be as given by* (1)*, and let its coefficients satisfy* (9)*. Let ρ_1, ρ_2 denote the two square roots of $\alpha_1^2 + 2\alpha_2$. Then all zeros of P_n lie in the union of the two circles centered at*

$$\frac{1}{2}(\alpha_1+\rho_1), \qquad \frac{1}{2}(\alpha_1+\rho_2),$$

with the same radius

$$\frac{1}{2}\left(\widehat{\beta}_1 + \left(2\widehat{\beta}_1|\alpha_1| + \left(\widehat{\beta}_1\right)^2 + 2\left(\sum_{k=2}^{n}\left|\alpha_k^{(1)}\right|^2\sum_{k=2}^{n}\left|\alpha_k^{(k)}\right|^2\right)^{1/2}\right)^{1/2}\right), \tag{10}$$

respectively.

Proof. Let $\widehat{n} := 1 + n(n-1)/2$, and let $\widehat{A}_c := [\widehat{\tau}_{i,j}]_{i,j=1}^{\widehat{n}}$ be the matrix of order $\widehat{n}$ with

$$\widehat{\tau}_{i,j} := \begin{cases} \alpha_\mu^{(1)}, & i = \widehat{n}, j = \widehat{n} - \frac{1}{2}\mu(\mu-1), 1 \le \mu \le n, \\ \alpha_\mu^{(\mu)}, & i = \widehat{n} - 1 - \frac{1}{2}(\mu-1)(\mu-2), j = \widehat{n}, 2 \le \mu \le n, \\ \alpha_\nu^{(\mu)}, & i = 3 - \mu + \frac{1}{2}\nu(\nu-1), j = i-1, 2 \le \mu < \nu \le n, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\det\left(xE_{\widehat{n}} - \widehat{A}_c\right) = x^{(n-1)(n-2)/2} P_n(x),$$

where $E_{\widehat{n}}$ is the identity matrix of order $\widehat{n}$. Thus $\widehat{A}_c$ is a generalized companion matrix of P_n. We decompose

$$\widehat{A}_c := \begin{bmatrix} \widehat{A}_{c1} & \widehat{B}_{c1} \\ \widehat{C}_{c1} & \widehat{D}_{c1} \end{bmatrix},$$

where $\widehat{A}_{c1} := [\widehat{\tau}_{i,j}]_{i,j=1}^{\widehat{n}-1}$, $\widehat{D}_{c1}$ is the matrix of order 1 given by $\widehat{D}_{c1} := [\alpha_1^{(1)}]$, $\widehat{B}_{c1}$ is the $(\widehat{n}-1) \times 1$ matrix given by $\widehat{B}_{c1} :=]\widehat{\tau}_{i,\widehat{n}}]_{i=1}^{\widehat{n}-1}$, and $\widehat{C}_{c1}$ is the $1 \times (\widehat{n}-1)$ matrix given by $\widehat{C}_{c1} := [\widehat{\tau}_{\widehat{n},j}]_{j=1}^{\widehat{n}-1}$. We apply Proposition 1 to this decomposition of $\widehat{A}_c$. Then the assertion follows immediately since the numerical range of the matrix $\widehat{A}_{c1}$ is a closed circular disk centered at the origin with radius less than or equal to $\widehat{\beta}_1$ (Theorem 3 in [3]). ∎

The analogue of Theorem 4 is a special case of Theorem 5. However, the method from Theorem 3 gives the following result.

THEOREM 6. *Let P_n be as given by* (1)*, and let its coefficients satisfy* (9)*. Then all zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2}\left(\widetilde{\beta}_2 + \left(\widetilde{\beta}_2|\alpha_1| + \widetilde{\beta}_2^2 + 2\left(\sum_{k=3}^{n}\left|\alpha_k^{(1)}\right|^2 \sum_{k=3}^{n}\left|\alpha_k^{(k)}\right|^2\right)^{1/2}\right)^{1/2}\right),$$

respectively, where

$$\widetilde{\beta}_2 = \widehat{\beta}_1 + \frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2 + 2\alpha_2\right| + \frac{1}{2}|\alpha_1|^2 + \left|\alpha_2^{(1)}\right|^2 + \left|\alpha_2^{(2)}\right|^2\right)^{1/2}.$$

Proof. We decompose

$$\widehat{A}_c := \begin{bmatrix} \widehat{A}_{c2} & \widehat{B}_{c2} \\ \widehat{C}_{c2} & \widehat{D}_{c2} \end{bmatrix},$$

where $\widehat{A}_{c2} := [\widehat{\tau}_{i,j}]_{i,j=1}^{\widehat{n}-2}$, $\widehat{D}_{c2}$ is the matrix of order 2 given by

$$\widehat{D}_{c2} := \begin{bmatrix} 0 & \alpha_2^{(2)} \\ \alpha_2^{(1)} & \alpha_1^{(1)} \end{bmatrix},$$

$\widehat{B}_{c2}$ is the $(\widehat{n}-2)\times 2$ matrix given by $\widehat{B}_{c2} := [\widehat{\tau}_{i,j}]_{\substack{i=1,\dots,\widehat{n}-2,\\ j=\widehat{n}-1,\widehat{n}}}$, and $\widehat{C}_{c2}$ is the $2\times(\widehat{n}-2)$ matrix given by $\widehat{C}_{c2} := [\widehat{\tau}_{i,j}]_{\substack{i=\widehat{n}-1,\widehat{n}\\ j=1,\dots,\widehat{n}-2}}$. The quadratic numerical range $W^2_{\widehat{A}_c}$ of $\widehat{A}_c$ with respect to this composition then is the set

$$\begin{aligned} W^2_{\widehat{A}_c} := \Bigg\{ \frac{1}{2} \Bigg(& \alpha_2^{(1)} g_1 \overline{g_2} + \alpha_2^{(2)} g_2 \overline{g_1} + \alpha_1^{(1)} |g_2|^2 + \sum_{k=3}^{n} \sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)} \overline{f_k^{(j)}} \\ & \pm \Bigg(\Bigg(\alpha_2^{(1)} g_1 \overline{g_2} + \alpha_2^{(2)} g_2 \overline{g_1} + \alpha_1^{(1)} |g_2|^2 - \sum_{k=3}^{n} \sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)} \overline{f_k^{(j)}} \Bigg)^2 \\ & + 4|g_2|^2 \sum_{k=3}^{n} \alpha_k^{(1)} f_k^{(1)} \sum_{k=3}^{n} \alpha_k^{(k)} \overline{f_k^{(k-1)}} \Bigg)^{1/2} \Bigg) : \\ & f = (f_n^{(1)}, \dots, f_n^{(n-1)}, \dots, f_3^{(2)})^T \in \mathbb{C}^{\widehat{n}-2}, \\ & g = (g_1, g_2)^T \in \mathbb{C}^2, \|f\| = 1 = \|g\| \Bigg\}. \end{aligned}$$

From this representation we derive a containment region for $W^2_{\widehat{A}_c}$, and thus we get the desired containment region for the zeros of P_n. However, in this case only the set

$$\left\{ 4|g_2|^2 \sum_{k=3}^{n} \alpha_k^{(1)} f_k^{(1)} \sum_{k=3}^{n} \alpha_k^{(k)} \overline{f_k^{(k-1)}} : f \in \mathbb{C}^{\widehat{n}-2}, g \in \mathbb{C}^2, \|f\| = 1 = \|g\| \right\}$$

has to be considered in more detail, because all other sets were already considered in the proof of Theorem 3. Now the set $\left\{ |g_2|^2 : g \in \mathbb{C}^2, 1 = \|g\| \right\}$ is the numerical range of the matrix $\left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, and thus it is the line segment from 0 to 1. Further, the set

$$\left\{ 4 \sum_{k=3}^{n} \alpha_k^{(1)} f_k^{(1)} \sum_{k=3}^{n} \alpha_k^{(k)} \overline{f_k^{(k-1)}} : f \in \mathbb{C}^{\widehat{n}-2}, \|f\| = 1 \right\}$$

is the numerical range of the matrix $4B_{c2}C_{c2}$, which is a matrix of rank 1. The numerical range of this matrix is a closed circular disk centered at the origin with radius

$$2 \left(\sum_{k=3}^{n} \left| \alpha_k^{(1)} \right|^2 \sum_{k=3}^{n} \left| \alpha_k^{(k)} \right|^2 \right)^{1/2}.$$

Taking all these sets together, the assertion follows. ∎

For the companion matrix used in the proofs of Theorems 5 and 6 a further decomposition is possible, which gives the following theorem.

THEOREM 7. *Let P_n be as given by* (1)*, and let its coefficients satisfy* (9)*. Then all zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2} \left(\widetilde{\beta}_3 + \left(\widetilde{\beta}_3 |\alpha_1| + \widetilde{\beta}_3^2 + 2 \left(\sum_{k=3}^{n} \left| \alpha_k^{(1)} \right|^2 \right)^{1/2} \left(\left(\sum_{k=4}^{n} \left| \alpha_k^{(k)} \right|^2 \right)^{1/2} + \left| \alpha_3^{(2)} \right| \right) \right)^{1/2} \right),$$

respectively, where

$$\widetilde{\beta}_3 = \widehat{\beta}_2 + \frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2 + 2\alpha_2\right| + \frac{1}{2}|\alpha_1|^2 + \left|\alpha_2^{(1)}\right|^2 + \left|\alpha_2^{(2)}\right|^2 + \left|\alpha_3^{(3)}\right|^2\right)^{1/2}.$$

Proof. We decompose

$$\widehat{A}_c := \begin{bmatrix} \widehat{A}_{c3} & \widehat{B}_{c3} \\ \widehat{C}_{c3} & \widehat{D}_{c3} \end{bmatrix},$$

where $\widehat{A}_{c3} := [\widehat{\tau}_{i,j}]_{i,j=1}^{\widehat{n}-3}$, $\widehat{D}_{c3}$ is the matrix of order 3 given by

$$\widehat{D}_{c3} := \begin{bmatrix} 0 & 0 & \alpha_3^{(3)} \\ 0 & 0 & \alpha_2^{(2)} \\ 0 & \alpha_2^{(1)} & \alpha_1^{(1)} \end{bmatrix},$$

$\widehat{B}_{c3}$ is the $(\widehat{n}-3)\times 3$ matrix given by $\widehat{B}_{c3} := [\widehat{\tau}_{i,j}]_{\substack{i=1,\dots,\widehat{n}-3 \\ j=\widehat{n}-2,\widehat{n}-1,\widehat{n}}}$, and $\widehat{C}_{c3}$ is the $3\times(\widehat{n}-3)$ matrix given by $\widehat{C}_{c3} := [\widehat{\tau}_{i,j}]_{\substack{i=\widehat{n}-2,\widehat{n}-1,\widehat{n} \\ j=1,\dots,\widehat{n}-3}}$. The quadratic numerical range $W^2_{\widehat{A}_c}$ of $\widehat{A}_c$ with respect to this decomposition is then the set

$$\begin{aligned} W^2_{\widehat{A}_c} := \Bigg\{ & \frac{1}{2}\Bigg(\alpha_3^{(3)} g_3\overline{g_1} + \alpha_2^{(1)} g_2\overline{g_3} + \alpha_2^{(2)} g_3\overline{g_2} + \alpha_1^{(1)}|g_3|^2 \\ & + \sum_{k=4}^{n}\sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)}\overline{f_k^{(j)}} \pm \Bigg(\Bigg(\alpha_3^{(3)} g_3\overline{g_1} + \alpha_2^{(1)} g_2\overline{g_3} + \alpha_2^{(2)} g_3\overline{g_2} \\ & + \alpha_1^{(1)}|g_3|^2 - \sum_{k=4}^{n}\sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)}\overline{f_k^{(j)}} \Bigg)^2 \\ & + 4|g_3|^2 \sum_{k=3}^{n} \alpha_k^{(1)} f_k^{(1)} \sum_{k=4}^{n} \alpha_k^{(k)}\overline{f_k^{(k-1)}} + 4\alpha_3^{(2)} g_1\overline{g_3} \sum_{k=3}^{n} \alpha_k^{(1)} f_k^{(1)}\overline{f_3^{(1)}} \Bigg)^{1/2} \Bigg) : \\ & f = (f_n^{(1)}, \dots, f_n^{(n-1)}, \dots, f_3^{(1)})^T \in \mathbb{C}^{\widehat{n}-3}, \\ & g = (g_1, g_2, g_3)^T \in \mathbb{C}^3, \|f\| = 1 = \|g\| \Bigg\}. \end{aligned}$$

From this representation we derive a containment region for $W^2_{\widehat{A}_c}$, and thus we get the desired containment region for the zeros of P_n. The set

$$\left\{ \alpha_3^{(3)} g_3\overline{g_1} + \alpha_2^{(1)} g_2\overline{g_3} + \alpha_2^{(2)} g_3\overline{g_2} + \alpha_1^{(1)}|g_3|^2 : g \in \mathbb{C}^3, 1 = \|g\| \right\}$$

is the numerical range of the matrix $\widehat{D}_{c3}$, and is therefore contained in the closed circular disk centered at $\frac{1}{2}\alpha_1$ with radius

$$\frac{1}{2}\left(2\left|\frac{1}{4}\alpha_1^2 + \alpha_2\right| + \frac{1}{2}|\alpha_1|^2 + \left|\alpha_2^{(1)}\right|^2 + \left|\alpha_2^{(2)}\right|^2 + \left|\alpha_3^{(3)}\right|^2\right)^{1/2}.$$

The set

$$\left\{ \sum_{k=4}^{n} \sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)} \overline{f_k^{(j)}} : 1 = \|f\| \right\}$$

is the numerical range of the matrix $\widehat{A}_{c3}$, and is therefore contained in the closed circular disk centered at the origin with radius $\widehat{\beta}_2$. Therefore, the set

$$\left\{ \alpha_3^{(3)} g_3 \overline{g_1} + \alpha_2^{(1)} g_2 \overline{g_3} + \alpha_2^{(2)} g_3 \overline{g_2} + \alpha_1^{(1)} |g_3|^2 + \sum_{k=4}^{n} \sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)} \overline{f_k^{(j)}} : \right.$$

$$\left. f \in \mathbb{C}^{\widehat{n}-3}, g \in \mathbb{C}^3, \|f\| = 1 = \|g\| \right\}$$

is contained in the closed circular disk centered at $\frac{1}{2}\alpha_1$ with radius

$$\widehat{\beta}_2 + \frac{1}{2} \left(\left| \frac{1}{2}\alpha_1^2 + 2\alpha_2 \right| + \frac{1}{2}|\alpha_1|^2 + \left|\alpha_2^{(1)}\right|^2 + \left|\alpha_2^{(2)}\right|^2 + \left|\alpha_3^{(3)}\right|^2 \right)^{1/2}.$$

Hence the set

$$\left\{ \left(\alpha_3^{(3)} g_3 \overline{g_1} + \alpha_2^{(1)} g_2 \overline{g_3} + \alpha_2^{(2)} g_3 \overline{g_2} + \alpha_1^{(1)} |g_3|^2 - \sum_{k=4}^{n} \sum_{j=1}^{k-2} \alpha_k^{(j+1)} f_k^{(j+1)} \overline{f_k^{(j)}} \right)^2 : \right.$$

$$\left. f \in \mathbb{C}^{\widehat{n}-3}, g \in \mathbb{C}^3, \|f\| = 1 = \|g\| \right\}$$

is contained in the closed circular disk centered at $\frac{1}{4}\alpha_1^2$ with radius $\widehat{\beta}|\alpha_1| + \widehat{\beta}^2$.

The set $\{|g_3|^2 : g \in \mathbb{C}^3, \|g\| = 1\}$ is the numerical range of the matrix

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and is therefore the line segment from 0 to 1. The set

$$\left\{ 4 \sum_{k=3}^{n} \alpha_k^{(1)} f_k^{(1)} \sum_{k=4}^{n} \alpha_k^{(k)} \overline{f_k^{(k-1)}} : f \in \mathbb{C}^{\widehat{n}-3}, \|f\| = 1 \right\}$$

is the numerical range of the matrix $4\widehat{B}_{c3}\widehat{C}_{c3}$, which is a matrix of rank 1. The numerical range of this matrix is a closed circular disk centered at the origin with radius

$$2 \left(\sum_{k=3}^{n} \left|\alpha_k^{(1)}\right|^2 \sum_{k=4}^{n} \left|\alpha_k^{(k)}\right|^2 \right)^{1/2}.$$

The set $\{4\alpha_3^{(2)} g_1 \overline{g_3} : g \in \mathbb{C}^3, \|g\| = 1\}$ is the numerical range of the matrix

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 4\alpha_3^{(2)} & 0 & 0 \end{bmatrix},$$

and is therefore a closed circular disk centered at the origin with radius $2|\alpha_3^{(2)}|$. The set

$$\left\{\sum_{k=3}^{n}\alpha_k^{(1)}f_k^{(1)}\overline{f_3^{(1)}} : f\in\mathbb{C}^{\widehat{n}-3}, \|f\|=1\right\}$$

is the numerical range of the square matrix of order $\widehat{n}-3$ given by

$$\begin{bmatrix} O \\ \widehat{C}_{c3} \end{bmatrix},$$

and thus (see [16]) it is the closed circular disk with center at 0 and radius

$$\frac{1}{2}\left(\sum_{k=3}^{n}\left|\alpha_k^{(1)}\right|^2\right)^{1/2}.$$

Combining the results the assertion follows. □

In the following two corollaries, we consider two special cases of the decompositions of the coefficients of P_n.

COROLLARY 8. *Let P_n be as given by* (1).

(a) *All zeros of P_n lie in the union of the two closed circles centered at*

$$\frac{1}{2}(\alpha_1+\rho_1), \qquad \frac{1}{2}(\alpha_1+\rho_2),$$

with the same radius

$$\frac{1}{2}\left(\cos\frac{\pi}{n}+\left(2|\alpha_1|\cos\frac{\pi}{n}+\cos^2\frac{\pi}{n}+2\sum_{k=2}^{n}|\alpha_k|\right)^{1/2}\right),$$

respectively.

(b) *All zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2}\left(\widetilde{\beta}_2+\left(\widetilde{\beta}_2|\alpha_1|+\widetilde{\beta}_2^2+2\sum_{k=3}^{n}|\alpha_k|\right)^{1/2}\right),$$

respectively, where

$$\widetilde{\beta}_2=\cos\frac{\pi}{n}+\frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2+2\alpha_2\right|+\frac{1}{2}|\alpha_1|^2+2\,|\alpha_2|\right)^{1/2}.$$

(c) *All zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2}\left(\widetilde{\beta}_3+\left(\widetilde{\beta}_3|\alpha_1|+\widetilde{\beta}_3^2+2\left(\sum_{k=3}^{n}|\alpha_k|\right)^{1/2}\left(\left(\sum_{k=4}^{n}|\alpha_k|\right)^{1/2}+1\right)\right)^{1/2}\right),$$

respectively, where

$$\widetilde{\beta}_3=\cos\frac{\pi}{n}+\frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2+2\alpha_2\right|+\frac{1}{2}|\alpha_1|^2+2\,|\alpha_2|+|\alpha_3|\right)^{1/2}.$$

Proof. In Theorems 5–7 we make the special choice $\alpha_k^{(1)} = \alpha_k^{(k)}$ to be equal to the same square root of $\alpha_k, k = 2, \dots, n$, and all other $\alpha_k^{(j)}$ to be equal to 1. □

The region from Remark 1(c) with $\beta = 1$ is contained in the region from Corollary 8(a).

COROLLARY 9. *Let P_n be as given by* (1).

(a) *All zeros of P_n lie in the union of the two closed circles centered at*

$$\frac{1}{2}(\alpha_1 + \rho_1), \qquad \frac{1}{2}(\alpha_1 + \rho_2),$$

with the same radius

$$\frac{1}{2}\left(\widehat{\beta}_1 + \left(2\widehat{\beta}_1|\alpha_1| + \widehat{\beta}_1^2 + 2\sum_{k=2}^{n}|\alpha_k|^{2/k}\right)^{1/2}\right),$$

respectively, where

$$\widehat{\beta}_1 = \max_{k=3,\dots,n}|\alpha_k|^{1/k}\cos\frac{\pi}{k}.$$

(b) *All zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2}\left(\widetilde{\beta}_2 + \left(\widetilde{\beta}_2|\alpha_1| + \widetilde{\beta}_2^2 + 2\sum_{k=3}^{n}|\alpha_k|^{2/k}\right)^{1/2}\right),$$

respectively, where

$$\widetilde{\beta}_2 = \max_{k=3,\dots,n}|\alpha_k|^{1/k}\cos\frac{\pi}{k} + \frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2 + 2\alpha_2\right| + \frac{1}{2}|\alpha_1|^2 + 2|\alpha_2|\right)^{1/2}.$$

(c) *All zeros of P_n lie in the union of the two closed circles centered at* 0 *and* $\alpha_1/2$ *with the same radius*

$$\frac{1}{2}\left(\widetilde{\beta}_3 + \left(\widetilde{\beta}_3|\alpha_1| + \widetilde{\beta}_3^2 + 2\left(\sum_{k=3}^{n}|\alpha_k|^{2/k}\right)^{1/2}\left(\left(\sum_{k=4}^{n}|\alpha_k|^{2/k}\right)^{1/2} + |\alpha_3|^{1/3}\right)\right)^{1/2}\right),$$

respectively, where

$$\widetilde{\beta}_3 = \max_{k=4,\dots,n}|\alpha_k|^{1/k}\cos\frac{\pi}{k} + \frac{1}{2}\left(\left|\frac{1}{2}\alpha_1^2 + 2\alpha_2\right| + \frac{1}{2}|\alpha_1|^2 + 2|\alpha_2| + |\alpha_3|^{2/3}\right)^{1/2}.$$

Proof. In Theorems 5–7 we make the special choice $\alpha_k^{(j)}, j = 1, \dots, k$, to be equal to the same kth root of α_k for $k = 2, \dots, n$. □

We consider some examples from [4, 5], where we compare some of our regions with the bounds from [4, 5]. For simplicity in the following examples let us denote $K(\alpha, \rho) := \{z \in \mathbb{C} : |z - \alpha| \le \rho\}$.

Example 1. Let $P_3(x) := x^3 - 0.5x^2 - 2x - 2$. Then all zeros of P_3 lie in $K(-0.78, 1.52) \cup K(1.28, 1.52)$ (Remark 1(a) with $\beta_1 = \beta_2 = 1$) and $K(0, 2.05) \cup$

$K(0.25, 2.05)$ (Corollary 9(c)), whereas the bound given in [4] is $K(0, 2.19)$. The exact zeros of P_3 are $2, (-3 \pm i\sqrt{7})/4$.

Example 2. Let $P_4(x) := x^4 - 4x^3 + 3x^2 + 2x - 1$. Then all zeros of P_4 lie in $K(0.42, 2.20) \cup K(3.58, 2.20)$ (Remark 1(a) with $\beta_1 = \beta_2 = 1$) and $K(0, 3.75) \cup K(2, 3.75)$ (Remark 3(b) with $\beta_1 = \beta_2 = \beta_3 = 1$), whereas the bound given in [4] is $K(0, 4.75)$. The exact zeros of P_4 are $(1 \pm \sqrt{5})/2, (3 \pm \sqrt{5})/2$.

Example 3. Let $P_4(x) := x^4 + 2x^3 - 13x^2 - 38x - 24$. Then all zeros of P_4 lie in $K(-3.74, 4.96) \cup K(1.74, 4.96)$ (Corollary 9(a)), whereas the bound given in [4] is $K(0, 6.14)$. The exact zeros of P_4 are $-3, -2, -1, 4$.

Example 4. Let $P_3(x) := x^3 - 8x - 3$. Then all zeros of P_3 lie in $K(-2, 2.35) \cup K(2, 2.35)$ (Remark 1(a) with $\beta_1 = \beta_2 = 1$), whereas the bound given in [5] is $K(0, 3.03)$. The exact zeros of P_3 are $-(3 \pm \sqrt{5})/2, 3$.

Example 5. Let $P_3(x) := x^3 - 3x - 18$. Then all zeros of P_3 lie in $K(-1.23, 2.40) \cup K(1.23, 2.40)$ (Remark 1(a) with $\beta_1 = \beta_2 = 1$) and $K(0, 3.23)$ (Corollary 9(c)), whereas the bound given in [5] is $K(0, 4.17)$. The exact zeros of P_3 are $-(3 \pm i\sqrt{15})/2, 3$.

Acknowledgments. The author thanks the anonymous referees for their helpful remarks.

REFERENCES

[1] A. A. Abdurakhmanov, *Geometry of a Hausdorff domain in problems of localization for the spectrum of arbitrary matrices*, Math. USSR-Sb., 59 (1988), pp. 39–51.

[2] Y. A. Alpin, M.-T. Chien, and L. Yeh, *The numerical radius and bounds for zeros of a polynomial*, Proc. Amer. Math. Soc., 131 (2003), pp. 725–730.

[3] M.-T. Chien, *On the numerical range of tridiagonal operators*, Linear Algebra Appl., 246 (1996), pp. 203–214.

[4] E. Deutsch, *Matricial norms and the zeros of polynomials*, Linear Algebra Appl., 3 (1970), pp. 483–489.

[5] E. Deutsch, *Matricial norms and the zeros of lacunary polynomials*, Linear Algebra Appl., 6 (1973), pp. 143–148.

[6] M. Fujii and F. Kubo, *Operator norms as bounds for roots of algebraic equations*, Proc. Japan Acad., 49 (1973), pp. 805–808.

[7] M. Fujii and F. Kubo, *Buzano's inequality and bounds for roots of algebraic equations*, Proc. Amer. Math. Soc., 117 (1993), pp. 359–361.

[8] K. E. Gustafson and K. M. Rao, *Numerical Range*, Springer, New York, 1997.

[9] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[10] D. S. Keeler, L. Rodman, and I. M. Spitkovsky, *The numerical range of* 3×3 *matrices*, Linear Algebra Appl., 252 (1997), pp. 115–139.

[11] H. Langer and C. Tretter, *Spectral decomposition of some nonselfadjoint block operator matrices*, J. Oper. Theory, 39 (1998), pp. 1–20.

[12] H. Langer, A. S. Markus, and C. Tretter, *Corners of numerical ranges*, Oper. Theory Adv. Appl., 124 (2001), pp. 385–400.

[13] H. Langer, A. S. Markus, V. Matsaev, and C. Tretter, *A new concept for block operator matrices: The quadratic numerical range*, Linear Algebra Appl., 330 (2001), pp. 89–112.

[14] H. Linden, *Bounds for the zeros of polynomials from eigenvalues and singular values of some companion matrices*, Linear Algebra Appl., 271 (1998), pp. 41–82.

[15] H. Linden, *Numerical radii of some companion matrices and bounds for the zeros of polynomials*, in Analytic and Geometric Inequalities and Applications, T. M. Rassias and H. M. Srivastava, eds., Kluwer, Dordrecht, 1999, pp. 205–229.

[16] H. Linden, *Containment regions for zeros of polynomials from numerical ranges of companion matrices*, Linear Algebra Appl., 350 (2002), pp. 125–145.

[17] M. Marden, *Geometry of Polynomials*, 2nd ed., Math. Surveys 3, AMS, Providence, RI, 1966.

[18] J. M. McNamee, *A bibliography on roots of polynomials*, J. Comput. Appl. Math., 47 (1993), pp. 391–394.

[19] J. M. McNamee, *A supplementary bibliography on roots of polynomials*, J. Comput. Appl. Math., 78 (1997), p. 1.
[20] J. M. McNamee, *An updated supplementary bibliography on roots of polynomials*, J. Comput. Appl. Math., 110 (1999), pp. 305–306.
[21] J. M. McNamee, *A* 2002 *updated supplementary bibliography on roots of polynomials*, J. Comput. Appl. Math., 142 (2002), pp. 433–434.
[22] G. V. Milovanović, D. S. Mitrinović, and Th. M. Rassias, *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*, World Scientific, Singapore, 1994.
[23] M. Parodi, *La Localisation des Valeurs Caractéristiques des Matrices et ses Applications*, Gauthier-Villars, Paris, 1959.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 285–294

MATRIX CLASSES THAT GENERATE ALL MATRICES WITH POSITIVE DETERMINANT*

C. R. JOHNSON†, D. D. OLESKY‡, AND P. VAN DEN DRIESSCHE§

Abstract. New factorization results dealing mainly with P-matrices and M-matrices are presented. It is proved that any matrix in $M_n(\mathbb{R})$ with positive determinant can be written as the product of three P-matrices (compared with the classical result that five positive definite matrices may be needed). It is also proved that a matrix A with positive determinant can be stabilized via multiplication by a P-matrix if and only if A is not a diagonal matrix with all diagonal entries negative. Factorization into two P-matrices is considered and characterized for $n = 2$. Using elementary bidiagonal factorization results, it is shown that the nonsingular M-matrices, or the nonsingular totally nonnegative matrices, generate all matrices in $M_n(\mathbb{R})$ with positive determinant. Further results on products of M-matrices and inverse M-matrices are given.

Key words. factorization, M-matrix, nested sequence of principal minors, P-matrix, positive determinant

AMS subject classification. 15A23

DOI. 10.1137/10.1137/S0895479802418446

1. Introduction.

Factorization of matrices into products of simple matrices or matrices of special type is fundamental to the theoretical development of matrix analysis and its applications, including numerical computation. We present here several new factorization results dealing with important classes of matrices. In the process several surprising facts are noted.

A set $\mathcal{U}$ of nonsingular matrices (multiplicatively) *generates* another set of matrices $\mathcal{F}$ if every $A \in \mathcal{F}$ is a finite product of matrices from $\mathcal{U}$ and matrices whose inverses are in $\mathcal{U}$. (Typically, we also mean that every such product lies in $\mathcal{F}$, so that $\mathcal{F}$ is a semigroup, but this is not essential.) A classical example is that the set $\mathcal{PD}$ of real positive definite matrices, which happens to be closed under inversion, generates the group of real matrices with positive determinant [2, 11]. Furthermore, Ballentine [2] (see also [11]) showed that (independent of dimension) at most five positive definite factors are needed to represent any matrix of positive determinant, and that if the matrix to be represented is not a negative scalar matrix, then at most four factors are needed. A related, and more classical, factorization result is that a matrix is a product of two positive definite matrices if and only if it has positive eigenvalues and is diagonalizable; see, for example [7, Thm. 7.6.3, p. 465]. Additional factorization results of this sort are given in [4].

Here, we consider several other possible generating sets for $\mathcal{U}$ and also give some mixed factor factorization results. One such set is the set $\mathcal{P}$ of P-matrices (all principal minors are positive), which is also closed under inversion. Since $\mathcal{PD} \subsetneq \mathcal{P}$, the above

*Received by the editors November 22, 2002; accepted for publication (in revised form) by R. Bhatia February 4, 2003; published electronically July 11, 2003.
http://www.siam.org/journals/simax/25-1/41844.html

†Department of Mathematics, College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (crjohnso@math.wm.edu).

‡Department of Computer Science, University of Victoria, Victoria, B.C., V8W 3P6 Canada (dolesky@cs.uvic.ca). The research of this author was partially supported by the Natural Sciences and Engineering Research Council of Canada.

§Department of Mathematics and Statistics, University of Victoria, Victoria, B.C., V8W 3P4 Canada (pvdd@math.uvic.ca). The research of this author was partially supported by the Natural Sciences and Engineering Research Council of Canada.

cited results imply that $\mathcal{P}$ generates all matrices with positive determinant. The maximum number of factors needed is at most five, and we answer the question of whether this number (the *generating number*) may be reduced for P-matrices (a larger class). Other inverse closed classes considered are the positive stable matrices $\mathcal{S}$ and the special subset $\mathcal{S}^+$, in which all eigenvalues are positive (real) and distinct. Mixed factor results are given involving these classes as well as $\mathcal{PD}$.

We also consider for $\mathcal{U}$ two well-known noninverse closed sets: the nonsingular M-matrices $\mathcal{M}$ (see, for example, [5]) and the nonsingular, totally nonnegative matrices $\mathcal{TN}$ (see, for example, [1]). The former are the P-matrices with nonpositive off-diagonal entries, and the latter are those with *all* minors nonnegative. The inverses of M-matrices $\mathcal{M}^{-1}$ are entrywise nonnegative and those of $\mathcal{TN}$ matrices have a checkerboard sign pattern. Throughout, we consider matrices in $M_n(\mathbb{R})$, $n \geq 2$, and we call a diagonal matrix with positive (negative) diagonal entries a positive (negative) diagonal matrix.

2. *P*-matrix factorizations. It is clear that the set of matrices generated from $\mathcal{P}$ must have positive determinant. Thus, we focus on the worst-case number of factors from $\mathcal{P}$ needed to represent a matrix with positive determinant, i.e., on the generating number of the set of matrices with positive determinant from $\mathcal{P}$. The following lemma states simple, but useful, observations.

LEMMA 2.1. *If A with $\det A > 0$ is a product of k P-matrices, then A^{-1}, any permutation similarity of A, any signature similarity of A, and any positive diagonal equivalence of A can each be factored as a product of k P-matrices. If $A = A_1 \oplus A_2$, and A_j can be factored as a product of k_j P-matrices, then A can be factored as a product of $\max_j \{k_j\}$ P-matrices.*

We first consider those matrices with positive determinant that are diagonal, and we begin with an important example. The identity matrix in $M_n(\mathbb{R})$ is denoted by I_n.

Example 2.2.

$$-I_2 = \begin{bmatrix} 1/3 & 2/3 \\ -4/3 & 1/3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 1/3 \\ -2/3 & 1/3 \end{bmatrix}$$

may be factored as a product of three P-matrices. Moreover, this number is best possible, as $-I_2$ cannot be factored as P_1P_2 with $P_j \in \mathcal{P}$. If this were possible, then $P_2 = -P_1^{-1}I_2$ with $P_1^{-1} \in \mathcal{P}$, but the right side product has $(1,1)$ entry negative, whereas P_2 has $(1,1)$ entry positive.

LEMMA 2.3. *If A is a real diagonal matrix with $\det A > 0$, then A can be factored as a product of (at most) three P-matrices.*

Proof. Since $\det A > 0$, there exists a permutation matrix Q and a positive diagonal matrix D so that

$$DQAQ^T = (-I_2) \oplus (-I_2) \oplus \cdots \oplus (-I_2) \oplus I_p,$$

where the number of $-I_2$ blocks may be zero. By Example 2.2, each $-I_2$ can be factored as a product of three P-matrices. Thus by Lemma 2.1 and the fact that $I_p \in \mathcal{P}$, A can be factored as the product of three P-matrices. ∎

It follows from the above lemma that 3 is the generating number of the set of real diagonal matrices with positive determinant from $\mathcal{P}$.

An $n \times n$ real matrix has a *nested sequence of positive principal minors* if it contains a sequence of positive principal minors of orders $1, 2, \ldots, n$, where each minor's

index set is properly contained in the next; this is equivalent to a permutation similarity having a leading sequence of positive principal minors. This concept is needed for the following lemma which is used to find the generating number of the set of matrices with positive determinant from $\mathcal{P}$.

LEMMA 2.4. *If $A \in M_n(\mathbb{R})$ is nondiagonal with $\det A > 0$, then there exists $P \in \mathcal{P}$ such that AP has a nested sequence of positive principal minors.*

Proof. The proof is by induction on n. Let $A = [a_{ij}]$ and, without loss of generality, assume that $a_{12} \neq 0$ (since this can be achieved by permutation similarity). Let

$$\hat{P} = \begin{bmatrix} 1 & 0 \\ w & I_{n-1} \end{bmatrix} \tag{2.1}$$

be the $n \times n$ matrix in which w^T consists of the first $(n-1)$ entries of $(p_{21}, p_{31}, 1, \ldots 1)$, and p_{21}, p_{31} are arbitrary. Thus $\hat{P} \in \mathcal{P}$.

For $n = 2$, the $(1,1)$ entry of $A\hat{P}$ is $a_{11} + a_{12}p_{21}$, which can be made positive by a suitable choice of p_{21}. Since $\det(A\hat{P}) > 0$, AP has a nested sequence of positive principal minors with $\hat{P} = P$. Assume the result is true for $n = k \geq 2$. Take $A \in M_{k+1}(\mathbb{R})$ with $\det A > 0$ and $\hat{P} \in \mathcal{P}$ as in (2.1) with $n = k+1$. Consider $(A\hat{P})^{-1} = \hat{P}^{-1}A^{-1}$, in which $A^{-1} = [\alpha_{ij}]$ and

$$\hat{P}^{-1} = \begin{bmatrix} 1 & 0 \\ -w & I_k \end{bmatrix}. \tag{2.2}$$

Partition

$$\hat{P}^{-1}A^{-1} = \begin{bmatrix} b & v^T \\ u & B \end{bmatrix} \tag{2.3}$$

with $B = [b_{ij}]$ for $i, j = 2, \ldots, k+1$.

We claim that by suitable choices of p_{21} and p_{31} we can make $\det B > 0$ and $b_{i2} \neq 0$ for some $i \geq 3$ (i.e., B nondiagonal). First, notice that if $\alpha_{12} = 0$, then the second column of A^{-1} cannot have α_{22} as the only nonzero entry (since $AA^{-1} = I_{k+1}$ and $a_{12} \neq 0$); thus $\alpha_{i2} \neq 0$ for some $3 \leq i \leq k+1$. From (2.3), $b_{i2} = \alpha_{i2} \neq 0$. Second, if $\alpha_{12} \neq 0$, then p_{31} can be chosen so that from (2.3)

$$b_{32} = -p_{31}\alpha_{12} + \alpha_{32} \neq 0.$$

The $(1,1)$ entry of $A\hat{P}$ is $a_{12}p_{21} + a_{13}p_{31} + \sum_{\substack{j=1 \\ j\neq 2,3}}^{k+1} a_{1j}$, and p_{21} can be chosen so that this entry is positive. This fact, together with $\det(\hat{P}^{-1}A^{-1}) > 0$, gives $\det B > 0$. Thus $\det\left(B^{-1}\right) > 0$, and B^{-1} is not diagonal. By the induction hypothesis there exists $P_1 \in \mathcal{P}$ so that $B^{-1}P_1$ has a nested sequence of positive principal minors, and consequently so does $P_1^{-1}B$. From (2.2) and (2.3)

$$\begin{bmatrix} 1 & 0 \\ 0 & P_1^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -w & I_k \end{bmatrix} A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & P_1^{-1} \end{bmatrix} \begin{bmatrix} b & v^T \\ u & B \end{bmatrix} = \begin{bmatrix} b & v^T \\ P_1^{-1}u & P_1^{-1}B \end{bmatrix}$$

has a nested sequence of positive principal minors of orders $1, \ldots, k+1$ since the determinant is positive. However, the product of the two P-matrices on the left is

$$P^{-1} = \begin{bmatrix} 1 & 0 \\ -P_1^{-1}w & P_1^{-1} \end{bmatrix} \in \mathcal{P}.$$

Consequently $P^{-1}A^{-1}$, and thus AP with $P \in \mathcal{P}$, has a nested sequence of positive principal minors. ∎

We note that if $\det A > 0$, it is also true that there exists $P \in \mathcal{P}$ so that PA has a nested sequence of positive principal minors.

LEMMA 2.5. *If $A \in M_n(\mathbb{R})$ has a nested sequence of positive principal minors, then A is the product of (at most) two P-matrices.*

Proof. There exists a permutation matrix Q so that QAQ^T has a nested sequence of leading positive principal minors. Then QAQ^T has an LU factorization, i.e., $QAQ^T = LU$, where L and U are lower and upper triangular P-matrices, respectively. By Lemma 2.1, A can be factored as the product of two P-matrices. ∎

The above results lead to the conclusion of the following theorem, namely that 3 is the generating number of all real matrices with positive determinant from the P-matrices. Note that this is smaller than the generating number 5 (4 for nonscalar matrices) obtained from the positive definite matrices [2, 11] and that the product of two positive definite matrices is not, in general, a P-matrix.

THEOREM 2.6. *Any real matrix A with $\det A > 0$ can be written as the product of (at most) three P-matrices.*

Proof. For diagonal A the result is given by Lemma 2.3. For nondiagonal A, Lemma 2.4 shows that there exists $P \in \mathcal{P}$ so that AP has a nested sequence of positive principal minors. By Lemma 2.5, AP is the product of two P-matrices. Thus A is a product of three P-matrices.

The fact that the generating number is exactly 3 follows from Example 2.2 for $-I_2$. Furthermore, for a nondiagonal example, it can be easily checked that any 2×2 matrix with every entry negative and positive determinant cannot be factored as the product of two P-matrices. ∎

The following theorem gives remarkable mixed factorizations for nondiagonal matrices with positive determinant. Parts (i) and (ii) can be interpreted as the fact that any nondiagonal matrix with positive determinant can be stabilized via multiplication by a P-matrix. This should be compared with the classical result [3, 6] that a matrix with a nested sequence of positive principal minors can be stabilized via multiplication by a positive diagonal matrix.

THEOREM 2.7. *If $A \in M_n(\mathbb{R})$ is nondiagonal with $\det A > 0$, then*

(i) *there exist $B \in \mathcal{S}^+$ and $C \in \mathcal{P}$ so that $A = BC$;*
(ii) *there exist $C \in \mathcal{P}$ and $B \in \mathcal{S}^+$ so that $A = CB$;*
(iii) *there exist $P_1, P_2 \in \mathcal{PD}$ and $C \in \mathcal{P}$ so that $A = P_1P_2C$; and*
(iv) *there exist $C \in \mathcal{P}$ and $P_1, P_2 \in \mathcal{PD}$ so that $A = CP_1P_2$.*

Proof. By Lemma 2.4, there exists $P \in \mathcal{P}$ such that AP has a nested sequence of positive principal minors. Then using a result of [6] (see also [3]), there is a positive diagonal matrix D so that $B = APD \in \mathcal{S}^+$. Since $PD = C^{-1} \in \mathcal{P}$, statement (i) follows. Statement (ii) follows by inversion, since $A^{-1} = C^{-1}B^{-1}$ is also nondiagonal with positive determinant and $\mathcal{P}$ and $\mathcal{S}^+$ are closed under inversion. The result from [7, Thm. 7.6.3, p. 465], given in the introduction, states that B can be written as P_1P_2, where $P_1, P_2 \in \mathcal{PD}$. Thus statements (iii) and (iv) follow from (i) and (ii), respectively. ∎

If $\det A > 0$ and A is not a negative diagonal matrix, then we can also give a mixed factorization.

LEMMA 2.8. *Let $A \in M_n(\mathbb{R})$ be a diagonal matrix with $\det A > 0$. Then $A = CB$ in which $C \in \mathcal{P}$ and $B \in \mathcal{S}$ if and only if A is not a negative diagonal matrix.*

Proof. Consider DA, where D is the positive diagonal matrix so that the diagonal entries of DA are ± 1. First, assume that all diagonal entries of DA are -1, so that

$DA = -I_n$ with n even. If $A = CB$ with $C \in \mathcal{P}$, then $DA = DCB$ with $DC \in \mathcal{P}$; thus $B = -(DC)^{-1}$. This means that each diagonal entry of B is negative, showing that B is not positive stable. Second, assume that DA has an even number of -1 entries and at least one $+1$ entry. By considering direct sums, we need only consider the factorization of $DA = -I_{n-1} \oplus [1]$; thus n is odd. For this case, let

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

with $\delta > 0$. Thus $\tilde{B}$ is a Schwarz matrix (see, e.g., [8, Prob. 9, p. 111]) and all eigenvalues of $\tilde{B}$ have positive real part. Now $\varepsilon > 0$ can be chosen sufficiently small so that $\tilde{B} - \varepsilon I$ is positive stable. Then

$$DA\left(\tilde{B} - \varepsilon I\right) = \begin{bmatrix} \varepsilon & -1 & & & & & & \\ 1 & \varepsilon & -1 & & & 0 & & \\ & 1 & \varepsilon & -1 & & & & \\ & & 1 & \varepsilon & -1 & & & \\ & & & \ddots & \ddots & -1 & & \\ & & & & \ddots & \ddots & \ddots & \\ & 0 & & & & 1 & \varepsilon & -1 \\ & & & & & & -\delta & (1-\varepsilon) \end{bmatrix} = \tilde{C}.$$

If $\delta = 0$, then from the digraph expansion of the determinant, it can be seen that all terms in the determinant of any principal minor of $\tilde{C}$ are positive, since each such term consists of a product of positive 1-cycles and negative 2-cycles. Thus, by continuity, for $\delta > 0$ sufficiently small, $\tilde{C} \in \mathcal{P}$. Thus $A = D^{-1}\tilde{C}(\tilde{B} - \varepsilon I)^{-1} = CB$ in which $C = D^{-1}\tilde{C} \in \mathcal{P}$ and $B = (\tilde{B} - \varepsilon I)^{-1} \in \mathcal{S}$. The case in which $AD = I_n$ is obvious, thus completing the proof. □

We note that, since $\mathcal{P}$ and $\mathcal{S}$ are closed under inversion, it is also true that if A is a diagonal matrix with $\det A > 0$, then $A = BC$ in which $B \in \mathcal{S}$ and $C \in \mathcal{P}$ if and only if A is not a negative diagonal matrix. Combining the results of Theorem 2.7 and Lemma 2.8 gives the following equivalent statements.

THEOREM 2.9. *If $A \in M_n(\mathbb{R})$ with $\det A > 0$, then $TFAE$.*

(i) *There exist $B \in \mathcal{S}$ and $C \in \mathcal{P}$ such that $A = BC$.*

(ii) *There exist $C \in \mathcal{P}$ and $B \in \mathcal{S}$ such that $A = CB$.*

(iii) *A is not a negative diagonal matrix.*

Note that if n is odd, then A cannot be a negative diagonal matrix; thus, in this case, any matrix with positive determinant can be factored as in (i) and (ii).

3. Products of two *P*-matrices. The results of the previous section raise the question of which nondiagonal matrices A can be factored as the product of (at most) two P-matrices. (Note that if A is diagonal, this is possible if and only if $A \in \mathcal{P}$, and only one factor is needed.) Lemma 2.5 provides one such sufficient condition; namely,

if A has a nested sequence of positive principal minors, then (at most) two factors are needed. However, this is not a necessary condition, as illustrated by

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

Here A has positive determinant and no nested sequence of principal minors but can be factored as a product of two P-matrices. Also, if A has all eigenvalues with positive real part, then Lyapunov's theorem implies that A can be factored as the product of a positive definite matrix and a matrix with positive definite symmetric part, and hence of two P-matrices.

The next result means that we can restrict our consideration to irreducible matrices.

LEMMA 3.1. *Suppose $A \in M_n(\mathbb{R})$ with $\det A > 0$. If A is reducible, then A is a product of two P-matrices if and only if every maximal irreducible submatrix of A is a product of two P-matrices.*

Proof. If $A_1 = P_1P_2$ and $A_2 = P_3P_4$ with $P_j \in \mathcal{P}$, then

$$A = \begin{bmatrix} A_1 & X \\ 0 & A_2 \end{bmatrix} = \begin{bmatrix} P_1 & XP_4^{-1} \\ 0 & P_3 \end{bmatrix} \begin{bmatrix} P_2 & 0 \\ 0 & P_4 \end{bmatrix},$$

in which each matrix on the right is a P-matrix. The converse follows similarly. ∎

THEOREM 3.2. *Suppose $\det A > 0$. If there exists an open orthant that is mapped into its negative by A or A^{-1}, then A is not the product of two P-matrices.*

Proof. Let $w \in \mathbb{R}^n$ be a vector with each component ± 1 and, using the Hadamard product, define $W = \{x \in \mathbb{R}^n : x \circ w > 0\}$. Assume that $AW \subseteq -W$ and that A is the product of two P-matrices, namely $A = P_1P_2$. Equivalently $AP_3 = P_1$ with $P_3 = P_2^{-1} \in \mathcal{P}$. It follows from [5, A_6 on p. 135] with the signature matrix $S = (s_{ij})$ and $sign(s_{ii}) = x_i$ that $SP_3Sz > 0$ for some $z > 0$. Thus $x = Sz \in W$ such that $y = P_3x \in W$ and $Ay = P_1x$. However, $Ay \in -W$ gives $P_1x \in -W$, which cannot be true for $P_1 \in \mathcal{P}$ [5, A_5 on p. 134]. Thus A is not the product of two P-matrices. The result for A^{-1} follows similarly. ∎

A restatement of this result is as follows. If A or A^{-1} reverses in sign a signed vector (for any magnitude), then A cannot be written as the product of two P-matrices. A useful special case follows.

COROLLARY 3.3. *Any matrix $A \in M_n(\mathbb{R})$ with $\det A > 0$ that is either nonpositive, signature similar to a nonpositive matrix, or has an inverse in either of these forms cannot be factored as a product of two P-matrices.*

Example 3.4. Let

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

with $\det A = 1$. To decide (based on the above results) whether A can be factored as a product of two P-matrices, compute

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

Clearly this is a negative matrix, and thus A cannot be factored as the product of two P-matrices.

For the case $n = 2$, by considering all sign patterns and exhibiting factorizations, we can see that the conditions of Corollary 3.3 are both necessary and sufficient.

THEOREM 3.5. *For nondiagonal $A \in M_2(\mathbb{R})$ with $\det A > 0$, if A does not have one of the sign patterns given by Corollary 3.3, then A can be factored as a product of two P-matrices.*

Since it is known that a P-matrix cannot have a negative eigenvalue (see, e.g., [8]), it is interesting to remark that the product of two P-matrices can have all eigenvalues negative. For example, the matrix

$$\begin{bmatrix} -1 & -4 \\ 4 & -9 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ \frac{22}{7} & \frac{6}{7} \end{bmatrix} \begin{bmatrix} 1 & -3 \\ 1 & \frac{1}{2} \end{bmatrix}$$

has -5 as a repeated eigenvalue.

4. Factorization involving matrix classes $\mathcal{M}$ and $\mathcal{TN}$. Consider now another set for $\mathcal{U}$, namely nonsingular M-matrices (which is a subset of the P-matrices not containing all of the positive definite matrices). The following result is proved using elementary bidiagonal (EB) factorizations. An EB matrix is a matrix with each diagonal entry equal to 1 and only one other nonzero entry, which is on either the sub- or the superdiagonal.

THEOREM 4.1. *The nonsingular M-matrices generate the real matrices with positive determinant.*

Proof. From the proof of Theorem 6 in [10], every nonsingular matrix has an EB factorization, in which each factor is either an EB matrix or a positive diagonal matrix (with at most one such diagonal matrix). Each EB matrix is either a nonsingular M-matrix or an inverse M-matrix. Since the determinant is positive, the result follows. ∎

Each EB matrix is also either a nonsingular totally nonnegative matrix or an inverse totally nonnegative matrix; thus a similar proof holds for the set of nonsingular totally nonnegative matrices (which is also a subset of the P-matrices).

THEOREM 4.2. *The nonsingular totally nonnegative matrices generate the real matrices with positive determinant.*

From results on elementary bidiagonal matrices, the generating number of the $n \times n$ real matrices with positive determinant from $\mathcal{M}$ or $\mathcal{TN}$ is at most $\mathcal{O}(n^2)$; see, e.g., [10]. For $n = 2$, the set of M-matrices is equal to the set of inverse totally nonnegative matrices, and this generating number is determined exactly.

THEOREM 4.3. *Any $A \in M_2(\mathbb{R})$ with $\det A > 0$ can be generated by the nonsingular M-matrices, with a generating number of 4. If A is nondiagonal, then at most three matrices are needed.*

Proof. If A is nondiagonal, perform one type 3 elementary row or column operation on A so that the new matrix $\tilde{A}$ has a positive $(1,1)$ entry. Such an operation is equivalent to pre- or postmultiplying A by an EB matrix, which is either an M-matrix or an inverse M-matrix. Since $\det \tilde{A} = \det A > 0$, $\tilde{A}$ has a nested sequence of positive principal minors. By the proof of Lemma 2.5, $\tilde{A}$ is the product of two triangular matrices that are either M- or inverse M-matrices. Thus A is a product of at most three M- and inverse M-matrices.

If $A = -I_2$, then four matrices from this set are needed, as is seen from the following factorization in terms of two M-matrices and two inverse M-matrices:

$$-I_2 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ -2 & 1/3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2/3 & 1/3 \end{bmatrix}.$$

A consideration of all sign patterns shows that $-I_2$ cannot be factored in terms of three M- and inverse M-matrices. ∎

The next theorem determines for $n = 2$ the set that can be generated by two matrices in $\mathcal{M}$. For its proof, we use an interesting result that we give in $M_n(\mathbb{R})$.

LEMMA 4.4. *If $A \in \mathcal{M}$ and $B \in \mathcal{M}^{-1}$ are $n \times n$ matrices, then AB (and BA) has at least one positive diagonal entry.*

Proof. Consider the Hadamard product $A \circ B^T$. This is an M-matrix [8, p. 359] and its row sums are the diagonal entries of AB. Thus taking e as the vector of all 1's, $(AB)_{ii} = \left[\left(A \circ B^T\right) e\right] > 0$ for at least one i [8, Exer. 13, p. 128]. The result for BA follows similarly since the column sums of $A \circ B^T$ are the diagonal entries of BA. ∎

THEOREM 4.5. *Any $A \in M_2(\mathbb{R})$ with $\det A > 0$ is generated by $\mathcal{M}$ with generating number 2 if and only if A has a positive diagonal entry.*

Proof. Assume that A with $\det A > 0$ can be factored as the product of two M- and inverse M-matrices. By Lemma 4.4 and considering the product of two M-matrices or two inverse M-matrices in $M_2(\mathbb{R})$, at least one diagonal entry must be positive. Thus A has a nested sequence of positive principal minors. As in the proof of Lemma 2.5, A is the product of two triangular M- or inverse M-matrices.

For the converse, given $\det A > 0$, assume that A has a positive diagonal entry, without loss of generality, $a_{11} > 0$. Thus $A = LU$ with L and U having positive diagonal entries, and each is either an M- or inverse M-matrix. ∎

The converse of Lemma 2.5 is obviously false for P-matrices, even with $n = 2$. However, as we now show, the converse statement does hold for a product of an M-matrix and an inverse M-matrix.

THEOREM 4.6. *If $A \in \mathcal{M}$ and $B \in \mathcal{M}^{-1}$ are $n \times n$ matrices, then AB (and BA) has a nested sequence of positive principal minors.*

Proof. The result follows immediately for $n = 2$ by Lemma 4.4 and the fact that $\det(AB) > 0$. To proceed by induction, note that an $(n-1)\times(n-1)$ principal minor of AB is positive since (by Lemma 4.4) a diagonal entry of $B^{-1}A^{-1}$ is positive and $\det(AB) > 0$. Assume without loss of generality that it is the leading principal minor, and partition A, B accordingly as

$$A = \begin{bmatrix} A_{11} & -a_{12} \\ -a_{21}^T & a_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & b_{12} \\ b_{21}^T & b_{22} \end{bmatrix},$$

in which a_{12}, a_{21}, b_{12}, $b_{21} \geq 0$ and $a_{22}, b_{22} > 0$. Here $A_{11} \in \mathcal{M}$, $B_{11} \in \mathcal{M}^{-1}$, and $\det\left(A_{11}B_{11} - a_{12}b_{21}^T\right) > 0$ by the above assumption.

We now show that the leading principal submatrix $A_{11}B_{11} - a_{12}b_{21}^T$ is a product of an M-matrix and an inverse M-matrix. Write

$$A_{11}B_{11} - a_{12}b_{21}^T = \left(A_{11} - a_{12}b_{21}^T B_{11}^{-1}\right) B_{11} = A_{11}\left(I_{n-1} - A_{11}^{-1}a_{12}b_{21}^T B_{11}^{-1}\right) B_{11}.$$

As $b_{21}^T B_{11}^{-1} \geq 0$ (a necessary condition for a matrix in $\mathcal{M}^{-1}$ from the partitioned form of the inverse [7, p. 18]) and $A_{11}^{-1}a_{12} \geq 0$, it follows that both $A_{11} - a_{12}b_{21}^T B_{11}^{-1}$ and $I_{n-1} - A_{11}^{-1}a_{12}b_{21}^T B_{11}^{-1}$ are matrices with the Z-sign pattern, which are nonpositive rank 1 perturbations of the M-matrices A_{11} and I_{n-1}, respectively. Also $\det\left(A_{11} - a_{12}b_{21}^T B_{11}^{-1}\right) > 0$, as $\det B_{11} > 0$ and $\det\left(A_{11}B_{11} - a_{12}b_{21}^T\right) > 0$. In addition as $\det A_{11} > 0$, it follows that $\det\left(I_{n-1} - A_{11}^{-1}a_{12}b_{21}^T B_{11}^{-1}\right) > 0$. This inequality, together with the fact that $A_{11}^{-1}a_{12}b_{21}^T B_{11}^{-1}$ has rank 1, shows that $I_{n-1} - A_{11}^{-1}a_{12}b_{21}^T B_{11}^{-1} \in \mathcal{M}$ [7, p. 19].

Now the Z-sign pattern matrix $A_{11} - a_{12}b_{21}^T B_{11}^{-1}$ can be written as the product of two matrices in $\mathcal{M}$, namely as $A_{11}\left(I_{n-1} - A_{11}^{-1}a_{12}b_{21}^T B_{11}^{-1}\right)$. Thus $A_{11} - a_{12}b_{21}^T B_{11}^{-1}$ is inverse nonnegative and so it is in $\mathcal{M}$, proving that $A_{11}B_{11} - a_{12}b_{21}^T$ is the product of a matrix in $\mathcal{M}$ with a matrix in $\mathcal{M}^{-1}$. The induction hypothesis then delivers a nested sequence of principal minors for this $(n-1)\times(n-1)$ principal submatrix of AB, and thus for AB, as $\det(AB) > 0$. The proof for BA follows by inversion. ∎

The following result follows from Theorem 4.6 and [6].

COROLLARY 4.7. *If $A \in \mathcal{M}$ and $B \in \mathcal{M}^{-1}$ are $n \times n$ matrices, then there exist positive diagonal matrices D_1, D_2 so that D_1AB and $D_2BA \in \mathcal{S}^+$.*

The example

$$\begin{bmatrix} 1 & -3 \\ -\frac{1}{4} & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -2 & -5 \\ \frac{3}{4} & \frac{7}{4} \end{bmatrix}$$

shows that, even for $n = 2$, it is not, in general, true that the product of an M-matrix and an inverse M-matrix is stable, since this product has a negative trace.

5. Concluding remarks and questions. The mixed factor factorization results in Theorems 2.7 and 2.9, and the M-matrix results in section 4, raise interesting questions, some of which we now record. Theorem 2.7 includes the condition that nondiagonal A with $\det A > 0$ can be factored in either of the two ways $A = P_1P_2C$ and CP_1P_2, in which $C \in \mathcal{P}$ and $P_1, P_2 \in \mathcal{PD}$. It is unknown whether such a matrix A can also be written as P_1CP_2. If the set product $\mathcal{P}$ times $\mathcal{PD}$ is inverse closed, then it is also $\mathcal{PD}$ times $\mathcal{P}$ and the factorization P_1CP_2 would follow from Theorem 2.7. It is also unknown whether or not the converse of Theorem 2.7 is true. If A is not a negative diagonal matrix, then Theorem 2.9 gives a mixed factorization with $B \in \mathcal{S}$. Does such a factorization also hold with $B \in \mathcal{S}^+$ (cf. Theorem 2.7)? Our proof of Lemma 2.8 does not provide such a factorization, since a Schwarz matrix may have nonreal eigenvalues.

It has been proved (Theorem 4.1) that the nonsingular M-matrices generate the real matrices with $\det A > 0$ and that, for $n = 2$, the generating number is 4. It is not known what the generating number is for $n \geq 3$. Is it bounded or does it grow as n or n^2?

Theorem 4.6 can be viewed as stating that a nested sequence of positive principal minors is a necessary condition for a matrix to be a product of an M-matrix and an inverse M-matrix. It is not known whether this is a necessary condition for a matrix to be a product of two M-matrices (equivalently, two inverse M-matrices) for $n \geq 4$. For $n \times n$ matrices A, $B \in \mathcal{M}$, it is easy to see that AB has all principal minors of orders 1, $n-1$, and n positive. In the special case $n = 4$, if A and B are, in addition, irreducible with the longest simple cycle in the digraph of $A + B$ having length ≤ 3, then AB has all order 2 principal minors positive; see [9, Lem. 3]. Thus in this special case $AB \in \mathcal{P}$. Even if A has a nested sequence of positive principal minors, it is unknown whether it can be written as a product of two matrices from $\mathcal{M} \cup \mathcal{M}^{-1}$ (cf. Lemma 2.5).

REFERENCES

[1] T. ANDO, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.
[2] C.S. BALLENTINE, *Products of positive definite matrices* III, J. Algebra, 10 (1968), pp. 174–182.
[3] C.S. BALLENTINE, *Stabilization by a diagonal matrix*, Proc. Amer. Math. Soc., 25 (1970), pp. 729–734.

[4] C.S. Ballentine and C.R. Johnson, *Accretive matrix products*, Linear Multilinear Algebra, 3 (1975), pp. 169–185.
[5] A. Berman and R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, London, 1979.
[6] M.E. Fisher and A.T. Fuller, *On the stabilization of matrices and the convergence of linear iterative processes*, Proc. Cambridge Philos. Soc., 54 (1958), pp. 417–425.
[7] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[8] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[9] C.R. Johnson, D.D. Olesky, and P. van den Driessche, *M-matrix products having positive principal minors*, Linear Multilinear Algebra, 16 (1984), pp. 29–38.
[10] C.R. Johnson, D.D. Olesky, and P. van den Driessche, *Elementary bidiagonal factorizations*, Linear Algebra Appl., 292 (1999), pp. 233–244.
[11] A.R. Sourour, *A factorization theorem for matrices*, Linear Multilinear Algebra, 19 (1986), pp. 141–147.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. vii–vii

CORRECTION TO "MATRIX CLASSES THAT GENERATE ALL MATRICES WITH POSITIVE DETERMINANT"

There is an editing error on page 289 of "Matrix Classes That Generate All Matrices with Positive Determinant," *SIAM Journal on Matrix Analysis and Applications*, 25 (2003), pp. 285–294, by C. R. Johnson, D. D. Olesky, and P. van den Driessche. The sentence after Theorem 2.9 should read "Note that if n is odd, then A cannot be a negative diagonal matrix; thus, in this case, any matrix with positive determinant can be factored as in (i) and (ii)."

SIAM regrets this error.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 1, pp. 295–300

ON THE DIGRAPH OF A UNITARY MATRIX*

SIMONE SEVERINI†

Abstract. Given a matrix M of size n, the digraph D on n vertices is said to be the *digraph of* M, when $M_{ij} \neq 0$ if and only if (v_i, v_j) is an arc of D. We give a necessary condition, called strong quadrangularity, for a digraph to be the digraph of a unitary matrix. With the use of such a condition, we show that a line digraph $\overrightarrow{L}D$ is the pattern of a unitary matrix if and only if D is Eulerian. It follows that, if D is strongly connected and $\overrightarrow{L}D$ is the digraph of a unitary matrix, then $\overrightarrow{L}D$ is Hamiltonian. We observe that strong quadrangularity is sufficient to show that disconnected strongly regular graphs are the digraphs of unitary matrices and that n-paths, n-paths with loops at each vertex, n-cycles, directed trees, and trees are not.

Key words. digraphs, unitary matrices, quantum random walks

AMS subject classifications. 05C20, 51F25, 81P68

DOI. 10.1137/S0895479802410293

1. Introduction. Let $D = (V, A)$ be a digraph on n vertices, with labeled vertex set $V(D)$, arc set $A(D)$, and adjacency matrix $M(D)$. We assume that D may have loops and multiple arcs. Let M be a matrix over any field. A digraph D is the *digraph of* M, or, equivalently, the *pattern of* M, if $|V(D)| = n$, and, for every $v_i, v_j \in V(D)$, $(v_i, v_j) \in A(D)$ if and only if $M_{ij} \neq 0$. The *support* ${}^sM$ of the matrix M is the $(0, 1)$-matrix with element

$$ {}^sM_{ij} = \begin{cases} 1 & \text{if } M_{ij} \neq 0, \\ 0 & \text{otherwise.} \end{cases} $$

Then the digraph of a matrix is the digraph whose adjacency matrix is the support of the matrix. The *line digraph* of a digraph D, denoted by $\overrightarrow{L}D$, is the digraph whose vertex set $V(\overrightarrow{L}D)$ is $A(D)$ and where $((v_i, v_j), (v_j, v_k)) \in A(\overrightarrow{L}D)$ if and only if $(v_i, v_j), (v_j, v_k) \in A(D)$.

A *discrete quantum random walk* on a digraph D is a discrete walk on D induced by a unitary transition matrix. The term *quantum random walk* was coined by Gudder (see, e.g., [G88]), who introduced the model and proposed to use it to describe the motion of a quantum object in discrete space-time and to describe the internal dynamics of elementary particles. Recently, quantum random walks have been rediscovered, in the context of quantum computation, by Ambainis et al. (see [ABNVW01] and [AKV01]). Since the notion of quantum random walks is analogous to the notion of random walks, interest in quantum random walks has been fostered by the successful use of random walks on combinatorial structures in probabilistic algorithms (see, e.g., [L93]). Clearly, a quantum random walk on a digraph D can be defined if and only if D is the digraph of a unitary matrix. Inspired by the work of Meyer on quantum cellular automata [M96], the authors of [ABNVW01] and [AKV01] overcame this obstacle in the following way. In order to define a quantum random walk on a simple digraph D, which is regular and is not the digraph of a unitary matrix, a quantum

*Received by the editors June 25, 2002; accepted for publication by H. Woerdeman December 20, 2002; published electronically July 11, 2003.
http://www.siam.org/journals/simax/25-1/41029.html
†Department of Computer Science, University of Bristol, Bristol, BS8 1UB, UK (severini@cs.bris.ac.uk). The author is supported by a University of Bristol research scholarship.

random walk on $\overrightarrow{L}D$ is defined. The digraph $\overrightarrow{L}D$ is the digraph of a unitary matrix. When we choose an appropriate labeling for $V(\overrightarrow{L}D)$, a quantum random walk on $\overrightarrow{L}D$ induces a probability distribution on $V(D)$. The quantum random walk on $\overrightarrow{L}D$ is called the *coined quantum random walk* on D.

With this scenario in mind, the question which this paper addresses is the following: On which digraphs can quantum random walks be defined? In more general language, we are interested in the combinatorial properties of the digraphs of unitary matrices. We give a simple necessary condition, called *strong quadrangularity*, for a digraph to be a digraph of a unitary matrix. While it seems too daring to conjecture that such a condition is sufficient in the general case, we discover "accidentally" that strong quadrangularity is sufficient when the digraph is a line digraph. We also prove that if a line digraph of a strongly connected digraph is the digraph of a unitary matrix, then it is Hamiltonian. We observe that strong quadrangularity is sufficient to show that certain strongly regular graphs are digraphs of unitary matrices and that n-paths, n-paths with loops at each vertex, n-cycles, directed trees, and trees are not. In [GZe88] and [M96] the fact that an n-path is not the digraph of a unitary matrix was called the *NO-GO lemma*. A consequence of the lemma was that there is no nontrivial, homogeneous, local, one-dimensional quantum cellular automaton. Proposition 2.21 below can be interpreted as a simple combinatorial version of the NO-GO lemma.

We refer to [T84] and [BR91] for notions of graph theory and matrix theory, respectively.

2. Digraphs of unitary matrices. Let $D = (V, A)$ be a digraph. A vertex of a digraph is called a *source* (*sink*) if it has no ingoing (outgoing) arcs. A vertex of a digraph is said to be *isolated* if it is not joined to another vertex. We assume that D has no sources, sinks, and disconnected loopless vertices. By this assumption, $A(D)$ has neither zero-rows nor zero-columns. For every $S \subset V(D)$, denote by

$$N^+[S] = \{v_j : (v_i, v_j) \in A(D), v_i \in S\}$$

and

$$N^-[S] = \{v_i : (v_i, v_j) \in A(D), v_j \in S\}$$

the *out-neighborhood* and *in-neighborhood* of S, respectively. Denote by $|X|$ the cardinality of a set X. The integers $|N^-[v_i]|$ and $|N^+[v_i]|$ are called *invalency* and *outvalency* of the vertex v_i, respectively. A digraph D is *Eulerian* if and only if every vertex of D has equal invalency and outvalency.

The notion defined in Definition 2.1 below is standard in combinatorial matrix theory (see, e.g., [BR91]). In graph theory, the term *quadrangular* was first used in [GZ98].

Definition 2.1. *A digraph D is said to be* quadrangular *if, for any two distinct vertices $v_i, v_j \in V(D)$, we have*

$$|N^+[v_i] \cap N^+[v_j]| \neq 1 \quad \textit{and} \quad |N^-[v_i] \cap N^-[v_j]| \neq 1.$$

Definition 2.2. *A digraph D is said to be* strongly quadrangular *if there does not exist a set $S \subseteq V(D)$ such that, for any two distinct vertices $v_i, v_j \in S$,*

$$N^+[v_i] \cap \bigcup_{j \neq i} N^+[v_j] \neq \emptyset \quad \textit{and} \quad N^+[v_i] \cap N^+[v_j] \subseteq T,$$

where $|T| < |S|$, and similarly for the in-neighborhoods.

Remark 2.3. Note that if a digraph is strongly quadrangular, then it is quadrangular.

LEMMA 2.4. *Let D be a digraph. If D is the digraph of a unitary matrix, then D is strongly quadrangular.*

Proof. Suppose that D is the digraph of a unitary matrix U and that D is not strongly quadrangular. Then there is a set $S \subseteq V(D)$ such that, for any two distinct vertices $v_i, v_j \in S$, $N^+[v_i] \cap \bigcup_{j \neq i} N^+[v_j] \neq \emptyset$ and $N^+[v_i] \cap N^+[v_j] \subseteq T$, where $|T| < |S|$. This implies that in U, there is a set S' of rows which contribute, with at least one nonzero entry, to the inner product with some other rows in S'. In addition, the nonzero entries of any two distinct rows in S', which contribute to the inner product of the two rows, are in the columns of the same set of columns T' such that $|T'| < |S'|$. Then the rows of S' form a set of orthonormal vectors of a dimension smaller than the cardinality of the set itself. This contradicts the hypothesis. The same reasoning holds for the columns of U. ∎

Two digraphs D and D' are *permutation equivalent* if there exist permutation matrices P and Q such that $M(D') = PM(D)Q$ (and hence also $P^{-1}M(D')Q^{-1} = M(D)$). If $Q = P^{-1}$, then D and D' are said to be *isomorphic*. We write $D \cong D'$ if D and D' are isomorphic. Denote by I_n the identity matrix of size n. Denote by $A^\intercal$ the transpose of a matrix A.

LEMMA 2.5. *Let D and D' be permutation equivalent digraphs. Then D is the digraph of a unitary matrix if and only if D' is.*

Proof. Suppose that D is the digraph of a unitary matrix U. Then, for permutation matrices P and Q, we have $PUQ = U'$, where U' is a unitary matrix of the digraph D'. The converse is similar. ∎

LEMMA 2.6. *For any n the complete digraph is the digraph of a unitary matrix.*

Proof. The lemma just means that for every n there is a unitary matrix without zero entries. An example is given by the Fourier transform on the group $\mathbb{Z}/n\mathbb{Z}$ (see, e.g., [T99]). ∎

A digraph D is said to be (k, l)*-regular* if, for every $v_i \in V(D)$, we have $|N^-[v_i]| = k$ and $|N^+[v_i]| = l$. If $k = l$, then D is said to be simply k*-regular*.

Remark 2.7. Not every digraph that is permutation equivalent to a k-regular digraph is the digraph of a unitary matrix. Let

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

Note that D is 2-regular and is not quadrangular.

Remark 2.8. Not every quadrangular digraph is the digraph of a unitary matrix. Let

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

Note that D is quadrangular and is not the digraph of a unitary matrix. In fact, D is not strongly quadrangular.

DEFINITION 2.9. *A digraph D is said to be* specular *when, for any two distinct vertices $v_i, v_j \in V(D)$, if $N^+[v_i] \cap N^+[v_j] \neq \emptyset$, then $N^+[v_i] = N^+[v_j]$, and, equivalently, if $N^-[v_i] \cap N^-[v_j] \neq \emptyset$, then $N^-[v_i] = N^-[v_j]$.*

DEFINITION 2.10. *An $n \times m$ matrix M is said to have* independent submatrices *M_1 and M_2 when, for every $1 \leq i, k \leq n$ and $1 \leq j, l \leq m$, if $M_{ij} \neq 0$ is an entry of M_1 and $M_{kl} \neq 0$ is an entry of M_2, then $i \neq k$ and $j \neq l$.*

THEOREM 2.11. *A specular and strongly quadrangular digraph is the digraph of a unitary matrix.*

Proof. Let D be a digraph. Note that if D is specular and strongly quadrangular, then $M(D)$ is composed of independent matrices. The theorem then follows from Lemma 2.6. □

The following theorem collects some classic results on line digraphs (see, e.g., [P96]).

THEOREM 2.12. *Let D be a digraph.*

(i) *Then, for every $(v_i, v_j) \in V(\overrightarrow{L}D)$,*

$$N^+[(v_i, v_j)] = N^+[v_j] \quad \text{and} \quad N^-[(v_i, v_j)] = N^-[v_i].$$

(ii) *A digraph D is a line digraph if and only if D is specular.*

(iii) *Let D be a strongly connected digraph. Then D is Eulerian if and only if $\overrightarrow{L}D$ is Hamiltonian.*

COROLLARY 2.13. *A strongly quadrangular line digraph is the digraph of a unitary matrix.*

Proof. The proof is obtained by point (i) of Theorem 2.12 together with Theorem 2.11. □

Remark 2.14. Not every line digraph that is the digraph of a unitary matrix is Eulerian. Let

$$M(D) = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad M(\overrightarrow{L}D) = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Note that $\overrightarrow{L}D$ is not Eulerian.

In a digraph, a *directed path of length r*, from v_1 to v_{r+1}, is a sequence of arcs of the form $(v_1, v_2), (v_2, v_3), \ldots, (v_r, v_{r+1})$. If the vertices and the arcs of a directed walk are all distinct, then the directed path is called a *Hamiltonian path*. A directed walk, in which $v_1 = v_{r+1}$, is called a *directed cycle*. A Hamiltonian path, in which $v_1 = v_{r+1} = v_n$ and $|V(D)| = n$, is called a *Hamiltonian cycle*. A digraph with a Hamiltonian cycle is said to be *Hamiltonian*.

THEOREM 2.15. *Let D be a digraph. Then $\overrightarrow{L}D$ is the digraph of a unitary matrix if and only if D is Eulerian or the disjoint union of Eulerian components.*

Proof. Suppose that $\overrightarrow{L}D$ is the digraph of a unitary matrix. By Corollary 2.13, $\overrightarrow{L}D$ is strongly quadrangular. If there exists $v_i \in V(\overrightarrow{L}D)$ such that $|N^+[v_i]| = 1$, then for every $v_j \in V(\overrightarrow{L}D)$, $N^+[v_i] \cap N^+[v_j] = \emptyset$. Suppose that, for every $v_i \in V(\overrightarrow{L}D)$, $|N^+[v_i]| = 1$. Since $\overrightarrow{L}D$ is strongly quadrangular, $A(D) = A(\overrightarrow{L}D)$ and is a permutation matrix. In general, for every $v_i \in V(\overrightarrow{L}D)$, if $|N^+[v_i]| = k > 1$, then there is a set $S \subset V(\overrightarrow{L}D)$ with $|S| = k - 1$ and not including v_i such that, for every $v_j \in S$, $N^+[v_j] = N^+[v_i]$. Writing $v_i = uv$, where $u, v \in V(D)$, by Theorem 2.12, $N^+[v_i] = N^+[v]$. It follows that $|N^+[v]| = k$. Then, because of S, it is easy to see that in $A(D)$ there are k arcs with head w. Hence $|N^+[v]| = |N^-[v]|$, and D is Eulerian. The proof of the sufficiency is immediate. □

COROLLARY 2.16. *Let D be a strongly connected digraph. Let $\overrightarrow{L}D$ be the digraph of a unitary matrix. Then $\overrightarrow{L}D$ is Hamiltonian.*

Proof. We obtain the proof by point (iii) of Theorem 2.12 together with Theorem 2.15. □

Let G be a group with generating set S. The *Cayley digraph* of G in respect to S is the digraph denoted by $Cay(G, S)$, with vertex set G and arc set including (g, h) if and only if there is a generator $s \in S$ such that $gs = h$.

COROLLARY 2.17. *The line digraph of a Cayley digraph is the digraph of a unitary matrix.*

Proof. The corollary follows from Theorem 2.15, since a Cayley digraph is regular. □

A *strongly regular graph* on n vertices is denoted by $srg(n, k, \lambda, \mu)$ and is a k-regular graph on n vertices, in which (1) two vertices are adjacent if and only if they have exactly λ common neighbors and (2) two vertices are nonadjacent if and only if they have exactly μ common neighbors (see, e.g., [CvL91]). The parameters of $srg(n, k, \lambda, \mu)$ satisfy the following equation: $k(k-\lambda-1) = (n-k-1)\mu$. The disjoint union of r complete graphs, each on m vertices, with $r, m > 1$, is denoted by rK_m. If $m = 2$, then rK_2 is called a *ladder graph*. A strongly regular graph is disconnected if and only if it is isomorphic to rK_m.

Remark 2.18. Not every strongly regular graph is the digraph of a unitary matrix. The graph $srg(10, 3, 0, 1)$ is called Petersen's graph. It is easy to check that $srg(10, 3, 0, 1)$ is not quadrangular.

Remark 2.19. By Theorem 2.11, if a digraph D is permutation equivalent to a disconnected strongly quadrangular graph, then D is the digraph of a unitary matrix.

The *complement* of a digraph D is a digraph denoted by $\overline{D}$ with the same vertex set of D and with two vertices adjacent if and only if the vertices are not adjacent in D. A digraph D is *self-complementary* if $D \cong \overline{D}$.

Remark 2.20. The fact that D is the digraph of a unitary matrix does not imply that $\overline{D}$ is. The digraph used in the proof of Proposition 2.22 provides a counterexample. Note that this does not hold in the case where D is self-complementary.

A digraph D is an *n-path* if $V(D) = \{v_1, v_2, \ldots, v_n\}$ and

$$A(D) = \{(v_1, v_2), (v_2, v_1), (v_2, v_3), (v_3, v_2), \ldots, (v_{n-1}, v_n), (v_n, v_{n-1})\},$$

where all the vertices are distinct. An n-path, in which $v_1 = v_n$, is called an *n-cycle*. A digraph D is a *directed n-cycle* if

$$A(D) = \{(v_1, v_2), (v_2, v_3), \ldots, (v_{n-1}, v_1)\}.$$

A connected (strongly connected) digraph that is disconnected (connected) if an arc is deleted is called a *directed tree* (*tree*).

PROPOSITION 2.21. *Let D be a digraph. If D is permutation equivalent to an n-path, then it is not the digraph of a unitary matrix.*

Proof. A digraph is strongly connected if and only if it is the digraph of an irreducible matrix. Since an n-path is strongly connected, it is the digraph of an irreducible matrix. Note that the number of arcs of an n-path is $2(n-1)$. The proposition is proved by Lemma 2.4 together with the following result (see, e.g., [BR91]). Let M be an irreducible matrix of size n and with exactly $2(n-1)$ nonzero

entries. Then there is a permutation matrix P such that

$$PMP^{\intercal} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 & 1 \\ 1 & a_{22} & \cdots & 0 & 0 \\ \vdots & 1 & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 & a_{nn} \end{bmatrix},$$

where a_{ii} can be equal to zero or one. It is easy to see that for any choice of the diagonal entries the digraph of $PMP^{\intercal}$ is not quadrangular. □

PROPOSITION 2.22. *Let D be a digraph. If D is permutation equivalent to one of the following digraphs, then D is not the digraph of a unitary matrix: n-path with a loop at each vertex, n-cycle, directed tree, or tree.*

Proof. When we choose any labeling of D, the proposition follows from Lemmas 2.4 and 2.5. □

Acknowledgments. The author thanks Peter Cameron, Richard Jozsa, Gregor Tanner, and Andreas Winter for their help.

REFERENCES

[ABNVW01] A. AMBAINIS, E. BACH, A. NAYAK, A. VISHWANATH, AND J. WATROUS, *One-dimensional quantum walks,* in Proceedings of the 33rd ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 50–59.

[AKV01] A. AMBAINIS, J. KEMPE, AND U. VAZIRANI, *Quantum walks on graphs,* in Proceedings of the 33rd ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 60–69.

[BR91] R. A. BRUALDI AND H. J. RYSER, *Combinatorial Matrix Theory*, Encyclopedia of Mathematics and Its Applications 39, Cambridge University Press, Cambridge, UK, 1991.

[CvL91] P. J. CAMERON AND J. H. VAN LINT, *Designs, Graphs, Codes and Their Links*, London Mathematical Society Student Texts 22, Cambridge University Press, Cambridge, UK, 1991.

[GZ98] P. M. GIBSON AND G.-H. ZHANG, *Combinatorially orthogonal matrices and related graphs,* Linear Algebra Appl., 282 (1998), pp. 83–95.

[GZe88] G. GRÖSSINGER AND A. ZEILINGER, *Quantum cellular automata,* Complex Systems, 2 (1988), pp. 197–208.

[G88] S. GUDDER, *Quantum Probability*, Academic Press, New York, 1988.

[L93] L. LOVÁSZ, *Random walks on graphs: A survey,* in Combinatorics, Paul Erdös is Eighty, Vol. 2, Bolyai Soc. Math. Stud. 2, János Bolyai Math. Soc., Budapest, 1996, pp. 353–397.

[M96] D. MEYER, *From quantum cellular automata to quantum lattice gases,* J. Statist. Phys., 85 (1996), pp. 551–574.

[P96] E. PRISNER, *Line graphs and generalizations: A survey,* in Surveys in Graph Theory, Congr. Numer. 116, G. Chartrand and M. Jacobson, eds., University of Manitoba, Winnipeg, Manitoba, Canada, 1996, pp. 193–230.

[T99] A. TERRAS, *Fourier Analysis on Finite Groups and Applications*, London Mathematical Society Student Texts 43, Cambridge University Press, Cambridge, UK, 1999.

[T84] W. T. TUTTE, *Graph Theory*, Encyclopedia of Mathematics and Its Applications 21, Cambridge University Press, Cambridge, UK, 1984.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 301–351

AN ORTHOGONAL HIGH RELATIVE ACCURACY ALGORITHM FOR THE SYMMETRIC EIGENPROBLEM*

FROILÁN M. DOPICO†, JUAN M. MOLERA†, AND JULIO MORO†

Abstract. We propose a new algorithm for the symmetric eigenproblem that computes eigenvalues and eigenvectors with high relative accuracy for the largest class of symmetric, definite and indefinite, matrices known so far. The algorithm is divided into two stages: the first one computes a singular value decomposition (SVD) with high relative accuracy, and the second one obtains eigenvalues and eigenvectors from singular values and vectors. The SVD, used as a first stage, is responsible both for the wide applicability of the algorithm and for being able to use only orthogonal transformations, unlike previous algorithms in the literature. Theory, a complete error analysis, and numerical experiments are presented.

Key words. symmetric eigenproblem, singular value decomposition, high relative accuracy

AMS subject classifications. 65F15, 65G50, 15A18

DOI. 10.1137/10.1137/S089547980139371X

1. Introduction.

An *orthogonal spectral decomposition* of a real symmetric $n \times n$ matrix A is a factorization $A = Q\,\Lambda\,Q^T$, where Q is real orthogonal and $\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_n]$ is diagonal. We assume that $\lambda_1 \geq \cdots \geq \lambda_n$. The columns q_i, $i = 1, \ldots, n$, of Q are the eigenvectors of A corresponding to the eigenvalues λ_i, $i = 1, \ldots, n$. In this paper we present an algorithm that computes an orthogonal spectral decomposition for the largest class (so far) of symmetric matrices with the following *high relative accuracy*:

- The error in each computed eigenvalue, $\widehat{\lambda}_i$, is

$$|\lambda_i - \widehat{\lambda}_i| = O(\kappa\,\epsilon)|\lambda_i|, \tag{1}$$

  where we assume that $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$, ϵ is the unit roundoff of the finite arithmetic employed in the computation and κ is a relevant condition number, usually of order $O(1)$, to be specified later in section 2.1.
- The angle $\Theta(q_i, \widehat{q}_i)$ between each computed eigenvector $\widehat{q}_i$ and the exact one q_i satisfies

$$\Theta(q_i, \widehat{q}_i) = \frac{O(\kappa\,\epsilon)}{relgap^*(|\lambda_i|)}, \tag{2}$$

  where

$$relgap^*(|\lambda_i|) = \min\left\{ \min_{\substack{j \in \mathcal{S} \\ j \neq i}} \left| \frac{|\lambda_j| - |\lambda_i|}{\lambda_i} \right|, 1 \right\}$$

  and the index set $\mathcal{S}$ is equal to $\{1, \ldots, n\}$ unless the eigenvalue, say λ_{j_0}, whose

*Received by the editors August 10, 2001; accepted for publication (in revised form) by I.C.F. Ipsen March 3, 2003; published electronically August 19, 2003. This research was partially supported by the Ministerio de Ciencia y Tecnología of Spain through grant BFM-2000-0008.
http://www.siam.org/journals/simax/25-2/39371.html
†Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Spain (dopico@math.uc3m.es, molera@math.uc3m.es, jmoro@math.uc3m.es).

absolute value is closest to $|\lambda_i|$ has opposite sign to λ_i. In that case, $\mathcal{S}$ is obtained from $\{1, \ldots, n\}$ by removing j_0 and the index k of any other eigenvalue with the sign of λ_{j_0} satisfying $|\lambda_{j_0} - \lambda_k| \leq O(\kappa\epsilon)|\lambda_{j_0}|$. In plain words, we remove from $\mathcal{S}$ the indices corresponding to eigenvalues with opposite sign to λ_i whose absolute value is closest to $|\lambda_i|$.

Expression (2) depends on the quantity $relgap^*$, not on the eigenvalue relative gap

$$relgap(\lambda_i) = \min\left\{\min_{j\neq i} \frac{|\lambda_j - \lambda_i|}{|\lambda_i|},\, 1\right\} \tag{3}$$

one would naturally expect. The reason is that the eigenvectors are computed via the singular value decomposition (SVD), which is closely related to the spectral decomposition for symmetric matrices. Postprocessing the singular vectors produces eigenvectors with the accuracy (2). At the cost of worsening this bound in a few cases, the error in the eigenvectors can be written in terms of (3): we will show in section 5 that the error is

$$\Theta(q_i, \widehat{q}_i) = \frac{O(\kappa\,\epsilon)}{relgap(\lambda_i)} \tag{4}$$

except in the case when λ_i and λ_{j_0}, the eigenvalue whose absolute value is closest to $|\lambda_i|$, have opposite sign, and $|\lambda_{j_0}|$ is much closer to $|\lambda_i|$ than any other $|\lambda_j|$ with $\lambda_j\lambda_i > 0$. In that case,

$$\Theta(q_i, \widehat{q}_i) = \frac{O(\kappa\,\epsilon)}{\min\{relgap(\lambda_{j_0}), relgap(\lambda_i)\}}. \tag{5}$$

For the sake of simplicity, both bounds (4) and (5) have been presented in their simplest forms, when no clusters of eigenvalues with close absolute values are present. General bounds, valid in the presence of clusters, will be derived in section 5 for bases of invariant subspaces.

Equations (1), (2) may allow a considerably more accurate outcome than that of a conventional eigenvalue method, such as QR, divide-and-conquer, or bisection with inverse iteration. Such algorithms produce results with high *absolute* accuracy, i.e., satisfying

$$|\lambda_i - \widehat{\lambda}_i| = O(\epsilon)\, \max_j |\lambda_j|,$$

instead of (1), and

$$\Theta(q_i, \widehat{q}_i) = \frac{O(\epsilon)}{\frac{\min_{j\neq i} |\lambda_i - \lambda_j|}{\max_j |\lambda_j|}},$$

instead of (2). Thus, a conventional algorithm may provide approximations for the small eigenvalues ($\frac{\max_j |\lambda_j|}{|\lambda_i|} \sim \frac{1}{\epsilon}$) with no correct significant digits, or even with the wrong sign. Moreover, if there are two or more small eigenvalues, their eigenvectors may be computed very inaccurately, even when the eigenvalues are well separated in the relative sense (e.g., $\lambda_i = 10^{-15}$ and $\lambda_j = 10^{-16}$ if $\lambda_1 = 1$). At present, high relative accuracy can be reached only for certain classes of *symmetric* matrices.

Identifying classes of matrices for which either an SVD or a spectral decomposition can be computed with high relative accuracy has been a very active area of

research in the last 15 years (see [6] and references therein for an overview). So far, high relative accuracy eigensolvers are available only for some symmetric matrices and are far less developed than accurate SVD algorithms (except, of course, in the related positive definite case [7]). To be more precise, several easily characterized classes of matrices have been identified in [6] for which high relative accuracy SVDs can be computed, while present symmetric indefinite eigensolvers deliver high relative accuracy for matrices which are not easy to recognize (with the exception of scaled diagonally dominant matrices [2]). As can be seen in [22, 27], the symmetric indefinite matrices deserving high relative accuracy spectral decompositions have been characterized through the structure of their positive semidefinite polar factors. This structure, however, is very difficult to relate with the structure of the matrix itself. In this regard, *the main contribution of the present paper is to prove that the proposed eigensolver achieves high relative accuracy* (1), (2) *for all symmetric matrices in any of the classes identified in* [6]. Moreover, it will do so, under very general assumptions, for any class of matrices eventually identified in the future for which high relative accuracy SVDs can be computed. To our knowledge, none of the present symmetric eigensolvers can guarantee high relative accuracy for the classes of matrices above.

The basic motivation for the algorithm we propose is to take advantage of the present knowledge of several classes of matrices for which an SVD can be computed with high relative accuracy (see [6] for a unified approach). The connection with our work lies in that the SVD and the spectral decomposition are closely related for symmetric[1] matrices: the singular values are the absolute values of the eigenvalues, and eigenvectors may be constructed from singular vectors. To be more precise, let $A = U\Sigma V^T$ be an SVD of $A = A^T$, where U, V are $n \times n$ orthogonal with columns u_i, v_i, $i = 1, \ldots, n$, and $\Sigma = \mathrm{diag}[\sigma_1, \ldots, \sigma_n]$ with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. In the simplest (and most frequent) case in which all singular values of A are distinct, the eigenvalues of A are

$$(v_i^T u_i)\, \sigma_i, \tag{6}$$

with $v_i^T u_i = \pm 1$ for all $i = 1, \ldots, n$, and the corresponding eigenvectors are v_i (u_i may be equally chosen). Hence, once an SVD is known, the only additional work to obtain the eigenvalues is to determine the sign ± 1 via the scalar product $v_i^T u_i$ of right and left singular vectors (the general case when groups of equal singular values appear is discussed in section 3.1). Notice that $v_i^T A v_i = v_i^T u_i \sigma_i$; i.e., the scalar product above can be thought of as a cheaper and indirect way of obtaining the sign from the Rayleigh quotient, avoiding the multiplication by the matrix A, which may give the wrong sign due to its large condition number (one example of this difficulty will be shown at the end of section 3.3). In fact, this particular way of assigning the signs through $v_i^T u_i$, together with the proof of its accuracy, is one of the crucial issues in this paper.

Therefore, given a computed high relative accuracy SVD of $A = A^T$ satisfying

$$|\sigma_i - \widehat{\sigma}_i| = O(\kappa\epsilon)\, |\sigma_i|, \tag{7}$$

$$\Theta(v_i, \widehat{v}_i) = \frac{O(\kappa\epsilon)}{relgap(\sigma_i)}, \qquad \Theta(u_i, \widehat{u}_i) = \frac{O(\kappa\epsilon)}{relgap(\sigma_i)}, \tag{8}$$

[1] All the results in this paper are valid for Hermitian matrices, although for the sake of simplicity we restrict the discussion to the real symmetric case.

with

$$relgap(\sigma_i) = \min\left\{\min_{j\neq i}\frac{|\sigma_i-\sigma_j|}{\sigma_i},\, 1\right\}, \tag{9}$$

if we prove that the pair $\widehat{v}_i$, $\widehat{u}_i$ approximates the pair v_i, u_i closely enough so that the computed value of the scalar product approximates ± 1 with an *absolute* error smaller than one (notice that this is no longer a high relative accuracy problem), then we will achieve the accuracy (1). For the eigenvectors this naive approach leads to $\Theta(q_i, \widehat{q}_i) = O(\kappa\epsilon)/relgap(\sigma_i)$, which can be improved to (2) by processing the singular vectors as described in section 5.

In this spirit we propose the following two-stage procedure to compute the eigenvalues and eigenvectors of a symmetric matrix:

Stage 1. Compute an SVD of A with accuracy (7) and (8).

Stage 2. Compute the eigenvalues of A by giving signs, according to (6), to the singular values computed in Stage 1. The corresponding eigenvectors are the right (or left) singular vectors computed in Stage 1. When groups of equal singular values are present, this step becomes more involved (see section 3.3 below).

We will show that Stage 2 provides high relative accuracy in the eigenvalues (1) and in the eigenvectors (2) as long as Stage 1 gives an SVD with small backward multiplicative error (as in formula (17) below, that in turn guarantees (7) and (8)). As to Stage 1, there are at present algorithms to perform it for several classes of matrices, summarized in [6]. These are the algorithms we are going to use, although any future high relative accuracy SVD algorithm may be employed for Stage 1.

One of the most remarkable contributions of Demmel et al. in [6] is the development of algorithms which compute high relative accuracy SVDs (i.e., satisfying (7) and (8)) for *any* matrix such that a so-called *rank-revealing decomposition* (RRD) can be computed with enough accuracy. An RRD of $G \in \mathbb{R}^{m\times n}$, $m \geq n$, is a factorization $G = \mathcal{X}\mathcal{D}\mathcal{Y}^T$ with $\mathcal{D} \in \mathbb{R}^{r\times r}$ diagonal and nonsingular and $\mathcal{X} \in \mathbb{R}^{m\times r}$, $\mathcal{Y} \in \mathbb{R}^{n\times r}$, where both matrices $\mathcal{X}$, $\mathcal{Y}$ have full column rank and are well conditioned (notice that this implies $r = \mathrm{rank}(G)$). According to the structure of the algorithms in [6], a more specific description of the *signed SVD (SSVD) algorithm* we propose here is the following.

ALGORITHM 1. (SSVD)

`Input:` Symmetric matrix A.

`Output:` Eigenvalues $\Lambda = \mathrm{diag}[\lambda_i]$ and eigenvectors $Q = [q_1 \ldots q_n]$; $A = Q\Lambda Q^T$.

1. Compute an RRD factorization XDY^T of A.
2. Compute SVD $XDY^T = U\Sigma V^T$ of RRD using algorithms from [6, section 3].
3. Compute the eigenvalues and eigenvectors of A from singular values and singular vectors using Algorithm 3 (see section 5).

We warn the reader that, before presenting Algorithm 3, we will discuss a simpler implementation of step 3 of Algorithm 1 which follows straightforwardly the ideas explained after (6). This procedure, Algorithm 2 (see section 3.3), is introduced for the sake of clarity; understanding Algorithm 3 is not easy starting from scratch, but it is elementary once the error analysis for Algorithm 2 is done in section 4. We will see there that Algorithm 2 delivers the announced accuracy (1) for eigenvalues but, in some cases, computes eigenvectors less accurately than (2). However, the error bound we obtain for eigenvectors suggests a modification in the eigenvector computation which, maintaining the validity of the error analysis, improves the accuracy in the eigenvectors to (2). This modification leads to Algorithm 3. We stress that both

versions compute the same eigenvalues and differ only in the eigenvector computation step, which is more accurate for Algorithm 3.

The accuracy required in [6] on the *computed* RRD matrices X, D, Y to guarantee that a high relative accuracy SVD can be obtained is given by the following forward error bounds:

$$\begin{aligned} |d_{ii} - \delta_{ii}| &= O(\epsilon)|\delta_{ii}|, \\ \|X - \mathcal{X}\| &= O(\epsilon)\|\mathcal{X}\|, \\ \|Y - \mathcal{Y}\| &= O(\epsilon)\|\mathcal{Y}\|, \end{aligned} \tag{10}$$

where $\|\cdot\|$ denotes the spectral norm and d_{ii}, δ_{ii} denote, respectively, the diagonal elements of D, $\mathcal{D}$. Once an RRD factorization XDY^T satisfying (10) is available, either Algorithm 3.1 or Algorithm 3.2 of [6] provides a high relative accuracy SVD of XDY^T with overall relative error (including the initial factorization stage) of order $O\left(\epsilon \max\{\kappa(X), \kappa(Y)\}\right)$ in the singular values, and $O\left(\epsilon \max\{\kappa(X), \kappa(Y)\}\right)$ over the relative gap (9) in the singular vectors, where $\kappa(\cdot)$ denotes the condition number in the spectral norm. The key to proving this high relative accuracy is that both the error (10) in the factorization and the errors introduced either by Algorithm 3.1 or by Algorithm 3.2 of [6] produce a backward error of multiplicative type, instead of the additive type usually produced by conventional algorithms (see section 2.1 for a more detailed discussion).

Several classes of matrices have been found in the last 10 years for which it is possible to compute an accurate RRD. They include bidiagonal, acyclic, Cauchy, Vandermonde, graded, and scaled diagonally dominant matrices, as well as all well-scalable symmetric positive definite matrices, some well-scalable symmetric indefinite matrices, and many others. Hence, for all symmetric matrices in any of the classes described in [6, pp. 26–27], Algorithm 1 will produce a spectral decomposition with the high relative accuracy given by (1) and (2) under the criteria given in [6] for computing accurate RRDs.

So far, the only general algorithm to compute high relative accuracy spectral decompositions of symmetric indefinite matrices is the so-called *implicit J-orthogonal* algorithm. It was introduced by Veselić in [26] and carefully analyzed by Slapničar in [22]. This algorithm begins by computing a *symmetric indefinite factorization* SJS^T of the matrix $A = A^T$, where J is square diagonal with diagonal elements ± 1, and S has full column rank.[2] If this factorization is computed with enough accuracy, the J-orthogonal algorithm yields the eigenvalues with relative error of order $O(\tilde{\kappa}\epsilon)$ for an appropriate condition number $\tilde{\kappa}$ which has been observed in practice to be of order $O(1)$. The eigenvectors are computed with error

$$\Theta(q_i, \widehat{q}_i) = \frac{O(\tilde{\kappa}\epsilon)}{relgap(\lambda_i)} \tag{11}$$

depending on the natural eigenvalue relative gap (3). This accuracy is better than the one obtained by Algorithm 1 in those cases in which the eigenvalue sign distribution is the one described right before (5). This is an advantage with respect to the algorithm proposed here. However, it should be stressed that, in view of both (4) and (5), whenever Algorithm 1 computes an eigenvector with error bound larger than the bound for

[2]Notice that, although SJS^T is not an RRD, its computation is equivalent to computing a symmetric RRD of the form XDX^T; see [23].

the J-orthogonal one, it must be due to the presence of some small eigenvalue relative gap. Thus, some other eigenvector is computed by the J-orthogonal algorithm with an error bound of similar magnitude. An illustrative example displaying this behavior will be shown in Experiment 4 in section 6.2.

An important advantage of Algorithm 1 over the J-orthogonal algorithm is that the latter does not guarantee high relative accuracy for the classes of symmetric matrices discussed in [6]. The reason is that RRDs with the accuracy (10) are obtained in [6] via Gaussian elimination with complete pivoting (GECP).[3] Moreover, a plain implementation of GECP does not guarantee accuracy (10) for most of the classes in [6]. This can be achieved only through special, nontrivial implementations of GECP, sometimes demanding a great deal of ingenuity (see [6, 5]). Since GECP leads, in general, to RRDs with $X \neq Y$, even if the matrix to be factorized is symmetric, the J-orthogonal algorithm cannot be directly applied because it begins with the *symmetric* indefinite factorization. Numerical experiments show that the usual algorithm [23] to compute the symmetric indefinite factorization does not provide, in general, the required accuracy for the symmetric matrices in those classes demanding special implementations of GECP. At present it is not known whether some modifications in the algorithm for the symmetric indefinite factorization would ensure that it is accurately computed in the sense of (10) for these matrices.

There are other important differences between the algorithm by Veselić and Slapničar and the one proposed below: the J-orthogonal algorithm uses *hyperbolic* transformations [17, section 12.5.4], which complicates the error analysis and increases the constants in the error bounds. The algorithm we propose here uses only *orthogonal* transformations. Also, the error bounds for the hyperbolic J-orthogonal algorithm are valid modulo a minor proviso (bounded growth of the scaled condition number of certain matrices appearing in each step of the iteration), while the new algorithm can be implemented in such a way that no proviso is needed to guarantee the error bounds. On the other hand, the J-orthogonal algorithm may be easily extended to matrix pencils, while this is not possible for the one presented here. There are also similarities: both algorithms require a previous factorization of the matrix, and both crucially depend on employing algorithms of one-sided Jacobi type.

Notice that the nonsymmetric character of Algorithm 1 is responsible both for making it valid for a large class of matrices and for being able to use only orthogonal transformations in step 2. The price to pay, however, is that by applying an SVD algorithm (valid for any matrix) to a symmetric matrix, we are not making any use of the symmetry of A. Thus, the algorithm is not backward stable, in the sense that one cannot guarantee that the computed eigenvalues and eigenvectors are the exact eigenvalues and eigenvectors of a close *symmetric* matrix. This is why Algorithm 1 produces an error bound in the eigenvectors which does not depend on the relative gap between the eigenvalues. This does not happen if we use a symmetric algorithm (such as the J-orthogonal algorithm) producing a *symmetric* backward error, since in that case the relative perturbation theory for symmetric matrices [16, 20, 27] leads to (11).

Concerning the computational cost of Algorithm 1, it is $O(n^3)$ provided the initial RRD costs $O(n^3)$ (some classes of matrices allow an accurate RRD, but not at $O(n^3)$ cost [6]). As is usual for high accuracy algorithms, Algorithm 1 is more expensive

[3] Some mention is also made in [6] of using QR with complete pivoting. This would open the possibility of using Algorithm 3.3 of [6], which is less costly than Algorithms 3.1–3.2 for step 2 of Algorithm 1.

than other $O(n^3)$ conventional eigenvalue methods, such as QR, divide-and-conquer, etc. The most expensive part of Algorithm 1 is the one-sided Jacobi method employed in step 2. However, some ways have been recently found [14] to speed up one-sided Jacobi which make it nearly as fast as the QR algorithm for SVD.

It is difficult to compare the cost of Algorithm 1 with that of the J-orthogonal algorithm. If in both cases we do not count the initial factorization, the difference between Algorithm 3.1 of [6] and Algorithm 3.3.1 of [22] seems to amount to two matrix multiplications and one QR factorization. However, numerical experience indicates that Algorithm 3.1 of [6] requires less Jacobi sweeps than Algorithm 3.3.1 of [22] (see section 6.2). Finally, step 3 of Algorithm 1 costs, in general, $O(n^2)$, but for every cluster with d close singular values corresponding to eigenvalues of both signs, and if eigenvectors need to be computed, there is an overhead cost of $O(d^3)+O(d^2n)$. Clearly, this is maximized when only one cluster of size $d = n$ is present. Then, the cost of step 3 is $O(n^3)$. As to the timing statistics, the run-times of both algorithms are comparable according to the numerical experiments below.

Both the comments on the computational cost and the numerical experiments in section 6.2 apply to a plain implementation of the one-sided Jacobi SVD algorithm included in Algorithm 3.1 of [6]. At present, fast and sophisticated implementations of the one-sided Jacobi SVD algorithm are being developed by Z. Drmač along the lines of [14]. We have tested a preliminary version of this routine in a few numerical experiments, and with this optimized Jacobi, Algorithm 1 was much faster than the J-orthogonal algorithm. Extensive numerical experiments will be done in the future.

The rest of the paper is organized as follows. Section 2 collects the mathematical results required to perform a complete error analysis of Algorithm 1. Section 3 describes in detail Algorithm 2, a preliminary implementation for step 3 of Algorithm 1, including the corresponding pseudocode. Section 4 contains a complete error analysis of a first, simpler implementation of Algorithm 1, using Algorithm 2 in step 3. This is done in the most general setting, allowing for the presence of clusters, which is why an entire section is devoted to discussing the error analysis. Otherwise, if the matrix has well-separated singular values, the error analysis is straightforward. We remind the reader that there are two reasons for doing the error analysis on this preliminary implementation: first, this error analysis gives the idea of how to design the final Algorithm 3 for step 3 of Algorithm 1. The second reason is that, once the error analysis is done with Algorithm 2, no new error analysis is required for Algorithm 3. Section 5 is devoted to developing and analyzing Algorithm 3, proving the error bounds (2), (4), and (5) in the most general setting, with any distribution of clusters. To keep the presentation within limits, most of the proofs in section 5 have been omitted (see [10] for complete proofs). However, in order to give a hint of the ideas and techniques employed we include in an appendix the proof of Theorem 5.7, one of the main results in section 5. Section 6 addresses the practical implementation of Algorithm 1, together with the numerical tests. Conclusions and discussion of open problems are presented in section 7.

2. Preliminary results. We collect in this section the mathematical results required to perform the error analysis of Algorithm 1. As stated in the introduction, the only requirement on the high relative accuracy SVD algorithm in step 2 of Algorithm 1 is producing a small multiplicative backward error when performed in finite arithmetic. A precise statement is given in section 2.1 for algorithms in [6]. We also show in section 2.1 that the error due to the initial RRD can be absorbed as an additional multiplicative backward error. Section 2.2 summarizes the multiplicative

perturbation theory for singular values and for bases of singular subspaces needed to guarantee the high relative accuracy of the overall algorithm.

2.1. Backward error of the SVD algorithm. The following theorem is essentially proved in [6].

THEOREM 2.1. *Algorithm* 3.1 *of* [6] *(see Algorithm* 4 *in section* 6.1 *below) produces a multiplicative backward error when executed with machine precision* ϵ*; i.e., if* $G = XDY^T \in \mathbb{R}^{m\times n}$ *is the RRD computed in step* 1 *of Algorithm* 1 *and* $\widehat{U}\widehat{\Sigma}\widehat{V}^T$ *is the SVD computed by the algorithm, then there exist matrices* $U' \in \mathbb{R}^{m\times r}$, $V' \in \mathbb{R}^{n\times r}$, $E \in \mathbb{R}^{m\times m}$, $F \in \mathbb{R}^{n\times n}$ *such that* U' *and* V' *have orthonormal columns,*

$$\begin{aligned} \|U' - \widehat{U}\| = O(\epsilon), \quad \|V' - \widehat{V}\| = O(\epsilon), \\ \|E\| = O(\epsilon\kappa(X)), \quad \|F\| = O(\epsilon\kappa(R')\kappa(Y)), \end{aligned} \tag{12}$$

where R' *is the best conditioned row diagonal scaling of the triangular matrix* R *appearing in* step 1 *of Algorithm* 3.1 *of* [6] *and*

$$(I + E)G(I + F) = U'\widehat{\Sigma}V'^T. \tag{13}$$

It is proved in [6] that $\kappa(R')$ is at most of order $O(n^{3/2}\kappa(X))$, but in practice we have observed in extensive numerical tests that $\kappa(R')$ behaves as $O(n)$. One can get rid of the factor $\kappa(R')$ at the price of using the more costly Algorithm 3.2 of [6].

We state Theorem 2.1 because the original result [6, Thm. 3.1] is not phrased as a backward error result, which is what we need for the subsequent error analysis. The only missing piece in the analysis of [6] is the fact that one-sided Jacobi [17, section 8.6.3] produces a small multiplicative backward error. This can be easily derived from Proposition 3.13 in [13] and, since it is not central to our argument, we omit its proof, together with that of Theorem 2.1. A full proof of both results will appear elsewhere [11] (and can be found in [10, Appendix A]). Two different versions of Algorithm 3.1 of [6] are analyzed in [11], depending on whether the right- or left-handed version of one-sided Jacobi is employed. One can show that the right-handed version, i.e., the one in which the Jacobi rotations are applied from the right, guarantees smaller error bounds and leads precisely to Theorem 2.1. For the left-handed version one can prove a result similar to Theorem 2.1, but with a weaker error bound for F, and requiring a minor proviso to ensure the accuracy. However, the left-handed version is still the one usually employed in practice since it is much faster and no significant difference has ever been observed in accuracy. This is why we use it in most of the experiments in section 6. Finally, it is crucial for the accuracy of one-sided Jacobi algorithms to impose as a stopping criterion that the cosines of the angles between the different columns (or rows, depending on the version of one-sided Jacobi) be smaller than ϵ times the dimension of the matrix.

Once the backward error of the SVD algorithm is shown to be multiplicative, the perturbation theory in section 2.2 below can be used to prove high relative accuracy, namely that the computed singular values and vectors of XDY^T satisfy

$$\begin{aligned} |\sigma_i - \widehat{\sigma}_i| &= O(\kappa\,\epsilon)\,\sigma_i, \\ \Theta(v_i, \widehat{v}_i) &= \frac{O(\kappa\,\epsilon)}{relgap(\sigma_i)}, \\ \Theta(u_i, \widehat{u}_i) &= \frac{O(\kappa\,\epsilon)}{relgap(\sigma_i)}, \end{aligned} \tag{14}$$

where

$$\kappa = \kappa(R') \max\{\kappa(X), \kappa(Y)\} \tag{15}$$

is the relevant condition number announced in the introduction.

As a matter of fact, one may even absorb into a backward error of the form (13) the error in the initial RRD, i.e., the one due to the fact that the SVD computation does not start from the symmetric matrix A itself but from its computed RRD: let $A = \mathcal{X}\mathcal{D}\mathcal{Y}^T$ be an exact RRD factorization of A and assume the starting decomposition XDY^T has been computed accurately enough so that the computed matrices X, D, Y satisfy conditions (10). Then, as shown in the proof of Theorem 2.1 in [6], there exist matrices E_f, F_f with $\|E_f\| = O(\epsilon\kappa(X))$, $\|F_f\| = O(\epsilon\kappa(Y))$ such that

$$(I + E_f)A(I + F_f) = XDY^T. \tag{16}$$

This, together with (13), implies that

$$U'\widehat{\Sigma}V'^T = (I + \widetilde{E})A(I + \widetilde{F}), \tag{17}$$

where the backward errors $\widetilde{E}$, $\widetilde{F}$ are of size $\|\widetilde{E}\| = O(\epsilon\kappa(X))$, $\|\widetilde{F}\| = O(\epsilon\kappa(R')\kappa(Y))$ and reflect that the errors produced by both the RRD factorization and the SVD algorithm are backward multiplicative.

We stress that all our error analysis is done in terms of the backward errors $\|\widetilde{E}\|$ and $\|\widetilde{F}\|$. Although we have focused on the case when $\|E_f\| = O(\epsilon\kappa(X))$ and $\|F_f\| = O(\epsilon\kappa(Y))$, any other more general backward errors for the factorization step can be trivially incorporated into the error analysis, since, up to first order,

$$\|\widetilde{E}\| \le \|E_f\| + O(\epsilon\kappa(X)), \quad \|\widetilde{F}\| \le \|F_f\| + O(\epsilon\kappa(R')\kappa(Y)).$$

2.2. Multiplicative perturbation theory. Let G be a real $m \times n$ matrix with SVD $G = U\Sigma V^T$ and singular values $\sigma_1 \ge \sigma_2 \ge \cdots$. We consider a multiplicative perturbation $\widetilde{G} = (I + E)G(I + F)$ of G with SVD $\widetilde{G} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ and singular values $\widetilde{\sigma}_i$, also ordered decreasingly.

THEOREM 2.2 (exactly Theorem 3.1 of [16]). *Let* $G \in \mathbb{R}^{m\times n}$, $\widetilde{G} = (I+E)G(I+F)$, *and set*

$$\eta = \max\{\|E\|, \|F\|\}, \qquad \eta' = 2\eta + \eta^2. \tag{18}$$

Then

$$\frac{|\,\sigma_i - \widetilde{\sigma}_i|}{\sigma_i} \le \eta'.$$

In addition to the change in the singular values, we also need to estimate the changes undergone by singular subspaces or, more precisely, by their bases. Although the following results are valid for rectangular matrices (see [20, 8]), we state them in the square case, the only case we deal with in section 4. Thus, G is now a real $n \times n$ matrix and $\widetilde{G} = (I + E)G(I + F)$. Let

$$G = [U_1\ U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}, \tag{19}$$

$$\widetilde{G} = \begin{bmatrix} \widetilde{U}_1\ \widetilde{U}_2 \end{bmatrix} \begin{bmatrix} \widetilde{\Sigma}_1 & 0 \\ 0 & \widetilde{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \widetilde{V}_1^T \\ \widetilde{V}_2^T \end{bmatrix} \tag{20}$$

be two conformally partitioned SVDs of G and $\widetilde{G}$; i.e., the four matrices $\Sigma_1, \widetilde{\Sigma}_1 \in \mathbb{R}^{q\times q}$ and $\Sigma_2, \widetilde{\Sigma}_2 \in \mathbb{R}^{(n-q)\times(n-q)}$ are diagonal. No particular order is assumed on the singular values. The change in the singular subspaces is usually measured through the sines of the canonical angles $\Theta(U_1, \widetilde{U}_1)$ between the column spaces of U_1 and $\widetilde{U}_1$, and $\Theta(V_1, \widetilde{V}_1)$ between the column spaces of V_1 and $\widetilde{V}_1$ (see [25]). It is well known that this change is governed (see, e.g., [20, Thm. 4.1]) by the singular value relative gap

$$rg(\Sigma_1, \widetilde{\Sigma}_2) = \min_{\substack{\sigma\in\sigma(\Sigma_1)\\ \widetilde{\sigma}\in\sigma(\widetilde{\Sigma}_2)}} \frac{|\sigma - \widetilde{\sigma}|}{\widetilde{\sigma}}, \tag{21}$$

where $\sigma(M)$ denotes the set of singular values of the matrix M.

This kind of result, however, is not enough for our purposes. The fact that the signs of the eigenvalues are obtained through scalar products like the one in (6) forces us to accurately compute not only the singular subspaces but also the corresponding *simultaneous* bases U_i and V_i. To ensure this, finer perturbation results are needed, dealing with the sensitivity of the bases themselves. It has been observed in [8] that simultaneous bases of singular subspaces do not have the same sensitivity under perturbation as their corresponding singular subspaces. More precisely, bases may be much more sensitive to *additive* perturbations than singular subspaces. Fortunately enough for our purposes, both sensitivities are essentially equal for multiplicative perturbations. A detailed discussion of these issues may be found in [8, 9], including a stronger version of the following result (we use the Frobenius norm $\|\cdot\|_F$, as is usual when the dimension of the subspaces is larger than 1).

THEOREM 2.3 (exactly Theorem 2.2 of [8]). *Let $G \in \mathbb{R}^{n\times n}$ and $\widetilde{G} = (I + E)G(I + F)$ with respective SVDs* (19) *and* (20). *Then there exists an orthogonal matrix $P \in \mathbb{R}^{q\times q}$ such that*

$$\sqrt{\|U_1P - \widetilde{U}_1\|_F^2 + \|V_1P - \widetilde{V}_1\|_F^2} \le 2\sqrt{q}\left[\eta + \frac{\eta'}{1-\eta}\ \frac{1}{relgap(\Sigma_1, \widetilde{\Sigma}_2)}\right], \tag{22}$$

where $relgap(\Sigma_1, \widetilde{\Sigma}_2)$ is given by

$$relgap(\Sigma_1, \widetilde{\Sigma}_2) = \min\{rg(\Sigma_1, \widetilde{\Sigma}_2), 1\}, \tag{23}$$

and η, η' are given by (18).

Although it is more usual in the literature [6, 5] to define the relative gap (21) with the roles of Σ_1 and Σ_2 reversed, we have chosen the definition above to conform to the cited perturbation theorems. However, this does not represent any significant difference in the error bounds, since a straightforward calculation shows that

$$2\, relgap(\widetilde{\Sigma}_2, \Sigma_1) \ge relgap(\Sigma_1, \widetilde{\Sigma}_2) \ge \frac{1}{2}\, relgap(\widetilde{\Sigma}_2, \Sigma_1). \tag{24}$$

On the other hand, as is usual in this kind of perturbation bounds, one can reformulate the definition of the gaps to make them depend only on the unperturbed singular values, at the cost of somewhat complicating the bounds.

The main point of Theorem 2.3 is that the orthogonal matrix P is the same for both left and right singular vectors. This will be enough to guarantee the accuracy of the sign assignment and of the computed bases of invariant subspaces.[4]

3. Computing spectral decompositions from SVDs. This section is divided into three parts. Section 3.1 outlines the mathematical basis for the main idea underlying Algorithm 1, namely that one can easily get a spectral decomposition of a symmetric matrix if one is given an SVD, even if the matrix has groups of equal singular values. Some practical details concerning clusters of close singular values in finite precision will be considered in section 3.2. The complete pseudocode for Algorithm 2 will be presented in section 3.3. This is the simplest implementation of step 3 in Algorithm 1.

3.1. Mathematical basis. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with SVD $A = U\Sigma V^T$. Then, $V^TAV = V^TU\Sigma$ is orthogonally similar to A with V^TU orthogonal. If A has distinct singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_p$ with respective multiplicities m_i, $i = 1, \ldots, p$ ($m_1 + \cdots + m_p = n$), and we partition U and V accordingly as

$$U = \left[\ \mathcal{U}_1 \mid \mathcal{U}_2 \mid \cdots \mid \mathcal{U}_p\ \right],$$
$$V = \left[\ \mathcal{V}_1 \mid \mathcal{V}_2 \mid \cdots \mid \mathcal{V}_p\ \right]$$

with $\mathcal{U}_i, \mathcal{V}_i \in \mathbb{R}^{n\times m_i}$ corresponding to each distinct singular value, then

$$\mathcal{V}_i^T\mathcal{U}_j = 0 \quad \text{whenever} \quad i \neq j \tag{25}$$

since, due to the symmetry of A, both its left and right singular vectors are eigenvectors of A^2. Consequently,

$$V^TU = \operatorname{diag}[\mathcal{V}_1^T\mathcal{U}_1, \ldots, \mathcal{V}_p^T\mathcal{U}_p] \tag{26}$$

is block-diagonal, where each diagonal block $\mathcal{V}_i^T\mathcal{U}_i \in \mathbb{R}^{m_i\times m_i}$ is itself orthogonal. Furthermore, since

$$V^TAV = \operatorname{diag}[\sigma_1\mathcal{V}_1^T\mathcal{U}_1, \ldots, \sigma_p\mathcal{V}_p^T\mathcal{U}_p] \tag{27}$$

is symmetric, we conclude that each $\mathcal{V}_i^T\mathcal{U}_i$ is not only orthogonal but also symmetric and its eigenvalues, ± 1, are precisely the signs of those eigenvalues of A having modulus σ_i. In the simplest case when $m_i = 1$, the eigenvalue is just $v_i^Tu_i\,\sigma_i$. In the general case, a simple calculation shows that if the spectrum of $\mathcal{V}_i^T\mathcal{U}_i$ contains m_i^+ eigenvalues equal to 1 and m_i^- equal to -1 ($m_i = m_i^+ + m_i^-$), then

$$m_i^{\pm} = \frac{m_i \pm \operatorname{trace}(\mathcal{V}_i^T\mathcal{U}_i)}{2}; \tag{28}$$

i.e., the multiplicity of the eigenvalues $\pm\sigma_i$ can be easily recovered from the trace of $\mathcal{V}_i^T\mathcal{U}_i$.

[4] Actually, Theorem 2.3 is stronger than the usual bounds on the canonical angles between singular subspaces, since one can easily show that $\|\sin(\Theta(U_1, \widetilde{U}_1))\|_F \leq \|U_1\,P - \widetilde{U}_1\|_F$, which holds similarly for V_1.

To obtain the eigenvectors of A, the simplest (and more frequent) case corresponds to $m_i = 1$. In that case, the right singular vector v_i itself is an eigenvector. When some m_i is larger than 1 and $\mathrm{trace}(\mathcal{V}_i^T\mathcal{U}_i) = m_i$ (resp., $\mathrm{trace}(\mathcal{V}_i^T\mathcal{U}_i) = -m_i$), the m_i eigenvalues are all equal to σ_i (resp., $-\sigma_i$), and the eigenvectors are the columns of $\mathcal{V}_i$. In the general case $m_i > 1$, $m_i \neq m_i^{\pm}$, consider for each $i = 1, \ldots, p$ an orthogonal diagonalization of $\mathcal{V}_i^T\mathcal{U}_i = \mathcal{W}_i J_i \mathcal{W}_i^T$, with $J_i = \mathrm{diag}[I_{m_i^+}, -I_{m_i^-}]$ and $\mathcal{W}_i = [\mathcal{W}_i^+ \,|\, \mathcal{W}_i^-] \in \mathbb{R}^{m_i \times m_i}$ partitioned conformally to J_i. Then, denoting $\mathcal{W} = \mathrm{diag}[\mathcal{W}_1, \ldots, \mathcal{W}_p]$, one can easily check that the matrix $Q = V\mathcal{W}$ is such that

$$Q^T A Q = \mathrm{diag}[\sigma_1 J_1, \ldots, \sigma_p J_p];$$

i.e., the set of columns of the submatrix $Q_i^+ = \mathcal{V}_i\mathcal{W}_i^+ \in \mathbb{R}^{n \times m_i^+}$ (resp., $Q_i^- = \mathcal{V}_i\mathcal{W}_i^- \in \mathbb{R}^{n \times m_i^-}$) is a basis of the eigenspace corresponding to the eigenvalue σ_i (resp., $-\sigma_i$) of A. In other words, $A = Q\Lambda Q^T$ with $\Lambda = \mathrm{diag}[\pm\sigma_i]$ is a spectral decomposition of A.

We conclude by noting that, although the right singular vectors $\mathcal{V}_i$ have been used throughout the argument, the symmetry of A implies that similar results hold using instead the left singular vectors $\mathcal{U}_i$.

3.2. Clusters in finite arithmetic. We have seen how to deal theoretically with groups of equal singular values. When working in finite precision, however, it is unlikely that some of the singular values in the output of step 2 of Algorithm 1 come out equal. But at the same time the expected accuracy (14) determines that some of the singular values should be considered as numerically indistinguishable and treated in the spirit of section 3.1. Thus we are forced to deal with, say, k different groups Σ_i of n_i close singular values ($i = 1, \ldots, k$, $n_1 + \cdots + n_k = n$), which we call *clusters*.[5] The criterion to divide the singular values into clusters is crucial for the final accuracy of Algorithm 1. This criterion will be carefully analyzed in section 4.4, where we show that to achieve the accuracy (1) (see Theorem 4.3) it is enough to include two contiguous singular values σ_j, σ_{j+1} in the same cluster whenever

$$\frac{|\sigma_j - \sigma_{j+1}|}{\sigma_j} \leq C\,\kappa\,\epsilon \tag{29}$$

for a suitable constant C, where

$$\kappa = \kappa(R')\max\{\kappa(X), \kappa(Y)\}$$

is the quantity (15) which came up in the error bound for the singular values computed in step 2 of Algorithm 1 (see section 4.4 for more on the choice of the constant C; we mention here that in the performed numerical experiments the choice $C = 1$ always gives very satisfactory results).

For each cluster Σ_i we take matrices U_i, $V_i \in \mathbb{R}^{n \times n_i}$ whose columns are, respectively, left and right singular vectors corresponding to the singular values in Σ_i. Since the singular values in Σ_i are, in general, different, each U_i and V_i is made up with several of the matrices $\mathcal{U}_j$ and $\mathcal{V}_j$ defined in section 3.1. Consequently, the products $\Delta_i = V_i^T U_i$ are symmetric, orthogonal, and block-diagonal matrices whose diagonal blocks are some of the blocks $\mathcal{V}_j^T\mathcal{U}_j$.

[5] For the sake of brevity, we use Σ_i to denote both the cluster of singular values and the corresponding $n_i \times n_i$ diagonal matrix.

We conclude by noting that the numbers n_i^+ of positive and n_i^- of negative eigenvalues with absolute values in the cluster Σ_i are still given by a formula such as (28). As to the eigenvectors, things are different from section 3.1, since the diagonalization of Δ_i does not lead, in general, to eigenvectors but just to two orthonormal bases, one for the invariant subspace corresponding to the positive eigenvalues in the cluster Σ_i and another for the negative ones. This is a fundamental issue in the error analysis for the eigenvector computations and will be carefully explained throughout the proof of Theorem 4.4.

3.3. A first version of step 3 of Algorithm 1. In this section we describe Algorithm 2, the first implementation of step 3 in Algorithm 1. The eigenvalue and the eigenvector computations are separated in the procedure into two independent parts. Doing this helps us to better understand the structure of Algorithm 3, our final implementation of step 3 in Algorithm 1, which will only insert a different cluster selection routine in between the eigenvalue and the eigenvector computations.

ALGORITHM 2.

Input: SVD of a symmetric matrix $A = U\Sigma V^T$.
Output: Eigenvalues $\Lambda = \mathrm{diag}[\lambda_i]$ and eigenvectors $Q = [q_1 \dots q_n]$; $A = Q\Lambda Q^T$.

1. Decide the singular value clusters, $\Sigma_i = \{\sigma_{i_0}, \dots, \sigma_{i_0+n_i-1}\}$, U_i, V_i, $i = 1, \dots, k$, according to (29).
2. Compute the eigenvalues using Algorithm 2.1 below.
3. Compute the eigenvectors using Algorithm 2.2 below.

ALGORITHM 2.1.

Input: SVD of $A = U\Sigma V^T$; Clusters $\Sigma_1, \Sigma_2, \dots, \Sigma_k$.
Output: Eigenvalues Λ.

1. **for** each cluster, $i = 1 : k$
2. compute the diagonal elements of $\Delta_i = V_i^T U_i$
3. **if** $n_i = 1$ **then**
4. $\lambda_{i_0} = sign(\Delta_i)\sigma_{i_0}$
5. **else**
6. **for** $j = i_0 : i_0 + n_i - 1$
7. $\lambda_j = sign[(\Delta_i)_{jj}]\sigma_j$
8. **endfor**
9. $t_i = \mathrm{trace}(\Delta_i)$, $n_i^- = \frac{n_i - t_i}{2}$
10. **if** $\#\{(\Delta_i)_{jj} < 0\} \neq n_i^-$ **then**
11. **for** $j = i_0 : i_0 + n_i^- - 1$
12. $\lambda_j = -\sigma_j$
13. **endfor**
14. **for** $j = i_0 + n_i^- : i_0 + n_i - 1$
15. $\lambda_j = \sigma_j$
16. **endfor**
17. **endif**
18. **endif**
19. **endfor**

ALGORITHM 2.2.

Input: SVD of $A = U\Sigma V^T$; Clusters $\Sigma_1, \Sigma_2, \dots, \Sigma_k$; Eigenvalues Λ.
Output: Eigenvectors $Q = [q_1 \dots q_n]$.
Notation: $Q_i^{\pm}$ denotes the eigenvector matrix corresponding to positive (resp., negative) eigenvalues in Σ_i.

```
1. for each cluster, i = 1 : k
2.     if n_i = 1 then
3.         q_{i_0} = v_{i_0}
4.     else
5.         n_i^- ≡ number of negative eigenvalues in Σ_i
6.         if n_i^- = 0 then
7.             Q_i^+ = V_i
8.         elseif n_i^- = n_i then
9.             Q_i^- = V_i
10.        else
11.            multiply Δ_i = V_i^T U_i
12.            diagonalize Δ_i = [W_i^+ W_i^-] J_i [W_i^+ W_i^-]^T
13.            Q_i^+ = V_i W_i^+,   Q_i^- = V_i W_i^-
14.        endif
15.    endif
16. endfor
```

Some comments on this code are in order. First, we have singled out the case $n_i = 1$, although it is not needed. This is done to highlight the fact that Algorithm 2 is extremely simple in this case, with all complications coming from the case $n_i > 1$.

Notice also that the code does not compute eigenvectors associated with zero eigenvalues in the case where $r = \mathrm{rank}(A) < n$. This is due to the fact that the SVD algorithms in [6] do not compute null vectors. However, if accurate null vectors are needed, they can be obtained as the last $n - r$ columns of the orthogonal factor in a complete QR factorization of the matrix V of right singular vectors.

If large clusters are present, one can save flops in steps 11 and 13 of Algorithm 2.2 by employing Strassen multiplication without spoiling the accuracy of the overall algorithm. As to the diagonalization step, step 12 of Algorithm 2.2, it is assumed that one performs it on a symmetrization of Δ_i. This is crucial to obtain orthonormal eigenvectors.

Notice that the eigenvalue sign assignment (steps 6–17 of Algorithm 2.1) is done in two stages when there are clusters: First (steps 6–8), we assign the signs given by the diagonal elements of $\Delta_i = V_i^T U_i$ as if the singular values in Σ_i were not a cluster. If the number of assigned negative eigenvalues coincides with $n_i^- = \frac{n_i - \mathrm{trace}(\Delta_i)}{2}$, the signs are kept. Otherwise, we proceed as described in steps 10–17 of Algorithm 2.1. The reason for this is that the random sign assignment inside each cluster in steps 10–17 proved to be too pessimistic in practice: although singular values inside each cluster are numerically indistinguishable according to (14), actual errors are frequently smaller than the error bounds. These smaller errors are lost if the signs of eigenvalues are randomly assigned. The modified procedure minimizes this loss of accuracy.

We finish this section with an interesting remark on the way the signs are assigned in Algorithm 2. One might think of obtaining the sign of each eigenvalue from the Rayleigh quotients $v_i^T A v_i$, one of the most common ways of approximating eigenvalues, instead of from $v_i^T u_i$. However, it is easy to construct examples for which the sign of $v_i^T A v_i$ is wrong, while the sign of $v_i^T u_i$ is right. We propose the following numerical example, easily reproducible in MATLAB 5.3: Generate a 100×100 symmetric Cauchy matrix with parameters $x_i = y_i \equiv r_i$, $i = 1 : 100$, where r_i is a random number chosen from a normal distribution with mean zero and variance one. Scale this matrix on both sides by the same diagonal matrix with diagonal elements $d_i = 10^{20 r_i'}$, where r_i' is a random number chosen from a uniform distribution on the interval $(0.0, 1.0)$. For

matrices of this kind Algorithm 3 in [5] can be used to obtain in a very simple way an RRD, $A = XDY^T$, with forward errors fulfilling (10). Finally, applying Algorithm 3.1 of [6] to this RRD yields an SVD of A with high relative accuracy. No clusters of singular values are present in general. For several of the computed singular vectors neither $v_i^T A v_i$ nor $(v_i^T X)D(Y^T v_i)$ have the same sign of $v_i^T u_i$, which is the correct one, as will be shown in section 4 (the reader also can check this by using a symbolic package such as Mathematica in very high precision). This example shows that using Rayleigh quotients may be dangerous, even in the case when the matrix is given as an RRD. Similar behavior is not rare in other Cauchy matrices or in random RRDs with very ill-conditioned diagonals. The use of Rayleigh quotients in the more favorable case when the matrix A is scaled in a certain particular way is covered in [15].

4. Error analysis. In this section we present the rounding error analysis for the eigenvalues and the eigenvectors computed by Algorithm 1 using Algorithm 2 in step 3. This error analysis remains valid for Algorithm 1 using Algorithm 3 in step 3: this is trivially true for the eigenvalues, since both versions of Algorithm 1 compute the same eigenvalues. It is also true for the eigenvectors, due to the generality of the error analysis, which allows us to use the new clusters of singular values appearing in Algorithm 3.

We stress that the error analysis applies *to the entire* Algorithm 1, since it relies on the backward multiplicative error formula (17), which absorbs the errors of the initial factorization in step 1. Although we focus on the case when the RRD is computed with the error (10), which ensures $\|E_f\| = O(\epsilon\kappa(X))$ and $\|F_f\| = O(\epsilon\kappa(Y))$, any other more general backward errors for the factorization step can be trivially incorporated into the error analysis, as explained at the end of section 2.1.

The main results in this section are the forward error bounds in Theorems 4.3 and 4.7. Both are expressed in big-O notation, without explicitly specifying the dimensional constants involved. There are two reasons for this. First, we rely on error bounds in [6], which are written in big-O notation without explicit mention of the constants. Second, it is well known that the precise value of the constant is, in general, not relevant for practical purposes.

This said, the reader should be aware that in the statements of the theorems in this section we absorb moderately growing functions of the dimensions (either n, of the whole matrix, or n_i, of the clusters) as constants inside the $O(\kappa\epsilon)$. Since none of them exceeds a moderate number times n^2, we choose not to write them explicitly in order not to complicate further the error bounds. However, the interested reader may find those corresponding to step 3 of Algorithm 1 explicitly stated in the proofs.

The error analysis is performed in the most general case when clusters of singular values are present. This somewhat complicates the analysis, which is almost straightforward in the simple (and most likely) case of matrices whose singular values are distinct enough. The practical criterion to decide when two singular values belong to the same cluster is also discussed in detail.

In the rest of this section we only deal with the error in nonzero eigenvalues and the corresponding eigenvectors. If the original matrix is singular, the number of zero eigenvalues is determined exactly, provided an RRD factorization fulfilling (10) is computed. As to the null vectors, it can be shown that they can be computed with error $O(\epsilon\,\kappa(R')\,\max\{\kappa(X), \kappa(Y)\})$ using the method already described following Algorithm 2.2. The relative gap does not appear because in this case it is equal to one.

We begin by fixing our model for floating point arithmetic and the notation.

4.1. Model of arithmetic. We use the conventional error model for floating point arithmetic,

$$\texttt{fl}(a \odot b) = (a \odot b)(1+\delta), \tag{30}$$

where a and b are real floating point numbers, $\odot \in \{+, -, \times, /\}$, and $|\delta| \leq \epsilon$, where ϵ is the machine precision. Moreover, we assume that neither overflow nor underflow occurs. We stress that the results proved in this section still hold under a weaker error model valid for arithmetic with no guard digit.

The error analysis below also remains valid for complex Hermitian matrices, since [18, Chapter 3] the equality (30) continues to hold for complex numbers with δ a small complex number bounded by $|\delta| = O(\epsilon)$. However, in order to simplify the presentation we consider only real symmetric matrices.

Finally, we will commit a slight abuse of notation, denoting by $\texttt{fl}(expr)$ the computed result in finite precision of expression $expr$, instead of its rigorous meaning of the closest floating point number to $expr$.

4.2. Notation. Letters with hats denote computed quantities appearing in any step of Algorithm 1. The same letters without hats denote their exact counterparts. It is assumed that the input of Algorithm 1 is a real symmetric $n \times n$ matrix A, for which an RRD factorization XDY^T with small multiplicative backward error (16) can be computed.

We assume that k different clusters $\widehat{\Sigma}_i$ of n_i $(n_1 + \cdots + n_k = n)$ close singular values are identified through criterion (29); thus, the usual decreasing order on singular values determines the unknown exact clusters Σ_i. The singular values of one particular cluster are supposed to be different from the singular values of any other cluster. Given an index $i \in \{1, \ldots, k\}$, we define

$$\Sigma_{\bar{i}} = \bigcup_{j \neq i} \Sigma_j. \tag{31}$$

For each cluster Σ_i we take matrices U_i, $V_i \in \mathbb{R}^{n \times n_i}$ whose columns are, respectively, left and right singular vectors corresponding to the singular values in Σ_i. Recall that the singular values in Σ_i may be different, so both U_i and V_i will, in general, contain singular vectors corresponding to different singular values. Therefore, the remarks in section 3.2 apply.

Many nontrivial choices are possible for the exact quantities U_i, V_i if A has multiple singular values in Σ_i. In that case, the results proved in this section are valid for *any* possible choice of U_i and V_i, provided their columns are singular vectors and not simply bases of the corresponding singular subspaces.

4.3. Fundamental lemma. The following lemma, which is a simple consequence of the fundamental perturbation theorem, Theorem 2.3, and the multiplicative backward error formula (17) for steps 1 and 2 of Algorithm 1, is the starting point of our error analysis. For the sake of brevity, the quantities K_i will be defined inside Lemma 4.1. These quantities play a relevant role in the error analysis.

LEMMA 4.1. *Let $\widehat{U}_i, \widehat{V}_i \in \mathbb{R}^{n \times n_i}$ be the matrices of computed left and right singular vectors corresponding to the cluster of singular values $\widehat{\Sigma}_i$ computed by steps* 1–2 *of Algorithm* 1 *applied to the symmetric matrix A. Let U_i, V_i, Σ_i be their exact*

counterparts. Then, there exists an exact orthogonal matrix P_i such that

$$K_i \equiv \sqrt{\|U_iP_i - \widehat{U}_i\|_F^2 + \|V_iP_i - \widehat{V}_i\|_F^2} \leq \frac{O(\kappa\,\epsilon)}{relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})} \tag{32}$$

with κ given by (15)*.*

Proof. Let U_i', V_i' be the submatrices corresponding to $\widehat{\Sigma}_i$ of the exact orthogonal matrices U' and V' appearing in (17). Then, Theorem 2.3 applied to (17) guarantees that there exists an orthogonal $n_i \times n_i$ matrix P_i such that

$$\sqrt{\|U_iP_i - U_i'\|_F^2 + \|V_iP_i - V_i'\|_F^2} = \left\| \begin{bmatrix} U_iP_i - U_i' \\ V_iP_i - V_i' \end{bmatrix} \right\|_F \leq \frac{O(\kappa\,\epsilon)}{relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})}.$$

Notice that

$$\sqrt{\|U_iP_i - \widehat{U}_i\|_F^2 + \|V_iP_i - \widehat{V}_i\|_F^2} = \left\| \begin{bmatrix} U_iP_i - \widehat{U}_i \\ V_iP_i - \widehat{V}_i \end{bmatrix} \right\|_F,$$

so the triangular inequality implies

$$\sqrt{\|U_iP_i - \widehat{U}_i\|_F^2 + \|V_iP_i - \widehat{V}_i\|_F^2} \leq \left\| \begin{bmatrix} U_iP_i - U_i' \\ V_iP_i - V_i' \end{bmatrix} \right\|_F + \left\| \begin{bmatrix} U_i' - \widehat{U}_i \\ V_i' - \widehat{V}_i \end{bmatrix} \right\|_F.$$

The last term in the right-hand side of this inequality is $O(\epsilon)$ by (12). This concludes the proof. ∎

Lemma 4.1 gives a forward error bound for simultaneous orthonormal bases of singular subspaces, which depends only on the quantities $\|\widetilde{E}\|$ and $\|\widetilde{F}\|$ appearing in (17). In other words, it only accounts for errors corresponding to steps 1 and 2 of Algorithm 1, i.e., to the SVD computation.

The rest of the bounds obtained in this section, i.e., those corresponding to step 3 of Algorithm 1, depend, for each cluster, on the quantities K_i on the left-hand side of (32). This allows us to write all subsequent error bounds as a function of K_i and to trace how each of the steps in Algorithm 2 contributes to the final error. From now on we assume that all quantities K_i for $i = 1, \ldots, k$ are sufficiently smaller than 1, which, according to Lemma 4.1, is the case whenever the clusters of singular values are properly chosen. More precisely, all we need is that K_i be small enough to make all bounds in sections 4.4 and 4.5 strictly smaller than one.

4.4. Error bounds for eigenvalues and cluster criterion. We begin by analyzing the error produced in the computation of $\mathrm{trace}(V_i^T U_i)$ using the standard inner product algorithm.

LEMMA 4.2. *Let $\widehat{U}_i, \widehat{V}_i \in \mathbb{R}^{n \times n_i}$ be the matrices of computed left and right singular vectors corresponding to the cluster of singular values $\widehat{\Sigma}_i$ computed by steps* 1–2 *of Algorithm* 1 *applied to the symmetric matrix A. Let U_i, V_i, Σ_i be their exact counterparts. Then,*

$$\begin{aligned} \left| \mathtt{fl}(\,\mathrm{trace}(\mathtt{fl}(\,\widehat{V}_i^T \widehat{U}_i\,))\,) - \mathrm{trace}(\,V_i^T U_i\,) \right| &\leq \sqrt{n_i}\left(\sqrt{2}K_i + \frac{K_i^2}{2}\right) + O(\epsilon) \\ &\leq \frac{O(\kappa\,\epsilon)}{relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})} \end{aligned} \tag{33}$$

with κ given by (15) *and K_i by* (32)*.*

Proof. First observe that

$$\left|\mathtt{fl}(\,\mathrm{trace}(\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,))\,) - \mathrm{trace}(\,V_i^TU_i\,)\right| \le \left|\mathtt{fl}(\,\mathrm{trace}(\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,))\,) - \mathrm{trace}(\,\widehat{V}_i^T\widehat{U}_i\,)\right| + \left|\mathrm{trace}(\,\widehat{V}_i^T\widehat{U}_i\,) - \mathrm{trace}(\,V_i^TU_i\,)\right|. \tag{34}$$

Taking into account that the norm of the columns of $\widehat{U}_i$ and $\widehat{V}_i$ is close to one by (12), a straightforward error analysis [18, Chapter 3] shows that the first term in the right-hand side of inequality (34) is $n_i(n+n_i)\epsilon + O(\epsilon^2)$. If P_i is the orthogonal matrix appearing in Lemma 4.1, the last term fulfills

$$\begin{aligned}\left|\mathrm{trace}(\,\widehat{V}_i^T\widehat{U}_i\,) - \mathrm{trace}(\,V_i^TU_i\,)\right| &= \left|\mathrm{trace}(\,\widehat{V}_i^T\widehat{U}_i\,) - \mathrm{trace}(\,P_i^TV_i^TU_iP_i\,)\right| \\ &\le \sqrt{n_i}\sqrt{\sum_{k=1}^{n_i}\left|(\,\widehat{V}_i^T\widehat{U}_i - P_i^TV_i^TU_iP_i\,)_{kk}\right|^2} \\ &\le \sqrt{n_i}\,\|\widehat{V}_i^T\widehat{U}_i - (V_iP_i)^TU_iP_i\|_F. \end{aligned} \tag{35}$$

Now define matrices Δ_u and Δ_v such that

$$\widehat{U}_i = U_iP_i + \Delta_u \quad \text{and} \quad \widehat{V}_i = V_iP_i + \Delta_v. \tag{36}$$

Combining (35) and (36) yields

$$\left|\mathrm{trace}(\,\widehat{V}_i^T\widehat{U}_i\,) - \mathrm{trace}(\,V_i^TU_i\,)\right| \le \sqrt{n_i}\,(\,\|\Delta_u\|_F + \|\Delta_v\|_F + \|\Delta_u\|_F\|\Delta_v\|_F\,),$$

where we have used that $\|CD\|_F \le \|C\|_2\|D\|_F$ for any matrices C, D, together with the fact that the spectral norm of any matrix with orthonormal columns is one. Finally, setting $K_i = \sqrt{\|\Delta_u\|_F^2 + \|\Delta_v\|_F^2}$ as in (32), we obtain, after some direct manipulations, the desired result. ∎

Notice that $\mathrm{trace}\left(V_i^TU_i\right)$ may only take the integer values $-n_i, -n_i+2, \ldots, n_i-4, n_i-2, n_i$, since $V_i^TU_i$ is symmetric and orthogonal. Thus, it is sufficient that the error bound in (33) be less than one to compute *exactly* the value of $\mathrm{trace}\left(V_i^TU_i\right)$. This can be done by obtaining t_i, the nearest integer to $\mathtt{fl}(\,\mathrm{trace}(\,\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,)\,)\,)$ *with the parity of* n_i. Then, the *integer* computation (with integer variables) of $(n_i - t_i)/2$ yields n_i^-, the *exact number of negative eigenvalues* included in the cluster Σ_i of singular values. The exact number of positive eigenvalues is obtained from the integer computation of $n_i - n_i^-$.

We stress that the conditions

$$\left|\mathtt{fl}(\,\mathrm{trace}(\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,)\,)\,) - \mathrm{trace}(\,V_i^TU_i\,)\right| < 1, \qquad i = 1, \ldots, k, \tag{37}$$

which ensure that signs are correctly assigned, determine the cluster criterion to be used in Algorithm 2. Giving a rigorous criterion would require an exact knowledge of the constants involved in the big-O bound in (33), which in any case are too pessimistic in practice. Instead, we assume that the singular values in each cluster $\widehat{\Sigma}_i$ satisfy

$$relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}}) \approx relgap(\widehat{\Sigma}_i, \widehat{\Sigma}_{\bar{i}}) > C\epsilon\kappa(R')\max(\kappa(X), \kappa(Y))$$

for a suitable constant C. This can be obtained by defining that two contiguous singular values $\widehat{\sigma}_j \ge \widehat{\sigma}_{j+1}$ belong to the same cluster whenever

$$\frac{|\widehat{\sigma}_j - \widehat{\sigma}_{j+1}|}{\widehat{\sigma}_j} \le C\,\kappa\,\epsilon,$$

i.e., whenever condition (29) above holds. Choosing a large C ensures (37) and, as a consequence, that the number of positive/negative eigenvalues is correctly computed. However, a large value for C favors the mixing of different singular values in the same cluster and, since the signs are assigned more or less randomly within each cluster, the error bound in the eigenvalues becomes roughly the product of C times the bound in the singular values (see (14)). Therefore, the choice of C is subject to a certain trade-off. A sensible choice might be choosing C between 1 and 10. All the numerical experiments in section 6 have been done with $C = 1$ and the results are very satisfactory.

In any case, notice that, on one hand, the singular values are computed with the accuracy given by (17) and Theorem 2.2. On the other hand, their signs as eigenvalues of A are correctly assigned whenever the bound (33) is less than one. With this we have proved the main result of this subsection.

THEOREM 4.3. *Let A be an $n \times n$ real symmetric matrix for which it is possible to compute an RRD fulfilling* (10). *Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of A and $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n$ be the approximations to the eigenvalues of A computed by Algorithm* 1. *Let $\widehat{U}_i, \widehat{V}_i \in \mathbb{R}^{n \times n_i}$ be the matrices of computed left and right singular vectors corresponding to the cluster of computed singular values $\widehat{\Sigma}_i$, and let U_i, V_i, Σ_i be their exact counterparts. Assume that all clusters have been chosen according to* (29), *so that conditions* (37) *hold. Then*

$$|\lambda_j - \widehat{\lambda}_j| = |\lambda_j| \; O(\epsilon \kappa(R') \max(\kappa(X), \kappa(Y))), \qquad j = 1, \ldots, n. \tag{38}$$

The error bound (38) holds even for zero eigenvalues, since the *exact* number of zero eigenvalues of A is known once an RRD factorization satisfying (10) is available.

4.5. Error bounds for eigenvectors. In this section we obtain bounds on the distance between *bases* of invariant subspaces. Although it is more common to bound the sines of the canonical angles between the exact and the computed invariant subspaces [25], we choose to compare the bases themselves because, as explained before Theorem 2.3, bases play an essential role both in Algorithm 2 and in its error analysis. However, usual $\sin \Theta$ bounds easily follow from Theorem 4.7, since distances between bases and canonical angles between subspaces are closely related [25, Thms. I.5.2 and II.4.11] and the same bounds hold for both, up to a factor $\sqrt{2}$ in Frobenius norm.

One important issue in the subsequent analysis comes from step 12 of Algorithm 2.2 in which the $n_i \times n_i$ matrix $\widehat{V}_i^T \widehat{U}_i$ is orthogonally diagonalized for each cluster $\widehat{\Sigma}_i$. Lemma 4.1 shows that the matrices $\widehat{U}_i, \widehat{V}_i$ of computed singular vectors are not reliable approximations of the matrices of exact singular vectors U_i, V_i, but just reliable approximations of $U_i P_i$ and $V_i P_i$, with P_i the unknown $n_i \times n_i$ orthogonal matrix in Lemma 4.1. Hence, we are forced in practice to diagonalize approximations to matrices $P_i^T V_i^T U_i P_i$. Theorem 4.4 shows that this is enough to get orthonormal bases of invariant subspaces, although not for obtaining eigenvectors.

THEOREM 4.4. *Let A be a symmetric $n \times n$ matrix and $U_i, V_i \in \mathbb{R}^{n \times n_i}$ be matrices of left and right singular vectors of A corresponding to a cluster of nonzero singular values Σ_i, different from the rest of the singular values of A. Let P_i be any $n_i \times n_i$ orthogonal matrix, and consider any orthogonal diagonalization of the $n_i \times n_i$ orthogonal and symmetric matrix $P_i^T V_i^T U_i P_i$ partitioned as*

$$P_i^T V_i^T U_i P_i = [W_i^+ \; W_i^-] \begin{bmatrix} I_{n_i^+} & 0 \\ 0 & -I_{n_i^-} \end{bmatrix} [W_i^+ \; W_i^-]^T, \tag{39}$$

where I_s denotes the $s \times s$ identity matrix and $n_i^+ + n_i^- = n_i$. Then the columns of $V_i P_i W_i^+$ (resp., $V_i P_i W_i^-$) form an orthonormal basis of the invariant subspace of A corresponding to the positive (resp., negative) eigenvalues whose absolute values are in Σ_i.

Proof. Without loss of generality, we may consider the SVD of A partitioned in only two blocks,

$$A = [U_1\ U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} [V_1\ V_2]^T, \tag{40}$$

where no special order is assumed on the singular values. Here Σ_1 corresponds to the cluster Σ_i to be studied and Σ_2 corresponds to the remaining clusters $\Sigma_{\bar{i}}$ defined as in (31). The matrix P_i will be denoted just by P, and the matrices $W_i^{\pm}$ in (39) will be denoted by $W_{\pm}$.

As mentioned in section 3.2, $V_1^T U_1$ is orthogonal, symmetric, and block-diagonal with the size of the blocks fixed by the groups of equal singular values inside Σ_1. The matrix $V_1^T U_1 \Sigma_1$ is also symmetric with the same block-diagonal structure of $V_1^T U_1$. An orthogonal diagonalization for each block of $V_1^T U_1$ leads to an orthogonal diagonalization of the full matrix $V_1^T U_1$ with eigenvectors which are also eigenvectors of $V_1^T U_1 \Sigma_1$. In this situation, the eigenvectors of $V_1^T U_1$ corresponding to the eigenvalue 1 (resp., -1) are the eigenvectors of $V_1^T U_1 \Sigma_1$ corresponding to positive (resp., negative) eigenvalues with absolute values in Σ_1. From this we deduce that the invariant subspaces corresponding to positive (resp., negative) eigenvalues of matrices $P^T V_1^T U_1 P$ and $P^T V_1^T U_1 \Sigma_1 P$ coincide. Once this is taken into account, the rest of the proof reduces to some easy block manipulations.

Combining (40) and $V_2^T U_1 = 0$ from (25), we obtain

$$AV_1P = U_1\Sigma_1 P = [V_1\ V_2] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} U_1\Sigma_1 P = V_1 P (P^T V_1^T U_1 \Sigma_1 P). \tag{41}$$

Splitting the spectrum into positive and negative eigenvalues, we orthogonally diagonalize

$$P^T V_1^T U_1 \Sigma_1 P = [Q_+\ Q_-] \begin{bmatrix} D_+ & 0 \\ 0 & D_- \end{bmatrix} [Q_+\ Q_-]^T,$$

and from (41) we obtain

$$A(\,V_1PQ_+\,) = (\,V_1PQ_+\,)D_+ \quad \text{and} \quad A(\,V_1PQ_-\,) = (\,V_1PQ_-\,)D_-. \tag{42}$$

Now, we know that $\mathrm{col}(Q_{\pm}) = \mathrm{col}(W_{\pm})$, and since the columns of $Q_{\pm}$ and $W_{\pm}$ are orthonormal, there exist square orthogonal matrices $T_{\pm}$ such that $W_{\pm} = Q_{\pm} T_{\pm}$. Combining this and (42) we obtain

$$A(\,V_1PW_{\pm}\,) = (\,V_1PW_{\pm}\,)\,(\,T_{\pm}^T D_{\pm} T_{\pm}\,),$$

which proves the theorem. ∎

Once the previous theorem is proved, the rest of the section is organized into the following three steps.

1. Although Lemma 4.1 guarantees that $\widehat{U}_i$ and $\widehat{V}_i$ are close to U_iP_i and V_iP_i, provided the clusters have been properly chosen, this does not mean that $\widehat{\Delta}_i =$

$\mathtt{fl}(\widehat{V}_i^T\widehat{U}_i)$ in step 11 of Algorithm 2.2 is symmetric. Let $\widehat{S}_i$ be the symmetric matrix obtained by replacing the upper triangular part of $\widehat{\Delta}_i$ with its lower triangular part. Lemma 4.5 bounds the difference between $\widehat{S}_i$ and the exact symmetric matrix $P_i^T V_i^T U_i P_i$. Notice that if any driver routine of LAPACK [1] for the symmetric eigenvalue problem is used in step 12 of Algorithm 2.2, just the upper (or lower) triangular part of $\widehat{\Delta}_i$ is stored. Hence, the symmetrization step does not require any additional work.

2. Lemma 4.6 relates the computed orthogonal eigendecomposition of $\widehat{S}_i$ with an exact eigendecomposition of $P_i^T V_i^T U_i P_i$. It is shown that exact matrices $W_i^{\pm}$ in (39) can be chosen close enough to the corresponding computed matrices $\widehat{W}_i^{\pm}$ in step 12 of Algorithm 2.2.

3. Finally, the main theorem, Theorem 4.7, bounds the difference between the $n \times n_i^{\pm}$ matrices $\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{\pm})$ computed in step 13 of Algorithm 2.2 and some orthonormal bases of exact invariant subspaces of A. This result is a simple consequence of Lemmas 4.1 and 4.6.

The bottom line after these three steps is that step 3 of Algorithm 1 produces errors of the order of K_i, the quantity defined in (32), whose upper bound (32) depends only on the errors in steps 1 and 2 of Algorithm 1.

LEMMA 4.5. *Let $\widehat{U}_i, \widehat{V}_i \in \mathbb{R}^{n\times n_i}$ be the matrices of computed left and right singular vectors corresponding to the cluster of singular values $\widehat{\Sigma}_i$ computed by steps* 1–2 *of Algorithm* 1 *applied to the symmetric matrix A. Let U_i, V_i, Σ_i be their exact counterparts. Let $\widehat{S}_i$ be a symmetrization of the floating point matrix $\widehat{\Delta}_i = \mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,)$ obtained by replacing the upper triangular part of $\widehat{\Delta}_i$ with its lower triangular part, or vice versa. Then an orthogonal $n_i \times n_i$ matrix P_i exists such that*

$$\begin{aligned} \|\widehat{S}_i - P_i^T V_i^T U_i P_i\|_F &\le 2K_i + \frac{K_i^2}{\sqrt{2}} + O(\epsilon) \\ &\le \frac{O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))}{relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})}. \end{aligned} \tag{43}$$

Proof. First observe that

$$\|\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,) - P_i^T V_i^T U_i P_i\|_F \le \|\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,) - \widehat{V}_i^T\widehat{U}_i\|_F + \|\widehat{V}_i^T\widehat{U}_i - P_i^T V_i^T U_i P_i\|_F,$$

where P_i is the orthogonal matrix appearing in Lemma 4.1. Standard error analysis of usual matrix multiplication [18], and the fact that the columns of $\widehat{U}_i$ and $\widehat{V}_i$ are almost orthonormal by (12), show that the first term in the right hand-side of the previous inequality is bounded by $n\, n_i\epsilon + O(\epsilon^2)$. The last term can be bounded as in the proof of Lemma 4.2, so we obtain

$$\|\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,) - P_i^T V_i^T U_i P_i\|_F \le \left(\sqrt{2}K_i + \frac{K_i^2}{2}\right) + O(\epsilon).$$

We write $\mathtt{fl}(\,\widehat{V}_i^T\widehat{U}_i\,) = \widehat{L} + \widehat{D} + \widehat{R}$ as the sum of its strict lower triangular part, its diagonal part, and its strict upper triangular part. The same is done for the symmetric matrix $P_i^T V_i^T U_i P_i = L + D + L^T$, so the previous equation yields

$$\sqrt{\|(\widehat{L}+\widehat{D}) - (L+D)\|_F^2 + \|\widehat{R} - L^T\|_F^2} \le \left(\sqrt{2}K_i + \frac{K_i^2}{2}\right) + O(\epsilon). \tag{44}$$

The same inequality holds for $\sqrt{\|\widehat{L}-L\|_F^2+\|\widehat{D}+\widehat{R}-(D+L^T)\|_F^2}$. On the other hand

$$\|\widehat{S}_i - P_i^T V_i^T U_i P_i\|_F = \sqrt{\|(\widehat{L}+\widehat{D})-(L+D)\|_F^2+\|\widehat{L}^T-L^T\|_F^2}.$$

Combining this equation with (44) proves the lemma. ∎

Errors in the diagonalization step, step 12, of Algorithm 2.2 are now analyzed. Notation and definitions of the previous lemma are used.

LEMMA 4.6. *Let $\widehat{W}_i\widehat{\Lambda}_i\widehat{W}_i^T$ be the computed orthogonal spectral decomposition of the symmetric $n_i \times n_i$ matrix $\widehat{S}_i$ using any* LAPACK *subroutine for the symmetric eigenproblem* [1, *section* 2.3.4.1]. *Then, there exists a matrix E_i, an orthogonal matrix Z_i, and an orthogonal matrix P_i such that*

$$P_i^T V_i^T U_i P_i + E_i = Z_i \widehat{\Lambda}_i Z_i^T, \tag{45}$$

where

$$\|Z_i - \widehat{W}_i\|_2 \le O(\epsilon) \quad \text{and} \quad \|E_i\|_F \le 2K_i + \frac{K_i^2}{\sqrt{2}} + O(\epsilon). \tag{46}$$

Moreover, if $\widehat{W}_i^+$ (resp., $\widehat{W}_i^-$) is the submatrix of $\widehat{W}_i$ with columns corresponding to the positive (resp., negative) elements of $\widehat{\Lambda}_i$, then there exist matrices W_i^+, W_i^- fulfilling (39) *such that*

$$\begin{aligned} \|\widehat{W}_i^{\pm} - W_i^{\pm}\|_F &\le 2\sqrt{2}K_i + K_i^2 + O(\epsilon) \\ &= \frac{O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))}{relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})}. \end{aligned} \tag{47}$$

Proof. Using the results in [1, section 4.7.1] we see that there exist an orthogonal matrix Z_i and a matrix E_i' such that

$$\widehat{S}_i + E_i' = Z_i \widehat{\Lambda}_i Z_i^T, \tag{48}$$

where

$$\|Z_i - \widehat{W}_i\|_2 \le O(\epsilon) \quad \text{and} \quad \|E_i'\|_2 \le O(\epsilon)\|\widehat{S}_i\|_2.$$

Let P_i be the orthogonal matrix appearing in Lemmas 4.1 and 4.5. The spectral norm of the orthogonal matrix $P_i^T V_i^T U_i P_i$ is equal to one, so (43) implies $\|\widehat{S}_i\|_2 = 1+\beta$, with $|\beta| \le 2K_i + K_i^2/\sqrt{2} + O(\epsilon)$. Thus $\|E_i'\|_2 = O(\epsilon)$. Now, expressions (45) and (46) are easily proved using Lemma 4.5, noting by (48) that

$$P_i^T V_i^T U_i P_i + \widehat{S}_i - P_i^T V_i^T U_i P_i + E_i' = Z_i \widehat{\Lambda}_i Z_i^T,$$

and defining

$$E_i = \widehat{S}_i - P_i^T V_i^T U_i P_i + E_i'.$$

We finally prove (47). Let $W_i^{\pm}$ be matrices fulfilling (39) and Z_i^+ (resp., Z_i^-) be a submatrix of Z_i corresponding to the positive (resp., negative) elements of $\widehat{\Lambda}_i$. We assume that K_i is small enough to imply $\|E_i\|_2 < 1$, so the eigenvalues equal

to 1 (resp., -1) of $P_i^T V_i^T U_i P_i$ remain positive (resp., negative) in $\widehat{\Lambda}_i$. This can be seen by applying Weyl's eigenvalue perturbation theorem to (45) (see, for instance, [25, Corollary IV.4.10]). Thus, Davis and Kahan's $\sin\Theta$ theorem for variations of invariant subspaces of Hermitian matrices [4] applied to (45) leads to

$$\|\sin\Theta(W_i^+, Z_i^+)\|_F \le \frac{\|E_i\|_F}{\min\limits_{\substack{\mu<0\\ \mu\in\widehat{\Lambda}_i}} |1-\mu|} \le \|E_i\|_F, \tag{49}$$

where the matrix $\Theta(W_i^+, Z_i^+)$ is the matrix of the canonical angles between the column space of W_i^+ and the column space of Z_i^+. Theorem II.4.11 in [25], (49), and (46) show that it is possible to choose W_i^+ such that

$$\begin{aligned}\|W_i^+ - Z_i^+\|_F &= \sqrt{\|\sin\Theta(W_i^+, Z_i^+)\|_F^2 + \|I - \cos\Theta(W_i^+, Z_i^+)\|_F^2}\\ &\le \sqrt{2}\,\|\sin\Theta(W_i^+, Z_i^+)\|_F\\ &\le \sqrt{2}\,\|E_i\|_F\\ &\le 2\sqrt{2}K_i + K_i^2 + O(\epsilon). \qquad (50)\end{aligned}$$

Similar results hold for W_i^- and Z_i^-. We finish the proof by noting that

$$\|\widehat{W}_i^{\pm} - W_i^{\pm}\|_F \le \|\widehat{W}_i^{\pm} - Z_i^{\pm}\|_F + \|Z_i^{\pm} - W_i^{\pm}\|_F.$$

The first term of the right-hand side is $O(\epsilon)$ by (46), and the second one is bounded in (50). ∎

We conclude with the main result on rounding errors for eigenvectors computed in step 13 of Algorithm 2.2. Previous notation and definitions are used.

THEOREM 4.7. *Let A be an $n\times n$ real symmetric matrix of rank r for which it is possible to compute an RRD fulfilling* (10). *Let $\widehat{\Sigma}_i$ be a cluster of nonzero computed singular values of A using steps* 1–2 *of Algorithm* 1 *and Σ_i be the corresponding cluster of exact singular values. Then there exist matrices Q_i^+ and Q_i^-, whose columns form orthonormal bases of the invariant subspaces of A corresponding, respectively, to the positive and negative eigenvalues of A with absolute values in Σ_i, such that*

$$\begin{aligned}\|\mathtt{fl}(\,\widehat{V}_i\widehat{W}_i^+\,) - Q_i^+\|_F &\le (2\sqrt{2}+1)(K_i + K_i^2) + K_i^3 + O(\epsilon)\\ &= \frac{O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))}{relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})}, \qquad (51)\end{aligned}$$

with an equal bound for $\|\mathtt{fl}(\,\widehat{V}_i\widehat{W}_i^-\,) - Q_i^-\|_F$.

Moreover, let $\widehat{Q} = [\mathtt{fl}(\,\widehat{V}_1\widehat{W}_1^+\,)\,\mathtt{fl}(\,\widehat{V}_1\widehat{W}_1^-\,)\ldots\mathtt{fl}(\,\widehat{V}_k\widehat{W}_k^+\,)\,\mathtt{fl}(\,\widehat{V}_k\widehat{W}_k^-\,)]$ be the $n\times r$ matrix whose columns are the bases of all considered invariant subspaces of A computed using Algorithm 1. *Then there exists an $n\times r$ matrix B with exact orthonormal columns such that*

$$\|\widehat{Q} - B\|_F = O(\epsilon). \tag{52}$$

Proof. Let $\widehat{V}_i$ be the matrix of computed right singular vectors corresponding to the cluster $\widehat{\Sigma}_i$, and let V_i be its exact counterpart. Let $W_i^{\pm}$, $\widehat{W}_i^{\pm}$, and P_i be the matrices appearing in Lemmas 4.6 and 4.1. By Theorem 4.4, the columns of $Q_i^+ \equiv V_i P_i W_i^+$ and $Q_i^- \equiv V_i P_i W_i^-$ are orthonormal bases of the invariant subspaces

of A corresponding, respectively, to the positive and negative eigenvalues of A with absolute values in Σ_i.

Note also that

$$\|\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{\pm}) - V_iP_iW_i^{\pm}\|_F \leq \|\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{\pm}) - \widehat{V}_i\widehat{W}_i^{\pm}\|_F + \|\widehat{V}_i\widehat{W}_i^{\pm} - V_iP_iW_i^{\pm}\|_F. \tag{53}$$

The first term of the right-hand side is bounded by $n_i\sqrt{n_i\,n_i^{\pm}}\,\epsilon + O(\epsilon^2)$ using the standard error analysis of usual matrix multiplication [18] and the fact that the columns of $\widehat{V}_i$ and $\widehat{W}_i^{\pm}$ are almost orthonormal by (12) and (46). For the second term we proceed as follows: Define matrices Δ_v and $\Delta_w^{\pm}$ by

$$\widehat{V}_i = V_iP_i + \Delta_v \quad \text{and} \quad \widehat{W}_i^{\pm} = W_i^{\pm} + \Delta_w^{\pm},$$

where $\|\Delta_v\|_F \leq K_i$ by (32) and $\|\Delta_w^{\pm}\|_F \leq 2\sqrt{2}K_i + K_i^2 + O(\epsilon)$ by (47). Thus

$$\begin{aligned}\|\widehat{V}_i\widehat{W}_i^{\pm} - V_iP_iW_i^{\pm}\|_F &\leq \|\Delta_v\|_F + \|\Delta_w^{\pm}\|_F + \|\Delta_v\|_F\|\Delta_w^{\pm}\|_F \\ &\leq (2\sqrt{2}+1)(K_i + K_i^2) + K_i^3 + O(\epsilon).\end{aligned}$$

Combining this with (53) proves (51).

Finally, (52) follows from the well-known fact that finite precision matrix multiplication of matrices with columns orthonormal up to $O(\epsilon)$ yields a matrix with columns orthonormal up to $O(\epsilon)$. ∎

As announced in the introduction, the eigenvector error bounds we derive suffer from an important drawback: they depend on *relgap* (23) between singular values, which is less than or equal to the natural relative gap between eigenvalues, the one expected for the symmetric eigenproblem. This is an unavoidable consequence of the nonsymmetric character of Algorithm 1. This drawback, however, can be partially solved applying Theorem 4.7 to certain new singular value clusters chosen as described in section 5.

It is worth observing that Theorem 4.7 does not guarantee that the columns of the matrices $\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{\pm})$ computed by Algorithm 1 approximate *eigenvectors* of A. This can only be ensured in three cases: when there is no cluster ($n_i = 1$), when all eigenvalues in the cluster have the same sign, and when the cluster contains eigenvalues of both signs with either $n_i^+ = 1$ or $n_i^- = 1$. In this last case, either $\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{+})$ or $\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{-})$ approximates an eigenvector of A. In any other situation, the columns of $\mathtt{fl}(\widehat{V}_i\widehat{W}_i^{\pm})$ do not approximate eigenvectors but just orthonormal bases of the invariant subspaces of A corresponding to either the positive or the negative eigenvalues with absolute values in the cluster. However, provided the clusters of singular values are chosen according to criterion (29), this does not represent any drawback, because the eigenvectors in the corresponding invariant subspaces are computed by any symmetric eigensolver (including the J-orthogonal algorithm [26, 22]) with large errors due to the presence of very small relative gaps between the eigenvalues inside the clusters. There is no need to say that the J-orthogonal algorithm also computes accurate bases of invariant subspaces, due to its backward stability.

We conclude with an interesting remark concerning the discussion in the previous paragraph. Consider, for simplicity, that according to criterion (29) a cluster of two singular values, one corresponding to a positive eigenvalue and the other to a negative one, has been found. Then the bound in Theorem 4.7 implies that Algorithm 1 computes *both* eigenvectors with an error governed by the relative gap between the cluster and the singular values outside the cluster. This can be much larger than

the relative gap between the singular values inside the cluster. Thus, the presence of clusters reduces the errors in the computed eigenvectors. We will take more advantage of this property in section 5.

5. Computing more accurate eigenvectors. The error in the eigenvectors computed by Algorithm 2 is governed (see Theorem 4.7) by the singular value relative gap, which is less than or equal to the natural eigenvalue relative gap. We present in this section Algorithm 3, an implementation of step 3 of Algorithm 1, which computes eigenvectors with the error (2) (see also (4) and (5)) announced in the introduction. As we will see, the underlying idea is very simple and does not require a new error analysis but simply takes advantage of the generality of the one performed in section 4. We stress that the eigenvalue computation (steps 1–2 in Algorithm 2) will stay the same; only the computation of the eigenvectors will be modified. The general case, when clusters of singular values of arbitrary dimension are present, will be considered.

First, note that Algorithm 2 computes the eigenvalues before computing the eigenvectors. The relative error in the eigenvalues is of order $O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))$ provided the clusters are chosen according to criterion (29). A second important remark is that the error analysis performed in section 4 for the eigenvectors is independent of the error analysis for the eigenvalues, both being valid under the hypothesis that the quantities K_i defined in (32) are sufficiently small. As Lemma 4.1 shows, this is achieved by defining clusters which yield large enough $relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})$, but whenever this condition is fulfilled different clusters, i.e., different K_i, can be chosen to compute the eigenvectors using Algorithm 2.2. Theorem 4.7 still applies and will provide a smaller error bound whenever the new clusters *for the eigenvector computation* have larger *relgaps* than the ones chosen according to (29). Consequently we present the following algorithm that is the final version of step 3 of Algorithm 1.

ALGORITHM 3.

`Input: SVD of a symmetric matrix` $A = U\Sigma V^T$.

`Output: Eigenvalues` $\Lambda = \mathrm{diag}[\lambda_i]$ `and eigenvectors` $Q = [q_1 \dots q_n]$; $A = Q\Lambda Q^T$.

1. `Decide the singular value clusters,` $\{\Sigma_i, U_i, V_i\}_{i=1}^{k}$`, according to (29).`
2. `Compute the eigenvalues using Algorithm 2.1.`
3. `Use Algorithm 3.1 in section 5.2 to merge, when necessary, some pairs of clusters to form a new set` $\{\Sigma_i, U_i, V_i\}_{i=1}^{q}$ `of clusters, according to the strategy developed in this section.`
4. `Compute the eigenvectors using Algorithm 2.2 on the new set of clusters.`

The difference with respect to Algorithm 2 is the presence of step 3, in which a new selection of clusters is made. The limit for improving the bound (51) in Theorem 4.7 by increasing $relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})$ is naturally the eigenvalue relative gap. With this in mind, the idea to be implemented is very simple: Let Σ_i be one of the singular value clusters chosen according to (29), and let Λ_i^+ (resp., Λ_i^-) be the corresponding clusters of positive (resp., negative) eigenvalues with absolute values in Σ_i. Then $relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})$ can be much worse than the minimum of the two eigenvalue relative gaps associated to Σ_i only in the case in which Σ_i is *signed* (all the eigenvalues of the same sign), and the closest (in the relative sense) cluster to Σ_i, let us say $\Sigma_{cl(i)}$, is oppositely signed. Without loss of generality, it can be assumed that $\Sigma_i = \Lambda_i^+$; therefore $\Sigma_{cl(i)} = -\Lambda_{cl(i)}^-$. If Σ_i and $\Sigma_{cl(i)}$ are joined to form a new cluster $\Lambda_i^+ \cup (-\Lambda_{cl(i)}^-)$ with a larger *relgap*, the bound (51) will improve *separately* for the bases of *exactly the same two invariant*

subspaces associated with Λ_i^+ and $\Lambda_{cl(i)}^-$, computed by Algorithm 2.2 applied to the new set of clusters. Therefore, nothing is lost by merging clusters of this kind, and the error bound (51) can improve by *joining* close adjacent clusters in such a way that $relgap$ increases.

It will be seen that in the other cases it is not necessary to join clusters, either because the singular value relative gap is already of the same order of the eigenvalue relative gap, or because joining clusters would mean increasing the number of eigenvalues of the same sign in the cluster, and consequently Algorithm 2.2 would compute bases of a larger invariant subspace, thus losing all the information about the original invariant subspaces.

The error bound for the eigenvectors computed by Algorithm 3 is given by (51) applied to the new set of clusters chosen in step 3. The formula (2) for individual eigenvectors follows easily from (51). The argument is as follows: Consider an individual eigenvalue λ_i, positive without loss of generality, belonging to a cluster Σ_k (chosen in step 3 of Algorithm 3). If λ_i is not the only positive eigenvalue in Σ_k, then (2) follows inmediately. If λ_i is the only positive eigenvalue in Σ_k and there are other negative eigenvalues in the cluster, then (2) follows because in Theorem 4.7 the bounds for the bases associated to positive and negative eigenvalues are independent of the relative gaps between the singular values inside Σ_k. The only remaining case is the one in which $\Sigma_k = \{\lambda_i\}$, i.e., the eigenvalue is by itself a cluster. If its closest cluster has not been joined to Σ_k by step 3 of Algorithm 3, it is either because it contains positive eigenvalues or because merging the two clusters would not improve the singular value relative gap. In any case, removing the closest (in absolute value) negative eigenvalues changes the singular value relative gap at most by a moderate factor. Therefore, (2) also holds in this case.

We will also relate our sharpest bound (51) with the eigenvalue relative gap. More precisely, we will show in this section that Algorithm 3 guarantees that the error in the computed basis of the invariant subspace corresponding to each cluster of eigenvalues $\widehat{\Lambda}_i$ of the symmetric matrix A is smaller than

$$\frac{O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))}{\min\{relgap(\widehat{\Lambda}_i), relgap(\widehat{\Lambda}_{cl(i)})\}}, \tag{54}$$

where the eigenvalue relative gap in the denominator corresponds to either the cluster $\widehat{\Lambda}_i$ under consideration or the cluster $\widehat{\Lambda}_{cl(i)}$ whose eigenvalues have different sign but are the closest (in relative sense) in absolute value. This result will be proved in Theorem 5.12 and generalizes to invariant subspaces the error bound (4), (5) appearing in the introduction for eigenvectors.

The rest of this section is organized as follows: Some relationships between eigenvalue and singular value relative gaps are proved in section 5.1. This is necessary if (54) has to be proved using Theorem 4.7, which only deals with singular value relative gaps. First we show in Theorem 5.5 that in the case of an *unsigned* cluster (a cluster containing singular values corresponding to positive and negative eigenvalues), the singular value relative gap of the cluster is not worse, up to a moderate constant, than an eigenvalue relative gap. Theorem 5.6 proves that this also happens to the relative gap of a signed cluster if the closest cluster is not signed of the opposite sign. Thus for clusters of these two kinds (54) holds, and it is not necessary to join them to any other cluster.

In the rest of section 5.1 we will study the case of a signed cluster whose closest cluster is oppositely signed. In all the theorems it will be assumed that the singular value relative gap is sufficiently smaller than the eigenvalue relative gap; otherwise it is trivial that (54) is reached. With these assumptions (54) is always achieved, either by joining clusters if the singular value relative gap improves (Theorem 5.7), or if not, by doing nothing (Theorem 5.9). Finally, Theorem 5.10 proves that it is not necessary to join more than two clusters. Let us remark that the only case in which Algorithm 2 has to be modified to get (54) is when the hypotheses of Theorem 5.7 are satisfied.

In subsection 5.2 we implement a routine, Algorithm 3.1, that merges pairs of adjacent singular value clusters, previously chosen according to (29), whenever the following conditions are met: (a) both clusters are signed with opposite sign, (b) the singular value relative gap is sufficiently smaller than the eigenvalue relative gap, and (c) the singular value relative gap increases after merging the two clusters. Algorithm 2.2 is then applied to these new clusters and Theorem 5.12 proves that (54) is achieved for the computed bases of the invariant subspaces.

Here, as in section 4, only clusters of nonzero singular values will be considered. Apart from the reasons stated in section 4, it should be remarked that a cluster of zero singular values is at the same time a cluster of zero eigenvalues, and both its eigenvalue and singular value relative gaps are equal to 1. Thus for such a cluster an error bound $O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))$ holds, and this cannot be improved. Moreover, a cluster of zero singular values is as far as possible, in relative distance, from any other cluster, thus joining it to another cluster makes no sense.

5.1. Eigenvalue versus singular value relative gaps. Throughout this section we consider a set of real numbers $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ decreasingly ordered, i.e., $\lambda_1 \geq \cdots \geq \lambda_n$, and the set of their moduli, $\Sigma = \{\sigma_1, \dots, \sigma_n\}$, also in decreasing order, i.e., $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Let Π be the index permutation such that $\sigma_i = |\lambda_{\Pi(i)}|$. Whenever we consider a subset $\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \dots, \sigma_{i+d_1}\}$ of Σ we will denote by $\Lambda_1 = \{\lambda_{\Pi(i+1)}, \dots, \lambda_{\Pi(i+d_1)}\}$ the corresponding subset of Λ; moreover, we will call Λ_1^+ (resp., Λ_1^-) the set of positive (resp., negative) elements of Λ_1. It is worth thinking of Λ and Σ as being, respectively, the set of eigenvalues and singular values of the real symmetric matrix A studied in section 4, but notice that the results in this subsection are proved using only elementary properties of real numbers, without any reference to spectral properties. Thus, the proofs of the theorems in this subsection are all elementary but sometimes long and involved, mainly due to dealing with clusters containing more than one element. This is why most of the proofs have been omitted. The proof of Theorem 5.7, one of the more intricate results in the section, is included in a final appendix, in order to give an idea of the techniques employed. The remaining proofs are similar, and those of a nonelementary character may be found in [10, Appendix B].

Our definitions of relative gaps (see (3) and (9)) are convenient and appealing in numerical analysis, but the lack of symmetry in relative errors of the type $|\sigma_j - \sigma_i|/\sigma_i$ is unpleasant from a mathematical point of view and complicates somewhat the statement of the results (see more on these questions and definitions of true relative mathematical distances in [19, 20]). In this sense, an effort has been made to state the theorems in such a way that they can be directly applied to the clusters chosen by Algorithm 3.1.

We begin with a general definition of cluster.

DEFINITION 5.1. *Let C_l be a real number such that $0 \leq C_l < 1$. The subset $\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \dots, \sigma_{i+d_1}\}$ of Σ is called a* cluster of tolerance C_l *if*

1. $(\sigma_j - \sigma_{j+1}) \leq C_l\,\sigma_j$ *for* $j = i+1, \ldots, i+d_1-1$,

2. $(\sigma_i - \sigma_{i+1}) > C_l\,\sigma_i$ *and* $(\sigma_{i+d_1} - \sigma_{i+d_1+1}) > C_l\,\sigma_{i+d_1}$, *whenever all the indices belong to* $\{1, 2, \ldots, n\}$*; otherwise the corresponding inequality does not appear in the definition.*

Notice that in the case of a cluster of dimension 1 ($d_1 = 1$) the first condition is empty. Notice also that this definition includes the clusters of singular values chosen in Algorithm 2, according to criterion (29), for $C_l = \epsilon\kappa(R')\max\{\kappa(X), \kappa(Y)\}$. The condition $C_l < 1$ appearing in Definition 5.1 is necessary—otherwise the whole set Σ would always be a trivial cluster, independently of the distribution of its elements.

Now we define relative gaps for subsets of Λ and Σ. For the sake of simplicity we will use only one argument.

DEFINITION 5.2. *Let* Λ_2 *and* Σ_1 *be any subsets of, respectively,* Λ *and* Σ. *We define the following relative gaps for both subsets:*

1.
$$rg(\Lambda_2) = \min_{\substack{\lambda_k \in \Lambda_2 \\ \lambda_q \notin \Lambda_2}} \frac{|\lambda_q - \lambda_k|}{|\lambda_k|}.$$

2.
$$relgap(\Lambda_2) = \min\{rg(\Lambda_2)\,,\, 1\}.$$

3.
$$rg(\Sigma_1) = \min_{\substack{\sigma_k \in \Sigma_1 \\ \sigma_q \notin \Sigma_1}} \frac{|\sigma_q - \sigma_k|}{\sigma_k}.$$

4.
$$relgap(\Sigma_1) = \min\{rg(\Sigma_1)\,,\, 1\}.$$

Given a subset Σ_1 of Σ, the relationship between the $relgap(\Sigma_1)$ appearing in Definition 5.2 and *relgap* as defined by (23) and (21) is

$$relgap(\Sigma_1) = relgap(\Sigma_{\bar{1}}, \Sigma_1), \tag{55}$$

where the notation introduced in (31) has been used. Similar comments apply to rg defined in (21) and rg defined above. Although $relgap(\Sigma_1, \Sigma_{\bar{1}})$ is the relative gap appearing in the error analysis of section 4, we have found it simpler, from both theoretical and computational points of view, to deal with $relgap(\Sigma_i)$, which has the elements of the cluster being analyzed in the denominators of the relative errors.[6] Both choices are equivalent, as shown in (24) and, on the other hand, it is possible to reformulate Theorem 2.3 using $relgap(\Sigma_i)$.

The error bounds for invariant subspaces computed using the J-orthogonal algorithm and Algorithm 1 are controlled by the relative gaps *relgap*, of eigenvalues and singular values, respectively, in the previous definition (see Theorem 4.7 and [22, p. 7]). However, in the following it is simpler and more general to use the relative gaps rg. At the end of this section it will be shown that theorems obtained for rg easily imply results for *relgap*.

[6] Notice that notation similar to Definition 5.2 has already been used in the introduction (see (3) and (9)).

We start with this simple lemma.

LEMMA 5.3. *Let* $\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i+d_1}\}$ *be a subset of consecutive elements of* Σ. *Then*

$$rg(\Sigma_1) = \min\left\{\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}, \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1}}\right\},$$

where if the index i *or* $i+d_1+1$ *does not belong to* $\{1, \ldots, n\}$ *the corresponding term does not appear in the minimum.*

This lemma allows a natural definition of the *closest cluster to* Σ_1 *in the relative sense.*

DEFINITION 5.4. *Let* $\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i+d_1}\}$ *be a cluster of tolerance* C_l. *We define its* relative closest cluster $\Sigma_{cl(1)}$ *as the cluster of tolerance* C_l *containing* σ_i *if* $rg(\Sigma_1) = (\sigma_i - \sigma_{i+1})/\sigma_{i+1}$, *or the one containing* σ_{i+d_1+1} *if* $rg(\Sigma_1) = (\sigma_{i+d_1} - \sigma_{i+d_1+1})/\sigma_{i+d_1}$.

It is seen from Lemma 5.3 that, with the possible exception of the cluster containing the smallest singular value, $rg(\Sigma_1) \leq 1$ and then $rg(\Sigma_1) = relgap(\Sigma_1)$. Obviously the last equality also holds whenever $rg(\Sigma_1) < 1$, a condition appearing frequently in the results of this section.

Our first result deals with the case of clusters containing singular values corresponding to positive and negative eigenvalues. This theorem shows that in this case the singular value relative gap of the cluster is not worse, up to a moderate constant, than an eigenvalue relative gap. Thus for clusters of singular values of this kind (54) holds, and it is not necessary to join them to any other cluster.

THEOREM 5.5. *Let* Σ_1 *be a cluster of singular values of tolerance* C_l *with* d_1 *elements such that* $(d_1 - 1)C_l < 1$, *and assume that* Λ_1 *contains both positive and negative elements. Then*

$$\min\{rg(\Lambda_1^+)\,,\, rg(\Lambda_1^-)\} \;\leq\; \frac{1}{1-(d_1-1)\,C_l}\left(1 + \frac{(d_1-1)\,C_l}{rg(\Sigma_1)}\right) rg(\Sigma_1).$$

Some remarks about the bound in the previous theorem are in order: the assumption $(d_1 - 1)\,C_l < 1$ is fulfilled for clusters of any size if we demand $C_l < 1/n$; this is really very mild because the clusters are chosen in practice according to (29) with $C = 1$, i.e., $C_l = \epsilon\kappa(R')\max(\kappa(X), \kappa(Y))$, which is smaller than $1/n$ for moderate values of $\max(\kappa(X), \kappa(Y))$. This has led us to set in the numerical experiments

$$C_l = \min\{\epsilon\kappa(R')\max(\kappa(X), \kappa(Y)), 1/n\}. \tag{56}$$

With this choice the factor $1/(1 - (d_1 - 1)\,C_l)$ is always less than n, but it is just a little greater than 1 when $C_l \approx \epsilon$. The presence of the ratio $C_l/rg(\Sigma_1)$ may look odd because we are bounding precisely the quotient $\min\{rg(\Lambda_1^+)\,,\, rg(\Lambda_1^-)\}/rg(\Sigma_1)$; however, notice that Definition 5.1 and Lemma 5.3 imply

$$C_l < rg(\Sigma_1) \quad \text{and} \quad C_l < relgap(\Sigma_1). \tag{57}$$

The ratio $C_l/rg(\Sigma_1)$ is kept in the bound because $C_l \ll rg(\Sigma_1)$ may often happen. It is convenient to bear in mind that these remarks also hold for the bounds appearing in the next theorems of this section. Notice also that all bounds are greatly simplified in the case of one-dimensional clusters.

Now we consider a signed cluster whose relative closest cluster has at least one singular value corresponding to an eigenvalue with the same sign. In this situation, the

next theorem shows that the singular value relative gap is equivalent to the eigenvalue relative gap up to a moderate constant.

THEOREM 5.6. *Let Σ_1 be a cluster of singular values and Σ_2 its relative closest cluster having d_2 elements, both of tolerance C_l. Let all the elements of Λ_1 have the same sign and at least one element of Λ_2 have the same sign as those of Λ_1. If $(d_2-1)C_l < 1$, then*

$$rg(\Lambda_1) \leq \left(1 + \frac{2}{1-(d_2-1)C_l}\,\frac{(d_2-1)\,C_l}{relgap(\Sigma_1)}\right) rg(\Sigma_1).$$

Theorems 5.5 and 5.6 guarantee that, in order to obtain (54) for all the singular value clusters, we need only deal with signed clusters whose relative closest cluster is oppositely signed. This will be the setting for the rest of the section. The following theorem proves that under mild conditions joining clusters of this kind leads to (54).

THEOREM 5.7. *Let Σ_1 be a cluster of d_1 elements and Σ_2 its relative closest cluster, having d_2 elements, both of tolerance C_l. Suppose that all the elements of Λ_1 have the same sign and all the elements of Λ_2 have the opposite sign. Moreover, assume that $(d-1)C_l < 1$, where $d = \max\{d_1, d_2\}$. If $rg(\Sigma_1) < t < 1$ and*

$$rg(\Sigma_1 \cup \Sigma_2) > \min\{rg(\Sigma_1), rg(\Sigma_2)\}, \tag{58}$$

then

$$\min\{rg(\Lambda_1)\,,\, rg(\Lambda_2)\} \leq \frac{1}{1-t}\left(1 + \frac{1}{1-(d-1)\,C_l} + \frac{1}{1-(d-1)\,C_l}\,\frac{(d-1)\,C_l}{rg(\Sigma_1 \cup \Sigma_2)}\right) rg(\Sigma_1 \cup \Sigma_2).$$

The assumption $rg(\Sigma_1) < t < 1$ means that only singular value clusters whose relative gaps are small enough need to be joined to other clusters in order to obtain (54). In practice we have set $t = relgap(\Lambda_1)/2$. Therefore, if $rg(\Sigma_1) \geq t$, the bound in Theorem 4.7 leads trivially to (54). The assumption (58), $rg(\Sigma_1 \cup \Sigma_2) > \min\{rg(\Sigma_1), rg(\Sigma_2)\}$, is imposed to guarantee that by joining clusters Σ_1 and Σ_2 when computing bases of invariant subspaces some improvement is achieved in the bound in Theorem 4.7. In this regard one may wonder what happens with $\max\{rg(\Sigma_1), rg(\Sigma_2)\}$; i.e., how much can the bound (51) worsen for the cluster with the maximum relative gap when Σ_1 and Σ_2 are joined? The next lemma shows that no significant worsening may occur.

LEMMA 5.8. *If both (58) and $rg\{\Sigma_1\} < t < 1$ are fulfilled, then*

$$\max\{rg(\Sigma_1), rg(\Sigma_2)\} < \frac{rg(\Sigma_1 \cup \Sigma_2)}{1-t}.$$

Notice that the difference between the maximum and the minimum values of $\{rg(\Sigma_1), rg(\Sigma_2)\}$ is in this case again a consequence of the lack of symmetry of the relative error.

In order to obtain (54) for all the clusters, we have to prove that if Σ_1 and its relative closest cluster Σ_2, defined as in Theorem 5.7, do not fulfill (58), they will not be joined because Σ_1 has a singular value relative gap not worse, up to a moderate constant, than either its eigenvalue relative gap or the eigenvalue relative gap of Σ_2. Proving this is the goal of the next theorem.

THEOREM 5.9. *Let Σ_1 be a cluster of d_1 elements and Σ_2 its relative closest cluster, having d_2 elements, both of tolerance C_l. Suppose that all the elements of Λ_1 have the same sign and all the elements of Λ_2 have the opposite sign. Moreover, assume that $(d-1)C_l < 1$, where $d = \max\{d_1, d_2\}$. If $rg(\Sigma_1) < t < 1$ and*

$$rg(\Sigma_1 \cup \Sigma_2) = \min\{rg(\Sigma_1), rg(\Sigma_2)\}, \tag{59}$$

then

$$\min\{rg(\Lambda_1)\,,\, rg(\Lambda_2)\} \leq \frac{1}{1-t}\left(1 + \frac{1}{1-(d-1)\,C_l} + \frac{1}{1-(d-1)\,C_l}\,\frac{(d-1)\,C_l}{rg(\Sigma_1)}\right) rg(\Sigma_1).$$

Observe that hypothesis (59) is simply the negation of (58) because we always have $rg(\Sigma_1 \cup \Sigma_2) \geq \min\{rg(\Sigma_1), rg(\Sigma_2)\}$.

Although similar, the bounds appearing in Theorems 5.7 and 5.9 are different in the following sense. While in Theorem 5.7 $\min\{rg(\Lambda_1)\,,\, rg(\Lambda_2)\} \approx rg(\Sigma_1 \cup \Sigma_2)$ always holds, in Theorem 5.9 $\min\{rg(\Lambda_1)\,,\, rg(\Lambda_2)\} \ll rg(\Sigma_1)$ might occur. Thus the error bounds obtained by replacing in (51) $rg(\Sigma_1)$ with $\min\{rg(\Lambda_1)\,,\, rg(\Lambda_2)\}$ may be pessimistic in the conditions of Theorem 5.9.

Our last result shows that in order to obtain (54), unions of more than two clusters are not necessary. In the following theorem three clusters are considered. Two of them satisfy the assumptions of Theorem 5.7, and the third cluster may be a candidate for joining the others. In this situation it will be proved that the relative singular value gap for the third cluster is equivalent, up to a moderate constant, to its eigenvalue relative gap.

THEOREM 5.10. *Let Σ_1 and Σ_2 be clusters satisfying the hypotheses of Theorem 5.7. Let Σ_3 be another cluster, of tolerance C_l, with all the elements of Λ_3 of the same sign and $rg(\Sigma_3) < t_3 < 1$. If Σ_1 (resp., Σ_2) is the relative closest cluster to Σ_3, and all the elements of Λ_3 have sign opposite to those of Λ_1 (resp., Λ_2), then*

$$rg(\Lambda_3) \;\leq\; \left(1 + \frac{1}{(1-t)(1-t_3)}\,\frac{1}{1-(d-1)\,C_l} + \frac{1+t_3}{1-(d-1)\,C_l}\,\frac{(d-1)\,C_l}{rg(\Sigma_3)}\right) rg(\Sigma_3).$$

As announced after Definition 5.2, all the bounds appearing in this section remain true if every rg is replaced by the corresponding *relgap*. This is easily understood as follows: the left-hand sides of the inequalities decrease if the rg's are replaced by the *relgap*'s, and the new left-hand sides are smaller than or equal to 1. The factors that multiply the rg's appearing in the right-hand sides are all greater than or equal to 1 and increase when quotients of the kind C_l/rg are replaced by $C_l/relgap$. Thus the left-hand sides are bounded simultaneously by 1 and by some factor greater than or equal to 1 times the corresponding rg. Then they are bounded by the factor times the *relgap*. Also notice that for testing the assumptions in the results in this section, it is equivalent to use rg's or *relgap*'s. First, it is trivial to see that $rg(\Sigma_1) < t < 1$ if and only if $relgap(\Sigma_1) < t < 1$. Second, in testing the condition (58), the following elementary lemma holds.

LEMMA 5.11. *Let*

$$\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i+d_1}\}, \quad \Sigma_2 = \{\sigma_{i+d_1+1}, \sigma_{i+d_1+2}, \ldots, \sigma_{i+d_1+d_2}\}$$

be any pair of consecutive clusters of nonzero singular values of tolerance C_l. Then

1. $rg(\Sigma_1 \cup \Sigma_2) = \min\{rg(\Sigma_1), rg(\Sigma_2)\}$ *if and only if* $relgap(\Sigma_1 \cup \Sigma_2) = \min\{relgap(\Sigma_1), relgap(\Sigma_2)\}$.

2. $rg(\Sigma_1 \cup \Sigma_2) > \min\{rg(\Sigma_1), rg(\Sigma_2)\}$ *if and only if* $relgap(\Sigma_1 \cup \Sigma_2) > \min\{relgap(\Sigma_1), relgap(\Sigma_2)\}$.

The key to proving this simple lemma is that $rg(\Sigma_1) \leq (\sigma_{i+d_1} - \sigma_{i+d_1+1})/\sigma_{i+d_1} < 1$; thus the 1 appearing in the *relgap*'s does not play any role. Taking into account the facts that $rg(\Sigma_1 \cup \Sigma_2) \geq \min\{rg(\Sigma_1), rg(\Sigma_2)\}$ and $relgap(\Sigma_1 \cup \Sigma_2) \geq \min\{relgap(\Sigma_1), relgap(\Sigma_2)\}$, statements 1 and 2 in the previous lemma are equivalent.

The final consequence of this section is that in order to get (54) only clusters fulfilling the hypotheses of Theorem 5.7 must be joined. Once a pair of clusters of this kind are joined, they can be disregarded in any other union processes as shown by Theorem 5.10. Otherwise, the rest of the results prove that union of clusters of different kinds is not needed. In the next subsection the task of developing a routine that selects and joins clusters according to this criterion will be undertaken.

5.2. Choosing a new set of clusters. Now we will present a routine for step 3 of Algorithm 3. Given a set of clusters as input, selected according to (29), a new set of clusters will come out according to the logic of the theorems in section 5.1; i.e., clusters will be joined only if the hypotheses of Theorem 5.7 are satisfied. All clusters of singular values appearing in the following algorithm are assumed to contain consecutive singular values. Moreover, we order the clusters $\{\Sigma_i\}$ in such a way that any singular value in Σ_i is smaller than any singular value in Σ_{i-1}.

ALGORITHM 3.1.

Input: Eigenvalues Λ; Clusters $\{\Sigma_i\}_{i=1}^{k}$; tolgap: parameter smaller than 1.
Output: New set of clusters: $\{\Sigma_i\}_{i=1}^{q}$ with $q \leq k$.
Notation: Λ_i denotes the set of eigenvalues whose absolute values are the elements of Σ_i.

1. $q = k$
2. for i=1:k
   $qrg(i) = \frac{relgap(\Sigma_i)}{relgap(\Lambda_i)}$
   if($\lambda_j > 0 \quad \forall\, \lambda_j \in \Sigma_i$) then
   $sign(\Sigma_i) = +1$
   elseif($\lambda_j < 0 \quad \forall\, \lambda_j \in \Sigma_i$)
   $sign(\Sigma_i) = -1$
   else
   $sign(\Sigma_i) = 0$
   $qrg(i) = 2$
   endif
   endfor
3. $qrgmin = \min_{1 \leq i \leq q} qrg(i) \equiv qrg(i_c)$
4. while $qrgmin <$ tolgap
   determine the relative closest[7] cluster to Σ_{i_c} according to Definition 5.4. Assume that it is Σ_{i_c+1}.
   if ($sign(\Sigma_{i_c}) * sign(\Sigma_{i_c+1}) = -1$) and
   ($relgap(\Sigma_{i_c} \cup \Sigma_{i_c+1}) > \min\{relgap(\Sigma_{i_c}), relgap(\Sigma_{i_c+1})\}$) then
   $q = q - 1$
   $relgap(\Sigma_{i_c}) = relgap(\Sigma_{i_c} \cup \Sigma_{i_c+1})$

[7] The same can be done if Σ_{i_c-1} is the relative closest cluster to Σ_{i_c}.

$sign(\Sigma_{i_c}) = 0$
$\Sigma_{i_c} = \Sigma_{i_c} \cup \Sigma_{i_c+1}$
`for` $j = i_c + 1 : q$
$\Sigma_j = \Sigma_{j+1}$
$relgap(\Sigma_j) = relgap(\Sigma_{j+1})$
$sign(\Sigma_j) = sign(\Sigma_{j+1})$
`endfor`
`endif`
$qrg(i_c) = 2$
$qrgmin = \min_{1 \le i \le q} qrg(i) \equiv qrg(i_c)$
`5. endwhile`

In practice we have set `tolgap` $= 1/2$, but other values are admissible. This choice leads to values $t = (relgap(\widehat{\Lambda}_i)/2) \le 1/2$ for the parameters t appearing in Theorems 5.7, 5.9, and 5.10.

For the new set of clusters selected by Algorithm 3.1, the error in the corresponding bases of invariant subspaces computed by Algorithm 2.2 *is given by Theorem* 4.7 *using the new singular value relative gaps, and these are the sharpest bounds we have for Algorithm* 3. Nevertheless, in the next theorem we will use the theorems in the previous subsection to give an upper bound for the inverse of the new singular value relative gaps in (51) in terms of inverses of the eigenvalue relative gaps. Therefore this theorem gives a precise statement of (54).

THEOREM 5.12. *Let A be a $n \times n$ real symmetric matrix of rank r for which it is possible to compute an RRD fulfilling* (10). *Let $\widehat{\Sigma}$ be the singular values of A computed using steps* 1–2 *of Algorithm* 1. *Let $\widehat{\Sigma}_i$, $i = 1, \ldots, q$, be the set of clusters of nonzero computed singular values of A selected by* step 3 *of Algorithm* 3, $\widehat{\Lambda}_i = \widehat{\Lambda}_i^+ \cup \widehat{\Lambda}_i^-$, $i = 1, \ldots, q$, *the corresponding set of clusters of eigenvalues, and $\widehat{Q}_i = [\widehat{Q}_i^+ \ \widehat{Q}_i^-]$, $i = 1, \ldots, q$, the matrices computed by* step 4 *of Algorithm* 3. *Let Σ_i (resp., Λ_i) , $i = 1, \ldots, q$, be the corresponding clusters of exact singular values (resp., eigenvalues).*

1. *If neither $\widehat{\Lambda}_i^+$ nor $\widehat{\Lambda}_i^-$ are empty, then there exist matrices Q_i^+ and Q_i^-, whose columns form orthonormal bases of the invariant subspaces of A corresponding, respectively, to the positive and negative eigenvalues of Λ_i, such that*

$$\|\widehat{Q}_i^+ - Q_i^+\|_F \le \frac{O(\epsilon \kappa(R') \max(\kappa(X), \kappa(Y)))}{\min\{relgap(\widehat{\Lambda}_i^+), relgap(\widehat{\Lambda}_i^-)\}}, \tag{60}$$

with a similar bound for $\|\widehat{Q}_i^- - Q_i^-\|_F$.

2. *If all the elements of $\widehat{\Lambda}_i$ have the same sign and $relgap(\widehat{\Sigma}_i) \ge$ `tolgap` $*$ $relgap(\widehat{\Lambda}_i)$, then there exists a matrix Q_i, whose columns form an orthonormal basis of the invariant subspace of A corresponding to the eigenvalues in Λ_i, such that*

$$\|\widehat{Q}_i - Q_i\|_F \le \frac{O(\epsilon \kappa(R') \max(\kappa(X), \kappa(Y)))}{relgap(\widehat{\Lambda}_i)}. \tag{61}$$

3. *If all elements of $\widehat{\Lambda}_i$ have the same sign, $relgap(\widehat{\Sigma}_i) <$ `tolgap` $* relgap(\widehat{\Lambda}_i)$, and the relative closest cluster $\widehat{\Sigma}_{cl(i)}$ to $\widehat{\Sigma}_i$ has all the corresponding eigenvalues with the opposite sign, then there exists a matrix Q_i, whose columns form an orthonormal*

basis of the invariant subspace of A corresponding to the eigenvalues in Λ_i, such that

$$\|\widehat{Q}_i - Q_i\|_F \leq \frac{O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))}{\min\{relgap(\widehat{\Lambda}_i), relgap(\widehat{\Lambda}_{cl(i)})\}}. \tag{62}$$

4. *If all elements of $\widehat{\Lambda}_i$ have the same sign, $relgap(\widehat{\Sigma}_i) < \texttt{tolgap} * relgap(\widehat{\Lambda}_i)$, and the relative closest cluster to $\widehat{\Sigma}_i$ does not have all the corresponding eigenvalues with the opposite sign, then there exists a matrix Q_i, whose columns form an orthonormal basis of the invariant subspace of A corresponding to the eigenvalues in Λ_i, such that*

$$\|\widehat{Q}_i - Q_i\|_F \leq \frac{O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))}{relgap(\widehat{\Lambda}_i)}. \tag{63}$$

Furthermore, let $\widehat{Q} = [\widehat{Q}_1^+ \ \widehat{Q}_1^- \ldots \widehat{Q}_q^+ \ \widehat{Q}_q^-]$ be the $n \times r$ matrix whose columns are the bases of all considered invariant subspaces of A computed using step 4 *of Algorithm* 3. *Then there exists an $n \times r$ matrix B with exact orthonormal columns such that*

$$\|\widehat{Q} - B\|_F = O(\epsilon). \tag{64}$$

Proof. The proof follows from Theorem 4.7 applied to the output clusters of Algorithm 3.1 (step 3 of Algorithm 3) and the theorems on gaps in section 5.1 with $C_l = \epsilon\kappa(R')\max(\kappa(X),\kappa(Y))$. As remarked after Theorem 5.10, *relgap*'s instead of *rg*'s can be used in these theorems.

We begin by replacing $relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})$ with $relgap(\widehat{\Sigma}_{\bar{i}}, \Sigma_i)$ in the bound (51). This does not significantly change the bound due to (24). Moreover, we assume that $relgap(\widehat{\Sigma}_{\bar{i}}, \Sigma_i) \approx relgap(\widehat{\Sigma}_{\bar{i}}, \widehat{\Sigma}_i)$. This is a fair assumption whenever steps 1–2 of Algorithm 1 compute singular values with high relative accuracy. Thus (55) allows us to apply (51), with $relgap(\Sigma_i, \widehat{\Sigma}_{\bar{i}})$ replaced by $relgap(\widehat{\Sigma}_i)$, to the clusters selected by Algorithm 3.1.

Consider a cluster $\widehat{\Sigma}_{i_c}$ of singular values corresponding to the quantity *qrgmin* in Algorithm 3.1. This cluster is joined to its relative closest cluster if and only if the following three conditions are simultaneously fulfilled:

(c1) $qrg(i_c) = \frac{relgap(\widehat{\Sigma}_{i_c})}{relgap(\widehat{\Lambda}_{i_c})} < \texttt{tolgap} < 1$.

(c2) $sign(\widehat{\Sigma}_{cl(i_c)}) * sign(\widehat{\Sigma}_{i_c}) = -1$, where $\widehat{\Sigma}_{cl(i_c)}$ is the closest cluster to $\widehat{\Sigma}_{i_c}$.

(c3) $relgap(\widehat{\Sigma}_{i_c} \cup \widehat{\Sigma}_{cl(i_c)}) > \min\{relgap(\widehat{\Sigma}_{i_c}), relgap(\widehat{\Sigma}_{cl(i_c)})\}$.

If all three conditions (c1), (c2), and (c3) are fulfilled, Algorithm 3.1 joins $\widehat{\Sigma}_{i_c}$ and $\widehat{\Sigma}_{cl(i_c)}$ in a new output cluster $\widehat{\Sigma}_{i_c} \cup \widehat{\Sigma}_{cl(i_c)}$. In this case Theorem 5.7 applies with $t = \texttt{tolgap} * relgap(\widehat{\Lambda}_{i_c})$. This together with (51) yields (60) for the eigenvectors corresponding to the new output cluster.

Now, suppose that at least one of the three conditions is not satisfied. Suppose first that (c1) is satisfied, which implies $sign(\widehat{\Sigma}_{i_c}) \neq 0$; otherwise $qrgmin = 2$. If (c2) is not verified and the closest cluster to $\widehat{\Sigma}_{i_c}$ is an input cluster, Theorem 5.6 can be applied to the bound (51) to obtain (63); on the other hand, if (c2) is not verified and the closest cluster is a new output cluster, (63) is achieved by using Theorem 5.6 or 5.10. If (c2) is verified and (c3) is also verified, we are in the previously studied case of joining clusters. If (c2) is verified and (c3) is not verified, Theorem 5.9 can be applied to (51) to yield (62).

Suppose from now on that (c1) is not satisfied. Then, Algorithm 3.1 stops and all the clusters existing at that moment verify

$$qrg(i) \geq \texttt{tolgap}, \quad i = 1, \ldots, q.$$

If $sign(\widehat{\Sigma}_i) = 0$, this is either because $sign(\widehat{\Sigma}_i) = 0$ on input or because $\widehat{\Sigma}_i$ is a new output cluster, i.e., union of two input clusters. Anyway, Theorem 5.5 or 5.7 leads to (60) by using (51). If $sign(\widehat{\Sigma}_i) \neq 0$ and $qrg(i) = 2$, then $\widehat{\Sigma}_i$ already has been analyzed inside the `while` loop and, according to the previous paragraph, either (62) or (63) is satisfied. If $sign(\widehat{\Sigma}_i) \neq 0$ but $\texttt{tolgap} \leq relgap(\widehat{\Sigma}_i)/relgap(\widehat{\Lambda}_i) \leq 1$, then (51) implies (61) at the cost of an additional factor $1/\texttt{tolgap}$. With this, all the possible cases on the decision tree for the conditions (c1), (c2), and (c3) have been studied. The proof of (64) is as in Theorem 4.7. □

We finish this section with two important remarks.

Remark 1. The *eigenvalue* clusters treated in the last theorem are exactly the same as the ones corresponding to the singular value clusters chosen according to (29). This is because Algorithm 3.1 only joins oppositely signed clusters and Algorithm 2.2 computes the bases separately.

Remark 2. The bounds in Theorem 5.12 have been obtained in two stages: first, applying Theorem 4.7 to the new set of clusters produces a bound depending on singular value relative gaps. Then, this bound is majorized by other ones, depending on certain eigenvalue relative gaps. This second stage never worsens significantly the first bound, except in case 3 of Theorem 5.12. Thus, the bound (62) may be pessimistic, because the quantity $\min\{relgap(\widehat{\Lambda}_i), relgap(\widehat{\Lambda}_{cl(i)})\}$ might be much smaller than $relgap(\widehat{\Sigma}_i)$. However, recall that the sharpest bound for Algorithm 3 is of the order of $\epsilon\kappa(R')\max(\kappa(X), \kappa(Y))/relgap(\widehat{\Sigma}_i)$.

6. Numerical experiments. In this section we present results of two types of numerical experiments. First, we test Algorithm 3, the third step of Algorithm 1, in a setting where the errors for steps 1 and 2 of Algorithm 1 are controlled. A second kind of experiment tests the entire Algorithm 1, including the computation of the RRD in two different ways, as either a symmetric RRD of the form $A = XDX^T$ or a nonsymmetric RRD of the form $A = XDY^T$. We also include experiments for Algorithm 1 with Algorithm 2 in step 3. Thus the reader can check that Algorithm 3 really improves the accuracy of the eigenvectors in the few cases in which Algorithm 2 delivers eigenvectors with large errors. When needed, we will distinguish between the two versions of Algorithm 1: the version with Algorithm 2 in step 3 will be called `SSVDO`, and the one with Algorithm 3 will be called simply `SSVD`. Besides, a first subsection describes some practical details of the implementation of the three steps of Algorithm 1.

As will be seen from the experiments in subsection 6.2, Algorithm 1 behaves as predicted by the error analysis in sections 4 and 5 and compares well in both the sense of accuracy and of speed with the J-orthogonal algorithm.

6.1. Implementation of Algorithm 1.

1. The RRD of the matrix A in step 1 of Algorithm 1 has been done in the following two ways:

- symmetric RRD, $A = XDX^T$, using a modification of the symmetric indefinite Bunch and Parlett (BP) decomposition [3]; more specifically, we have used an adapted version of the routine `SGJGT` in [22].

- a nonsymmetric RRD, $A = XDY^T$, by means of an LU factorization with complete pivoting (Gaussian elimination with complete pivoting (GECP)). We have used a modification of the LAPACK procedure `SGETF2`.

2. The SVD in step 2 of Algorithm 1 has been done using Algorithm 3.1 of [6]. Only LAPACK and BLAS routines have been used, as in [6], except for the one-sided Jacobi code in which we have used a routine developed by Z. Drmač according to the ideas in [12]. The implementation of the procedure (called `SGEPSV` in [6] in single precision) has the following steps.

ALGORITHM 4. `(SGEPSV)` (ALGORITHM 3.1 IN [6].)
`Input:` $X, D, Y : A = XDY^T$.
`Output:` $U, \Sigma, V : A = U\Sigma V^T$.
1. `QR factorization with column pivoting of` XD, $XDP = QR$; $A = QRP^TY^T$
   `LAPACK Routine: SGEQPF`
2. `Multiply to get` $W = R(YP)^T$; $A = QW$
   `BLAS Routine: STRMM`
3. `SVD of W with one-sided Jacobi;` $W = \overline{U}\Sigma V^T$; $A = Q\overline{U}\Sigma V^T$
   `Routine: S_SGESVDJ developed by Z. Drmač [12]`
4. `Multiply` $U = Q\overline{U}$; $A = U\Sigma V^T$
   `LAPACK Routine: SORMQR`

Two versions of this algorithm have been used, depending on whether right-Jacobi (right multiplication on W by Jacobi plane rotations) or left-Jacobi (right multiplication on W^T by Jacobi plane rotations) is employed in the one-sided Jacobi step 3 of Algorithm 4 in [6]. The left-Jacobi version has the advantage of speeding up the convergence. Although the error bounds for this version are weaker than for the other version (see [11] or [10, Appendix A]), no significant difference in accuracy has ever been observed in practice. Our experiments confirm this.

In any case the routine that has been used computes one of the singular vector matrices by a product of Jacobi plane rotations. There exist much faster, equally accurate, versions of one-sided Jacobi algorithms which do not accumulate rotations [14], and which could also be used. Nevertheless, with the present implementation the timing statistics of Algorithm 1 are comparable to the J-orthogonal algorithm (see the timing data in the last paragraph of Experiment 2 in subsection 6.2 below).

3. Algorithm 2 in step 3 of Algorithm 1 has been implemented as described in subsection 3.3. Algorithm 3, the final version of step 3 in Algorithm 1, has been implemented as described in section 5. Some additional specific details are the following:

(i) Recall that steps 1 and 2 are the same in both Algorithms 2 and 3, and therefore the eigenvalues computed by both algorithms are the same.

(ii) The choice of clusters in step 1 of Algorithms 2 and 3 has been done using (29) by taking $C = 1$ and using the $O(n^2)$ estimator LAPACK routine `STRCON` to estimate $\kappa(R')$, or $\kappa(X)$, $\kappa(Y)$, when starting from a nonfactorized matrix. When generating matrices in RRD form $A = XDX^T$, some matrices X producing values of $\epsilon\kappa(R')\kappa(X)$ larger than 1 have appeared. This means that the SVD routine, Algorithm 4, guarantees no significant digits of precision in the computation of the singular values. Moreover, using (29) produces in this case that all singular values are contained in just one cluster. Our discussion after Theorem 5.5 has led us to establish in practice the criterion to include two contiguous singular values σ_j, σ_{j+1} in the same cluster

whenever

$$\frac{|\sigma_j - \sigma_{j+1}|}{\sigma_j} \leq \min\{\epsilon\kappa(R')\max\{\kappa(X),\kappa(Y)\}, 1/n\}. \tag{65}$$

(iii) The product $\Delta_i = V_i^T U_i$ in step 11 of Algorithm 2.2 has been done using the BLAS routine `SGEMM`.

(iv) The diagonalization of $\Delta_i = [W_i^+ \, W_i^-] J_i [W_i^+ \, W_i^-]^T$ (step 12 of Algorithm 2.2) has been done using the LAPACK routine `SSYEV` applied only to the triangular upper half of the matrix, as assumed in Lemma 4.5. Finally, the eigenvector matrices $Q_i^{\pm} = V_i W_i^{\pm}$ (step 13 of Algorithm 2.2) are obtained using the BLAS multiplication routine `SGEMM`.

(v) In all the experiments the value for the parameter `tolgap` appearing in Algorithm 3.1 has been set to `tolgap = 1/2`.

6.2. Numerical results. The following experiments were done using an AMD K7 ATHLON processor with IEEE arithmetic, and the routines were implemented with Fortran PowerStation 4.0 from Microsoft. All numerical experiments in this section have been done with nonsingular matrices, although as pointed out in sections 3 and 4, Algorithm 1 also can be applied to rank-deficient matrices.

In the first experiment we start from matrices already in factorized RRD form $A = XDX^T$, directly generating the matrices X and D. This has helped us to focus on the accuracy of step 3 in Algorithm 1 since, given the RRD, the work by Demmel et al. in [6] allows us to control the error in step 2 of Algorithm 1.

In the second group of experiments, two different kinds of nonfactorized test matrices have been generated: graded matrices and matrices specifically designed in [22] to guarantee a good performance of the J-orthogonal algorithm. The reason for choosing graded matrices is that it is known, under the conditions given in [6, section 4], that an accurate RRD, in the sense of (10), can be computed using a *plain implementation* of GECP. For the rest of the classes of matrices treated in [6, pp. 26–27], special implementations of GECP are needed to get the desired accuracy, and it is unfair to compare in these cases Algorithm 1 with the J-orthogonal algorithm, since at present no special implementations of the symmetric indefinite factorization are known to guarantee the accuracy. The reason for choosing the matrices designed in [22] is to compare Algorithm 1 and the J-orthogonal algorithm on matrices where the accuracy of the J-orthogonal algorithm of the latter is guaranteed.

To test Algorithm 1 we have used as reference the eigenvalues and eigenvectors computed by the routine `DSYEVJ`, developed by I. Slapničar, that implements the implicit one-sided J-orthogonal algorithm[8] [22] in double precision ($\epsilon = \epsilon_D \approx 1.11 \times 10^{-16}$). From now on these eigenvalues and eigenvectors are denoted, respectively, simply by λ_i and q_i. These are compared with the eigenvalues and eigenvectors, $\lambda_i^{(S)}$ and $q_i^{(S)}$, computed in single precision ($\epsilon = \epsilon_s \approx 5.96 \times 10^{-8}$) by the following routines: `SSVDO` (Algorithm 1, using Algorithm 2 in step 3), `SSVD` (Algorithm 1, using Algorithm 3 in step 3), `SSYEVJ` (J-orthogonal algorithm, denoted simply by J-O in the tables and figure), and, only when we start from a full (not already in rank-

[8] `DSYEVJ` is a driver routine formed by two routines that implement the two steps of the J-orthogonal algorithm: subroutine `DGJGT` (symmetric indefinite decomposition with complete pivoting) and subroutine `DJGJF` (implicit J-orthogonal Jacobi method with the same stopping criterion as one-sided Jacobi). `DSYEVJ` has been used when starting with the full matrix A. When starting from a factorized matrix $A = XDX^T$ only the subroutine `DJGJF` has been used. Similar remarks apply to the single precision driver routine `SSYEVJ`.

revealing form) matrix A, `SJAC` (standard Jacobi algorithm with the new stopping criterion introduced in [7, p. 1206] with `tol` $= \epsilon_s$) and `SSYEV` (LAPACK driver routine that implements tridiagonalization followed by QR iteration). For these methods the following quantities have been measured for each test matrix:

1. The maximum relative error in the eigenvalues:

$$e_\lambda^{(\mathrm{S})} = \max_i \left| \frac{\lambda_i - \lambda_i^{(\mathrm{S})}}{\lambda_i} \right|. \tag{66}$$

2. A control quantity for eigenvalues:

$$\vartheta^{(\mathrm{S})} = \frac{e_\lambda^{(\mathrm{S})}}{\kappa\, \epsilon_s}, \tag{67}$$

where $\kappa = \kappa(R') \max\{\kappa(X), \kappa(Y)\}$, as in (15). Observe that when referring to symmetric RRDs κ is just $\kappa(R')\kappa(X)$. According to the bound (38), the quantity $\vartheta^{(\mathrm{S})}$ is expected to be $O(1)$ for Algorithm 1. For the J-orthogonal algorithm the error $e_\lambda^{(\mathrm{S})}$ is essentially bounded by $O(\epsilon_s\, \kappa(XD_X))$, where XD_X is the best conditioned column diagonal scaling of matrix X [22]. However, we have checked that $\kappa(X) \approx \kappa(XD_X)$ in our tests. This is due to the fact that the matrices X appearing in our experiments do not have any special structure. Furthermore, the extra factor $\kappa(R')$ in the denominator that we have observed is $O(n)$ in the numerical tests in this section (see also [6, Thm. 3.2]) renders $\vartheta^{(\mathrm{S})}$ inadequate to check how well the bounds for the J-orthogonal algorithm behave, although it is still valid to compare the accuracy of Algorithm 1 and the J-orthogonal algorithm. For the other two considered algorithms, Jacobi and QR, $\vartheta^{(\mathrm{S})}$ is just the maximum error in the eigenvalues normalized in the same way as for both Algorithm 1 and the J-orthogonal algorithm. Similar remarks apply to the eigenvector computations.

3. Corresponding to each cluster of eigenvalues, the sine of the maximum of canonical angles between the subspaces spanned by the computed basis, Q_i, in double precision and the computed basis, $Q_i^{(\mathrm{S})}$, in single precision:

$$E_{\Lambda_i}^{(\mathrm{S})} = \| \sin \Theta(Q_i, Q_i^{(\mathrm{S})}) \|_2. \tag{68}$$

In the case of clusters with one single element we have computed just the Euclidean norm of the difference between the computed eigenvectors in double, q_i, and single $q_i^{(\mathrm{S})}$, precision,

$$e_{q_i}^{(\mathrm{S})} = \| q_i - q_i^{(\mathrm{S})} \|_2. \tag{69}$$

Actually, the quantities $e_{q_i}^{(\mathrm{S})}$ are always computed, even in the presence of clusters of dimension larger than one. We do this in order to check that clusters are only chosen whenever no accuracy can be guaranteed for individual computed eigenvectors.

4. The control quantities for bases of invariant subspaces are

$$\Xi_\Sigma^{(\mathrm{S})} = \max_i \frac{E_{\Lambda_i}^{(\mathrm{S})}\ \mathit{relgap}(\Sigma_i^{(\mathrm{S})})}{\kappa\, \epsilon_s}, \qquad \Xi_\Lambda^{(\mathrm{S})} = \max_i \frac{E_{\Lambda_i}^{(\mathrm{S})}\ \mathit{relgap}(\Lambda_i^{(\mathrm{S})})}{\kappa\, \epsilon_s}, \tag{70}$$

and the corresponding ones for individual eigenvectors are

$$\xi_\sigma^{(\mathrm{S})} = \max_i \frac{\|q_i - q_i^{(\mathrm{S})}\|_2 \;\; relgap(\sigma_i^{(\mathrm{S})})}{\kappa\, \epsilon_s},$$
$$\xi_\lambda^{(\mathrm{S})} = \max_i \frac{\|q_i - q_i^{(\mathrm{S})}\|_2 \;\; relgap(\lambda_i^{(\mathrm{S})})}{\kappa\, \epsilon_s}. \tag{71}$$

According to Theorem 4.7, $\Xi_\Sigma^{(\mathrm{S})}$ and $\xi_\sigma^{(\mathrm{S})}$ are expected to be $O(1)$ for Algorithms `SSVD` and `SSVD0`. Also $\Xi_\Lambda^{(\mathrm{S})}$ and $\xi_\lambda^{(\mathrm{S})}$ are expected to be $O(1)$ for the J-orthogonal algorithm, but not for Algorithms `SSVD` and `SSVD0`, because the accuracy of `SSVD` is governed by Theorem 5.12. However, the quantities $\Xi_\Lambda^{(\mathrm{S})}$ and $\xi_\lambda^{(\mathrm{S})}$ will be computed by `SSVD` and `SSVD0` to check in practice how the `SSVD` algorithm improves the accuracy of `SSVD0` and how it compares with the J-orthogonal algorithm. Notice that the quantities $relgap(\Sigma_i^{(\mathrm{S})})$ correspond either to the set of cluster chosen according to (65) for Algorithm `SSVD0` or to the output clusters of Algorithm 3.1 for Algorithm `SSVD`. The quantities $relgap(\Lambda_i^{(\mathrm{S})})$ are always the same because the clusters for eigenvalues do not change (see the remarks at the end of subsection 5.2). The $relgap$s in (71) are the ones defined in (3) and (9) for any of the algorithms.

For the sake of brevity, values of $\xi_\sigma^{(\mathrm{S})}$ or $\xi_\lambda^{(\mathrm{S})}$ are not shown for routines `SJAC` and `SSYEV`; we simply report that extremely large errors are obtained for these algorithms.

To do our experiments we have generated matrices in single precision in different ways. All the random matrices needed have been generated using the LAPACK routines `SLATM1`, for diagonal matrices, and `SLATMR`, for full matrices. When we have generated matrices with a fixed condition number $\mathcal{K}$, it has been done by producing diagonal matrices with elements of absolute values in the range from 1 to $1/\mathcal{K}$, and after that multiplying by random single precision orthogonal matrices. The distribution of the diagonal elements is controlled by the parameter `MODE` of the routine `SLATM1`: $|$`MODE`$| = 3$, geometrically distributed; $|$`MODE`$| = 4$, arithmetically distributed; `MODE` $= 5$, with logarithms uniformly distributed. If `MODE` is positive (resp., negative) the elements are set in decreasing (resp., increasing) order.

Experiment 1. This experiment is designed to test Algorithms 2 and 3. We have generated $n \times n$ matrices X and D (diagonal), factors of a matrix $A = XDX^T$, as done in [6]. Parameters have been chosen as follows: $\kappa(X) = 10^{[2:1:6]}$; $\kappa(D) = 10^{[2:2:16]}$; $MODE_X = 3, 4, 5$; $MODE_D = \pm 3, \pm 4, 5$. For each set of parameters we have run 20 matrices for $n = 50, 100$ (total 12000 matrices for each n), 2 for $n = 250$ (total 1200 matrices), 2 for $n = 500$ (total 1200 matrices), 1 for $n = 1000$, and only for 2 combinations of the $MODE$s (total 80 matrices).

Figure 6.1 shows the maximum, minimum, and average (over all $MODE$s, samples, and $\kappa(D)$s) of the quantity $log_{10} e_\lambda^{(\mathrm{S})}$, roughly the number of correct digits in the computed eigenvalues, as a function of $\kappa(X)$ for $n = 100$ for Algorithm 1 (`SSVD` or `SSVD0`) and for the J-orthogonal algorithm. The line $\epsilon_s \kappa(X) \kappa(R')$ is plotted as a guide to the eye; the quantity $\kappa(R')$ in this line is really the average of $\kappa(R')$ over all the matrices with that value of $\kappa(X)$. The results confirm the theoretical error bounds for eigenvalues.

Table 6.1 shows the statistical data corresponding to the quantity $\vartheta^{(\mathrm{S})}$. The aim is to check the bound (38) for Algorithm 1 and compare its accuracy against the J-orthogonal algorithm. The most significant data in Table 6.1 appear under the columns labeled "max" where the maximum values of each magnitude (the ones bounded by the error analysis) are shown. In particular, the fact that the quantities

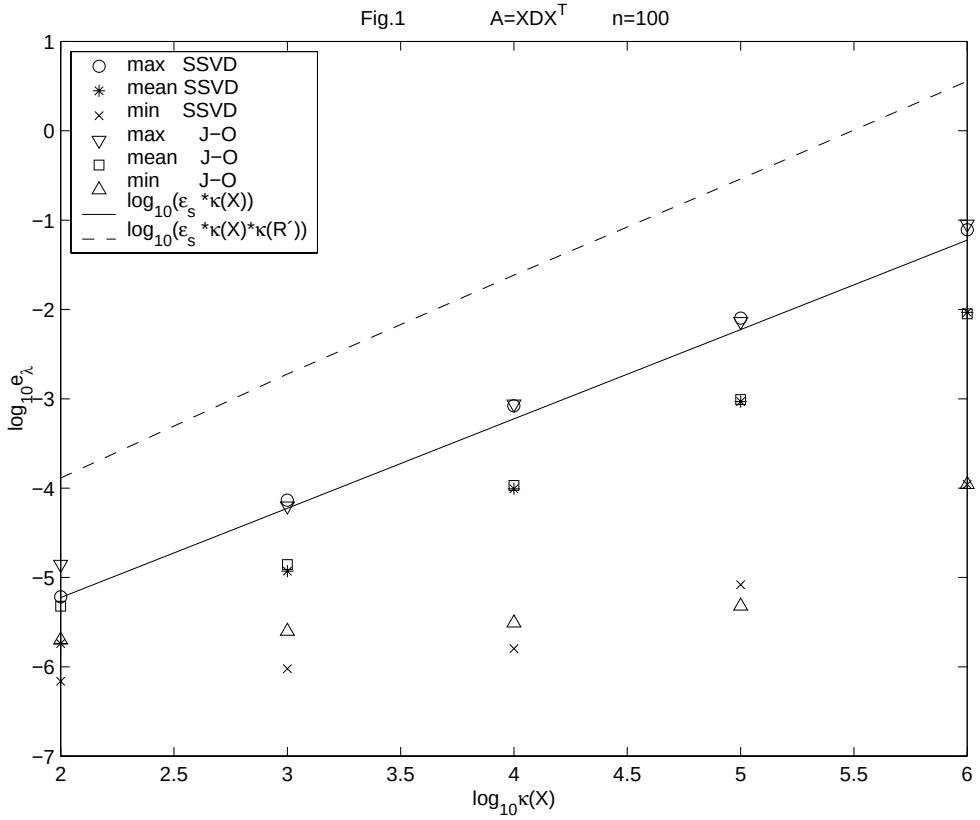


FIG. 6.1. *Experiment* 1. *Maximum relative error for eigenvalues:* $\log_{10} e_\lambda^{(\mathrm{S})}$ *against* $\log_{10} \kappa(X)$.

TABLE 6.1
Experiment 1. *Statistical data for accuracy in eigenvalues:* $\vartheta^{(\mathrm{S})}$.

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
$\vartheta^{(\mathtt{SSVD})}$	.030	.40	.022	.31	.015	.17	.013	.22	.013	.20
$\vartheta^{(\mathtt{J-O})}$	.041	.58	.037	.44	.039	.47	.044	.63	.050	.65
$\vartheta^{(\mathtt{SVD})}$	.030	.40	.022	.31	.015	.17	.013	.22	.012	.20

in the first row are smaller than 1 confirms that Algorithm 1 satisfies the bound (38). In addition, the third row itself is the control quantity ϑ calculated for the singular values computed in step 2 of Algorithm 1. The comparison of the first and third rows shows that Algorithm 1 never misses a sign and always gives eigenvalues with the same precision as the singular values, except for five matrices of dimension 1000. These cases have $\kappa(X) = 10^6$ and $\epsilon_s \kappa(X)\kappa(R')$ greater than 100. Therefore *whenever* $\epsilon_s \kappa(X)\kappa(R') < 1$ *Algorithm* 1 *has given the eigenvalues with the same precision as the singular values computed by Algorithm* 3.1 *in* [6]. It can be seen, from both Figure 6.1 and Table 6.1, that Algorithm 1 performs for eigenvalues as well (even a little better, especially for small values of $\kappa(X)$) as the J-orthogonal algorithm, with the maximum errors in Algorithm 1 adjusting very well to the predicted behavior $\epsilon\kappa(X)\kappa(R')$. It can be observed also that the data do not depend on n.

TABLE 6.2
Experiment 1. *Statistical data for the number of sweeps.*

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
Sweeps (SSVD)	5.5	10	6.3	12	7.4	12	8.4	14	9.3	15
Sweeps (J-O)	10.5	20	11.7	22	13.0	22	13.9	24	13.1	24

Moreover, for a significant portion of all the matrices (4144 matrices out of 12000 for $n = 50$; 6693 matrices out of 12000 for $n = 100$; 974 matrices out of 1200 for $n = 250$; 1105 matrices out of 1200 for $n = 500$; 79 matrices out of 80 for $n = 1000$), clusters of singular values of dimension greater than 1, according to criterion (65), have been found, with the maximum dimension of a cluster being 5. The average number of clusters has ranged from almost no clusters for $n = 50$ to approximately 40 clusters for $n = 1000$, with a typical dimension of 2. This shows that criterion (65) chooses clusters which determine perfectly in practice the signs of the eigenvalues. After applying Algorithm 3.1 all the considered matrices have clusters. The average number of clusters in this case is approximately $0.3n$ for all n.

In Table 6.2 we show the statistics for the number of orthogonal Jacobi sweeps for Algorithm SSVD and the number of hyperbolic Jacobi sweeps for the J-orthogonal algorithm. These data correspond to the use of left-Jacobi in step 3 of Algorithm 4. If right-Jacobi is used, the average number of sweeps for Algorithm SSVD is 13.8, with a maximum of 28 for $n = 100$, while the accuracy is the same. For these reasons, we have used in the rest of our experiments the left-handed version of the algorithm. It can be seen that the J-orthogonal algorithm uses more sweeps than Algorithm SSVD: on average, from 5 more for $n = 50$ to almost 4 for $n = 1000$.

Now we focus on the analysis of data both for eigenvectors and for bases of invariant subspaces. Table 6.3 shows the quantities $\Xi_\Sigma^{(S)}$ and $\Xi_\Lambda^{(S)}$ defined in (70) for Algorithm 1, in both versions: SSVD0, using Algorithm 2, and SSVD, using Algorithm 3. For the J-orthogonal algorithm we only show the quantity that governs its error: $\Xi_\Lambda^{(S)}$. When comparing the results of routines SSVD0 and SSVD with the corresponding relative gaps of singular values (rows 1 and 3), it can be seen that both methods behave as expected. When comparing the errors in the bases computed using the routine SSVD0 with the relative gap between eigenvalues, the results can go rather poorly (see row 2).[9] When using SSVD these results improve significantly (compare rows 4 and 2), showing that the method computes the bases for these test matrices with errors depending on the relative gap between eigenvalues, as the J-orthogonal algorithm does. It can be observed that the control quantities increase mildly with n for all the algorithms. Since this effect is not observed in the accuracy of the eigenvalues, this lead us to question if it is a real effect of the eigenvector bounds or is simply reflecting the fact that the quantities Ξ are computed from n-dimensional vectors.

Table 6.4 shows the quantities $\xi_\sigma^{(S)}$ and $\xi_\lambda^{(S)}$ defined in (71). These are the quantities referring to the errors eigenvector by eigenvector. It can be seen that the accuracy of the eigenvectors is not spoiled by the clustering processes implicit in Algorithms

[9] However, as can be deduced from the mean value of $\Xi_\Lambda^{(\text{SSVD0})}$, matrices for which SSVD0 computes eigenvectors with a large error with respect to the relative gap between eigenvalues are quite infrequent.

TABLE 6.3
Experiment 1. *Statistical data for accuracy in bases of invariant subspaces.*

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
$\Xi_{\Sigma}^{(\texttt{SSVDO})}$	.032	.46	.051	1.2	.084	2.5	.12	4.5	.17	4.4
$\Xi_{\Lambda}^{(\texttt{SSVDO})}$	.37	320	1.1	3300	2.4	500	6.5	1700	5.6	150
$\Xi_{\Sigma}^{(\texttt{SSVD})}$	.034	.50	.056	1.2	.095	2.5	.13	4.5	.18	4.4
$\Xi_{\Lambda}^{(\texttt{SSVD})}$	.041	.65	.075	4.6	.15	3.2	.23	6.0	.37	7.3
$\Xi_{\Lambda}^{(\texttt{J-O})}$	.044	.64	.076	1.5	.15	2.6	.21	5.7	.32	7.3

TABLE 6.4
Experiment 1. *Statistical data for accuracy in eigenvectors:* $\xi_{\sigma}^{(\mathrm{S})}$ *and* $\xi_{\lambda}^{(\mathrm{S})}$.

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
$\xi_{\sigma}^{(\texttt{SSVDO})}$	.033	.74	.057	1.3	.092	2.5	.13	4.5	.19	4.4
$\xi_{\lambda}^{(\texttt{SSVDO})}$	.37	320	1.1	3300	2.4	500	6.5	1700	5.6	150
$\xi_{\sigma}^{(\texttt{SSVD})}$	.035	.90	.063	1.6	.10	2.5	.14	4.5	.20	4.4
$\xi_{\lambda}^{(\texttt{SSVD})}$	.045	.90	.089	4.6	.17	3.2	.26	6.0	.42	7.3
$\xi_{\lambda}^{(\texttt{J-O})}$	.044	.64	.076	1.5	.15	2.6	.21	5.7	.32	7.3

`SSVD` and `SSVDO`. Comments similar to those made with respect to Table 6.3 apply here.

To conclude, we show other quantities of numerical interest. The minimum singular value and eigenvalue relative gaps for clusters selected in Algorithm 2 have exceeded, respectively, 10^{-5} and 10^{-4}, and after the clustering process in Algorithm 3 both relative gaps, for eigenvalues and singular values, have been bigger than 10^{-4}. The minimum relative gap for individual eigenvalues has been greater than 10^{-5}, and for singular values greater than 10^{-8}. The maximum values of $\kappa(R')$ have been 190 for $n = 50$, 270 for $n = 100$, 600 for $n = 250$, 1300 for $n = 500$, and 2200 for $n = 1000$, showing that it increases roughly as some constant times n.

EXPERIMENT 2. We have generated $n \times n$ graded matrices $A = DBD$ by multiplying random well-conditioned matrices, B, and random ill-conditioned diagonal matrices, D, to test the accuracy of the complete Algorithm 1 including the factorization in step 1. Not always can an accurate RRD fulfilling (10) be computed for graded matrices [6, section 4]: the accuracy that can be guaranteed at best (and is frequently achieved in practice) is $O(\epsilon_s \kappa(B))$. Thus, high relative accuracy is expected when computing eigenvalues and eigenvectors for the matrices generated in this experiment. As mentioned in section 6.1, the initial RRD in Algorithm 1 has been done in two ways: using either a modification of the symmetric indefinite BP decomposition or a nonsymmetric LU factorization with complete pivoting. We have obtained similar results for both decompositions. Parameters have been chosen as follows: $\kappa(B) = 10^{[0:1:3]}$, $\kappa(D) = 10^{[2:2:10]}$, $MODE_B = 3, 4, 5$, $MODE_D = \pm 3, \pm 4, 5$. For each set of parameters we have run 50 matrices for $n = 50, 100$ (total 15000 matrices for each n), 5 for $n = 250, 500$ (total 1500 matrices for each n), 1 for $n = 1000$, and only for 5 combinations of the $MODE$s (total 100 matrices). As announced, Jacobi and QR also have been applied on these test matrices.

The same quantities as in Experiment 1 are shown in Table 6.5 for eigenvalues and in Table 6.6 for individual eigenvectors. The results for bases of invariant subspaces

TABLE 6.5
Experiment 2. *Statistical data for accuracy in eigenvalues:* $\vartheta^{(\mathrm{S})}$.

n	50		100		250	
Method	mean	max	mean	max	mean	max
$\vartheta^{(\texttt{SSVD})}$	1.8	2600	.82	1100	.21	52
$\vartheta^{(\texttt{J-O})}$	1.5	1100	.80	1200	.21	64
$\vartheta^{(\texttt{JAC})}$	$3 \cdot 10^{15}$	$3 \cdot 10^{19}$	$1 \cdot 10^{14}$	$3 \cdot 10^{17}$	$1 \cdot 10^{13}$	$7 \cdot 10^{15}$
$\vartheta^{(\texttt{QR})}$	$2 \cdot 10^{13}$	$2 \cdot 10^{17}$	$7 \cdot 10^{11}$	$5 \cdot 10^{15}$	$5 \cdot 10^{10}$	$4 \cdot 10^{13}$
$\vartheta^{(\texttt{SVD})}$	1.8	2600	.82	1100	.21	52

n	500		1000	
Method	mean	max	mean	max
$\vartheta^{(\texttt{SSVD})}$	.22	140	.014	.24
$\vartheta^{(\texttt{J-O})}$	.31	320	.019	.33
$\vartheta^{(\texttt{JAC})}$	$7 \cdot 10^{12}$	$5 \cdot 10^{15}$	$2 \cdot 10^{11}$	$8 \cdot 10^{12}$
$\vartheta^{(\texttt{QR})}$	$2 \cdot 10^{10}$	$1 \cdot 10^{13}$	$2 \cdot 10^{3}$	$4 \cdot 10^{4}$
$\vartheta^{(\texttt{SVD})}$	.22	140	.014	.24

TABLE 6.6
Experiment 2. *Statistical data for accuracy in eigenvectors:* $\xi_\sigma^{(\mathrm{S})}$ *and* $\xi_\lambda^{(\mathrm{S})}$.

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
$\xi_\sigma^{(\texttt{SSVDO})}$	.47	11	.28	4.6	.17	1.1	.064	.55	.023	.16
$\xi_\lambda^{(\texttt{SSVDO})}$	3.6	3300	2.8	1900	1.2	1600	.30	14	.067	.51
$\xi_\sigma^{(\texttt{SSVD})}$	.47	11	.31	5.2	.20	1.1	.076	1.3	.024	.16
$\xi_\lambda^{(\texttt{SSVD})}$	.56	12	.34	5.8	.25	2.4	.091	1.3	.030	.16
$\xi_\lambda^{(\texttt{J-O})}$	.60	21	.37	4.3	.17	1.2	.090	.67	.039	.20

are almost the same as those in Table 6.6 and, therefore, are not shown. In these tables we show only the data corresponding to symmetric RRDs obtained by the BP method. The corresponding data for these tables using the unsymmetric RRD based on GECP are so similar that they are omitted. Nevertheless for other quantities (see Tables 6.7 and 6.8) we show the results for both decompositions (GECP is abbreviated as CP in the tables).

Notice that the maximum values in Table 6.5 are greater than in Experiment 1, for both Algorithm 1 and the J-orthogonal algorithm. This is due to the error in the initial factorization step, which is roughly bounded by $O(\epsilon_s \kappa(B))$. In any case, they behave much better than the classical methods, Jacobi and QR. An interesting remark is that the quantities $\vartheta^{(\mathrm{S})}$ decrease in Table 6.5 as n increases. This is because in this experiment (see Table 6.7) the condition number κ increases with the dimension n faster than the relative errors $e_\lambda^{(S)}$ in the eigenvalues. The control quantities for eigenvectors in Table 6.6 also decrease with n for the same reason. However, the maximum values of the control quantities for eigenvalues (Table 6.5) are much bigger than those of eigenvectors (Table 6.6). This is not explained by the error bounds.

As in Experiment 1, for a good number of the generated matrices (310 matrices out of 15000 for $n = 50$; 4821 matrices out of 15000 for $n = 100$; 1019 matrices out of 1500 for $n = 250$; 1454 matrices out of 1500 for $n = 500$; 100 matrices out of 100 for $n = 1000$), there are clusters of singular values of dimension greater than 1, according to criterion (65), with a maximal dimension of 5. The average number of clusters has ranged from almost no clusters for $n = 50$ to approximately 60 clusters

TABLE 6.7
Experiment 2. *Table for* $\kappa(R')$ *and* $M_\kappa = \max\{\kappa(X), \kappa(Y)\}$.

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
$\kappa(R')$(BP)	11	39	23	84	67	220	150	430	330	960
$\kappa(R')$(CP)	11	37	24	80	71	201	160	450	360	860
$\kappa(X)$ (BP)	100	500	300	1300	1400	5000	4300	16000	14000	40000
M_κ (CP)	78	320	230	1000	1000	3200	2900	7900	5000	20000

TABLE 6.8
Experiment 2. *Statistical data for the number of sweeps.*

n	50		100		250		500		1000	
Method	mean	max	mean	max	mean	max	mean	max	mean	max
$Sweeps^{\texttt{(SSVD)BP}}$	5.0	7	5.6	8	6.4	9	7.3	9	8.1	9
$Sweeps^{\texttt{(SSVD)CP}}$	5.0	7	5.5	8	6.4	9	7.2	9	8.0	9
$Sweeps^{\texttt{(J-O)}}$	6.3	8	7.1	10	8.5	11	9.6	12	11.0	13

for $n = 1000$ with a typical dimension of 2. This shows again that criterion (65) determines perfectly in practice the signs of the eigenvalues, even when clusters are present. After applying Algorithm 3.1 all the considered matrices have clusters. The average number of clusters has been in this case around $0.3n$ for all n.

In addition, we show other quantities of numerical interest. The minimum singular value and eigenvalue relative gaps for clusters selected in Algorithm 2 are, respectively, 10^{-5} and $3.3 \cdot 10^{-4}$; and after the clustering process in Algorithm 3 both relative gaps, for eigenvalues and singular values, have reached the minimum $3.3 \cdot 10^{-4}$. The minimum relative gap for individual eigenvalues has been $4.1 \cdot 10^{-5}$, and for singular values greater than $9.1 \cdot 10^{-8}$. With respect to the condition numbers $\kappa(X)$, $\max\{\kappa(X), \kappa(Y)\}$ and $\kappa(R')$, they are shown in Table 6.7. The maximum values of $\epsilon\kappa(X)\kappa(R')$ are $8 \cdot 10^{-4}$ for $n = 50$, $4 \cdot 10^{-3}$ for $n = 100$, $5 \cdot 10^{-2}$ for $n = 250$, $3 \cdot 10^{-1}$ for $n = 500$, and 1.8 for $n = 1000$, showing that it increases roughly as some constant times n.

Table 6.8 shows that the J-orthogonal algorithm uses again more sweeps than Algorithm 1: on average, from one more for $n = 50$ to three more for $n = 1000$. This is reflected in the run-time used by the different routines. Taking as a reference the time employed by the QR routine (`SSYEV` of LAPACK), we have the following average results for our experiments: For $n = 100$, Algorithm `SSVD` (with symmetric RRD factorization) employs 200% more time than QR, the J-orthogonal algorithm employs 250% more time, and the Jacobi algorithm `SJAC` employs 190% more time; for $n = 500$, Algorithm `SSVD` (with symmetric RRD factorization) employs 380% more time, the J-orthogonal algorithm employs 350% more time, and the Jacobi algorithm `SJAC` employs 340% more time. These numbers can be explained as coming from two opposite effects: `SSVD` uses less Jacobi sweeps, but the number of clusters increases with the size of the matrix.

EXPERIMENT 3. We have also generated full matrices in another form to compare the accuracy of Algorithms 1 and J-orthogonal. We have used the matrix generator developed in [22], which is specifically designed to test the performance of the J-orthogonal algorithm on matrices for which the error bounds of this algorithm are controlled (see [22] for details).

The set of parameters has been chosen as follows: $n = 100$; `ASCAL` $= [1 : 1 : 3]$;

TABLE 6.9
Experiment 3. *Statistical data.*

	ϑ		ξ_σ		ξ_λ		*Sweeps*	
Method	mean	max	mean	max	mean	max	mean	max
`SSVD`	.27	2.2	2.1	14	2.9	21	4.6	6
`J-O`	.47	2.8	–	–	3.1	20	5.5	8

`HSCAL`$= [5 : 2 : 25]$.[10] For each set of parameters we have run 50 matrices, in total 1650 matrices.

The results confirm that Algorithm `SSVD` performs very well also for matrices of this type. The results for eigenvalues, eigenvectors, and number of sweeps are summarized in Table 6.9. As in the other experiments, the results for individual eigenvectors, $\xi_{\sigma,\lambda}^{(\mathtt{S})}$, are similar to those for bases. For this set of matrices, no clusters of singular values with dimension greater than 1 were found in the sense of criterion (65).

EXPERIMENT 4. The results for testing the accuracy of computed eigenvectors in previous experiments seem to show that the errors for the `SSVD` and J-orthogonal algorithms are comparable (see rows 4 and 5 of Tables 6.4, 6.6 and columns 6–7 of Table 6.9 in Experiment 3), both depending on the relative gap between eigenvalues. However, it should not be forgotten that the error bound for eigenvectors in the `SSVD` algorithm is given by the expressions (4) and (5) (or, more precisely, Theorem 5.12) and not (11). It is not difficult to think of situations in which Algorithm `SSVD` can calculate single eigenvectors much worse than the J-orthogonal algorithm. Take for example the following 3×3 very well conditioned matrix generated in single precision:

$$A = \begin{bmatrix} .1804019 & .9148742 & -.3611555 \\ .9148742 & -.2908984 & -.2799287 \\ -.3611555 & -.2799287 & -.8894936 \end{bmatrix}$$

with eigenvalues $\lambda_1 = 0.9999904633563307$, $\lambda_2 = -0.9999802814301686$, and $\lambda_3 = -1.000000302456291$ in double precision. The corresponding computed eigenvectors in single precision have the following errors for the `SSVD` algorithm:

$$[\|q_i - q_i^{(\mathrm{SSVD})}\|_2]_{i=1,2,3} = [3.12,\ 5.25,\ 4.23] \times 10^{-3}$$

and

$$[\|q_i - q_i^{(\mathrm{J-O})}\|_2]_{i=1,2,3} = [3.79 \times 10^{-5},\ 1.43,\ 1.43] \times 10^{-3}$$

for the J-orthogonal algorithm. Notice that the J-orthogonal algorithm computes the eigenvector corresponding to the positive eigenvalue λ_1 with full machine precision, while with the `SSVD` algorithm five significant decimal digits are lost. The reason for this is easily understood, because the eigenvalue relative gap for λ_1 is 1, while the corresponding singular value relative gap is near 10^{-5} (in this case relative or absolute gaps are equivalent). This cannot be improved by the clustering process done in Algorithm 3.1, because any of the two possible clusters of singular values containing one positive and one negative eigenvalue has a close singular value at a distance of order 10^{-5}, and the minimum of the eigenvalue relative gaps is also of order 10^{-5}.

[10] The routine `GENSYM` generates a nonsingular symmetric matrix H of order n, with $\kappa(H) \approx 10^{\mathtt{HSCAL}}$ and the measure $C(A, \widehat{A}) \approx 10^{\mathtt{ASCAL}}$ (see [22] for details).

However, notice that the SSVD algorithm is able to compute all the eigenvectors with three correct decimal digits and that $\max_i e_{q_i}^{(\mathtt{SSVD})} / \max_i e_{q_i}^{(\mathtt{J\text{-}O})} = 3.7$, of order 1 as predicted by the bound (5); i.e., the J-orthogonal algorithm also computes some eigenvectors with three correct significant digits.

Finally, notice that if all the eigenvalues of the matrix A are considered inside the same cluster, the SSVD algorithm computes the eigenvector corresponding to λ_1 with full machine precision, according to the bound (51). However, the eigenvectors corresponding to the negative eigenvalues are computed with errors of order 1, although according to (51) they form a very accurate orthonormal basis of the invariant subspace associated with the negative eigenvalues.

EXPERIMENT 5. Our last experiment is designed to show how the SSVD algorithm, like the J-orthogonal one, is able to compute accurate bases of invariant subspaces, even when the gaps between eigenvalues are very small.

We generate a 10×10 matrix $A = QDQ^T$ by multiplying, in single precision, a single precision random orthogonal matrix Q by the diagonal matrix $D = \mathrm{diag}[-1, 1, 1, 1, 1, 0.1, 0.1, 0.1, 0.1, 0.1]$. Due to roundoff errors, the absolute values of all the eigenvalues of A become different. But two clusters of singular values are found according to criterion (65), one around 1, of dimension 5, and another around 0.1, of the same dimension. Since one of the clusters is unsigned, Algorithm 3.1 does not change these clusters. The absolute gaps between the singular values inside each cluster exceed 10^{-7}. Thus the double precision routine DSYEVJ computes the eigenvectors with at least eight correct decimal digits. The SSVD and J-orthogonal algorithms, in single precision, compute all the eigenvectors with errors of $O(1)$, except the eigenvector corresponding to the negative eigenvalue which is computed, in both cases, with an error near 10^{-7}. This error is predicted by bound (51) for the SSVD algorithm (see also the remarks after the proof of Theorem 4.7). The errors in the invariant subspaces can be estimated using $E_{\Lambda_i}^{(\mathrm{S})}$ in (68). These, for SSVD and J-orthogonal algorithms, are of order 10^{-7} for the following invariant subspaces: the subspace corresponding to the four positive eigenvalues close to 1; the subspace corresponding to the five positive eigenvalues close to 0.1; and the subspace corresponding to the negative eigenvalue. Moreover, the same errors appear if we consider the invariant subspace corresponding to all the eigenvalues of absolute value around 1 (including the negative one). This shows in practice that, as studied in the error analysis leading to Theorem 4.7, once a cluster of singular values is chosen, we obtain two bases, one for the invariant subspace corresponding to the positive eigenvalues in the cluster and another for the negative ones, with an error of the same order as the one appearing in the basis of the singular subspace corresponding to the whole cluster of singular values.

7. Conclusions and future work. In this paper we have presented formal error analysis and numerical experiments of a new algorithm which computes eigenvalues and eigenvectors with high relative accuracy for the largest class of symmetric matrices known so far—in particular for all symmetric matrices belonging to the classes of general matrices studied in [6]. This high relative accuracy is achieved for a given symmetric matrix A whenever an accurate rank-revealing decomposition (RRD) of A can be computed.

The new algorithm is based on computing, in a first stage, a singular value decomposition (SVD) of the symmetric matrix A. This is the reason for its wide applicability, because in this stage the symmetry of A is not used. Thus, we can compute nonsymmetric RRDs of A and apply the theory developed in [6].

It is not known if accurate symmetric RRDs can be computed for all symmetric

matrices in any of the classes described in [6]. The J-orthogonal algorithm [26, 22] computes eigenvalues and eigenvectors with high relative accuracy *only* if symmetric RRDs that are accurate enough are available. The authors are presently studying this interesting question.

Appendix. Proof of Theorem 5.7. We begin with some previous elementary results that will be frequently used.

Let a and a' be any two real numbers. Then

$$\frac{a-a'}{a'} = \frac{\frac{a-a'}{a}}{1-\frac{a-a'}{a}} \quad \text{and} \quad \frac{a}{a'} = \frac{1}{1-\frac{a-a'}{a}}. \tag{72}$$

The following lemma bounds the relative distance between the maximum and the minimum elements in a cluster of tolerance C_l.

LEMMA A.1. *Let $\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i+d_1}\}$ be a cluster of tolerance C_l with d_1 elements. Then*

$$\frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+1}} \leq (d_1 - 1)\, C_l.$$

Proof. Notice that

$$\frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+1}} = \frac{\sigma_{i+1} - \sigma_{i+2}}{\sigma_{i+1}} + \frac{\sigma_{i+2} - \sigma_{i+3}}{\sigma_{i+1}} + \cdots + \frac{\sigma_{i+d_1-1} - \sigma_{i+d_1}}{\sigma_{i+1}}.$$

Thus

$$\frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+1}} \leq \frac{\sigma_{i+1} - \sigma_{i+2}}{\sigma_{i+1}} + \frac{\sigma_{i+2} - \sigma_{i+3}}{\sigma_{i+2}} + \cdots + \frac{\sigma_{i+d_1-1} - \sigma_{i+d_1}}{\sigma_{i+d_1-1}} \leq (d_1 - 1)\, C_l.$$

∎

Proof of Theorem 5.7. Let

$$\Sigma_1 = \{\sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i+d_1}\}, \quad \Sigma_2 = \{\sigma_{i+d_1+1}, \sigma_{i+d_1+2}, \ldots, \sigma_{i+d_1+d_2}\} \tag{73}$$

be the two clusters of singular values appearing in the statement of the theorem. Although in this setting the elements of Σ_1 are greater than the elements of Σ_2, the opposite case can be proved with the notation in (73) by interchanging the roles of Σ_1 and Σ_2.

Lemma 5.3 implies

$$rg(\Sigma_1 \cup \Sigma_2) = \min\left\{\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}, \frac{\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1+d_2}}\right\}, \tag{74}$$

and

$$\begin{aligned} &\min\{rg(\Sigma_1), rg(\Sigma_2)\} \\ &\quad = \min\left\{\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}, \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1}}, \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}}, \frac{\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1+d_2}}\right\}, \end{aligned} \tag{75}$$

where if some of the subindices do not belong to $\{1, \ldots, n\}$, the corresponding fraction does not appear. Therefore $rg(\Sigma_1 \cup \Sigma_2) \geq \min\{rg(\Sigma_1), rg(\Sigma_2)\}$, and the assumption (58) appearing in Theorem 5.7 leads to the following results:

1.

$$\min\{rg(\Sigma_1), rg(\Sigma_2)\} = \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1}}. \tag{76}$$

2.

$$rg(\Sigma_1) = \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1}}. \tag{77}$$

Thus in the setting (73), condition (58) implies that Σ_2 is the relative closest cluster to Σ_1 and it is not necessary to impose this condition explicitly. This has been done in the statement of Theorem 5.7 for the sake of clarity. Recall that one of the hypotheses of Theorem 5.7 is

$$rg(\Sigma_1) < t < 1. \tag{78}$$

The previous setting also allows us to prove Theorem 5.7 in the case in which the elements of Σ_1 are smaller than the elements of Σ_2 just by interchanging the roles of Σ_1 and Σ_2 in the statement of the theorem. Notice that condition $\mathrm{rg}\{\Sigma_1 \cup \Sigma_2\} > \min\{\mathrm{rg}\{\Sigma_1\}, \mathrm{rg}\{\Sigma_2\}\}$ remains unchanged, and therefore its consequences (76), (77) still hold. This, together with $rg(\Sigma_2) < t < 1$, leads to $rg(\Sigma_1) < t$, i.e., condition (78). Therefore, in the rest of the proof we will focus on the situation in (73) with assumptions (58) (and its consequences (76)–(77)) and (78).

Suppose that $(i + d_1 + d_2 + 1) \in \{1, \ldots, n\}$. If $\lambda_{\Pi(i+d_1+d_2+1)}$ is either zero or has the same sign as the elements of Λ_2, then $rg(\Lambda_2) \leq (\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1})/\sigma_{i+d_1+d_2}$. Otherwise $\lambda_{\Pi(i+d_1+d_2+1)}$ has the same sign as the elements of Λ_1, and then $rg(\Lambda_1) \leq (\sigma_{i+d_1} - \sigma_{i+d_1+d_2+1})/\sigma_{i+d_1}$. In any case

$$\begin{aligned} \min\{rg(\Lambda_1), rg(\Lambda_2)\} &\leq \max\left\{ \frac{\sigma_{i+d_1} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1}}, \frac{\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1+d_2}} \right\} \\ &= \frac{\sigma_{i+d_1} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1}}. \end{aligned} \tag{79}$$

Suppose now that i belongs to the set $\{1, \ldots, n\}$. If $\lambda_{\Pi(i)}$ has the same sign as the elements of Λ_1, then $rg(\Lambda_1) \leq (\sigma_i - \sigma_{i+1})/\sigma_{i+1}$. Otherwise $\lambda_{\Pi(i)}$ has the same sign as the elements of Λ_2, and then $rg(\Lambda_2) \leq (\sigma_i - \sigma_{i+d_1+1})/\sigma_{i+d_1+1}$. In any case

$$\min\{rg(\Lambda_1), rg(\Lambda_2)\} \leq \max\left\{ \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}, \frac{\sigma_i - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}} \right\} = \frac{\sigma_i - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}}. \tag{80}$$

Once (79) and (80) have been established, it only remains to prove

$$\frac{\sigma_{i+d_1} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1}} \leq R \, \frac{\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1+d_2}} \tag{81}$$

and

$$\frac{\sigma_i - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}} \leq R \, \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}, \tag{82}$$

where

$$R = \frac{1}{1-t}\left(1 + \frac{1}{1-(d-1)\,C_l} + \frac{1}{1-(d-1)\,C_l}\,\frac{(d-1)\,C_l}{rg(\Sigma_1 \cup \Sigma_2)}\right).$$

If these two inequalities hold, then (79) and (80) imply that $\min\{rg(\Lambda_1), rg(\Lambda_2)\}$ is bounded simultaneously by the right-hand side of (81) and the right-hand side of (82). Thus using (74), Theorem 5.7 is finally proved.

Proof of (81). Notice that

(83)
$$\frac{\sigma_{i+d_1} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1}} = \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1}} + \frac{\sigma_{i+d_1+1} - \sigma_{i+d_1+d_2}}{\sigma_{i+d_1}} + \frac{\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1}}.$$

The first term of the right-hand side in the previous equation is less than $(\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1})/\sigma_{i+d_1+d_2}$, due to (76) and (75). The third term is trivially bounded by the same quantity, since $\sigma_{i+d_1} > \sigma_{i+d_1+d_2}$. For the second term,

$$\frac{\sigma_{i+d_1+1} - \sigma_{i+d_1+d_2}}{\sigma_{i+d_1}} < \frac{\sigma_{i+d_1+1} - \sigma_{i+d_1+d_2}}{\sigma_{i+d_1+1}} \leq (d_2 - 1)C_l,$$

where the last inequality is just Lemma A.1 applied to Σ_2. Plugging these bounds into (83) and using $rg(\Sigma_1 \cup \Sigma_2) \leq (\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1})/\sigma_{i+d_1+d_2}$, we obtain

$$\frac{\sigma_{i+d_1} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1}} \leq \left(2 + \frac{(d_2 - 1)C_l}{rg(\Sigma_1 \cup \Sigma_2)}\right) \frac{\sigma_{i+d_1+d_2} - \sigma_{i+d_1+d_2+1}}{\sigma_{i+d_1+d_2}}.$$

The first factor of the right-hand side is bounded by R and (81) follows. ∎

Proof of (82). Notice that

$$\frac{\sigma_i - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}} = \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+d_1+1}} + \frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+d_1+1}} + \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}}. \tag{84}$$

Now we will bound the three terms in the right-hand side of (84). We begin with the last one: using the first equality in (72), (77), (78), and (76), we get

$$\frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1+1}} < \frac{1}{1-t} \frac{\sigma_{i+d_1} - \sigma_{i+d_1+1}}{\sigma_{i+d_1}} < \frac{1}{1-t} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}. \tag{85}$$

For the second term, the first equality in (72) and Lemma A.1 yield

$$\frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+d_1+1}} = \frac{\sigma_{i+d_1}}{\sigma_{i+d_1+1}} \frac{\frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+1}}}{1 - \frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+1}}} \leq \frac{\sigma_{i+d_1}}{\sigma_{i+d_1+1}} \frac{(d_1 - 1)C_l}{1 - (d_1 - 1)C_l}.$$

The factor $\sigma_{i+d_1}/\sigma_{i+d_1+1}$ can be bounded by $1/(1-t)$, using the second equality in (72), (77), and (78). Therefore, the following bound for the second term of the right-hand side of (84) is obtained:

$$\frac{\sigma_{i+1} - \sigma_{i+d_1}}{\sigma_{i+d_1+1}} \leq \frac{1}{1-t} \frac{(d_1 - 1)C_l}{1 - (d_1 - 1)C_l}. \tag{86}$$

Finally, the first term verifies

$$\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+d_1+1}} = \frac{\sigma_{i+d_1}}{\sigma_{i+d_1+1}} \frac{\sigma_{i+1}}{\sigma_{i+d_1}} \frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}}.$$

The factor $\sigma_{i+d_1}/\sigma_{i+d_1+1}$ already has been bounded by $1/(1-t)$, while the factor $\sigma_{i+1}/\sigma_{i+d_1}$ is bounded by $1/(1-(d_1-1)C_l)$ by the second equality in (72) and Lemma A.1. Thus

$$\frac{\sigma_i-\sigma_{i+1}}{\sigma_{i+d_1+1}} \leq \frac{1}{1-t}\,\frac{1}{1-(d_1-1)C_l}\,\frac{\sigma_i-\sigma_{i+1}}{\sigma_{i+1}}. \tag{87}$$

Replacing (87), (86), and (85) in (84), and taking into account that $rg(\Sigma_1 \cup \Sigma_2) \leq (\sigma_i-\sigma_{i+1})/\sigma_{i+1}$,

$$\frac{\sigma_i-\sigma_{i+d_1+1}}{\sigma_{i+d_1+1}} \leq \frac{1}{1-t}\left(1+\frac{1}{1-(d_1-1)\,C_l}+\frac{1}{1-(d_1-1)\,C_l}\,\frac{(d_1-1)\,C_l}{rg(\Sigma_1\cup\Sigma_2)}\right)\frac{\sigma_i-\sigma_{i+1}}{\sigma_{i+1}}$$

is obtained. Now inequality (82) is easily proved. □

Acknowledgments. The authors thank Professor Zlatko Drmač, who provided the source code for the one-sided Jacobi SVD routine employed in the experiments. As can be seen in the numerical tests in section 6, the performance of his code is excellent. The authors thank also Professor J. W. Demmel for providing the source code of the routines used in [5].

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[3] 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.

[4] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[5] J. Demmel, *Accurate singular value decompositions of structured matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 562–580.

[6] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[7] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[8] F. M. Dopico, *A note on sin Θ theorems for singular subspace variations*, BIT, 40 (2000), pp. 395–403.

[9] F. M. Dopico and J. Moro, *Perturbation theory for simultaneous bases of singular subspaces*, BIT, 42 (2002), pp. 84–109.

[10] F. M. Dopico, J. M. Molera, and J. Moro, *An Orthogonal High Relative Accuracy Algorithm for the Symmetric Eigenproblem*, Tech. report, available online at http://www.uc3m.es/uc3m/dpto/MATEM/molera/indice.html.

[11] F. M. Dopico, J. M. Molera, and J. Moro, *A note on multiplicative backward errors of accurate SVD algorithms*, submitted.

[12] Z. Drmač, *Implementation of Jacobi rotations for accurate singular value computation in floating point arithmetic*, SIAM J. Sci. Comput., 18 (1997), pp. 1200–1222.

[13] Z. Drmač, *Accurate computation of the product-induced singular value decomposition with applications*, SIAM J. Numer. Anal., 35 (1998), pp. 1969–1994.

[14] Z. Drmač, *A posteriori computation of the singular vectors in a preconditioned Jacobi SVD algorithm*, IMA J. Numer. Anal., 19 (1999), pp. 191–213.

[15] Z. Drmač and K. Veselić, *Approximate eigenvectors as preconditioner*, Linear Algebra Appl., 309 (2000), pp. 191–215.

[16] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
[18] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[19] R.-C. Li, *Relative perturbation theory:* I. *Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.
[20] R.-C. Li, *Relative perturbation theory:* II. *Eigenspace and singular subspace variations*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 471–492.
[21] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.
[22] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph.D. thesis, Fachbereich Mathematik Fernuniversität, Gesamthochschule Hagen, Germany, 1992.
[23] I. Slapničar, *Componentwise analysis of direct factorization of real symmetric and Hermitian matrices*, Linear Algebra Appl., 272 (1998), pp. 227–275.
[24] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.
[25] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[26] K. Veselić, *A Jacobi eigenreduction algorithm for definite matrix pairs*, Numer. Math., 64 (1993), pp. 241–269.
[27] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 352–361

THE PARTER–WIENER THEOREM: REFINEMENT AND GENERALIZATION*

CHARLES R. JOHNSON†, ANTÓNIO LEAL DUARTE‡, AND CARLOS M. SAIAGO§

Abstract. An important theorem about the existence of principal submatrices of a Hermitian matrix whose graph is a tree, in which the multiplicity of an eigenvalue increases, was largely developed in separate papers by Parter and Wiener. Here, the prior work is fully stated, then generalized with a self-contained proof. The more complete result is then used to better understand the eigenvalue possibilities of reducible principal submatrices of Hermitian tridiagonal matrices. Sets of vertices, for which the multiplicity increases, are also studied.

Key words. eigenvalues, Hermitian matrix, multiplicity, parter vertices, tree, vertex degrees

AMS subject classifications. 15A18, 15A57, 05C50

DOI. 10.1137/10.1137/S0895479801393320

Let T be a tree on n vertices $1, 2, \ldots, n$, and suppose that $\mathcal{S}(T)$ is the set of all $n \times n$ complex Hermitian matrices whose graph is T; the diagonal of $A \in \mathcal{S}(T)$ is not restricted by T. (All results also apply to $n \times n$ matrices $A = (a_{ij})$ for which $a_{ij}a_{ji} > 0$ when $\{i, j\}$ is an edge of T and the diagonal of A is real. Such matrices are diagonally similar to Hermitian matrices with the same graph.) For a complex Hermitian $n \times n$ matrix A, we denote the multiplicity of λ as an eigenvalue of A by $m_A(\lambda)$, and if $\alpha \subseteq N = \{1, \ldots, n\}$ is an index set, we denote the principal submatrix of A resulting from deletion (retention) of the rows and columns α by $A(\alpha)$ ($A[\alpha]$). Often, α will consist of a single index i, and we abbreviate $A(\{i\})$ by $A(i)$. If $A = (a_{ij})$, identify $A[\{i\}]$ with a_{ii}. Note that when $A \in \mathcal{S}(T)$, the subgraph of T induced by deletion of vertex v, $T - v$, corresponds, in a natural way, to $A(v)$. In particular, $A(v)$ is a direct sum whose summands correspond to components of $T - v$ (which we call *branches* of T at v), the number of summands or components being the degree of v ($\deg v$) in T. We will often identify (such parts of) T with (such parts of) A for convenience. Throughout, for $\deg v = k + 1$, we identify the neighbors of a vertex v in T as $u_0, u_1, \ldots, u_k$, and we denote the branch of T resulting from deletion of v and containing u_i by T_i, $i = 0, 1, \ldots, k$.

According to the interlacing theorem for Hermitian eigenvalues [2], there is a simple relationship between $m_{A(i)}(\lambda)$ and $m_A(\lambda)$ when A is Hermitian:

$$m_{A(i)}(\lambda) = m_A(\lambda) - 1 \text{ or } m_{A(i)}(\lambda) = m_A(\lambda) \text{ or } m_{A(i)}(\lambda) = m_A(\lambda) + 1.$$

It is natural to imagine that the first possibility is generic, and, for sufficiently full Hermitian matrices A, it probably is. However, in [8] a very surprising observation

*Received by the editors August 6, 2001; accepted for publication (in revised form) by R. Nabben December 4, 2002; published electronically August 19, 2003.
http://www.siam.org/journals/simax/25-2/39332.html

†Department of Mathematics, College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (crjohnso@math.wm.edu).

‡Departmento de Matemática, Universidade de Coimbra, Apartado 3008, 3001-454 Coimbra, Portugal (leal@mat. uc.pt). The research of this author was supported by Centro de Matemática da Universidade de Coimbra.

§Departmento de Matemática, Facultad de Ciências e Tecnologia da Universidade Nova de Lisboa, 2829-516 Monte Caparica, Portugal (cls@fct.unl.pt). The research of this author was supported in part by Fundação para a Ciência e a Tecnologia, Portugal through the research grant SFRH/BD/899/2000. Part of this research was done while the author was visiting the College of William and Mary.

was made: If T is a tree and $A \in \mathcal{S}(T)$ and $m_A(\lambda) \geq 2$, then there is a vertex i such that $m_{A(i)}(\lambda) \geq 3$ and λ is an eigenvalue of at least three components (branches) of $A(i)$. In particular, if $m_A(\lambda) = 2$, $m_{A(i)}(\lambda) = m_A(\lambda) + 1$! In [10] it was further shown that if $m_A(\lambda) \geq 2$, then there is a vertex i such that $m_{A(i)}(\lambda) = m_A(\lambda) + 1$.

We note that the principal results of neither [8] nor [10] apply when T is a path, as, then, $m_A(\lambda) > 1$ cannot occur. For self-containment, we give a simple proof of this known fact later, and our generalization of [8] and [10] will apply to this case.

It is curious that Parter did not identify the multiplicity increase for all values of $m_A(\lambda) \geq 2$ and Wiener did not explicitly identify the distribution of the eigenvalue among at least three branches, both of which are important, though it appears that each author might have, given the machinery they developed. When just one vertex is removed, we note that the "three branches" cannot generally be improved upon, as there are trees with maximum degree 3 and arbitrarily high possible multiplicities [1], [3]. However, as we shall see in Theorem 14, the "three branches" may be improved by removing more vertices.

These results have been important to us in our recent works on possible multiplicities of eigenvalues among matrices in $\mathcal{S}(T)$ [3], [4], [5], [6], [7]. Although not explicitly stated by either, we feel it appropriate to attribute the following theorem to Parter and Wiener.

THEOREM 1 (PW-theorem). *Let T be a tree on n vertices and suppose that $A \in \mathcal{S}(T)$ and that $\lambda \in \mathbb{R}$ is such that $m_A(\lambda) \geq 2$. Then, there is a vertex i of T such that $m_{A(i)}(\lambda) = m_A(\lambda) + 1$ and λ occurs as an eigenvalue in direct summands of A that correspond to at least three branches of T at i.*

Besides focusing attention on the complete statement of Theorem 1, our purpose here is to give a generalization of it (the PW-theorem will be a special case) and to apply the generalization in a few ways. We give new and rather complete information about the relationship between the eigenvalues of a tridiagonal Hermitian matrix and those of a principal submatrix of size one smaller. Our approach also gives a clear identification of the elements necessary in a proof of the original observations.

We call a vertex i in T a (*weak*) *Parter vertex* for $\lambda \in \mathbb{R}$ and $A \in \mathcal{S}(T)$ when $m_{A(i)}(\lambda) = m_A(\lambda) + 1$ and call a collection $\alpha \subseteq N$ a *Parter set* when $m_{A(\alpha)}(\lambda) = m_A(\lambda) + |\alpha|$. We also examine when a collection of Parter vertices is a Parter set, and related issues. We also have used the term (*strong*) *Parter vertex* for one satisfying the conclusion of Theorem 1 elsewhere, but this will not be needed here. That a collection of Parter vertices need not be a Parter set is noted by example in [9].

Our generalization of the PW-theorem follows.

THEOREM 2. *Let A be a Hermitian matrix whose graph is a tree T, and suppose that there exists a vertex v of T and a real number λ such that $\lambda \in \sigma(A) \cap \sigma(A(v))$. Then*

(a) *there is a vertex v' of T such that $m_{A(v')}(\lambda) = m_A(\lambda) + 1$;*
(b) *if $m_A(\lambda) \geq 2$, then v' may be chosen so that $\deg v' \geq 3$ and so that there are at least three components T_1, T_2, and T_3 of $T - v'$ such that $m_{A[T_i]}(\lambda) \geq 1$, $i = 1, 2, 3$;*
(c) *if $m_A(\lambda) = 1$, then v' may be chosen so that $\deg v' \geq 2$ and so that there are two components T_1 and T_2 of $T - v'$ such that $m_{A[T_i]}(\lambda) = 1$, $i = 1, 2$.*

Before continuing, we note that, even when $m_A(\lambda) \geq 2$, it can happen that $\deg v' = 1$ or $\deg v' = 2$ or λ appears in only one or two components of $T - v'$ even when $\deg v' \geq 3$. Of course, it also can happen that v does not qualify as a v' (v need

not increase the multiplicity of λ). Examples are easily constructed and some appear in [9].

Naturally, in the PW-theorem case ($m_A(\lambda) \geq 2$), $m_{A(v)}(\lambda) \geq 1$, so that our hypothesis is automatically satisfied for any v. Thus, Theorem 1 is a special case of Theorem 2.

The proof of Theorem 2 rests, in part, on two key lemmas, but first we record (and prove, for completeness) a well-known fact that we shall use.

LEMMA 3. *If A is a Hermitian matrix whose graph is a path on n vertices, then for any $\lambda \in \sigma(A)$, $m_A(\lambda) = 1$.*

Proof. Up to permutation similarity, A, and thus $A - \lambda I$, is tridiagonal. Since A is irreducible, the result of deletion of the first column and last row of $A - \lambda I$ is an $(n-1) \times (n-1)$ lower triangular matrix with nonzero diagonal, which is, therefore, nonsingular and rank $n-1$. Since rank cannot increase by extracting a submatrix, $\operatorname{rank}(A - \lambda I) = n - 1$, and, as A is Hermitian, $m_A(\lambda) = 1$. ∎

LEMMA 4. *Let A be a Hermitian matrix whose graph is a tree T. If there is a vertex v of T and a real number λ such that $\lambda \in \sigma(A) \cap \sigma(A(v))$, then there are adjacent vertices v' and u of T such that the component T_0 of $T - v'$ containing u satisfies $m_{A[T_0]}(\lambda) = m_{A[T_0 - u]}(\lambda) + 1$.*

Proof. We argue by induction on the number n of vertices of T. For convenience, we actually prove a slightly stronger statement by adding to the induction hypothesis the statement that v is not a vertex of T_0. If $n = 1$ or $n = 2$, the claimed implication is correct because it is not possible for the hypothesis to be satisfied, as may be easily checked. If $n = 3$, then T is a path and it can be easily checked that the hypothesis is satisfied only if A is a tridiagonal matrix whose first and last diagonal entries are both λ and v is the middle vertex. Then, taking v' to be the middle vertex v and u to be either the first or last vertex shows that the conclusion is satisfied (as the empty matrix cannot have λ as an eigenvalue).

Now, suppose that the claim is valid for all trees on fewer than n vertices, $n > 3$, and consider a tree on n vertices and a Hermitian matrix A such that there is a vertex v such that $\lambda \in \sigma(A) \cap \sigma(A(v))$. First, try letting v' be the vertex v. If there is a neighbor u_j of v such that $m_{A[T_j]}(\lambda) \geq 1$ and $m_{A[T_j - u_j]}(\lambda) = m_{A[T_j]}(\lambda) - 1$, we are done. If not, there are, by the hypothesis, neighbors u_j such that $m_{A[T_j]}(\lambda) \geq 1$, and, by replacing v with u_j and applying induction, the claim follows. ∎

The second lemma may be proven in two different ways, each giving different insights. We give one proof here, and another may be found in [9]. Both proofs rely on an expansion of the characteristic polynomial for Hermitian matrices whose graphs are trees. First, focus on a particular vertex v of T with neighbors $u_0, \ldots, u_k$ and expand $p_A(t)$ along the corresponding row of $A = (a_{ij})$ to obtain

$$p_A(t) = (t - a_{vv}) \prod_{j=0}^{k} p_{A[T_j]}(t) - \sum_{j=0}^{k} |a_{vu_j}|^2 p_{A[T_j - u_j]}(t) \prod_{\substack{l=0 \\ l \neq j}}^{k} p_{A[T_l]}(t), \tag{1}$$

and also

$$p_{A(T_i)}(t) = (t - a_{vv}) \prod_{\substack{j=0 \\ j \neq i}}^{k} p_{A[T_j]}(t) - \sum_{\substack{j=0 \\ j \neq i}}^{k} |a_{vu_j}|^2 p_{A[T_j - u_j]}(t) \prod_{\substack{l=0 \\ l \neq i,j}}^{k} p_{A[T_l]}(t). \tag{2}$$

(Here, we observe the standard convention that the characteristic polynomial of the empty matrix is identically 1.)

It will be convenient to focus on the identified neighbor u_0 of v and rewrite (1) and (2) by letting

$$f(t) = \sum_{j=1}^{k} |a_{vu_j}|^2 p_{A[T_j - u_j]}(t) \prod_{\substack{l=1 \\ l \neq j}}^{k} p_{A[T_l]}(t)$$

and

$$g(t) = \prod_{j=1}^{k} p_{A[T_j]}(t)$$

to obtain

$$p_A(t) = (t - a_{vv}) p_{A[T_0]}(t) g(t) - |a_{vu_0}|^2 g(t) p_{A[T_0 - u_0]}(t) - f(t) p_{A[T_0]}(t) \tag{3}$$

and

$$p_{A(T_0)}(t) = (t - a_{vv}) g(t) - f(t). \tag{4}$$

We also have a useful form for $p_A(t)$ when we focus on the edge connecting v and u_0. Denote by T_v the tree $T - T_0$. We have $A(T_0) = A[T_v]$ and $g(t) = p_{A[T_v - v]}(t)$. From (3),

$$p_A(t) = \left[(t - a_{vv}) g(t) - f(t)\right] p_{A[T_0]}(t) - |a_{vu_0}|^2 g(t) p_{A[T_0 - u_0]}(t),$$

and with (4) we obtain

$$p_A(t) = p_{A[T_v]}(t) p_{A[T_0]}(t) - |a_{vu_0}|^2 p_{A[T_v - v]}(t) p_{A[T_0 - u_0]}(t). \tag{5}$$

Using these expansions we now prove the following lemma.

LEMMA 5. *Let A be a Hermitian matrix, whose graph is a tree T. If there is a vertex v of T and a real number λ for which*

$$\lambda \in \sigma(A) \cap \sigma(A(v)),$$

and there is a branch T_0 of T at v such that

$$m_{A[T_0 - u_0]}(\lambda) = m_{A[T_0]}(\lambda) - 1,$$

in which u_0 is the neighbor of v in T_0, then

$$m_{A(v)}(\lambda) = m_A(\lambda) + 1.$$

Proof. We employ (3) and (4) above, with v and u_0 corresponding to the hypothesis of the lemma. First, note that $p_{A(v)}(t) = p_{A[T_0]}(t) g(t)$. Let $m = m_A(\lambda)$ and $m_0 = m_{A[T_0]}(\lambda)$, so that $m_{A[T_0 - u_0]}(\lambda) = m_0 - 1$. (We note that if it happens that $m_0 = m + 1$, the conclusion is immediate. Although the proof is technically correct in any event, it may be convenient to assume $m_0 \leq m$.) Also, let m_f and m_g be the multiplicities of λ as a root of f and g, respectively. Since removal of u_0 from T leaves $A(T_0) \oplus A[T_0 - u_0]$, by the interlacing inequalities and the assumption that $m_{A[T_0 - u_0]}(\lambda) = m_0 - 1$, λ is a root of $p_{A(T_0)}$ at least $m - m_0$ times. Also by interlacing, $m - m_0 - 1 \leq m_g \leq m - m_0 + 1$. If $m_g = m - m_0 + 1$, we would be done;

so consider the other two possibilities. In either event, $m_f \geq m_g$ by a divisibility argument applied to (4). Returning to (3), we find that if $m_g = m - m_0$, a divisibility argument would contradict our hypothesis, as all terms except $|a_{vu_0}|^2 p_{A[T_0-u_0]}(t)g(t)$ would be divisible by $(t-\lambda)^m$. Suppose $m_g = m - m_0 - 1$. Then $m_f = m_g$, or else a divisibility argument would contradict the fact that λ is a root of $p_{A(T_0)}(t)$ at least $m - m_0$ times. However, then a divisibility argument applied to (3) leads to a contradiction, as λ is a root of the left-hand side and the first and third terms on the right at least $m-1$ times each, but only $m-2$ times in the second term on the right. □

Although we have made the statement in the form we wish to apply it, we note that the statement of Lemma 5 remains correct (trivially) if the hypothesis "$\lambda \in \sigma(A) \cap \sigma(A(v))$" is replaced by the weaker "$\lambda \in \sigma(A(v))$."

Another proof of this key lemma is given in [9]. This proof uses (5) and focuses primarily on the nature of the neighbor u_0. See also [10] for a variant of Lemma 5 and a different approach.

We next turn to a proof of Theorem 2.

Proof of Theorem 2. If $m_A(\lambda) \geq 2$, the first part of the hypothesis of Lemma 5 is satisfied for any vertex of T, in particular the vertex v' guaranteed by Lemma 4. In this event, the entire hypothesis of Lemma 5 holds, verifying part (a) of the theorem. If $m_A(\lambda) = 1$ (and $m_{A(v)}(\lambda) \geq 1$), we still have from Lemma 4 the existence of v', and since $m_{A(v')}(\lambda) \geq m_{A[T_0]}(\lambda) = m_{A[T_0-u]}(\lambda) + 1 \geq 1$, we have $m_A(\lambda) \geq 1$ and $m_{A(v')}(\lambda) \geq 1$. Thus, v' in place of v satisfies the hypothesis of Lemma 5 and part (a) of the theorem still holds.

For part (b) we argue by induction on the number n of vertices of T. If $n \leq 3$, the claimed implication is correct because it is not possible that the hypothesis is satisfied, as may be easily checked, or simply apply Lemma 3, as any tree on $n \leq 3$ vertices is a path.

If $n = 4$, the only tree on four vertices that is not a path is a star (one vertex of degree 3 and three pendant vertices). Since $m_A(\lambda) = m \geq 2$, there is a vertex v in T such that $m_{A(v)}(\lambda) = m + 1$. In that case, v must be the central vertex (the vertex of degree 3), since for any other vertex u, $T - u$ is a path. Thus, $T - v$ is a graph consisting of three isolated vertices with $m_{A(v)}(\lambda) \leq 3$. Therefore, $m = 2$ and $m_{A(v)}(\lambda) = 3$; i.e., λ is an eigenvalue of three components of $T - v$.

Now, suppose that the claimed result is valid for all trees on fewer than n vertices, $n > 4$, and consider a tree T on n vertices and a Hermitian matrix $A \in \mathcal{S}(T)$ such that λ is an eigenvalue of A with $m_A(\lambda) = m \geq 2$. By part (a) of Theorem 2, there is a vertex v in T such that $m_{A(v)}(\lambda) = m + 1$. If λ is an eigenvalue of at least three components of $T - v$, we are done. If not, there are two possible situations: λ is an eigenvalue of two components of $T - v$ (case 1) or λ is an eigenvalue of one component of $T - v$ (case 2).

In case 1, there is a component T' of $T - v$ with λ as an eigenvalue of $A[T']$ and $m_{A[T']}(\lambda) \geq 2$. Applying induction to T', we have a vertex u in T' such that $m_{A[T'-u]}(\lambda) = m_{A[T']}(\lambda) + 1$ and λ is an eigenvalue of at least three components of $T' - u$. Observe that $m_{A(\{v,u\})}(\lambda) = m + 2$; thus, by interlacing, $m_{A(u)}(\lambda) = m + 1$. Consider the (unique) shortest path between v and u in T, P_{vu}, and let (v, u) denote the component of $T - \{v, u\}$ containing vertices of P_{vu}. Note that (v, u) is one of the components of $T' - u$ (if not empty). If there are three components of $T' - u$ having λ as an eigenvalue and none of these is (v, u), then these three components are also components of $T - u$, and we are done. If there are only three components of $T' - u$

having λ as an eigenvalue and one of them is (v, u) then, by interlacing applied to the component of $T - u$ containing v, since $T - v$ has another component with λ as an eigenvalue, $T - u$ still has three components with λ as an eigenvalue.

In case 2, there is a component T' of $T - v$ with λ as an eigenvalue of $A[T']$ and $m_{A[T']}(\lambda) = m_A(\lambda) + 1$. Applying induction to T', we have a vertex u in T' such that $m_{A[T'-u]}(\lambda) = m_{A[T']}(\lambda) + 1$ and λ is an eigenvalue of at least three components of $T' - u$. By interlacing, $m_{A(u)}(\lambda) = m+1$. Thus, if there are three components of $T' - u$ having λ as an eigenvalue and none of these is (v, u), then these three components are also components of $T - u$ and we are done. If there are only three components of $T' - u$ having λ as an eigenvalue and one of these components is (v, u), we may apply case 1 to complete the consideration of case 2.

For part (c), the only contrary possibility is that λ is an eigenvalue of multiplicity 2 of one of the direct summands of $A(v')$. But, in this event, we may replace v' with the vertex adjacent to it in the corresponding branch of $T - v'$ and continue such replacement until a v' of the desired sort is found. □

COROLLARY 6. *Let T be a tree and A be a matrix of $\mathcal{S}(T)$. If for some vertex v of T, λ is an eigenvalue of $A(v)$, then there is a vertex v' of T such that $m_{A(v')}(\lambda) = m_A(\lambda) + 1$.*

Proof. Suppose that λ is an eigenvalue of $A(v)$ for some vertex v of T. If λ is not an eigenvalue of A, then setting $v' = v$, $m_{A(v')}(\lambda) = m_A(\lambda) + 1$. If λ is also an eigenvalue of A, by Theorem 2, there is a vertex v' of T such that $m_{A(v')}(\lambda) = m_A(\lambda) + 1$. □

It has been mentioned in several prior works (e.g., [1], [4], [5]) that for a tree, the multiplicities of the largest and smallest eigenvalues are 1. It is an interesting question to characterize those trees for which there is a matrix with just two eigenvalues of multiplicity 1, and to determine for each tree the minimum number of eigenvalues of multiplicity 1 that can occur (it may be much more than two). Here, we give another (simple) proof about the multiplicities of the largest and smallest eigenvalues.

COROLLARY 7. *If T is a tree, the largest and smallest eigenvalues of each $A \in \mathcal{S}(T)$ have multiplicity 1. Moreover, the largest or smallest eigenvalue of a matrix $A \in \mathcal{S}(T)$ cannot occur as an eigenvalue of a submatrix $A(v)$, for any vertex v of T.*

Proof. Let T be a tree and λ be the smallest (largest) eigenvalue of a matrix $A \in \mathcal{S}(T)$. Suppose that there is a vertex v of T such that λ is an eigenvalue of $A(v)$. By Theorem 2, there is a vertex v' of T such that $m_{A(v')}(\lambda) = m_A(\lambda) + 1$. But, from the interlacing inequalities, since λ is the smallest (largest) eigenvalue of A, for any vertex i of T, $m_{A(i)}(\lambda) \leq m_A(\lambda)$, which gives a contradiction. Thus, λ cannot occur as an eigenvalue of any submatrix $A(i)$ of A. Therefore, $m_A(\lambda) = 1$. □

Lemma 5 indicates that a neighboring vertex, in whose branch the multiplicity goes down, is important for the existence of a Parter vertex. We call a branch at v in the direction of u_0, satisfying the requirement $m_{A[T_0]}(\lambda) = m_{A[T_0-u_0]}(\lambda) + 1$, of Lemma 5 a *downer branch* at v for the eigenvalue λ; the vertex u_0 is called a *downer vertex*. According to Lemma 5, the existence of a downer branch is sufficient for a vertex to be Parter. Importantly, the existence of a downer branch is also necessary for a vertex to be Parter, which provides a precise structural mechanism for recognition of Parter vertices. Notice that, even when $m_A(\lambda) = 0$, if there is a downer branch at v, then $m_{A(v)}(\lambda) = 1$. It cannot be more by interlacing, nor less because $A[T_0]$ is a direct summand of $A(v)$.

THEOREM 8. *For $A \in \mathcal{S}(T)$, T a tree, and v a vertex of T, $m_{A(v)}(\lambda) = m_A(\lambda)+1$*

if and only if there is a downer branch at v for λ.

Proof. The sufficiency follows from Lemma 5 and the comment preceding the statement of this theorem.

For necessity, return to (1). Suppose that none of $u_0, u_1, \ldots, u_k$ is a downer vertex. Then, the number of times λ is a root of the second term on the right is at least the number of times that λ is a root of $p_{A(v)}(t)$ (i.e., $\prod_{i=0}^{k} p_{A[T_i]}(t)$). Thus, $m_A(\lambda)$ is, at least, $m_{A(v)}(\lambda)$, and v could not be Parter. ∎

By Corollary 7, a branch of T at v having λ as the smallest (largest) eigenvalue is automatically a downer branch, so that we may make the following observation using Theorem 8.

COROLLARY 9. *Let A be a Hermitian matrix whose graph is a tree T. If λ is the smallest (largest) eigenvalue of at least one of the direct summands of $A(v)$, then $m_{A(v)}(\lambda) = m_A(\lambda) + 1$.*

We note that, extending the divisibility argument of the proof of Theorem 8, if each neighbor of v is Parter in its branch, then v cannot be Parter. We note also that if u_i is Parter in its branch, then it is Parter in T, as its downer branch within its branch will be a downer branch in T.

Let T be a path on n vertices and $A \in \mathcal{S}(T)$. Theorem 2 allows us to give information about the relationship between the eigenvalues of A and those of a principal submatrix of size one smaller. A path on n vertices admits a labeling $1, 2, \ldots, n$ such that, for $i = 1, \ldots, n-1$, $\{i, i+1\}$ is an edge. Without loss of generality, if T is a path, we shall assume this labeling of the vertices, giving an irreducible tridiagonal matrix, in terms of which, for convenience, we now make several observations. The first is a classical fact that now follows here in quite a different way.

COROLLARY 10. *If A is an $n \times n$ irreducible tridiagonal Hermitian matrix, then the eigenvalues of $A(1)$ and $A(n)$ strictly interlace those of A.*

Proof. We induct on n. For $n \leq 3$, the validity of the claim was mentioned in the proof of Lemma 4. Assume now the claim for tridiagonal matrices of size less than n. By symmetry, we need only verify the claim for $A(n)$. Suppose to the contrary that $\lambda \in \sigma(A) \cap \sigma(A(n))$. By the induction hypothesis, $\lambda \notin \sigma(A(\{n-1, n\}))$, so that $n-1$ is a downer vertex for λ at n. By Theorem 8, then, $m_{A(n)}(\lambda) = 1 + 1$, a contradiction to Lemma 3, as the graph of $A(n)$ is again a path. ∎

By Corollary 10, a pendant path with λ as an eigenvalue is also a downer branch, so that we may make the following observation using Theorem 8.

COROLLARY 11. *Let A be a Hermitian matrix whose graph is a tree T. If at least one of the direct summands of $A(v)$ has λ as an eigenvalue, and its graph is a path and a neighbor of v is a pendant vertex of this path, then $m_{A(v)}(\lambda) = m_A(\lambda) + 1$.*

A new observation is now immediate.

COROLLARY 12. *If A is an $n \times n$ irreducible, tridiagonal, Hermitian matrix, then $\lambda \in \sigma(A) \cap \sigma(A(i))$ if and only if $1 < i < n$ and $m_{A(i)}(\lambda) = 2$, with $\lambda \in \sigma(A[\{1, \ldots, i-1\}])$ and $\lambda \in \sigma(A[\{i+1, \ldots, n\}])$.*

From Corollary 12 and Lemma 3, we immediately have the following.

COROLLARY 13. *Let A be an $n \times n$ irreducible, tridiagonal, Hermitian matrix. Then there are at most $\min\{i-1, n-i\}$ different eigenvalues that are common to both A and $A(i)$, i.e., at most $\min\{i-1, n-i\}$ equalities in the interlacing inequalities.*

We note that Corollary 13 is sharp. If $A[\{1, \ldots, i-1\}]$ and $A[\{i+1, \ldots, n\}]$ have $\min\{i-1, n-i\}$ eigenvalues in common (which may always be arranged), then the upper bound on the number of the interlacing inequalities will be attained. Smaller numbers also may be designed.

Remark. We note that if A is an irreducible, tridiagonal, Hermitian matrix, then an interpretation of Corollary 12 is the following. If any common eigenvalues of A and $A(i)$ are deleted from $\sigma(A)$ and $\sigma(A(i))$ (only once each in the latter case), then the latter strictly interlaces the former.

We now turn to a more detailed discussion of the structure and size of Parter sets.

THEOREM 14. *Let A be a Hermitian matrix whose graph is a tree T and let λ be an eigenvalue of A. Then, there is a vertex v of T such that $\lambda \in \sigma(A) \cap \sigma(A(v))$ if and only if there is a Parter set S of cardinality $k \geq 1$ such that λ is an eigenvalue of $m_A(\lambda) + k$ direct summands of $A(S)$.*

Proof. Let λ be an eigenvalue of A. Suppose that $S = \{v_1, \ldots, v_k\}$, $k \geq 1$, is a Parter set of λ such that λ is an eigenvalue of $m_A(\lambda) + k$ direct summands of $A(S)$. By the interlacing inequalities, for the multiplicity to increase by k, it would have to increase by 1 with the removal of each vertex, starting with any one; i.e., each vertex of S is a Parter vertex. Thus if $v \in S$, then $m_{A(v)}(\lambda) = m_A(\lambda) + 1$. Therefore, there is a vertex v of T such that $\lambda \in \sigma(A) \cap \sigma(A(v))$.

For the converse, suppose that v is a vertex of T such that $\lambda \in \sigma(A) \cap \sigma(A(v))$. By Theorem 2, there is a vertex v_1 of T such that $m_{A(v_1)}(\lambda) = m_A(\lambda) + 1$ and, if $m_A(\lambda) = 1$, λ is an eigenvalue of two direct summands of $A(v_1)$ or, if $m_A(\lambda) \geq 2$, then λ is an eigenvalue of at least three direct summands of $A(v_1)$. So, if $m_A(\lambda) = 1$ or $m_A(\lambda) = 2$, the claimed result follows directly from Theorem 2. Now, suppose that $m_A(\lambda) \geq 3$. If λ is an eigenvalue of $m_A(\lambda) + 1$ direct summands of $A(v_1)$, we are done. If not, λ is an eigenvalue of less than $m_A(\lambda) + 1$ direct summands of $A(v_1)$. This means that λ is still a multiple eigenvalue of some direct summands of $A(v_1)$. Since each direct summand of $A(v_1)$ is a Hermitian matrix whose graph is a subtree of T, applying recursively Theorem 2 we find vertices $v_2, \ldots, v_k$ of T such that $m_{A(\{v_1,\ldots,v_k\})}(\lambda) = m_{A(\{v_1,\ldots,v_{k-1}\})}(\lambda) + 1$ and λ is not a multiple eigenvalue of any direct summands of $A(\{v_1, \ldots, v_k\})$; i.e., setting $S = \{v_1, \ldots, v_k\}$, $m_{A(S)}(\lambda) = m_A(\lambda) + k$ and λ is an eigenvalue of $m_A(\lambda) + k$ direct summands of $A(S)$. □

In Corollary 11, we noted that if $\lambda \in \sigma(A)$, $A \in \mathcal{S}(T)$, and there is a pendant path in T with λ as an eigenvalue, then that pendant path is a downer branch for λ in T. Of course, by Lemma 3, λ has multiplicity 1 in this downer branch. It is possible to show by example that there may be no multiplicity 1 downer branch in T that is a path, but it is not difficult to show, using Theorem 14 and induction, that there is always a multiplicity 1 downer branch for λ in T, $A \in \mathcal{S}(T)$, $\lambda \in \sigma(A) \cap \sigma(A(v))$. We have the following.

COROLLARY 15. *Let $A \in \mathcal{S}(T)$ and suppose that there is a vertex v of T such that $\lambda \in \sigma(A) \cap \sigma(A(v))$. Then, there is a Parter vertex v' of T such that for at least one of its downer branches T_0 for λ at v', $m_{A[T_0]}(\lambda) = 1$.*

If λ is a multiple eigenvalue of A, there is a Parter vertex v for λ such that $m_{A(v)}(\lambda) = m_A(\lambda) + 1$. It can occur that λ is an eigenvalue of $m_A(\lambda) + 1$ direct summands of $A(v)$ but, for example, if $\deg v < m_A(\lambda)+1$, necessarily λ is an eigenvalue of less than $m_A(\lambda) + 1$ direct summands of $A(v)$.

COROLLARY 16. *Let A be a Hermitian matrix whose graph is a tree T and let λ be an eigenvalue of A. If S is a Parter set for λ of cardinality k such that λ is an eigenvalue of $m_A(\lambda) + k$ direct summands of $A(S)$, and $v \in S$ is a Parter vertex for λ of degree less than $m_A(\lambda) + 1$, then $k > 1$.*

THEOREM 17. *Let A be a Hermitian matrix whose graph is a tree T and let λ be an eigenvalue of A. Also, let $d_1 \geq \cdots \geq d_n$ be the vertex degree sequence of T and S be a Parter set for λ of cardinality k such that λ is an eigenvalue of $m_A(\lambda) + k$ direct*

summands of $A(S)$ (each exactly once). Then, for $1 \leq p \leq r$, in which $d_r > 1$ and $d_{r+1} = 1$,

$$\sum_{i=1}^{p} d_i \leq m_A(\lambda) + 2(p-1)$$

implies $k > p$.

Proof. By hypothesis, λ is an eigenvalue of $m_A(\lambda) + k$ direct summands of $A(S)$. This means that the number of components of $T - S$ is, at least, $m_A(\lambda) + k$. Let $v_1, \ldots, v_k$ be the vertices of S. The number of components of $T - S$, c_S, is $1 + \sum_{i=1}^{k}(\deg v_i - 1) - e$, in which e is the number of edges in the subgraph of T induced by S. It is clear that $0 \leq e \leq k-1$. Therefore, $c_S \leq 1 + \sum_{i=1}^{k}(\deg v_i - 1)$. Recall that c_S must be, at least, $m_A(\lambda) + k$ and, observe that, since $d_1 \geq \cdots \geq d_n$, $1 + \sum_{i=1}^{k}(\deg v_i - 1) \leq 1 + \sum_{i=1}^{k}(d_i - 1)$. Thus, if for $p \geq 1$ ($d_p > 1$),

$$1 + \sum_{i=1}^{p}(d_i - 1) \leq m_A(\lambda) + p - 1,$$

i.e.,

$$\sum_{i=1}^{p} d_i \leq m_A(\lambda) + 2(p-1),$$

then $k > p$. ∎

We conclude with a general lower bound for the cardinality of a Parter set of the special type guaranteed in Theorem 14.

COROLLARY 18. *Let A be a Hermitian matrix whose graph is a tree T and let λ be an eigenvalue of A. Let $d_1 \geq \cdots \geq d_n$ be the degree sequence of the vertices of T and S be a Parter set for λ of cardinality k such that λ is an eigenvalue of $m_A(\lambda) + k$ direct summands of $A(S)$. Then, $k \geq q$, in which q is the first integer such that $\sum_{i=1}^{q} d_i > m_A(\lambda) + 2(q-1)$.*

If we let K_q be a maximum number of components remaining after removal of q vertices, then in the language of [5], a lower bound on the cardinality of such a Parter set is the first value of q such that $K_q \geq m_A(\lambda) + q$.

REFERENCES

[1] J. GENIN AND J. MAYBEE, *Mechanical vibration trees*, J. Math. Anal. Appl., 45 (1974), pp. 746–763.

[2] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[3] C. R. JOHNSON AND A. LEAL-DUARTE, *The maximum multiplicity of an eigenvalue in a matrix whose graph is a tree*, Linear Multilinear Algebra, 46 (1999), pp. 139–144.

[4] C. R. JOHNSON AND A. LEAL-DUARTE, *On the possible multiplicities of the eigenvalues of an Hermitian matrix whose graph is a given tree*, Linear Algebra Appl., 348 (2002), pp. 7–21.

[5] C. R. JOHNSON, A. LEAL-DUARTE, C. M. SAIAGO, B. D. SUTTON, AND A. J. WITT, *On the relative position of multiple eigenvalues in the spectrum of an Hermitian matrix with a given graph*, Linear Algebra Appl., 363 (2003), pp. 147–159.

[6] C. R. JOHNSON AND C. M. SAIAGO, *Estimation of the maximum multiplicity of an eigenvalue in terms of the vertex degrees of the graph of a matrix*, Electron. J. Linear Algebra, 9 (2002), pp. 27–31.

[7] A. LEAL-DUARTE AND C. R. JOHNSON, *On the minimum number of distinct eigenvalues for a symmetric matrix whose graph is a given tree*, Math. Inequal. Appl., 5 (2002), pp. 175–180.

[8] S. Parter, *On the eigenvalues and eigenvectors of a class of matrices*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 376–388.

[9] C. M. Saiago, *The Possible Multiplicities of the Eigenvalues of an Hermitian Matrix Whose Graph Is a Tree,* Ph.D. Thesis, *Universidade Nova de Lisboa*, Monte Caparica, Portugal, 2003.

[10] G. Wiener, *Spectral multiplicity and splitting results for a class of qualitative matrices*, Linear Algebra Appl., 61 (1984), pp. 15–29.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 362–372

NEW PERTURBATION BOUNDS FOR UNITARY POLAR FACTORS*

WEN LI† AND WEIWEI SUN‡

Abstract. In this paper, we present some new perturbation bounds for polar decompositions in the Frobenius norm. We prove that under any active condition of the perturbation being small, our bounds always improve previous bounds. Some perturbation bounds in the spectral norm and general unitarily invariant norms are also given.

Key words. perturbation bound, polar decomposition, singular value decomposition

AMS subject classifications. 65F10, 15A45

DOI. 10.1137/S0895479802413625

1. Introduction.

Let $\mathcal{C}^{m\times n}$ ($\mathcal{R}^{m\times n}$) be the set of $m\times n$ complex (real) matrices and $\mathcal{C}_r^{m\times n}$ ($\mathcal{R}_r^{m\times n}$) be the set of $m\times n$ complex (real) matrices having rank r. Here we always assume that $m\geq n$. We denote by $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|$ the spectral norm, the Frobenius norm, and a general unitarily invariant norm, respectively. Let

$$A = U\begin{pmatrix}\Sigma_1 & 0\\ 0 & 0\end{pmatrix}V^* \tag{1.1}$$

be the singular value decomposition (SVD) of A and

$$H = V_1\Sigma_1 V_1^*, \qquad Q = U_1V_1^*, \tag{1.2}$$

where $U=(U_1,U_2)\in\mathcal{C}^{m\times m}$ and $V=(V_1,V_2)\in\mathcal{C}^{n\times n}$ are unitary, $U_1\in\mathcal{C}_r^{m\times r}$, $V_1\in\mathcal{C}_r^{n\times r}$, $\Sigma_1=\mathrm{diag}(\sigma_1,\ldots,\sigma_r)$, and $\sigma_i, i=1,2,\ldots,r$, are the singular values of A with $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r>0$. The polar decomposition of the matrix A is defined by

$$A = QH. \tag{1.3}$$

The matrix Q is called the unitary polar factor of A. When $r=m=n$, we have $U=U_1$ and $V=V_1$.

In this paper, we are mainly concerned with perturbation bounds for the unitary polar factor. This problem has been studied by many authors; e.g., see Barrlund [1], Bhatia [2, 4], Bhatia and Mukherjea [3], Chatelin and Gratton [5], Li [9], Li and Sun [10], Mathias [11], and Sun and Chen [13]. Let $A\in\mathcal{C}_r^{m\times n}$ and $\widetilde{A}\in\mathcal{C}_r^{m\times n}$ with

$$A = QH, \qquad \widetilde{A} = \widetilde{Q}\widetilde{H}.$$

*Received by the editors August 28, 2002; accepted for publication (in revised form) by R. Bhatia February 18, 2003; published electronically August 19, 2003.
http://www.siam.org/journals/simax/25-2/41362.html

†Department of Mathematics, South China Normal University, Guangzhou, 510631, People's Republic of China (liwen@scnu.edu.cn). The work of this author was supported in part by the CityU research grant 7001331, Nature Science Foundation of University Teachers of Guangdong Province, and Excellent Talent Foundation of Guangdong Province, China.

‡Department of Mathematics, City University of Hong Kong, Hong Kong, People's Republic of China (maweiw@math.cityu.edu.hk). The work of this author was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project CityU 1084/02P) and the CityU research grant 7001331.

The general form of these perturbation bounds is

$$\|Q - \widetilde{Q}\| \le C\|A - \widetilde{A}\|.$$

Here, C depends heavily upon the number field, the rank, and the dimension of A. For complex matrices with $m = n = r$, Mathias [11] gave the bound

$$\|Q - \widetilde{Q}\| \le -\frac{\|A - \widetilde{A}\|}{\|A - \widetilde{A}\|_2} \times \log\left(1 - \frac{\|A - \widetilde{A}\|_2}{\sigma_n}\right). \tag{1.4}$$

A sharper bound is

$$\|Q - \widetilde{Q}\| \le \frac{2}{\sigma_r + \widetilde{\sigma}_r}\|A - \widetilde{A}\|, \tag{1.5}$$

which was proved by Li [9] for the case $r = m = n$ and by Li and Sun [10] for the general case with the Frobenius norm. It was claimed in [9, 10] that the bound in (1.5) is optimal in some general sense and some examples were given to confirm the sharpness of the bound. The motivation of this paper is to present some new perturbation bounds for complex matrices. The bound in (1.5) is optimal for general matrices A and $\widetilde{A}$. However, the bound can be improved for many practical perturbation problems. We prove that any active condition of the perturbation being small results in an improvement of (1.5). In particular, we show that

$$\|Q - \widetilde{Q}\|_F^2 \le \frac{2}{\sigma_r^2 + \widetilde{\sigma}_r^2}\|A - \widetilde{A}\|_F^2 \tag{1.6}$$

for small perturbations. Some perturbation bounds in the spectral norm and general unitarily invariant norms are also given.

2. Notation and some lemmas. Let A, $\widetilde{A} \in \mathcal{C}_r^{m\times n}$, $m \ge n$, have the SVDs

$$A = U\Sigma V^* \quad \text{and} \quad \widetilde{A} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^*, \tag{2.1}$$

where

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{C}_r^{m\times n} \quad \text{and} \quad \widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{C}_r^{m\times n},$$

$\Sigma_1 = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$, $\widetilde{\Sigma}_1 = \operatorname{diag}(\widetilde{\sigma}_1, \dots, \widetilde{\sigma}_r)$, $\sigma_1 \ge \cdots \ge \sigma_r > 0$, and $\widetilde{\sigma}_1 \ge \cdots \ge \widetilde{\sigma}_r > 0$.

Let I_p be the $p \times p$ identity matrix and let

$$I_{m,n}^{(p)} \equiv \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

For simplicity we replace $I_{m,n}^{(p)}$ with $I^{(p)}$. Let $S = \widetilde{U}^*U$ and $T = \widetilde{V}^*V$ have the block forms

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \in \mathcal{C}^{m\times m} \quad \text{and} \quad T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \in \mathcal{C}^{n\times n},$$

where both S_{11} and T_{11} are $r \times r$. Then S and T are unitary matrices. Let

$$\begin{aligned} M &= 2I - S_{11}^*T_{11} - T_{11}^*S_{11}, & \widetilde{M} &= 2I - T_{11}S_{11}^* - S_{11}T_{11}^*, \\ W &= I^{(r)} - S^*I^{(r)}T, & \widetilde{W} &= I^{(r)} - SI^{(r)}T^*, \end{aligned} \tag{2.2}$$

and m_{ij}, $\widetilde{m}_{ij}$, w_{ij}, and $\widetilde{w}_{ij}$ denote the (i,j) entry of M, $\widetilde{M}$, W, and $\widetilde{W}$, respectively. Since

$$\begin{aligned} WW^* &= (I^{(r)} - S^* I^{(r)} T)(I^{(r)} - S^* I^{(r)} T)^* \\ &= \begin{bmatrix} I + S_{11}^* S_{11} - S_{11}^* T_{11} - T_{11}^* S_{11} & S_{11}^* S_{12} - T_{11}^* S_{12} \\ S_{12}^* S_{11} - S_{12}^* T_{11} & S_{12}^* S_{12} \end{bmatrix}, \end{aligned}$$

we have

$$\begin{aligned} \operatorname{tr}(WW^*) &= \operatorname{tr}(I + S_{11}^* S_{11} - S_{11}^* T_{11} - T_{11}^* S_{11}) + \operatorname{tr}(S_{12}^* S_{12}) \\ &= \operatorname{tr}(I + S_{11} S_{11}^* - S_{11}^* T_{11} - T_{11}^* S_{11} + S_{12} S_{12}^*) \\ &= \operatorname{tr}(M) \end{aligned}$$

by noting that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and $S^*S = SS^* = I$, where $\operatorname{tr}(A)$ denotes the trace of A. It follows that

$$\begin{aligned} &\|A - \widetilde{A}\| = \|S\Sigma - \widetilde{\Sigma} T\|, \\ &\|Q - \widetilde{Q}\|_F^2 = \|SI^{(r)} - I^{(r)} T\|_F^2 = \|W\|_F^2 = \operatorname{tr}(M) = \operatorname{tr}(\widetilde{M}), \\ &w_{ii} + w_{ii}^* = m_{ii}, \qquad \widetilde{w}_{ii} + \widetilde{w}_{ii}^* = \widetilde{m}_{ii}, \quad i \le r. \end{aligned} \tag{2.3}$$

Let

$$\Gamma = \Sigma - \sigma I^{(r)} \quad \text{and} \quad \widetilde{\Gamma} = \widetilde{\Sigma} - \sigma I^{(r)}.$$

Then

$$\begin{aligned} \|S\Sigma - \widetilde{\Sigma} T\|_F &= \sigma^2 \|Q - \widetilde{Q}\|_F^2 + \|S\Gamma - \widetilde{\Gamma} T\|_F^2 \\ &\quad + 2\sigma \mathcal{R}e \operatorname{tr}[(SI^{(r)} - I^{(r)} T)(S\Gamma - \widetilde{\Gamma} T)^*], \end{aligned} \tag{2.4}$$

where $\mathcal{R}e$ denotes the real part of a complex number.

Equation (2.4) has been studied by several authors. Li [8] proved that $\mathcal{R}e \operatorname{tr}[(SI^{(r)} - I^{(r)} T)(S\Gamma - \widetilde{\Gamma} T)^*]$ is nonnegative when $\sigma \le \min\{\sigma_r, \widetilde{\sigma}_r\}$. Based on this analysis, he gave the bound

$$\|Q - \widetilde{Q}\|_F \le \frac{1}{\min\{\sigma_r, \widetilde{\sigma}_r\}} \|A - \widetilde{A}\|_F.$$

Equation (2.4) was studied more precisely in our recent work [10], where we obtain

$$\|A - \widetilde{A}\|_F^2 = \sigma^2 \|Q - \widetilde{Q}\|_F^2 + \|S\Gamma - \widetilde{\Gamma} T\|_F^2 + \sigma \left(\sum_{i=1}^{r} (\sigma_i - \sigma) m_{ii} + \sum_{i=1}^{r} (\widetilde{\sigma}_i - \sigma) \widetilde{m}_{ii} \right). \tag{2.5}$$

The last term is nonnegative for general complex matrices when $\sigma \le (\sigma_r + \widetilde{\sigma}_r)/2$ and for real matrices with $r = n = m$ when $\sigma \le (\sigma_n + \widetilde{\sigma}_n + \sigma_{n-1} + \widetilde{\sigma}_{n-1})/4$. The bound in (1.5) with $r \le n \le m$ can be obtained by choosing $\sigma = (\sigma_r + \widetilde{\sigma}_r)/2$ in (2.4) and by noting that

$$\sum_{i=1}^{n} (\sigma_i - \sigma) m_{ii} + \sum_{i=1}^{n} (\widetilde{\sigma}_i - \sigma) \widetilde{m}_{ii} \ge (\sigma_{r-1} + \widetilde{\sigma}_{r-1} - 2\sigma) \operatorname{tr}(M) - (\sigma_{r-1} - \sigma_r) m_{rr} - (\widetilde{\sigma}_{r-1} - \widetilde{\sigma}_r) \widetilde{m}_{rr}.$$

Moreover, a simple bound for real matrices was also given in [10]. However, in all these works, the term $\|S\Gamma - \widetilde{\Gamma} T\|_F^2$ was ignored. Let

$$B = U \Sigma^b V^* \in \mathcal{C}^{m \times n}, \qquad \widetilde{B} = \widetilde{U} \widetilde{\Sigma}^b \widetilde{V}^* \in \mathcal{C}^{m \times n}$$

be two SVDs, where

$$\Sigma^b = \begin{bmatrix} \Sigma_1^b & 0 \\ 0 & 0 \end{bmatrix}, \qquad \Sigma_1^b = \operatorname{diag}(\sigma_1^b, \ldots, \sigma_r^b), \qquad \sigma_i^b \geq 0,$$

and $\widetilde{\Sigma}^b$ is defined similarly. The following lemma is an extension of Lemma 2.3 in [10] and can be proved along the same lines as in [10].

LEMMA 2.1. *Let B, $\widetilde{B} \in \mathcal{C}^{m\times n}$ be defined as above. Then*

(2.6)

$$\|B - \widetilde{B}\|_F^2 = \sigma^2\|Q - \widetilde{Q}\|_F^2 + \|S\Gamma^b - \widetilde{\Gamma}^b T\|_F^2 + \sigma\left(\sum_{i=1}^{r}(\sigma_i^b - \sigma)m_{ii} + \sum_{i=1}^{r}(\widetilde{\sigma}_i^b - \sigma)\widetilde{m}_{ii}\right),$$

where $\Gamma^b = \Sigma^b - \sigma I^{(r)}$ and $\widetilde{\Gamma}^b = \widetilde{\Sigma}^b - \sigma I^{(r)}$.

For a given $\sigma > 0$, we assume

$$\begin{aligned} &\sigma_1 - \sigma \geq \sigma_2 - \sigma \geq \cdots \geq \sigma_{k_1} - \sigma > 0 \geq \sigma_{k_1+1} - \sigma \geq \sigma_r - \sigma, \\ &\widetilde{\sigma}_1 - \sigma \geq \widetilde{\sigma}_2 - \sigma \geq \cdots \geq \widetilde{\sigma}_{k_2} - \sigma > 0 \geq \widetilde{\sigma}_{k_2+1} - \sigma \geq \widetilde{\sigma}_r - \sigma. \end{aligned}$$

Let $\epsilon_i = |\sigma_i - \sigma|$, $\widetilde{\epsilon}_i = |\widetilde{\sigma}_i - \sigma|$, and

$$\begin{aligned} &\epsilon_{i_1} \geq \epsilon_{i_2} \geq \cdots \geq \epsilon_{i_r} \geq 0, \\ &\widetilde{\epsilon}_{i_1} \geq \widetilde{\epsilon}_{i_2} \geq \cdots \geq \widetilde{\epsilon}_{i_r} \geq 0. \end{aligned}$$

It is easy to see that

$$\|S\Gamma - \widetilde{\Gamma}T\|_F = \|U\Gamma V^* - \widetilde{U}\widetilde{\Gamma}\widetilde{V}^*\|_F = \|(UD)\Gamma' V^* - (\widetilde{U}\widetilde{D})\widetilde{\Gamma}'\widetilde{V}^*\|_F, \tag{2.7}$$

where

$$\begin{aligned} &D = \operatorname{diag}(\overbrace{1, \ldots, 1}^{k_1}, -1, \ldots, -1, \overbrace{1, \ldots, 1}^{m-r}), \\ &\widetilde{D} = \operatorname{diag}(\overbrace{1, \ldots, 1}^{k_2}, -1, \ldots, -1, \overbrace{1, \ldots, 1}^{m-r}), \\ &\Gamma' = \begin{bmatrix} \Gamma_1' & 0 \\ 0 & 0 \end{bmatrix}, \qquad \widetilde{\Gamma}' = \begin{bmatrix} \widetilde{\Gamma}_1' & 0 \\ 0 & 0 \end{bmatrix}, \\ &\Gamma_1' = \operatorname{diag}(\sigma_1 - \sigma, \ldots, \sigma_{k_1} - \sigma, \sigma - \sigma_{k_1+1}, \ldots, \sigma - \sigma_r) = \operatorname{diag}(\epsilon_1, \ldots, \epsilon_r), \\ &\widetilde{\Gamma}_1' = \operatorname{diag}(\widetilde{\sigma}_1 - \sigma, \ldots, \widetilde{\sigma}_{k_2} - \sigma, \widetilde{\sigma} - \sigma_{k_2+1}, \widetilde{\sigma}_r - \sigma) = \operatorname{diag}(\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_r). \end{aligned}$$

We consider the two SVDs

$$A' = (UD)\Gamma' V^*, \qquad \widetilde{A}' = (\widetilde{U}\widetilde{D})\widetilde{\Gamma}'\widetilde{V}^*. \tag{2.8}$$

The corresponding perturbation is defined by $(UD)I^{(r)}V^* - (\widetilde{U}\widetilde{D})I^{(r)}\widetilde{V}^*$. An estimate for the perturbation is given in the following lemma.

LEMMA 2.2. *With the above notation, we have*

$$\begin{aligned} \|(UD)I^{(r)}V^* - (\widetilde{U}\widetilde{D})I^{(r)}\widetilde{V}^*\|_F^2 &= \|Q - \widetilde{Q}\|_F^2 + 4(r - k_1) + 4(r - k_2) \\ &\quad - \operatorname{tr}(ES^*\widetilde{E}T + T^*\widetilde{E}SE) - 2\sum_{i=k_1+1}^{r} m_{ii} - 2\sum_{i=k_2+1}^{r}\widetilde{m}_{ii}, \end{aligned} \tag{2.9}$$

where $E = I - D$ and $\widetilde{E} = I - \widetilde{D}$.

Proof. By the properties of the Frobenius norm,

$$\begin{aligned}&\|(UD)I^{(r)}V^* - (\widetilde{U}\widetilde{D})I^{(r)}\widetilde{V}^*\|_F^2 = \|I^{(r)} - S^*I^{(r)}T - (E - S^*\widetilde{E}T)\|_F^2\\ &= \|W\|_F^2 + \|E - S^*\widetilde{E}T\|_F^2 - \mathrm{tr}(W^*(E - S^*\widetilde{E}T) + (E - S^*\widetilde{E}T)^*W),\end{aligned}$$

where W is defined in (2.2). Moreover,

$$\|E - S^*\widetilde{E}T\|_F^2 = \|E\|_F^2 + \|S^*\widetilde{E}T\|_F^2 - \mathrm{tr}(ES^*\widetilde{E}T + T^*\widetilde{E}SE),$$

and since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we have

$$\begin{aligned}&\mathrm{tr}(W^*(E - S^*\widetilde{E}T) + (E - S^*\widetilde{E}T)^*W)\\ &\qquad = \mathrm{tr}((W + W^*)E) + \mathrm{tr}((\widetilde{W} + \widetilde{W}^*)\widetilde{E})\\ &\qquad = 2\sum_{i=k_1+1}^{r}(w_{ii} + w_{ii}^*) + 2\sum_{i=k_2+1}^{r}(\widetilde{w}_{ii} + \widetilde{w}_{ii}^*).\end{aligned}$$

Equation (2.9) follows immediately by noting (2.3). The proof is complete. ∎

3. Perturbation bounds. Let

$$S' = \begin{pmatrix} S'_{11} & S'_{12}\\ S'_{21} & S'_{22}\end{pmatrix} = (\widetilde{U}\widetilde{D})^*(UD), \qquad T' = \widetilde{V}^*V = T,$$

and

$$\begin{aligned}M' &= (m'_{ij}) = 2I - S'^*_{11}T_{11} - T^*_{11}S'_{11},\\ \widetilde{M}' &= (\widetilde{m}'_{ij}) = 2I - S'_{11}T^*_{11} - T_{11}S'^*_{11}.\end{aligned}$$

It is easy to see that

$$\mathrm{tr}(M') = \mathrm{tr}(\widetilde{M}') = \|(UD)I^{(r)}V^* - (\widetilde{U}\widetilde{D})I^{(r)}\widetilde{V}^*\|_F^2$$

and

$$\begin{aligned}M' &= 2I - D^*S^*_{11}\widetilde{D}T_{11} - T^*_{11}\widetilde{D}^*S_{11}D,\\ \widetilde{M}' &= 2I - \widetilde{D}^*S_{11}\widetilde{D}T^*_{11} - T_{11}D^*S^*_{11}\widetilde{D}.\end{aligned}$$

Applying Lemma 2.1 for the decompositions in (2.6) gives

$$\|(UD)\Gamma'V^* - (\widetilde{U}\widetilde{D})\widetilde{\Gamma}'\widetilde{V}^*\|_F^2 \geq \epsilon^2\mathrm{tr}(M') + \epsilon\left(\sum_{i=1}^{r}(\epsilon_i - \epsilon)m'_{ii} + \sum_{i=1}^{r}(\widetilde{\epsilon}_i - \epsilon)\widetilde{m}'_{ii}\right), \tag{3.1}$$

where $\epsilon > 0$.

Here we consider the case

$$\sigma = (\sigma_r + \widetilde{\sigma}_r)/2, \qquad \epsilon = (\epsilon_{i_r} + \widetilde{\epsilon}_{i_r})/2.$$

For simplicity, we assume that $\sigma_r < \widetilde{\sigma}_r$. In this case, $\widetilde{\sigma}_i - \sigma \geq 0$, $i = 1, 2, \ldots, r$, i.e., $k_2 = r$. Then

$$\begin{aligned}&\widetilde{D} = I, \qquad \widetilde{E} = 0,\\ &m'_{ii} = m_{ii}, \quad i \leq k_1, \qquad m'_{ii} = 4 - m_{ii}, \quad i > k_1,\\ &\epsilon = \frac{\left(\frac{\widetilde{\sigma}_r - \sigma_r}{2} + \tau\right)}{2},\end{aligned} \tag{3.2}$$

where $\tau = \min\{\sigma_{k_1} - \sigma, \sigma - \sigma_{k_1+1}\}$. By (2.5), (2.7), and (3.1),

$$\|A - \widetilde{A}\|_F^2 \geq \sigma^2 \mathrm{tr}(M) + \sigma \left(\sum_{i=1}^{r} (\sigma_i - \sigma) m_{ii} + \sum_{i=1}^{r} (\widetilde{\sigma}_i - \sigma) \widetilde{m}_{ii} \right) + \epsilon^2 \mathrm{tr}(M') + \epsilon \sum_{i=1}^{r} (\epsilon_i - \epsilon) m'_{ii} + \epsilon \sum_{i=1}^{r} (\widetilde{\epsilon}_i - \epsilon) \widetilde{m}'_{ii}$$

and noting that $m_{ii} \geq 0, m'_{ii} \geq 0$, we have

$$\|A - \widetilde{A}\|_F^2 \geq \sigma^2 \mathrm{tr}(M) - \sigma \sum_{i=k_1+1}^{r} (\sigma - \sigma_i) m_{ii} + \sigma(\widetilde{\sigma}_r - \sigma) \mathrm{tr}(\widetilde{M}) - \epsilon^2 \mathrm{tr}(M') + \epsilon \sum_{i=1}^{r} \epsilon_i m'_{ii} + \epsilon \sum_{i=1}^{r} \widetilde{\epsilon}_i \widetilde{m}'_{ii}.$$

Let $\alpha_0 = \epsilon\widetilde{\epsilon}_r - \epsilon^2 = (\frac{\widetilde{\sigma}_r - \sigma_r}{4})^2 - \left(\frac{\tau}{2}\right)^2 \geq 0$. By (3.2),

$$\begin{aligned} \|A - \widetilde{A}\|_F^2 &\geq \sigma\widetilde{\sigma}_r \mathrm{tr}(M) - \sum_{i=k_1+1}^{n} (\sigma - \sigma_i)(\sigma m_{ii} - \epsilon m'_{ii}) + (\epsilon\epsilon_r - \epsilon^2)\mathrm{tr}(M') \\ &\geq \sigma\widetilde{\sigma}_r \mathrm{tr}(M) + \alpha_0 \mathrm{tr}(M') - \sum_{i=k_1+1}^{r} (\sigma - \sigma_i)((\sigma + \epsilon) m_{ii} - 4\epsilon), \end{aligned}$$

and by Lemma 2.2,

$$\|A - \widetilde{A}\|_F^2 \geq \sigma\widetilde{\sigma}_r \mathrm{tr}(M) + \alpha_0 \left(\mathrm{tr}(M) + 4(r - k_1) - 2 \sum_{i=k_1+1}^{r} m_{ii} \right) - \sum_{i=k_1+1}^{r} (\sigma - \sigma_i)((\sigma + \epsilon) m_{ii} - 4\epsilon).$$

Let $K = \{k_1 + 1 \leq i \leq r | (\sigma + \epsilon) m_{ii} - 4\epsilon > 0\}$, and assume that

$$\max_i m_{ii} \leq \alpha^2, \tag{3.3}$$

where $\alpha > 0$. Then

$$\begin{aligned} \sum_{i=k_1+1}^{n} (\sigma - \sigma_i)((\sigma + \epsilon) m_{ii} - 4\epsilon) &\leq \sum_{i \in K} (\sigma - \sigma_i)((\sigma + \epsilon) m_{ii} - 4\epsilon) \\ &\leq (\sigma - \sigma_r) \sum_{i \in K} ((\sigma + \epsilon) m_{ii} - 4\epsilon) \\ &\leq (\sigma - \sigma_r) \sum_{i \in K} \left(\sigma + \epsilon - \frac{4\epsilon}{\alpha^2} \right) m_{ii} \\ &\leq (\sigma - \sigma_r) \sum_{i=k_1+1}^{r} \left(\sigma + \epsilon - \frac{4\epsilon}{\alpha^2} \right) m_{ii} \end{aligned}$$

since $\sigma+\epsilon-\frac{4\epsilon}{\alpha^2}\geq 0$. Therefore,

$$\begin{aligned}\|A-\widetilde{A}\|_F^2 &\geq (\sigma\widetilde{\sigma}_r+\alpha_0)\mathrm{tr}(M)+\alpha_0\left(4(r-k_1)-2\sum_{i=k_1+1}^{r}m_{ii}\right)\\ &\quad-(\sigma-\sigma_r)\sum_{i=k_1+1}^{r}\left(\sigma+\epsilon-\frac{4\epsilon}{\alpha^2}\right)m_{ii}\\ &\geq (\sigma\widetilde{\sigma}_r+\alpha_0)\mathrm{tr}(M)-\beta\sum_{i=k_1+1}^{r}m_{ii},\end{aligned}$$

where

$$\beta=(\sigma-\sigma_r)\left(\sigma+\epsilon-\frac{4\epsilon}{\alpha^2}\right)+\alpha_0\left(2-\frac{4}{\alpha^2}\right).$$

A straightforward calculation gives

$$\sigma\widetilde{\sigma}_r+\alpha_0-\beta=\left(\frac{\widetilde{\sigma}_r+\sigma_r}{2}\right)^2+3\left(\frac{\widetilde{\sigma}_r-\sigma_r}{2}\right)^2\left(\frac{1}{\alpha^2}-\frac{1}{4}\right)$$

and

$$\sigma\widetilde{\sigma}_r+\alpha_0=\left(\frac{\widetilde{\sigma}_r+\sigma_r}{2}\right)^2+\frac{\widetilde{\sigma}_r^2-\sigma_r^2}{4}+\alpha_0.$$

Since $\sum_{i=k_1+1}^{r}m_{ii}\leq\mathrm{tr}(M)$, we have

$$\|A-\widetilde{A}\|_F^2\geq\left(\left(\frac{\widetilde{\sigma}_r+\sigma_r}{2}\right)^2+\beta_0\frac{\widetilde{\sigma}_r-\sigma_r}{4}\right)\mathrm{tr}(M), \tag{3.4}$$

where

$$\beta_0=\min\left\{3(\widetilde{\sigma}_r-\sigma_r)\left(\frac{1}{\alpha^2}-\frac{1}{4}\right),\ \widetilde{\sigma}_r+\sigma_r+\alpha_0/(\widetilde{\sigma}_r-\sigma_r)\right\}. \tag{3.5}$$

A new perturbation bound is given in the following theorem.

THEOREM 3.1. *Let A and $\widetilde{A}\in\mathcal{C}_n^{n\times n}$ have the SVDs in* (2.1). *If*

$$\|A-\widetilde{A}\|_2\leq\alpha(\widetilde{\sigma}_n+\sigma_n)/2, \tag{3.6}$$

then

$$\left(\left(\frac{\widetilde{\sigma}_n+\sigma_n}{2}\right)^2+\beta_0\frac{\widetilde{\sigma}_n-\sigma_n}{4}\right)\|Q-\widetilde{Q}\|_F^2\leq\|A-\widetilde{A}\|_F^2. \tag{3.7}$$

In particular, when $\alpha\leq\sqrt{12/7}$,

$$\|Q-\widetilde{Q}\|_F^2\leq\frac{2}{\sigma_n^2+\widetilde{\sigma}_n^2}\|A-\widetilde{A}\|_F^2. \tag{3.8}$$

Proof. When $r = n = m$, we obtain (3.7) using the fact [9, 10] that

$$m_{ii} \le \|Q - \widetilde{Q}\|_2^2 \le \left(\frac{2}{\sigma_n + \widetilde{\sigma}_n}\right)^2 \|A - \widetilde{A}\|_2^2.$$

When $\alpha \le \sqrt{12/7}$, $\beta_0 \ge \widetilde{\sigma}_n - \sigma_n$ and (3.8) follows immediately. ∎

Remark. It is obvious that the perturbation bounds in Theorem 3.1 are always better than the previous bound in (1.5) under the condition of the perturbation being small, which results in

$$\|Q - \widetilde{Q}\|_2 \le \alpha. \tag{3.9}$$

Since both Q and $\widetilde{Q}$ are unitary,

$$\|Q - \widetilde{Q}\|_2 \le 2$$

is always true. Condition (3.9) becomes an active constraint if $\alpha < 2$. For any $\alpha < 2$, $\beta_0 > 0$ and the bound in (3.7) improves the bound in (1.5). In fact, many previous results were obtained under such conditions. For complex matrices with $m = n = r$, Mathias [11] proved (1.4) under the condition $\|A - \widetilde{A}\|_2 < \sigma_n$. For real matrices, Mathias presented two bounds (Theorems 2.3 and 2.4 of [11]) as follows: (i) If $\|A - \widetilde{A}\|_2 < \sigma_n$,

$$\|Q - \widetilde{Q}\| \le -\frac{2\|A - \widetilde{A}\|}{\sigma_1(A - \widetilde{A}) + \sigma_2(A - \widetilde{A})} \log\left(1 - \frac{\sigma_1(A - \widetilde{A}) + \sigma_2(A - \widetilde{A})}{\sigma_n + \sigma_{n-1}}\right) \tag{3.10}$$

and (ii) if $A + tE$ is nonsingular for all $t \in [0, 1]$,

$$\|Q - \widetilde{Q}\| \le \max_{0 \le t \le 1} \left\{\frac{2}{\sigma_n(A + tE) + \sigma_{n-1}(A + tE)}\right\} \|A - \widetilde{A}\|. \tag{3.11}$$

In our recent work [10], we presented the bound

$$\|Q - \widetilde{Q}\|_F \le \frac{4}{\sigma_n + \widetilde{\sigma}_n + \sigma_{n-1} + \widetilde{\sigma}_{n-1}} \|A - \widetilde{A}\|_F \tag{3.12}$$

under the condition $\|A - \widetilde{A}\|_2 < \widetilde{\sigma}_n + \sigma_n$. The bound in (3.12) is sharper than both (3.10) and (3.11).

Two examples are given below. The first shows the sharpness of our bound and the second shows that the condition of the perturbation being small is necessary to get (3.8).

Example 3.1. Let

$$A = U\Sigma V^* \quad \text{and} \quad \widetilde{A} = U\widetilde{\Sigma}(V(I - D))^*, \tag{3.13}$$

i.e., $\widetilde{V} = V(I - D)$, where

$$\Sigma = \sigma_2 I, \qquad \widetilde{\Sigma} = \sigma_2 \begin{pmatrix} 1 + \delta & 0 \\ 0 & 1 - \delta \end{pmatrix}, \qquad D = \begin{pmatrix} a & -\sqrt{2a - a^2} \\ \sqrt{2a - a^2} & a \end{pmatrix},$$

and $a = \frac{\delta^2}{2(1-\delta)^2}$. A straightforward calculation gives $\|Q - \widetilde{Q}\|_F^2 = 4a$ and $\|A - \widetilde{A}\|_F^2 = (2\delta^2 + 4a)\sigma_2^2$. Condition (3.6) is always satisfied for small δ. Then

$$4a = \|Q - \widetilde{Q}\|_F^2 = \frac{2}{\sigma_2^2 + (1 - \delta)^2\sigma_2^2} \|A - \widetilde{A}\|_F^2 < \frac{4}{(\sigma_2 + (1 - \delta)\sigma_2)^2} \|A - \widetilde{A}\|_F^2.$$

Example 3.2. Let A and $\widetilde{A}$ be defined in (3.13) with

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \sigma_n \end{pmatrix},\ \widetilde{\Sigma} = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 2\sigma_n, \end{pmatrix},\ \text{and } D = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

with $\Sigma_1 \geq 2\sigma_n I$. Clearly, $\|A-\widetilde{A}\|_F = \|A-\widetilde{A}\|_2 = 3\sigma_n$ and $\|Q-\widetilde{Q}\|_F^2 = 4$. Then $\|A-\widetilde{A}\|_2 = 3\sigma_n > \sqrt{\frac{12}{7}}\frac{\sigma_n+\widetilde{\sigma}_n}{2} = \sqrt{\frac{12}{7}}\frac{3\sigma_n}{2}$; i.e., condition (3.6) with $\alpha = \sqrt{\frac{12}{7}}$ is not satisfied. In this case,

$$\left(\frac{2}{\sigma_n+\widetilde{\sigma}_n}\right)^2 \|A-\widetilde{A}\|^2 = 4 = \|Q-\widetilde{Q}\|_F^2 > \frac{2}{\sigma_n^2+(2\sigma_n)^2}\|A-\widetilde{A}\|_F^2 = \frac{18}{5};$$

i.e., (3.8) does not hold while (1.5) holds.

For $r = n \leq m$, it was proved in [9] that

$$\|Q-\widetilde{Q}\|_2 \leq \left(\frac{2}{\sigma_n+\widetilde{\sigma}_n} + \frac{1}{\max\{\sigma_n,\widetilde{\sigma}_n\}}\right)\|A-\widetilde{A}\|_2.$$

Our next theorem provides additional information.

THEOREM 3.2. *Let A and $\widetilde{A} \in \mathcal{C}_n^{m\times n}$ have the SVDs in* (2.1). *If*

$$\left(\frac{2}{\sigma_n+\widetilde{\sigma}_n} + \frac{1}{\max\{\sigma_n,\widetilde{\sigma}_n\}}\right)\|A-\widetilde{A}\|_2 \leq \alpha, \tag{3.14}$$

then (3.7) *holds. In particular, when $\alpha \leq \sqrt{12/7}$,* (3.8) *holds.*

For the general problems in $\mathcal{C}_r^{m\times n}$, no perturbation bound in the spectral norm has been published until now. It is easy to give some improved perturbation bounds with the restriction on $\|A-\widetilde{A}\|_F$ instead of $\|A-\widetilde{A}\|_2$. However, when the dimension of matrices is very large, such a restriction seems less efficient.

Since

$$\widetilde{U}^*(A-\widetilde{A})V = S\Sigma - \widetilde{\Sigma}T = \begin{bmatrix} S_{11}\Sigma_1 - \widetilde{\Sigma}_1 T_{11} & -\widetilde{\Sigma}_1 T_{12} \\ S_{21}\Sigma_1 & 0 \end{bmatrix}$$

and

$$\widetilde{V}^*(\widetilde{A}^*-A^*)U = \widetilde{\Sigma}^t S - T\Sigma^t = \begin{bmatrix} \widetilde{\Sigma}_1 S_{11} - T_{11}\Sigma_1 & \widetilde{\Sigma}_1 S_{12} \\ -T_{21}\Sigma_1 & 0 \end{bmatrix},$$

by some basic properties of unitarily invariant norms [7, 12], we get

$$\begin{aligned}\|A-\widetilde{A}\| \geq \max\{&\|S_{11}\Sigma_1 - \widetilde{\Sigma}_1 T_{11}\|, \|\widetilde{\Sigma}_1 S_{11} - T_{11}\Sigma_1\|, \\ &\|\widetilde{\Sigma}_1 T_{12}\|, \|S_{21}\Sigma_1\|, \|S_{12}\widetilde{\Sigma}_1\|, \|T_{21}\Sigma_1\|\}.\end{aligned} \tag{3.15}$$

Similarly we have

$$\|Q-\widetilde{Q}\| = \left\|\begin{bmatrix} S_{11}-T_{11} & -T_{12} \\ S_{21} & 0 \end{bmatrix}\right\| = \left\|\begin{bmatrix} S_{11}-T_{11} & S_{12} \\ -T_{21} & 0 \end{bmatrix}\right\|,$$

which leads to

$$\|Q-\widetilde{Q}\| \leq \|S_{11}-T_{11}\| + \|S_{12}\| + \|T_{21}\|. \tag{3.16}$$

For the spectral norm, we have

$$(Q-\widetilde{Q})(Q-\widetilde{Q})^* = \begin{pmatrix} S_{11}-T_{11} \\ -T_{21} \end{pmatrix}\begin{pmatrix} S_{11}-T_{11} \\ -T_{21} \end{pmatrix}^* + \begin{pmatrix} S_{12} \\ 0 \end{pmatrix}\begin{pmatrix} S_{12} \\ 0 \end{pmatrix}^*,$$

and therefore,

$$\|Q-\widetilde{Q}\|_2^2 \le \left\|\begin{pmatrix} S_{11}-T_{11} \\ -T_{21} \end{pmatrix}\right\|_2^2 + \left\|\begin{pmatrix} S_{12} \\ 0 \end{pmatrix}\right\|_2^2.$$

Hence

$$\|Q-\widetilde{Q}\|_2^2 \le \|S_{11}-T_{11}\|_2^2 + \|S_{12}\|_2^2 + \|T_{21}\|_2^2. \tag{3.17}$$

It has been proved [12, p. 260] that

$$\|S_{12}\| = \|S_{21}\| \tag{3.18}$$

and, similarly, that

$$\|T_{12}\| = \|T_{21}\|. \tag{3.19}$$

The following lemma given in [9] is a special case of Davis and Kahan [6, Theorem 5.2].

LEMMA 3.3. *Let B_1 and B_2 be two Hermitian matrices and let P be a complex matrix. Suppose there are two disjoint intervals separated by a gap of width at least η, where one interval contains the spectrum of B_1 and the other contains that of B_2. If $\eta > 0$, then there exists a unique solution X to the matrix equation $B_1X - XB_2 = P$ and, moreover, $\|X\| \le \frac{1}{\eta}\|P\|$.*

Let $X = S_{11} - T_{11}$, $B_1 = \Sigma_1$, and $B_2 = -\widetilde{\Sigma}_1$. By Lemma 3.3,

$$\|S_{11}-T_{11}\| \le \frac{1}{\sigma_r+\widetilde{\sigma}_r}(\|S_{11}\Sigma_1 - \widetilde{\Sigma}_1 T_{11}\| + \|\widetilde{\Sigma}_1 S_{11} - T_{11}\Sigma_1\|). \tag{3.20}$$

Combining (3.15)–(3.20), we have Theorem 3.4.

THEOREM 3.4. *Let A and $\widetilde{A} \in \mathcal{C}_r^{m\times n}$ have SVDs as in* (2.1). *Then*

$$\|Q-\widetilde{Q}\| \le \left(\frac{2}{\sigma_r+\widetilde{\sigma}_r} + \frac{2}{\max\{\sigma_r,\widetilde{\sigma}_r\}}\right)\|A-\widetilde{A}\| \tag{3.21}$$

and

$$\|Q-\widetilde{Q}\|_2 \le \sqrt{\left(\frac{2}{\sigma_r+\widetilde{\sigma}_r}\right)^2 + \frac{2}{\max\{\sigma_r^2,\widetilde{\sigma}_r^2\}}}\,\|A-\widetilde{A}\|_2. \tag{3.22}$$

Noting that S and T are unitary, we obtain

$$\begin{aligned} M &= 2I^{(r)} - S_{11}^*T_{11} - T_{11}^*S_{11} = (S_{11}-T_{11})^*(S_{11}-T_{11}) + S_{21}^*S_{21} + T_{21}^*T_{21}, \\ \widetilde{M} &= 2I - S_{11}T_{11}^* - T_{11}^*S_{11}^* = (S_{11}-T_{11})(S_{11}-T_{11})^* + S_{12}S_{12}^* + T_{12}T_{12}^*. \end{aligned}$$

Hence

$$\begin{aligned} m_{ii} &\le \|S_{11}-T_{11}\|_2^2 + \|S_{21}\|_2^2 + \|T_{21}\|_2^2, \\ \widetilde{m}_{ii} &\le \|S_{11}-T_{11}\|_2^2 + \|S_{12}\|_2^2 + \|T_{12}\|_2^2, \end{aligned} \tag{3.23}$$

and by (3.15), we obtain

$$\max\{m_{ii}, \widetilde{m}_{ii}\} \le \left(\left(\frac{2}{\sigma_r + \widetilde{\sigma}_r} \right)^2 + \frac{2}{\max\{\sigma_r^2, \widetilde{\sigma}_r^2\}} \right) \|A - \widetilde{A}\|_2^2. \tag{3.24}$$

THEOREM 3.5. *Let A and $\widetilde{A} \in \mathcal{C}_r^{m\times n}$ have SVDs as in* (2.1)*. If*

$$\sqrt{\left(\frac{2}{\sigma_r + \widetilde{\sigma}_r} \right)^2 + \frac{2}{\max\{\sigma_r^2, \widetilde{\sigma}_r^2\}}}\ \|A - \widetilde{A}\|_2 \le \alpha, \tag{3.25}$$

then (3.7) *holds. In particular, when $\alpha \le \sqrt{12/7}$,* (3.8) *holds.*

Acknowledgments. The authors would like to thank the editor, Professor R. Bhatia, and an anonymous referee for their valuable comments, which improved the presentation.

REFERENCES

[1] A. BARRLUND, *Perturbation bounds on the polar decomposition*, BIT, 30 (1989), pp. 101–113.
[2] R. BHATIA, *Matrix Analysis*, Springer, New York, 1997.
[3] R. BHATIA AND K. MUKHERJEA, *Variation of the unitary part of a matrix*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1007–1014.
[4] R. BHATIA, *Matrix factorizations and their perturbations*, Linear Algebra Appl., 197/198 (1994), pp. 245–276.
[5] F. CHATELIN AND S. GRATTON, *On the condition numbers associated with the polar factorization of a matrix*, Numer. Linear Algebra Appl., 7 (2000), pp. 337–354.
[6] C. DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a perturbation*. III, SIAM J. Numer. Anal., 1 (1970), pp. 1–46.
[7] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[8] R. C. LI, *A perturbation bound for the generalized polar decomposition*, BIT, 33 (1993), pp. 304–308.
[9] R.-C. LI, *New perturbation bounds for the unitary polar factor*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 327–332.
[10] W. LI AND W. SUN, *Perturbation bounds of unitary and subunitary polar factors*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1183–1193.
[11] R. MATHIAS, *Perturbation bounds for the polar decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 588–597.
[12] J. G. SUN, *Matrix Perturbation Analysis*, 2nd ed., Science Press, Beijing, 2001.
[13] J. G. SUN AND C. H. CHEN, *Generalized polar decomposition*, Math. Numer. Sinica, 11 (1989), pp. 262–273.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 373–384

FAST AND STABLE ALGORITHMS FOR BANDED PLUS SEMISEPARABLE SYSTEMS OF LINEAR EQUATIONS*

S. CHANDRASEKARAN† AND M. GU‡

Abstract. We present fast and numerically stable algorithms for the solution of linear systems of equations, where the coefficient matrix can be written in the form of a banded plus semiseparable matrix. Such matrices include banded matrices, banded bordered matrices, semiseparable matrices, and block-diagonal plus semiseparable matrices as special cases. Our algorithms are based on novel matrix factorizations developed specifically for matrices with such structures. We also present interesting numerical results with these algorithms.

Key words. banded matrix, bordered matrix, semiseparable matrix, $\mathcal{H}$-matrix, fast algorithms, stable algorithms

AMS subject classifications. 15A09, 15A23, 65F05, 65L10, 65R20

DOI. 10.1137/S0895479899353373

1. Introduction.

In this paper we consider fast and numerically stable solutions of the $n \times n$ linear system of equations

$$A\,x = b, \tag{1.1}$$

where A is the sum of a banded matrix and a semiseparable matrix (see (1.2) below for a definition).

This class of matrices includes banded bordered matrices, which have been discussed in van Huffel and Park [12] and Govaerts [6]. It also includes block-diagonal plus semiseparable matrices, which appear in the numerical solution of boundary-value problems for ordinary differential equations (ODEs) and certain integral equations (see Greengard and Rokhlin [7], Starr [14], and Lee and Greengard [8]). The coefficient matrices generated from domain decomposition methods for partial differential equations (PDEs) tend to have block-diagonal plus bordered structure. Some related work on (1.1) can also be found in Eidelman and Gohberg [3, 4].

1.1. Contributions.

The most important feature of problem (1.1) is that A is a generally dense but highly structured matrix. When A is symmetric positive definite, such a structure can be fully exploited in computing the Cholesky factorization of A. However, the picture changes completely when A is not symmetric positive definite. Although direct methods have been developed for efficient and numerically stable LU and QR factorizations of banded matrices (see Demmel [2, Chap. 2]), such methods do not currently exist for semiseparable matrices, let alone banded plus semiseparable matrices. The main difficulty is that LU and QR factorizations have tremendous difficulties in maintaining both numerical stability and banded plus semiseparable

*Received by the editors March 8, 1999; accepted for publication (in revised form) by L. Reichel July 31, 2002; published electronically August 19, 2003.
http://www.siam.org/journals/simax/25-2/35337.html

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 (shiv@ece.ucsb.edu).

‡Department of Mathematics, University of California, Berkeley, CA 94720-3840 (mgu@math.berkeley.edu). The research of this author was supported in part by NSF Career Award CCR-9702866 and by Alfred Sloan Research Fellowship BR-3720.

structure and, consequently, require $O(n^3)$ flops[1] to stably compute such factorizations of A in (1.1).

In this paper, we present a number of fast and numerically stable direct methods for solving (1.1). For banded bordered linear systems of equations, these methods work equally well and are more stable and efficient than those of [12, 6]. Our methods are based on some new matrix factorizations we developed specifically for banded plus semiseparable matrices. We also present interesting results from our numerical experiments with these methods in MATLAB.

1.2. Notation. To describe the problem precisely, we first introduce some notation. Reminiscent of MATLAB notation, we use $\mathbf{triu}\,(\mathbf{A}, \mathbf{k})$ to denote the matrix which is identical to the matrix A on and *above* the kth diagonal. $k = 0$ is the main diagonal, $k > 0$ is above the main diagonal, and $k < 0$ is below the main diagonal. Similarly, $\mathbf{tril}\,(\mathbf{A}, \mathbf{k})$ denotes the matrix which is identical to the matrix A on and *below* the kth diagonal. For example,

$$\mathbf{triu}\left(\begin{pmatrix} \alpha & \beta & \gamma \\ \delta & \zeta & \eta \\ \theta & \lambda & \mu \end{pmatrix}, \mathbf{1}\right) = \begin{pmatrix} 0 & \beta & \gamma \\ 0 & 0 & \eta \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{tril}\left(\begin{pmatrix} \alpha & \beta & \gamma \\ \delta & \zeta & \eta \\ \theta & \lambda & \mu \end{pmatrix}, -\mathbf{1}\right) = \begin{pmatrix} 0 & 0 & 0 \\ \delta & 0 & 0 \\ \theta & \lambda & 0 \end{pmatrix}.$$

As a banded plus semiseparable matrix, the matrix A in (1.1) can be written as

$$A = D + \mathbf{triu}\left(\mathbf{u}\,\mathbf{v}^{\mathbf{T}}, \mathbf{b_u} + \mathbf{1}\right) + \mathbf{tril}\left(\mathbf{p}\,\mathbf{q}^{\mathbf{T}}, -\mathbf{b_l} - \mathbf{1}\right), \tag{1.2}$$

where D is an $n \times n$ banded matrix, with b_u nonzero diagonals strictly above the main diagonal and b_l nonzero diagonals strictly below the main diagonal; u and v are $n \times r_u$ matrices and p and q are $n \times r_l$ matrices.

When $b_u = b_l = 0$, D is a diagonal matrix, and A is a diagonal plus semiseparable matrix. When $r_u = r_l = 0$, $A = D$ is a banded matrix. We are interested in the numerical solution of the linear system (1.1). The rest of this paper provides a set of numerically backward stable algorithms which take $O\left(n(b_u + b_l + r_u + r_l)^2\right)$ flops to solve (1.1) as opposed to $O(n^3)$ by using traditional methods involving LU and QR factorizations. The exact constant hidden in the $O(\cdot)$ notation varies among our algorithms.

Throughout this paper, we will take the liberty of using I to denote an identity matrix of any dimension.

The rest of this paper is organized as follows. In section 2 we illustrate the basic ideas behind our algorithms through a simple example. In section 3 we describe the algorithms in some detail. In section 4 we present our numerical results with these algorithms.

2. Basic idea. In this section we give a description of the basic idea in the simple case when D is a diagonal matrix ($b_u = b_l = 0$), and u, v, p, and q have only one column ($r_u = r_l = 1$).

The idea is to compute a two-sided decomposition of the form

$$A = W\,L\,H, \tag{2.1}$$

where W and H can be written as the product of elementary matrices, and L is a lower triangular matrix. The three matrices W, L, H themselves are never explicitly

[1] A *flop* is a floating point operation such as $+, -, \times$, or $\div$.

formed but inverted efficiently online as the algorithm proceeds. In this section, we will choose the matrices W and H to be the products of elementary Givens rotations. When we discuss our algorithms in full detail in section 3, we will allow ourselves the additional freedom of choosing W and H to be products of elementary Householder reflections or Gaussian elimination matrices with column and/or row permutations.

More specifically, consider the 5×5 case,

$$A = \begin{pmatrix} D_0 & u_0v_1 & u_0v_2 & u_0v_3 & u_0v_4 \\ p_1q_0 & D_1 & u_1v_2 & u_1v_3 & u_1v_4 \\ p_2q_0 & p_2q_1 & D_2 & u_2v_3 & u_2v_4 \\ p_3q_0 & p_3q_1 & p_3q_2 & D_3 & u_3v_4 \\ p_4q_0 & p_4q_1 & p_4q_2 & p_4q_3 & D_4 \end{pmatrix}.$$

For future convenience we also assume that the right-hand side is of the form

$$\bar{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} - \tau_{-1} \begin{pmatrix} 0 \\ 0 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix},$$

where $\tau_{-1} = 0$. (Of course, the second term on the right-hand side has no effect at this stage, but it will capture the general form of the recursion as we proceed.)

Now suppose that W_0 is a Givens rotation such that

$$W_0 \begin{pmatrix} u_0 \\ u_1 \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ \sqrt{u_0^2 + u_1^2} \\ w \end{pmatrix} \equiv \begin{pmatrix} 0 \\ \widehat{u}_1 \\ w \end{pmatrix} \tag{2.2}$$

for any vector w. Then if we apply W_0 from the left to A, we obtain

$$\widehat{A} = W_0\, A = \begin{pmatrix} \widehat{A}_{00} & \widehat{A}_{01} & 0 & 0 & 0 \\ \widehat{A}_{10} & \widehat{A}_{11} & \widehat{u}_1v_2 & \widehat{u}_1v_3 & \widehat{u}_1v_4 \\ p_2q_0 & p_2q_1 & D_2 & u_2v_3 & u_2v_4 \\ p_3q_0 & p_3q_1 & p_3q_2 & D_3 & u_3v_4 \\ p_4q_0 & p_4q_1 & p_4q_2 & p_4q_3 & D_4 \end{pmatrix}.$$

We also apply W_0 to $\bar{b}$ to obtain

$$\widehat{b} = W_0\, \bar{b} = W_0 \left(\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} - \tau_{-1} \begin{pmatrix} 0 \\ 0 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} \right) = \begin{pmatrix} \widehat{b}_0 \\ \widehat{b}_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} - \tau_{-1} \begin{pmatrix} 0 \\ 0 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix},$$

where we have deliberately written the formula in such a way that it would be correct even if τ_{-1} had not been zero.

We next choose a Givens rotation, H_0, such that

$$H_0^T \begin{pmatrix} \widehat{A}_{00} \\ \widehat{A}_{01} \\ w \end{pmatrix} = \begin{pmatrix} \sqrt{\widehat{A}_{00}^2 + \widehat{A}_{01}^2} \\ 0 \\ w \end{pmatrix} \equiv \begin{pmatrix} \widetilde{A}_{00} \\ 0 \\ w \end{pmatrix} \tag{2.3}$$

for any vector w. Further let

$$H_0^T \begin{pmatrix} q_0 \\ q_1 \end{pmatrix} = \begin{pmatrix} \widetilde{q}_0 \\ \widetilde{q}_1 \end{pmatrix} \quad \text{and} \quad H_0^T \begin{pmatrix} \widehat{A}_{10} \\ \widehat{A}_{11} \end{pmatrix} = \begin{pmatrix} \widetilde{A}_{10} \\ \widetilde{A}_{11} \end{pmatrix}.$$

Then

$$\widetilde{A} = W_0\, A\, H_0 = \widehat{A}\, H_0 = \begin{pmatrix} \widetilde{A}_{00} & 0 & 0 & 0 & 0 \\ \widetilde{A}_{10} & \widetilde{A}_{11} & \widehat{u}_1 v_2 & \widehat{u}_1 v_3 & \widehat{u}_1 v_4 \\ p_2 \widetilde{q}_0 & p_2 \widetilde{q}_1 & D_2 & u_2 v_3 & u_2 v_4 \\ p_3 \widetilde{q}_0 & p_3 \widetilde{q}_1 & p_3 q_2 & D_3 & u_3 v_4 \\ p_4 \widetilde{q}_0 & p_4 \widetilde{q}_1 & p_4 q_2 & p_4 q_3 & D_4 \end{pmatrix}.$$

Now let

$$H_0^{-1}\, x = H_0^{-1} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \equiv \begin{pmatrix} \widetilde{\chi}_0 \\ \widetilde{x}_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \equiv \widetilde{x}. \tag{2.4}$$

Then it follows from $W_0\, A\, H_0\, H_0^{-1}\, x = \widetilde{A}\, \widetilde{x} = W_0\, b = \widehat{b}$ that

$$\widetilde{\chi}_0 = \frac{\widehat{b}_0}{\widetilde{A}_{00}}.$$

Also let

$$\tau_0 = \tau_{-1} + \widetilde{\chi}_0\, \widetilde{q}_0, \quad \widetilde{b}_1 = \widehat{b}_1 - \widetilde{\chi}_0\, \widetilde{A}_{10}, \quad \text{and} \quad \widetilde{b}_2 = b_2 - \tau_0\, p_2. \tag{2.5}$$

To reach this stage we needed to compute all the "tilde" and "hat" quantities except $\widetilde{x}_1$. They can be computed in *constant* time, independent of the size of the matrix A.

Now we can proceed to solve the smaller 4×4 system of equations,

$$\begin{pmatrix} \widetilde{A}_{11} & \widehat{u}_1 v_2 & \widehat{u}_1 v_3 & \widehat{u}_1 v_4 \\ p_2 \widetilde{q}_1 & D_2 & u_2 v_3 & u_2 v_4 \\ p_3 \widetilde{q}_1 & p_3 q_2 & D_3 & u_3 v_4 \\ p_4 \widetilde{q}_1 & p_4 q_2 & p_4 q_3 & D_4 \end{pmatrix} \begin{pmatrix} \widetilde{x}_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \widetilde{b}_1 \\ \widetilde{b}_2 \\ b_3 \\ b_4 \end{pmatrix} - \tau_0 \begin{pmatrix} 0 \\ 0 \\ p_3 \\ p_4 \end{pmatrix},$$

which is exactly like the original 5×5 system of equations in form. That is, the coefficient matrix is a diagonal matrix plus a semiseparable matrix, and the right-hand side is also of the requisite form. Hence we can use this recursion ((2.2) through (2.5)) three times until the problem size becomes two, at which point we solve the system directly. Let the five numbers obtained by this recursion, $\widetilde{\chi}_0$, $\widetilde{\chi}_1$, $\widetilde{\chi}_2$, $\widetilde{\chi}_3$, and $\widetilde{\chi}_4$, be the components of the five-dimensional vector χ. Then it follows from (2.4) that the actual solution x to the original 5×5 system of equations is given by

$$x = H_0\, H_1\, H_2\, \chi, \tag{2.6}$$

where the H_i's are the successive Givens transforms computed from the recursion (2.3) but set up in such a way that they only affect rows i and $i+1$. Since there are only three of these transforms, we retain the linear time complexity of the algorithm.

The backward stability of the algorithm follows from the fact that we use only orthogonal transforms and a single forward substitution.

Our factorization is similar in form to the ULV factorization proposed by Stewart [15]. However, the ULV factorization of Stewart is developed primarily to reveal potential numerical rank-deficiency in a general matrix and can take $O(n^3)$ flops to compute, whereas our factorization is designed primarily to take advantage of the banded plus semiseparable structure for large savings in computational cost without sacrificing numerical stability.

There are two places in the recursion where elimination is necessary. In (2.2) we chose W_0 to be a Givens rotation to eliminate u_0, and in (2.3) we chose P_0 to be another Givens rotation to eliminate $\widehat{A}_{01}$. These transformations can be replaced by Householder transformations or Gaussian elimination matrices with row or column pivoting for general banded plus semiseparable matrices. This results in several algorithms with different efficiency and numerical stability properties. In the next section, we describe a general procedure for solving (1.1) via the computation of the factorization (2.1). We also discuss efficiency and numerical stability issues for different choices of W and H in (2.1).

3. The algorithms. We now describe fast algorithms for solving (1.1), where A is a general banded plus semiseparable matrix of the form (1.2).

3.1. Preprocessing and basic linear algebra procedures. Some preprocessing is needed before the algorithms formally start. We make u lower triangular by computing a QR factorization $u^T = Q\,R$ and resetting

$$u := R^T \quad \text{and} \quad v := v\,Q. \tag{3.1}$$

This operation takes roughly $6nr_u^2$ flops using the fact that Q is computed in factored form [5, Chap. 5].

We also review a few well-known basic linear algebra routines needed in our algorithms. Let L be an $m \times s$ lower triangular matrix with $m > s$,

$$L = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{s1} & l_{s2} & \cdots & l_{ss} \\ \vdots & \vdots & & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{ms} \end{pmatrix}.$$

Algorithm 3.1 below is a standard procedure for efficiently zeroing out entries $l_{11}, l_{22}, \ldots, l_{ss}$ on the main diagonal of L by using s Givens rotations (see [5, Chap. 12]).

ALGORITHM 3.1. *Elimination with Givens rotations.*

for $i := s$ **to** 1 **step** -1 **do**

- Choose $c_i^2 + s_i^2 = 1$ such that

$$\begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix} \begin{pmatrix} l_{i,i} \\ l_{i+1,i} \end{pmatrix} = \begin{pmatrix} 0 \\ \rho_i \end{pmatrix}, \quad \text{where} \quad \rho_i = \sqrt{l_{i,i}^2 + l_{i+1,i}^2}.$$

- Set $l_{i,i} := 0$, $l_{i+1,i} := \rho_i$, and compute

$$\begin{pmatrix} l_{i,1} & \cdots & l_{i,i-1} \\ l_{i+1,1} & \cdots & l_{i+1,i-1} \end{pmatrix} := \begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix} \begin{pmatrix} l_{i,1} & \cdots & l_{i,i-1} \\ l_{i+1,1} & \cdots & l_{i+1,i-1} \end{pmatrix}.$$

endfor

Let W be the product of all the Givens rotations used in the above algorithm. Then its output can be written via a matrix-matrix product as $L := W\,L$.

Similarly, we can zero out the main diagonal of L by using a banded Gaussian elimination procedure with row pivoting. See Golub and Van Loan [5, Chap. 4] for details.

Let $G \in \mathbf{R}^{m\times s}$ be a general dense matrix. Then we can choose a Householder transformation

$$H = I - 2u\,u^T \quad \text{with} \quad \|u\|_2 = 1$$

to zero out all the entries in the first row of G except the $(1,1)$ entry as follows:

$$G\,H = \begin{pmatrix} \widehat{\gamma} & 0 \\ \widehat{g} & \widehat{G} \end{pmatrix}. \tag{3.2}$$

The cost for computing u is $O(s)$, and the cost for computing $G\,H$ is about $4ms$ flops (see [5, Chap. 5]).

Alternatively, we can choose H in (3.2) to be a Gaussian elimination matrix of the form

$$H = \begin{pmatrix} 1 & -h^T \\ 0 & I \end{pmatrix}.$$

Column pivoting can be used to enhance numerical stability. The cost for computing $G\,H$ is about $2ms$ flops (see Golub and Van Loan [5, Chap. 3]).

3.2. New algorithms. Let $\ell = b_u + r_u + 1$ and $m = \ell + b_l$; we begin by writing A in the form

$$A = \begin{pmatrix} G & E \\ C & F \end{pmatrix}, \tag{3.3}$$

where $G \in \mathbf{R}^{m\times\ell}$ is a dense matrix (its banded plus semiseparable structure will be ignored); $F \in \mathbf{R}^{(n-r_u-1)\times(n-\ell)}$ is a banded plus semiseparable rectangular matrix; and both $C \in \mathbf{R}^{(n-m)\times\ell}$ and $E \in \mathbf{R}^{(r_u+1)\times(n-\ell)}$ are low rank matrices. We caution that, strictly speaking, (3.3) is *not* a block partitioning of A, since the row dimension of G is larger than that of E in general.

In further detail, we write $C = \bar{p}\,\bar{q}^T$, where $\bar{p} \in \mathbf{R}^{(n-m)\times r_l}$ and $\bar{q} \in \mathbf{R}^{\ell\times r_l}$ contain the last $n-m$ rows of p and the first ℓ rows of q, respectively. Similarly, $E = \bar{u}\,\bar{v}^T$, where $\bar{u} \in \mathbf{R}^{(r_u+1)\times r_u}$ and $\bar{v} \in \mathbf{R}^{(n-\ell)\times r_u}$ contain the first r_u+1 rows of u and the last $n-\ell$ rows of v, respectively. As suggested in section 3.1, we assume that $\bar{u}$ is a lower triangular matrix.

As in section 2, we will solve (1.1) by recursively solving the linear system

$$A\,x = \bar{b} \equiv b - \begin{pmatrix} 0 \\ \bar{p}\,\tau \end{pmatrix}, \tag{3.4}$$

where $\tau_{-1} = 0 \in \mathbf{R}^{r_l}$ is an auxiliary vector that will play the role of scalar τ_{-1} in the example in section 2. As before, we will compute a two-sided decomposition (2.1) of A and invert matrices W, L, and H on the fly.

To start the recursion, we use Algorithm 3.1 to compute a matrix W_0 so that $W_0\,\bar{u}$ is a lower triangular matrix with zeros on its main diagonal. Compute

$$\begin{pmatrix} 0 \\ \widehat{u} \end{pmatrix} := W_0\,\bar{u}, \quad \widehat{G} := \begin{pmatrix} W_0 & 0 \\ 0 & I \end{pmatrix} G, \quad \text{and} \quad \widehat{b} := \begin{pmatrix} W_0 & 0 \\ 0 & I \end{pmatrix} b - \begin{pmatrix} 0 \\ \bar{p}\,\tau_{-1} \end{pmatrix}. \tag{3.5}$$

Linear system (3.4) now becomes

$$\begin{pmatrix} \widehat{G} & \begin{pmatrix} 0 \\ \widehat{u} \end{pmatrix} \bar{v}^T \\ \bar{p}\,\bar{q}^T & F \end{pmatrix} x = \widehat{b}.$$

We further choose H_0 to zero out the first row of $\widehat{G}$ except the $(1,1)$ entry. Compute

$$\begin{pmatrix} \widetilde{\gamma} & 0 \\ \widetilde{g} & \widetilde{G} \end{pmatrix} := \widehat{G}\,H_0 \quad \text{and} \quad \begin{pmatrix} \widetilde{\rho}^T \\ \widetilde{q} \end{pmatrix} := H_0^T\,\bar{q}. \tag{3.6}$$

Linear system (3.4) now has the following form:

$$\begin{pmatrix} \widetilde{\gamma} & 0 & 0 \\ \widetilde{g} & \widetilde{G} & \widehat{u}\,\bar{v}^T \\ \bar{p}\,\widetilde{\rho} & \bar{p}\,\widetilde{q}^T & F \end{pmatrix} \widetilde{x} = \widehat{b}, \tag{3.7}$$

where

$$\widetilde{x} = \begin{pmatrix} H_0^{-1} & 0 \\ 0 & I \end{pmatrix}, \quad x = \begin{pmatrix} \widetilde{\chi}_0 \\ \widetilde{x}_1 \\ \widetilde{x}_2 \end{pmatrix}, \quad \text{and} \quad \widehat{b} = \begin{pmatrix} \widehat{\beta} \\ \widehat{b}_1 \\ \widehat{b}_2 - \bar{p}\,\tau_{-1} \end{pmatrix}.$$

Now we can perform one step of forward substitution in (3.7) to get $\widetilde{\chi}_0 = \widehat{\beta}/\widetilde{\gamma}$ and

$$\begin{pmatrix} \widetilde{G} & \widehat{u}\,\bar{v}^T \\ \bar{p}\,\widetilde{q}^T & F \end{pmatrix} \widetilde{x} = \begin{pmatrix} \widehat{b}_1 - \widetilde{\chi}_0\,\widetilde{g} \\ \widehat{b}_2 - \bar{p}\,(\tau_{-1} + \widetilde{\chi}_0\,\widetilde{\rho}) \end{pmatrix} \equiv \begin{pmatrix} \widetilde{b}_1 \\ \widehat{b}_2 - \bar{p}\,\tau_0 \end{pmatrix}. \tag{3.8}$$

This is a system smaller in dimension than (3.4). To complete the recursion, in the following we rewrite it in the form of (3.4):

$$\bar{v} = \begin{pmatrix} \widetilde{\nu}^T \\ \widetilde{v} \end{pmatrix}, \quad \bar{p} = \begin{pmatrix} \widetilde{\pi}^T \\ \widetilde{p} \end{pmatrix}, \quad \text{and} \quad F = \begin{pmatrix} \widetilde{f}_1 & \widetilde{f}_2^T \\ \widetilde{f}_3 & \widetilde{F} \end{pmatrix},$$

where $\widetilde{\nu}^T$ and $\widetilde{\pi}^T$ are the first rows of $\bar{v}$ and $\bar{p}$, respectively; $\widetilde{f}_1 \in \mathbf{R}^{m-r_u}$; and $\widetilde{f}_2$ and $\widetilde{f}_3$ are column vectors of appropriate dimensions. Similarly to (3.3), the block form of F above is, strictly speaking, *not* a block partitioning of F, since the length of $\widetilde{f}_1$ is larger than 1 in general. $\widetilde{F}$ is itself a banded plus semiseparable rectangular matrix.

With this notation, we can now rewrite (3.8) in the form of (3.4) as

$$\begin{pmatrix} \dot{G} & \dot{E} \\ \dot{C} & \tilde{F} \end{pmatrix} \widetilde{x} = \dot{b} - \begin{pmatrix} 0 \\ \widetilde{p}\,\tau_0 \end{pmatrix}, \tag{3.9}$$

where

$$\dot{G} = \begin{pmatrix} \widetilde{G} & \widehat{u}\,\widetilde{\nu} \\ \widetilde{\pi}^T \widetilde{q}^T & \widetilde{f}_1 \end{pmatrix}, \quad \dot{C} = \widetilde{p} \begin{pmatrix} \widetilde{q}^T & \widetilde{\phi} \end{pmatrix}, \quad \dot{E} = \begin{pmatrix} \widehat{u} \\ \widehat{\mu}^T \end{pmatrix} \widetilde{v}^T, \quad \dot{b} = \begin{pmatrix} \widetilde{b}_1 \\ \widehat{b}_2 - \begin{pmatrix} \widetilde{\pi}^T \tau_0 \\ 0 \end{pmatrix} \end{pmatrix},$$

with $\widetilde{\phi}^T$ and $\widehat{\mu}^T$ being the $(\ell+1)$th and (r_u+2)th rows of q and u, respectively. Once again the block form of $\dot{G}$ is not a block partitioning.

As in section 2, we can perform elimination and forward substitution steps using formulas (3.4) through (3.9) recursively for some k times to obtain solution components $\widetilde{\chi}_0, \widetilde{\chi}_1, \ldots, \widetilde{\chi}_{k-1}$. We stop the recursion when the problem size $n-k$ in (3.9) becomes so small that $n-k \approx m$, at which point we solve it directly to get a solution $\widetilde{\chi}$.

To recover the solution to our original problem (1.1), let $H_0, H_1, \ldots, H_{k-1}$ be the elimination matrices used at the second elimination step defined by (3.6) and (3.7). We compute the solution to (1.1) as

$$x = \begin{pmatrix} I & 0 & 0 \\ 0 & H_0 & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & H_1 & 0 \\ 0 & 0 & I \end{pmatrix} \cdots \begin{pmatrix} I & 0 & 0 \\ 0 & H_{k-1} & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} \widetilde{\chi}_0 \\ \vdots \\ \widetilde{\chi}_{k-1} \\ \widetilde{\chi} \end{pmatrix}, \tag{3.10}$$

where the various identity matrices I are, in general, of different dimensions.

3.3. Efficiency and numerical stability considerations. In this section we consider special choices of matrices W and H in the recursion and how they affect the efficiency and numerical stability of the procedure. To make flop counting simpler, in this section we assume that $1 \ll r_l, r_u, b_l, b_u \ll n$ even though our algorithms work for general banded plus semiseparable matrices.

For complete backward stability, we can choose the W_0 matrices in (3.5) to be the product of r_u Givens rotations as suggested by Algorithm 3.1. The costs for computing $\widehat{u}$, $\widehat{G}$, and $\widehat{b}$ are about $3r_u^2$ flops, $6r_u\ell$ flops, and $6r_u$ flops, respectively. Hence the total cost for one step of (3.5) is about $3r_u(r_u+2\ell)$ flops.

We then choose H_0 in (3.6) as a Householder transformation. The costs for computing $\widehat{G}\,H_0$ and $H_0^T\,\bar{q}$ are about $4m\ell$ flops and $4r_l\ell$ flops, respectively. Hence the total cost for one step of (3.6) is about $4(m+r_l)\ell$ flops.

In (3.8), the costs for computing $\widetilde{b}_1$ and τ_0 are about $2m$ flops and $2r_l$ flops, respectively, leading to a total of $2(m+r_l)$ flops.

In (3.9), the main cost is to explicitly form the last row and column of $\dot{G}$. The costs for computing $\widehat{u}\,\widetilde{\nu}$ and $\widetilde{\pi}^T\,\widetilde{q}^T$ are about $2r_u^2$ flops and $2r_l\ell$ flops, respectively. There is essentially no cost for $\widetilde{f}_1$, which consists of the nonzero components of a column in the banded matrix D. Hence the total cost in (3.9) is about $2(r_u^2+r_l\ell)$ flops.

Since there are $k \approx n$ steps of recursion, the total cost for the procedure is about

$$\begin{aligned} &\left(3r_u(r_u+2\ell)+4(m+r_l)\ell+2(r_u^2+r_l\ell)\right)n \\ &\quad=\left(5r_u^2+2\left(2b_u+2b_l+3r_l+5r_u\right)\left(b_u+r_u\right)\right)n \quad \text{flops.} \end{aligned} \tag{3.11}$$

Additionally, there is a cost of about $6r_u^2 n$ flops for the preprocessing step (3.1).

With such choices of W_0 and H_0, we obtain a factorization (2.1) with orthogonal matrices W and H. Since only orthogonal transformations and one forward substitution are used for the solution of (1.1), this algorithm is backward stable.

To reduce computational cost, we can also choose W_0 via the banded Gaussian elimination procedure with row pivoting in Golub and Van Loan [5, Chap. 4]. Also, we can choose H_0 as a Gaussian elimination matrix with column pivoting. This choice of W_0 and H_0 leads to a factorization (2.1) with upper triangular matrices W and H. It is quite interesting to note that factorizations of this form do not seem to have been previously discussed in the literature.

With this choice of W_0 and H_0, the cost for one step of (3.5) is about $r_u(r_u+2\ell)$ flops; the cost for one step of (3.6) is about $2(m+r_l)\ell$ flops; and the total cost in (3.9) is about $2(r_u^2+r_l\ell)$ flops. With $k \approx n$ steps of recursion, the total cost for the procedure is about

$$\begin{aligned} &\left(r_u(r_u+2\ell)+2(m+r_l)\ell+2(r_u^2+r_l\ell)\right)n \\ &\quad=\left(3r_u^2+2\left(b_u+2b_l+2r_l+2r_u\right)\left(b_u+r_u\right)\right)n \quad \text{flops.} \end{aligned} \tag{3.12}$$

Additionally, there is a cost of about $6r_u^2 n$ flops for the preprocessing step (3.1).

It is well known that Gaussian elimination with partial pivoting could occasionally become numerically unstable if certain element growth is too large (see Golub and Van Loan [5, Chap. 3]). Thus, the numerical stability of Gaussian elimination procedures in (3.5) and (3.6) could not be guaranteed for a large value of r_u or b_u. In fact, the above procedure will be unstable for the case where $r_l = r_u = 0$, $b_l = b_u = k \gg 1$, and the first k rows of A are all zero, except leading k columns which contain the $k \times k$ matrix A_1 where (see Golub and Van Loan [5, Chap. 3])

$$A_1' = \begin{pmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ \vdots & \ddots & \ddots & & \vdots \\ -1 & \cdots & -1 & 1 & 1 \\ -1 & \cdots & -1 & -1 & 1 \end{pmatrix}.$$

Alternatively, we can choose only one of W_0 and H_0 to be orthogonal, leading to a factorization (2.1) with one of W and H orthogonal and the other upper triangular. Furthermore, the choices of W_0 and H_0 can change from one recursion step to another, leading to a factorization (2.1) with no obvious structures in W and H.

While our algorithms were presented in such a way that only one variable in (3.4) is eliminated in forward substitution at every recursion step, it is possible to reorganize the computation to develop a block version where a number of variables are all eliminated at the same time. Given the success of blocking in the recent linear algebra package LAPACK [1], it seems clear that when the dimension becomes very large, the problem (1.1) can be solved more efficiently by block versions of our algorithms.

Finally, we note that the problem (1.1) can be rewritten in the form

$$B\,y = S\,b, \quad B = S\,A\,S, \quad \text{and} \quad x = S\,y, \tag{3.13}$$

where S is the matrix with ones on the main antidiagonal and zero elsewhere.[2] It is easy to verify that

$$B = (S\,D\,S) + \mathbf{tril}\left((\mathbf{S\,u})\ (\mathbf{S\,v})^{\mathbf{T}}, -\mathbf{b_u} - \mathbf{1}\right) + \mathbf{triu}\left((\mathbf{S\,p})\ (\mathbf{S\,q})^{\mathbf{T}}, \mathbf{b_l} + \mathbf{1}\right).$$

It can be verified that $(S\,D\,S)$ is a banded matrix with b_u nonzero diagonals strictly below the main diagonal and b_l nonzero diagonals strictly above the main diagonal. Hence, B is itself a banded plus semiseparable matrix with the banded plus semiseparable structure of A'. Applying the two algorithms we just discussed to solve (3.13), we see that the total costs are

$$\left(5r_l^2 + 2\,(2b_l + 2b_u + 3r_u + 5r_l)\,(b_l + r_l)\right) n \quad \text{flops} \tag{3.14}$$

and

$$\left(3r_l^2 + 2\,(b_l + 2b_u + 2r_u + 2r_l)\,(b_l + r_l)\right) n \quad \text{flops}, \tag{3.15}$$

respectively. This suggests that one should choose among the two forms (1.1) and (3.13) according to formulas (3.11), (3.12), (3.14), and (3.15) to reduce computational cost.

4. Numerical experiments. In this section, we summarize the results from our numerical experiments with the algorithms that were presented in section 3. These experiments were performed on an UltraSparc 2 workstation in MATLAB with double precision $\epsilon \approx 2 \times 10^{-16}$.

We tested the following two algorithms:

- Algorithm I: Only Gaussian elimination steps with partial pivoting were used in computing (2.1).
- Algorithm II: Only Givens rotations and Householder reflections were used in computing (2.1).

In all of the test matrices, we chose $r_l = n/10$, $r_u = n/250$, $b_u = 10$, and $b_l = 10$. The matrix entries were generated randomly.

In Table 4.1, we compared Algorithms I and II in terms of the numbers of flops required to solve (1.1). The column marked GEPP is the number of flops required for Gaussian elimination with partial pivoting to solve (1.1) by treating A as a dense matrix. We see that Algorithm I requires less flops than Algorithm II, and both Algorithms I and II require significantly fewer flops than GEPP to solve (1.1).

In Table 4.2, we compared Algorithms I and II in terms of execution times and backward errors. The execution times are in seconds, and the backward error is defined as

$$\frac{\|A\,\widehat{x} - b\|_\infty}{\|A\|_\infty\,\|\widehat{x}\|_\infty},$$

where $\widehat{x}$ is the computed solution to (1.1). This backward error is the smallest relative backward error in the ∞-norm (see [5, Chap. 3]). Clearly Algorithm I is faster than Algorithm II as expected. Both are comparable in terms of backward errors. However, as we mentioned in section 3, the numerical stability of Algorithm I could not be guaranteed for a large value of b_u or r_u.

[2] For example, when $n = 2$, we have

$$S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

TABLE 4.1
Numbers of flops.

n	Algorithm I	Algorithm II	GEPP
250	5.5×10^5	8.7×10^5	1.0×10^7
500	2.1×10^6	3.2×10^6	8.3×10^7
750	4.7×10^6	7.2×10^6	2.8×10^8
1000	8.7×10^6	1.3×10^7	6.7×10^8
1250	1.4×10^7	2.2×10^7	1.3×10^9
1500	2.2×10^7	3.3×10^7	2.3×10^9
1750	3.1×10^7	4.7×10^7	3.6×10^9
2000	4.3×10^7	6.4×10^7	5.3×10^9
2250	5.6×10^7	8.4×10^7	7.6×10^9
2500	7.3×10^7	1.1×10^8	1.0×10^{10}

TABLE 4.2
Execution times and backward errors.

	Time (seconds)		Backward Error	
n	Algorithm I	Algorithm II	Algorithm I	Algorithm II
250	7.6×10^{-1}	1.0×10^0	6.1×10^{-19}	1.6×10^{-18}
500	2.2×10^0	3.2×10^0	1.5×10^{-19}	5.8×10^{-19}
750	4.5×10^0	5.7×10^0	3.6×10^{-20}	2.0×10^{-19}
1000	8.4×10^0	1.1×10^1	6.1×10^{-20}	2.0×10^{-19}
1250	1.3×10^1	1.6×10^1	4.8×10^{-20}	6.3×10^{-20}
1500	1.9×10^1	2.3×10^1	5.4×10^{-20}	3.8×10^{-19}
1750	2.5×10^1	3.1×10^1	2.8×10^{-20}	2.9×10^{-20}
2000	3.3×10^1	4.1×10^1	4.3×10^{-20}	5.3×10^{-20}
2250	4.1×10^1	5.1×10^1	5.0×10^{-20}	3.4×10^{-19}
2500	5.2×10^1	6.3×10^1	5.5×10^{-20}	2.2×10^{-19}

5. Conclusions and future work. In this paper we presented fast and numerically stable algorithms for the solution of linear systems of equations, where the coefficient matrix has the banded plus semiseparable structure (1.2). We also presented numerical results that clearly showed the stability and efficiency of these methods. It turns out that the two-sided elimination approach developed in this paper can be applied to a much broader class of matrices, including the $\mathcal{H}$-matrices of Hackbusch and his colleagues [9, 10, 11]. Our future work will concentrate on generalizing our methods to efficiently and stably solve linear systems of equations involving these and other structured matrices.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[3] Y. Eidelman and I. Gohberg, *Inversion formulas and linear complexity algorithm for diagonal plus semiseparable matrices*, Comput. Math. Appl., 33 (1997), pp. 69–79.

[4] Y. Eidelman and I. Gohberg, *A look-ahead block Schur algorithm for diagonal plus semiseparable matrices*, Comput. Math. Appl., 35 (1998), pp. 25–34.

[5] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[6] W. Govaerts, *Stable solvers and block elimination for bordered systems*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 469–483.

[7] L. Greengard and V. Rokhlin, *On the numerical solution of two-point boundary value problems*, Comm. Pure Appl. Math., 44 (1991), pp. 419–452.

[8] J.-Y. Lee and L. Greengard, *A fast adaptive numerical method for stiff two-point boundary value problems*, SIAM J. Sci. Comput., 18 (1997), pp. 403–429.

[9] W. Hackbusch and B. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic: General complexity estimates*, J. Comput. Appl. Math., 125 (2000), pp. 479–501.

[10] W. Hackbusch, B. N. Khoromskij, and S. A. Sauter, *On $\mathcal{H}^2$-matrices*, in Lectures on Applied Mathematics, H.-J. Bungartz, R. H. W. Hoppe, and C. Zenger, eds., Springer-Verlag, Berlin, 2000, pp. 9–29.

[11] W. Hackbusch and B. N. Khoromskij, *Towards $\mathcal{H}$-matrix approximation of linear complexity*, in Problems and Methods in Mathematical Physics, J. Elschner, I. Hohberg, and B. Silbermann, eds., Birkhäuser, Basel, 2001, pp. 194–220.

[12] S. van Huffel and H. Park, *Efficient reduction algorithms for bordered band matrices*, J. Numer. Linear Algebra Appl., 2 (1995), pp. 95–113.

[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[14] H. P. Starr, Jr., *On the Numerical Solution of One-Dimensional Integral and Differential Equations*, Ph.D. thesis, Department of Computer Science, Yale University, New Haven, CT, 1992.

[15] G. W. Stewart, *Updating a rank-revealing ULV decomposition*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 494–499.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 385–404

A STABLE DIVIDE AND CONQUER ALGORITHM FOR THE UNITARY EIGENPROBLEM*

MING GU†, ROBERT GUZZO‡, XUE-BIN CHI§, AND XING-QIN CAO¶

Abstract. We present a divide and conquer algorithm for computing the eigendecomposition of a unitary upper Hessenberg matrix H. Previous divide and conquer approaches suffer a potential loss of orthogonality among the computed eigenvectors of H. Using a backward stable method based on previous work by Gu and Eisenstat in the rank-one modification of the symmetric eigenproblem, our algorithm provides a backward stable method for computing the eigenvectors. The method also compares well against the efficiency of other available methods.

Key words. eigendecomposition, unitary eigenproblem, divide and conquer

AMS subject classification. 65F15

DOI. 10.1137/S0895479899359539

1. Introduction.

1.1. Problem defined. In this paper, we describe a stable and efficient method for determining the spectral resolution of a unitary[1] upper Hessenberg matrix H of order n,

$$H = \begin{bmatrix} -\bar{\gamma}_0\gamma_1 & -\bar{\gamma}_0\sigma_1\gamma_2 & -\bar{\gamma}_0\sigma_1\sigma_2\gamma_3 & \cdots & -\bar{\gamma}_0\sigma_1\cdots\sigma_{n-1}\gamma_n \\ \sigma_1 & -\bar{\gamma}_1\gamma_2 & -\bar{\gamma}_1\sigma_2\gamma_3 & \cdots & -\bar{\gamma}_1\sigma_2\cdots\sigma_{n-1}\gamma_n \\ & \sigma_2 & -\bar{\gamma}_2\gamma_3 & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \sigma_{n-1} & -\bar{\gamma}_{n-1}\gamma_n \end{bmatrix}, \tag{1.1}$$

where σ_k are real and positive, $|\gamma_k|^2 + \sigma_k^2 = 1$ for $1 \le k < n, \gamma_0 = 1$, and $|\gamma_n| = 1$ [17]. We call the γ_k's the *Schur parameters* of H and the σ_k's the *complementary parameters* of H.

We seek the spectral resolution of H,

$$H = W\,\Omega\,W^*, \tag{1.2}$$

where the columns of the matrix W are the eigenvectors of H, and Ω is a diagonal matrix whose diagonal entries are the eigenvalues corresponding to the eigenvectors

*Received by the editors July 28, 1999; accepted for publication (in revised form) by L. Reichel March 26, 2002; published electronically August 19, 2003.
http://www.siam.org/journals/simax/25-2/35953.html

†Department of Mathematics, University of California, Berkeley, CA 94720-3840 (mgu@math.berkeley.edu). This research was supported in part by NSF Career Award CCR-9702866 and by Alfred Sloan Research Fellowship BR-3720.

‡Department of Mathematics, University of California, Los Angeles, CA 90095-1555 (rguzzo@math.ucla.edu). This research was supported in part by NSF Career Award CCR-9702866.

§Computer Network Information Center, Chinese Academy of Sciences, P.O. Box 349, Beijing 100080, China (chi@jupiter.cnc.ac.cn). This research was supported in part by Major State Basic Research Project G1999032805.

¶Department of Computer Science, Huazhong University of Science and Technology, Wuhan, China.

[1]A unitary matrix H satisfies that $H^*H = HH^* = I$, where $*$ denotes complex conjugate transposition.

in W. It is easy to show that since H is unitary, W can also be chosen to be unitary, and the eigenvalues of H must have unit modulus [13].

There are two general methods available for calculating the spectral resolution of H: QR algorithms and divide and conquer algorithms. Various QR algorithms have been developed which compute solutions to the eigenproblem in a stable fashion. Recent work by Ammar, Gragg, and Reichel [1], Gragg [10, 11], and Stewart [17] has shown certain QR algorithms to be quite stable. However, there are certain advantages to divide and conquer strategies proposed by Ammar, Reichel, and Sorensen [5] and Gragg and Reichel [13]. Namely, such methods can be implemented much more efficiently and are better suited to parallel implementation. In fact, such strategies have been used to solve the symmetric tridiagonal eigenvalue problem with great success (see Cuppen [7], Dongarra and Sorensen [8], and Gu and Eisenstat [14, 15]).

The traditional problem with divide and conquer methods is numerical instability, especially in regard to calculating the eigenvectors of H (see Ammar, Reichel, and Sorensen [5] and Stewart [17]). On the contrary, the method presented here will be numerically stable, guaranteeing that the columns of W are numerically orthogonal and that the eigenvalues of H all lie on the unit circle in the complex plane. Our extensive numerical experiments indicate that our method compares very well against existing methods in both efficiency and accuracy (see section 5).

It is helpful to note that the interest surrounding this problem arises out of signal processing applications [4]—more specifically, in frequency estimation, including Pisarenko's method [2]. The applications to signal processing are closely related to the computation of Gauss–Szegö quadrature rules, which is discussed more fully in [13].

Throughout the paper, we use the usual model of floating point arithmetic,

$$\mathbf{fl}(x \circ y) = (x \circ y)(1 + \xi),$$

where x and y are floating point numbers; $\circ$ is one of $+, -, \times, \div$; $\mathbf{fl}(x \circ y)$ is the floating point result of the operation; and $|\xi| \leq \epsilon$ is the machine precision. We also require that

$$\mathbf{fl}(\sqrt{x}) = \sqrt{x}(1 + \xi)$$

for any positive floating point number x.

Let $\hat{x}$ be an approximation to $x \neq 0$. For the purpose of this paper, we say that $\hat{x}$ is close to x (to high absolute accuracy) if $x - \hat{x} = O(\epsilon)$, and we say that $\hat{x}$ is close to $x \neq 0$ to high relative accuracy if $(x - \hat{x})/x = O(\epsilon)$. Finally, we shall let $\|\cdot\|$ denote the vector 2-norm.

The rest of the paper is organized as follows. In section 2, we introduce the unitary divide and conquer (UDC) algorithm presented in [5, 12], which is referred to as "old UDC" or "original UDC" in this paper. This algorithm is a FORTRAN implementation of the method introduced by Gragg and Reichel [12, 13]. In the same section, we will also introduce our new method, referred to as "new UDC" or "our UDC algorithm." The new UDC is a modification of the old UDC, extending previous work by Gu and Eisenstat in [14, 15]. In section 3, we discuss the nature of the rootfinder used in the new method as well as provide a specific way to handle eigenvalues. In section 4 we prove the numerical stability of our method. Finally, in section 5, we will present some numerical results for various types of eigenproblems.

2. Solving the unitary eigenproblem recursively. From the Schur parameters and complementary parameters of H in (1.1), we can uniquely represent H in

its *Schur parametric form* [17],

$$H = H(\gamma_1, \gamma_2, \ldots, \gamma_n) = G_1 G_2 \cdots G_{n-1} \widetilde{G}_n, \tag{2.1}$$

where each $G_k \in \mathbf{C^{n\times n}}, 1 \le k < n$, is a Givens matrix,

$$G_k = \begin{bmatrix} I_{k-1} & & & \\ & -\gamma_k & \sigma_k & \\ & \sigma_k & \bar{\gamma}_k & \\ & & & I_{n-k-1} \end{bmatrix}, \quad \gamma_k \in \mathbf{C},\ \sigma_k \in \mathbf{R},\ \sigma_k \ge 0,\ |\gamma_k|^2 + \sigma_k^2 = 1,$$

and $\widetilde{G}_n$ is the diagonal matrix

$$\widetilde{G}_n = \begin{bmatrix} I_{n-1} & \\ & -\gamma_n \end{bmatrix}, \quad \gamma_n \in \mathbf{C}, |\gamma_n| = 1.$$

Given the matrix H in upper Hessenberg form, it is easy to compute the Schur parameters (for details, see [13]). Working with the Schur parameters and complementary parameters of H, instead of with H itself, will greatly reduce the computational complexity of the algorithm. It would appear that we can further reduce the amount of storage necessary by storing only the γ_k values and calculating the σ_k values as needed. However, this calculation could lead to numerical instability should any of the $|\gamma_k|$ be close to one (see Stewart [17]).

2.1. The divide phase. The idea behind divide and conquer is to obtain the spectral resolution of H from the spectral resolution of two subproblems. As described in [5] (for details, see [13]), we will make use of the fact that a complex Givens matrix G_s is diagonally unitarily equivalent with a real Givens reflector, which can be written as a Householder transformation. Define

$$\gamma_s' = \begin{cases} \gamma_s/|\gamma_s|, & \gamma_s \ne 0, \\ 1, & \gamma_s = 0. \end{cases}$$

Then $|\gamma_s'| = 1$ and

$$\begin{bmatrix} I_{s-1} & & \\ & \bar{\gamma}_s' & \\ & & I_{n-s} \end{bmatrix} G_s \begin{bmatrix} I_s & & \\ & \gamma_s' & \\ & & I_{n-s-1} \end{bmatrix} = \begin{bmatrix} I_{s-1} & & & \\ & -|\gamma_s| & \sigma_s & \\ & \sigma_s & |\gamma_s| & \\ & & & I_{n-s-1} \end{bmatrix}.$$

The right-hand side above can be written as a Householder transformation $I - 2ww^*$, where $w \in \mathbf{R}^n$ satisfies

$$w = \omega_s e_s + \omega_{s+1} e_{s+1} \ \text{ with } \ \omega_s = ((1 + |\gamma_s|)/2)^{1/2}, \ \ \omega_{s+1} = -\sigma_s/(2(1 + |\gamma_s|))^{1/2},$$

and e_j denotes the jth axis vector whose length may vary depending on the context. We can now express (2.1) as two subproblems "pasted" together by the Householder transformation,

$$H = \begin{bmatrix} H_1 & \\ & I_{n-s} \end{bmatrix} (I - 2ww^*) \begin{bmatrix} I_s & \\ & H_2 \end{bmatrix}, \tag{2.2}$$

where, using the notation in (2.1),

$$H_1 = H(\gamma_1, \ldots, \gamma_{s-1}, -\gamma_s^{'}) \in \mathbf{C}^{s\times s},$$
$$H_2 = H(\bar{\gamma}_s^{'}\gamma_{s+1}, \bar{\gamma}_s^{'}\gamma_{s+2}, \ldots, \bar{\gamma}_s^{'}\gamma_n) \in \mathbf{C}^{(n-s)\times(n-s)}.$$

For the purposes of divide and conquer, we assume that we know the spectral resolutions of the two submatrices,

$$H_k = W_k \Lambda_k W_k^*, \qquad k = 1, 2,$$

where the W_k are unitary and the Λ_k are diagonal. Now we seek the spectral resolution of the original matrix, H. Define

$$\widetilde{W} = \begin{bmatrix} W_1 & \\ & W_2 \end{bmatrix}, \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix}, \quad z = \begin{bmatrix} W_1^* e_s \omega_s \\ \bar{\Lambda}_2 W_2^* e_1 \omega_{s+1} \end{bmatrix}.$$

Note that $z^*z = 1$, in exact arithmetic. Substitution of the above into (2.2) renders the following formulation:

$$H = \widetilde{W}\Lambda(I - 2zz^*)\widetilde{W}^*. \tag{2.3}$$

Since $\widetilde{W}$ is unitary, (2.3) reveals that H and the core matrix

$$A = \Lambda(I - 2zz^*) \quad \text{with} \quad z^*z = 1 \tag{2.4}$$

have the same eigenvalues. Let $A = U\Lambda' U^*$ be the eigendecomposition of A. Then the eigendecomposition (1.2) for H is simply

$$\Omega = \Lambda' \quad \text{and} \quad W = \widetilde{W}\,U. \tag{2.5}$$

Since A is a rank-one modification on diagonal Λ, we will determine the spectral resolution of A by using a similar strategy as in [14].

To complete the divide phase, we choose, say, $s = \lfloor n/2 \rfloor$. We then recursively apply the dividing strategy (2.2) to H_1 and H_2, respectively, until their dimensions are sufficiently small (less than 10, for example), resulting in $O(\log_2 n)$ levels of recursion. We can obtain the spectral resolution of the sufficiently small problems at the bottom of the recursion tree directly with little effort. To obtain the spectral resolution of the original matrix H in (1.1), we solve the core problems of the form (2.4) at every level of the recursion tree in a bottom-up fashion. The eigenvectors of H can be recursively computed as $\widetilde{W}\,U$ (see (2.5)). The total cost for this algorithm is $O(n^3)$ flops.[2] Note that the actual cost of this algorithm can sometimes be much lower due to deflation (see section 2.3).

Similar to the divide and conquer methods for the symmetric tridiagonal eigenvalue problem, the above recursion can also be simplified into a divide and conquer method for computing the eigenvalues of H only, with a total cost of $O(n^2)$ flops (see Ammar, Reichel, and Sorensen [5], Cuppen [7], Dongarra and Sorensen [8], Gragg and Reichel [13], and Gu and Eisenstat [14, 15]).

[2] A *flop* is a floating point operation $x \circ y$, where x and y are floating point numbers and $\circ$ is one of $+, -, \times$, or $\div$.

2.2. Roots of the spectral function. To determine the eigenvalues of A, we must find the roots of its characteristic polynomial,

$$\chi(\lambda) = \det(A - \lambda I) = \det(\Lambda - \lambda I)\left(1 - 2z^*(\Lambda - \lambda I)^{-1}\Lambda z\right) = 0.$$

Thus, the eigenvalues of A include the roots of the spectral function,

$$\phi(\lambda) = 1 - 2z^*(\Lambda - \lambda I)^{-1}\Lambda z = 0.$$

Since $z^*z = 1$ according to (2.4), we can rewrite the spectral function as

$$\phi(\lambda) = \sum_{j=1}^{n} |z_j|^2 \frac{\lambda + \lambda_j}{\lambda - \lambda_j} = 0. \tag{2.6}$$

Recall that the eigenvalues for a unitary upper Hessenberg matrix all lie on the unit circle. Hence, the eigenvalues of A and Λ can be written as $\lambda = \exp(i\theta)$ and $\lambda_j = \exp(i\theta_j)$, respectively, where $i = \sqrt{-1}$, and we restrict $-\pi < \theta_j \leq \pi$. Substitution into (2.6) renders

$$\Phi(\theta) = -i\phi(\lambda) = \sum_{j=1}^{n} |z_j|^2 \cot\left(\frac{\theta - \theta_j}{2}\right) = 0. \tag{2.7}$$

Thus, finding the roots of the spectral function is equivalent to finding the roots of $\Phi(\theta)$. Inspection of this function shows that Φ has n poles on the interval $(-\pi, \pi]$, occurring at each of the θ_j's. Also, Φ is a monotonically decreasing function on any interval between two adjacent poles. Following Golub [9], we call (2.7) the *secular equation*. We will talk about how to find roots of this equation in section 3.2.

2.3. Deflation. The work of divide and conquer methods can be reduced (sometimes dramatically) by the deflation procedure described in [5]. If two diagonal entries of Λ, λ_j, and λ_k are identical or have very close arguments, then λ_j can be regarded as an eigenvalue of A. If some component z_j of z is zero or has very small magnitude, then λ_j can again be regarded as an eigenvalue of A. In both cases, A can be reduced to a matrix with similar structure but smaller dimension. This will reduce the amount of computation involved in finding the eigenvalues, since there are fewer roots of Φ. Used to full advantage, deflation can also reduce the amount of calculation involved in computing the eigenvectors of A and H. More fundamentally, the stability of our method relies on the assumption that deflation has been done (see section 4). Similar deflation procedures have also been used in the numerical solution of the symmetric tridiagonal eigenproblem (see Cuppen [7], Dongarra and Sorensen [8], and Gu and Eisenstat [14, 15]).

From now on, we will assume that the deflation procedure of [5] has been applied to A in (2.4). We assume that $n > 1$ and that the θ's are ordered in the following way:

$$-\pi < \theta_1 < \theta_2 < \cdots < \theta_n \leq \pi. \tag{2.8}$$

Our implementation of the deflation procedure ensures that

$$|z_j| \geq \epsilon^{'}, \;\; \theta_{k+1} - \theta_k \geq \epsilon^{''}, \;\; \text{and} \;\; (\pi + \theta_1) + (\pi - \theta_n) \geq \epsilon^{''}, \tag{2.9}$$

where $\epsilon^{'}$ and $\epsilon^{''}$ are some specified deflation tolerances to be discussed in more detail in section 4. The last condition in (2.9) ensures that angles between any two eigenvalues, both in the clockwise direction and counterclockwise direction around the circle, are at least as big as $\epsilon^{''}$.

The conditions in (2.9) imply that the eigenvalues of H, $\lambda_j^{'}$ for $1 \le j \le n$ strictly interlace on the unit circle with the λ_j [5]. Let $\hat{\lambda}_j$ be the computed eigenvalues of H. Since all eigenvalues of H are on the unit circle, we can further write

$$\lambda_j^{'} = \exp\left(i\theta_j^{'}\right) \quad \text{and} \quad \hat{\lambda}_j = \exp\left(i\hat{\theta}_j\right). \tag{2.10}$$

2.4. Unstable eigenvector formulas. It follows from (2.4) that the normalized eigenvector of A associated with $\lambda_j^{'}$ satisfies the following formulas:

$$v_j = \frac{u}{\|u\|}, \qquad u = (I - \Lambda^* \lambda_j^{'})^{-1} z. \tag{2.11}$$

The UDC algorithm in [5, 12] computes approximations $\hat{\theta}_j$ by solving (2.7), and computes the eigenvectors of A and H using (2.11), with $\lambda_j^{'}$ replaced by $\hat{\lambda}_j$. However, due to the potential ill-conditioning in the eigenvectors, the vectors computed this way can often lose mutual orthogonality in finite precision, leading to inaccurate spectral resolution of A and H (see Ammar, Reichel, and Sorensen [5], Gragg and Reichel [13], and Sun [18]). Similar instability problems also occurred in the old divide and conquer methods for the symmetric tridiagonal eigenproblem (see Dongarra and Sorensen [8] and Gu and Eisenstat [14, 15]).

2.5. Stable eigenvector formulas. To develop a stable method for computing the eigenvectors, we first rewrite the kth component of u as follows:

$$\begin{aligned} u_k &= \frac{z_k}{1 - \lambda_j^{'}\bar{\lambda}_k} = \frac{z_k}{1 - \lambda_j^{'}/\lambda_k} \\ &= \frac{z_k}{1 - \cos(\theta_j^{'} - \theta_k) - i\sin(\theta_j^{'} - \theta_k)}. \end{aligned}$$

Making use of the double-angle formulas, we find that

$$\begin{aligned} u_k &= \frac{z_k}{2\sin^2\left(\frac{\theta_j^{'} - \theta_k}{2}\right) - 2i\sin\left(\frac{\theta_j^{'} - \theta_k}{2}\right)\cos\left(\frac{\theta_j^{'} - \theta_k}{2}\right)} \\ &= \frac{i}{2}\exp\left(i\left(\frac{\theta_k - \theta_j^{'}}{2}\right)\right) \cdot \left(z_k \Big/ \sin\left(\frac{\theta_j^{'} - \theta_k}{2}\right)\right). \end{aligned} \tag{2.12}$$

From the above formulation, we observe that the eigenvectors of H can be directly calculated in terms of the poles and roots of the spectral function and the components of z. Furthermore, if $\theta_j^{'}$ were known exactly and could be exactly represented as a floating point number, then we would be able to compute u_k to full accuracy using (2.12).

Of course, the angles $\hat{\theta}_j$ computed by our rootfinder by solving (2.7) are only *approximations* to $\theta_j^{'}$. If $\hat{\theta}_j$ is used in place of $\theta_j^{'}$ in (2.12) to compute u_k, the computed u_k can incur a very large relative error, which can lead to loss of orthogonality among computed eigenvectors. In other words, (2.12) is still an unstable way to compute the eigenvectors.

It turns out that a stable method for computing the eigenvectors can be developed by constructing a new matrix,

$$\widehat{A} = \Lambda(I - 2\gamma\hat{z}\hat{z}^*), \tag{2.13}$$

where γ is a complex scalar. It is clear that $\widehat{A}$ has a structure similar to A. We choose the scalar γ and vector $\hat{z}$ so that the exact eigenvalues of $\widehat{A}$ are the eigenvalues we computed for A. In section 2.6 we will show that this matrix does exist and is in fact unitary with distinct eigenvalues. Hence the eigenvectors of $\widehat{A}$ are always mutually orthogonal.

Similar to formulas (2.11) and (2.12), the eigenvector of $\widehat{A}$ associated with $\hat{\lambda}_j$ can be computed as

$$\hat{v}_j = \frac{\hat{u}}{\|\hat{u}\|}, \tag{2.14}$$

where the kth component of $\hat{u}$ satisfies

$$\hat{u}_k = \frac{i}{2} \exp\left(i\left(\frac{\theta_k - \hat{\theta}_j}{2}\right)\right) \cdot \left(\hat{z}_k \Big/ \sin\left(\frac{\hat{\theta}_j - \theta_k}{2}\right)\right). \tag{2.15}$$

Note that γ does not appear in (2.14) and (2.15). It also follows from (2.15) that

$$\|\hat{u}\| = \frac{1}{2}\sqrt{\sum_{k=1}^{n}\left(\hat{z}_k \Big/ \sin\left(\frac{\hat{\theta}_j - \theta_k}{2}\right)\right)^2}. \tag{2.16}$$

In section 2.6, we show that the vector $\hat{z}$ can be computed to high relative accuracy from the $\hat{\theta}$'s, and in section 3.2 we show that the denominators in formulas (2.15) and (2.16), $\sin(\frac{\hat{\theta}_j - \theta_k}{2})$, can also be computed to high relative accuracy. Consequently, we can compute $\|\hat{u}\|$ to high relative accuracy as well. In addition, it is clear that we can compute the unit modulus term $\exp(i(\frac{\theta_k - \hat{\theta}_j}{2}))$ in (2.15) to high relative accuracy. It now follows from (2.14) and (2.15) that we can compute the eigenvector $\hat{v}_j$ to high relative accuracy.

The above analysis implies that we can compute all the the eigenvectors of $\widehat{A}$ to high relative accuracy regardless of its eigenvalue distribution. Since $\widehat{A}$ is itself unitary, these computed eigenvectors will be numerically orthogonal. In our UDC algorithm, we use these vectors as approximations to the eigenvectors of A. In section 4, we justify this approach by showing that the matrix $\widehat{A}$ is very close to A in finite precision, and hence the spectral resolution for $\widehat{A}$ is a good approximation of that of A. A similar approach has been taken by Gu and Eisenstat in the rank-one modification of the symmetric eigenproblem [14].

2.6. Building $\widehat{A}$. In the following we construct the matrix $\widehat{A}$ by deriving formulas for γ and the components of the vector $\hat{z}$. We assume that the deflation procedure has been performed on A, and thus the deflation criteria (2.9) hold.

It has already been stated that the n roots of Φ strictly interlace its poles on the unit circle. Our rootfinder discussed in section 3.2 guarantees that the computed angles $\hat{\theta}$'s satisfy

$$-\pi < \theta_1 < \hat{\theta}_1 < \theta_2 < \hat{\theta}_2 < \cdots < \hat{\theta}_{n-1} < \theta_n < \hat{\theta}_n < \theta_1 + 2\pi. \tag{2.17}$$

Note that the unique $\hat{\theta}_n$ that satisfies (2.17) may actually be greater than π. We will have further discussion on $\hat{\theta}_n$ in section 3.2 (see (3.6)).

The characteristic polynomial for $\widehat{A}$ can be written as follows:

$$\begin{aligned}\widehat{\chi}(\lambda) &= \det(\widehat{A} - \lambda I) = \det(\Lambda - \lambda I)\left(1 - 2\gamma \hat{z}^*(\Lambda - \lambda I)^{-1}\Lambda \hat{z}\right) \\ &= \det(\Lambda - \lambda I)\left(1 - 2\gamma \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j - \lambda}|\hat{z}_j|^2\right) \\ &= \prod_{j=1}^{n}(\lambda_j - \lambda) - 2\gamma \sum_{j=1}^{n}\left(\frac{\lambda_j |\hat{z}_j|^2 \prod_{l=1}^{n}(\lambda_l - \lambda)}{\lambda_j - \lambda}\right). \end{aligned} \tag{2.18}$$

On the other hand, the fact that the $\hat{\lambda}_j$'s are the eigenvalues of $\widehat{A}$ implies that

$$\hat{\chi}(\lambda) = \prod_{j=1}^{n}(\hat{\lambda}_j - \lambda). \tag{2.19}$$

Combining (2.18) and (2.19), and setting $\lambda = \lambda_k$ for $k = 1, 2, \ldots, n$, we obtain

$$\prod_{j=1}^{n}(\hat{\lambda}_j - \lambda_k) = -2\gamma\lambda_k|\hat{z}_k|^2 \prod_{j=1,\ j\neq k}^{n}(\lambda_j - \lambda_k). \tag{2.20}$$

Solving for $\gamma|\hat{z}_k|^2$, and using the same calculations as in (2.12), we get

$$\begin{aligned}\gamma|\hat{z}_k|^2 &= -\frac{\prod_{j=1}^{n}(\hat{\lambda}_j - \lambda_k)}{2\lambda_k \prod_{j\neq k}(\lambda_j - \lambda_k)} \quad = \quad -\frac{\prod_{j=1}^{n}(\hat{\lambda_j}/\lambda_k - 1)}{2\prod_{j\neq k}(\lambda_j/\lambda_k - 1)} \\ &= -i\exp\left(i\sum_{j=1}^{n}\frac{\hat{\theta}_j - \theta_j}{2}\right)\frac{\prod_{j=1}^{n}\sin((\hat{\theta}_j - \theta_k)/2)}{\prod_{j\neq k}\sin((\theta_j - \theta_k)/2)}. \end{aligned} \tag{2.21}$$

In the following, we discuss how to choose γ and $\hat{z}_k$ according to (2.21). To this end, we rewrite the ratio of products in (2.21) as

$$\frac{\prod_{j=1}^{n}\sin((\hat{\theta}_j - \theta_k)/2)}{\prod_{j\neq k}\sin((\theta_j - \theta_k)/2)} = \sin((\hat{\theta}_k - \theta_k)/2) \cdot \left(\prod_{j\neq k}\frac{\sin((\hat{\theta}_j - \theta_k)/2)}{\sin((\theta_j - \theta_k)/2)}\right). \tag{2.22}$$

The interlacing property (2.17) implies that

$$\begin{aligned} 0 &< \frac{\hat{\theta}_j - \theta_k}{2}, \quad \frac{\theta_j - \theta_k}{2} < \pi \quad \text{if } j > k, \\ -\pi &< \frac{\hat{\theta}_j - \theta_k}{2}, \quad \frac{\theta_j - \theta_k}{2} < 0 \quad \text{if } j < k, \end{aligned} \tag{2.23}$$

and

$$0 < (\hat{\theta}_k - \theta_k)/2 < \pi. \tag{2.24}$$

It follows from these relations that the first term in (2.22) must be positive; it also follows that every ratio in the product in (2.22) must be positive. Hence the left-hand side in (2.22) must be positive for every k.

This discussion suggests that the following choice of γ and $\hat{z}_k$ satisfies (2.21):

$$|\hat{z}_k| = \sqrt{\frac{\prod_{j=1}^{n} \sin((\hat{\theta}_j - \theta_k)/2)}{\prod_{j \neq k} \sin((\theta_j - \theta_k)/2)}} \quad \text{and} \quad \gamma = -i \exp\left(i \sum_{j=1}^{n} \frac{\hat{\theta}_j - \theta_j}{2} \right). \tag{2.25}$$

Furthermore, since z_k is usually a complex number, we choose the phase angle of $\hat{z}_k$ to be that of z_k. Hence

$$\hat{z}_k = |\hat{z}_k| \, \frac{z_k}{|z_k|} \quad \text{for} \quad 1 \leq k \leq n. \tag{2.26}$$

To complete the construction for $\widehat{A}$, we note that by working the above steps backward it is straightforward to verify that (2.13), (2.26), and (2.25) indeed uniquely define a matrix $\widehat{A}$ that has the $\hat{\lambda}_j$'s as its eigenvalues. Formulas similar to (2.25) and (2.26) have also been derived in [3] in the context of an algorithm for an inverse eigenvalue problem for unitary Hessenberg matrices.

In the following, we show that $\widehat{A}$ is unitary. According to (2.10),

$$\begin{aligned} 1 &= \left| \prod_{j=1}^{n} \hat{\lambda}_j \right| = \left| \det(\widehat{A}) \right| = \left| \det\left(\Lambda(I - 2\gamma \hat{z}\hat{z}^*) \right) \right| \\ &= |\det(\Lambda)| \cdot |\det(I - 2\gamma \hat{z}\hat{z}^*)| = |1 - 2\gamma \hat{z}^* \hat{z}| \,. \end{aligned}$$

The last equation implies that

$$\gamma + \bar{\gamma} - 2|\gamma|^2 \hat{z}^* \hat{z} = 0.$$

Consequently,

$$\begin{aligned} \widehat{A}^* \, \widehat{A} &= (I - 2\gamma \hat{z}\hat{z}^*)^* \Lambda^* \Lambda (I - 2\gamma \hat{z}\hat{z}^*) \\ &= (I - 2\gamma \hat{z}\hat{z}^*)^* (I - 2\gamma \hat{z}\hat{z}^*) = I - 2\left(\gamma + \bar{\gamma} - 2|\gamma|^2 \hat{z}^* \hat{z} \right) \hat{z}\hat{z}^* = I. \end{aligned}$$

Finally, we note that the components of the vector z can also be expressed in terms of Λ and the eigenvalues of A. Indeed, (2.21) now becomes

$$|z_k|^2 = -i \exp\left(i \sum_{j=1}^{n} \frac{\theta_j' - \theta_j}{2} \right) \frac{\prod_{j=1}^{n} \sin((\theta_j' - \theta_k)/2)}{\prod_{j \neq k} \sin((\theta_j - \theta_k)/2)}. \tag{2.27}$$

Since the λ_j' are the eigenvalues of H and, since $z^* z = 1$,

$$\begin{aligned} \exp\left(i \sum_{j=1}^{n} \theta_j' \right) &= \prod_{j=1}^{n} \lambda_j' = \det(H) = \det\left(\Lambda(I - 2zz^*) \right) \\ &= -\det(\Lambda) = -\prod_{j=1}^{n} \lambda_j = -\exp\left(i \sum_{j=1}^{n} \theta_j \right). \end{aligned}$$

It follows that

$$\exp\left(i \sum_{j=1}^{n} \frac{\theta_j' - \theta_j}{2} \right) = i. \tag{2.28}$$

In light of the above discussion, we can rewrite (2.27) as

$$|z_k| = \sqrt{\frac{\prod_{j=1}^{n} \sin((\theta_j' - \theta_k)/2)}{\prod_{j\neq k} \sin((\theta_j - \theta_k)/2)}}. \tag{2.29}$$

3. Some computational issues.

3.1. The FORTRAN sine function. Formulas (2.15), (2.16), and (2.25) all involve the sine function. To guarantee numerical stability, we would like to compute every sine term as accurately as we can. Throughout this paper, we assume the following.

Assumption 3.1. The FORTRAN sine function computes $\sin(\psi)$ to high relative accuracy for $|\psi| \leq \pi/2$.

It is not realistic to require the FORTRAN sine function to compute $\sin(\psi)$ to high relative accuracy for any ψ. In fact, since $\sin(\pm\pi) = 0$, large relative errors are hard to avoid for any FORTRAN sine function if ψ is very close to $\pm\pi$.

In the following, we show that for $|\psi| \leq \pi/2$, a small relative change in ψ can only imply a small relative change in $\sin(\psi)$. This is trivially true for $\psi = 0$. For $\psi \neq 0$ and any $|\varepsilon| \ll 1$,

$$\begin{aligned}\sin(\psi(1+\varepsilon)) - \sin\psi &= \sin\psi \cdot (\cos(\psi\varepsilon) - 1) + \sin(\psi\varepsilon)\cdot\cos\psi \\ &= -2\sin\psi\cdot\sin^2(\psi\varepsilon/2) + \sin(\psi\varepsilon)\cdot\cos\psi.\end{aligned}$$

Taking absolute value, we have

$$\begin{aligned}|\sin(\psi(1+\varepsilon)) - \sin\psi| &\leq 2\left|\sin\psi\cdot\sin^2(\psi\varepsilon/2)\right| + |\sin(\psi\varepsilon)\cdot\cos\psi| \\ &\leq 2\,|\sin(\psi\varepsilon/2)| + |\sin(\psi\varepsilon)| \leq 2\cdot|\psi\varepsilon/2| + |\psi\varepsilon| \\ &= 2\,|\psi\varepsilon| \leq \pi\,|\varepsilon\sin\psi|,\end{aligned}$$

where we have used the fact that

$$\frac{2}{\pi} \leq \frac{\sin\psi}{\psi} \leq 1 \quad \text{for} \quad 0 < |\psi| \leq \pi/2. \tag{3.1}$$

It now follows that

$$\frac{|\sin(\psi(1+\varepsilon)) - \sin\psi|}{|\sin\psi|} \leq \pi|\varepsilon| \quad \text{for} \quad 0 < |\psi| \leq \pi/2. \tag{3.2}$$

Assumption 3.1 and relation (3.2) imply that the sine terms in formulas (2.15), (2.16), and (2.25) can be computed to high relative accuracy if their arguments are between $-\pi/2$ and $\pi/2$ and are computed to high relative accuracy. In section 3.2, we will further discuss how to compute the sine terms in these formulas accurately when the arguments are not between $-\pi/2$ and $\pi/2$.

3.2. The rootfinder and computing angles. Our rootfinder for finding the roots of (2.7) is basically the cubically convergent rootfinder developed in [5, 13], with a number of modifications aimed at improving numerical accuracy. We assume that the deflation procedure has been performed on A, and thus relations (2.9) and (2.8) hold. The n roots θ_j' of Φ satisfy strict interlacing properties similar to (2.17).

In each interval, (θ_j, θ_{j+1}) for $j < n$, denote

$$\alpha_j = \theta_j' - \theta_j \quad \text{and} \quad \beta_j = \theta_{j+1} - \theta_j'. \tag{3.3}$$

If $\theta_j^{'}$ is closer to θ_j, the rootfinder computes an approximation $\hat{\alpha}_j$ to α_j. It then computes $\hat{\beta}_j$, the approximation to β_j, according to the following:

$$\hat{\beta}_j = (\theta_{j+1} - \theta_j) - \hat{\alpha}_j. \tag{3.4}$$

If $\theta_j^{'}$ is closer to θ_{j+1}, then the rootfinder computes an approximation $\hat{\beta}_j$ to β_j; it then computes the approximation $\hat{\alpha}_j$ from (3.4). We will postpone discussion on the computation of $\theta_n^{'}$ to the end of section 3.2.

By computing the smaller of the two angles between the root $\theta_j^{'}$ and its two nearest poles, we prevent any catastrophic cancellation when the root is extremely close to one of the poles. With $\hat{\alpha}_j$ and $\hat{\beta}_j$, the difference between $\theta_j^{'}$ and any pole θ_k can be approximated as

$$\hat{\theta}_j - \theta_k = \begin{cases} \hat{\alpha}_j + (\theta_j - \theta_k) & \text{for } \theta_k \le \theta_j, \\ (\theta_{j+1} - \theta_k) - \hat{\beta}_j & \text{for } \theta_k > \theta_j. \end{cases}$$

This way, we can compute $\hat{\theta}_j - \theta_k$ to high relative accuracy, given $\hat{\alpha}_j$ and $\hat{\beta}_j$. In particular, we avoid any catastrophic cancellation in the event that $\theta_j^{'}$ is very close to one of the poles. According to (3.1), we can also compute $\sin((\hat{\theta}_j - \theta_k)/2)$ to high relative accuracy if $|\hat{\theta}_j - \theta_k|/2 \le \pi/2$. A similar result holds for $\sin((\theta_j - \theta_k)/2)$.

To accurately compute the sine terms in (2.15), (2.16), and (2.25) when the arguments are not between $-\pi/2$ and $\pi/2$, we recall that the eigenvalues all lie on the unit circle. Therefore, calculating angles between eigenvalues can be done in either the clockwise or counterclockwise direction around the circle. If the angle between two points on the unit circle is calculated in the counterclockwise direction to be close to 2π, then in the clockwise direction, the angle is close to zero. We achieve this effect when we make the following alternate formulation:

$$\sin\left(\frac{\theta_j - \theta_k}{2}\right) = \begin{cases} -\sin\left(\dfrac{\nu_1 + (\theta_j - \theta_1) + (\theta_n - \theta_k) + \nu_n}{2}\right) & \text{if } (\theta_j - \theta_k)/2 < -\pi/2, \\ \sin\left(\dfrac{\nu_n + (\theta_n - \theta_j) + (\theta_k - \theta_1) + \nu_1}{2}\right) & \text{if } (\theta_j - \theta_k)/2 > \pi/2, \end{cases} \tag{3.5}$$

where $\nu_1 = \pi + \theta_1$ and $\nu_n = \pi - \theta_n$. Given ν_1 and ν_n, the arguments on the right-hand side of (3.5) can be computed to high relative accuracy, as can the sine function. We also make a similar reformulation to $\sin((\hat{\theta}_j - \theta_k)/2)$. However, since π is not a floating point number, it sometimes may not be possible to compute ν_1 and ν_n to high relative accuracy. See section 4 for further discussion on their computation.

Now, we address the issue of computing $\theta_n^{'}$. In the spirit of the above discussion, let α_n and β_n be the smaller angles on the circle between $\theta_n^{'}$ and its nearest poles, θ_n and $\theta_1 + 2\pi$. If $\alpha_n \le \beta_n$, then our rootfinder computes an approximation $\hat{\alpha}_n$ by solving (2.7) and computes $\hat{\beta}_n$ from $\hat{\alpha}_n$ using the following formula: $\hat{\beta}_n = (\theta_1 + 2\pi - \theta_n) - \hat{\alpha}_n = (\nu_1 + \nu_n) - \hat{\alpha}_n$. Otherwise, it computes $\hat{\beta}_n$ by solving (2.7) and computes $\hat{\alpha}_n$ from $\hat{\beta}_n$. After $\hat{z}$ and the eigenvectors for $\widehat{A}$ are computed from $\hat{\alpha}$'s, $\hat{\beta}$'s, and θ's, we compute

$$\hat{\theta}_n = \begin{cases} \theta_n + \hat{\alpha}_n & \text{if } \theta_n + \hat{\alpha}_n \le \pi, \\ \theta_n + \hat{\alpha}_n - 2\pi & \text{if } \theta_n + \hat{\alpha}_n > \pi. \end{cases} \tag{3.6}$$

This formula ensures that $\hat{\theta}_n$ will satisfy $-\pi < \hat{\theta}_n \leq \pi$ after the eigendecomposition of A is computed.

3.3. The stopping criterion. In practice a rootfinder cannot be expected to make progress at a point λ, where it is impossible to determine the sign of $\Phi(\theta)$. Motivated by [14], we use the following stopping criterion in the rootfinder:

$$\sum_{k=1}^{n} |z_k|^2 \left| \mathbf{fl}\left(\cot\left(\frac{\hat{\theta}_j - \theta_k}{2} \right) \right) \right| \leq \eta \sum_{k=1}^{n} \frac{|z_k|^2}{\left| \mathbf{fl}\left(\sin((\hat{\theta}_j - \theta_k)/2) \right) \right|}, \tag{3.7}$$

where η is some appropriately chosen multiple of machine precision, and

$$\mathbf{fl}\left(\cot\left(\frac{\hat{\theta}_j - \theta_k}{2} \right) \right) \quad \text{and} \quad \mathbf{fl}\left(\sin((\hat{\theta}_j - \theta_k)/2) \right)$$

are the floating point results of computing the cot and sin functions by computing the arguments with the procedure described in section 3.2. Similar to [14], the right-hand side in (3.7) is an upper bound on the round-off error in evaluating $\Phi(\hat{\theta}_j)$. Using arguments similar to those in [14], it can be shown that the set of approximate solutions satisfying (3.7) is nonempty in finite precision for any j. We would expect a good rootfinder to be able to compute such approximate solutions. In our FORTRAN implementation, we used a modified version of the rootfinder in [5, 13].

4. Numerical stability of the method. In this section we show that $\widehat{A}$ is close to A. Consider the following:

$$\begin{aligned} A - \widehat{A} &= \Lambda(I - 2zz^*) - \Lambda(I - 2\gamma\hat{z}\hat{z}^*) = 2\Lambda\left(\gamma\hat{z}\hat{z}^* - zz^*\right) \\ &= 2\Lambda\left((\gamma - 1)\hat{z}\hat{z}^* + (\hat{z} - z)\hat{z}^* + z(\hat{z} - z)^*\right). \end{aligned}$$

So, to show that $\widehat{A}$ is close to A, we need only show that γ and $\hat{z}$ are close to 1 and z, respectively.

Before our formal analysis, we note that the secular equation (2.7) is derived under the condition that $\|z\|$ be exactly 1 (cf. (2.4)), which rarely holds in practice. In addition, our analysis below will require that ν_1 and ν_n be computed to high relative accuracy, which may not be possible if ν_1 is close to $-\pi$ or ν_n close to π. To simplify the analysis, we assume for the moment that vector z in (2.4) satisfies $\|z\| = 1$ exactly and that scalars ν_1 and ν_n in (3.5) are known to high relative accuracy. We will come back to these assumptions at the end of section 4.

Under our assumption on the high relative accuracy in ν_1 and ν_n, the formulas established in section 3.2 for computing the sine function guarantee that we can compute $\sin((\theta_j - \theta_k)/2)$ and $\sin((\hat{\theta}_j - \theta_k)/2)$ to high relative accuracy for any j and k. Let us denote

$$d_{jk} = (\theta_j - \theta_k)/2, \quad d'_{jk} = (\theta'_j - \theta_k)/2, \quad \text{and} \quad \hat{d}_{jk} = (\hat{\theta}_j - \theta_k)/2.$$

Since $\Phi(\theta'_j) = 0$, and

$$\sin\left((\hat{\theta}_j - \theta'_j)/2\right) = \sin(\hat{d}_{jk})\cos(d'_{jk}) - \sin(d'_{jk})\cos(\hat{d}_{jk}),$$

we have

$$-\Phi(\hat{\theta}_j) = \Phi(\theta'_j) - \Phi(\hat{\theta}_j) = \sin\left((\hat{\theta}_j - \theta'_j)/2\right) \sum_{k=1}^{n} \frac{|z_k|^2}{\sin(d'_{jk})\sin(\hat{d}_{jk})}.$$

The rootfinder guarantees that $\hat{\theta}_j$ and $\theta_j^{'}$ are in the same interval (θ_j, θ_{j+1}) for $j < n$ and that $\hat{\theta}_n$ and $\theta_n^{'}$ are in the same interval $(\theta_n, \theta_1 + 2\pi)$, which ensures that the product $\sin(\hat{d}_{jk})\sin(d_{jk}^{'})$ is always positive.

Combining the above equation with stopping criterion (3.7), we have

$$\begin{aligned}\left|\sin\left((\hat{\theta}_j - \theta_j^{'})/2\right)\right| \sum_{k=1}^{n} \frac{|z_k|^2}{|\sin(d_{jk}^{'})\sin(\hat{d}_{jk})|} &= \left|\sin\left((\hat{\theta}_j - \theta_j^{'})/2\right)\right| \left|\sum_{k=1}^{n} \frac{|z_k|^2}{\sin(d_{jk}^{'})\sin(\hat{d}_{jk})}\right| \\ &= |\Phi(\hat{\theta}_j)| \le \eta \sum_{k=1}^{n} \frac{|z_k|^2}{|\sin(\hat{d}_{jk})|} \le \eta \sum_{k=1}^{n} \frac{|z_k|^2}{|\sin(d_{jk}^{'})\sin(\hat{d}_{jk})|}. \end{aligned} \tag{4.1}$$

This yields the following result:

$$\left|\sin\left((\hat{\theta}_j - \theta_j^{'})/2\right)\right| \le \eta. \tag{4.2}$$

Hence

$$\left|\lambda_j^{'} - \hat{\lambda}_j\right| = \left|\exp(\theta_j^{'}) - \exp(\hat{\theta}_j)\right| = 2\left|\sin\left((\hat{\theta}_j - \theta_j^{'})/2\right)\right| \le 2\eta, \tag{4.3}$$

which is to say that the eigenvalues are computed to full accuracy.

Note that the third condition in (2.9) guarantees that $\hat{\theta}_j - \theta_j^{'}$ cannot be too close to $\pm 2\pi$ for $j \le n$:

$$\left|\left(\hat{\theta}_j - \theta_j^{'}\right)/2\right| \le \pi - \epsilon^{''}/2.$$

We choose $\sin(\epsilon^{''}/2) > \eta$. These two conditions guarantee that $\hat{\theta}_j$ and $\theta_j^{'}$ in (4.2) must satisfy

$$\left|\hat{\theta}_j - \theta_j^{'}\right| \le 2\sin^{-1}\eta. \tag{4.4}$$

Now we use the above inequality to show that γ is close to 1. It follows from (2.25) and (2.28) that $\gamma = \exp(i\sum_{j=1}^{n} \frac{\hat{\theta}_j - \theta_j^{'}}{2})$. Combining this with (4.4), we have

$$\begin{aligned} |\gamma - 1| &= \left|\prod_{j=1}^{n} \exp\left(i\left(\frac{\hat{\theta}_j - \theta_j^{'}}{2}\right)\right) - 1\right| \le \prod_{j=1}^{n}\left(1 + \left|\exp\left(i\left(\frac{\hat{\theta}_j - \theta_j^{'}}{2}\right)\right) - 1\right|\right) - 1 \\ &= \prod_{j=1}^{n}\left(1 + 2\left|\sin\left(\frac{\hat{\theta}_j - \theta_j^{'}}{4}\right)\right|\right) - 1 \le \prod_{j=1}^{n}\left(1 + \left|\frac{\hat{\theta}_j - \theta_j^{'}}{2}\right|\right) - 1 \\ &\le \left(1 + \sin^{-1}\eta\right)^n - 1 \le e^{n\sin^{-1}\eta} - 1 \le (e-1)n\sin^{-1}\eta, \end{aligned} \tag{4.5}$$

where we have used the fact that $(e^x - 1)/x \le e - 1$ for $0 \le x \le 1$.

To show that $\hat{z}$ is close to z, we need the following lemma.

LEMMA 4.1. *Let* $\sin y \ne 0$. *Then*

$$\sqrt{\left|\frac{\sin x}{\sin y}\right|} + \left|\frac{\sin(x \pm y)}{\sin y}\right| \ge \frac{2}{\pi} \quad \textit{and} \quad \sqrt{|\sin x\ \sin y|} \le |\sin x| + |\sin(x \pm y)|. \tag{4.6}$$

Proof. To prove the first inequality in (4.6), we first consider the case $0 \le x \le \pi/2$ and $0 < y \le \pi/2$. Since it holds trivially if $x \ge y$, we further assume that $x < y$. Using (3.1), we get that

$$\sqrt{\left|\frac{\sin x}{\sin y}\right|} + \left|\frac{\sin(x \pm y)}{\sin y}\right| \ge \sqrt{\frac{\sin x}{\sin y}} + \frac{\sin(y-x)}{\sin y} \ge 2/\pi \left(\sqrt{\frac{x}{y}} + \frac{y-x}{y}\right)$$
$$= 2/\pi \left(1 + \sqrt{\frac{x}{y}}\left(1 - \sqrt{\frac{x}{y}}\right)\right) \ge 2/\pi.$$

Hence the first relation in (4.6) holds when both $0 \le x \le \pi/2$ and $0 < y \le \pi/2$. Replacing x by $\pi - x$ in the inequality, the resulting inequality is exactly the same, and hence it holds when $\pi/2 < x \le \pi$. Similarly, the inequality still holds when $\pi/2 \le y < \pi$. Thus, It holds for any value of x and any $\sin y \ne 0$ due to periodicity.

To prove the second inequality in (4.6), we also restrict our attention to the special case $0 \le x < y \le \pi/2$. Let $a = y - x$. Then

$$|\sin x| + |\sin(x \pm y)| \ge \sin x + \sin a \ge \sqrt{\sin^2 x + \sin x \sin a}$$
$$\ge \sqrt{\sin^2 x \cos a + \sin x \sin a \cos x} = \sqrt{\sin x \sin(a + x)} = \sqrt{\sin x \sin y}.$$

Hence the second inequality in (4.6) holds when $0 \le x < y \le \pi/2$. By arguments used earlier in the proof, it is straightforward to further conclude that this inequality holds for any x and y. □

Letting $x = \hat{d}_{jk}$ and $y = d'_{jk}$ in (4.6), we have

$$\frac{1}{|\sin(\hat{d}_{jk})|} \le \frac{\pi/2}{\sqrt{|\sin(\hat{d}_{jk})\sin(d'_{jk})|}} + \frac{\pi/2\,|\sin(\hat{d}_{jk} - d'_{jk})|}{|\sin(\hat{d}_{jk})\sin(d'_{jk})|}.$$

Note that $\hat{d}_{jk} - d'_{jk} = (\hat{\theta}_j - \theta'_j)/2$. Plugging the above into the first inequality in (4.1) and simplifying, we have

$$\left|\sin\left((\hat{\theta}_j - \theta'_j)/2\right)\right| \sum_{k=1}^{n} \frac{|z_k|^2}{|\sin(d'_{jk})\sin(\hat{d}_{jk})|} \le \frac{\pi\,\eta/2}{1 - \pi\,\eta/2} \sum_{k=1}^{n} \frac{|z_k|^2}{\sqrt{|\sin(d'_{jk})\sin(\hat{d}_{jk})|}}$$
$$\le \frac{\|z\|\,\pi\,\eta/2}{1 - \pi\,\eta/2} \sqrt{\sum_{k=1}^{n} \frac{|z_k|^2}{|\sin(d'_{jk})\sin(\hat{d}_{jk})|}},$$

where we have used the Cauchy–Schwarz inequality. Further simplifying, we have

$$\left|\sin\left((\hat{\theta}_j - \theta'_j)/2\right)\right| \le \frac{\|z\|\,\pi\,\eta/2}{1 - \pi\,\eta/2} \Bigg/ \sqrt{\sum_{k=1}^{n} \frac{|z_k|^2}{|\sin(d'_{jk})\sin(\hat{d}_{jk})|}}$$
$$\le \frac{\|z\|\,\pi\,\eta/2}{(1 - \pi\,\eta/2)|z_k|} \sqrt{|\sin(d'_{jk})\sin(\hat{d}_{jk})|}. \tag{4.7}$$

Setting $x = \hat{d}_{jk}$ and $y = d'_{jk}$ in the second inequality in (4.6), we have

$$\sqrt{|\sin(d'_{jk})\sin(\hat{d}_{jk})|} \le |\sin(d'_{jk})| + |\sin(\hat{d}_{jk} - d'_{jk})|$$
$$= |\sin(d'_{jk})| + |\sin((\hat{\theta}_j - \theta'_j)/2)|.$$

Plugging this into (4.7) and simplifying, we have

$$\left|\sin\left((\hat{\theta}_j - \theta_j^{'})/2\right)\right| \leq \frac{\|z\|\,\pi\,\eta/2}{(1-\pi\,\eta/2)|z_k| - \|z\|\,\pi\,\eta/2}|\sin(d_{jk}^{'})|. \tag{4.8}$$

Similar to (4.4), and in light of (3.1), we get from (4.8) that

$$\begin{aligned}\left|\hat{\theta}_j - \theta_j^{'}\right| &\leq 2\sin^{-1}\left(\frac{\|z\|\,\pi\,\eta/2}{(1-\pi\,\eta/2)|z_k| - \|z\|\,\pi\,\eta/2}|\sin(d_{jk}^{'})|\right)\\ &\leq \frac{\|z\|\,\pi^2\,\eta/2}{(1-\pi\,\eta/2)|z_k| - \|z\|\,\pi\,\eta/2}|\sin(d_{jk}^{'})| \stackrel{\text{def}}{=} \frac{2\delta_k\,\eta}{|z_k|}|\sin(d_{jk}^{'})|. \end{aligned}\tag{4.9}$$

In (2.9), we choose the deflation tolerance

$$\epsilon^{'} \geq \pi^2 n\eta/4 \geq \pi\eta/(1-\pi\eta).$$

This implies that

$$\delta_k = \frac{\|z\|\,\pi^2/4}{(1-\pi\,\eta/2) - \|z\|\,\pi\,\eta/(2|z_k|)} \leq \pi^2/2.$$

We are now in a position to show that $\hat{z}$ is close to z. Using (2.25), (2.26), and (2.29),

$$\begin{aligned}|\hat{z}_k - z_k| &= \left||\hat{z}_k|\,\frac{z_k}{|z_k|} - |z_k|\,\frac{z_k}{|z_k|}\right| = ||\hat{z}_k| - |z_k||\\ &= \left|\left|\frac{\prod_{j=1}^{n}\sin(\hat{d}_{jk})}{\prod_{j\neq k}\sin(d_{jk})}\right|^{1/2} - \left|\frac{\prod_{j=1}^{n}\sin(d_{jk}^{'})}{\prod_{j\neq k}\sin(d_{jk})}\right|^{1/2}\right|\\ &= |z_k|\left|\left|\prod_{j=1}^{n}\frac{\sin(\hat{d}_{jk})}{\sin(d_{jk}^{'})}\right|^{1/2} - 1\right|. \end{aligned}\tag{4.10}$$

We seek bounds on each factor of the product. Using (3.1) and the identity,

$$\sin(x+y) - \sin(x-y) = 2\sin(y)\cos(x),$$

we get

$$\left|\frac{\sin(\hat{d}_{jk})}{\sin(d_{jk}^{'})} - 1\right| = \left|\frac{2\sin(\frac{\hat{\theta}_j - \theta_j^{'}}{4})\ \cos(\frac{\hat{\theta}_j + \theta_j^{'}}{4} - \frac{\theta_k}{2})}{\sin(d_{jk}^{'})}\right| \leq \left|\frac{\frac{\hat{\theta}_j - \theta_j^{'}}{2}}{\sin(d_{jk}^{'})}\right|.$$

With (4.9) and the upper bound on δ_k, we have

$$\left|\frac{\sin(\hat{d}_{jk})}{\sin(d_{jk}^{'})} - 1\right| \leq \frac{\delta_k\,\eta}{|z_k|} \leq \frac{\pi^2\,\eta}{2|z_k|}. \tag{4.11}$$

Plugging this into (4.10), we obtain

$$|\hat{z}_k - z_k| \leq |z_k|\left(\left(1 + \frac{\pi^2\,\eta}{2|z_k|}\right)^{n/2} - 1\right) \leq |z_k|\left(e^{\pi^2 n\,\eta/(4|z_k|)} - 1\right).$$

Using the fact that $\pi^2 n\eta/(4|z_k|) \le 1$ and that $(e^x - 1)/x \le e - 1$ for $0 \le x \le 1$, we have

$$|\hat{z}_k - z_k| \le |z_k|(e-1)\pi^2 n\eta/(4|z_k|) \le \pi^2 n\eta/2. \tag{4.12}$$

This last relation implies that $\hat{z}$ is indeed close to z, and hence we conclude that $\widehat{A}$ is indeed close to A.

Finally, we address the issues regarding $\|z\|$ and the computed accuracy in ν_1 and ν_n. We note that since $\|z\|$ is close to 1, the matrix

$$\widetilde{A} \stackrel{\text{def}}{=} \Lambda\,(I - \tilde{z}\,\tilde{z}^*)\,, \quad \tilde{z} = z/\|z\|,$$

is close to A with $\|\tilde{z}\| = 1$. The stopping criteria (3.7) for $\widetilde{A}$ and A differ by a common factor of $1/\|z\|$ on both sides of the inequality, and hence are equivalent. Repeating the above analysis leading to (4.12), we conclude that $\widehat{A}$ is close to $\widetilde{A}$, and thus to A. We point out that $\widetilde{A}$ is constructed for the above analysis only and not for actual computation.

We have also developed a somewhat more detailed analysis paralleling the one in this section to show that $\widehat{A}$ is still close to $\widetilde{A}$ even if ν_1 and ν_n are computed only to high absolute accuracy. We omit it in our paper for the following reasons: first, this analysis is quite technical and does not provide additional insight; and second, ν_1 and ν_n can be easily computed to high relative accuracy with emulated extra precision techniques (see Priest [16]).

5. Numerical experiments. We now present some experimental results to compare the performance of our method against the "old UDC" described in [5] and the HQR methods in [6]. To make easy comparison with the FORTRAN subroutines UDC and CHSEQR, we have implemented our algorithm in FORTRAN as well. Below are four graphs representing the performance of the three algorithms. All three were run on a Sparc-20 workstation in single precision arithmetic, roughly corresponding to seven significant digits. Deflation tolerance was set to 10^{-6}.

We considered 20 matrices ranging from size 50 to size 1000, measuring the speed of the algorithms, the accuracy of the spectral resolution compared to the original matrix, and the orthogonality of the eigenvectors. For Figure 5.1, we measured the speed of the algorithms in seconds. To calculate how close the computed spectral resolution came to the approximating H, we took the infinity norm of $H\,W - W\,\Lambda$. If all arithmetic was done in exact precision, this residual should equal zero. Figure 5.3 illustrates the numerical value of the residual. Similarly, to calculate how close the eigenvectors come to being orthogonal, we took the infinity norm of $W^*W - I$. Figure 5.4 illustrates this numerical value.

We experimented on three kinds of matrices. In Figures 5.1, 5.3, and 5.4, the dotted line marked off with asterisks represents the performance of the HQR from LAPACK; the dashed line marked off with x's represents the performance of the "old UDC" code; and the solid line marked off with o's represents the performance of our method, the "new UDC" code.

Type I. In our first experiment, we simply considered randomly generated unitary upper Hessenberg matrices. Such an H is constructed by inputting the Schur parameters, $\gamma_j = \rho_j \exp(i\alpha_j), 1 \le j \le n$, where the α_j are uniformly distributed random variables on $[0, 2\pi]$ and the ρ_j are uniformly distributed on $[0, 1]$ and $\rho_n = 1$. This Schur decomposition ensures that H is unitary.

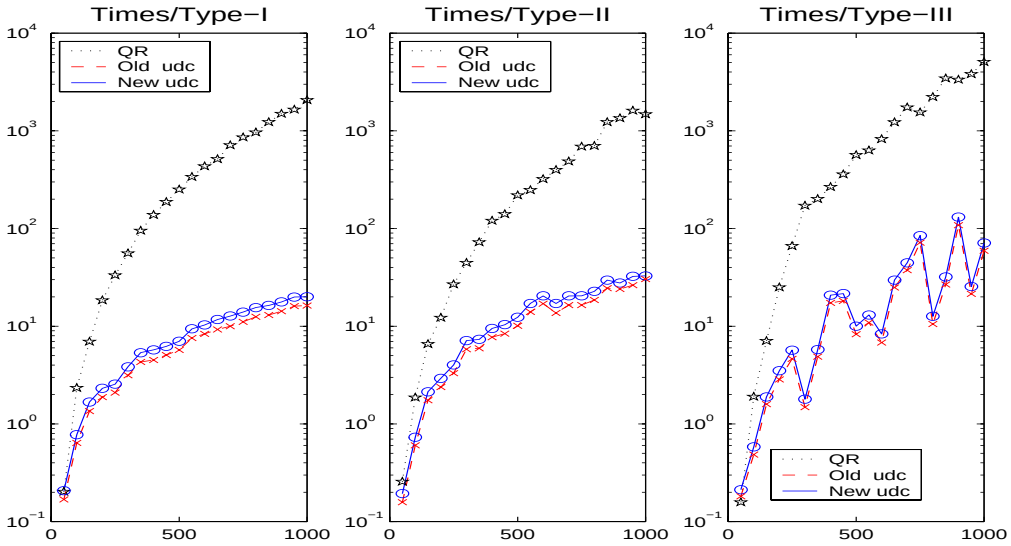


FIG. 5.1. *Efficiency of method.*

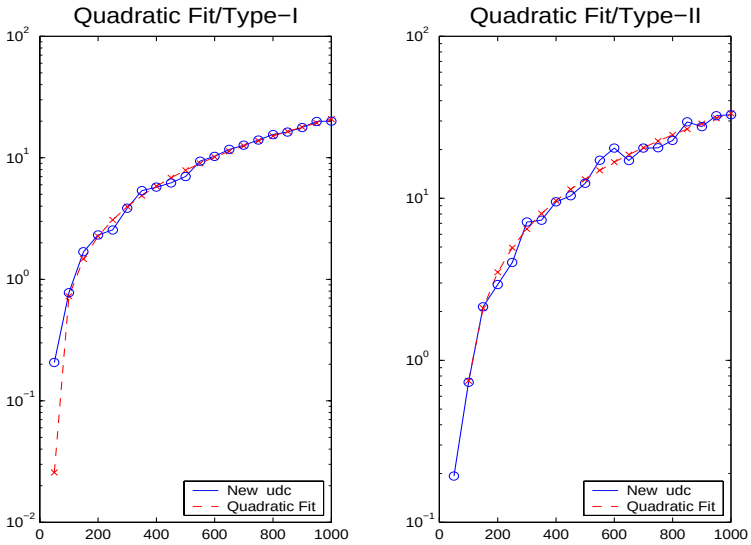


FIG. 5.2. *Fitting efficiency with quadratic least squares.*

The results of this experiment showed that our method improved upon the original UDC method by roughly a factor of 10 or more with regard to both the residual and orthogonality of the eigenvectors. It is also much faster and significantly more accurate than the HQR code. The original UDC performs only slightly faster then our method. Additionally, the speed of our method seemed to be on the order of the square of the size of the matrix, since the data seems to fit a quadratic polynomial of n quite well (see Figure 5.2).

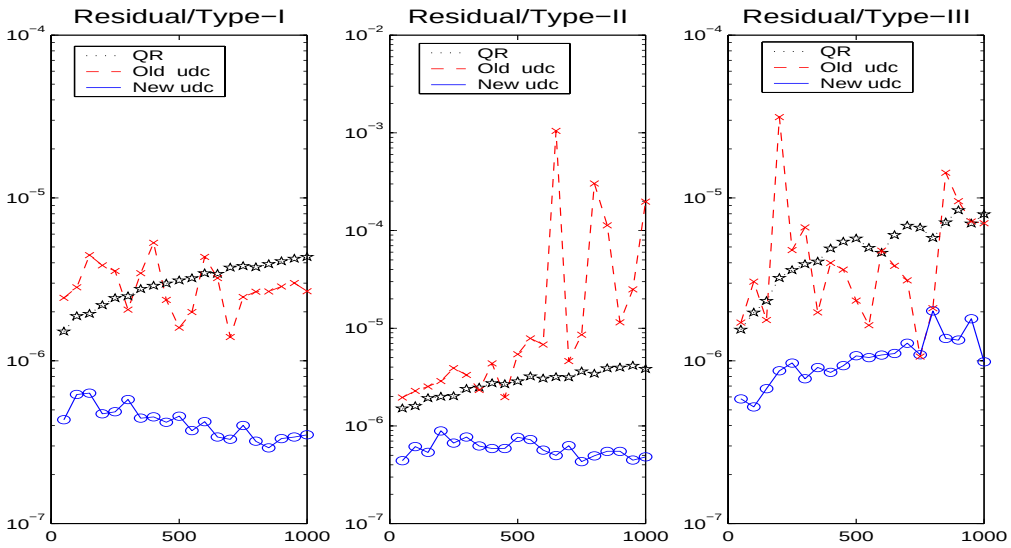


FIG. 5.3. $||W^*W - I||_\infty/\sqrt{n}$.

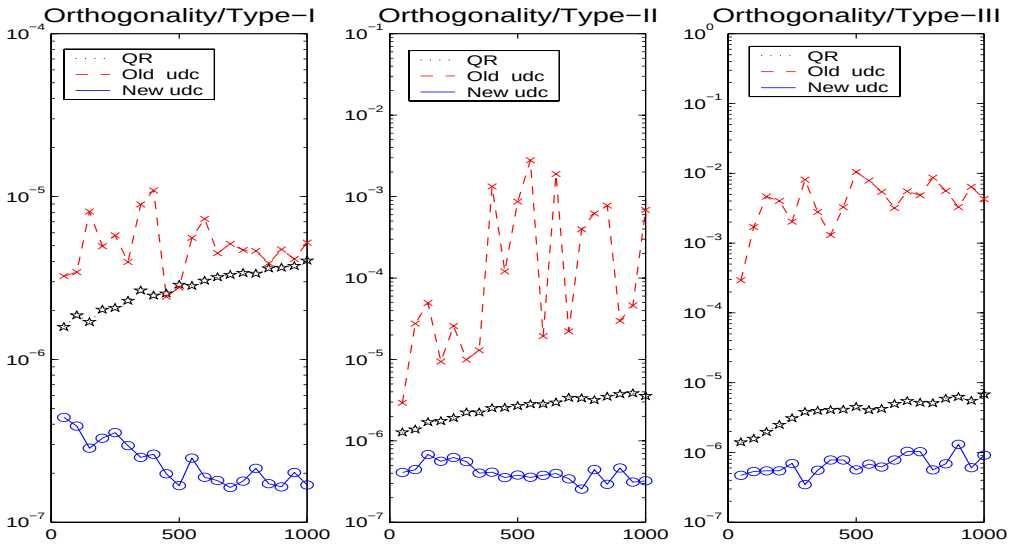


FIG. 5.4. $||HW - W\Lambda||_\infty/\sqrt{n}$.

Type II. In our next experiment, we considered matrices which have one or more eigenvalues whose arguments are near $\pm\pi$. This experiment is constructed by creating a real-valued matrix H with odd size. Then, one of the eigenvalues cannot have a distinct complex conjugate, thereby forcing that eigenvalue to equal 1 or -1. By setting $\gamma_n = -1$, we force the real eigenvalue to have an argument at π.

The results of this experiment showed remarkable improvement on the original UDC algorithm. For sizable matrices, the original method becomes highly unstable, producing inaccurate results. An examination of the results

for a matrix of size 651 reveals the residual and orthogonality results on the "old UDC" to be somewhere in the neighborhood of 10^{-3}, whereas the results for our method stayed stable around 10^{-6}, the deflation tolerance. Again, our method compares very favorably with the HQR code. Similar to the previous experiment, the efficiency of the new method is only slightly worse than the original method, and Figure 5.2 still suggests that the speed of our UDC is quadratic with respect to the size of the problem.

Type III. In the third experiment, we designed H to have nearly multiple eigenvalues. We do this by making H nearly block diagonal with identical blocks. As described in [5], we let $n = pk$. Generate the first $p-1$ Schur parameters as in the first experiment. Then set σ_p equal to some small constant. The remaining parameters are given by $\gamma_{lp+j} = \gamma_j, \sigma_{lp+j} = \sigma_j, 1 \leq j \leq p, 1 \leq l < k$. Then set $\gamma_n = 1$. If $\sigma_p = 0$, then the eigenvalues of H occur with multiplicity k. Otherwise, for small σ_p we get nearly multiple eigenvalues. For our experiments we chose $p = 40$.

Experimental results on the third experiment once again show a vast improvement over the original UDC method and HQR with regard to stability of the eigenvector calculations. Figure 5.1 indicates that the efficiency effect of deflation on Type III matrices is both dramatic and erratic, making it difficult to predict the speed of our UDC for Type III matrices.

6. Conclusion. This paper has outlined a stable algorithm for computing the spectral resolution of a unitary upper Hessenberg matrix. We showed that our algorithm is stable regardless of eigenvalue distribution of the given problem. The computed eigenvalues are all unit modulus, and the computed eigenvectors are all numerically orthogonal.

This method relied on several delicate techniques. First, as in all divide and conquer methods, we required a deflation procedure to ensure that we could find the roots of the spectral function. Additionally, in the calculation of the eigenvectors of H, special attention is given to the way that angles are handled. Finally, we used a matrix reconstruction idea from [14, 15] to guarantee that the computed eigenvectors are automatically orthogonal.

Future work includes parallelization of the new UDC algorithm and developing a simplified version for the special case where the input data are all real.

Acknowledgments. The authors are grateful to the anonymous referees for many suggestions that improved the presentation of the paper and for pointing out reference [3].

REFERENCES

[1] G. S. Ammar, W. B. Gragg, and L. Reichel, *On the eigenproblem for orthogonal matrices*, in Proceedings of the 25th Annual IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 1963–1966.

[2] G. S. Ammar, W. B. Gragg, and L. Reichel, *Determination of Pisarenko frequency estimates as eigenvalues of an orthogonal matrix*, in Advanced Algorithms and Architectures for Signal Processing II, Proceedings of SPIE, Vol. 826, F. T. Luk, ed., SPIE, San Diego, 1987, pp. 143–145.

[3] G. S. Ammar, W. B. Gragg, and L. Reichel, *Constructing a unitary Hessenberg matrix from spectral data,* in Numerical Linear Algebra, Digital Signal Processing, and Parallel Algorithms, G. H. Golub and P. van Dooren, eds., Springer-Verlag, New York, 1991, pp. 385–396.

[4] G. S. Ammar, W. B. Gragg, and L. Reichel, *Direct and inverse unitary eigenvalue problems in signal processing: An overview,* in Linear Algebra for Large Scale and Real Time Applications, NATO Adv. Sci. Inst. Ser. E Appl. Sci. 323, M.S. Moonen, G. H. Golub, and B. L. R. De Moor, eds., Kluwer, Dordrecht, 1993, pp. 341–343.

[5] G. S. Ammar, L. Reichel, and D.C. Sorensen, *An implementation of a divide and conquer algorithm for the unitary eigenproblem*, ACM Trans. Math. Software, 18 (1992), pp. 292–307.

[6] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, PA, 1999.

[7] J. J. M. Cuppen, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[8] 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.

[9] G. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[10] W. B. Gragg, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[11] W. B. Gragg, *Stabilization of the uhqr algorithm,* in Advances in Computational Mathematics, Lecture Notes in Pure and Appl. Math. 202, Z. Chen, Y. Li, C. Micchelli, and Y. Xu, eds., Marcel Dekker, Hong Kong, 1999, pp. 139–154.

[12] W. B. Gragg and L. Reichel, *A divide and conquer algorithm for the unitary eigenproblem*, in Hypercube Multiprocessors 1987, M. T. Heath., ed., SIAM, Philadelphia, PA, 1987, pp. 639–647.

[13] W. B. Gragg and L. Reichel, *A divide and conquer method for unitary and orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.

[14] M. Gu and S. C. Eisenstat, *A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1266–1276.

[15] M. Gu and S. C. Eisenstat, *A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 172–192.

[16] D. Priest, *Algorithms for arbitrary precision floating point arithmetic*, in Proceedings of the 10th Symposium on Computer Arithmetic, Grenoble, France, June 26–28, 1991, P. Kornerup and D. Matula, eds., IEEE Computer Society Press, Los Alamitos, CA, pp. 132–145.

[17] M. Stewart, *An error analysis of a unitary Hessenberg QR algorithm*, in Joint CS Technical Report Series, Australian National University, Canberra, Australia, 1998, pp. 17–29.

[18] J.-G. Sun, *Residual bounds on approximate solutions for the unitary eigenproblem*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 69–82.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 405–428

A KRYLOV SUBSPACE METHOD FOR QUADRATIC MATRIX POLYNOMIALS WITH APPLICATION TO CONSTRAINED LEAST SQUARES PROBLEMS*

REN-CANG LI† AND QIANG YE†

Abstract. We present a Krylov subspace–type projection method for a quadratic matrix polynomial $\lambda^2 I - \lambda A - B$ that works directly with A and B without going through any linearization. We discuss a special case when one matrix is a low rank perturbation of the other matrix. We also apply the method to solve quadratically constrained linear least squares problem through a reformulation of Gander, Golub, and von Matt as a quadratic eigenvalue problem, and we demonstrate the effectiveness of this approach. Numerical examples are given to illustrate the efficiency of the algorithms.

Key words. quadratic matrix polynomial, Krylov subspace, quadratic eigenvalue problem, least squares problem, quadratic constraint

AMS subject classifications. 65F15, 65F20, 15A18

DOI. 10.1137/S0895479802409390

1. Introduction. Krylov subspace techniques are widely used for solving linear systems of equations and eigenvalue problems involving large and sparse matrices [7, 14]. It has found applications in many other large scale matrix problems such as model reductions of linear input-output systems. The basic idea of the techniques is to extract information of an $n \times n$ matrix A most relevant to the underlying computational problem through utilizing the so-called *Krylov* subspace

$$\mathcal{K}_k(A, v) = \text{span}\{v, Av, \ldots, A^{k-1}v\}$$

or through utilizing two (row and column) Krylov subspaces $\mathcal{K}_k(A, v)$ and $\mathcal{K}_k(A^*, w)$ simultaneously, where v and w are vectors of dimension n and A^* is the conjugate transpose. This is realized by the *Lanczos/Arnoldi process* [1, 18]. See also [7, 14, 22, 28, 29].

The quadratic eigenvalue problem (QEP) in its generality takes the form

$$(\lambda^2 M + \lambda C + K)z = 0, \tag{1.1}$$

where M, C, K are $n \times n$ matrices, scalar λ is called an *eigenvalue*, and n-dimensional $0 \neq z$ is a corresponding (right) *eigenvector*. In solving it when n is large and M, C, K are sparse, it is often transformed implicitly into a mathematically equivalent *monic* QEP

$$(\lambda^2 I_n - \lambda A - B)x = 0, \tag{1.2}$$

where A and B stay in some factored forms so that the matrix-vector multiplications by A and B are cheap. (It is possible that λ in (1.2) differs from the one in the

*Received by the editors June 7, 2002; accepted for publication (in revised form) by H. A. van der Vorst March 10, 2003; published electronically August 19, 2003.
http://www.siam.org/journals/simax/25-2/40939.html

†Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rcli@ms.uky.edu, qye@ms.uky.edu). The research of the first author was supported in part by the National Science Foundation CAREER award under grant CCR-9875201 and by the National Science Foundation under grant ACI-9721388. The research of the second author was supported in part by the National Science Foundation under grant CCR-0098133.

original (1.1) but relates to it by a shifting transformation.) For this reason, we shall focus in this paper on monic QEPs.

A related problem is the approximation of the transfer function

$$f(s) = c^*(s^2 I_n - sA - B)^{-1} b,$$

which arises in a single input single output system as governed by a second order initial value problem.

For these problems, a typical approach is to reduce them to an equivalent linear problem for the $2n \times 2n$ matrix [13],

$$A_{\text{LIN}} = \begin{pmatrix} 0 & I \\ B & A \end{pmatrix},$$

to which well-established methods can be applied (e.g., ARPACK [19]). This is called *linearization*. For the eigenvalue problem or the model reduction problem, one can use the Lanczos or the Arnoldi algorithm to produce a small projection of A_{LIN} on a Krylov subspace, which is then used to approximate A_{LIN}. This, however, increases the computational complexity by doubling the problem size. Furthermore, the projection of A_{LIN} is usually not a linearization of any QEP and thus loses its intrinsic physical connection to the problem that it approximates. As a result, for example, certain spectral properties of the original problem are not preserved in the projection and the approximations so obtained may not possess certain desirable properties such as the Galerkin condition. For the model reduction problem, the reduced model that is obtained by applying the Arnoldi or the Lanczos process to the linearization problem A_{LIN} cannot be synthesized with a physical model of QEP [2].

It is thus desirable to approximate a large scale QEP with another QEP of smaller size. The objective of this paper is to extend the *standard Arnoldi process* (and the *standard Lanczos process*) to cover matrix polynomials without going through any linearization. Namely, we develop a Krylov-type projection process applied simultaneously to A and B so as to obtain a projected lower-dimensional matrix polynomial to approximate the original one. With two matrices involved, the projections will no longer be in the upper Hessenberg (or tridiagonal) form, but rather a lower banded form with a growing lower bandwidth as the process progresses. However, in the case when some combination of the coefficient matrices A and B is of low rank, the projection matrix simplifies to a banded form and the algorithm becomes more efficient. We note that several other methods [20, 25] have been developed that do not rely on the linearization processes (see also [3, 30]).

As an application, we shall study the following quadratically constrained least squares problem

$$\min_{\|x\|_2 = \delta} \|Cx - b\|_2, \tag{1.3}$$

which arises, for example, in the regularization solution of discretized ill-posed problem (see [15, 16] and [23]), where all numbers are real, C is $m \times n$, and x and b are vectors of dimensions n and m, respectively. It can be formulated as the constrained minimization problem

$$\min_{x^T x = \delta^2} x^T H x - 2g^T x,$$

where $H = C^T C$, $g = C^T b$, and C^T is the transpose of C. A slightly more general form that uses the inequality constraint $x^T x \le \delta^2$ is called a trust region subproblem

(see [23], for example). We note that the problem with the inequality constraint will be more general (i.e., it will have no solution satisfying the equality constraint) only when H is invertible and $x = H^{-1}g$ (the solution to the unconstrained problem) lies in the interior of the constraint region [21]. To solve the above constrained minimization problem, several factorization-based methods have been developed [9, 11, 12, 21, 26], which typically apply to small or moderate size problems. For large problems, however, iterative methods are usually considered; see [4, 5, 6, 16, 23, 24, 27] for various methods developed.

In [11], Gander, Golub, and von Matt show that the above minimization problem can be transformed to the QEP

$$(\lambda^2 I - 2\lambda H + H^2 - \delta^{-2} g g^T) y = 0.$$

With the structure of this eigenvalue problem, the Krylov-type method can be adapted to solve it efficiently. This turns out to be a very efficient approach for solving the above constrained minimization problem, and the process of the Krylov-type method itself has a regularization effect for discrete ill-posed problems (1.3). We shall discuss various theoretical and numerical issues concerning this approach.

The paper is organized as follows. We present the Arnoldi-type algorithm for the quadratic matrix polynomial in section 2 and then the low rank perturbed case in section 3. We study the constrained least squares problem via the Arnoldi-type algorithm in section 4. We present some numerical examples in section 5 to illustrate the efficiency of the algorithms, and we give our concluding remarks in section 6.

Notation. Throughout, $\|\cdot\|$ refers to the 2-norm, i.e., $\|v\|^2 = v^*v$. I_n is the $n \times n$ identity matrix or simply I whenever its dimension is clear from the context; e_j is its jth column. $\lambda(X)$ is the spectrum of X. We use MATLAB-like notation $X_{(i:j,k:\ell)}$ to denote the submatrix of X, consisting of the intersections of rows i to j and columns k to ℓ, and when $i : j$ is replaced by :, it means all rows, similarly for columns. We shall use generic notation `x` for a possibly nonzero scalar or vector and `X` for a possibly nonzero matrix.

2. Arnoldi-type process for monic quadratic matrix polynomials. We first develop an Arnoldi-type process for monic quadratic matrix polynomial $I\lambda^2 - A\lambda - B$. Our algorithm will be based on a simultaneous orthogonal reduction of A and B. For the sake of generality, we state all results in the field of complex numbers. However, when all numbers involved are real, the only changes needed to be made are to replace $\mathbb{C}$ by $\mathbb{R}$ and asterisk superscripts $\cdot^*$ by $\cdot^T$.

2.1. Decomposition theorem. Our proofs below rely on the ability to transform a vector to a scalar multiplier of e_1 by an orthogonal transformation. This can be realized by at least two ways: by a Householder transformation or by a sequence of Givens rotations [7, 14, 31].

LEMMA 2.1. *There is a unitary matrix $Q \in \mathbb{C}^{n\times n}$ with $Qe_1 = e_1$ such that*

$$Q^*AQ = H_a \equiv (h_{a;ij}), \quad Q^*BQ = H_b \equiv (h_{b;ij})^1$$

satisfy

$$h_{a;ij} = 0 \text{ for } i \geq 2j+1, \quad h_{b;ij} = 0 \text{ for } i \geq 2j+2.$$

[1] $h_{a;ij}$ denotes the (i,j) entry of H_a, but we shall also use $h_{a;i,j}$ to denote the same when i and j are not clearly separated.

Proof. Our proof is constructive. It goes as follows. Partition

$$A = \begin{array}{c} \\ 1 \\ n-1 \end{array} \begin{array}{c} \begin{array}{cc} 1 & n-1 \end{array} \\ \begin{pmatrix} a_{11} & \mathtt{x} \\ a_1 & \mathtt{X} \end{pmatrix} \end{array},$$

and then find a unitary $\widehat{Q}_{1a} \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $\widehat{Q}_{1a}^* a_1 = \alpha_1 e_1$. Let $Q_{1a} = \mathrm{diag}(1, \widehat{Q}_{1a})$. We have

$$Q_{1a}^* A Q_{1a} = \left(\begin{array}{c|c} a_{11} & \mathtt{x} \\ \hline \alpha_1 & \\ 0 & \mathtt{X} \end{array}\right), \quad Q_{1a}^* B Q_{1a} = \begin{array}{c} \\ 1 \\ 1 \\ n-2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & n-1 \end{array} \\ \begin{pmatrix} b_{11} & \mathtt{x} \\ b_{21} & \mathtt{x} \\ b_1 & \mathtt{X} \end{pmatrix} \end{array}.$$

Now find a unitary $\widehat{Q}_{1b} \in \mathbb{C}^{(n-2)\times(n-2)}$ such that $\widehat{Q}_{1b}^* b_1 = \beta_1 e_1$. Let $Q_{1b} = \mathrm{diag}(I_2, \widehat{Q}_{1b})$ and $Q_1 \stackrel{\mathrm{def}}{=} Q_{1a} Q_{1b}$. We have

$$Q_1^* A Q_1 = \left(\begin{array}{c|c} a_{11} & \mathtt{x} \\ \hline \alpha_1 & \\ 0 & \mathtt{X} \end{array}\right), \quad Q_1^* B Q_1 = \left(\begin{array}{c|c} b_{11} & \mathtt{x} \\ b_{21} & \mathtt{x} \\ \hline \beta_1 & \\ 0 & \mathtt{X} \end{array}\right).$$

This puts the first columns of A and B into the desired forms. Next we work on their second columns. Partition

$$Q_1^* A Q_1 = \begin{array}{c} \\ 1 \\ 1 \\ n-3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 1 & n-2 \end{array} \\ \begin{pmatrix} \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ 0 & a_{32} & \mathtt{x} \\ 0 & a_2 & \mathtt{X} \end{pmatrix} \end{array},$$

and then find a unitary $\widehat{Q}_{2a} \in \mathbb{C}^{(n-3)\times(n-3)}$ such that $\widehat{Q}_{2a}^* a_2 = \alpha_2 e_1$. Let $Q_{2a} = \mathrm{diag}(I_3, \widehat{Q}_{2a})$. We have

$$Q_{2a}^* Q_1^* A Q_1 Q_{2a} = \left(\begin{array}{c|c|c} \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ 0 & a_{32} & \mathtt{x} \\ \hline 0 & \alpha_2 & \\ 0 & 0 & \mathtt{X} \end{array}\right), \quad Q_{2a}^* Q_1^* B Q_1 Q_{2a} = \begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \\ n-4 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 1 & n-2 \end{array} \\ \begin{pmatrix} \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & b_{32} & \mathtt{x} \\ 0 & b_{42} & \mathtt{x} \\ 0 & b_2 & \mathtt{X} \end{pmatrix} \end{array}.$$

Now find a unitary $\widehat{Q}_{2b} \in \mathbb{C}^{(n-4)\times(n-4)}$ such that $\widehat{Q}_{2b}^* b_2 = \beta_2 e_1$. Let $Q_{2b} = \mathrm{diag}(I_4, \widehat{Q}_{2b})$ and $Q_2 \stackrel{\mathrm{def}}{=} Q_{2a} Q_{2b}$. We have

$$Q_2^* Q_1^* A Q_1 Q_2 = \left(\begin{array}{c|c|c} \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ 0 & a_{32} & \mathtt{x} \\ \hline 0 & \alpha_1 & \\ 0 & 0 & \mathtt{X} \end{array}\right), \quad Q_2^* Q_1^* B Q_1 Q_2 = \left(\begin{array}{c|c|c} \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & \mathtt{x} & \mathtt{x} \\ \mathtt{x} & b_{32} & \mathtt{x} \\ 0 & b_{42} & \mathtt{x} \\ \hline 0 & \beta_2 & \\ 0 & 0 & \mathtt{X} \end{array}\right).$$

By now the first two columns of A and B are put into the desired forms. The process proceeds in a similar fashion from here. At the end, the jth column of transformed A has $2j$ possible nonzero entries at the top, and the jth column of transformed B has $2j+1$ possible nonzero entries also at the top. Taking $Q = Q_1 Q_2 \cdots Q_k$ completes the reduction, where at most $k \le n/2$. It is easy to see $Qe_1 = e_1$. ∎

THEOREM 2.2. *Given $q_1 \in \mathbb{C}^n$ with $\|q_1\|_2 = 1$, there is a unitary matrix $Q \in \mathbb{C}^{n\times n}$ with $Qe_1 = q_1$ such that*

$$Q^*AQ = H_a \equiv (h_{a;ij}), \quad Q^*BQ = H_b \equiv (h_{b;ij}) \tag{2.1}$$

satisfy

$$h_{a;ij} = 0 \text{ for } i \ge 2j+1, \quad h_{b;ij} = 0 \text{ for } i \ge 2j+2. \tag{2.2}$$

Proof. Find a unitary $Q_0 \in \mathbb{C}^{n\times n}$ with $Q_0 e_1 = q_1$. Then apply Lemma 2.2 to $Q_0^*AQ_0$ and $Q_0^*BQ_0$ to get a unitary $\widehat{Q} \in \mathbb{C}^{n\times n}$ with $\widehat{Q}e_1 = e_1$ such that

$$\widehat{Q}^*(Q_0^*AQ_0)\widehat{Q} \equiv H_a, \quad \widehat{Q}^*(Q_0^*BQ_0)\widehat{Q} \equiv H_b$$

have the desired forms. Now letting $Q = Q_0\widehat{Q}$ completes the proof. ∎

2.2. Arnoldi-type process. Although the proofs for Lemma 2.1 and Theorem 2.2 are constructive, they are of little use when it comes to numerical computations with large and sparse A and B for which we can only afford to generate Q, H_a, and H_b partially. In what follows, we shall present an Arnoldi-type process to do so. Rewrite (2.1) to get

$$AQ = QH_a, \quad BQ = QH_b. \tag{2.3}$$

Inspecting the jth column, we see

$$Aq_j = \sum_{i=1}^{2j-1} q_i h_{a;ij} + q_{2j} h_{a;2j,j}, \tag{2.4}$$

$$Bq_j = \sum_{i=1}^{2j} q_i h_{b;ij} + q_{2j+1} h_{b;2j+1,j}. \tag{2.5}$$

Equation (2.4) and the orthogonality among $q_1, \ldots, q_{2j}$ yield

$$h_{a;ij} = q_i^* A q_j \quad \text{for } i \le 2j-1,$$

and then we have

$$h_{a;2j,j} = \left\| Aq_j - \sum_{i=1}^{2j-1} q_i h_{a;ij} \right\|_2,$$

$$q_{2j} = \left(Aq_j - \sum_{i=1}^{2j-1} q_i h_{a;ij} \right) / h_{a;2j,j},$$

where we assume also $h_{a;2j,j} \neq 0$. Similarly, (2.5) implies

$$h_{b;ij} = q_i^* B q_j \quad \text{for } i \le 2j,$$

and then

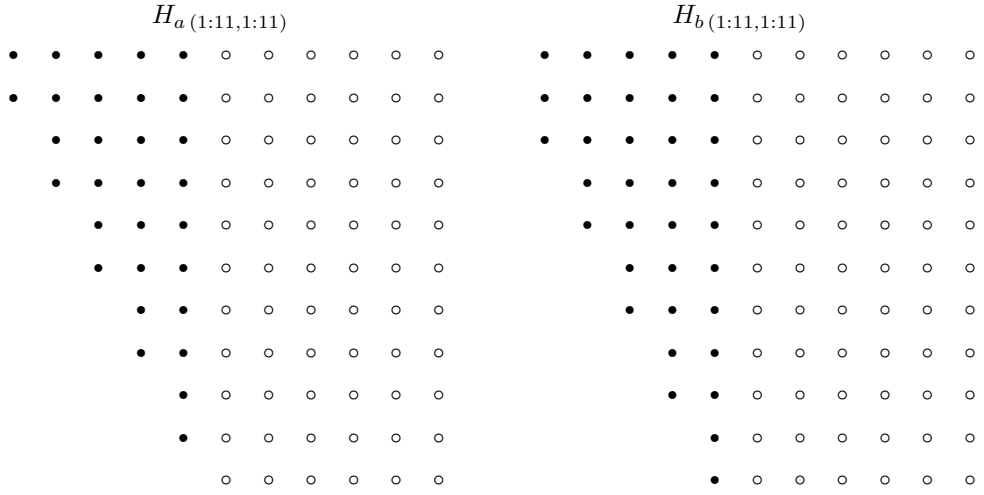
$$h_{b;2j+1,j} = \left\| Bq_j - \sum_{i=1}^{2j} q_i h_{b;ij} \right\|_2,$$
$$q_{2j+1} = \left(Bq_j - \sum_{i=1}^{2j} q_i h_{b;ij} \right) / h_{b;2j+1,j},$$

where we assume $h_{b;2j+1,j} \neq 0$. This leads to a process that constructs q_{2j}, q_{2j+1} from $q_1, q_2, \ldots, q_{2j-1}$. After k steps of construction, we obtain $q_1, q_2, \ldots, q_{2k+1}$ such that

$$AQ_{(:,1:k)} = Q_{(:,1:2k)} H_{a\,(1:2k,1:k)}, \tag{2.6}$$
$$BQ_{(:,1:k)} = Q_{(:,1:2k+1)} H_{b\,(1:2k+1,1:k)}. \tag{2.7}$$

The following figures in (2.8) show what the computed parts of H_a and H_b look like for $k = 5$, where the entries marked by *unfilled circles* are not computed yet.

(2.8)



With those computed entries, $H_{a\,(1:k,1:k)}$ and $H_{b\,(1:k,1:k)}$ provide the projections of A and B on $\mathrm{span}\{Q_{(:,1:k)}\}$. To fully utilize those unused computed entries, we can complete $H_{a\,(1:2k+1,1:2k+1)}$ and $H_{b\,(1:2k+1,1:2k+1)}$ by computing

$$h_{a;ij} = q_i^* A q_j, \quad h_{b;ij} = q_i^* B q_j$$

for $1 \le i \le 2k+1$ and $k+1 \le j \le 2k+1$ (i.e., the entries marked by *unfilled circles* above), which will then give the projections on a bigger subspace $\mathrm{span}\{Q_{(:,1:2k+1)}\}$. This requires computing Aq_j and Bq_j for $k+1 \le j \le 2j+1$. Therefore, to construct a $(2k+1) \times (2k+1)$ projection, we still need $2k+1$ matrix-vector multiplications by both A and B, but the number of vector operations required will be less.

So far, we have assumed that $h_{a;2j,j}$ and $h_{b;2j+1,j}$ are nonzero. When an $h_{a;2j,j}$ or $h_{b;2j+1,j}$ vanishes, no new q-vector can be generated, but we will show that the process can be continued. This is actually a welcome situation.

In the process, we apply A and B alternately on each vector in the sequence to construct new q-vectors. At any given point, let N be the number of q-vectors already

constructed. At the beginning of the process, $N = 1$ and there is only q_1, which has not yet been applied by A and B. For the first step ($j = 1$), we apply A to q_1, which may or may not generate a new q-vector, and if it does, $N \leftarrow N + 1$ (which is 2) and q_N is constructed. We then apply B to q_1, which again may or may not generate a new q-vector, and if it does, $N \leftarrow N + 1$ (which is either 2 or 3) and we have constructed a new q_N. Then, N q-vectors have been constructed, and if $N = 1$, the process can be terminated with $\text{span}\{q_1\}$ being invariant under both A and B. If $N \geq 2$, we then proceed to apply A and B to q_2 in the same way. In general, at the beginning of step j, among $q_1, \ldots, q_N$ that have been constructed, $q_1, \ldots, q_{j-1}$ have been applied by A and B. If $N = j - 1$, $\text{span}\{q_1, \ldots, q_N\}$ is invariant under both A and B and we can terminate the process. If $N \geq j$, we apply A to q_j (the next vector that has not been applied yet), and if a new vector is generated, $N \leftarrow N + 1$ and q_N is added to the q-vector list. We then apply B to q_j similarly. The process continues until $N = j - 1$, which must occur at $j = n + 1$, or a preselected k number of steps is reached. Thus, N may be much smaller than $2k + 1$. To fully utilize the information provided by the generated subspace $\text{span}\{Q(:, 1 : N)\}$, in our later numerical examples we compute the fully projected $H_{a(1:N,1:N)}$ and $H_{b(1:N,1:N)}$. Algorithm 2.1 summarizes our new process.

ALGORITHM 2.1 (Arnoldi-type process).

1. Given q_1 with $\|q_1\|_2 = 1$;
2. $N = 1$;
3. For $j = 1, 2, \ldots, k$ do
4. If $j > N$, **BREAK**;
5. $\hat{q} = Aq_j$;
6. For $i = 1, 2, \ldots N$ do
7. $h_{a;ij} = q_i^* \hat{q}$; $\hat{q} = \hat{q} - q_i h_{a;ij}$;
8. EndDo
9. $h_{a;N+1,j} = \|\hat{q}\|_2$;
10. If $h_{a;N+1,j} > 0$,
11. $N = N + 1$, $q_N = \hat{q}/h_{a;Nj}$;
12. EndIf
13. $\hat{q} = Bq_j$;
14. For $i = 1, 2, \ldots N$ do
15. $h_{b;ij} = q_i^* \hat{q}$; $\hat{q} = \hat{q} - q_i h_{b;ij}$;
16. EndDo
17. $h_{b;N+1,j} = \|\hat{q}\|_2$;
18. If $h_{b;N+1,j} > 0$,
19. $N = N + 1$, $q_N = \hat{q}/h_{a;N,j}$;
20. EndIf
21. EndDo

We point out that an appropriate tolerance must be used in practical implementations of line 10 and line 18 of Algorithm 2.1 as, e.g., $h_{a;N+1,j} > n\epsilon\|A\|$ and $h_{b;N+1,j} > n\epsilon\|B\|$, where ϵ is the machine roundoff unit. Define

$$\alpha_j = \text{ value of } N \text{ at line 12 at step } j, \tag{2.9}$$

$$\beta_j = \text{ value of } N \text{ at line 20 at step } j, \tag{2.10}$$

with $\alpha_0 = \beta_0 = 1$. Then,

$$Aq_j = \sum_{i=1}^{\alpha_j} h_{a;ij} q_i, \qquad Bq_j = \sum_{i=1}^{\beta_j} h_{b;ij} q_i.$$

Thus, upon completion of the above process, we have in general

$$AQ_{(:,1:k)} = Q_{(:,1:\alpha_k)}H_{a\,(1:\alpha_k,1:k)}, \tag{2.11}$$

$$BQ_{(:,1:k)} = Q_{(:,1:\beta_k)}H_{b\,(1:\beta_k,1:k)}, \tag{2.12}$$

unless the j-loop is forced to **BREAK** out at line 4, in which case we have obtained an invariant subspace of both A and B with

$$AQ_{(:,1:N)} = Q_{(:,1:N)}H_{a\,(1:N,1:N)}, \tag{2.13}$$

$$BQ_{(:,1:N)} = Q_{(:,1:N)}H_{b\,(1:N,1:N)}, \tag{2.14}$$

where N takes its value when the j-loop is terminated.

It is clear that

$$\beta_{j-1} \le \alpha_j \le \beta_{j-1} + 1 \text{ and } \alpha_j \le \beta_j \le \alpha_j + 1.$$

Furthermore, the nonzeros of the jth column of H_a (and H_b, resp.) are contained in the first α_j (β_j, resp.) entries only. α_j (and β_j as well) can increase at most by 2 at each step. So, the nonzero patterns in H_a and H_b are contained in those as described in (2.8).

We can use the reduced matrices $H_{a(1:N,1:N)}$ and $H_{b(1:N,1:N)}$ to approximate A and B. For example, we can use the eigenvalues of $\lambda^2 I - \lambda H_{a(1:N,1:N)} - H_{b(1:N,1:N)}$ to approximate those of the original quadratic problem. However, as the lower bandwidth of H_a and H_b grows very fast in general, the convergence is expected to be slow in general; see [17] for an analysis on the relation between the bandwidth and the speed of convergence. There are some special cases where the lower bandwidth can be bounded by a constant or grows at a much slower pace than in general. We shall discuss two such cases in the next two sections.

Similar to our derivation here, a (nonsymmetric) Lanczos-type process can be derived. The details will be presented in [17]. Finally, we remark that the way that the subspace $\text{span}\{q_1, \ldots, q_N\}$ are generated here bears some resemblance to the so-called generalized Krylov subspace in [33].

2.3. Hermitian case. When A and B are Hermitian, H_a and H_b will also be Hermitian. In that case, their upper triangular parts need not be computed and it is easy to prove that the recurrences are simplified to

$$h_{a;\alpha_j j} q_{\alpha_j} = Aq_j - \sum_{1 \le i < \alpha_j, \text{ and } \alpha_i \ge j} h_{a;ij} q_i,$$

$$h_{b;\beta_j j} q_{\beta_j} = Bq_j - \sum_{1 \le i < \beta_j, \text{ and } \beta_i \ge j} h_{b;ij} q_i.$$

We call the corresponding algorithm the symmetric Lanczos-type process. We omit the details here.

It is worth mentioning that the reduction process here also preserves other structural properties such as skew-symmetry or positive-definiteness in A or B.

3. Low rank case. In this section, we consider the case when some linear combination of A and B is of low rank, i.e.,

$$\zeta B + \xi A = E,$$

where E is a matrix of rank p and ζ and ξ are some, possibly unknown, scalars, at least one of which is nonzero. This includes the cases when one matrix is of low rank or is a low rank perturbation of the other matrix. We show that the Arnoldi-type process will be greatly simplified to yield a reduction with a lower bandwidth at most $p+1$ throughout the process. The resulting algorithm will be much more efficient.

Apply the Arnoldi-type process (Algorithm 2.1), we obtain at step k (see (2.11) and (2.12))

$$\begin{aligned} AQ_{(:,1:k)} &= Q_{(:,1:\alpha_k)}H_{a\,(1:\alpha_k,1:k)} = Q_{(:,1:\beta_k)}H_{a\,(1:\beta_k,1:k)}, \\ BQ_{(:,1:k)} &= Q_{(:,1:\beta_k)}H_{b\,(1:\beta_k,1:k)}. \end{aligned}$$

Therefore,

$$EQ_{(:,1:k)} = Q_{(:,1:\beta_k)}(\zeta H_{b\,(1:\beta_k,1:k)} + \xi H_{a\,(1:\beta_k,1:k)}).$$

This shows $\zeta H_{b\,(1:\beta_k,1:k)} + \xi H_{a\,(1:\beta_k,1:k)}$ has at most rank p. We consider now the case that $\zeta \neq 0$; the case that $\xi \neq 0$ follows similarly. From the structures of H_a and H_b, it can be seen that there are at most p columns in which H_b has more nonzeros than H_a, which is the time in the process that the lower bandwidth is increased. Thus, the lower bandwidth of H_a and H_b can grow at most p times throughout the process and is therefore bounded by $p+1$. To be more rigorous, let $i_1 < i_2 < \cdots < i_\ell$ be the index j between 1 and k such that $\beta_j = \alpha_j + 1$, in which case $h_{b;\beta_j,j} \neq 0$. For such $j \in \{i_1, i_2, \ldots, i_\ell\}$, $\beta_j > \alpha_j \geq \beta_{j-1}$ and therefore $h_{a;\beta_j,j} = 0$. Furthermore, $\beta_{i_1} < \beta_{i_2} < \cdots < \beta_{i_\ell}$. It follows from examining the $i_1, i_2, \ldots, i_\ell$th columns of $\zeta H_{b\,(1:\beta_k,1:k)} + \xi H_{a\,(1:\alpha_k,1:k)}$ that its rank is at least ℓ. Thus,

$$\ell \leq \text{ rank } (EQ_{(:,1:k)}) \leq p.$$

This demonstrates that there are at most p indexes j for which $\beta_j = \alpha_j + 1$. Hence there are at most p indexes j for which $\alpha_{j+1} = \alpha_j + 2$. For the same reason, there are at most p indexes j for which $\beta_{j+1} = \beta_j + 2$. Thus,

$$\alpha_j \leq j + 1 + p \text{ and } \beta_j \leq j + 1 + p.$$

So, $H_{a\,(1:\alpha_k,1:k)}$ and $H_{b\,(1:\beta_k,1:k)}$ are banded matrices with lower bandwidth at most $p+1$. We state this result as the following theorem.

THEOREM 3.1. *In Algorithm* 2.1*, if* $\zeta B + \xi A = E$ *(either* ζ *or* $\xi \neq 0$*) and* E *is a matrix of rank* p*, then* $\alpha_j \leq j + 1 + p$ *and* $\beta_j \leq j + 1 + p$*. In particular,* $H_{a\,(1:\alpha_k,1:k)}$ *and* $H_{b\,(1:\beta_k,1:k)}$ *are banded with lower bandwidth at most* $p+1$*.*

We note that it is not necessary to know the explicit combination $\zeta B + \xi A = E$ or the rank of E in advance. The algorithm will produce a reduction with the lower bandwidth limited by the rank of E. In practice, we may need to implement some reorthogonalization technique and use an appropriate tolerance in line 10 and line 18 of Algorithm 2.1. Then, the lower bandwidth will also be limited by the rank of E (see numerical examples in subsection 5.1.

3.1. Quadratic eigenvalue problems. The Arnoldi-type method can be used to find some eigenvalues and eigenvectors of the quadratic matrix polynomial $I\lambda^2 - A\lambda - B$. If Algorithm 2.1 produces $Q_{(:,1:k)}$, $H_{a\,(1:k,1:k)}$, and $H_{b\,(1:k,1:k)}$, let θ be an eigenvalue and u a right eigenvector of

$$I\lambda^2 - H_{a\,(1:k,1:k)}\lambda - H_{b\,(1:k,1:k)}. \tag{3.1}$$

We use (θ, y) as an approximate eigenvalue and eigenvector for the original problem, where

$$y = Q_{(:,1:k)}u. \tag{3.2}$$

θ will be called a Ritz value and y a Ritz vector. We note that the method works for general A and B, but the convergence may be slow [17]. For this reason, we shall consider the current case that $\zeta B + \xi A = E$ is of low rank.

In the next theorem, we present an a posteriori residual bound and show that the Ritz values and the Ritz vectors satisfy a Galerkin-type condition.

THEOREM 3.2. *Let $H_{a\,(1:k,1:k)}$ and $H_{b\,(1:k,1:k)}$ be obtained from k steps of the Arnoldi-type process (Algorithm 2.1), and let θ be an eigenvalue and u be a unit right eigenvector of* (3.1). *Then the Ritz value θ and the Ritz vector $y = Q_{(:,1:k)}u$ satisfy the following Galerkin-type condition:*

$$r \equiv (\theta^2 I - \theta A - B)y \perp \operatorname{span}\{Q_{(:,1:k)}\}. \tag{3.3}$$

Furthermore,

$$\|r\| \le (|\theta|\,\|A\| + \|B\|)\|u_{(k-p:k)}\|.$$

Proof. First, from (2.6) and (2.7), we have

$$\begin{aligned} AQ_{(:,1:k)} &= Q_{(:,1:k)}H_{a\,(1:k,1:k)} + Q_{(:,k+1:k+1+p)}H_{a\,(k+1:k+1+p,1:k)},\\ BQ_{(:,1:k)} &= Q_{(:,1:k)}H_{b\,(1:k,1:k)} + Q_{(:,k+1:k+1+p)}H_{b\,(k+1:k+1+p,1:k)}. \end{aligned}$$

Then

$$\begin{aligned} r &= (\theta^2 Q_{(:,1:k)} - \theta AQ_{(:,1:k)} - BQ_{(:,1:k)})u\\ &= Q_{(:,1:k)}(\theta^2 I - \theta H_{a\,(1:k,1:k)} - H_{b\,(1:k,1:k)})u\\ &\quad - \theta Q_{(:,k+1:k+1+p)}H_{a\,(k+1:k+1+p,1:k)}u - Q_{(:,k+1:k+1+p)}H_{b\,(k+1:k+1+p,1:k)}u\\ &= -Q_{(:,k+1:k+1+p)}(\theta H_{a\,(k+1:k+1+p,k-p:k)} + H_{b\,(k+1:k+1+p,k-p:k)})u_{(k-p:k)}. \end{aligned}$$

The orthogonality among q-vectors implies (3.3). Furthermore,

$$\|H_{a\,(k+1:k+1+p,k-p:k)}\| = \|Q^*_{(k+1:k+1+p,:)}AQ_{(:,k-p:k)}\| \le \|A\|.$$

Similarly, $\|H_{b\,(k+1:k+1+p,k-p:k)}\| \le \|B\|$. Taking the norm on r above, we obtain the bound. ∎

The theorem shows that if the last $p+1$ entries of an approximate eigenvector u become small, then the corresponding approximate eigenvalue will be a good approximation. This is usually the case for extreme eigenvalues of tridiagonal matrices produced by the standard Lanczos algorithm, and we observe that the banded matrices here appear to have a similar property.

We next derive an a priori convergence analysis similar to that of [32]. Here, we establish a relationship between the Ritz values and the eigenvalues of the original QEP through the linearizations. Let

$$L = \begin{pmatrix} 0 & I\\ H_b & H_a \end{pmatrix} \quad \text{and} \quad L_k = \begin{pmatrix} 0 & I\\ H_{b\,(1:k,1:k)} & H_{a\,(1:k,1:k)} \end{pmatrix},$$

where H_a and H_b are $n \times n$ as obtained by continuing the reduction process to the end. The following lemma can be verified by induction.

LEMMA 3.3. *Let S_ℓ and $\widetilde{S}_\ell$ be recursively defined by*

$$S_0 = 0, \quad \widetilde{S}_0 = 0,$$
$$S_1 = H_b, \quad \widetilde{S}_1 = H_{b\,(1:k,1:k)},$$
$$S_\ell = H_a S_{\ell-1} + H_b S_{\ell-2} \widetilde{S}_\ell = H_{a\,(1:k,1:k)} \widetilde{S}_{\ell-1} + H_{b\,(1:k,1:k)} \widetilde{S}_{\ell-2}$$

for $\ell \geq 2$. Then

$$L^\ell = \begin{pmatrix} S_{\ell-1} & \mathtt{X} \\ S_\ell & \mathtt{X} \end{pmatrix} \quad \textit{and} \quad L_k^\ell = \begin{pmatrix} \widetilde{S}_{\ell-1} & \mathtt{X} \\ \widetilde{S}_\ell & \mathtt{X} \end{pmatrix}.$$

As H_a and H_b are banded with lower bandwidth $p+1$, it is clear that S_ℓ and $\widetilde{S}_\ell$ are also banded but with lower bandwidth $\ell(p+1)$.

LEMMA 3.4. *Suppose $k \geq 3$, and let $m = \lfloor \frac{k}{p+1} \rfloor$ (the largest integer $\leq \frac{k}{p+1}$). Then*

1. $S_\ell e_1 = \begin{smallmatrix} k \\ n-k \end{smallmatrix} \begin{pmatrix} \widetilde{S}_\ell e_1 \\ 0 \end{pmatrix}$ *for $\ell = 0, 1, \ldots, m$,*
2. $S_{m+1} e_1 = \begin{smallmatrix} k \\ n-k \end{smallmatrix} \begin{pmatrix} \widetilde{S}_{m+1} e_1 \\ \mathtt{x} \end{pmatrix}$.

Proof. We shall prove claim 1 by induction on ℓ. It holds true for $\ell = 0, 1$. Suppose $m \geq \ell \geq 2$ and that the claim holds for $0, 1, \ldots, \ell-1$. Then $\ell(p+1) \leq k$ and

$$\begin{aligned} S_\ell e_1 &= H_a S_{\ell-1} e_1 + H_b S_{\ell-2} e_1 \\ &= H_a \begin{pmatrix} \widetilde{S}_{\ell-1} e_1 \\ 0 \end{pmatrix} + H_b \begin{pmatrix} \widetilde{S}_{\ell-2} e_1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} H_{a\,(1:k,1:k)} \widetilde{S}_{\ell-1} e_1 \\ 0 \end{pmatrix} + \begin{pmatrix} H_{b\,(1:k,1:k)} \widetilde{S}_{\ell-2} e_1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \widetilde{S}_\ell e_1 \\ 0 \end{pmatrix}, \end{aligned}$$

where we note that $\widetilde{S}_{\ell-1} e_1$ and $\widetilde{S}_{\ell-2} e_1$ have at most the first $(\ell-1)(p+1)$ entries nonzero and H_a and H_b have lower bandwidth $p+1$. Claim 1 is therefore proved. With claim 1 proved, setting $\ell = m+1$ in the above equations leads to claim 2. ∎

It follows from the above lemma that $e_1^* S_\ell e_1 = e_1^* \widetilde{S}_\ell e_1$ for $\ell = 0, 1, \ldots, m+1$ ($m = \lfloor \frac{k}{p+1} \rfloor$). (Recall that e_1 is the first column of I of appropriate dimension.) Then,

$$e_1^* L^{\ell+1} e_1 = e_1^* L_k^{\ell+1} e_1.$$

Therefore, for any polynomial f of degree $m+2$,

$$e_1^* f(L) e_1 = e_1^* f(L_k) e_1. \tag{3.4}$$

We now derive from this equation some relations between the eigenvalues of L and L_k. For the sake of simplicity, we assume that L and L_k are diagonalizable and write

$$L_k = U^* \Theta V \quad \text{and} \quad L = X^* \Lambda Y, \tag{3.5}$$

where Θ =diag$(\theta_1, \ldots, \theta_{2k})$, $U^*V = I$, and Λ =diag$(\lambda_1, \ldots, \lambda_{2n})$, $X^*Y = I$. Write $U = (u_{ij}), V = (v_{ij})$, $X = (x_{ij})$, and $Y = (y_{ij})$. Substituting (3.5) into (3.4), we obtain

$$e_1^* X^* f(\Lambda) Y e_1 = e_1^* U^* f(\Theta) V e_1.$$

Thus

$$\sum_{i=1}^{2n} f(\lambda_i)\bar{x}_{i1}y_{i1} = \sum_{i=1}^{2k} f(\theta_i)\bar{u}_{i1}v_{i1}.$$

Without loss of generality, we consider approximation of λ_1 and assume that $|\lambda_1-\theta_1| = \min_j |\lambda_1-\theta_j|$. Then, for any polynomial p of degree $m+1$, we use $f(t) = (t-\theta_1)p(t)$ in the above and obtain

$$\lambda_1 - \theta_1 = \frac{1}{p(\lambda_1)\bar{x}_{11}y_{11}} \left[-\sum_{i=2}^{2n} (\lambda_i-\theta_1)p(\lambda_i)\bar{x}_{i1}y_{i1} + \sum_{i=2}^{2k} (\theta_i-\theta_1)p(\theta_i)\bar{u}_{i1}v_{i1} \right].$$

Bounding $p(\lambda_i)$, $p(\theta_i)$ by their maximum, we obtain

$$|\lambda_1-\theta_1| \le \frac{\max_{i\ne 1}\{|p(\lambda_i)|, |p(\theta_i)|\}}{|p(\lambda_1)|} \frac{\sum_{i=2}^{2n} |(\lambda_i-\theta_1)\bar{x}_{i1}y_{i1}| + \sum_{i=2}^{2k} |(\theta_i-\theta_1)\bar{u}_{i1}v_{i1}|}{|\bar{x}_{11}y_{11}|},$$

which leads to the following theorem.

THEOREM 3.5. *Let* $|\lambda_1-\theta_1| = \min_j |\lambda_1-\theta_j|$. *Then we have*

$$|\lambda_1-\theta_1| \le K\epsilon_{m+1} \frac{\sqrt{\sum_{i\ne 1}(|x_{i1}|^2+|u_{i1}|^2)}}{|x_{11}|} \cdot \frac{\sqrt{\sum_{i\ne 1}(|y_{i1}|^2+|v_{i1}|^2)}}{|y_{11}|},$$

where

$$\epsilon_\ell = \min_{\deg p=\ell, p(\lambda_1)=1} \max_{i\ne 1} \{|p(\lambda_i)|, |p(\theta_i)|\},$$

$m = \lfloor \frac{k}{p+1} \rfloor$, *and* $K = \max_{i\ne 1}\{|\lambda_i-\theta_1|; |\theta_i-\theta_1|\}$.

ϵ_{m+1} is the dominating factor in the bound and can be bounded with the Chebyshev polynomials under some assumptions of the eigenvalue distribution (see [29, p. 191] for details). Essentially, if λ_1 and θ_1 are well separated from the other λ_i and θ_i, then ϵ_{m+1} can be made small and the bound shows that a good approximation of λ_1 is expected. The last two factors in the bound are related to the angle between q_1 and the right and left eigenvectors corresponding to λ_1 and show the dependence of convergence on the initial vector.

3.2. Shift-and-invert transform. The Arnoldi-type algorithm is often combined with a shift-and-invert transformation to accelerate convergence [8]. For example, to compute the eigenvalues near λ_0, a transformation of the form $\mu = (\lambda-\lambda_0)^{-1}$ is usually used, but this would destroy the low rank perturbation property. It turns out that the transformation

$$1/\lambda = 1/\mu + 1/\lambda_0 \tag{3.6}$$

also maps the eigenvalues λ close to λ_0 to large and well-separated μ, and more importantly it preserves the low rank perturbation property. Indeed,

$$\begin{aligned} \lambda^2 I - \lambda A - B &= \lambda^2 (I - (1/\lambda)A - (1/\lambda)^2 B) \\ &= \lambda^2 \left[I - (1/\mu + 1/\lambda_0)A - (1/\mu+1/\lambda_0)^2 B \right] \\ &= \lambda^2 \left[I - (1/\lambda_0)A - (1/\lambda_0)^2 B - (1/\mu)(A + 2/\lambda_0 B) - (1/\mu)^2 B \right] \\ &= (\lambda/\mu)^2 M(\mu^2 I - \mu\hat{A} - \hat{B}), \end{aligned} \tag{3.7}$$

where

$$M = I - (1/\lambda_0)A - (1/\lambda_0)^2 B,$$
$$\hat{A} = M^{-1}(A + 2/\lambda_0 B),$$
$$\hat{B} = M^{-1}B.$$

For $\zeta B + \xi A = E$, we have

$$(\zeta - 2\xi/\lambda_0)\hat{B} + \xi\hat{A} = M^{-1}E,$$

which is still of low rank.

4. A constrained least squares problem. Let[2] $H \in \mathbb{R}^{n\times n}$ be symmetric and $g \in \mathbb{R}^n$. We consider the constrained minimization problem

$$\min_{x^T x = \delta^2} x^T H x - 2g^T x. \tag{4.1}$$

As pointed out in the introduction, this problem arises in the regularization of discretized ill-posed problems and trust-region subproblems. The Lagrangian equations for (4.1) are

$$Hx - g = \lambda x, \tag{4.2}$$
$$x^T x = \delta^2, \tag{4.3}$$

where λ is the Lagrangian multiplier. It is shown in Gander [10] that the solution (λ, x) to the Lagrange equation (4.2), (4.3) with the smallest λ solves (4.1). Furthermore, it is shown by Gander, Golub, and von Matt [11] that (4.2) and (4.3) can be reduced to the QEP

$$(\lambda^2 I - 2\lambda H + H^2 - \delta^{-2} g g^T)y = 0. \tag{4.4}$$

Specifically, it is proved that if (λ, x) solves (4.2) and (4.3), then λ is an eigenvalue of (4.4). Conversely, for an eigenpair (λ, y) of (4.4), if $\lambda \notin \lambda(H)$, then (λ, x) with $x = (H - \lambda I)^{-1}g$ solves (4.2) and (4.3); if $\lambda \in \lambda(H)$, then λ is a solution to (4.2) and (4.3) if and only if $x = (H - \lambda I)^{\dagger} g$ satisfies $(H - \lambda I)x = g$ and $x^T x \le \delta^2$, where $(H - \lambda I)^{\dagger}$ is the pseudo-inverse [7, 14].

For small problems, it appears that the solution through (4.4) is not competitive when compared with other direct methods; see [11]. For large scale problems, however, we will show that (4.4) can be solved efficiently by the Arnoldi-type process, and thus it offers a very promising approach to solving (4.1).

In the setting of large scale problems, the eigenvalue problem (4.4) is usually solved only approximately by an iterative method that reduces the residual of the approximate solution to certain threshold. Here we first consider when an approximate solution of (4.4) leads to an approximate solution of (4.2) and (4.3). The following theorem is an inexact version of the result presented in [11] and reveals an interesting numerical issue associated with using (4.4).

THEOREM 4.1. *Let (θ, y) with $\|y\| = 1$ be an approximate eigenpair of* (4.4) *and let*

$$r = (\theta^2 I - 2\theta H + H^2 - \delta^{-2} g g^T)y. \tag{4.5}$$

Assume $g^T y \neq 0$.

[2]We restrict our discussion in this section to real matrices so as to be consistent with existing related literature. Obviously the section can be extended to cover the complex case in which H is Hermitian.

1. *Let* $z = \frac{\delta^2}{g^T y}(H - \theta I)y$. *We have*

$$(H - \theta I)z - g = \frac{\delta^2}{g^T y} r, \tag{4.6}$$

$$\frac{z^T z - \delta^2}{\delta^2} = \frac{\delta^2}{(g^T y)^2} y^T r. \tag{4.7}$$

   In particular, if $y^T r = 0$ *(which is the case if* (θ, y) *is obtained from the Arnoldi-type process), then* $z^T z - \delta^2 = 0$ *by* (4.7).
2. *If* $\theta \notin \lambda(H)$, *let* $\hat{z} = (H - \theta I)^{-1} g$. *We have*

$$(H - \theta I)\hat{z} - g = 0,$$
$$\frac{\hat{z}^T \hat{z} - \delta^2}{\delta^2} = \frac{\hat{z}^T (H - \theta I)^{-1} r}{g^T y}.$$

Proof. From (4.5), it follows that

$$(H - \theta I)^2 y = \delta^{-2} g g^T y + r,$$

which implies $(H - \theta I)z - g = \frac{\delta^2}{g^T y} r$. Using the definition of z, we have

$$\begin{aligned} z^T z &= \frac{\delta^4}{(g^T y)^2} y^T (H - \theta I)^2 y \\ &= \frac{\delta^4}{(g^T y)^2} y^T (\delta^{-2} g g^T y + r) \\ &= \delta^2 + \frac{\delta^4}{(g^T y)^2} y^T r. \end{aligned}$$

This proves (4.7). For part 2, $(H - \theta I)\hat{z} - g = 0$ follows directly from the definition of $\hat{z}$. Furthermore, from (4.5), $\frac{g^T y}{\delta^2}(H - \theta I)^{-2} g = y - (H - \theta I)^{-2} r$. Thus

$$\begin{aligned} \hat{z}^T \hat{z} &= g^T (H - \theta I)^{-2} g \\ &= \frac{\delta^2}{g^T y}(g^T y - g^T (H - \theta I)^{-2} r) \\ &= \delta^2 - \frac{\delta^2}{g^T y} \hat{z} (H - \theta I)^{-1} r, \end{aligned}$$

which leads to the second equation. ∎

Once an approximation to the smallest eigenpair is found, then either $x \approx z$ or $x \approx \hat{z}$ gives an approximate solution to (4.1). However, $\hat{z}$ requires solving $(H - \theta I)\hat{z} = g$, and the constraint error $(\hat{z}^T \hat{z} - \delta^2)/\delta^2$ can be large. On the other hand, taking $x \approx z$ is more straightforward. We will consider $z = \frac{\delta^2}{g^T y}(H - \theta I)y$ only.

The theorem illustrates a potential difficulty to construct a solution of (4.2) and (4.3) from an approximate eigenpair. The error for the constraint equation (4.3) is inversely proportional to $(g^T y)^2$ and, in discretized ill-posed problems, $g^T y$ is typically very small. Thus, an approximate eigenpair with small residual r does not necessarily lead to a good approximate solution to the Lagrange equations. Fortunately, the theorem also shows that this problem is eliminated if we have $y^T r = 0$. For (θ, y) as obtained from the Arnoldi-type algorithm, we have $y^T r = 0$ since $r \perp \text{span}\{Q_{(:,1:k)}\}$

and $y \in \text{span}\{Q_{(:,1:k)}\}$ (see Theorem 3.2). Hence z will always satisfy the constraint, but this is valid in theory only. In practice, we have only near orthogonality between y and r, but this orthogonality can be further improved by recomputing θ to enforce orthogonality $y^T r = 0$. Namely, if (θ, y) is an approximate eigenpair, we recompute θ as the Rayleigh quotient by solving

$$\theta^2 I - 2\theta y^T H y + y^T (H^2 - \delta^{-2} g g^T) y = 0. \tag{4.8}$$

This will lead to much improved orthogonality $y^T r = 0$ and will hence keep the error in the constraint equation small (see examples in section 5.2). The importance of the orthogonality $y^T r = 0$ can be highlighted by considering the QR algorithm. If (θ, y) is obtained from the QR algorithm, we know $r \approx \mathcal{O}(\epsilon)$ but cannot say anything about the direction of r, which implies $y^T r$ is of order ϵ only. Using (θ, y) directly to compute z, the error in the constraint equation (4.7) can be very large, even when $g^T y$ is modestly small (e.g., of order $\sqrt{\epsilon}$); see [11] for some numerical results. This problem can be corrected by recomputing θ through (4.8) to enforce the orthogonality.

The theorem is valid only when $g^T y \neq 0$. If $g^T y = 0$ and y is an exact eigenvector (i.e., $r = 0$), then $(H - \theta I)^2 y = 0$. Since $H - \theta I$ is real symmetric, we have $(H - \theta I) y = 0$, and hence θ is an eigenvalue of H with y a corresponding eigenvector. In this case, θ is a solution to the Lagrange equation if and only if $x = (H - \theta I)^{\dagger} g$ satisfies $(H - \theta I)x = g$ and $x^T x \leq \delta^2$. This is indeed an extreme situation called *the hard case* of (4.1) (see [23]). In the hard case, the solution does not depend continuously on g.

We now show that the QEP (4.4) can be efficiently solved by the Arnoldi-type algorithm. While theoretically we can apply the Arnoldi-type process directly to H and $H^2 - \delta^{-2} g g^T$, it is easier to do it indirectly by using Algorithm 2.1 on H and $g g^T$ first, from which a reduction of $H^2 - \delta^{-2} g g^T$ can be derived.

Let Algorithm 2.1 (or the symmetric version) be applied to $A = H$ and $B = g g^T$ for k steps; we obtain

$$\begin{aligned} A Q_{(:,1:k)} &= Q_{(:,1:k+2)} H_{a\,(1:k+2,1:k)}, \\ B Q_{(:,1:k)} &= Q_{(:,1:k+2)} H_{b\,(1:k+2,1:k)}. \end{aligned}$$

Since A and B are symmetric and B is of rank 1, H_a and H_b are symmetric banded with bandwidth 2. Indeed, $B q_1 - q_1 h_{b;11} - q_2 h_{b;21} = q_3 h_{b;31}$, i.e., $g(g^T q_1) = q_1 h_{b;11} + q_2 h_{b;21} + q_3 h_{b;31} = Q_{(:,1:3)} H_{b\,(1:3,1)}$. Then

$$\begin{aligned} B Q_{(:,1:k)} = g g^T Q_{(:,1:k)} &= \frac{1}{(g^T q_1)^2} Q_{(:,1:3)} (H_{b\,(1:3,1)} H_{b\,(1:3,1)}^T) Q_{(:,1:3)}^T Q_{(:,1:k)} \\ &= \frac{1}{(g^T q_1)^2} Q_{(:,1:3)} [H_{b\,(1:3,1)} H_{b\,(1:3,1)}^T, 0]. \end{aligned}$$

Thus,

$$H_{b\,(1:k+2,1:k)} = \begin{pmatrix} H_{b\,(1:3,1)} H_{b\,(1:3,1)}^T / (g^T q_1)^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

In fact, with H_b as defined above, the algorithm can be implemented with the B part (line 13 to line 20 of Algorithm 2.1) omitted after $j > 2$. Furthermore,

$$A^2 Q_{(:,1:k)} = A Q_{(:,1:k+2)} H_{a\,(1:k+2,1:k)} = Q_{(:,1:k+4)} H_{a\,(1:k+4,1:k+2)} H_{a\,(1:k+2,1:k)}.$$

Thus

$$(\delta^{-2} g g^T - H^2) Q_{(:,1:k)} = Q_{(:,1:k+4)} \hat{H}_{b\,(1:k+4,1:k)},$$

where $\hat{H}_b = \delta^{-2} H_b - H_a^2$. Clearly $\hat{H}_b$ has a bandwidth 4. We can now approximate (4.1) by solving the reduced problem

$$\theta^2 I - 2\theta H_{a\,(1:k,1:k)} - \hat{H}_{b\,(1:k,1:k)},$$

where

$$\hat{H}_{b\,(1:k,1:k)} = \delta^{-2} H_{b\,(1:k,1:k)} - H^T_{a\,(1:k+2,1:k)} H_{a\,(1:k+2,1:k)}.$$

Noting that A and B are symmetric, we can use the symmetric version of Algorithm 2.1 here. We observe that the approximate eigenpair (θ_k, y_k) as obtained from this algorithm still satisfies the Galerkin-type condition (3.3). We summarize the process into the following algorithm for solving (4.1).

ALGORITHM 4.1 (Lanczos-type process for constrained minimization problem).

1. Input: H, g, and q_1 with $\|q_1\|_2 = 1$;
2. $\hat{q} = Hq_1$;
3. $h_{a;11} = q_1^T \hat{q}$; $\hat{q} = \hat{q} - q_1 h_{a;11}$;
4. $h_{a;21} = \|\hat{q}\|_2$; $q_2 = \hat{q}/h_{a;21}$;
5. $h_{b;11} = (g^T q_1)^2$; $h_{b;21} = q_2^T g(g^T q_1)$;
6. $\hat{q} = g(g^T q_1) - q_1 h_{b;11} - q_2 h_{b;21}$;
7. $h_{b;31} = \|\hat{q}\|_2$; $q_3 = \hat{q}/h_{b;31}$;
8. $N = 3$
9. For $j = 2, \ldots, k$
10. $\quad \hat{q} = Hq_j$;
11. $\quad$ For $i = \max\{1, j-2\} : N$ do
12. $\quad\quad h_{a;ij} = q_i^T \hat{q}$; $\hat{q} = \hat{q} - q_i h_{a;ij}$;
13. $\quad$ EndDo
14. $\quad h_{a;N+1,j} = \|\hat{q}\|_2$;
15. $\quad$ If $h_{a;N+1,j} > 0$,
16. $\quad\quad N = N + 1$, $q_N = \hat{q}/h_{a;Nj}$;
17. $\quad$ EndIf;
18. $\quad$ If $N \le j$, break;
19. EndDo
20. $H_{b\,(1:k,1:k)} = \begin{pmatrix} H_{b\,(1:3,1)} H^T_{b\,(1:3,1)}/(g^T q_1)^2 & 0 \\ 0 & 0 \end{pmatrix}$;
21. $\hat{H}_{b\,(1:k,1:k)} = \delta^{-2} H_{b\,(1:k,1:k)} - H^T_{a\,(1:k+2,1:k)} H_{a\,(1:k+2,1:k)}$;
22. Find the smallest real eigenpair (θ_k, v_k) of $I\theta^2 - 2H_{a\,(1:k,1:k)}\theta - \hat{H}_{b\,(1:k,1:k)}$;
23. $y_k = Q_{(:,1:k)} v_k$;
24. θ = the root of $\theta^2 I - 2\theta y_k^T H y_k + \|Hy_k\|^2 - \delta^{-2}(y_k^T g)^2 = 0$ that is closer to θ_k;
25. $z_k = \frac{\delta^2}{g^T y_k}(H - \theta I) y_k$.

In the algorithm, the iteration number k can be determined by requiring that the solution z_k satisfies, for example, $\|Hz_k - \theta_k z_k - g\|/\|g\| \le \texttt{tol}_1$ and $|z_k^T z_k - \delta^2|/\delta^2 \le \texttt{tol}_2$ for some given tolerances $\texttt{tol}_1$ and $\texttt{tol}_2$.

Finally, we note that with its special structure, the QEP (4.4) can also be solved by using the standard Lanczos algorithm; namely, we can apply k step of the Lanczos algorithm to an initial vector q_1 to produces $Q_{k+1} = [q_1, q_2, \ldots, q_k, q_{k+1}]$ with orthonormal columns such that

$$AQ_k = Q_{k+1} T_{(1:k+1,1:k)},$$

where T is $n \times n$ tridiagonal. Then, if $h = Q_k^T g$, we can approximate (4.4) by its projection $Q_k^T(\lambda^2 I - 2\lambda H + H^2 - \delta^{-2} g g^T)Q_k$, which is

$$\lambda^2 I - 2\lambda T_{(1:k,1:k)} + T^T_{(1:k+1,1:k)} T_{(1:k+1,1:k)} - \delta^{-2} h h^T. \tag{4.9}$$

In this case, the choice of q_1 plays an important role, as we need the Krylov subspace $\mathrm{span}\{Q_k\}$ to approximate well both g and the eigenvector sought. If q_1 is chosen to be a random vector, g may not be well approximated by its projection onto $\mathrm{span}\{Q_k\}$. On the other hand, if $q_1 = g/\|g\|$, then $g \in \mathrm{span}\{Q_k\}$ but the eigenvector sought is not necessarily well approximated by $\mathrm{span}\{Q_k\}$. We note that the choice of g works out quite well compared to our process with a random q_1 for discrete ill-posed problems that we tested.

5. Numerical examples. In this section we shall present two sets of numerical examples. In the first set, we use random sparse matrices as generated by MATLAB. The second set is for the constrained least squares problems (4.1) as arising in the regularization solution of discretized ill-posed problems [23].

5.1. QEP with random matrices. We start by testing on QEP $\lambda^2 I - \lambda A - B$ with no relation between A and B assumed, where A and B are generated by MATLAB commands

$$\mathtt{n} = 500; \quad \mathtt{A} = \mathtt{sprandn}(\mathtt{n}, \mathtt{n}, 0.05); \quad \mathtt{B} = \mathtt{sprandn}(\mathtt{n}, \mathtt{n}, 0.05);$$

initial vector q_1 is a random vector. A direct application of Krylov-type methods to random matrices gives poor convergence results. Instead, we use a shift-and-invert transformation with the shift $\lambda_0 = -1.0 + 3i$, which gives a much more favorable spectral distribution. Then applying Algorithm 2.1 with $k = 8$ on the transformed problems as in (3.6) and (3.7), an approximate eigenvalue $\lambda_1 \approx -0.9549 + 2.8519i$ is computed. Figure 1 plots the normalized residual

$$\gamma_j \equiv \frac{\|(\lambda_j^2 I - A\lambda_j - B)x_j\|}{\max\{|\lambda_j|^2\|x_j\|, |\lambda_j|\|Ax_j\|, \|Bx_j\|\}} \tag{5.1}$$

for all eigenvalues obtained, where λ_j is a computed eigenvalue and x_j is a corresponding computed eigenvector. Notice that since both A and B are randomly generated and thus unrelated, every application of A or B on q-vectors produces new directions, and consequently $N = 2k + 1 = 17$ and there are 34 approximate eigenvalues.

Next we test Algorithm 2.1 on the low rank cases. The matrices A and B are generated as

```
n=500; A=sprandn(n,n,0.05);
X=randn(n,2); Y=randn(n,2); B=1.1*A+2.3*X*Y'.
```

Thus $-1.1\mathtt{A} + \mathtt{B} = 2.3\mathtt{XY}'$, of rank 2. But in running Algorithm 2.1, we do not assume knowing `X` and `Y`. Without shifting and with a random q_1 and $k = 30$, Algorithm 2.1 outputs $N = 33$ and $H_{a\,(1:N,1:N)}$ and $H_{b\,(1:N,1:N)}$. Figure 2 plots the residual errors for the 66 Ritz values obtained, where computed $\lambda_{51}, \lambda_{52} = -1.1345 \pm 0.0307i$ and $\lambda_{65}, \lambda_{66} = -1.0561 \pm 0.0168i$. The sparsity patterns $H_{a\,(1:N,1:N)}$ and $H_{b\,(1:N,1:N)}$ are displayed in Figure 3.

Now we apply the shift-and-invert transformation of (3.6), which will preserve the low rank perturbation property (see section 3.2). We take $\lambda_0 = -1.2 + i$ and apply Algorithm 2.1 with $k = 15$ and an random q_1 on the transformed problems as in (3.6) and (3.7). Figure 4 plots the residual errors of the computed approximate

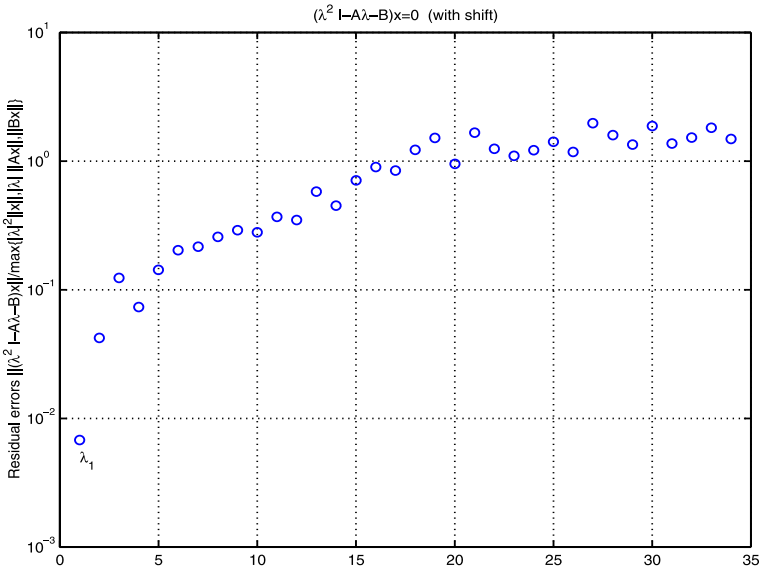


FIG. 1. *Residual errors of computed eigenvalues: A and B unrelated.*

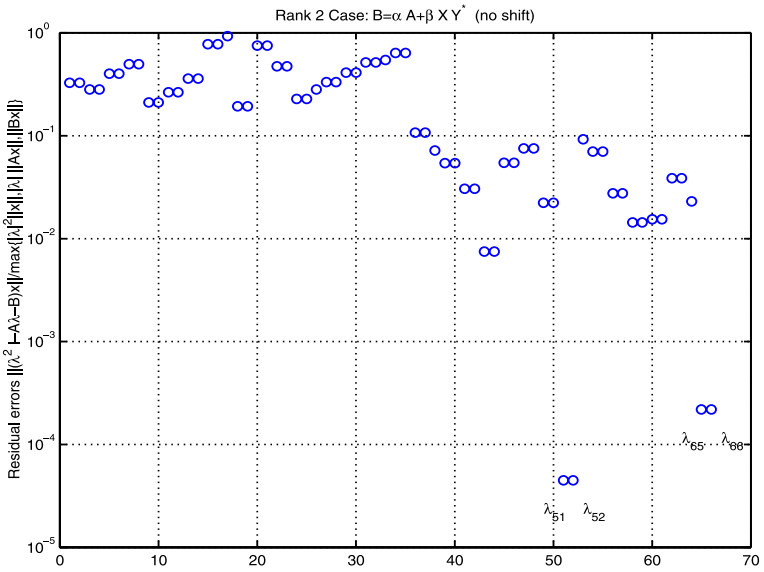


FIG. 2. *Residual errors of computed eigenvalues: A rank 2 case.*

eigenvalues, where computed $\lambda_1 = -1.1415 + 0.9082i$ and $\lambda_{35} = -1.1725 - 1.3274i$. With $N = 18$, the projections have the same sparsity structure as in Figure 3, while the convergence is clearly accelerated.

5.2. Constrained least squares problems. We now consider some constraint least squares testing problems (1.3) taken from the regularization tool of Hansen [15]. They are discretizations of some integral equations (see [15] for more detailed

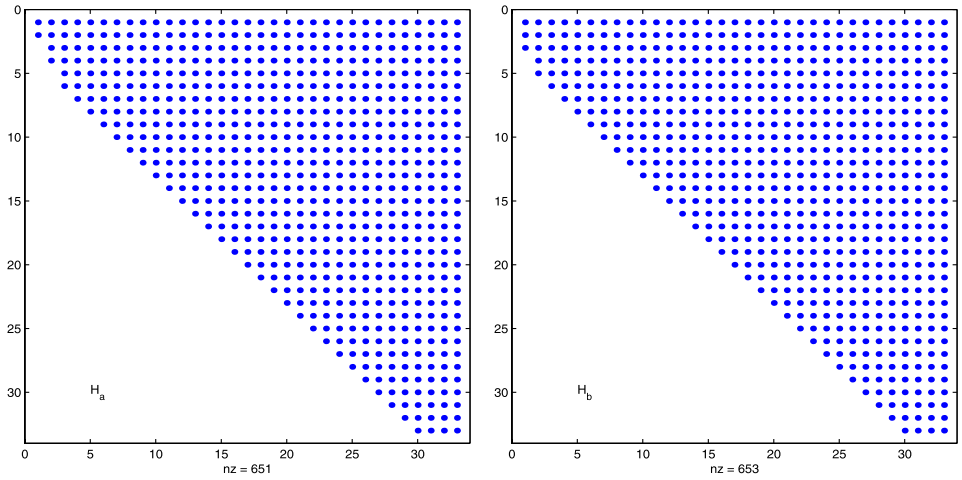


FIG. 3. *Sparsity patterns of H_a and H_b: A rank 2 case.*

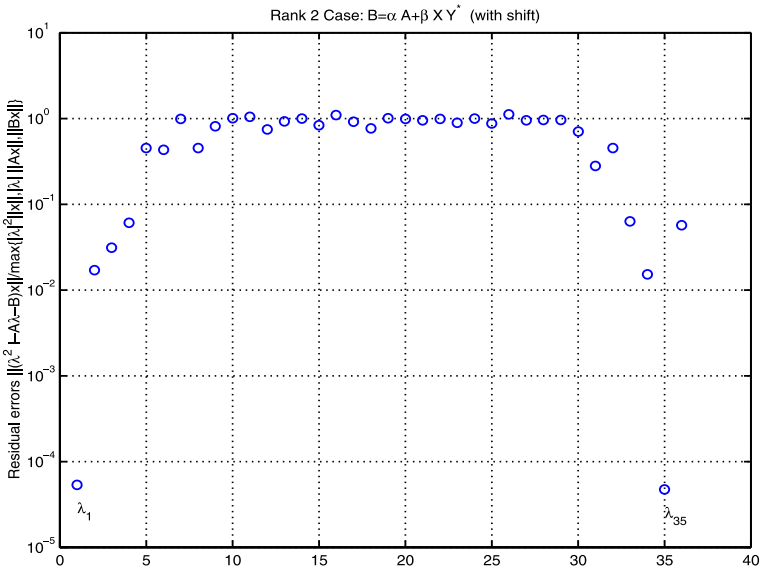


FIG. 4. *Residual errors of computed eigenvalues: A rank 2 case with shift.*

description of the matrices). In all test problems except `parallax` and `ursell`, a reference solution $x_{\rm IP}$ is provided by the routine and, in that case, we set $\delta = \|x_{\rm IP}\|$. We also set the dimension $n = 1000$ for all tests except for `blur` (image deblurring problem) for which $n = 32^2$ due to the problem's characteristic. Typically, the matrix H is either of low rank (with a rectangular C) or numerically of low rank (with a large number of tiny singular values). This appears to be one of the reasons for very fast convergence that we will see.

We first test the convergence of the eigenvalue with the smallest real part. Here we use a random vector as the initial vector and terminate the iteration when the

TABLE 1
QEP from constraint least squares problems (γ_k—normalized residual).

Problem	θ_k	γ_k	$\lvert\theta_k - \lambda_{\rm QR}\rvert$	k
`barrt`	$4.66188e-08$	$2.1e-13$	$4.8e-08$	4
`ill heat`	$7.47851e-08$	$3.0e-09$	$7.6e-08$	18
`well heat`	$1.19617e-08$	$9.9e-09$	$1.3e-08$	188
`blur`	$1.34996e-12$	$9.3e-09$	$1.3e-12$	347
`deriv2` (1)	$4.42695e-08$	$3.5e-09$	$4.5e-08$	8
`deriv2` (2)	$4.59196e-08$	$2.5e-09$	$4.6e-08$	8
`deriv2` (3)	$6.68204e-08$	$2.9e-09$	$6.7e-08$	7
`foxgood`	$2.22965e-09$	$5.2e-13$	$6.9e-09$	3
`parallax`	$-1.34982e-01$	$9.1e-10$	$1.3e-15$	10
`phillips`	$3.17750e-05$	$1.7e-09$	$5.0e-05$	11
`shaw`	$1.01446e-04$	$3.2e-09$	$1.0e-04$	6
`spikes`	$1.64716e-02$	$7.1e-09$	$1.6e-02$	10
`ursell`	$-2.42031e-01$	$1.0e-12$	$3.1e-16$	4
`wing`	$1.13499e-07$	$3.4e-11$	$1.1e-07$	3

normalized residual (5.1) satisfies $\gamma_k < 10^{-8}$. Table 1 lists the results obtained, where we include the computed Ritz value θ_k, the normalized residual γ_k, the errors $|\theta_k - \lambda_{\rm QR}|$ ($\lambda_{\rm QR}$ is the leftmost eigenvalue returned by the QR algorithm (`eig` of MATLAB) on $A_{\rm LIN}$), and the required number of iterations k. We note that for those problems where $Cx_{\rm IP} \approx b$ (cf. (1.3)), $x_{\rm IP}$ is a solution to (4.1) because it satisfies the constraint. Then, $Hx_{\rm IP} - g \approx 0$, and therefore the eigenvalue is nearly 0 for those problems.

In all problems, the residual falls below the given threshold within a small number of iterations. For the problems where the smallest eigenvalue is 0 or nearly 0, the absolute error of eigenvalue is approximately equal to θ and is approximately 10^{-5} or smaller except for the `spike` problem. For the other problems (`parallax` and `ursell`, in which $x_{\rm IP}$ is not given and $\delta = \|b\|$), eigenvalues are of $\mathcal{O}(1)$, and the absolute errors are then of $\mathcal{O}(10^{-15})$ as compared with the QR algorithm. For the `spike` problem, the large eigenvalue error is due to the fact that the norm of H^2 is so large ($\|H^2\|_1 \approx \mathcal{O}(10^{10})$) that the absolute residual $\|r_k\|$ is only reduced to $\mathcal{O}(1)$.

In Figure 5, we present the residual convergence history for the inverse heat problem (ill-conditioned `heat` with $\kappa = 1$). The solid line is for the normalized residual γ_k (5.1) and the dotted line for the error $|\theta_k - \lambda_{\rm QR}|$.

We also present a comparison among Algorithm 4.1, (4.9) with $q_1 = g$, and (4.9) with random q_1 that directly use the projection onto the Krylov subspace generated by H and q_1. Figure 6 compares convergence history of normalized residuals for the `heat` problem. It appears that with the choice $q_1 = g$, the direct approach (4.9) and Algorithm 4.1 have a very similar convergence characteristic, with the former converging a few steps faster and the latter being slightly more stable after the residual has converged to the level of machine precision. The random choice of q_1, on the other hand, can result in slower convergence, as expected.

We next test convergence of the approximate solution z_k to the Lagrange equations. Here we terminate the iteration whenever both the relative residual and the constraint error are below 10^{-6}, i.e., when

$$\zeta_k \equiv \frac{\|Hz_k - \theta_k z_k - g\|}{\|g\|} < 10^{-6} \quad \text{and} \quad \eta_k \equiv \frac{z_k^T z_k - \delta^2}{\delta^2} < 10^{-6}.$$

In addition to the residual ζ_k, the constraint error η_k, the relative error $\|z_k - x_{\rm IP}\|/\|x_{\rm IP}\|$

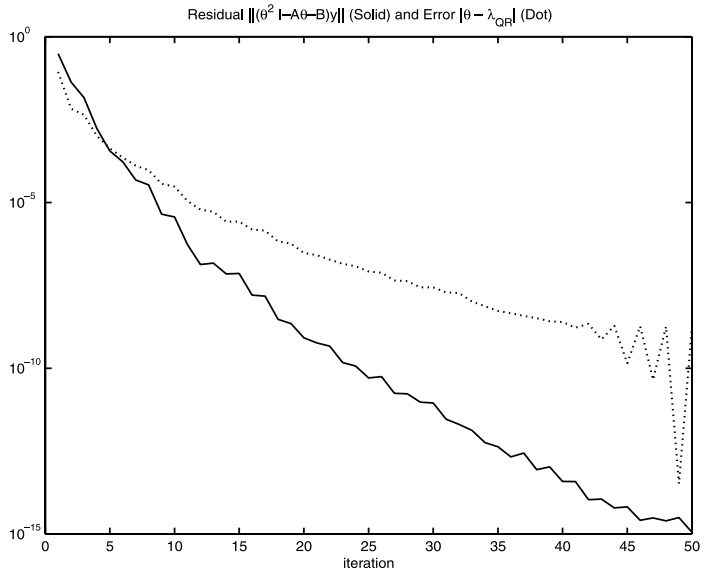


FIG. 5. *Eigenvalue convergence history for the* `heat` *problem with* $\kappa = 1$.

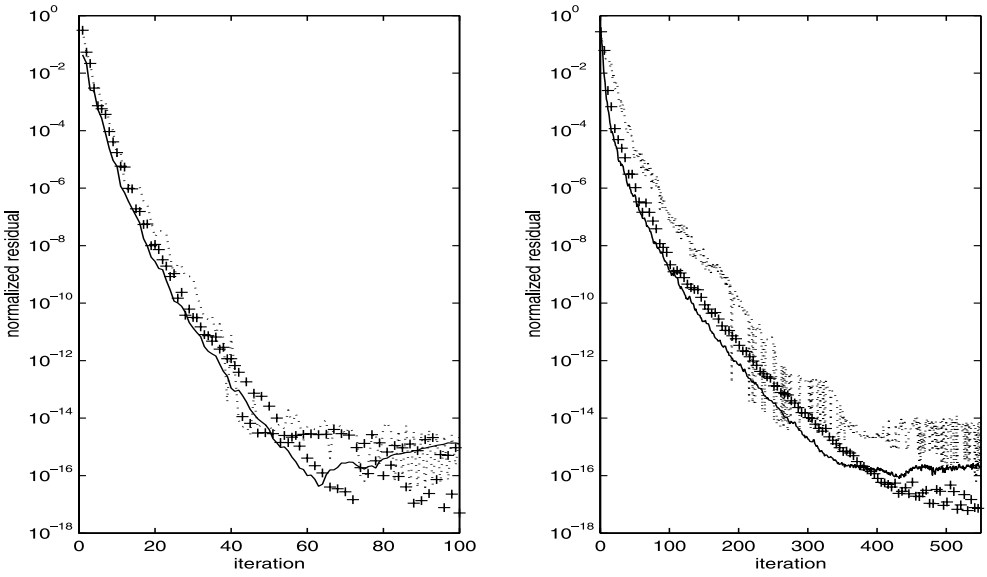


FIG. 6. *Comparisons on the* `heat` *problem.* Left: $\kappa = 1$. Right: $\kappa = 3$. *Here +–line: Algorithm* 4.1; *dashed:* (4.9) *with* $q_1 = g$; *dotted:* (4.9) *with random* q_1.

(where x_{IP} is available), and iteration number k, Table 2 also displays $\|r_k\|$, $|y_k^T r_k|$, and $|g^T y_k|$, which relates η_k and ζ_k to the eigenvalue residual $\|r_k\|$ (see Theorem 4.1).

Figure 7 plots the convergence history of z_k for the inverse heat problem.

These numerical results show that a solution to the Lagrange equations to the desired accuracy is obtained within a small number of iterations k for all but the `spike` problem. The speed of convergence compares favorably with that of the LSTRS

TABLE 2
$\eta_k = \frac{z_k^T z_k - \delta^2}{\delta^2}$, $\zeta_k = \frac{\|Hz_k - \theta_k z_k - g\|}{\|g\|}$.

Problem	δ	η_k	γ_k	$\frac{\|z_k - x_{\rm IP}\|}{\|x_{\rm IP}\|}$	k	$\|r_k\|$	$\|y_k^T r_k\|$	$\|g^T y_k\|$
`barrt`	1.2	$2e-12$	$3e-7$	$1e-1$	4	$3e-08$	$6e-17$	$1e-2$
`ill heat`	7.7	$9e-16$	$4e-7$	$2e-2$	26	$5e-13$	$3e-23$	$1e-4$
`well heat`	7.7	$2e-15$	$9e-7$	$4e-4$	112	$6e-08$	$5e-17$	$1e+0$
`blur`	36.5	$1e-15$	$9e-7$	$1e-5$	327	$1e-08$	$6e-19$	$7e-1$
`deriv2` (1)	0.6	$1e-16$	$8e-7$	$2e-1$	19	$1e-15$	$1e-26$	$1e-7$
`deriv2` (2)	1.7	$2e-16$	$9e-7$	$1e-1$	19	$1e-15$	$1e-26$	$3e-7$
`deriv2` (3)	0.3	$5e-16$	$2e-7$	$9e-3$	10	$4e-14$	$8e-24$	$6e-6$
`foxgood`	18.2	$8e-16$	$1e-8$	$7e-3$	4	$1e-11$	$3e-20$	$2e-2$
`parallax`	18.1	$1e-16$	$2e-8$	–	13	$9e-15$	$1e-23$	$3e-05$
`phillips`	2.9	$1e-15$	$5e-7$	$8e-3$	11	$1e-06$	$1e-16$	$2e-1$
`shaw`	31.5	$1e-15$	$2e-7$	$4e-2$	7	$4e-09$	$9e-20$	$1e-1$
`spikes`	40.6	$1e-14$	$7e-4$	$1e+1$	200	$1e-06$	$1e-19$	$1e-5$
`ursell`	1.0	$5e-16$	$1e-7$	–	3	$4e-08$	$1e-16$	$5e-1$
`wing`	0.6	$6e-16$	$4e-7$	$6e-1$	3	$4e-11$	$7e-21$	$5e-4$

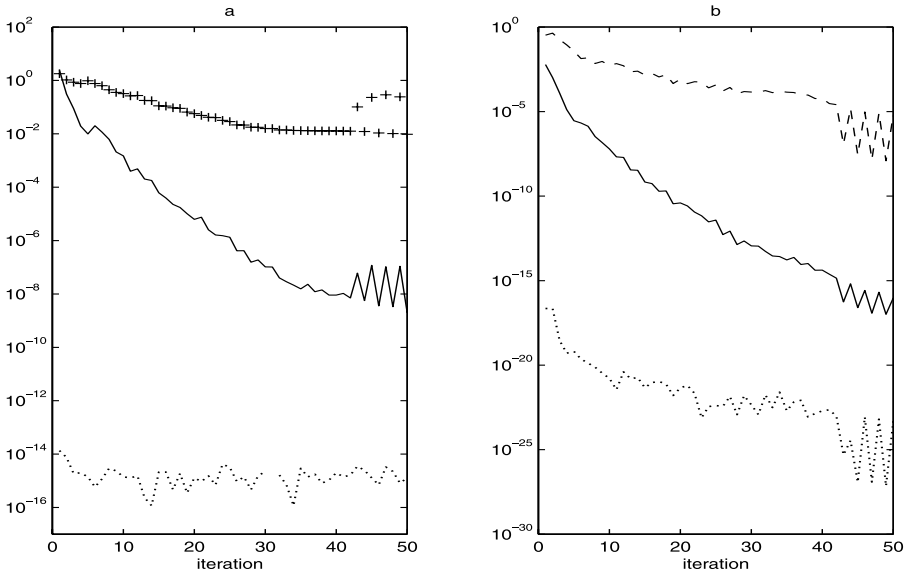


FIG. 7. *Convergence of least squares solution for* `heat` *problem.* Left: $||(H-\lambda)z_k - g||/||g||$ *(solid),* $(z_k^T z_k - \delta^2)/\delta^2$ *(dot),* $||z_k - x_{\rm IP}||/||x_{\rm IP}||$ *(+).* Right: $||r_k||$ *(solid),* $|y_k^T r_k|$ *(dot),* $|g^T y_k|$ *(dash).*

method due to Rojas and Sorenson [23]. The results also show improvement in accuracy in these tests. For the `spike` problem, $\|r_k\|$ is of $\mathcal{O}(1)$ throughout because of the large norm of H^2 and hence γ_k, η_k are not reduced to the given thresholds.

We further observe from Table 2 that γ_k is proportional to $\|r_k\|/|g^T y_k|$ and η_k is nearly proportional to $|y^T r_k|/|g^T y_k|^2$, as suggested by Theorem 4.1. With $|g^T y|$ being very small in such problems, a typical iteration will see $\|r_k\|$ gradually decreased, while $|g^T y_k|$ is also decreased. Then, $\gamma_k = \|(H-\theta I)z_k - g\|$ will stagnate at a level given by $\delta^2\|r_k\|/|g^T y_k|$. On the other hand, with θ_k computed through a Rayleigh quotient, a very good orthogonality $|y_k^T r_k|$ is achieved and this in turn keeps $\delta^2|y_k^T r_k|/|g^T y_k|^2$ and hence the constraint error $(z_k^T z_k - \delta^2)/\delta^2$ usually in the order of machine precision. So, z_k nearly satisfies the constraint throughout.

6. Conclusions. We have presented a basic Arnoldi-type process for a large monic quadratic matrix polynomial. The process is particularly efficient when some combination of the coefficient matrices A and B is of low rank, or one of them, say B, is a polynomial of A plus a low rank matrix. We have applied it to the quadratic eigenvalue problem arising in the quadratically constrained least squares problem. Our testing demonstrates its effectiveness for this class of problems.

Acknowledgments. The authors would like to acknowledge many fruitful conversations they had with Prof. Zhaojun Bai of the University of California at Davis during the course of this work. They are also indebted to the referees for their constructive and detailed suggestions that improved the paper significantly both in terms of presentations and technical details. In particular, they thank an anonymous referee for suggesting the Rayleigh quotient approach (4.8) to enforce the orthogonality between y and r from using the QR algorithm [11]. The approach also actually improves the solutions by the new process here. They also thank the other referee for suggesting the ordinary Lanczos process for (4.4) (see (4.9)).

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[2] Z. Bai, *personal communication*, 2000.
[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.
[4] D. Calvetti, G. H. Golub, and L. Reichel, *Estimation of the L-curve via Lanczos bidiagonalization*, BIT, 39 (1999), pp. 603–619.
[5] D. Calvetti, B. Lewis, and L. Reichel, *On the regularizing properties of the GMRES method*, Numer. Math., 91 (2002), pp. 605–625.
[6] D. Calvetti, L. Reichel, and Q. Zhang, *Iterative exponential filtering for large discrete ill-posed problems*, Numer. Math., 83 (1999), pp. 535–556.
[7] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[8] 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.
[9] G. E. Forsythe and G. H. Golub, *On the stationary values of a second-degree polynomial on the unit sphere*, SIAM J. Soc. Indust. Appl. Math., 13 (1965), pp. 1050–1068.
[10] W. Gander, *Least squares with a quadratic constraint*, Numer. Math, 36 (1981) pp. 291–307.
[11] W. Gander, G. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114/115 (1989), pp. 815–839.
[12] G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991). pp. 561–580.
[13] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[14] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[15] P. C. Hansen, *Regularization Tools: A MATLAB package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[16] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1997.
[17] L. Hoffnung, R.-C. Li, and Q. Ye, *Krylov Type Subspace Methods for Matrix Polynomials*, Research report 2002-08, Department of Mathematics, University of Kentucky, 2002, available online from http://www.ms.uky.edu/∼math/MAreport; Linear Algebra Appl., to appear.
[18] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research National Bureau Standards, 45 (1950), pp. 255–282.
[19] R. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK User's Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998.
[20] K. Meerbergen, *Locking and restarting quadratic eigenvalue solvers*, SIAM J. Sci. Comput.,

22 (2001), pp. 1814–1839.
[21] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM. J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[22] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1997.
[23] M. Rojas and D. C. Sorensen, *A trust-region approach to the regularization of large-scale discrete forms of ill-posed problems*, SIAM J. Sci. Comput., 23 (2002), pp. 1842–1860.
[24] M. Rojas, S. A. Santos, and D. C. Sorensen, *A new matrix-free algorithm for the large-scale trust-region subproblem*, SIAM J. Optim., 11 (2000), pp. 611–646
[25] G. Sleijpen, J. Booten, D. Fokkema, and H. van der Vorst, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 593–633.
[26] D. C Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426.
[27] D. C. Sorensen, *Minimization of a large-scale quadratic function subject to a spherical constraint*, SIAM J. Optim., 7 (1997). pp. 141–161.
[28] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.
[29] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, MA, 1996.
[30] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–386.
[31] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, England, 1965.
[32] Q. Ye, *A convergence analysis for nonsymmetric Lanczos algorithms*, Math. Comp., 56 (1991), pp. 677–691.
[33] T. Zhang, G. H. Golub, and K. H. Law, *Eigenvalue perturbation and generalized Krylov subspace method*, Appl. Numer. Math., 27 (1998), pp. 185–202.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 429–444

A MATRIX PERTURBATION VIEW OF THE SMALL WORLD PHENOMENON*

DESMOND J. HIGHAM†

Abstract. We use techniques from applied matrix analysis to study small world cutoff in a Markov chain. Our model consists of a periodic random walk plus uniform jumps. This has a direct interpretation as a teleporting random walk, of the type used by search engines to locate web pages, on a simple ring network. More loosely, the model may be regarded as an analogue of the original small world network of Watts and Strogatz [*Nature*, 393 (1998), pp. 440–442]. We measure the small world property by expressing the mean hitting time, averaged over all states, in terms of the expected number of shortcuts per random walk. This average mean hitting time is equivalent to the expected number of steps between a pair of states chosen uniformly at random. The analysis involves nonstandard matrix perturbation theory and the results come with rigorous and sharp asymptotic error estimates. Although developed in a different context, the resulting cutoff diagram agrees closely with that arising from the mean-field network theory of Newman, Moore, and Watts [*Phys. Rev. Lett.*, 84 (2000), pp. 3201–3204].

Key words. Google, Markov chain, matrix perturbation, mean hitting time, optional sampling theorem, partially random graph, random walk, Sherman–Morrison formula, teleporting, web search engine

AMS subject classifications. 65F35, 65C40, 60J27

DOI. 10.1137/S0895479802406142

1. Introduction.

We show here that a small world cutoff arises in a simple random walk setting that is amenable to rigorous analysis via matrix perturbation theory. Our model is derived by adding uniform jumps to a periodic, one-dimensional random walk. Increasing the jump probability allows us to interpolate between completely local and completely global behavior. The small world property is then quantified by the average or maximum of the mean hitting times.

Although it is simplistic, we believe that this model is relevant to many physical, sociological, epidemiological, and computational applications, as it combines the traditional notion of diffusion on a lattice [3, 4, 16, 20] with the type of partially random connectivity that has recently been used to describe complex, real-life networks [6, 12, 15, 17, 18, 22, 23, 24]. In particular, we mention that the original work of Watts and Strogatz [25] included a disease simulation that is in a similar spirit to our model.

More specifically, the idea of taking a "random walk plus shortcuts" is used by web search engines. Here, the fundamental task is to locate all web pages by following hyperlinks. A simple random walk—finding all links out of the current page and choosing one of them uniformly—is liable to reach a dead end or to cycle. To avoid this, it is common to jump occasionally to a page chosen uniformly at random. Adding jumps in this way is known as *teleporting* [10]. The search engine Google uses just such an algorithm [14]. Our results apply directly to the case of teleporting on a ring lattice and quantify, in terms of the teleporting parameter, the expected number of links that must be followed to reach a given target.

*Received by the editors April 22, 2002; accepted for publication (in revised form) by M. Chu April 4, 2003; published electronically September 9, 2003. This work was supported by a Research Fellowship from the Leverhulme Trust.

http://www.siam.org/journals/simax/25-2/40614.html

†Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK (djh@maths.strath.ac.uk).

Although the correspondence is not exact, we originally developed this model by analogy with the randomized network approach of Watts and Strogatz [25]. In that work, the authors showed experimentally that by replacing a small number of connections by new connections between randomly chosen nodes, that is, by *randomly rewiring* a few times, the small world property is roused before the clustering property is lost. They coined the phrase *small world phenomenon* to describe the unlikely alliance of the clustering and small world properties. Initially, the small world property was verified through numerical simulation. More recently, Newman, Moore, and Watts [19] gave a semiheuristic analysis of a closely related model in the limit of large network size. Here random links do not replace existing links but instead are added to the network and are thus referred to as shortcuts. The resulting mean-field expression for the path length is shown as a dashed line in Figure 3.3 below. An unsatisfactory feature of the treatment in [19] is that it is designed to be valid for either a large or small number of shortcuts, that is, $x \ll 1$ or $x \gg 1$ on the x-axis in Figure 3.3. This does not cover the interesting cutoff region where the average path length sharply decreases as a function of the average number of shortcuts added. However, simulations reported in [19] showed that the mean-field expression continues to give a reasonable fit in this range. A fully rigorous analysis that applies only for a large number of shortcuts ($x \gg 1$) has been given in [1].

For our random walk model, we measure the small world property as the maximum or average mean hitting time, rather than the expected path length. One of the key advantages of this approach is that it permits a rigorous analysis in the asymptotic limit of a large number of states. Further, the analysis is sharp; we obtain exact expressions for the leading terms in the expansions. Our results include what appears to be the first rigorous analysis of a small world cutoff effect for the interesting $O(1)$ shortcuts regime. Quite remarkably, the analytical cutoff diagram that we derive is in close agreement with the one that has been found experimentally for the Watts–Strogatz network model.

In the next section we set up the random walk as a Markov chain and state results about the mean hitting times. The results are interpreted in section 3. We show that a certain scaling of the interpolation parameter (in terms of the chain length) has a particular physical significance. For this scaling, we obtain a cutoff diagram that illustrates the small world phenomenon and may be compared with that of the Watts–Strogatz network model. Section 4 is the heart of the paper. Here we prove the key results using techniques from numerical analysis to capture the effect of a certain structured perturbation on a linear algebraic system. Because the perturbation depends on the dimension of the system, the usual "$(I+E)^{-1} = I - E + O(\|E\|^2)$" expansion cannot be employed in general. Section 5 points to possible future work.

2. The Markov chain approach.

2.1. The model. We begin by setting up the relevant mathematical concepts. A discrete time, finite state Markov chain is a stochastic process $\{X_n\}_{n\geq 0}$ that can be characterized by a *transition matrix* P. We suppose that there are N states, labeled 1 to N, so $P \in \mathbb{R}^{N\times N}$. The value p_{ij} specifies the probability that $X_{n+1} = j$ given that $X_n = i$, that is,

$$\mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij},$$

with all $p_{ij} \geq 0$ and $\sum_{j=1}^{N} p_{ij} = 1$. We will always make the process start at state 1, so $X_0 = 1$ with probability 1. The *mean hitting time* for state j is the average

number of steps taken by the process before first reaching state j. More precisely, the mean hitting time for state j is the expected value of the random variable $h^j(\omega) := \inf\{n \geq 0 : X_n(\omega) = j\}$. We let $\mathbf{z} \in \mathbb{R}^{N-1}$ denote the vector of mean hitting times for states $2, 3, \ldots, N$, so z_j is the mean hitting time for state $j+1$. A standard result that may be found, for example, in [20, Theorem 1.3.5] shows that $\mathbf{z}$ is the minimal nonnegative solution to the system of linear equations

$$\left(I - \widehat{P}\right)\mathbf{z} = \mathbf{e}. \tag{2.1}$$

Here, $\widehat{P} \in \mathbb{R}^{(N-1)\times(N-1)}$ is formed by removing the first row and column from P, so $\widehat{p}_{ij} = p_{i+1,j+1}$, and $\mathbf{e} := [1, 1, \ldots, 1]^T \in \mathbb{R}^{N-1}$. We find it natural to use the mean hitting time as an analogue of the path length in order to measure the "small world" size of the Markov chain. We will consider the *maximum mean hitting time*

$$\mathrm{mht}_{\max}(P) := \max_{1\leq i\leq N-1} z_i \tag{2.2}$$

and the *average mean hitting time*

$$\mathrm{mht}_{\mathrm{ave}}(P) := \frac{1}{N-1}\sum_{i=1}^{N-1} z_i. \tag{2.3}$$

We note that $\mathrm{mht}_{\mathrm{ave}}(P)$ has the agreeable interpretation as the expected number of steps between a pair of sates chosen uniformly at random. There are, of course, other hitting time measures, such as the expected value of $\max_{1\leq j\leq N} h^j(\omega)$, that may be of interest. We focus on (2.2) and (2.3) because we believe them to be natural choices and because they can be studied via matrix analysis.

By analogy with the basic ring network in [25], we consider the Markov chain with transition matrix

$$P_0 = \begin{bmatrix} 0 & \frac{1}{2} & & & & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & & & \\ & \frac{1}{2} & 0 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \frac{1}{2} \\ \frac{1}{2} & & & & \frac{1}{2} & 0 \end{bmatrix} \in \mathbb{R}^{N\times N}. \tag{2.4}$$

Here, at each step the process moves to either of the two neighboring states with equal probability (with 1 and N regarded as neighbors). This could also be described as a symmetric, one-dimensional, periodic random walk. With this choice of transition matrix, the system (2.1) becomes

$$T\mathbf{z} = \mathbf{e}, \tag{2.5}$$

where $T := \mathrm{tridiag}(-\frac{1}{2}, 1, -\frac{1}{2})$ and $\mathrm{tridiag}(a, b, c)$ denotes a tridiagonal Toeplitz matrix of the form

$$\begin{bmatrix} b & c & & & \\ a & b & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c \\ & & & a & b \end{bmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}.$$

It is well known, and easily verified, that (2.5) has the unique solution $z_i = i(N-i)$. Hence,

$$\text{mht}_{\max}(P_0) := \frac{N^2}{4} + O(1) \quad \text{and} \quad \text{mht}_{\text{ave}}(P_0) := \frac{N(N+1)}{6}. \tag{2.6}$$

Now we perturb the basic transition matrix P_0 in (2.4) by resetting all zero entries to $\epsilon > 0$. This gives the transition matrix

$$P_\epsilon = \begin{bmatrix} \epsilon & \frac{1}{2}-\widehat{\epsilon} & \epsilon & \dots & \epsilon & \frac{1}{2}-\widehat{\epsilon} \\ \frac{1}{2}-\widehat{\epsilon} & \epsilon & \frac{1}{2}-\widehat{\epsilon} & \epsilon & \dots & \epsilon \\ \epsilon & \frac{1}{2}-\widehat{\epsilon} & \epsilon & \frac{1}{2}-\widehat{\epsilon} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \epsilon \\ \epsilon & \dots & \epsilon & \frac{1}{2}-\widehat{\epsilon} & \epsilon & \frac{1}{2}-\widehat{\epsilon} \\ \frac{1}{2}-\widehat{\epsilon} & \epsilon & \dots & \epsilon & \frac{1}{2}-\widehat{\epsilon} & \epsilon \end{bmatrix} \in \mathbb{R}^{N\times N}, \tag{2.7}$$

where in order to keep the row sums equal to one we require

$$\widehat{\epsilon} = \frac{N-2}{2}\epsilon, \tag{2.8}$$

and in order to maintain nonnegative entries we require

$$\epsilon \le \frac{1}{N-2}. \tag{2.9}$$

On each step the Markov chain with transition matrix P_ϵ moves to either of the two neighboring states with equal probability $\frac{1}{2} - \frac{N-2}{2}\epsilon$ and to each nonneighboring state with probability ϵ. This is precisely the teleporting idea described in section 1 applied to a ring network and, more loosely, is analogous to the rewiring operation used to generate small world networks. We may regard ϵ as a parameter that allows us to interpolate between a local and a global process.

The main issue that we address in this work is how the mean hitting times are reduced as ϵ is increased from zero. This leads to an interesting problem in matrix perturbation theory. We will compute expressions for the *maximum mean hitting time reduction ratio*

$$\frac{\text{mht}_{\max}(P_\epsilon)}{\text{mht}_{\max}(P_0)} \tag{2.10}$$

and the *average mean hitting time reduction ratio*

$$\frac{\text{mht}_{\text{ave}}(P_\epsilon)}{\text{mht}_{\text{ave}}(P_0)} \tag{2.11}$$

for small ϵ and large N. The constraint (2.9) shows that ϵ must scale with N, and hence we consider the power law relationship

$$\epsilon := \frac{K}{N^\alpha} \quad \text{for fixed } K > 0 \text{ and } \alpha > 1 \quad \text{in the limit } N \to \infty. \tag{2.12}$$

For reference, note that in the case where $\epsilon = 1/N$, all entries of P_ϵ are equal. This is the fully global regime where the process moves to any other state with equal probability. In this case, it follows from (2.1) (or from basic probabilistic arguments) that the mean hitting time vector has all entries $z_j = N$. Hence, from (2.6),

$$\frac{\text{mht}_{\max}(P_{1/N})}{\text{mht}_{\max}(P_0)} = \frac{4}{N} + O(N^{-3}) \quad \text{and} \quad \frac{\text{mht}_{\text{ave}}(P_{1/N})}{\text{mht}_{\text{ave}}(P_0)} = \frac{6}{N+1}.$$

2.2. Results. Theorems 2.1, 2.2, and 2.3 below completely characterize the reduction ratios for all ϵ in (2.12), to leading order in N. Proofs are given in section 4.

THEOREM 2.1. *For $\alpha > 3$,*

$$\frac{\mathrm{mht}_{\max}(P_\epsilon)}{\mathrm{mht}_{\max}(P_0)} = 1 + O(N^{3-\alpha}) \quad \text{and} \quad \frac{\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)}{\mathrm{mht}_{\mathrm{ave}}(P_0)} = 1 + O(N^{3-\alpha}).$$

THEOREM 2.2. *For $\alpha = 3$,*

$$\frac{\mathrm{mht}_{\max}(P_\epsilon)}{\mathrm{mht}_{\max}(P_0)} = \frac{2\sqrt{2}}{\sqrt{K}} \tanh \tfrac{\sqrt{K}}{2\sqrt{2}} + O(N^{-1}) \tag{2.13}$$

and

$$\frac{\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)}{\mathrm{mht}_{\mathrm{ave}}(P_0)} = \frac{6}{K}\left(\frac{\sqrt{2K}}{2\tanh \frac{\sqrt{2K}}{2}} - 1\right) + O(N^{-1}). \tag{2.14}$$

THEOREM 2.3. *For $1 < \alpha < 3$,*

$$\frac{\mathrm{mht}_{\max}(P_\epsilon)}{\mathrm{mht}_{\max}(P_0)} = \frac{2\sqrt{2}}{\sqrt{K}} N^{\frac{\alpha-3}{2}} + O(N^{-1}) \tag{2.15}$$

and

$$\frac{\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)}{\mathrm{mht}_{\mathrm{ave}}(P_0)} = \frac{3\sqrt{2}}{\sqrt{K}} N^{\frac{\alpha-3}{2}} - \frac{6}{K} N^{\alpha-3} + O(N^{-1}). \tag{2.16}$$

(We remark that the second term on the right-hand side of (2.16) *can be absorbed into the final $O(N^{-1})$ term for $\alpha \le 2$.)*

3. Interpretation and discussion. The theorems show that there is a threshold at $\alpha = 3$. For larger α values, the ϵ perturbation has no effect on the mean hitting time reduction ratios in the $N \to \infty$ limit. For $\alpha = 3$, the reduction ratio has a fixed, nonzero value for each K. For α below 3, the ϵ perturbation dominates the process, giving a reduction ratio that is asymptotically zero.

In the case of networks, the small world phenomenon has been characterized by expressing some measure of the average path length in terms of the expected number of shortcuts added [19, 25]. An appropriate characterization in our Markov chain setting is to measure the average mean hitting time, $\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)$, as a function of the expected number of shortcuts (teleportings) *taken* per random walk. (We say that a shortcut takes place from step n to step $n+1$ if $X_{n+1} \neq (X_n \pm 1) \bmod N$.) Now, on each step, the probability of a shortcut is $\epsilon(N-2)$. Define the process M_n by

$$M_n := (\text{number of shortcuts up to step } n) - n\epsilon(N-2). \tag{3.1}$$

Subtracting the drift in this way produces a martingale, that is, $\mathbb{E}M_n = 0$. Since h^j is a stopping time, the optional sampling theorem [11, Chapter 3, Corollary 3.1] may be applied to give $\mathbb{E}M_{h^j} = \mathbb{E}M_0 = 0$. Using this in (3.1), we find that the expected number of shortcuts up to the hitting time for state j is given by $\epsilon(N-2)\mathbb{E}h^j$. So if we let W_ϵ denote the average over all states of the expected number of shortcuts taken per random walk, then $W_\epsilon = \epsilon(N-2)\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)$. Applying Theorems 2.1–2.3 leads immediately to the following corollary.

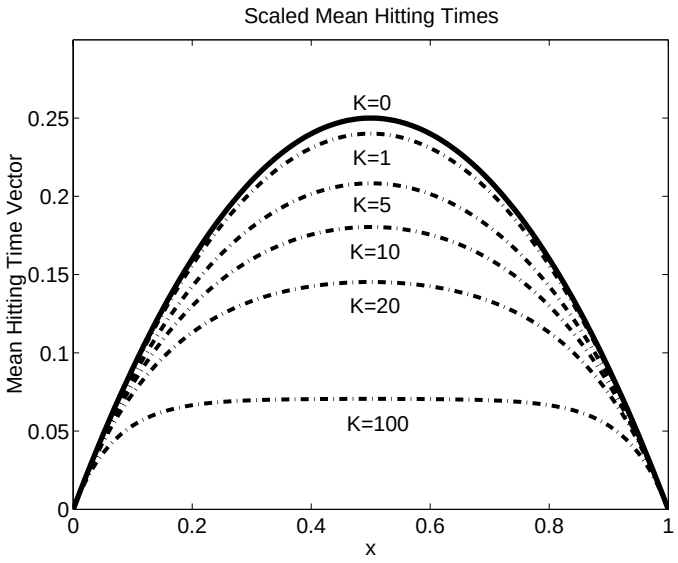


FIG. 3.1. *Curves describing the vector of scaled mean hitting times with $\alpha = 3$ for different values of K in* (2.12).

COROLLARY 3.1. *The average over all states of the expected number of shortcuts taken per random walk, W_ϵ, has the following properties:*

1. *For $\alpha > 3$, $W_\epsilon \to 0$ as $N \to \infty$.*
2. *For $\alpha = 3$,*

$$W_\epsilon = \frac{\sqrt{2K}}{2\tanh\frac{\sqrt{2K}}{2}} - 1 + O(N^{-1}).$$

3. *For $1 < \alpha < 3$, $W_\epsilon \to \infty$ as $N \to \infty$.*

Corollary 3.1 distinguishes $\alpha = 3$ as the appropriate regime in which to search for the small world phenomenon—it is only in this case that the ϵ perturbation introduces a nonzero but bounded number of shortcuts. So henceforth we consider only the case $\alpha = 3$. Note that this scaling is easily arrived at via the following heuristic arguments. Typical excursions on the basic ring take $O(N^2)$ steps. For the P_ϵ model, the probability of a shortcut on each step is $(N-2)\epsilon = O(N^{1-\alpha})$. Hence, if the $O(N^2)$ excursion length is preserved and a finite number of shortcuts are to be taken, then a reasonable guess is to set $O(N^2) \times O(N^{1-\alpha}) = O(1)$, giving $\alpha = 3$. However, since our analysis provides the coefficients associated with the leading order asymptotics, we are able to investigate the model more closely.

Returning our attention to the individual mean hitting times, for $\alpha = 3$ it follows from the analysis in section 4 (more precisely, from (4.5), (4.11), and (4.15)–(4.17)) that z_j is perturbed to $z_{\epsilon j}$, where

$$z_{\epsilon j} = \frac{N^2}{\sqrt{2K}\tanh\frac{\sqrt{2K}}{2}}\left[1 - \frac{\cosh\sqrt{2K}(x_j - \frac{1}{2})}{\cosh\frac{\sqrt{2K}}{2}}\right] + O(N), \text{ with } x_j := \frac{j}{N}. \tag{3.2}$$

In Figure 3.1 we plot curves for the mean hitting time vector, as given by the first term on the right-hand side of (3.2), scaled by N^2. We show the cases $K = 1, 5, 10, 20, 100$.

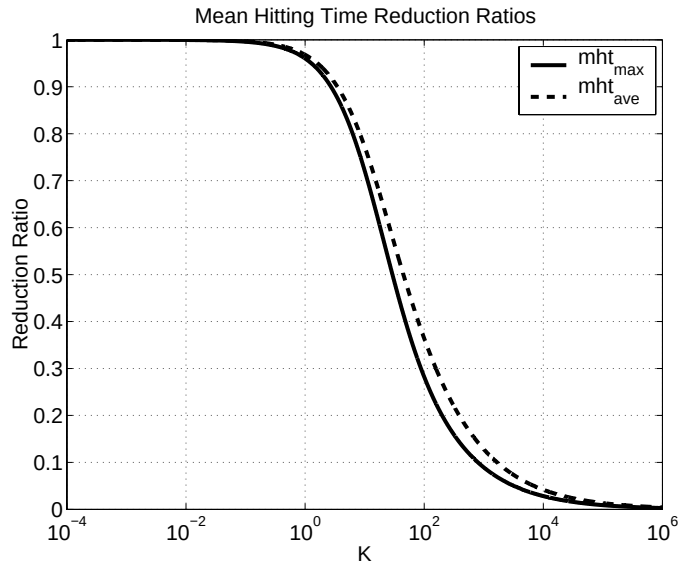


FIG. 3.2. *Leading term in the maximum (solid line) and average (dashed line) mean hitting time reduction ratios as a function of K, from Theorem* 2.2.

(Note that components in the scaled vector $\mathbf{z}_\epsilon/N^2$ are found by evaluating the curves at the equally spaced points $1/N, 2/N, \ldots, (N-1)/N$ along the x-axis.) We have also plotted the $K = 0$ case, that is, $\epsilon = 0$, which corresponds to the parabola $x(1-x)$. The outcome is intuitively reasonable—the mean hitting times decrease and the profile flattens as K, and hence ϵ, increases.

Separate numerical testing indicated that (3.2) is sharp—the remainder term behaves like a nonzero multiple of N.

Turning now to the mean hitting time reduction ratios, (2.10) and (2.11), Figure 3.2 plots the leading terms in (2.13) and (2.14) as functions of K. (Note that the horizontal axis is logarithmically scaled in order to zoom in on the region of interest.) We see that there is a rapid decay when K is increased beyond ≈ 1.

To look for the small world phenomenon, we now plot the average mean hitting time reduction ratio, $\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)$, as a function of the the average over all states of the expected number of shortcuts taken per random walk, W_ϵ. From Theorem 2.2 and Corollary 3.1, these may be computed via the parametric form

$$W_\epsilon = \frac{\sqrt{2K}}{2\tanh\frac{\sqrt{2K}}{2}} - 1 \quad \text{and} \quad \mathrm{mht}_{\mathrm{ave}}(P_\epsilon) = \frac{6}{K} W_\epsilon, \tag{3.3}$$

with an error of $O(N^{-1})$. The solid line in Figure 3.3 shows the resulting curve. A sharp cutoff is noticeable as the number of shortcuts increases from around $\frac{1}{2}$ to 50—the small world effect kicks in abruptly when only a small number of shortcuts are taken.

It is possible to compare the behavior of this model with that of the $k = 1$ version of the Newman–Moore–Watts network model [19], which is closely related to the corresponding Watts–Strogatz model [25]. In the network model, we begin with a ring of N nodes, where node i is connected to node j if $|i-j| = 1 \bmod N$. This "local" network is interpolated toward the "global" by adding shortcuts between randomly

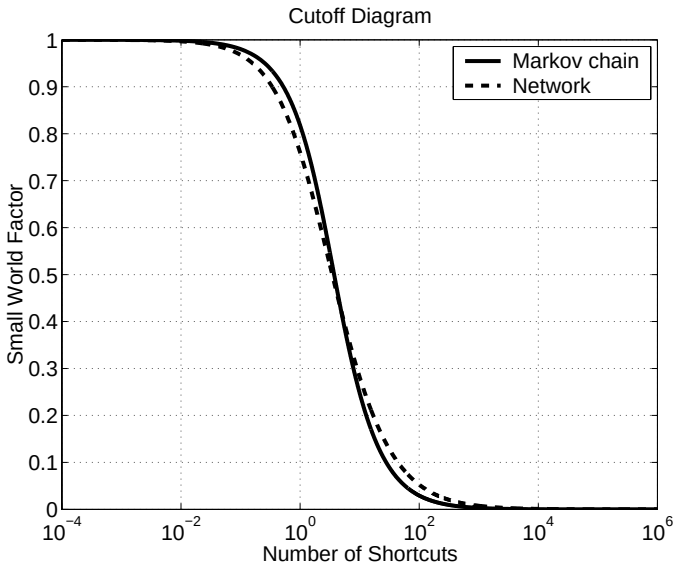


FIG. 3.3. *Solid line is* (3.3)*: x-axis is average number of shortcuts per excursion,* W_ϵ*, y-axis is reduction ratio for average mean hitting time,* $\mathrm{mht}_{\mathrm{ave}}(P_\epsilon)/\mathrm{mht}_{\mathrm{ave}}(P_0)$*. Dashed line is, from* [19]*,* $y = 4f(x)$ *with* f *defined in* (3.4)*: x-axis is average number of shortcuts per network, y-axis is reduction ratio for average path length.*

selected nodes. The mean-field expression in [19] for the average path length as a function of the expected number of shortcuts is $y = Nf(x)$, where

$$f(x) = \frac{1}{2\sqrt{x^2+2x}} \tanh^{-1} \frac{x}{\sqrt{x^2+2x}}. \tag{3.4}$$

Since the average path length when there are no shortcuts is, to leading order, $N/4$, the curve $y = 4f(x)$ gives the reduction in the average path length in terms of the average number of shortcuts. This is plotted with a dashed line in Figure 3.3. As we mentioned in section 1, the authors note in [19] that their mean-field approximation involves assumptions that are valid only for values that correspond to $x \ll 1$ and $x \gg 1$ on the x-axis of Figure 3.3. However, simulations show that the curve also gives quite an accurate description of the cutoff region around $x = 1$; see [19, Figure 2] or [22, Figure 5].

Overall, although the two measures are fundamentally different, Figure 3.3 shows that there is a remarkable qualitative and quantitative agreement between the small world cutoff behavior in the Markov chain and randomized network models. In particular, the mean-field theory predicts that an average of 3.5 shortcuts per network are needed to give a reduction of $\frac{1}{2}$ in the average path length (consistent with the simulations of [25]). The average number of shortcuts per random walk required to give an average mean hitting time reduction ratio of $\frac{1}{2}$ is 3.7. The corresponding figures for a reduction of $\frac{1}{10}$ are 44 and 28, respectively.

So far, we have focused on measuring mean hitting times by analogy with path length. Is there a corresponding analogue of the clustering property? One possibility is to consider how rapidly the Markov chain converges to its equilibrium distribution. We may regard the chain as *not* being clustered if it tends quickly to equilibrium—that is, transient behavior rapidly gives way to steady state behavior. The rate at

which equilibrium is approached can be bounded above and below in terms of the spectrum of the transition matrix; see, for example, [2, Theorem 10.3]. In our case P_ϵ in (2.7) is circulant, and hence its eigenvalues can be calculated explicitly. For $\alpha = 3$ we find that with N even there is an eigenvalue of modulus $1 - K/N^2$ and with N odd there is a repeated eigenvalue of modulus $1 - (K + \pi^2)/N^2 + O(N^{-4})$. We may conclude that for large N the process is slow to approach equilibrium and hence remains clustered. In this sense, when the cutoff in Figure 3.3 takes place we have captured the small world phenomenon.

4. Proofs.

4.1. Preliminaries. Theorems 2.1–2.3 concern matrix perturbation theory. The vector $\mathbf{z}_\epsilon \in \mathbb{R}^{N-1}$ is the (minimal nonnegative) solution to

$$T_\epsilon \mathbf{z}_\epsilon = \mathbf{e}, \tag{4.1}$$

where $T_\epsilon = T + \text{tridiag}(\frac{N\epsilon}{2}, 0, \frac{N\epsilon}{2}) - \epsilon \mathbf{e}\mathbf{e}^T$. We have $\text{mht}_{\max}(P_\epsilon) = \| \mathbf{z}_\epsilon \|_\infty$ and $\text{mht}_{\text{ave}}(P_\epsilon) = \| \mathbf{z}_\epsilon \|_1/(N-1)$, where $\| \cdot \|_\infty$ and $\| \cdot \|_1$ are used to denote the vector ∞ and 1 norms and their induced matrix norms, respectively. We are thus concerned with the normwise effect on the size of the solution when (2.5) is perturbed to (4.1). For the $\alpha > 3$ case, a standard expansion can be used; see section 4.2. However, for $\alpha \leq 3$ this approach is no longer applicable—special care is needed to deal with the dependence of the perturbation on the dimension N; see section 4.3.

We find it useful to let

$$\widehat{T}_\epsilon = T + \Delta T_\epsilon, \tag{4.2}$$

with

$$\Delta T_\epsilon = \text{tridiag}(\tfrac{N\epsilon}{2}, 0, \tfrac{N\epsilon}{2}). \tag{4.3}$$

Note that $\widehat{T}_\epsilon$ is diagonally dominant and hence nonsingular. We also let $\mathbf{y}_\epsilon \in \mathbb{R}^{N-1}$ satisfy

$$\widehat{T}_\epsilon \mathbf{y}_\epsilon = \mathbf{e}. \tag{4.4}$$

Now, we may use the Sherman–Morrison formula [9, p. 490] to deal with the rank one perturbation that converts $\widehat{T}_\epsilon$ to T_ϵ. First note that the inequality $1 - \epsilon \mathbf{e}^T \mathbf{y}_\epsilon \neq 0$ follows from the analysis below. (More precisely, it follows from (4.9) for $\alpha > 3$ and from (4.16) for $1 < \alpha \leq 3$.) Hence, by the Sherman–Morrison formula, $\widehat{T}_\epsilon$ in (4.2) is nonsingular and

$$\begin{aligned} \mathbf{z}_\epsilon = T_\epsilon^{-1}\mathbf{e} &= \left(\widehat{T}_\epsilon - \epsilon \mathbf{e}\mathbf{e}^T\right)^{-1} \mathbf{e} \\ &= \left(\widehat{T}_\epsilon^{\,-1} + \frac{\epsilon \widehat{T}_\epsilon^{\,-1} \mathbf{e}\mathbf{e}^T \widehat{T}_\epsilon^{\,-1}}{1 - \epsilon \mathbf{e}^T \widehat{T}_\epsilon^{\,-1} \mathbf{e}}\right) \mathbf{e} \\ &= \frac{1}{1 - \epsilon \mathbf{e}^T \mathbf{y}_\epsilon} \mathbf{y}_\epsilon. \end{aligned} \tag{4.5}$$

We also note a few more facts. First, recall that A is defined to be a *Stieltjes matrix*, that is, a symmetric M-matrix, if $A^{-1} \geq 0$ and $a_{ij} \leq 0$ for $i \neq j$. (Inequalities between vectors or matrices are understood to hold for all components.) Further, any

strictly or irreducibly diagonally dominant symmetric A with $a_{ij} \leq 0$ for $i \neq j$ and $a_{ii} > 0$ for all i is a Stieltjes matrix; see, for example, [21, Theorem 6.2.17]. It follows that T and $\widehat{T}_\epsilon$ are Stieltjes matrices, and hence $T^{-1} \geq 0$ and $\widehat{T}_\epsilon^{\,-1} \geq 0$. Further,

$$\| T^{-1} \|_1 = \| T^{-1} \|_\infty = \| T^{-1}\mathbf{e} \|_\infty = \frac{N^2}{4} + O(1). \tag{4.6}$$

4.2. Proof of Theorem 2.1.

Proof. First, note that

$$\| \Delta T_\epsilon \|_\infty = \| \Delta T_\epsilon \|_1 = O(N^{1-\alpha})$$

and hence

$$\| T^{-1}\Delta T_\epsilon \|_\infty = \| T^{-1}\Delta T_\epsilon \|_1 \leq \| T^{-1} \|_1 \| \Delta T_\epsilon \|_1 = O(N^{3-\alpha}).$$

Since $\alpha > 3$ we have $\| T^{-1}\Delta T_\epsilon \|_\infty \to 0$ and $\| T^{-1}\Delta T_\epsilon \|_1 \to 0$. We may thus appeal to standard perturbation theory and expand $(I + T^{-1}\Delta T_\epsilon)^{-1}$ in powers of $T^{-1}\Delta T_\epsilon$; see, for example, [5, Lemma 2.1]. We have

$$\begin{aligned} \| \mathbf{y}_\epsilon \|_\infty &= \| (T + \Delta T_\epsilon)^{-1}\mathbf{e} \|_\infty \\ &= \| (I + T^{-1}\Delta T_\epsilon)^{-1}T^{-1}\mathbf{e} \|_\infty \\ &= \| \left[I - T^{-1}\Delta T_\epsilon + O(\| T^{-1}\Delta T_\epsilon \|_\infty^2) \right] T^{-1}\mathbf{e} \|_\infty) \\ &= \| T^{-1}\mathbf{e} \|_\infty + O(\| T^{-1}\Delta T_\epsilon \|_\infty \| T^{-1}\mathbf{e} \|_\infty) \\ &= \| \mathbf{z} \|_\infty \left(1 + O(N^{3-\alpha}) \right). \end{aligned} \tag{4.7}$$

Similarly,

$$\| \mathbf{y}_\epsilon \|_1 = \| (T + \Delta T_\epsilon)^{-1}\mathbf{e} \|_1 = \| \mathbf{z} \|_1 \left(1 + O(N^{3-\alpha}) \right). \tag{4.8}$$

Since $\mathbf{y}_\epsilon = (T + \Delta T_\epsilon)^{-1}\mathbf{e} \geq 0$, this also shows that $\mathbf{e}^T\mathbf{y}_\epsilon = O(N^3)$, and hence

$$1 - \epsilon\mathbf{e}^T\mathbf{y}_\epsilon = 1 + O(N^{3-\alpha}). \tag{4.9}$$

Using (4.5), this gives

$$\mathbf{z}_\epsilon = \left(1 + O(N^{3-\alpha}) \right) \mathbf{y}_\epsilon.$$

So, from (4.7)

$$\| \mathbf{z}_\epsilon \|_\infty = \| \mathbf{z} \|_\infty \left(1 + O(N^{3-\alpha}) \right)$$

and from (4.8)

$$\| \mathbf{z}_\epsilon \|_1 = \| \mathbf{z} \|_1 \left(1 + O(N^{3-\alpha}) \right),$$

as required. ∎

4.3. Proofs of Theorems 2.2 and 2.3. We begin this subsection by discussing the main ideas in the proofs of Theorems 2.2 and 2.3 and introducing some notation before proving a lemma that formalizes the key steps.

Note that by using the Sherman–Morrison formula to establish (4.5) we have essentially reduced the problem to the study of $\mathbf{y}_\epsilon$ in (4.4). This system may be written as

$$\left(\frac{1}{\Delta x^2}\text{tridiag}(1,-2,1) - 2\epsilon N^3 \text{tridiag}(\tfrac{1}{2},0,\tfrac{1}{2})\right)\mathbf{y}_\epsilon = -2N^2\mathbf{e},$$

where $\Delta x = 1/N$. This may be interpreted as a finite difference formula applied to the boundary value problem

$$y''(x) - 2\epsilon N^3 y(x) = -2N^2, \quad 0 \le x \le 1, \quad y(0) = y(1) = 0. \tag{4.10}$$

The finite difference formula applies standard central differences to the $y''(x)$ term but uses slightly unusual symmetric averaging for the $y(x)$ term. The boundary value problem (4.10) has solution

$$y_{\text{bvp}}(x) = \frac{1}{\epsilon N}\left[1 - \frac{\cosh\gamma(x-\frac{1}{2})}{\cosh\frac{\gamma}{2}}\right], \tag{4.11}$$

where $\gamma := \sqrt{2\epsilon N^3}$. We let $\mathbf{y}_{\text{bvp}} \in \mathbb{R}^{N-1}$ denote the vector whose ith component is given by $y_{\text{bvp}}(x_i)$, where $x_i = i\Delta x$.

If the finite difference method is successful, then we would expect $\mathbf{y}_{\text{bvp}}$ to form a good approximation to $\mathbf{y}_\epsilon$, and this is the basis of our analysis. We note, however, that some care is required since, unlike in the scenario normally studied by numerical analysts, the underlying problem (4.10) depends on the discretization parameter Δx (through N). However, by carefully adapting the traditional M-matrix type analysis (see, for example [21, Chapter 6]) and exploiting the special structure of the problem, it is possible to obtain a useful result. (As an aside, we mention that the original system (4.1) could be analyzed through a finite difference framework by regarding $\mathbf{e}\mathbf{e}^T$ as approximating a scaled integral operator. However, we found it more convenient to invoke Sherman–Morrison.)

To proceed, we therefore define the truncation error vector $\boldsymbol{\tau} \in \mathbb{R}^{N-1}$ by

$$\begin{aligned}\tau_i := {} & \frac{1}{\Delta x^2}\left[y_{\text{bvp}}(x_i - \Delta x) - 2y_{\text{bvp}}(x_i) + y_{\text{bvp}}(x_i + \Delta x)\right] \\ & - \frac{\gamma^2}{2}\left[y_{\text{bvp}}(x_i - \Delta x) + y_{\text{bvp}}(x_i + \Delta x)\right] + 2N^2.\end{aligned} \tag{4.12}$$

Equivalently, we may write

$$\widehat{T}_\epsilon \mathbf{y}_{\text{bvp}} = \mathbf{e} - \frac{\Delta x^2}{2}\boldsymbol{\tau}. \tag{4.13}$$

LEMMA 4.1. *Suppose $1 < \alpha \le 3$. Then the truncation error $\boldsymbol{\tau}$ satisfies $\tau_i > 0$ for all i (for sufficiently large N) with*

$$\|\boldsymbol{\tau}\|_\infty = O(N^{3-\alpha}) \quad \text{and} \quad \|\boldsymbol{\tau}\|_1 = O(N^{\frac{5-\alpha}{2}}). \tag{4.14}$$

Further,

$$\|\mathbf{y}_\epsilon - \mathbf{y}_{\text{bvp}}\|_\infty = O(1) \tag{4.15}$$

and

$$1 - \epsilon \mathbf{e}^T \mathbf{y}_\epsilon = \frac{2 \tanh \frac{\gamma}{2}}{\gamma} + O(N^{-1}). \tag{4.16}$$

Proof. It follows from (4.11) that

$$\| \mathbf{y}_{\rm bvp} \|_\infty = O\left(\frac{1}{\epsilon N} \right) = O(N^{\alpha-1}). \tag{4.17}$$

Also

$$\mathbf{e}^T \mathbf{y}_{\rm bvp} = \frac{N-1}{\epsilon N} - \frac{2}{\epsilon N} \frac{\sum_{i=1}^{N-1} e^{\gamma(i\Delta x - \frac{1}{2})}}{e^{\gamma/2} + e^{-\gamma/2}}.$$

By summing the geometric series and exploiting the fact that $\gamma = O(N^{(3-\alpha)/2})$ we find

$$\mathbf{e}^T \mathbf{y}_{\rm bvp} = \frac{1}{\epsilon} - \frac{2 \tanh \frac{\gamma}{2}}{\epsilon \gamma} + O(N^{\alpha-1}). \tag{4.18}$$

(Note that for $1 < \alpha < 3$ the $\tanh \frac{\gamma}{2}$ factor in (4.18) may be replaced by 1.)

To estimate $\boldsymbol{\tau}$, we note that since $y_{\rm bvp} \in C^4[0,1]$, Taylor expansions give

$$y_{\rm bvp}(x_i - \Delta x) - 2y_{\rm bvp}(x_i) + y_{\rm bvp}(x_i + \Delta x) = \Delta x^2 y''_{\rm bvp}(x_i) + \frac{\Delta x^4}{4!} \left[y^{\rm IV}_{\rm bvp}(\xi_i^1) + y^{\rm IV}_{\rm bvp}(\xi_i^2) \right],$$

$$y_{\rm bvp}(x_i - \Delta x) + y_{\rm bvp}(x_i + \Delta x) = 2y_{\rm bvp}(x_i) + \frac{\Delta x^2}{2} \left[y''_{\rm bvp}(\zeta_i^1) + y''_{\rm bvp}(\zeta_i^2) \right],$$

where $\xi_i^1, \zeta_i^1 \in [x_{i-1}, x_i]$ and $\xi_i^2, \zeta_i^2 \in [x_i, x_{i+1}]$. It follows from (4.12) that

$$\tau_i = \frac{\Delta x^2}{4!} \left[y^{\rm IV}_{\rm bvp}(\xi_i^1) + y^{\rm IV}_{\rm bvp}(\xi_i^2) \right] - \frac{\gamma^2 \Delta x^2}{4} \left[y''_{\rm bvp}(\zeta_i^1) + y''_{\rm bvp}(\zeta_i^2) \right].$$

Now, since $y^{\rm IV}_{\rm bvp}(x) = \gamma^2 y''_{\rm bvp}(x)$, we have

$$\tau_i = \frac{\gamma^2 \Delta x^2}{4} \left[\frac{y''_{\rm bvp}(\xi_i^1) + y''_{\rm bvp}(\xi_i^2)}{6} - \left(y''_{\rm bvp}(\zeta_i^1) + y''_{\rm bvp}(\zeta_i^2) \right) \right]. \tag{4.19}$$

Another Taylor expansion gives

$$|y''_{\rm bvp}(\xi_i^1) - y''_{\rm bvp}(x_i)| \le \Delta x |y'''(\xi_i^{1,1})| \le \gamma \Delta x |y''(\xi_i^{1,1})|$$

for some $\xi_i^{1,1} \in [x_{i-1}, x_i]$, and thus

$$\begin{aligned} |y''_{\rm bvp}(\xi_i^1) - y''_{\rm bvp}(x_i)| &\le \gamma \Delta x \left(|y''_{\rm bvp}(x_i)| + \Delta x |y'''(\xi_i^{1,2})| \right) \\ &\le \gamma \Delta x \left(|y''_{\rm bvp}(x_i)| + \gamma \Delta x |y''(\xi_i^{1,2})| \right) \end{aligned}$$

for some $\xi_i^{1,2} \in [x_{i-1}, x_i]$. Continuing this argument we find

$$\begin{aligned} |y''_{\rm bvp}(\xi_i^1) - y''_{\rm bvp}(x_i)| &\le |y''_{\rm bvp}(x_i)| \sum_{k=1}^{l} (\gamma \Delta x)^k + (\gamma \Delta x)^{l+1} \max_{[x_{i-1}, x_i]} |y''_{\rm bvp}(x)| \\ &\le |y''_{\rm bvp}(x_i)| \frac{\gamma \Delta x}{1 - \gamma \Delta x} + (\gamma \Delta x)^{l+1} \max_{[x_{i-1}, x_i]} |y''_{\rm bvp}(x)| \end{aligned} \tag{4.20}$$

for any $l \geq 1$. By taking l sufficiently large, we can make the second term in (4.20) negligible, and hence

$$y''_{\rm bvp}(\xi_i^1) = y''_{\rm bvp}(x_i)\,(1 + O(\gamma\Delta x))\,.$$

Similarly, this expansion holds for $y''_{\rm bvp}(\xi_i^2)$, $y''_{\rm bvp}(\zeta_i^1)$, and $y''_{\rm bvp}(\zeta_i^2)$, so, in (4.19),

$$\tau_i = \gamma^2\Delta x^2 y''_{\rm bvp}(x_i)\left(\frac{-5}{12} + O(\gamma\Delta x)\right).$$

Since $y''_{\rm bvp}(x_i) < 0$ for all i, the positivity of τ_i follows. Using $\max_{[0,1]}|y''_{\rm bvp}(x)| = O(\gamma^2/(\epsilon N)) = O(N^2)$ we then find that $\|\,\boldsymbol{\tau}\,\|_\infty = O(N^{3-\alpha})$.

To bound $\|\,\boldsymbol{\tau}\,\|_1$ we note from (4.11) that

$$\sum_{i=0}^{N}|y''_{\rm bvp}(x_i)| = \gamma^2\sum_{i=0}^{N}\left[\frac{1}{\epsilon N} - y_{\rm bvp}(x_i)\right] \leq \frac{\gamma^2(N+1)}{\epsilon N} + \gamma^2\|\,\mathbf{y}_{\rm bvp}\,\|_1.$$

From (4.18) we have $\|\,\mathbf{y}_{\rm bvp}\,\|_1 = O(N^{3(\alpha-1)/2})$, so

$$\sum_{i=0}^{N}|y''_{\rm bvp}(x_i)| = O(N^{(3+\alpha)/2}). \tag{4.21}$$

Since $|y''_{\rm bvp}(x)|$ takes its extreme value over $[x_i, x_{i+1}]$ at an endpoint, we have, from (4.19) and (4.21),

$$\|\,\boldsymbol{\tau}\,\|_1 \leq \frac{\gamma^2\Delta x^2}{4}\left(\frac{1}{6}+\frac{1}{6}+1+1\right)\sum_{i=0}^{N}|y''_{\rm bvp}(x_i)| = O(N^{(5-\alpha)/2}).$$

Now from (2.5) and (4.4) we have $\mathbf{z} - \mathbf{y}_\epsilon = T^{-1}\Delta T_\epsilon\mathbf{y}_\epsilon$. We know that $T^{-1} \geq 0$, $\Delta T_\epsilon \geq 0$, and $\mathbf{y}_\epsilon \geq 0$ (because $T + \Delta T_\epsilon$ is Stieltjes). Hence $\mathbf{z} - \mathbf{y}_\epsilon \geq 0$, that is,

$$T^{-1}\mathbf{e} \geq \widehat{T}_\epsilon^{\,-1}\mathbf{e}. \tag{4.22}$$

Then from (4.4) and (4.13) we have

$$\mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} = \frac{\Delta x^2}{2}\widehat{T}_\epsilon^{\,-1}\boldsymbol{\tau}, \tag{4.23}$$

so, using (4.22),

$$\begin{aligned}|\mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp}| &\leq \frac{\Delta x^2}{2}\|\,\boldsymbol{\tau}\,\|_\infty\widehat{T}_\epsilon^{\,-1}\mathbf{e}\\ &\leq \frac{\Delta x^2}{2}\|\,\boldsymbol{\tau}\,\|_\infty T^{-1}\mathbf{e}.\end{aligned} \tag{4.24}$$

Hence, using (4.6) and (4.14),

$$\|\,\mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp}\,\|_\infty \leq \frac{\Delta x^2}{2}\|\,\boldsymbol{\tau}\,\|_\infty\|\,T^{-1}\mathbf{e}\,\|_\infty = O(N^{3-\alpha}). \tag{4.25}$$

We now refine this bound for $\alpha < 3$. From (4.17) and (4.25) we have

$$\|\,\widehat{T}_\epsilon^{\,-1}\mathbf{e}\,\|_\infty = \|\,\mathbf{y}_\epsilon\,\|_\infty \leq \|\,\mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp}\,\|_\infty + \|\,\mathbf{y}_{\rm bvp}\,\|_\infty = O(N^{3-\alpha}) + O(N^{\alpha-1}). \tag{4.26}$$

For $2 \leq \alpha < 3$ the $O(N^{\alpha-1})$ term dominates, and so after taking norms in (4.24) we have

$$\| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty = O(N^{-2} N^{3-\alpha} N^{\alpha-1}) = O(N^0). \tag{4.27}$$

For $1 < \alpha < 2$, in (4.26) we have $\| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty = O(N^{3-\alpha})$. Using this in (4.24) gives

$$\| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty = O(N^{-2} N^{3-\alpha} N^{3-\alpha}) = O(N^{4-2\alpha}).$$

Hence,

$$\| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty \leq \| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty + \| \mathbf{y}_{\rm bvp} \|_\infty = O(N^{4-2\alpha}) + O(N^{\alpha-1}). \tag{4.28}$$

For $\alpha \geq 5/3$, the $O(N^{\alpha-1})$ term dominates and we may use (4.24) to recover (4.27). For $1 < \alpha < 5/3$, in (4.28) we have $\| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty = O(N^{4-2\alpha})$. Using this in (4.24) gives

$$\| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty = O(N^{-2} N^{3-\alpha} N^{4-2\alpha}) = O(N^{5-3\alpha}).$$

Hence

$$\| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty \leq \| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty + \| \mathbf{y}_{\rm bvp} \|_\infty = O(N^{5-3\alpha}) + O(N^{\alpha-1}). \tag{4.29}$$

For $\alpha \geq 6/4$, the $O(N^{\alpha-1})$ term dominates and we may use (4.24) to recover (4.27). For $1 < \alpha < 6/4$, in (4.29) we have $\| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty = O(N^{5-3\alpha})$. Using this in (4.24) gives

$$\| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty = O(N^{-2} N^{3-\alpha} N^{5-3\alpha}) = O(N^{6-4\alpha}).$$

Hence

$$\| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty \leq \| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty + \| \mathbf{y}_{\rm bvp} \|_\infty = O(N^{6-4\alpha}) + O(N^{\alpha-1}). \tag{4.30}$$

For $\alpha \geq 7/5$, the $O(N^{\alpha-1})$ term dominates and we may use (4.24) to recover (4.27).

The pattern is now clear. Given any integer $k \geq 1$ we can establish $\| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty = O(N^0)$ for $(k+2)/k \leq \alpha \leq 3$, which confirms (4.15).

From (4.14), (4.15), (4.17), and (4.23) we have

$$\begin{aligned}
\| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_1 &\leq \frac{\Delta x^2}{2} \| \widehat{T}_\epsilon^{-1} \|_1 \| \boldsymbol{\tau} \|_1 \\
&= \frac{\Delta x^2}{2} \| \widehat{T}_\epsilon^{-1} \mathbf{e} \|_\infty \| \boldsymbol{\tau} \|_1 \\
&= \frac{\Delta x^2}{2} \| \mathbf{y}_\epsilon \|_\infty \| \boldsymbol{\tau} \|_1 \\
&\leq \frac{\Delta x^2}{2} \left[\| \mathbf{y}_{\rm bvp} \|_\infty + \| \mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} \|_\infty \right] \| \boldsymbol{\tau} \|_1 \\
&= O(N^{-2} N^{\alpha-1} N^{(5-\alpha)/2}) \\
&= O(N^{(\alpha-1)/2}).
\end{aligned}$$

Since $\mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp} = \frac{\Delta x^2}{2} \widehat{T}_\epsilon^{-1} \boldsymbol{\tau} \geq 0$, this is equivalent to

$$\mathbf{e}^T (\mathbf{y}_\epsilon - \mathbf{y}_{\rm bvp}) = O(N^{(\alpha-1)/2}),$$

which gives, using (4.18),

$$1 - \epsilon \mathbf{e}^T \mathbf{y}_\epsilon = 1 - \epsilon \mathbf{e}^T \mathbf{y}_{\text{bvp}} + O(N^{(-\alpha-1)/2}) = \frac{2 \tanh \frac{\gamma}{2}}{\gamma} + O(N^{-1}),$$

completing the proof. ∎

We are now in a position to prove Theorems 2.2 and 2.3.

Proof of Theorems 2.2 *and* 2.3. From (4.5), (4.15), (4.16), and (4.17) we obtain

$$\| \mathbf{z}_\epsilon \|_\infty = \frac{\gamma}{2 \tanh \frac{\gamma}{2}} \| \mathbf{y}_{\text{bvp}} \|_\infty + O(N).$$

For $1 < \alpha < 3$, $\| \mathbf{y}_{\text{bvp}} \|_\infty$ equals $y_{\text{bvp}}(\frac{1}{2})$ plus exponentially small terms, and for $\alpha = 3$, $\| \mathbf{y}_{\text{bvp}} \|_\infty = y_{\text{bvp}}(\frac{1}{2}) + O(1)$. So

$$\| \mathbf{z}_\epsilon \|_\infty = \frac{\gamma}{2 \tanh \frac{\gamma}{2}} y_{\text{bvp}}(\tfrac{1}{2}) + O(N)$$

for $1 < \alpha \leq 3$. This simplifies to

$$\| \mathbf{z}_\epsilon \|_\infty = \frac{\gamma \tanh \frac{\gamma}{4}}{2\epsilon N} + O(N).$$

Using $\| \mathbf{z} \|_\infty = N^2/4 + O(1)$ we thus have

$$\frac{\| \mathbf{z}_\epsilon \|_\infty}{\| \mathbf{z} \|_\infty} = \frac{2\gamma \tanh \frac{\gamma}{4}}{\epsilon N^3} + O(N^{-1}). \tag{4.31}$$

For $\alpha = 3$ we have $\epsilon = KN^{-3}$ and $\gamma = \sqrt{2K}$. Inserting this into (4.31) gives (2.13). For $1 < \alpha < 3$ we have $\gamma = \sqrt{2K} N^{(3-\alpha)/2}$ and (2.15) follows.

For the 1-norm result, we first note that $\mathbf{y}_\epsilon \geq 0$, $\mathbf{y}_{\text{bvp}} \geq 0$ and $\mathbf{y}_\epsilon - \mathbf{y}_{\text{bvp}} \geq 0$, so that

$$\| \mathbf{y}_\epsilon \|_1 = \| \mathbf{y}_{\text{bvp}} \|_1 + \| \mathbf{y}_\epsilon - \mathbf{y}_{\text{bvp}} \|_1 = \| \mathbf{y}_{\text{bvp}} \|_1 + O(N \| \mathbf{y}_\epsilon - \mathbf{y}_{\text{bvp}} \|_\infty).$$

Using (4.5), (4.15), (4.16), and (4.18) we find

$$\| \mathbf{z}_\epsilon \|_1 = \frac{\gamma}{2\epsilon \tanh \frac{\gamma}{2}} - \frac{1}{\epsilon} + O(N^2).$$

Scaling by $\| \mathbf{z} \|_1 = \frac{(N-1)N(N+1)}{6}$ and inserting $\gamma = \sqrt{2\epsilon N^3}$ gives the estimates (2.14) and (2.16). ∎

5. Final remarks. Our aim in this work was to show via matrix perturbation theory that the small world phenomenon arises in the context of Markov chains. The results are fully rigorous, with sharp error estimates that vanish as the system size increases.

There are, of course, many ways in which the Markov chain model may be extended or altered. The two most obvious directions are perhaps moving to higher dimensions and considering more complex underlying lattice topologies. Further, instead of giving equal weight to all nonneighboring states we could, for example, introduce range-dependent perturbations to the transition matrix of the form $f(|i-j|)$ for some suitable function f. Grindrod [7, 8] has recently produced some elegant results for analogous network models. Alternatively, we could perturb only a small, fixed number of zeros in (2.4). We note that Liu, Strang, and Ott [13] have characterized the effect of this type of modification on the spectrum of a structured matrix. In all cases, the techniques developed here form a useful starting point for further analysis.

Acknowledgments. I thank Neal Madras, Jeff Rosenthal, Tony Shardlow, Andrew Stuart, and the referees for valuable input.

REFERENCES

[1] A. D. Barbour and G. Reinert, *Small worlds*, Random Structures Algorithms, 19 (2001), pp. 54–74.
[2] E. Behrends, *Introduction to Markov Chains with Special Emphasis on Rapid Mixing*, Vieweg, Braunschweig, 2000.
[3] H. C. Berg, *Random Walks in Biology*, Princeton University Press, Princeton, NJ, 1983.
[4] H. Caswell, *Matrix Population Models*, 2nd ed., Sinauer Associates, Sunderland, MA, 2001.
[5] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[6] P. M. Gleiss, P. F. Stadler, A. Wagner, and D. A. Fell, *Relevant cycles in chemical reaction networks*, Adv. Complex Systems, 4 (2001), pp. 207–226.
[7] P. Grindrod, *Range-dependent random graphs and their application to modeling large small-world proteome datasets*, Phys. Rev. E, 66 (2002), article 066702.
[8] P. Grindrod, *Modeling proteome networks with range-dependent graphs*, Amer. J. Pharmaco Genomics, 3 (2003), pp. 1–4.
[9] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[10] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, *Extrapolation methods for accelerating pagerank computations*, in Proceedings of the Twelfth International World Wide Web Conference, Budapest, Hungary, 2003.
[11] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, San Diego, 1975.
[12] J. Kleinberg, *Navigation in a small world*, Nature, 406 (2000), p. 845.
[13] X. Liu, G. Strang, and S. Ott, *Localized eigenvectors from widely spaced matrix modifications*, SIAM J. Discrete Math., 16 (2003), pp. 479–498.
[14] C. Moler, *The world's largest matrix computation*, MATLAB News and Notes, October 2002.
[15] J. M. Montoya and R. V. Solé, *Small world patterns in food webs*, J. Theoret. Biology, 214 (2002), pp. 405–412.
[16] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, UK, 1995.
[17] M. E. J. Newman, *The structure of scientific collaboration networks*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 404–409.
[18] M. E. J. Newman, *The structure and function of networks*, Comput. Phys. Comm., 147 (2002), pp. 40–45.
[19] M. E. J. Newman, C. Moore, and D. J. Watts, *Mean-field solution of the small-world network model*, Phys. Rev. Lett., 84 (2000), pp. 3201–3204.
[20] J. R. Norris, *Markov Chains*, Cambridge University Press, Cambridge, UK, 1997.
[21] J. M. Ortega, *Numerical Analysis: A Second Course*, SIAM, Philadelphia, 1990.
[22] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.
[23] A. Wagner and D. A. Fell, *The small world inside large metabolic networks*, Proc. Roy. Soc. London Ser. B., 268 (2001), pp. 1803–1810.
[24] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ, 1999.
[25] D. J. Watts and S. H. Strogatz, *Collective dynamics of "small-world" networks*, Nature, 393 (1998), pp. 440–442.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 445–463

RECURSIVE CALCULATION OF DOMINANT SINGULAR SUBSPACES*

Y. CHAHLAOUI†, K. GALLIVAN‡, AND P. VAN DOOREN†

Abstract. In this paper we show how to compute recursively an approximation of the left and right dominant singular subspaces of a given matrix. In order to perform as few as possible operations on each column of the matrix, we use a variant of the classical Gram–Schmidt algorithm to estimate this subspace. The method is shown to be particularly suited for matrices with many more rows than columns. Bounds for the accuracy of the computed subspace are provided. Moreover, the analysis of error propagation in this algorithm provides new insights in the loss of orthogonality typically observed in the classical Gram–Schmidt method.

Key words. singular value decomposition, Gram–Schmidt, dominant subspace

AMS subject classifications. 15A18, 15A42, 65Y20

DOI. 10.1137/S0895479803374657

1. Introduction.

In many problems one needs to compute the projector on the dominant subspace of a given data matrix A of dimension $m \times n$. The type of application we are thinking of here implies $m \gg n$, and for the sake of simplicity we will assume A to be real. In addition, we assume that the matrix A is produced incrementally, so all of the columns are not available simultaneously. Several applications have this property. For example, approximating a matrix A in which each column represents an image of a given sequence amounts to an SVD-based compression [5]. Such an approximation is also used in the context of observation-based model reduction for dynamical systems. The so-called proper orthogonal decomposition (POD) approximation uses the dominant left space of a matrix A where a column consists of a time instance of the solution of an evolution equation, e.g., the flow field from a fluid dynamics simulation. Since these flow fields tend to be very large only a small number can be stored efficiently during the simulation, and therefore an incremental approach is useful [11]. Finally, the dominant space approximation is also used in text retrieval to encode document/term information and avoid certain types of semantic noise. The incremental form is required when documents are added or when the entire matrix is not available at one point in time and space [3].

In each of these applications, one can interpret the columns of the matrix A as "data vectors" with some "energy" equal to their 2-norm. Finding the dominant space of dimension $k < \min(m, n)$ amounts to finding the k first columns of the matrix U

*Received by the editors June 29, 2000; accepted for publication (in revised form) by J. M. Hyman March 3, 2003; published electronically September 9, 2003. This paper presents research supported by NSF contract CCR-99-12415 and by the Belgian Programme on Inter-University Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors.

http://www.siam.org/journals/simax/25-2/37465.html

†Department of Mathematical Engineering, Université Catholique de Louvain, CESAME, avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium (chahlaoui@csam.ucl.ac.be, vdooren@csam.ucl.ac.be). The work of the first author was partially carried out within the framework of a collaboration agreement between CESAME (Université Catholique de Louvain, Belgium) and LINMA of the Faculty of Sciences (Université Chouaib Doukkali, Morocco), funded by the Secretary of the State for Development Cooperation and by the CIUF (Conseil Interuniversitaire de la Communauté Française, Belgium).

‡School of Computational Science and Information Technology, Florida State University, Tallahassee, FL 32306 (gallivan@csit.fsu.edu).

in the singular value decomposition of A:

$$A = U\Sigma V^T, \quad U^TU = I_n, \quad VV^T = V^TV = I_n, \quad \Sigma = \mathrm{diag}\{\sigma_1, \ldots \sigma_n\}, \tag{1.1}$$

and where the diagonal elements σ_i of Σ are nonnegative and nonincreasing. This decomposition in fact expresses that the orthogonal transformation V applied to the columns of A yields a new matrix $AV = U\Sigma$ with orthogonal columns of nonincreasing norm. The "dominant" columns of this transformed matrix are obviously the k leading ones. A block version of this decomposition makes this more explicit:

$$A = U\Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_{1,1} & \\ & \Sigma_{2,2} \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T, \tag{1.2}$$

where U_1 and V_1 have k columns and $\Sigma_{1,1}$ is $k \times k$. An orthogonal basis for the corresponding space is then given by U_1, which is also equal to $AV_1\Sigma_{1,1}^{-1}$. The cost of this decomposition including the construction of U is $14mn^2 + O(n^3)$. For an additional $O(n^3)$ operations it is also possible to compute an orthogonal basis for the columns of V_1, which is required in several applications.

A cheaper procedure is to first perform a QR decomposition of A, followed by a singular value decomposition of the smaller matrix R [4]:

$$A = QR, \quad R = U\Sigma V^T. \tag{1.3}$$

From these equations it is easy to see that $AV = QU\Sigma$, and again this has orthogonal columns of nonincreasing norms. This decomposition costs typically $6mn^2 + O(n^3)$ [8]. It is even more economical to use the normal equations (or covariance matrix) of A. Its eigenvalue decomposition gives

$$A^TA = V\Lambda V^T, \tag{1.4}$$

and comparing this with (1.1) shows that the same matrix V is constructed and that

$$(AV)^T(AV) = \Lambda = \Sigma^T\Sigma.$$

This algorithm requires mn^2 operations to construct A^TA and $mnk + O(n^3)$ operations to obtain $U_1 = AV_1\Sigma_{1,1}^{-1}$. Unfortunately, using the covariance matrix is not recommended because it is more sensitive to rounding errors [8].

In this paper we consider applications where m is huge, and where every column operation on A or on the basis U not only is costly in operations but also involves swapping data from the main memory, which will slow down the algorithm significantly. We present an algorithm that yields an approximate decomposition but requires only $8mnk + O(nk^3)$ operations and also works recursively on the columns of A; i.e., the columns of A (or data vectors) can be produced recursively and A need not be stored in its entirety.

The paper is organized as follows. In sections 2 and 3 we derive an economical sequential procedure to approximate a matrix A by a low-rank factorization. In section 4 we derive bounds for the residual error and compare our method with the "optimal" singular value decomposition approach. In section 5 we illustrate these bounds via numerical experiments. In section 6 we study the effect of round-off and prove backward stability as well as preservation of orthogonality of our computed basis vectors under some mild conditions. This surprising feature (of a classical Gram–Schmidt-like method) is explained and illustrated numerically in the last section.

2. A recursive procedure. In this section we propose a recursive procedure to estimate the dominant subspace of a given matrix A using a sequential (and incremental) processing of the columns of A. Bounds for the accuracy of this decomposition are derived later. The algorithm is based on an efficient calculation of the dominant k-dimensional space of an $m \times (k+1)$ matrix M. Assume that a QR decomposition of M is available:

$$M = QR. \tag{2.1}$$

Then compute the smallest singular vector u_{k+1} of R (i.e., $Rv_{k+1} = u_{k+1}\mu_{k+1}$) and construct an orthogonal transformation G_u such that $G_u^T u_{k+1} = e_{k+1}$. Now apply G_u^T to the rows of R and let G_v be an orthogonal transformation putting $G_u^T R$ back in triangular form:

$$G_u^T R G_v = R_{up}.$$

In this new coordinate system the right singular vector u_{k+1} becomes e_{k+1}, a unit vector with 1 in the $(k+1)$ element, and v_{k+1} is transformed to a new vector $\hat{v}_{k+1}$. Therefore,

$$R_{up} e_{k+1} = \mu_{k+1} \hat{v}_{k+1}, \quad R_{up}^T \hat{v}_{k+1} = \mu_{k+1} e_{k+1}.$$

It easily follows that R_{up} has the form

$$R_{up} = \begin{bmatrix} R_{1,1} & 0 \\ 0 & \mu_{k+1} \end{bmatrix}. \tag{2.2}$$

We therefore have the updated QR decomposition

$$MG_v = Q_{up} R_{up} = (QG_u)(G_u^T R G_v),$$

and since R_{up} has the required block form (1.2) we have found a basis for the dominant k-dimensional subspace of M in the form of the first k columns of Q_{up}.

Both matrices G_u and G_v can be constructed as a product of k 2×2 Givens transformations, allowing an elegant update of R using only $O(k^2)$ operations. But the costly part of the algorithm is the update of Q, and hence it is preferable to choose G_u to be a Householder transformation. When retriangularizing $G_u^T R$ one then needs to perform again a QR factorization, which requires $O(k^3)$ operations, but since $k < n \ll m$, this is of no concern. The cost of the update of Q to Q_{up} is that of a Householder transformation applied to an $m \times (k+1)$ matrix and is thus $4m(k+1)$ operations. The vector u_{k+1} can be computed with a few steps of inverse iteration or with a shifted inverse iteration. The cost of this calculation as well as the update of R is thus $O(k^3)$ and hence negligible with respect to the update of Q. A more involved technique uses modified Givens transformations since their complexity is the same as that of Householder transformations for the product QG_u, and is of $O(k^2)$ when used for forming the product $G_u^T R G_v$. Unfortunately, this requires storing and updating additional diagonal scaling matrices, which typically hurt the performance of codes used for parallel machines.

How is this now applied to finding the dominant subspace of A? We start with a QR factorization of the first k columns of A:

$$A(:, 1\colon k) = Q_{(k)} R_{(k)}. \tag{2.3}$$

Then we recursively apply the following update and downdate of this decomposition. For $i = k+1$ to n, append the next column $a_i \doteq A(:,i)$ to the current matrix decomposition and perform a QR decomposition of it. The formulas for this are standard. Define $r_i = Q_{(i-1)}^T a_i$; then $\hat{a}_i \doteq a_i - Q_{(i-1)} r_i$ is orthogonal to $Q_{(i-1)}$. Define ρ_i as its norm, and $\hat{q}_i = \hat{a}_i/\rho_i$. Then

$$\begin{bmatrix} Q_{(i-1)}R_{(i-1)} & a_i \end{bmatrix} = \begin{bmatrix} Q_{(i-1)} & \hat{q}_i \end{bmatrix} \begin{bmatrix} R_{(i-1)} & r_i \\ 0 & \rho_i \end{bmatrix}. \tag{2.4}$$

Update this matrix decomposition to "deflate" its smallest singular value as above,

$$\begin{bmatrix} Q_{(i-1)} & \hat{q}_i \end{bmatrix} G_u \cdot G_u^T \begin{bmatrix} R_{(i-1)} & r_i \\ 0 & \rho_i \end{bmatrix} G_v = \begin{bmatrix} Q_{(i)} & q_i \end{bmatrix} \cdot \begin{bmatrix} R_{(i)} & 0 \\ 0 & \mu_i \end{bmatrix}, \tag{2.5}$$

and delete the last columns to obtain the new $Q_{(i)}$ and $R_{(i)}$. The complexity of this algorithm is $10mkn + O((n-k)k^3)$ when using Givens transformations for G_u and $8mkn + O((n-k)k^3)$ when using a Householder transformation or modified Givens transformations for G_u. This is clearly cheaper than all earlier algorithms if $m \gg n \gg k$.

The algorithm thus computes at each step a decomposition that "deflates" the smallest singular vector of the current $m \times (k+1)$ matrix and then appends to it the next column of A. All columns of A therefore are passed through once and compared with the current best estimate of this dominant subspace. At first sight this is a very heuristic algorithm, but in the next section we show that quite good bounds can be obtained for the quality of this basis.

REMARK 2.1. *Although we do not consider in this paper the updating problem to dimension $k+l$ for $l > 1$, it can be done in a very similar manner. If appropriately implemented, this "block" version still has $\theta(mkn)$ complexity. Convergence results are essentially the same and good performance can be expected on parallel architectures (see also* [2]*).*

3. Updating a two-sided decomposition. The algorithm above yields at step i an approximation $Q_{(i)}$ of the dominant left singular subspace of $A(:, 1:i)$, but in several applications it makes sense to update simultaneously an approximation of the corresponding right singular subspace of this matrix. This can be done with little extra cost.

We start from the notation introduced in (2.3), which we rewrite as

$$A(:, 1:k)V_{(k)} = Q_{(k)}R_{(k)}, \tag{3.1}$$

where $V_{(k)} = I_k$. We show by induction that at each step $i \geq k$ we have a decomposition

$$A(:, 1:i)V_{(i)} = Q_{(i)}R_{(i)}, \tag{3.2}$$

where $V_{(i)} \in \mathbb{R}^{i \times k}$ satisfies $V_{(i)}^T V_{(i)} = I_k$. From (3.1) it is obvious that this holds for $i = k$. For the induction step we start by assuming that it holds for $i-1$:

$$A(:, 1:(i-1))V_{(i-1)} = Q_{(i-1)}R_{(i-1)}.$$

We then append a column a_i to $A(:, 1:i-1)$ to get $A(:, 1:i)$ and obviously

$$A(:, 1:i) \begin{bmatrix} V_{(i-1)} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} Q_{(i-1)}R_{(i-1)} & a_i \end{bmatrix}. \tag{3.3}$$

Now use (2.4), (2.5) to update this to

$$A(:,1\colon i)\begin{bmatrix} V_{(i-1)} & 0 \\ 0 & 1 \end{bmatrix} G_v = \begin{bmatrix} Q_{(i)}R_{(i)} & q_i\mu_i \end{bmatrix}. \tag{3.4}$$

Taking the first k columns of both sides of this equation yields (3.1) with

$$V_{(i)} = \begin{bmatrix} V_{(i-1)} & 0 \\ 0 & 1 \end{bmatrix} G_v \begin{bmatrix} I_k \\ 0 \dots 0 \end{bmatrix} \in \mathbb{R}^{i\times k}, \tag{3.5}$$

which obviously satisfies $V_{(i)}^T V_{(i)} = I_k$. The additional work for updating the approximation $V_{(i)}$ is just the multiplication (3.5), which requires $6ik$ flops and hence leads to a total of

$$\sum_{i=k}^{n} 6ik \approx 3k(n+k)(n-k+1)$$

additional flops for the full decomposition. This additional work can be neglected if $m \gg k$.

We terminate this section by writing a decomposition for the matrix $A(:,1:i)$ if we would not delete the last column at each step. There exists an orthogonal matrix $V_i \in \mathbb{R}^{i\times i}$ embedding $V_{(i)}$:

$$V_i = \begin{bmatrix} V_{(i)} & V_{(i)}^{\perp} \end{bmatrix}.$$

Choosing appropriate basis vectors for $V_{(i)}^{\perp}$, we obtain a decomposition of the type

$$A(:,1\colon i)V_i = \begin{bmatrix} Q_{(i)}R_{(i)} & \tilde{q}_i & \dots & \tilde{q}_n \end{bmatrix}, \tag{3.6}$$

where $\tilde{q}_j = q_j\mu_j$ and $\|\tilde{q}_j\|_2 = \mu_j$. From this we obtain the additive decomposition

$$A(:,1\colon i) = Q_{(i)}R_{(i)}V_{(i)}^T + \begin{bmatrix} \tilde{q}_i & \dots & \tilde{q}_n \end{bmatrix} V_{(i)}^{\perp T}, \tag{3.7}$$

which will be used later on to derive error bounds.

4. Accuracy bounds. It is clear that after the first step $i = k+1$ we obtain a decomposition

$$[A(:,1\colon k{+}1)]\, G_v^T = \begin{bmatrix} Q_{(k+1)} & q_{k+1} \end{bmatrix} \cdot \begin{bmatrix} R_{(k+1)} & 0 \\ 0 & \mu_{k+1} \end{bmatrix}. \tag{4.1}$$

Let σ_i, $i = 1,\dots,n$, be the singular values of A and $\hat{\sigma}_i^{(j)}$, $i = 1,\dots,k$, those of $R(j)$. Then according to the above decomposition, $A(:,1\colon k+1)$ has singular values

$$\hat{\sigma}_1^{(k+1)}, \dots, \hat{\sigma}_k^{(k+1)}, \mu_{k+1}.$$

But since this is a submatrix of A obtained by deleting a number of columns, we have the inequalities [8]

$$\hat{\sigma}_1^{(k+1)} \le \sigma_1, \quad \dots, \quad \hat{\sigma}_k^{(k+1)} \le \sigma_k, \quad \mu_{k+1} \le \sigma_{k+1}. \tag{4.2}$$

Similarly one easily shows that each intermediate matrix

$$\begin{bmatrix} Q_{(i)} & q_i \end{bmatrix} \cdot \begin{bmatrix} R_{(i)} & 0 \\ 0 & \mu_i \end{bmatrix} \tag{4.3}$$

with singular values

$$\hat{\sigma}_1^{(i)}, \ldots, \hat{\sigma}_k^{(i)}, \mu_i$$

is also orthogonally equivalent to a submatrix of A. Therefore we have in general

$$\hat{\sigma}_1^{(i)} \leq \sigma_1, \quad \ldots, \quad \hat{\sigma}_k^{(i)} \leq \sigma_k, \quad \mu_i \leq \sigma_{k+1}. \tag{4.4}$$

Finally, since the matrix

$$\begin{aligned} \left[\begin{array}{cc} A(:, 1\colon (i-1)) & a_i \end{array} \right] &= \left[\begin{array}{cc} Q_{(i-1)} R_{(i-1)} & Q_{(i-1)} r_i + \hat{q}_i \rho_i \end{array} \right] \\ &= \left[\begin{array}{cc} Q_{(i)} & q_i \end{array} \right] \left[\begin{array}{cc} R_{(i)} & 0 \\ 0 & \mu_i \end{array} \right] G_v^T \end{aligned} \tag{4.5}$$

has singular values $\hat{\sigma}_1^{(i)}, \ldots, \hat{\sigma}_k^{(i)}, \mu_i$ and $Q_{(i-1)} R_{(i-1)}$ is its submatrix, we have the inequalities

$$\hat{\sigma}_1^{(i-1)} \leq \hat{\sigma}_1^{(i)}, \quad \ldots \ , \quad \hat{\sigma}_k^{(i-1)} \leq \hat{\sigma}_k^{(i)}. \tag{4.6}$$

All this says that the singular values μ_i that are dismissed at each step are all smaller than σ_{k+1} and that the singular values $\hat{\sigma}_j^{(i)}$, $j = 1, \ldots, k$, that are updated increase monotonically towards the first k singular values of A. To obtain bounds at the end of the iterative procedure we need to relate A to the computed quantities. For this, we point out that there exists an orthogonal column transformation V which relates A and the intermediate results of the recursive algorithm:

$$AV_n = \left[\begin{array}{cccc} Q_{(n)} R_{(n)} & \mu_{k+1} q_{k+1} & \ldots & \mu_n q_n \end{array} \right]. \tag{4.7}$$

The transformation V_n indeed consists of all the smaller transformations G_v and appropriately chosen permutations to obtain (4.7). Using the singular value decomposition of $R_{(n)}$,

$$R_{(n)} = \hat{U}_n \Sigma \hat{V}_n^T,$$

one then constructs orthogonal transformations such that

$$AV_n \left[\begin{array}{cc} \hat{V}_n & 0 \\ 0 & I \end{array} \right] = \left[\begin{array}{cc} Q_{(n)} \hat{U}_n & Q_{(n)}^{\perp} \end{array} \right] \left[\begin{array}{cc} \hat{\Sigma} & A_{1,2} \\ 0 & A_{2,2} \end{array} \right], \tag{4.8}$$

where $Q_{(n)}^{\perp}$ is orthogonal to $Q_{(n)}$ and where the columns of $A_2 \doteq \left[\begin{smallmatrix} A_{1,2} \\ A_{2,2} \end{smallmatrix} \right]$ have 2-norms μ_i. The Frobenius norm of this submatrix is therefore equal to $\| [\mu_{k+1}, \ldots, \mu_n] \|_2$. From (4.8) one already finds a bound for the accuracy of the computed singular values. The singular values of A are also those of $M \doteq \left[\begin{smallmatrix} \hat{\Sigma} & A_{1,2} \\ 0 & A_{2,2} \end{smallmatrix} \right]$. Applying the Wielandt–Hoffman theorem for singular values to this [8] yields

$$\sum_{i=1}^{k} (\sigma_i - \hat{\sigma}_i^{(n)})^2 \leq \|A_2\|_F^2 = \sum_{i=k+1}^{n} (\mu_i)^2 \leq (n-k) \cdot \sigma_{k+1}^2. \tag{4.9}$$

If we know the singular values have a considerable gap $\gamma \doteq \sigma_k - \sigma_{k+1}$, then this bound says that the k largest singular values are well approximated. If γ is large, the space

spanned by the corresponding singular vectors is also insensitive to perturbations. Moreover, one can improve the bounds for the singular value perturbations provided by the Wielandt–Hoffman theorem. To analyze this in more detail we use the following theorem proven in [10].

THEOREM 4.1. *Let $\hat{H}$ and E be square Hermitian matrices partitioned as*

$$\hat{H} = \begin{bmatrix} \hat{H}_{1,1} & 0 \\ 0 & \hat{H}_{2,2} \end{bmatrix}, \quad E = \begin{bmatrix} E_{1,1} & E_{1,2} \\ E_{2,1} & E_{2,2} \end{bmatrix},$$

and define $\epsilon = \|E_{1,2}\|_2$ and $\delta = \min|\lambda(\hat{H}_{1,1}) - \lambda(\hat{H}_{2,2})| - \|E_{1,1}\|_2 - \|E_{2,2}\|_2$.

If $\delta > 2\epsilon$, then there exists a unitary matrix X of the form

$$X = \begin{bmatrix} I_k & -P^T \\ P & I_{n-k} \end{bmatrix} \begin{bmatrix} (I + P^T P)^{-1/2} & 0 \\ 0 & (I + PP^T)^{-1/2} \end{bmatrix}$$

such that

$$H \doteq X^T(\hat{H} + E)X = \begin{bmatrix} H_{1,1} & 0 \\ 0 & H_{2,2} \end{bmatrix},$$

where $\|P\|_2 < 2\epsilon/\delta$.

This theorem is used to estimate the accuracy of both the left and right dominant subspaces of A as follows. Suppose

$$\hat{H}_u = \begin{bmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} \tag{4.10}$$

is the current "approximation" of the eigenvalue decomposition of

$$H_u \doteq MM^T = \begin{bmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \begin{bmatrix} A_{1,2}^T & A_{2,2}^T \end{bmatrix}. \tag{4.11}$$

The left dominant "singular" subspace of M is also the dominant eigensubspace of H_u. The dominant eigensubspace of the nearby matrix $\hat{H}_u$ is clearly $\mathrm{Im}\left[\begin{smallmatrix} I_k \\ 0 \end{smallmatrix}\right]$ and the corresponding eigenvalues are the diagonal elements $\hat{\sigma}_1^{(n)}, \ldots, \hat{\sigma}_k^{(n)}$ of $\hat{\Sigma}^2$. But due to the perturbations $A_{1,2}$ and $A_{2,2}$ these are incorrect. After transforming MM^T to $X_u^T MM^T X_u$ we obtain its true eigenvalues (i.e., the squared singular values of M) in the matrix $H_{1,1}$ and the true dominant subspace as $\mathrm{Im}\left[\begin{smallmatrix} I_k \\ P_u \end{smallmatrix}\right]$. The norm of P_u is a measure for the angular rotation of this subspace, and it is bounded by $2\epsilon_u/\delta_u$. The largest canonical angle θ_k between the spaces $\mathrm{Im}\left[\begin{smallmatrix} I_k \\ 0 \end{smallmatrix}\right]$ and $\mathrm{Im}\left[\begin{smallmatrix} I_k \\ P_u \end{smallmatrix}\right]$ in fact satisfies [10]

$$\cos\theta_k = 1/\sqrt{1 + \|P_u\|^2}, \quad \sin\theta_k = \|P_u\|/\sqrt{1 + \|P_u\|^2}, \quad \tan\theta_k = \|P_u\|$$

and measures the "rotation" of the dominant subspace with respect to its approximation.

Clearly here $\epsilon_u = \|A_{1,2}A_{2,2}^T\|_2$ and $\delta_u = (\hat{\sigma}_k^{(n)})^2 - \|A_{1,1}\|_2^2 - \|A_{2,2}\|_2^2$. Notice that $\|A_2\|_F^2 = \sum_i \mu_i^2$ and that we actually compute these values during our recursive calculations. It would therefore be convenient to bound $2\epsilon_u/\delta_u$ in terms of these "discarded" singular values μ_i. One easily derives the bounds

$$\|A_{1,2}A_{2,2}^T\|_2 \leq \frac{1}{2}\left\| \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \begin{bmatrix} A_{1,2}^T & A_{2,2}^T \end{bmatrix} \right\|_2 = \frac{1}{2}\left\| \underbrace{\begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix}}_{A_2} \right\|_2^2$$

and

$$\|A_2\|_2^2 \leq \|A_{1,2}A_{1,2}^T\|_2 + \|A_{2,2}A_{2,2}^T\|_2 = \|A_{1,2}^T A_{1,2}\|_2 + \|A_{2,2}^T A_{2,2}\|_2 \leq 2\|A_2\|_2^2.$$

Defining

$$\mu \doteq \left\| \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \right\|_2 \tag{4.12}$$

we then have

$$\epsilon_u \leq \mu^2/2, \quad (\hat{\sigma}_k^{(n)})^2 - \mu^2 \geq \delta_u \geq (\hat{\sigma}_k^{(n)})^2 - 2\mu^2, \tag{4.13}$$

and provided that $\hat{\sigma}_k^{(n)} \geq \sqrt{3}\mu$ we obtain

$$\delta_u \geq 2\epsilon_u \Rightarrow \|P_u\|_2 \leq 2\epsilon_u/\delta_u.$$

For the right dominant singular subspace of M we must consider

$$H_v \doteq M^T M = \begin{bmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & \hat{\Sigma}A_{1,2} \\ A_{1,2}^T\hat{\Sigma} & A_{1,2}^T A_{1,2} + A_{2,2}^T A_{2,2} \end{bmatrix}. \tag{4.14}$$

For the quantities ϵ_v and δ_v corresponding to Theorem 4.1, we find

$$\epsilon_v \doteq \|\hat{\Sigma}A_{1,2}\|_2 \leq \mu\|A\|_2, \quad \delta_v \doteq \min|\lambda(\hat{\Sigma}^2)| - \|A_2\|_2^2 = (\hat{\sigma}_k^{(n)})^2 - \mu^2.$$

Provided that $(\hat{\sigma}_k^{(n)})^2 \geq \frac{16}{7}\mu\|A\|_2$ we obtain

$$\delta_v \geq 2\epsilon_v \Rightarrow \|P_v\|_2 \leq 2\epsilon_v/\delta_v.$$

Applying the same reasoning as above we denote the true dominant subspace as $\mathrm{Im}\left[\begin{smallmatrix} I_k \\ P_v \end{smallmatrix}\right]$. The norm of P_v is then a measure for the angular rotation of this subspace, and it is bounded by $2\epsilon_v/\delta_v$. The corresponding largest canonical angle ϕ_k satisfies again [10]

$$\cos\phi_k = 1/\sqrt{1+\|P_v\|^2}, \quad \sin\phi_k = \|P_v\|/\sqrt{1+\|P_v\|^2}, \quad \tan\phi_k = \|P_v\|$$

and measures the "rotation" of the right dominant singular subspace with respect to its approximation. We summarize this discussion in the following theorem.

THEOREM 4.2. *Let*

$$\hat{M} = \begin{bmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} \hat{\Sigma} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix}, \quad \mu \doteq \left\| \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \right\|_2.$$

Then the angles θ_k and ϕ_k between the k-dimensional left and right singular subspaces of M and $\hat{M}$, respectively, satisfy the bounds

$$\tan\theta_k < \mu^2/((\sigma_k^{(n)})^2 - 2\mu^2) \qquad \text{if} \quad \mu < \sigma_k^{(n)}/\sqrt{3}$$

and

$$\tan\phi_k < \mu\|M\|_2/((\sigma_k^{(n)})^2 - \mu^2) \quad \text{if} \quad \mu < 7(\sigma_k^{(n)})^2/16\|A\|_2.$$

These are also the angles of the left and right singular subspaces of $Q_{(i)}R_{(i)}V_{(i)}^T$ and A.

Unfortunately, we do not compute the matrices $A_{1,2}$ and $A_{2,2}$, and so we have to estimate μ. Bounding μ^2 in terms of the Frobenius norm

$$\mu^2 \leq \sum_i \mu_i^2$$

would yield serious overestimates since δ may become negative. Therefore we have to make some simplifying assumptions. The ith column of A_2 at step i of the recursive calculation contains what could be considered "residual noise vectors," and we assume therefore that they are randomly distributed. It is shown in [7] that an $(n-k) \times n$ matrix B with elements chosen independently from a standard Gaussian distribution has column norms tending to $\sqrt{n}$ and a spectral norm $\|B\|_2$ tending to $\sqrt{n}(1 + \sqrt{(n-k)/n})$ as n becomes large. If our matrix A_2 has equal column norms (hence equal to $\max_i \mu_i$ rather than $\sqrt{n}$), we then obtain the approximation

$$\max_i \mu_i \leq \mu \leq c. \max_i \mu_i, \quad c \approx (1 + \sqrt{(n-k)/n}).$$

On the other hand, if the columns are of very different norm, one gets closer to the lower bound since the number of relevant columns entering the above analysis becomes smaller than $(n-k)$, and thus c tends to 1. We will simply use $\hat{\mu} = \max_i \mu_i$ and $\hat{\sigma}_1^{(n)}$, respectively, as estimates of μ and $\|A\|_2$, which leads to the following approximations for our bounds:

$$\hat{\epsilon}_u \approx \hat{\mu}^2/2, \quad \hat{\delta}_u \approx (\hat{\sigma}_k^{(n)})^2 - \hat{\mu}^2, \quad \hat{\epsilon}_v \approx \hat{\mu}\hat{\sigma}_1^{(n)}, \quad \hat{\delta}_v \approx (\hat{\sigma}_k^{(n)})^2 - \hat{\mu}^2.$$

Notice that these approximations have the advantage that $\hat{\delta}_u$ and $\hat{\delta}_v$ will always be positive since $\sigma_k^{(n)} \geq \sigma_{k+1}^{(i)} = \mu_i$. The resulting estimates for the norm of P_u and P_v then become

$$\|P_u\|_2 \approx \tan\hat{\theta}_k \doteq 2\frac{\hat{\epsilon}_u}{\hat{\delta}_u} = \frac{\hat{\mu}^2}{(\hat{\sigma}_k^{(n)})^2 - \hat{\mu}^2}, \tag{4.15}$$

$$\|P_v\|_2 \approx \tan\hat{\phi}_k \doteq 2\frac{\hat{\epsilon}_u}{\hat{\delta}_u} = \frac{\hat{\mu}\hat{\sigma}_1^{(n)}}{(\hat{\sigma}_k^{(n)})^2 - \hat{\mu}^2}. \tag{4.16}$$

It is possible to estimate the quality of the computed singular values using a simpler analysis. From Theorem 4.1 it follows that

$$N \begin{bmatrix} I + P^T \end{bmatrix} \left(\begin{bmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \begin{bmatrix} A_{1,2}^T & A_{2,2}^T \end{bmatrix} \right) \begin{bmatrix} I \\ P \end{bmatrix} N = H_{1,1}, \tag{4.17}$$

where

$$N = (I + P^T P)^{-\frac{1}{2}}, \quad N = N^T \leq I.$$

This yields the residual equation

$$H_{1,1} - N\hat{\Sigma}^2 N = R \doteq N \begin{bmatrix} I & P^T \end{bmatrix} \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \begin{bmatrix} A_{1,2}^T & A_{2,2}^T \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} N,$$

and since

$$N\hat{\Sigma}^2 N \leq \hat{\Sigma}^2$$

we have

$$H_{1,1} - \hat{\Sigma}^2 \leq H_{1,1} - N\hat{\Sigma}^2 N = R.$$

But

$$\|R\|_2 = \left\| \begin{bmatrix} A_{1,2} \\ A_{2,2} \end{bmatrix} \right\|_2^2 = \mu^2,$$

from which we obtain the strict bound

$$|\sigma_i^2 - (\hat{\sigma}_i^{(n)})^2| \leq \|H_{1,1} - \hat{\Sigma}^2\|_2 \leq \mu^2.$$

This analysis is very simple and does not take into account any information about P, which can be used to improve the bound. Instead, we replace μ by its estimate $\hat{\mu}$, which yields

$$|\sigma_i - \hat{\sigma}_i^{(n)}| \approx \hat{\mu}^2/(\sigma_i + \hat{\sigma}_i^{(n)}) \leq \hat{\mu}^2/2\hat{\sigma}_i^{(n)}. \tag{4.18}$$

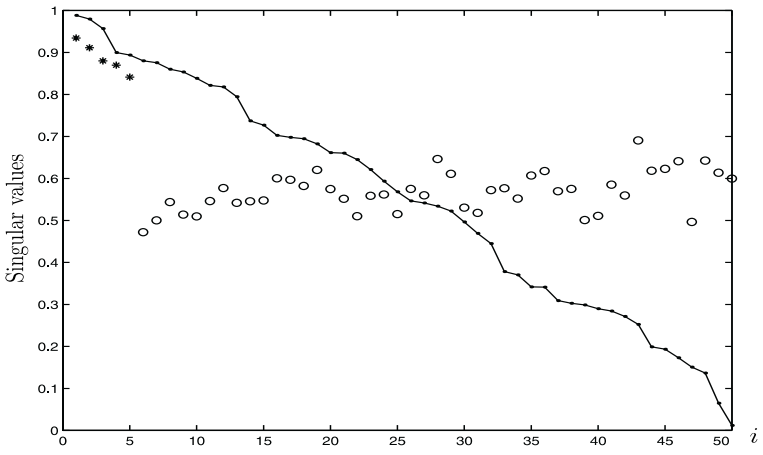
We point out that all of the estimates are quadratic in $\hat{\mu}$, which should give very accurate results if $\hat{\mu} \ll \hat{\sigma}_i^{(n)}$. This is the case if the gap γ at the kth singular value is large, and the quality of the estimate should be expected to deteriorate when this gap becomes small. We illustrate the quality of these bounds in the examples of the next section.

REMARK 4.1. *If A has rank k, then this approach produces an exact decomposition since each submatrix $A_{(i)}$ has rank less than or equal to k and hence $\mu_i = 0$ at each step.*

5. Numerical tests of the approximation. We generated random matrices of dimension $m = 1000$ by $n = 50$ and attempted to track the $k = 5$ dominant singular values and vectors. At every step we keep at most $k + 1 = 6$ vectors in our basis. We thus update to a subspace of dimension 6 and then deflate the smallest singular value to fall back to a space of dimension 5 at each step.

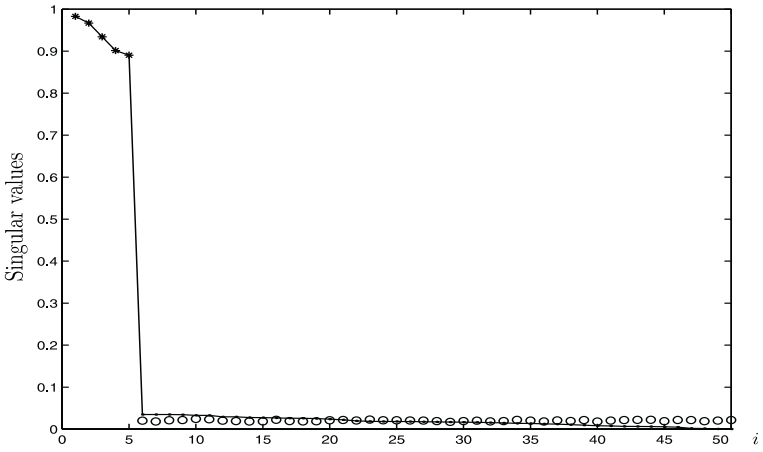
In Figures 1 and 2, the true singular values σ_i $(i = 1, \ldots, n)$ are represented by the solid line, the approximations $\sigma_i^{(n)}$ of the $i = 1, \ldots, k$ leading singular values are the asterisks, and the dismissed singular values μ_i $(i = k + 1, \ldots, n)$ are the circles. Two different gaps are used to illustrate the trend of a larger gap improving the quality of the approximations. Both figures are accompanied by a table (see Tables 1 and 2) listing the singular values σ_i, their approximations $\hat{\sigma}_i^{(n)}$, the corresponding errors $|\sigma_i - \hat{\sigma}_i^{(n)}|$ and their estimate $\hat{\mu}^2/(2\hat{\sigma}_i^{(n)})$, and finally the cosines of the canonical angles $\cos\theta_i$ and $\cos\phi_i$, the smallest of which indicate the rotation of the dominant left and right singular subspaces versus their approximation, and the estimated angles $\cos\hat{\theta}_k$ and $\cos\hat{\phi}_k$. We also give the true value of μ, its estimate $\hat{\mu}$, and finally the $k + 1$ singular value.

From these examples it appears that the method works reasonably well. It should be pointed out that Theorem 4.2 applies only to the second example and that the estimates are very good. Nevertheless the estimates are still acceptable even when the conditions of this theorem do not apply, as is shown by the first example, which has virtually *no* gap! Notice that $\mu/\hat{\mu}$ remains smaller than 2, as suggested by the statistical arguments of section 4. We also analyzed intermediate values of γ, which confirmed the remarks made above.



—— true sv's $\sigma_i(A)$, $*$ approximated sv's $\hat{\sigma}_1^{(n)}, \ldots, \hat{\sigma}_k^{(n)}$, $\circ$ dismissed sv's $\mu_{k+1}, \ldots, \mu_n$

FIG. 1. *Matrix with small gap* $\gamma = 0.01375$.



—— true sv's $\sigma_i(A)$, $*$ approximated sv's $\hat{\sigma}_1^{(n)}, \ldots, \hat{\sigma}_k^{(n)}$, $\circ$ dismissed sv's $\mu_{k+1}, \ldots, \mu_n$

FIG. 2. *Matrix with large gap* $\gamma = 0.85541$.

6. The effect of round-off. In this section we analyze the propagation of round-off in the proposed algorithm. The first aim is to prove some kind of backward stability of the algorithm. We show that at each step i the algorithm produces "approximate" matrices $\bar{V}_{(i)}, \bar{Q}_{(i)}$, and $\bar{R}_{(i)}$ that satisfy exactly the perturbed equations

$$[A(:, 1 \ : i) + E]\bar{V}_{(i)} = \bar{Q}_{(i)}\bar{R}_{(i)}, \quad (\bar{V}_{(i)} + F)^T(\bar{V}_{(i)} + F) = I_k, \tag{6.1}$$

where

$$\|E\|_F \leq \epsilon_e \|A\|_2, \ \ \epsilon_e \approx u, \quad \|F\|_F \leq \epsilon_f \approx u,$$

TABLE 1

σ_i	$\hat{\sigma}_i^{(n)}$	$\lvert\sigma_i - \hat{\sigma}_i\rvert$	$\frac{\hat{\mu}^2}{(2\hat{\sigma}_i^{(n)})}$	$\cos\theta_i$	$\cos\hat{\theta}_i$	$\cos\phi_i$	$\cos\hat{\phi}_i$
0.98833	0.93436	0.05398	0.27320	0.97419	0.36164	0.95272	0.34189
0.97975	0.91122	0.06852	0.28725	0.94833	0.11482	0.91511	0.10679
0.95684	0.87986	0.07698	0.30809	0.88082	0.04148	0.84415	0.03815
0.89977	0.86969	0.03008	0.31534	0.80644	0.11320	0.75753	0.10941
0.89390	0.84136	0.05253	0.33693	0.16487	0.27966	0.14274	0.26322

$\mu = 0.97905$	$\hat{\mu} = 0.69067$	$\sigma_{k+1} = 0.88014$

TABLE 2

σ_i	$\hat{\sigma}_i^{(n)}$	$\lvert\sigma_i - \hat{\sigma}_i\rvert$	$\frac{\hat{\mu}^2}{(2\hat{\sigma}_i^{(n)})}$	$\cos\theta_i$	$\cos\hat{\theta}_i$	$\cos\phi_i$	$\cos\hat{\phi}_i$
0.98299	0.98299	$2.0\ 10^{-7}$	0.00030	0.99999	0.99999	0.99999	0.99999
0.96689	0.96689	$1.0\ 10^{-7}$	0.00032	0.99999	0.99999	0.99999	0.99999
0.93424	0.93424	$1.0\ 10^{-7}$	0.00034	0.99999	0.99999	0.99999	0.99999
0.90161	0.90161	$0.5\ 10^{-7}$	0.00036	0.99999	0.99999	0.99999	0.99999
0.89032	0.89032	$1.5\ 10^{-7}$	0.00037	0.99999	0.99999	0.99999	0.99999

$\mu = 0.03491$	$\hat{\mu} = 0.02430$	$\sigma_{k+1} = 0.03491$

in which u is the so-called unit round-off of the IEEE floating point standard (see, e.g., [9]). This is used to prove that the effect of round-off remains small despite the fact that this is a classical Gram–Schmidt procedure.

The proof of the following theorem is given in the appendix.

THEOREM 6.1. *The recursive algorithm described in sections* 2 *and* 3 *produces "approximate" matrices* $\bar{V}_{(i)}, \bar{Q}_{(i)}$, *and* $\bar{R}_{(i)}$ *that satisfy* exactly *the perturbed equation* (6.1) *with the bounds (up to* $O(u^2)$ *terms)*

$$\|E\|_F \leq \epsilon_e \|A\|_2, \quad \epsilon_e \leq 26k^{\frac{3}{2}}nu, \quad \|F\|_F \leq \epsilon_f \leq 9k^{\frac{3}{2}}nu.$$

We point out here that these bounds do *not* depend on m, the largest dimension of A. Moreover, if one uses Householder transformations rather than Givens transformations, the results are very similar.

REMARK 6.1. *Although Theorem* 6.1 *indicates that the error* $\|E\|_F$ *grows with the number of columns* n, *it does not seem to grow in actual experiments. This can be explained as follows. Assume that at step* i *we have the perturbed equation*

$$\begin{bmatrix} Q_{(i-1)} + E_{(i-1)} & \hat{q}_i + e_i \end{bmatrix} G_u = \begin{bmatrix} Q_{(i)} + E_{(i)} & q_i + g_i \end{bmatrix}, \tag{6.2}$$

where $E_{(i)}$ *accounts for the loss of orthogonality in* $Q_{(i)}$, *and* e_i *is the local error in the vector* $\hat{q}_i$, *and* g_i *is the resulting error in the vector* q_i. *If we assume the errors in the right-hand side of* (6.2) *to be evenly distributed over the matrix, then it follows that*

$$\|E_{(i)}\|_F^2 \leq \frac{k}{(k+1)}\|E_{(i-1)}\|_F^2 + \|e_i\|_2^2, \tag{6.3}$$

which for growing i *tends to a limit*

$$\|E\|_F^2 \leq (k+1)\max_i \|e_i\|_2^2$$

that is independent of n. The same reasoning can be applied to the error $\|F\|_F$. *The corresponding bounds of Theorem* 6.1 *become*

$$\epsilon_e \le 26k^2u, \quad \epsilon_f \le 9k^2u.$$

We now turn our attention to the loss of orthogonality in the computed matrix $\bar{Q}$. This can be bounded using a perturbation result for the QR factorization of

$$(A+E)\bar{V} = A\bar{V} + E\bar{V} \doteq A\bar{V} + G,$$

where, using the bounds of Theorem 6.1, we have

$$\|G\|_F = \epsilon_g\|A\|_2, \quad \epsilon_g \le \epsilon_e + O(\epsilon_e\epsilon_f) \approx u.$$

THEOREM 6.2. *Let (a given matrix)* $\bar{V} \in \mathcal{R}^{n\times k}$ *"select" k columns of the matrix* $A \in \mathcal{R}^{m\times n}$, *and let*

$$A\bar{V} = QR, \quad Q^TQ = I_k,$$

with R upper triangular, be its exact QR factorization. Let

$$A\bar{V} + G = \bar{Q}\bar{R}, \quad \|G\|_F = \epsilon_g\|A\|_2 \approx u\|A\|_2 \tag{6.4}$$

be a "computed" version, where $\bar{Q} = Q + \Delta_Q$, $\bar{R} = R + \Delta_R$. *Then under a mild assumption, namely, condition* (6.6), *we can bound the loss of orthogonality in* $\bar{Q}$ *as follows:*

$$\|\bar{Q}^T\bar{Q} - I_k\|_F \le \sqrt{2}\epsilon_g\kappa_2(R)\kappa_R(A\bar{V}) \le 2\epsilon_g\kappa_2^2(R), \quad \epsilon_g \approx u.$$

Proof. Since $\bar{Q}$ is not necessarily orthogonal we first compute its QR factorization:

$$\bar{Q} = Q_0R_0, \quad Q_0^TQ_0 = I_k.$$

So we can consider the perturbation of the QR decomposition of $A\bar{V}$:

$$A\bar{V} = QR, \quad A\bar{V} + G = Q_0(R_0\bar{R}). \tag{6.5}$$

The loss of orthogonality in $\bar{Q}$ can be measured by R_0 since

$$\bar{Q}^T\bar{Q} - I_k = R_0^TQ_0^TQ_0R_0 - I_k = R_0^TR_0 - I_k.$$

To measure this, we first use a perturbation analysis of [6] for (6.5) to obtain

$$\|R_0\bar{R} - R\|_F \le \epsilon_g\kappa_R(A\bar{V})\|R\|_2,$$

where $\kappa_R(A\bar{V})$ is the "refined" condition number of the factor R of the QR factorization (6.5) of $A\bar{V}$ [6]. If we define $\Delta_0 \doteq R_0 - I_k$, we then have

$$R_0\bar{R} - R = (I_k + \Delta_0)(R + \Delta_R) - R = \Delta_0\bar{R} + \Delta_R \approx \Delta_0R + \Delta_R$$

and, hence,

$$\|\Delta_0R + \Delta_R\|_F \approx \|\Delta_0\bar{R} + \Delta_R\|_F \le \epsilon_g\kappa_2(R)\|R\|_2.$$

We now assume that there are no strong cancellations between $\|\Delta_R\|_F$ (measuring the perturbation of R) and $\|\Delta_0 R\|_F$ (measuring the perturbation in Q) and hence that $\|\Delta_0 R\|_F$ and $\|\Delta_R + \Delta_0 R\|_F$ are of the same order of magnitude:

$$\|\Delta_0 R\|_F \approx \|\Delta_R + \Delta_0 R\|_F. \tag{6.6}$$

From $\|\Delta_0 R\|_F \leq \epsilon_g \kappa_R(A\bar{V})\|R\|_2$ it then follows that

$$\|\Delta_0\|_F \leq \epsilon_g \kappa_R(A\bar{V})\|R\|_2\|R^{-1}\|_2.$$

This can now be used to bound $\|R_0^T R_0 - I_k\|_F = \|\Delta_0 + \Delta_0^T + \Delta_0^T\Delta_0\|_F \approx \sqrt{2}\|\Delta_0\|_F$, which yields

$$\|R_0^T R_0 - I_k\|_F \leq \sqrt{2}\epsilon_g \kappa_2(R)\kappa_R(A\bar{V}). \tag{6.7}$$

Using the overestimate $\kappa_R(A\bar{V}) \leq \sqrt{2}\kappa_2(R)$ of [6] we approximate this finally by

$$\|R_0^T R_0 - I_k\|_F \leq 2\epsilon_g \kappa_2^2(R). \qquad \square \tag{6.8}$$

REMARK 6.2. *Assumption* (6.6) *is crucial to the proof of Theorem* 6.2. *It is easy to see that any factorization of the type* (6.4) *will not yield the bounds* (6.7) *or* (6.8)*: consider, e.g., the factorization*

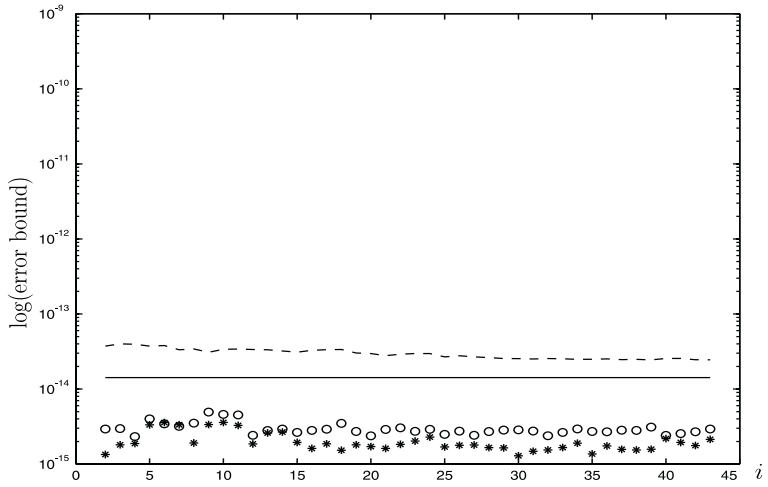
$$A\bar{V} + G = (\bar{Q}U)(U^{-1}\bar{R}),$$

where U is any invertible upper triangular matrix. This clearly satisfies the conditions of the theorem, except for assumption (6.6). *The critical quantity for this new factorization then becomes $\|U^T R_0^T R_0 U - I_k\|_F$, and since U can be chosen arbitrarily, it is impossible to bound it. Assumption* (6.6) *is therefore crucial, and we show in the next section that it indeed holds in practice.*

7. Numerical tests for the error propagation. In this section we present numerical evidence that the analysis of the previous section can be applied to the tracking problem of the dominant spaces of a given matrix. The numerical experiments we ran show that the loss of orthogonality in the computed matrix $\bar{Q}_{(i)}$ of (6.1) remains bounded by the condition number squared of the matrix R that we are "tracking."

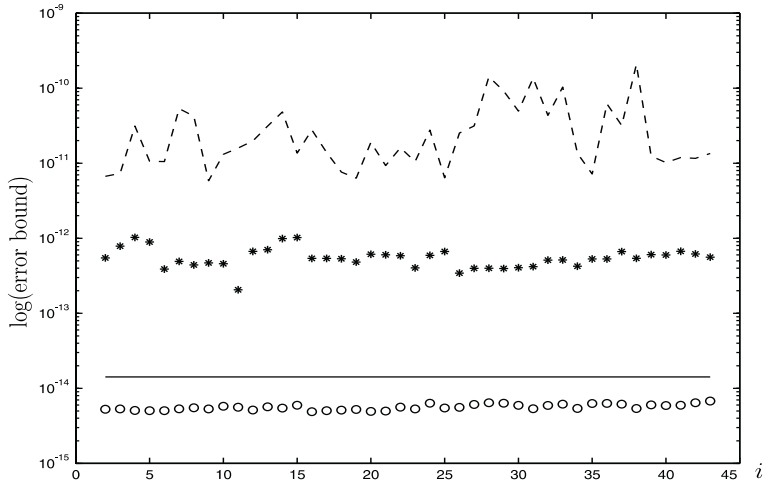
We show in Figures 3 and 4 two plots that compare the loss of orthogonality in the proposed algorithm based on the classical Gram–Schmidt method (labeled CGS) and a "fully orthogonal" method, which we obtain by performing *two* steps of CGS, rather than one, at each iteration. This second method, labeled CGS2, was analyzed in [1] and shown to yield a Q factor that is close to orthogonal. We chose this as an alternative to the Householder method because in the iterative scheme considered in this paper, CGS2 involves significantly fewer operations than the Householder method.

As suggested by Remark 6.1, the backward error $E_{(i)}$ and the quantity ϵ_e can be bounded independently of the step i. We therefore compare the loss of orthogonality $\|R_0^T R_0 - I_k\|_F$ with the quantities $uk^2\kappa_2(R_{(i)})\kappa_R(A(:, 1:i)\bar{V}_{(i)})$ for the CGS method and uk^2 for the CGS2 method. These "simplified" quantities are indicators to show that the loss of orthogonality is of the order of magnitude predicted by our error analysis. To show the effect of the condition number of the triangular factor $R_{(i)}$, we let it grow in the two examples by choosing a growing condition number for A.



– – CGS bound $uk^2\kappa_2(R_{(i)})\kappa_R(A(:,1:i)\bar{V}_{(i)}))$, —— CGS2 bound uk^2,
$*$ loss of orthogonality in CGS method, $\circ$ loss of orthogonality in CGS2 method

FIG. 3. $\kappa_2(A) = 41.806$, $\kappa_2(R_{(n)}) = 1.156$, $\kappa_R(A\bar{V}_{(n)})) = 1.492$.



– – CGS bound $uk^2\kappa_2(R_{(i)})\kappa_R(A(:,1:i)\bar{V}_{(i)}))$, —— CGS2 bound uk^2,
$*$ loss of orthogonality in CGS method, $\circ$ loss of orthogonality in CGS2 method

FIG. 4. $\kappa_2(A) = 6928$, $\kappa_2(R_{(n)}) = 134.7$, $\kappa_R(A\bar{V}_{(n)}) = 7.028$.

The following observations can be derived from these experiments:

- The condition numbers $\kappa_2(R_{(i)})$ and $\kappa_R(A(:,1:i)\bar{V}_{(i)})$ do not affect the loss of orthogonality of the CGS2 method, as expected from the analysis of [1]. (The product $\kappa_2(R_{(i)})\kappa_R(A(:,1:i)\bar{V}_{(i)})$ can be inferred from the gap between the CGS and CGS2 bounds.)
- The statistical assumption of Remark 6.1 seems to hold since there is no growth in the loss of orthogonality of the computed matrices $\bar{Q}_{(i)}$: this should

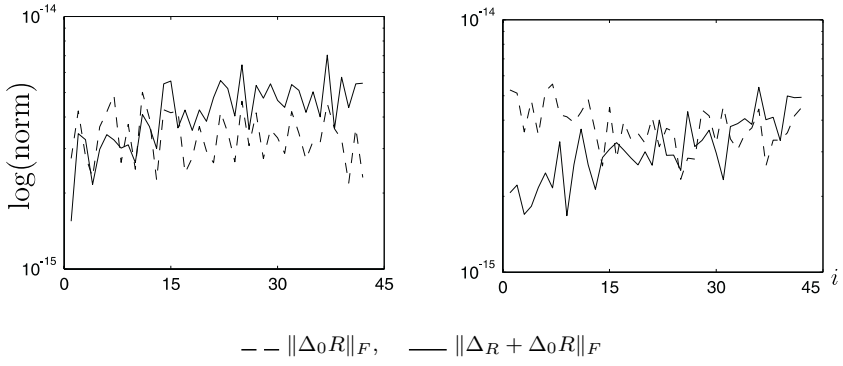


FIG. 5. *Verification of assumption* (6.6) *for examples Figures* 3 *and* 4.

depend on the backward error $E_{(i)}$, which does not depend on i if the assumption of Remark 6.1 holds

- Assumption (6.6) made in Theorem 6.2 was verified in these experiments and validates the resulting bounds (6.7), (6.8) of that theorem; the graphs in Figure 5 give the norms of the two quantities for the two examples given earlier and illustrate that the assumption that those quantities are of the same order of magnitude is reasonable.
- The loss of orthogonality remains very reasonable when the condition number $\kappa_2(R_{(i)})$ is not too large, which is a reasonable assumption in applications where a "dominant matrix" $R_{(i)}$ is being tracked.

We observed no difference in the computed spaces for the CGS or CGS2 methods. We conclude from our analysis and the experimental evidence that the cheapest version of the algorithm (CGS) can be used safely for the applications represented by the experiments and mentioned in section 1. By this we mean that the angles $\cos\theta_k$ and $\cos\phi_k$ for both methods were equal in the first four digits despite a very small loss of orthogonality in the CGS method.

8. Conclusions. In this paper we presented an analysis of an efficient incremental algorithm to compute the dominant subspace of a given matrix A. Although similar algorithms have been discussed in the literature [5], we have given here a more efficient implementation along with a fairly tight bound on its accuracy and estimators that can be used in practice to monitor that accuracy.

The contributions of this paper are the following:

- A CGS-like algorithm of complexity close to $8mnk$ flops was derived for computing a rank k approximation of an $m \times n$ matrix A.
- A posteriori bounds for the accuracy of the approximation error were presented and their reliability was illustrated.
- The effect of round-off was studied, and it was shown that the algorithm behaves much better than what can be expected for CGS. An explanation of this phenomenon was given and illustrated by numerical experiments. The effect of propagation of round-off errors was also analyzed and shown to be negligible for the applications considered in this paper.

Appendix. In this section we give the proof of Theorem 6.1. This result is obtained by analyzing one step i of the recursive algorithm. We first analyze the local errors in that step and hence assume all quantities at the beginning of step i to be

exact. For the computations of step i we use $\bar{x}$ to denote the "computed version" of x that is actually stored in the computer.

The first part of step i is the Gram–Schmidt update, which corresponds to

$$\bar{r}_i = fl(\bar{Q}_{(i-1)}^T a_i), \tag{A.1}$$

$$\bar{\tilde{q}}_i = fl(a_i - \bar{Q}_{(i-1)}^T \bar{r}_i), \tag{A.2}$$

$$\bar{\rho}_i = fl\left(\sqrt{\bar{\tilde{q}}_i^T \bar{\tilde{q}}_i}\right), \tag{A.3}$$

$$\bar{q}_i = fl(\bar{\tilde{q}}_i / \bar{\rho}_i). \tag{A.4}$$

From (A.2), (A.4), and standard error analysis results it follows that

$$\bar{\tilde{q}}_i = a_i + d_i - [\bar{Q}_{(i-1)} + \delta Q_{(i-1)}]\bar{r}_i = \bar{\rho}_i[\bar{q}_i + f_i], \tag{A.5}$$

where (up to order u^2) we have the elementwise inequalities

$$|[f_i]_j| \le u|[\bar{q}_i]_j|, \quad |[d_i]_j| \le ku|[a_i]_j|, \quad |[\delta Q_{(i-1)}]_{jl}| \le (k-l+2)u|[\bar{Q}_{(i-1)}]_{jl}|.$$

To obtain this result we assumed that the loop on the columns of the Gram–Schmidt orthogonalization (A.2) progresses from left to right. We can then equate this as follows:

$$a_i + e_i = \begin{bmatrix} \bar{Q}_{(i-1)} & \bar{q}_i \end{bmatrix} \begin{bmatrix} \bar{r}_i \\ \rho_i \end{bmatrix}, \quad e_i = d_i - \delta Q_{(i-1)}\bar{r}_i + f_i\bar{\rho}_i. \tag{A.6}$$

We also assume that

$$\| \begin{bmatrix} \bar{Q}_{(i-1)} & \bar{q}_i \end{bmatrix} - \begin{bmatrix} Q_{(i-1)} & q_i \end{bmatrix} \|_2 = K.u \ll 1, \tag{A.7}$$

i.e., there is no *complete* loss of orthogonality, which allows us to approximate the 2-norm of $\begin{bmatrix} \bar{Q}_{(i-1)} & \bar{q}_i \end{bmatrix}$ or any of its columns by $1 + O(u)$. We then obtain the inequalities

$$\begin{aligned}
\|e_i\|_2 &\le \|d_i\|_2 + \|f_i\bar{\rho}_i\|_2 + \sum_l \||\delta Q_{(i-1)}|_{:,l}\|_2 . |\bar{r}_i|_l + O(u^2) \\
&\le u\left[k\|a_i\|_2 + \|\bar{q}_i\|\bar{\rho}_i + \sum_l \||Q_{(i-1)}|_{:,l}\|_2 . (k-l+2)|r_i|_l\right] + O(u^2) \\
&\le u\left[k\|a_i\|_2 + \left(|\bar{\rho}_i| + \sum_l (k-l+2)|\bar{r}_i|_l\right)\right] + O(u^2) \\
&\le u(k\|a_i\|_2 + \|[1,2,\ldots,k+1]\|_2\|a_i\|_2) + O(u^2) \\
&\le u\left(k + \sqrt{\frac{(k+2)^3}{3}}\right)\|a_i\|_2 + O(u^2),
\end{aligned} \tag{A.8}$$

where the next-to-last line was obtained by Cauchy–Schwarz. Notice that all errors due to this part are superposed on column a_i. Therefore the error matrix E_1 of this first part satisfies $\|E_1\|_F = \|e_i\|_2$.

The second part of step i consists of the transformations G_v and G_u in (9), which we assume are each implemented with a sequence of k Givens rotations. For this we

will use Lemma 18.8 of [9], which we recall in a slightly modified form. (We refer to [9] for the details of the implementation and construction of each Givens rotation.)

LEMMA A.1. *Consider the sequence of Givens transformations*

$$M_k = G_k \cdot \ldots \cdot G_1 M = G \cdot M.$$

Then there exists a perturbation ΔM of M such that the computed matrix $\bar{M}_k$ satisfies

$$\bar{M}_k = G(M + \Delta M), \quad \|\Delta M\|_F \leq 6k\sqrt{2}u\|M\|_F + O(u^2).$$

Applying this to the products $Q_{up} \cdot R_{up} = (QG_u^T) \cdot (G_u R G_v^T)$ and $V_{up} = (VG_v^T)$ we obtain

$$\begin{aligned} \bar{Q}_{up}\bar{R}_{up} &= (Q + \Delta Q)G_u^T \cdot G_u(R + \Delta R)G_v^T \doteq QRG_v^T + E_2, \\ \bar{V}_{up} &= (V + \Delta V)G_v^T \doteq VG_v^T + F, \end{aligned}$$

where

$$\begin{aligned} E_2 &\doteq (\Delta QR + Q\Delta R + \Delta Q\Delta R)G_v^T, \\ \|\Delta Q\|_F &\leq 6\sqrt{2}ku\|Q\|_F + O(u^2) = 6k\sqrt{2(k+1)}u + O(u^2), \\ \|\Delta R\|_F &\leq 12\sqrt{2}ku\|R\|_F + O(u^2) = 12k\sqrt{2(k+1)}u\|A\|_2 + O(u^2), \end{aligned}$$

and

$$\begin{aligned} F &\doteq (\Delta V)G_v^T, \\ \|\Delta V\|_F &\leq 6\sqrt{2}ku\|V\|_F + O(u^2) = 6k\sqrt{2(k+1)}u + O(u^2). \end{aligned}$$

The norms of E_2 and F can then be bounded by

$$\begin{aligned} \|E_2\|_F &\leq \|Q\|_2\|\Delta R\|_F + \|R\|_2\|\Delta Q\|_F + O(u^2) \\ &\leq 18k\sqrt{2(k+1)}u\|A\|_2 + O(u^2), \\ \|F\|_F &\leq 6k\sqrt{2(k+1)}u + O(u^2). \end{aligned}$$

Combining the bounds for E_1 and E_2 yields the bound

$$\|E\|_F \ \leq \ 26uk^{\frac{3}{2}}\|A\|_2 + O(u^2)$$

for the local error E in step i. Similarly, the error matrix F on $V_{(i)}$ corresponding to the local errors of step i can be bounded by

$$\|F\|_F \ \leq \ 9uk^{\frac{3}{2}} + O(u^2).$$

In order to sum up these errors over the $n-k$ steps of the algorithm, we can neglect the second order effects and then only need to multiply these bounds by $(n-k)$. This then yields the bounds of Theorem 6.1.

Acknowledgment. We would like to thank A. Edelman for pointing out the work of Gemam to us.

REFERENCES

[1] N. Abdelmalek, *Round-off error analysis for Gram-Schmidt method and solution of linear least squares problems*, BIT, 11 (1971), pp. 45–68.

[2] C. G. Baker, *An Incremental Block Algorithm for Tracking Dominant Singular Subspaces*, Technical Report FSU-CSIT-03-03, CSIT, Florida State University, Tallahassee, FL, 2003.

[3] Y. Chahlaoui, K. Gallivan, and P. Van Dooren, *An incremental method for computing dominant singular spaces*, in Computational Information Retrieval, M. W. Berry, ed., SIAM, Philadelphia, 2001, pp. 53–62.

[4] T. Chan, *An improved algorithm for computing the singular value decomposition*, ACM Trans. Math. Software, 8 (1982), pp. 72–83.

[5] S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkeler, and H. Zhang, *An eigenspace update algorithm for image analysis*, Graph. Models Image Process., 59 (1997), pp. 321–332.

[6] X.-W. Chang, C. C. Paige, and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 775–791.

[7] S. Geman, *A limit theorem for the norm of random matrices*, Ann. Probab., 8 (1980), pp. 252–261.

[8] G. H. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[9] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[10] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.

[11] P. Van Dooren, *Gramian based model reduction of large-scale dynamical systems*, in Numerical Analysis 1999, Chapman Hall/CRC Press, London, 2000, pp. 231–247.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 464–485

A SCHUR–PARLETT ALGORITHM FOR COMPUTING MATRIX FUNCTIONS*

PHILIP I. DAVIES† AND NICHOLAS J. HIGHAM†

Abstract. An algorithm for computing matrix functions is presented. It employs a Schur decomposition with reordering and blocking followed by the block form of a recurrence of Parlett, with functions of the nontrivial diagonal blocks evaluated via a Taylor series. A parameter is used to balance the conflicting requirements of producing small diagonal blocks and keeping the separations of the blocks large. The algorithm is intended primarily for functions having a Taylor series with an infinite radius of convergence, but it can be adapted for certain other functions, such as the logarithm. Novel features introduced here include a convergence test that avoids premature termination of the Taylor series evaluation and an algorithm for reordering and blocking the Schur form. Numerical experiments show that the algorithm is competitive with existing special-purpose algorithms for the matrix exponential, logarithm, and cosine. Nevertheless, the algorithm can be numerically unstable with the default choice of its blocking parameter (or in certain cases for all choices), and we explain why determining the optimal parameter appears to be a very difficult problem. A MATLAB implementation is available that is much more reliable than the function `funm` in MATLAB 6.5 (R13).

Key words. matrix function, matrix exponential, matrix logarithm, matrix cosine, Taylor series, Schur decomposition, Parlett recurrence, sep function, LAPACK, MATLAB

AMS subject classification. 65F30

DOI. 10.1137/S0895479802410815

1. Introduction.

Matrix functions play a diverse role in science and engineering. They arise most frequently in connection with the solution of differential equations, with application areas including control theory [2], nuclear magnetic resonance [6], [15], Lie group methods for geometric integration [22, sect. 8], and the numerical solution of stiff ordinary differential equations [9]. A large body of theory on matrix functions exists, with comprehensive treatments available in [12] and [21], for example. In this work a function $f(A)$ of a matrix $A \in \mathbb{C}^{n\times n}$ has the usual meaning, which can be defined in terms of a Cauchy integral formula, a Hermite interpolating polynomial, or the Jordan canonical form, and we assume that f is "defined on the spectrum of A" (see any of the above references for details). The main property we need is that for each A, $f(A)$ is expressible as a polynomial in A (and of course that polynomial depends on A).

A wide variety of computational methods have been proposed, most of them geared to particular functions such as the exponential, the logarithm, and the square root. However, apart from the method of Kågström [24] discussed below, no numerically reliable method exists for computing $f(A)$ for a general function f. Such a method is needed for several reasons. First, software packages cannot provide special-purpose routines for all the functions that might be required. For example, MATLAB 6.5 (R13) provides routines to evaluate the matrix functions e^A (`expm`) and $A^{1/2}$ (`sqrtm`), but the matrix logarithm and matrix cosine, for example, must be computed via the routine `funm` for general f. (MATLAB has a routine `logm` that computes

*Received by the editors July 4, 2002; accepted for publication (in revised form) by B. Kågström March 31, 2003; published electronically September 9, 2003. This research was supported by Engineering and Physical Sciences Research Council grant GR/R22612.
http://www.siam.org/journals/simax/25-2/41081.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (pdavies@ma.man.ac.uk, http://www.ma.man.ac.uk/~ieuan, higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/).

the matrix logarithm, but it calls `funm`). MATLAB's `funm` has the capabilities that we are arguing for, but it is not numerically reliable, as is shown by our numerical experiments in section 7. The second benefit of a general purpose routine is that it provides a benchmark for comparison. Methods for specific f can be rejected if they offer no advantage over the best general method.

A general approach to compute $f(A)$ for $A \in \mathbb{C}^{n\times n}$ is to employ a similarity transformation

$$A = ZBZ^{-1}, \tag{1.1}$$

where $f(B)$ is easily computable. Then

$$f(A) = Zf(B)Z^{-1}. \tag{1.2}$$

If A is diagonalizable, for example, we can take $B = \mathrm{diag}(\lambda_i)$ and then $f(B) = \mathrm{diag}(f(\lambda_i))$ is trivially obtained. The drawback with this approach is that errors in evaluating $f(B)$ are multiplied by as much as $\kappa(Z) = \|Z\|\|Z^{-1}\| \geq 1$, yet the conditioning of $f(A)$ is not necessarily related to $\kappa(Z)$, so this approach may be numerically unstable. It is therefore natural to restrict to well conditioned transformations Z. Two ways do so are to take (1.1) to be a Schur decomposition, so that Z is unitary and B triangular, and to block diagonalize A using well conditioned transformations. We consider these two possibilities in the next two subsections.

1.1. Schur method. Computation of a Schur decomposition $A = QTQ^*$, where Q is unitary and T is upper triangular, is achieved with perfect backward stability by the QR algorithm [13, Chap. 7], so in computing $f(A) = Qf(T)Q^*$ the interest is in how to obtain $F = f(T)$. Since T is upper triangular, so is F (since it is a polynomial in T). Parlett [33] proposed using the following recurrence, which comes from equating (i, j) elements $(i < j)$ in the commutativity relation $FT = TF$:

$$f_{ij} = t_{ij}\frac{f_{ii} - f_{jj}}{t_{ii} - t_{jj}} + \sum_{k=i+1}^{j-1} \frac{f_{ik}t_{kj} - t_{ik}f_{kj}}{t_{ii} - t_{jj}}. \tag{1.3}$$

From (1.3) we see that any element of F can be calculated so long as all the elements to the left and below it are known. Thus the recurrence allows us to compute F a superdiagonal at a time, starting with the diagonal elements $f_{ii} = f(t_{ii})$. MATLAB's `funm` implements this Schur method.

Unfortunately, Parlett's recurrence breaks down when $t_{ii} = t_{jj}$ for some $i \neq j$, that is, when T has repeated eigenvalues, and it can give inaccurate results in floating point arithmetic when T has close eigenvalues. For example, if all the elements of F and T are $O(1)$ but T has two close eigenvalues with $t_{ii} - t_{jj} = O(\epsilon)$ (a not unreasonable scenario), then $t_{ij}(f_{ii} - f_{jj}) + \sum_{k=i+1}^{j-1}(f_{ik}t_{kj} - t_{ik}f_{kj}) = O(\epsilon)$, so that the sum suffers massive, and probably very damaging, cancellation.

Parlett [32] notes that if $T = (T_{ij})$ is block upper triangular, then $F = (F_{ij})$ has the same block structure and, for $i < j$,

$$T_{ii}F_{ij} - F_{ij}T_{jj} = F_{ii}T_{ij} - T_{ij}F_{jj} + \sum_{k=i+1}^{j-1}(F_{ik}T_{kj} - T_{ik}F_{kj}). \tag{1.4}$$

This recurrence can be used to compute F a block superdiagonal at a time, provided we can evaluate the blocks $F_{ii} = f(T_{ii})$ and solve the Sylvester equations (1.4) for the

F_{ij}. For the Sylvester equation (1.4) to be nonsingular we need that T_{ii} and T_{jj} have no eigenvalue in common. Moreover, for the Sylvester equations to be well conditioned a necessary condition is that the eigenvalues of T_{ii} and T_{jj} are well separated. Therefore to implement this block form of Parlett's recurrence we need first to reorder the Schur factor T into a block triangular matrix having two properties: distinct diagonal blocks have "sufficiently distinct" eigenvalues and, to aid the evaluation of f on the diagonal blocks, the eigenvalues within a block are "close." A parameter is required to define "close" and "sufficiently distinct."

1.2. Block diagonalization. An alternative approach is first to compute $A = XDX^{-1}$, where X is well conditioned and D is block diagonal. Then $f(A) = Xf(D)X^{-1}$ and the problem reduces to computing $f(D)$. The usual way to compute a block diagonalization is first to compute the Schur form and then to eliminate off-diagonal blocks by solving Sylvester equations [4], [13, sect. 7.6.3], [28]. In order to guarantee a well-conditioned X a bound must be imposed on the condition of the individual transformations; this bound will be a parameter in the algorithm.

Computing $f(D)$ reduces to computing $f(D_{ii})$ for each diagonal block D_{ii}. The D_{ii} are triangular but, unlike for the Schur method, no particular eigenvalue distribution is guaranteed, because of the limitations on the condition of the transformations; therefore $f(D_{ii})$ is still a nontrivial calculation.

1.3. Choice of method. The Schur method and the block diagonalization method are closely related. Both employ a Schur decomposition, both solve Sylvester equations, and both must compute $f(T_{ii})$ for atomic triangular blocks T_{ii} ("atomic" refers to the fact that these blocks cannot be further reduced). Parlett and Ng [34, sect. 5] show that the two methods are mathematically equivalent, differing only in the order in which two commuting Sylvester operators are applied. In this work we have chosen to use the Schur method, because it has the advantage that it produces atomic blocks with "close" eigenvalues—a property that we can exploit.

Our algorithm for computing $f(A)$ consists of several stages. The Schur decomposition $A = QTQ^*$ is computed, T is reordered to $\widetilde{T}$, the diagonal blocks $f(\widetilde{T}_{ii})$ are computed, the rest of $f(\widetilde{T})$ is computed using the block form of the Parlett recurrence, and finally the unitary similarity transformations from the Schur decomposition and the reordering are applied. We consider first, in section 2, the evaluation of f on the atomic blocks, for which we use a Taylor series expansion. This approach is mainly intended for functions whose Taylor series have an infinite radius of convergence, such as the exponential and the trigonometric and hyperbolic functions, but for some other functions, such as the logarithm, this step can be adapted or replaced by another technique. In section 3 we analyze the use of Parlett's recurrence. Based on the conflicting requirements of these two stages we describe our Schur reordering strategy in section 4.

Our algorithm is summarized in section 5 and the relevance of several preprocessing techniques is discussed in section 6. An extensive set of numerical experiments is described in section 7.

For real matrices, it is natural to use the real Schur decomposition in the first step of the algorithm and to attempt to work entirely in real arithmetic. However, the algorithm's strategy of placing eigenvalues that are not close in different blocks requires splitting complex conjugate pairs of eigenvalues having large imaginary parts, forcing complex arithmetic, so the algorithm does not lend itself to exploitation of the real Schur form.

We note that an attraction of the algorithm developed here is that it allows a function of the form $f(A) = \sum_i f_i(A)$ (e.g., $f(A) = \sin A + \cos A$) to be computed with less work than is required to compute each $f_i(A)$ separately, since the Schur decomposition and its reordering need only be computed once.

We emphasize that our goal is to develop a method applicable for a wide range of f. For particular f it will usually be possible to produce a more efficient or a more accurate algorithm. For example, for matrix pth roots reordering the Schur form is not necessary—the Schur-based methods of [5], [17], and [36] achieve essentially perfect numerical stability by exploiting elegant recurrences for pth roots of triangular matrices. In the case of the logarithm function our algorithm in its general form is not applicable, but we will specialize it to the logarithm and thereby obtain a method that is a candidate for the best general purpose $\log A$ method.

Ours is not the first work to exploit reordered Schur decompositions or the Parlett recurrence for computing matrix functions. Parlett's recurrence was used by Kågström in his thesis [24]. There are three main differences between Kågström's approach and ours. First, he used an initial block diagonalization, carried out with the method of Kågström and Ruhe [26], whereas we compute a Schur decomposition and reorder the triangular form. Second, Kågström uses the scalar rather than the block form of the Parlett recurrence and when t_{ii} and t_{jj} are sufficiently close he uses an explicit formula for f_{ij} involving derivatives (this formula is given in [13, Thm. 11.1.3], for example). Finally, we use a combination of Taylor series and the Parlett recurrence, whereas Kågström investigated the separate use of these two tools upon his block diagonal form. More recently, Parlett and Ng [34] developed an algorithm specifically for the matrix exponential that employs the Schur form with reordering and two levels of blocking, exponentiates the diagonal blocks using the Newton divided difference form of the interpolating polynomial, and uses the Parlett recurrence to obtain the off-diagonal blocks.

2. Evaluating functions of the atomic blocks. Given an upper triangular matrix $T \in \mathbb{C}^{n\times n}$ whose eigenvalues are "close" and an arbitrary function f, we need a method for evaluating $f(T)$ efficiently and accurately. One approach, suggested by Stewart [30, Method 18] for the matrix exponential and investigated for general f by Kågström [24], is to expand f in a Taylor series about the mean of the eigenvalues of T. Write

$$T = \sigma I + M, \qquad \sigma = \operatorname{trace}(T)/n, \tag{2.1}$$

which defines M as T shifted by the mean of its eigenvalues, and let $\lambda(T)$ denote the set of eigenvalues of T. If f has a Taylor series representation

$$f(\sigma + z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\sigma)}{k!} z^k \tag{2.2}$$

for z in an open disk containing $\lambda(T - \sigma I)$, then

$$f(T) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\sigma)}{k!} M^k. \tag{2.3}$$

If T has just one eigenvalue, so that $t_{ii} \equiv \sigma$, then M is strictly upper triangular and hence is nilpotent with $M^n = 0$; the series (2.3) is then finite. More generally, if the eigenvalues of T are sufficiently close, then the powers of M can be expected to

decay quickly after the $(n-1)$st, and so a suitable truncation of (2.3) should yield good accuracy. We make this notion precise in the following lemma, in which we represent $M = D + N$, with D diagonal and N strictly upper triangular (that is, having zero diagonal) and hence nilpotent. For matrices, absolute values and inequalities are defined componentwise.

LEMMA 2.1. *Let $D \in \mathbb{C}^{n\times n}$ be diagonal with $|D| \le \delta I$ and let $N \in \mathbb{C}^{n\times n}$ be strictly upper triangular. Then*

$$|(D+N)^k| \le \sum_{i=0}^{\min(k,n-1)} \binom{k}{i} \delta^{k-i} |N|^i$$

and the same inequality holds with absolute values replaced by any consistent matrix norm.

Proof. The bound follows from

$$|(D+N)^k| \le (|D| + |N|)^k \le (\delta I + |N|)^k,$$

followed by a binomial expansion of the last term. Since $|N|^{n-1} = 0$ we can drop the terms involving $|N|^i$ for $i \ge n-1$. An analogous argument holds for any consistent matrix norm. ∎

If $\delta < 1$ and $\delta \ll \|N\|$ in Lemma 2.1, then, for $k \ge n-1$,

$$\|(D+N)^k\| = O(\delta^{k+1-n}\|N\|^{n-1}),$$

and hence the powers of $D+N$ decay rapidly after the $(n-1)$st, irrespective of N.

This analysis shows that as long as the scalar multipliers $f^{(k)}(\sigma)/k!$ in (2.3) are not too large we should be able to truncate the series (2.3) soon after the $(n-1)$st term (and possibly much earlier if M is small).

We need a reliable criterion for deciding when to truncate the Taylor series. When summing a series whose terms decrease monotonically it is safe to stop as soon as a term is smaller than the desired error. Unfortunately, our matrix Taylor series can exhibit very nonmonotonic convergence. Indeed, when $n = 2$, $M = T - \sigma I$ always has the form

$$M = \begin{bmatrix} \epsilon & \alpha \\ 0 & -\epsilon \end{bmatrix}, \tag{2.4}$$

and its powers are

$$M^{2k} = \begin{bmatrix} \epsilon^{2k} & 0 \\ 0 & \epsilon^{2k} \end{bmatrix}, \quad M^{2k+1} = \begin{bmatrix} \epsilon^{2k+1} & \alpha\epsilon^{2k} \\ 0 & -\epsilon^{2k+1} \end{bmatrix}.$$

For $|\epsilon| < 1$, $\|M^k\| \to 0$ as $k \to \infty$, but $\|M^{2k+1}\| \gg \|M^{2k}\|$ for $\alpha \gg 1$. The next theorem shows that this phenomenon of the "disappearing nonnormal part" is connected with the fact that f can map distinct λ_i into the same value.

THEOREM 2.2. *Let $D \in \mathbb{C}^{n\times n}$ be diagonal with distinct eigenvalues $\lambda_1, \dots, \lambda_p$ $(1 \le p \le n)$ of multiplicity $k_1, \dots, k_p$, respectively, and let $f(z)$ be an analytic function on an open set containing $\lambda_1, \dots, \lambda_p$. Then $f(D+N) = f(D)$ for all strictly triangular $N \in \mathbb{C}^{n\times n}$ if and only if $f(D) = f(\lambda_1)I$ and*

$$f^{(j)}(\lambda_i) = 0, \quad j = 1\colon k_i - 1. \tag{2.5}$$

Note that (2.5) *is vacuous when $k_i = 1$.*

Proof. ($\Leftarrow$) For any strictly triangular N let $D+N = Z\mathrm{diag}(J_1,\dots,J_q)Z^{-1}$ ($q \geq p$) be the Jordan canonical form of $D+N$ with Jordan blocks

$$J_i = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{m_i \times m_i},$$

where, necessarily, m_i does not exceed the k_j corresponding to λ_i. Then

$$f(D+N) = Z\mathrm{diag}(f(J_1),\dots,f(J_q))Z^{-1},$$

where (from (2.3), for example)

$$f(J_i) = \begin{bmatrix} f(\lambda_i) & f'(\lambda_i) & \dots & \dots & \frac{f^{(m_i-1)}(\lambda_i)}{(m_i-1)!} \\ & f(\lambda_i) & f'(\lambda_i) & \dots & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & f'(\lambda_i) \\ & & & & f(\lambda_i) \end{bmatrix}. \tag{2.6}$$

Since the derivatives of f are zero on any repeated eigenvalue and $f(\lambda_i) = f(\lambda_1)$ for all i, $f(D+N) = Zf(D)Z^{-1} = Zf(\lambda_1)IZ^{-1} = f(\lambda_1)I = f(D)$.

($\Rightarrow$) Let $F = f(D+N)$, and note that by assumption $F = f(D)$ and hence F is diagonal. The equation $F(D+N) = (D+N)F$ reduces to $FN = NF$, and equating (i,j) elements for $j > i$ gives $(f_{ii} - f_{jj})n_{ij} = 0$. Since this equation holds for all strictly triangular N, it follows that $f_{ii} = f_{jj}$ for all i and j and hence that $F = f(\lambda_1)I$.

If at least one of the λ_i is repeated, then we can find a permutation matrix P and a strictly upper bidiagonal matrix B such that $PDP^T + B = P(D + P^TBP)P^T$ is nonderogatory and is in Jordan canonical form, and $N = P^TBP$ is strictly upper triangular. We have $\lambda(D) = \lambda(D+N)$ and so the requirement $f(D+N) = f(D)$ implies that $f(PDP^T + B) = Pf(D)P^T = f(\lambda_1)I$, and hence, in view of (2.6), (2.5) holds. ∎

Applying Theorem 2.2 to the function $f(x) = x^k$ we obtain the following corollary.

COROLLARY 2.3. *Let $D \in \mathbb{C}^{n\times n}$ be a nonzero diagonal matrix and let $k \geq 2$. Then $(D+N)^k = D^k$ for all strictly triangular matrices $N \in \mathbb{C}^{n\times n}$ if and only if*

$$D = \beta\,\mathrm{diag}(e^{2k_1\pi i/k}, e^{2k_2\pi i/k}, \dots, e^{2k_n\pi i/k}),$$

where $\beta \neq 0$, $k_i \in \{0,1,\dots,k-1\}$ and the k_i are distinct (and hence $k \geq n$).

Proof. By Theorem 2.2, all the diagonal elements of D must be kth roots of the same number, β^k say. The condition (2.5) implies that any repeated diagonal element d_{ii} must satisfy $f'(d_{ii}) = kd_{ii}^{k-1} = 0$, which implies $d_{ii} = 0$ and hence $D = 0$; therefore D has distinct diagonal elements. ∎

As a check, we note that the diagonal of M in (2.4) is of the form in the corollary for even powers k. The corollary shows that this phenomenon of very nonmonotonic convergence of the Taylor series can occur when the eigenvalues are a constant multiple of kth roots of unity. As is well known, the computed approximations to multiple

eigenvalues occurring in a single Jordan block tend to have this distribution. We will see in Experiment 4 in section 7 that this eigenvalue distribution also causes problems in finding a good blocking.

We now develop a strict bound for the truncation error of the Taylor series, which we will use to decide when to terminate the series.

THEOREM 2.4 ([13, Thm. 11.2.2]). *Let $Q^*AQ = T = \mathrm{diag}(\lambda_i) + N$ be a Schur decomposition of $A \in \mathbb{C}^{n\times n}$, where N is strictly upper triangular. If $f(z)$ is analytic on a closed convex set Ω whose interior contains $\lambda(A)$, then*

$$\|f(A)\|_\infty \le \left\| \sum_{r=0}^{n-1} \omega_r \frac{|N|^r}{r!} \right\|_\infty \le \max_{0\le r\le n-1} \frac{\omega_r}{r!} \ \|(I-|N|)^{-1}\|_\infty,$$

where

$$\omega_r = \sup_{z\in\Omega} |f^{(r)}(z)|.$$

THEOREM 2.5 ([29, Cor. 2]). *If f has the Taylor series*

$$f(\sigma + y) = \sum_{k=0}^{\infty} \alpha_k y^k, \quad \alpha_k = \frac{f^{(k)}(\sigma)}{k!}$$

for y in an open disk containing the eigenvalues of $Y \in \mathbb{C}^{n\times n}$, then

$$\left\| f(\sigma I + Y) - \sum_{k=0}^{s-1} \alpha_k Y^k \right\|_\infty \le \frac{1}{s!} \max_{0\le t\le 1} \|Y^s f^{(s)}(\sigma I + tY)\|_\infty. \tag{2.7}$$

We need to apply Theorem 2.5 with $Y = M$ in (2.1), and so we need to be able to bound $\max_{0\le t\le 1} \|M^s f^{(s)}(\sigma I + tM)\|_\infty$. The term M^s is needed anyway if we form the next term of the series. To bound $\max_{0\le t\le 1} \|f^{(s)}(\sigma I + tM)\|_\infty$ we can use Theorem 2.4 to show that

$$\max_{0\le t\le 1} \|f^{(s)}(\sigma I + tM)\|_\infty \le \max_{0\le r\le n-1} \frac{\omega_{s+r}}{r!} \ \|(I-|N|)^{-1}\|_\infty, \tag{2.8}$$

where $\omega_{s+r} = \sup_{z\in\Omega} |f^{(s+r)}(z)|$. By using (2.8) in (2.7) we can therefore bound the truncation error. The term $\|(I-|N|)^{-1}\|_\infty$ can be evaluated in just $O(n^2)$ flops[1] for the ∞-norm, since $I-|N|$ is an M-matrix: we solve the triangular system $(I-|N|)y = e$, where $e = [1,\dots,1]^T$, and then $\|y\|_\infty = \|(I-|N|)^{-1}\|_\infty$ [20, sect. 8.3].

We now state our algorithm for evaluating a function of an atomic block via the Taylor series. We denote by u the unit roundoff.

ALGORITHM 2.6 (evaluating function of atomic block). *Given a triangular matrix $T \in \mathbb{C}^{n\times n}$ whose eigenvalues $\lambda_1,\dots,\lambda_n$ are "close," a function f having the Taylor series (2.2) for z in an open disk containing $\lambda_i - \sigma$, $i = 1\colon n$, where $\sigma = n^{-1}\sum_{i=1}^n \lambda_i$, and the ability to evaluate derivatives of f, this algorithm computes $F = f(T)$ using a truncated Taylor series.*

$\sigma = n^{-1}\sum_{i=1}^n \lambda_i$, $M = T - \sigma I$, $\mathrm{tol} = u$

$\mu = \|y\|_\infty$, where y solves $(I-|N|)y = e$ and N is the strictly

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

upper triangular part of T.
$F_0 = f(\sigma)I_n$
$P = M$
for $s = 1\colon\infty$
  $F_s = F_{s-1} + f^{(s)}(\sigma)P$
  $P = PM/(s+1)$
  if $\|F_s - F_{s-1}\|_\infty \le \mathrm{tol}\|F_s\|_\infty$
    % Successive terms are close so check the truncation error bound.
    Estimate or bound $\Delta = \max_{0\le r\le n-1}\omega_{s+r}/r!$, where
    $\omega_{s+r} = \sup_{z\in\Omega}|f^{(s+r)}(z)|$, with Ω a closed convex set containing $\lambda(T)$.
    if $\mu\Delta\|P\|_\infty \le \mathrm{tol}\|F_s\|_\infty$, quit, end if
  end if
end for

Unless we are able to exploit particular properties of f, we can in practice take $\omega_{s+r} = \max\{\,|f^{(s+r)}(\lambda_i)| : \lambda_i \in \lambda(T)\,\}$.

Algorithm 2.6 costs $O(n^4)$ flops, since even if T has constant diagonal, so that M is nilpotent, the algorithm may need to form the first $n-1$ powers of M. Although we usually insist on $O(n^3)$ flops algorithms in numerical linear algebra, this higher order operation count is mitigated by three factors. First, n here is the size of a block, and in most cases the blocks will be of much smaller dimension than the original matrix. Second, M is an upper triangular matrix, so forming all the powers $M^2, \dots, M^{n-1}$ costs $n^4/3$ flops—a factor 6 less than the flop count for multiplying full matrices. Third, for certain particular f the function of the atomic blocks can be evaluated in $O(n^3)$ flops by a method particular to that f.

Since in our overall $f(A)$ algorithm we are not able to impose a fixed bound on the spread $\max_{i,j}|t_{ii} - t_{jj}|$ of the diagonal of T, Algorithm 2.6 is suitable in its stated form only for functions that have a Taylor series with an infinite radius of convergence, such as exp, cos, sin, cosh, and sinh.

We now turn to the effects of rounding errors on Algorithm 2.6. Ignoring truncation errors, standard error analysis [20] shows that the best possible forward error bound is of the form

$$|F - \widehat{F}| \le \frac{nu}{1-nu}\sum_{k=0}^{\infty}\frac{|f^{(k)}(\lambda)|}{k!}|M|^k.$$

If there is heavy cancellation in the sum (2.3), then a large relative error $\|F-\widehat{F}\|/\|F\|$ is possible. This danger is well known, particularly in the case of the matrix exponential [30]. A mitigating factor here is that our matrix T is chosen to have eigenvalues that are clustered, which tends to limit the amount of cancellation in the sum. However, for sufficiently far from normal T, damaging cancellation can take place. For general functions there is little we can do to improve the accuracy; for particular f we can of course apply alternative methods, as illustrated in the next subsection for the logarithm.

2.1. Matrix logarithm. We show how Algorithm 2.6 can be adapted in the important case of the matrix logarithm. We need to evaluate $\log T$, where log denotes the principal logarithm [8] and T is triangular with close eigenvalues. The basic approximation tools at our disposal are a Taylor series and a Padé approximation, both of which are applicable to $\log(I+E)$ with $\|E\| < 1$. We write $\log T = \log(I+E)$, with $E = T-I$. If $\|E\|_\infty \le \theta$, for some tolerance $\theta < 1$, then we will compute a degree

m diagonal Padé approximation to $\log(I+E)$ for a suitable m. If $\|E\|_\infty > \theta$, then we compute the principal square root of T, using the method of Björck and Hammarling [5], and make the same test on the square root. Since $T^{1/2^k} \to I$ as $k \to \infty$, we will eventually be able to apply the Padé approximation, after which we recover the desired logarithm from the relation (see, e.g., [8])

$$\log T = 2^k \log T^{1/2^k}. \tag{2.9}$$

The method we have described is the inverse scaling and squaring method introduced by Kenney and Laub [27]. Note that this method does not exploit the clustered nature of the eigenvalues of T. We might hope to exploit this property by writing $\log T = \log(\alpha \cdot \alpha^{-1}T) = \log(\alpha^{-1}T) + (\log \alpha)I$, where $\alpha = n^{-1}\sum_i t_{ii}$ (say), so that $\operatorname{diag}(\alpha^{-1}T) \approx I$. However, the multivalued nature of the log function can cause the second equality to fail (more precisely, it holds only if some of the logarithms are interpreted as a nonprincipal logarithm) and so we have not pursued this approach.

3. Evaluating the upper triangular part of $f(A)$. We evaluate the upper triangular part of $F = f(T)$ using Parlett's recurrence (1.4), which we rewrite here as

$$T_{ii}F_{ij} - F_{ij}T_{jj} = F_{ii}T_{ij} - T_{ij}F_{jj} + \sum_{k=i+1}^{j-1} (F_{ik}T_{kj} - T_{ik}F_{kj}). \tag{3.1}$$

We assume that T has been reordered and blocked so that T_{ii} and T_{jj} have no eigenvalue in common for all $i \neq j$. This Sylvester equation is therefore nonsingular and it is easy to see that F_{ij} can be computed a column at a time, with each column obtained as the solution of a triangular system. Of particular concern is the propagation of errors in the recurrence. These errors are of two sources: errors in the evaluation of the diagonal blocks F_{ii}, and rounding errors in the formation and solution of (3.1). To gain insight into both types of error we consider the residual of the computed solution $\widehat{F}$:

$$T\widehat{F} - \widehat{F}T =: R, \tag{3.2}$$

where R_{ij} is the residual from the solution of the Sylvester equation (3.1). Although it is possible to obtain precise bounds on R, these are not important to our argument. Writing $\widehat{F} = F + \Delta F$, on subtracting $TF - FT = 0$ from (3.2) we obtain

$$T\Delta F - \Delta F T = R.$$

As for the original equation $TF - FT = 0$, this equation uniquely determines the off-diagonal blocks ΔF in terms of the diagonal blocks. Equating (i,j) blocks yields

$$\begin{aligned} T_{ii}\Delta F_{ij} - \Delta F_{ij}T_{jj} &= R_{ij} + \Delta F_{ii}T_{ij} - T_{ij}\Delta F_{jj} + \sum_{k=i+1}^{j-1} (\Delta F_{ik}T_{kj} - T_{ik}\Delta F_{kj}) \\ &=: B_{ij}, \end{aligned} \tag{3.3}$$

and these equations can be solved to determine ΔF_{ij} a block superdiagonal at a time.

It is straightforward to show that

$$\|\Delta F_{ij}\|_F \leq \operatorname{sep}(T_{ii}, T_{jj})^{-1} \|B_{ij}\|_F, \tag{3.4}$$

where sep is the separation of T_{ii} and T_{jj} [13, sect. 7.2.4], [38],

$$\operatorname{sep}(T_{ii}, T_{jj}) = \min_{X \neq 0} \frac{\|T_{ii}X - XT_{jj}\|_F}{\|X\|_F}.$$

It follows that rounding errors introduced during the stage at which F_{ij} is computed (i.e., represented by R_{ij}) can lead to an error ΔF_{ij} of norm proportional to $\operatorname{sep}(T_{ii}, T_{jj})^{-1}\|R_{ij}\|$. Moreover, earlier errors (represented by the ΔF_{ij} terms on the right-hand side of (3.3)) can be magnified by a factor $\operatorname{sep}(T_{ii}, T_{jj})^{-1}$. It is also clear from (3.3) that even if $\operatorname{sep}(T_{ii}, T_{jj})^{-1}$ is not large, serious growth of errors in the recurrence (3.3) is possible if some off-diagonal blocks T_{ij} are large.

To maximize the accuracy of the computed $f(T)$ we clearly need the blocks T_{ii} to be as well separated as possible in the sense of sep. However, trying to maximize the separations between the diagonal blocks T_{ii} tends to produce larger blocks with less tightly clustered eigenvalues, which increases the difficulty of evaluating $f(T_{ii})$, so any strategy for reordering the Schur form is necessarily a compromise. Moreover, the unitary transformations that produce and then reorder the Schur form may be ill-determined functions of the original matrix A and can be the dominant source of error in the whole computation (see Experiment 9 in section 7), making attempts to maximize the separations ineffective.

Computing $\operatorname{sep}(T_{ii}, T_{jj})$ exactly when both blocks are $m \times m$ costs $O(m^4)$ flops, while condition estimation techniques allow an estimate to be computed at the cost of solving a few Sylvester equations, that is, in $O(m^3)$ flops [7], [18], [25]. It is unclear how to develop a reordering and blocking strategy for producing "large seps" at reasonable cost; in particular, it is unclear how to define "large." Indeed the maximal separations are likely to be connected with the conditioning of $f(T)$, but little or nothing is known about any such connections. More generally, how to characterize matrices for which the condition number of f is large is not well understood, even for the matrix exponential [13, sect. 11.3.1], [23], [37]. Recalling the equivalence mentioned in section 1.3 between block diagonalization and the use of the Parlett recurrence, a result of Gu [14] provides further indication of the difficulty of maximizing the seps: he shows that, given a constant τ, finding a similarity transformation with condition number bounded by τ that block diagonalizes a triangular matrix is NP-hard.

In the next section we will adopt a reordering and blocking strategy that bounds the right-hand side of the approximation

$$\operatorname{sep}(T_{ii}, T_{jj})^{-1} \approx \frac{1}{\min\{\, |\lambda - \mu| : \lambda \in \lambda(T_{ii}),\, \mu \in \lambda(T_{jj}) \,\}}$$

by the reciprocal of a given tolerance. The right-hand side is a lower bound for the left that can be arbitrarily weak, but it is a reasonable approximation for matrices not too far from being normal.

It is natural to look for ways of improving the accuracy of the computed $\widehat{F}$ from the Parlett recurrence. One candidate is fixed precision iterative refinement of the systems (3.1). However, these systems are essentially triangular, and standard error analysis shows that the backward error is already small componentwise [20, Thm. 8.5]; fixed precision iterative refinement therefore cannot help. The only possibility is to use extended precision when solving the systems.

4. Reordering and blocking the Schur form. Given the upper triangular Schur factor T we will reorder it into a partitioned upper triangular matrix $\widetilde{T} = U^*TU = (\widetilde{T}_{ij})$, where U is unitary and two conditions hold:

1. *separation between blocks*:

$$\min\{\,|\lambda-\mu| : \lambda \in \lambda(\widetilde{T}_{ii}),\, \mu \in \lambda(\widetilde{T}_{jj}),\, i \neq j\,\} > \delta, \tag{4.1}$$

2. *separation within blocks*: for every block $\widetilde{T}_{ii}$ with dimension bigger than 1, for every $\lambda \in \lambda(\widetilde{T}_{ii})$ there is a $\mu \in \lambda(\widetilde{T}_{ii})$ with $\mu \neq \lambda$ such that $|\lambda - \mu| \leq \delta$.

Here, $\delta > 0$ is a tolerance. The second property implies that for $\widetilde{T}_{ii} \in \mathbb{R}^{m\times m}$ $(m > 1)$

$$\max\{\,|\lambda-\mu| : \lambda, \mu \in \lambda(\widetilde{T}_{ii}),\, \lambda \neq \mu\,\} \leq (m-1)\delta,$$

and this bound is attained when, for example, $\lambda(\widetilde{T}_{ii}) = \{\delta, 2\delta, \ldots, m\delta\}$.

The following algorithm is the first step in obtaining the ordering. It can be interpreted as finding the connected components of the graph on the eigenvalues of T in which there is an edge between two nodes if the corresponding eigenvalues are a distance at most δ apart.

ALGORITHM 4.1 (block pattern). *Given a triangular matrix $T \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_i \equiv t_{ii}$ and a tolerance $\delta > 0$, this algorithm produces a block pattern, defined by an integer vector q, for the block version of Parlett's method: the eigenvalue λ_i is assigned to the set S_{q_i}, and it satisfies the conditions that $\min\{|\lambda_i - \lambda_j| : \lambda_i \in S_p, \lambda_j \in S_q, p \neq q\} > \delta$ and, for each set S_i with more than one element, every element of S_i is within distance at most δ from some other element in the set. For each such set S_q, all the eigenvalues in S_q are intended to appear together in an upper triangular block $\widetilde{T}_{ii}$ of $\widetilde{T} = U^*TU$.*

```
p = 1
Initialize the S_p to empty sets.
for i = 1:n
    if λ_i ∉ S_q for all 1 ≤ q < p
      Assign λ_i to S_p.
      p = p + 1
    end if
    for j = i + 1:n
        Denote by S_{q_i} the set that contains λ_i.
        if λ_j ∉ S_{q_i}
          if |λ_i − λ_j| ≤ δ
            if λ_j ∉ S_k for all 1 ≤ k < p
              Assign λ_j to S_{q_i}.
            else
              Move the elements of S_{max(q_i,q_j)} to S_{min(q_i,q_j)}.
              Reduce by 1 the indices of sets S_q for q > max(q_i, q_j).
              p = p − 1
            end if
          end if
        end if
    end for
end for
```

Algorithm 4.1 provides a mapping from each eigenvalue λ_i of T to an integer q_i such that the set S_{q_i} contains λ_i. Our remaining problem is equivalent to finding a method for swapping adjacent elements in q to obtain a confluent permutation q'. A confluent permutation of n integers, $q_1, \ldots, q_n$, is a permutation such that any repeated integers q_i are next to each other. For example, there are 3! confluent

permutations of $(1,2,1,3,2,1)$ which include $(1,1,1,3,2,2)$ and $(3,2,2,1,1,1)$. Ideally we would like a confluent permutation that requires a minimal number of swaps to transform q to q'. Ng [31] notes that finding such a permutation is an NP-complete problem. He proves that the minimum number of swaps required to obtain a given confluent permutation is bounded above by $\frac{n^2}{2}(1-\frac{1}{k})$, where k is the number of distinct q_i, and that this bound is attainable [31, Thm. A.1]. In practice, since the QR algorithm tends to order the eigenvalues by absolute value in the Schur form, complicated strategies for determining a confluent permutation are not needed. The following method works well in practice: find the average index of the integers in q and then order the integers in q' in ascending average index. If we take our example $(1,2,1,3,2,1)$ and let g_k denoted the average index of the integer k, we see that $g_1 = (1+3+6)/3 = 3\frac{1}{3}$, $g_2 = (2+5)/2 = 3\frac{1}{2}$, and $g_3 = 4$. Therefore we try to obtain the confluent permutation $q' = (1,1,1,2,2,3)$ by a sequence of swaps of adjacent elements:

$$
\begin{aligned}
q &= (1,2,1,3,2,1) \to (1,1,2,3,2,1) \\
&\to (1,1,2,3,1,2) \qquad (4.2)\\
&\to (1,1,2,1,3,2) \qquad (4.3)\\
&\to (1,1,1,2,3,2) \qquad (4.4)\\
&\to (1,1,1,2,2,3) = q'.
\end{aligned}
$$

Swapping adjacent diagonal elements of T requires $20n$ flops, plus another $20n$ flops to update the Schur vectors, so the cost of the swapping is $40n$ times the number of swaps. The total cost is usually small compared with the overall cost of the algorithm.

Having determined the blocking and the desired confluent permutation we can make repeated calls to the LAPACK routine `xTREXC` [1] to obtain it. This routine applies a unitary similarity transformation to move the diagonal element of T with row index $j =$ `IFST` to row $i =$ `ILST`, which is achieved by performing a sequence of $|j-i|$ swaps of adjacent diagonal elements. For example, if $j > i$, the diagonal of T has the ordering

$$\dots, \lambda_{i-1}, \lambda_j, \lambda_i, \lambda_{i+1}, \dots, \lambda_{j-1}, \lambda_{j+1} \qquad (4.5)$$

after application of `xTREXC`. Notice that swaps (4.2)–(4.4) can be achieved through one call to the LAPACK routine `xTREXC` by requesting that $\lambda_6 \in S_1$ be moved to row 3. The following algorithm is expressed with MATLAB indexing notation for conciseness.

ALGORITHM 4.2 (obtaining a confluent permutation). *Given a vector $q \in \mathbb{R}^n$ containing all the integers $1,\dots,k$ (some repeated if $k < n$), this algorithm obtains a confluent permutation according to the average indices of the integers in q. Returned is a swapping strategy, stored in vectors* `ILST` *and* `IFST`*, to be used by the LAPACK routine* `xTREXC` *to obtain a block form of T.*

Let $\phi(j)$ denote the number of j's in q. $\beta = 1$.
for $i = 1\colon k$
  $g_i = (\sum_{q_j = i} j)/\phi(i)$
end for
Sort g into ascending order $g_{y_1} \le \cdots \le g_{y_k}$, where y is an index vector.
for $i = y$
  if any$(q(\beta\colon \beta + \phi(i) - 1) \ne i)$

$f = \text{find}(q == i); g = \beta\colon \beta + \phi(i) - 1$
Concatenate $g(f \sim= g)$ and $f(f \sim= g)$ to the end of `ILST` and `IFST`, respectively.
Let $v = \beta\colon f(\text{end})$ and delete all elements of v that are elements of f.
$q(g(\text{end}) + 1\colon f(\text{end})) = q(v)$
$q(g) = [i, \ldots, i]$
$\beta = \beta + \phi(i)$
end if
end for

The routine `xTREXC` implements the swapping algorithm of Bai and Demmel [3], which has guaranteed backward stability and, since we are swapping only 1×1 blocks, always succeeds.

5. Overall algorithm. Our complete Schur algorithm for computing $f(A)$ is as follows.

ALGORITHM 5.1 (Computing $f(A)$). *Given $A \in \mathbb{C}^{n\times n}$, a function f analytic on a closed convex set Ω whose interior contains the eigenvalues of A, and the ability to evaluate derivatives of f, this algorithm computes $F = f(A)$.*

Compute the Schur decomposition $A = QTQ^*$ (Q unitary, T upper triangular).
If T is diagonal, $F = f(T)$, goto $(*)$, end if
Using Algorithm 4.1 with $\delta = 0.1$, assign each eigenvalue λ_i to a set S_{q_i}.
Apply Algorithm 4.2 to the vector q to produce a swapping strategy
in `ILST` and `IFST`.
for $k = 1\colon \text{length}(\texttt{ILST})$
call `xTREXC(V, n, T, n, Q, n, IFST(k), ILST(k), info)`
end for
% Now $A = QTQ^*$ is our reordered Schur decomposition, with block $m \times m$ T.
for $i = 1\colon m$
Use Algorithm 2.6 to evaluate $F_{ii} = f(T_{ii})$.
for $j = i - 1\colon -1\colon 1$
Solve the Sylvester equation in (3.1) for F_{ij}.
end for
end for
$(*)$ $F = QFQ^*$

The cost of Algorithm 5.1 depends greatly on the eigenvalue distribution of A, and is roughly between $28n^3$ flops and $n^4/3$ flops. Note that Q, and hence F, can be kept in factored form, with a significant computational saving. This is appropriate if F needs just to be applied to a few vectors, for example.

Note that we have set the blocking parameter $\delta = 0.1$, which our experiments indicate is as good a default choice as any. This optimal choice of δ in terms of cost or accuracy is problem-dependent.

Algorithm 5.1 has a property noted as being desirable by Parlett and Ng [34]: it acts simply on simple cases. Specifically, if A is normal, so that the Schur decomposition is $A = QDQ^*$ with D diagonal, the algorithm simply evaluates $f(A) = Qf(D)Q^*$. At another extreme, if A has just one eigenvalue of multiplicity n, then the algorithm works with a single block, $T_{11} \equiv T$, and evaluates $f(T_{11})$ via its Taylor series expanded about the eigenvalue.

If we specialize to the matrix logarithm and use the inverse scaling and squaring method in place of Algorithm 2.6, as described in section 2, Algorithm 5.1 is similar to a Schur method for the matrix logarithm proposed by Dieci, Morini, and Papini

[10]. The main difference is that in the latter paper the eigenvalues are ordered in the Schur form by increasing modulus and then the Schur form is blocked, without any further reordering, so that (4.1) holds; this tends to lead to larger blocks than Algorithm 4.1. (Consider, for example, the case where $\delta = 0.1$ and the diagonal of T is 1, i, $-i$, 1.1, for which the ordering of [10] produces one 4×4 block, whereas Algorithm 2.6 produces one 2×2 block and two 1×1 blocks.)

6. Preprocessing. In an attempt to improve the accuracy of Algorithm 5.1 we might try to preprocess the data before applying a particular stage of the algorithm. Two techniques that have been used in the past, notably in Ward's implementation of the scaling and squaring algorithm for computing the matrix exponential [39], are translation and diagonal scaling, and in [39] their purpose is to reduce the norm of the matrix.

Translation has no effect on our algorithm. Algorithm 2.6 for evaluating the Taylor series already translates the diagonal blocks, and further translations before applying the Parlett recurrence are easily seen to have no effect, because (3.1) is invariant under translations $T \to T - \alpha I$ and $F \to F - \beta I$.

A diagonal similarity transformation could be applied at any stage of the algorithm and then undone later. For example, such a transformation could be used in conjunction with Parlett's recurrence in order to make $U := D^{-1}TD$ less nonnormal than T and to increase the separations between diagonal blocks. In fact, by choosing D of the form $D = \mathrm{diag}(\theta^{n-1}, \dots, 1)$ we can make U arbitrarily close to diagonal form. Unfortunately, no practical benefit is gained: Parlett's recurrence involves solving triangular systems and the substitution algorithm is invariant under diagonal scalings (at least, as long as they involve only powers of the machine base). Similar comments apply to the evaluation of the Taylor series in Algorithm 2.6.

A diagonal similarity transformation may be beneficial at the outset, prior to computing the Schur decomposition. One can balance A with the aid of the standard balancing algorithm used in conjunction with the QR algorithm (function `balance` in MATLAB); this algorithm computes $B = D^{-1}AD$, where D is chosen so that the norm of the ith row and ith column are of similar magnitude for all i. Ward's algorithm [39] uses an initial balancing. Balancing is a heuristic that is not guaranteed to lead to a more accurate result. We omit balancing from Algorithm 5.1, while recognizing that it is potentially useful when we are dealing with badly scaled matrices.

7. Numerical experiments. Our experiments were carried out in MATLAB 6.5 (R13) on a Pentium IV, for which the unit roundoff $u \approx 1.1 \times 10^{-16}$. Our implementation of Algorithm 5.1 comprises several M-files and a MEX file that calls the LAPACK routine `ZTREXC` (we call the LAPACK binary supplied with MATLAB). Unless otherwise stated, $\delta = 0.1$ in Algorithm 4.1.

In computing errors we take for the "exact" $f(A)$ an approximation X computed at high precision using MATLAB's Symbolic Math Toolbox (which invokes the Maple kernel). The (relative or forward) error in $\widehat{X}$ is defined to be

$$\|X - \widehat{X}\|_\infty / \|X\|_\infty.$$

In certain applications the componentwise relative error $\max_{i,j}(|x_{ij} - \widehat{x}_{ij}|/|x_{ij}|)$ might be of interest. However, while componentwise accuracy is potentially achievable in evaluating $f(T)$, the subsequent similarity transformation by Q will, in general, destroy any special structure in the error and lead at best to a small normwise error.

TABLE 7.1
Errors for Experiment 1: $A =$ `gallery('triw',8)`.

	Algorithm 5.1	`funm`
A	4.5e-16	7.0e-1
$A +$ `rand(n)*1e-8`	6.4e-15	1.2e-10
$A +$ `triu(rand(n))*1e-8`	3.4e-16	2.2e44

We also quote the (relative) condition number

$$\operatorname{cond}(A, f) = \lim_{\epsilon \to 0} \max_{\|E\|_2 \le \epsilon \|A\|_2} \frac{\|f(A+E) - f(A)\|_2}{\epsilon \|f(A)\|_2},$$

which we estimate using the finite-difference power method proposed by Kenney and Laub [27].

We present ten experiments that give insight into the many facets of the $f(A)$ problem and our particular algorithm.

Experiment 1. Our first experiment shows the importance of using a block form of the Parlett recurrence. We take A to be the 8×8 triangular matrix with $a_{ii} \equiv 1$ and $a_{ij} \equiv -1$ for $j > i$, which is MATLAB's `gallery('triw',8)`. With f the exponential, Table 7.1 shows the errors for A and two small perturbations of A, one full and one triangular. The condition number of $f(A)$ is about 2 in each case, so we would expect to be able to compute $f(A)$ accurately. Algorithm 5.1 provides very good accuracy. MATLAB 6.5's `funm`, which employs the point version of the Parlett recurrence, performs badly, as expected in view of the repeated or close eigenvalues. This is an extreme example, in that Algorithm 5.1 takes just one block, the whole Schur factor T, and so reduces to evaluating the Taylor series of T.

Experiment 2. It is easy to show numerically the need for the safeguard in the test in Algorithm 2.6 for terminating the Taylor series. For the matrix

$$T = \begin{bmatrix} 0.5 & 10^{12} \\ 0 & -0.5 \end{bmatrix}$$

Algorithm 5.1 evaluates the exponential with error less than u, treating the matrix as one block and taking 10 terms of the Taylor series. If the Taylor series evaluation is terminated solely based on comparison of successive terms, thus omitting the derivative test in Algorithm 2.6, then only 4 terms are taken and the error is 5×10^{-8}.

Experiment 3. We give an example to show that for the exponential function Algorithm 5.1 can be much more accurate than the scaling and squaring method implemented in MATLAB's `expm`. We take the upper triangular matrix

```
T = gallery('triw',4,2^(60)) - diag([17 17 2 2]),
```

which has diagonal elements $-16, -16, -1, -1$ and off-diagonal elements $2^{60} \approx 11 \times 10^{18}$. This badly scaled matrix causes great difficulty for `expm`, which yields a relative error of order 100. Algorithm 5.1 chooses the blocking (1: 2), (3: 4) (with no reordering) and produces a result correct to machine precision. We note that the more sophisticated scaling strategy proposed in [11] would improve the accuracy of the scaling and squaring method. The significance of this experiment is that it shows that our general purpose method can be significantly more accurate than one of the best available e^A implementations.

Experiment 4. The next experiment shows how Algorithm 5.1 can behave in an unstable manner. We compute e^T, where the upper triangular T is generated by the MATLAB code

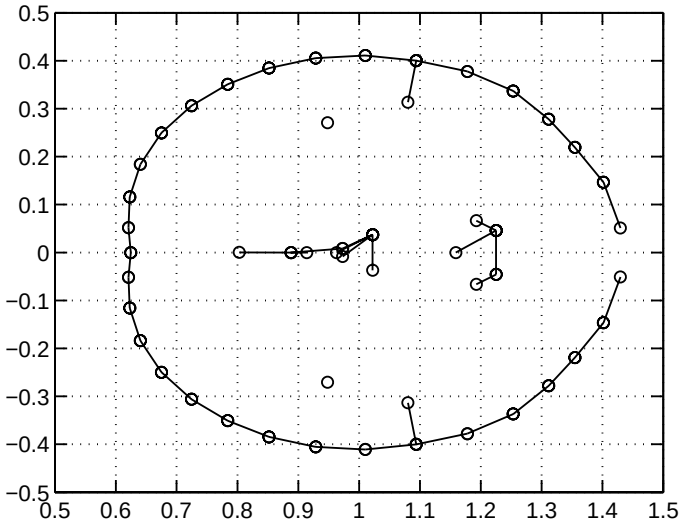


FIG. 7.1. *Eigenvalue distribution for Experiment* 4. *Circles denote eigenvalues and eigenvalues in the same block are joined by lines.*

```
n = 50; randn('state',1)
B = triu(randn(n),1) + eye(n);
Q = gallery('orthog',n);
B = Q*B*Q'; T = schur(B,'complex')
```

Although T has n eigenvalues 1 if formed in exact arithmetic, the computed T has eigenvalues mainly lying on and in an approximate circle of radius 0.4 centered on $(1, 0)$. Algorithm 5.1 requires just 2 swaps to produce the block pattern

$$(1\colon 35) : 25 \text{ terms}, \quad (36\colon 36), \quad (37\colon 37), \quad (38\colon 42) : 11 \text{ terms}, \quad (43\colon 50) : 13 \text{ terms},$$

where the number of terms required in the Taylor series evaluation of each nontrivial diagonal block is shown. Figure 7.1 shows the eigenvalues and the blocking: the eigenvalues are represented by circles and a path is drawn between two eigenvalues if they belong to the same block. The condition number is $\mathrm{cond}(T, f) \approx 293$, but the error is 7×10^{-4}. Some insight is provided by Tables 7.2 and 7.3, which show, with T now the *reordered* Schur form, the blockwise errors $\|X_{ij} - \widehat{X}_{ij}\|_\infty / \|X_{ij}\|_\infty$ and the separations $\mathrm{sep}(T_{ii}, T_{jj})$ for $i \neq j$. The blocks with largest errors lie off the diagonal in the first block row and correspond to very small values of sep. This is not surprising in view of the bound (3.4).

An interesting feature of this example is that if we increase δ to 0.2, then Algorithm 5.1 chooses just one block and so calculates the exponential by a Taylor series of the whole of T, giving a result with error $1.4 \times 10^{-14} < \mathrm{cond}(T, f)u$. Figure 7.2 gives further insight by showing δ plotted against the error. The data for this plot was generated in such a way that all values of δ at which the blocking changes are included. The error is of order 10^{-4} for all δ until the first δ for which only one block is chosen. It seems that for this example any attempt to split eigenvalues into different blocks has a disastrous effect on the error.

Experiment 5. The previous experiment might suggest that it is better to overestimate δ. However, the graph of δ versus error can be U-shaped. Consider the exponential of minus the upper triangular Schur factor of the 50×50 Frank ma-

TABLE 7.2
Errors in blocks X_{ij} computed by Algorithm 5.1 *in Experiment* 4.

			j		
i	1	2	3	4	5
1	1.6e-14	1.9e-6	2.9e-6	2.3e-5	2.0e-3
2		1.1e-14	5.4e-15	6.6e-15	2.2e-12
3			2.1e-14	1.1e-14	5.4e-13
4				1.0e-14	4.8e-13
5					4.5e-14

TABLE 7.3
Values of $\operatorname{sep}(T_{ii}, T_{jj})$ *for Experiment* 4.

			j		
i	1	2	3	4	5
1		2.2e-12	2.2e-12	3.4e-12	2.0e-13
2			5.4e-1	4.4e-2	8.3e-3
3				4.4e-2	8.3e-3
4					3.4e-5

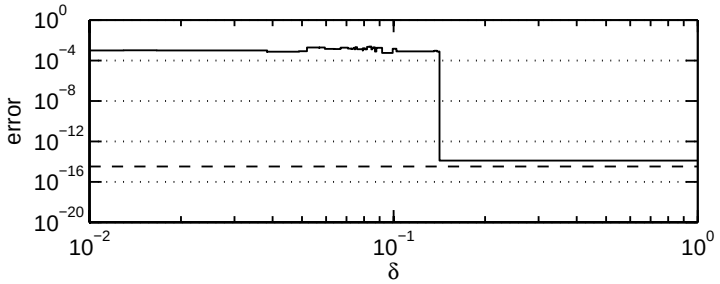


FIG. 7.2. *Blocking parameter δ versus error for Experiment* 4. *Dotted line denotes level of* $\operatorname{cond} u$.

trix (MATLAB's `gallery('frank',50)`), for which $\operatorname{cond}(A, f) \approx 2 \times 10^3$. Figure 7.3 shows the corresponding δ versus error plot; the error is near minimal for $2.3 \lesssim \delta \lesssim 5.2$ and increases rapidly outside this range.

Experiment 6. The next experiment shows that Algorithm 5.1 can fail to behave in a stable way for *all* choices of δ. The matrix is a 65×65 upper triangular matrix T constructed in MATLAB by

```
A = -schur(gallery('frank',125),'complex')/2;
i = [26:60 96:125]; T = A(i,i)
```

Figure 7.4 plots δ versus the error for the exponential function; the error is always at least 10^{-10}, which is three orders of magnitude greater than $\operatorname{cond}(T, f)u$. Note, however, that varying δ does not generate all possible blockings, so we cannot rule out the possibility that the Schur–Parlett method is stable on this example for some other blocking. The following experiment provides further insight.

Experiment 7. For any particular matrix, it is of interest to know which blocking produces the most accurate computed result. We can answer this question experimentally by testing all possible blockings. The number S_n of blocking patterns for an $n \times n$ matrix can be shown to be

$$S_n = \sum_{k=1}^{n} S_n^{(k)},$$

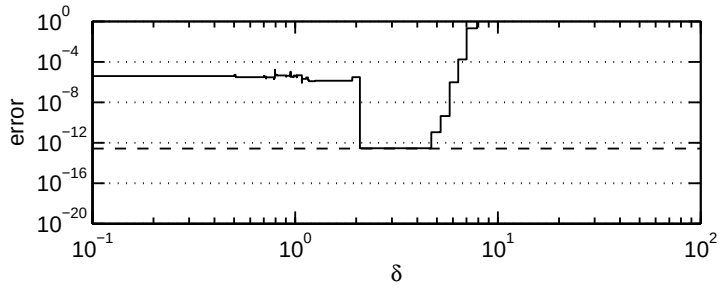


FIG. 7.3. *Blocking parameter δ versus error for Experiment* 5. *Dotted line denotes level of* cond u.

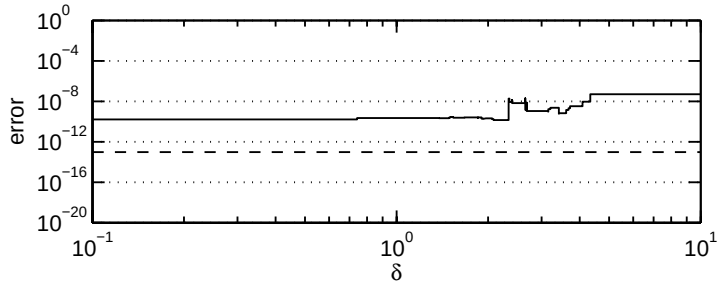


FIG. 7.4. *Blocking parameter δ versus error for Experiment* 6. *Dotted line denotes level of* cond u.

where $S_n^{(k)}$ is the number of ways a set of n elements can be partitioned into k disjoint, nonempty subsets. The numbers S_n and $S_n^{(k)}$ are known as Bell numbers and Stirling numbers of the second kind, respectively. The S_n grow very quickly, so it is feasible to try all orderings only for small n. We describe an example with $n = 10$, for which there are $S_{10} = 115,975$ different blockings. We generate an upper triangular T with the MATLAB code

```
n = 10;
mu = 0.2; phi = 5;
randn('state',0)
B = phi*triu(randn(n),1) + eye(n);
Q = gallery('orthog',n); B = Q*B*Q';
[U,T] = schur(B,'complex');
d = diag(T - eye(n)); delta = abs(d(1)-d(2));
T(1:n+1:n^2) = mu/delta*d + ones(n,1).
```

The computed eigenvalues of T lie approximately equally spaced on a circle center 1, radius 0.3.

Again, the function is the exponential, for which the condition number for this problem is 1.1×10^2. The results can be summarized as follows.

- Algorithm 5.1 chooses all 1×1 blocks and produces an error 2.8×10^{-10}.
- For the trivial blocking $\{1\colon 10\}$, the error is 3.9×10^{-16}. This blocking is produced by Algorithm 5.1 when δ is increased to 0.2.
- The other 115,974 nontrivial blockings produce errors ranging from 8.7×10^{-12} (for the blocking $(1\colon 5)$, $(6\colon 10)$) to 3.0×10^{-9} (for the blocking $(1\colon 2)$, $(3\colon 4)$, $(5\colon 6)$, 7, $(8\colon 10)$).

- For comparison, MATLAB's `expm` produces an error 1.4×10^{-14}.

Thus only the trivial blocking produces a computed result with error bounded by a small multiple of $\mathrm{cond}\, u$. This example shows that the block Parlett recurrence can fail to behave in a forward stable way for *all* nontrivial blockings.

Experiment 8. Now we consider the matrix cosine function. A method specialized to this function is proposed by Serbin and Blalock [35] (see also [13, sect. 11.2.3]). The idea is to compute $\cos(A)$ by scaling by a power of 2 to produce a matrix with norm of order 1, approximate the cosine of the scaled matrix, then use the double-angle formula to recover the cosine of the original matrix:

$C_0 =$ Taylor series approximation to $\cos(A/2^k)$
for $j = 1\colon k$,
  $C_j = 2C_{j-1}^2 - I$
end.

Here, we have specified a Taylor series approximation, though alternatives such as Padé approximants could also be used. Although some analysis of the method is given in [35], how to choose k and the degree of the Taylor approximation to strike a balance between minimizing the truncation error, rounding errors, and the computational effort is not understood. We have therefore implemented the following approach: we run the method with $k = 0\colon \lceil 2\log_2 \|A\|_1 \rceil$ and with the Taylor series evaluated with convergence tolerance u and record the smallest error observed. In other words, we find the most accurate solution that the method can provide for a wide range of k.

For the 6×6 Pascal matrix (MATLAB's `pascal(6)`), which has ∞-norm 462 Algorithm 5.1 produces a computed solution with error 9.0×10^{-15}; since this matrix is symmetric Algorithm 5.1 simply evaluates the cosine function on the diagonal matrix of eigenvalues. The double-angle method produces minimum error 8.5×10^{-13}, which is achieved for $k = 6$ and using 35 terms of the Taylor series.

For the MATLAB matrix $A =$ `gallery('invol',8)*pi`, which has ∞-norm of order 10^6 and eigenvalues $\pm\pi$, so that $\cos(A) = I$, the relative error for Algorithm 5.1 is 4.73×10^{-11}, resulting from the blocking $(1\colon 4)$, $(5\colon 8)$ with 4 Taylor series terms for each block (with no reordering). If just one block is taken, then 35 Taylor series terms are required and the error is about 6 times larger. The minimum error from the double-angle method is 8.6×10^{-14}, achieved for $k = 15$ and using 3 terms of the Taylor series. Interestingly, the error for $k = 0$, which evaluates $\cos(A)$ directly from the Taylor series, is 9.0×10^{-14}, while $k = 20 \approx \log_2(\|A\|_\infty)$ (which is suggested in [13]) produces a much larger error 2.0×10^{-11}. The condition number $\mathrm{cond}(A, f)$ is of order 10^8.

Our conclusion from this experiment is that Algorithm 5.1 is competitive in accuracy with the double-angle method, even when the optimal k is chosen for the latter method.

Experiment 9. Next we consider the matrix logarithm. In Algorithm 5.1 we use the inverse scaling and squaring method in place of Algorithm 2.6, as described in section 2; we take $\theta = 0.25$ and $m = 8$ and evaluate the Padé approximant by a partial fraction expansion, as recommended in [19]. We take the matrix $A = ZJZ^{-1}$ from [4], where

$$J = \mathrm{diag}(1, J_3(1), 0.3, 0.4, 0.5, 0.6, 0.7, 0.8),$$

with $J_m(\lambda)$ an $m \times m$ Jordan block with eigenvalue λ, and Z is a random matrix with condition number 10^8. The reordered Schur triangular factor, denoted by T, is blocked $(1\colon 1)$, $(2\colon 2)$, $(3\colon 3)$, $(4\colon 4)$, $(5\colon 5)$, $(6\colon 7)$, $(8\colon 10)$. The error in the computed

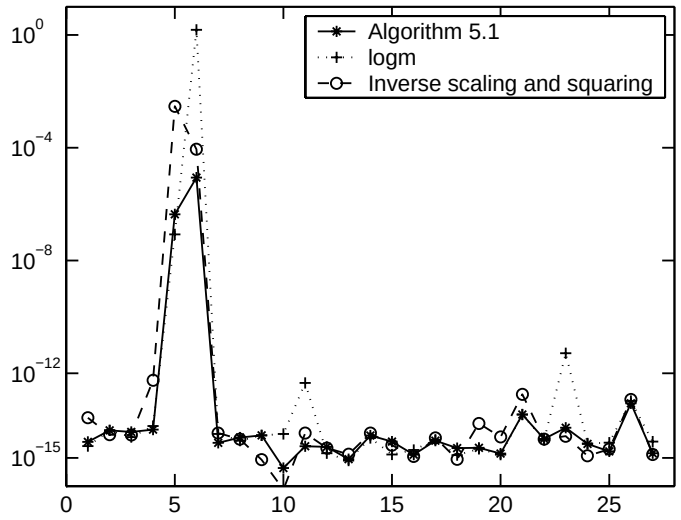


FIG. 7.5. *Error measure* (7.1) *for* 26 13×13 *test matrices.*

$X = \log A$ is $8 \times 10^{-4} \approx \mathrm{cond}(A, f)u$. However the error in the computed $\log T$ is only 1×10^{-14}, which is consistent with the fact that $\min_{i \neq j} \mathrm{sep}(T_{ii}, T_{jj}) = 1 \times 10^{-4}$. In this example, then, the error is dominated by the error introduced by the unitary transformations, and the error in the evaluation of the $\log T_{ii}$ and in the block Parlett recurrence is negligible, by comparison. Even if we evaluate $\log T$ to full working accuracy, the unitary back-transformations increase the error to the level 10^{-3} once again. This illustrates that although unitary transformations are perfectly backward stable, they can be the dominant source of forward error in Algorithm 5.1.

Experiment 10. In the final experiment we use the quantity

$$\beta = \frac{\|A - e^{\widehat{X}}\|_\infty}{\|A\|_\infty}, \tag{7.1}$$

where $\widehat{X}$ is the computed logarithm of A, to test the quality of three matrix logarithm methods: Algorithm 5.1 specialized to the logarithm as in the previous experiment, MATLAB 6.5's `logm` (which is essentially `funm` applied to the log function), and an implementation of the inverse scaling and squaring method that computes a Schur decomposition, takes square roots as necessary of the full triangular factor, and then computes a degree 8 diagonal Padé approximation. We use 27 13×13 matrices obtained from the function `matrix` in the Matrix Computation Toolbox [16]; these matrices include test matrices from MATLAB itself. The results, in Figure 7.5, show that Algorithm 5.1 performs at least as well as the other two logarithm methods for these test matrices.

8. Conclusions. Algorithm 5.1 is applicable to a wide range of functions and imposes no restrictions on the matrix. It requires $O(n^3)$ flops unless close or repeated eigenvalues force large blocks to be chosen when the Schur form is blocked, in which case the operation count can be up to $n^4/3$ flops. The algorithm needs to evaluate derivatives of the function when there are blocks of dimension greater than 1. This is a price to be paid for catering for general functions and nonnormal matrices with possibly repeated eigenvalues.

The algorithm has a parameter δ that is used to determine the reordering and blocking of the Schur form. This parameter serves to balance the conflicting requirements of producing small diagonal blocks and keeping the separations of the blocks large. It is unclear how to choose δ to (nearly) maximize the accuracy of the computed $f(A)$. Indeed it is an open problem to understand fully the conditioning of general matrix functions, and a good choice of δ is likely to require knowledge of the conditioning. Our default choice of $\delta = 0.1$ performs well much of the time. The most difficult cases for our algorithm are when a substantial subset of the computed eigenvalues are approximately equally spaced on a circle in the complex plane, in which case the default δ may yield an unnecessarily inaccurate result. The option of running the algorithm with several different δ is not usually helpful in practice, because for most f we have no way to judge the quality of a computed $f(A)$ without comparing it with the exact answer. Moreover, it is possible that for all choices of δ the error is greater than the condition of the problem warrants (see Experiment 6). Nevertheless, as our numerical experiments make clear, even specialized methods, such as the scaling and squaring method for the matrix exponential, can behave unstably on certain examples, and Algorithm 5.1 is competitive with all the specialized algorithms to which we have compared it experimentally.

Our MATLAB implementation of Algorithm 5.1 is more robust and numerically reliable than MATLAB 6.5's `funm`, which ignores the dangers of close or repeated eigenvalues and always uses the point version of the Parlett recurrence. We hope that this implementation will serve as a benchmark with which to compare both specific $f(A)$ routines and other general purpose routines.

9. Acknowledgments. We thank Françoise Tisseur and Alessandra Papini for their comments on a draft manuscript. The comments of the referees were also very helpful.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, PA, 1999.

[2] K. J. Åström and B. Wittenmark, *Computer-Controlled Systems: Theory and Design*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 1997.

[3] Z. Bai and J. W. Demmel, *On swapping diagonal blocks in real Schur form*, Linear Algebra Appl., 186 (1993), pp. 73–95.

[4] C. A. Bavely and G. W. Stewart, *An algorithm for computing reducing subspaces by block diagonalization*, SIAM J. Numer. Anal., 16 (1979), pp. 359–367.

[5] Å. Björck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[6] B. A. Borgias, M. Gochin, D. J. Kerwood, and T. L. James, *Relaxation matrix analysis of 2D NMR data*, Progress in NMR Spectroscopy, 22 (1990), pp. 83–100.

[7] R. Byers, *A LINPACK-style condition estimator for the equation* $AX - XB^T = C$, IEEE Trans. Automat. Control, 29 (1984), pp. 926–928.

[8] S. H. Cheng, N. J. Higham, C. S. Kenney, and A. J. Laub, *Approximating the logarithm of a matrix to specified accuracy*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1112–1125.

[9] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.

[10] L. Dieci, B. Morini, and A. Papini, *Computational techniques for real logarithms of matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 570–593.

[11] L. Dieci and A. Papini, *Padé approximation for the exponential of a block triangular matrix*, Linear Algebra Appl., 308 (2000), pp. 183–202.

[12] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[14] M. Gu, *Finding well-conditioned similarities to block-diagonalize nonsymmetric matrices is NP-hard*, J. Complexity, 11 (1995), pp. 377–391.
[15] T. F. Havel, I. Najfeld, and J. Yang, *Matrix decompositions of two-dimensional nuclear magnetic resonance spectra*, Proc. Natl. Acad. Sci. USA, 91 (1994), pp. 7962–7966.
[16] N. J. Higham, *The Matrix Computation Toolbox*, http://www.ma.man.ac.uk/~higham/mctoolbox (5 September 2002).
[17] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.
[18] N. J. Higham, *FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm* 674*)*, ACM Trans. Math. Software, 14 (1988), pp. 381–396.
[19] N. J. Higham, *Evaluating Padé approximants of the matrix logarithm*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1126–1135.
[20] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, PA, 2002.
[21] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[22] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numer., 9 (2000), pp. 215–365.
[23] B. Kågström, *Bounds and perturbation bounds for the matrix exponential*, BIT, 17 (1977), pp. 39–57.
[24] B. Kågström, *Numerical computation of matrix functions*, Report UMINF-58.77, Department of Information Processing, University of Umeå, Sweden, 1977.
[25] B. Kågström and P. Poromaa, *Distributed and shared memory block algorithms for the triangular Sylvester equation with* sep^{-1} *estimators*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 90–101.
[26] B. Kågström and A. Ruhe, *An algorithm for numerical computation of the Jordan normal form of a complex matrix*, ACM Trans. Math. Software, 6 (1980), pp. 398–419.
[27] C. Kenney and A. J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.
[28] P.-F. Lavallée, A. Malyshev, and M. Sadkane, *Spectral portrait of matrices by block diagonalization*, in Numerical Analysis and Its Applications, L. Vulkov, J. Waśniewski, and P. Yalamov, eds., Lecture Notes in Comput. Sci. 1196, Springer-Verlag, Berlin, 1997, pp. 266–273.
[29] R. Mathias, *Approximation of matrix-valued functions*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1061–1063.
[30] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.
[31] K. C. Ng, *Contributions to the Computation of the Matrix Exponential*, Ph.D. thesis, Technical report PAM-212, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, 1984.
[32] B. N. Parlett, *Computation of functions of triangular matrices*, Memorandum ERL-M481, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, CA, 1974.
[33] B. N. Parlett, *A recurrence among the elements of functions of triangular matrices*, Linear Algebra Appl., 14 (1976), pp. 117–121.
[34] B. N. Parlett and K. C. Ng, *Development of an accurate algorithm for* $\exp(Bt)$, Technical report PAM-294, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, 1985.
[35] S. M. Serbin and S. A. Blalock, *An algorithm for computing the matrix cosine*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 198–204.
[36] M. I. Smith, *A Schur algorithm for computing matrix pth roots*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 971–989.
[37] C. Van Loan, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971–981.
[38] J. M. Varah, *On the separation of two matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 216–222.
[39] R. C. Ward, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600–610.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 486–494

LEAST SQUARES SOLUTION OF $BXA^T = T$ OVER SYMMETRIC, SKEW-SYMMETRIC, AND POSITIVE SEMIDEFINITE X*

YUAN-BEI DENG†, XI-YAN HU†, AND LEI ZHANG†

Abstract. An efficient method based on the quotient singular value decomposition (QSVD) is used to solve the constrained least squares problem $\min \|T - BXA^T\|_F$ over symmetric, skew-symmetric, and positive semidefinite (maybe asymmetrical) X. The general expression of the solution is given and some necessary and sufficient conditions are derived about the solvability of the matrix equation $BXA^T = T$. In each case, an algorithm is given for the unique solution when B and A are of full column rank.

Key words. matrix equation, least squares problem, QSVD factorization, Hadamard product

AMS subject classifications. 65F15, 65F20

DOI. 10.1137/S0895479802402491

1. Introduction.

In this paper, we consider the constrained least squares approximation problem: Find

$$\min_{X \in S} \|T - BXA^T\|_F \tag{1.1}$$

and the consistency of the related linear matrix equation

$$BXA^T = T, \qquad X \in S, \tag{1.2}$$

where B, A, T are given matrices, $S \subseteq R^{n \times n}$, and $\|Y\|_F$ denotes the Frobenius norm of a real matrix Y, defined as

$$\|Y\|_F^2 = \langle Y, Y \rangle = \sum_{i,j} y_{ij}^2,$$

where the inner product is given by $\langle A, B \rangle = trace(A^T B)$.

This work is concerned with the least squares solution of (1.1) when S is the set of symmetric, skew-symmetric, or positive semidefinite (maybe asymmetrical) matrices. By using the quotient singular value decomposition (QSVD) factorization, we obtain the general expression of the solution, and some necessary and sufficient conditions are derived about the solvability of (1.2) with a given structure. Besides, in each case an algorithm is proposed for the unique solution when B and A are of full column rank.

In [10], the least squares solution of the equation

$$(A \otimes B)x = t$$

is considered, where $x = vec(X)$, $t = vec(T)$, the vec operator stacks the columns of a matrix, and a method based on the QR decomposition is developed by using

*Received by the editors February 11, 2002; accepted for publication (in revised form) by L. Eldén February 7, 2003; published electronically September 17, 2003. The work was supported by the National Natural Science Foundation of China.
http://www.siam.org/journals/simax/25-2/40249.html
†College of Mathematics and Econometrics, Hunan University, Changsha 410082, People's Republic of China (ybdeng@hnu.net.cn, xyhu@hnu.net.cn).

the Kronecker product. The least squares problem is equivalent to the unconstrained least squares problem (see [22])

$$\min_X \|T - BXA^T\|_F$$

when B and A are assumed to be of full column rank.

The unconstrained and constrained least squares problems have been of interest for many applications, including particle physics and geology [11], control theory, the inverse Sturm–Liouville problem [13], inverse problems of vibration theory [17], digital image and signal processing, photogrammetry, finite elements, and multidimensional approximation [10].

Don [7], Magnus [18], Chu [5], and Hua [16] discussed (1.2), where the solution matrix is known to have a given structure (e.g., symmetric, triangular, diagonal), either directly from the matrix equation or indirectly from the equivalent vector equation, but they did not consider the least squares problem of the equation. For the least squares problem, the case $B = A = I$ was treated for some matrices by [14] and [8], and in the case $A = I$, Higham [15], Allwright and Woodgate [1, 2], and Woodgate [20] obtained the symmetric and symmetric positive semidefinite solution and derived some algorithms, respectively; also, [9] and [3] considered the case $A = I$ for several types of convex cones, and [12] discussed the case on sphere.

Our notation is as follows: $R^{m\times n}$ is the set of all $m \times n$ real matrices, $SR^{n\times n}$, $AR^{n\times n}$, and $OR^{n\times n}$ are the sets of all symmetric, skew-symmetric, and orthogonal matrices in $R^{n\times n}$, respectively. We denote the set of positive semidefinite matrices by

$$R_0^{n\times n} = \{A \in R^{n\times n} | x^T Ax \geq 0 \quad \text{for all} \quad x \in R^n\}$$

and the set of symmetric positive semidefinite matrices by

$$SR_0^{n\times n} = \{A \in R^{n\times n} | A = A^T, \quad x^T Ax \geq 0 \ \text{ for all } \ x \in R^n\}.$$

It is obvious that $SR_0^{n\times n}$ is a proper subset of $R_0^{n\times n}$. The Moore–Penrose generalized inverse of matrix A is denoted by A^+, and $A * B$ and $A \bigotimes B$ represent the Hadamard product and the Kronecker product of A and B, respectively.

In the following sections, we always suppose $B \in R^{m\times n}$, $A \in R^{p\times n}$, $T \in R^{m\times p}$ are given, and $X \in R^{n\times n}$.

The QSVD of a pair of matrices B and A is related in the following theorem.

QSVD Theorem (see [14]). *Let $B \in R^{m\times n}$, $A \in R^{p\times n}$. Then there exist orthogonal matrices $U \in R^{m\times m}$, $V \in R^{p\times p}$ and a nonsingular matrix $Y \in R^{n\times n}$ such that*

$$U^T BY = \Sigma_B, \quad V^T AY = \Sigma_A, \tag{1.3}$$

where

$$\Sigma_B = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & S_{AB} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} r \\ s \\ m-r-s \end{matrix} \\ \begin{matrix} r & s & k-r-s & n-k \end{matrix}, \tag{1.4}$$

$$\Sigma_A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_{AB} & 0 & 0 \\ 0 & 0 & I_A & 0 \end{pmatrix} \begin{matrix} p+r-k \\ s \\ k-r-s \end{matrix} \\ \begin{matrix} r & s & k-r-s & n-k \end{matrix}, \tag{1.5}$$

$$k = \operatorname{rank}(B^T, A^T), \quad r = k - \operatorname{rank}(A),$$
$$s = \operatorname{rank}(B) + \operatorname{rank}(A) - k, \quad S_{AB} = \operatorname{diag}(\sigma_1, \ldots, \sigma_s),$$
$$\sigma_i > 0 \quad (i = 1, \ldots, s).$$

When B and A are of full column rank, i.e., $r(B) = r(A) = n$, then $r = 0, s = n, k = n$, and

$$\Sigma_B = \begin{pmatrix} S_{AB} \\ 0 \\ n \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix}, \qquad \Sigma_A = \begin{pmatrix} 0 \\ I_{AB} \\ n \end{pmatrix} \begin{matrix} p-n \\ n \end{matrix}. \tag{1.6}$$

2. The solutions of (1.1) when $S = SR^{n\times n}$ or $AR^{n\times n}$. In this section we start with a lemma in order to prove the main results (cf. [19]).

LEMMA 2.1. *Suppose that $G \in R^{s\times s}, \Sigma_0 = \operatorname{diag}(\sigma_1, \ldots, \sigma_s), \sigma_i > 0$ $(i = 1, \ldots, s)$; then there exist a unique $S_s \in SR^{s\times s}$ and a unique $S_a \in AR^{s\times s}$ such that*

$$\|\Sigma_0 S - G\|_F = \min \tag{2.1}$$

and

$$S_s = \phi * (G^T\Sigma_0 + \Sigma_0 G), \tag{2.2}$$

$$S_a = \phi * (\Sigma_0 G - G^T\Sigma_0), \tag{2.3}$$

where

$$\phi = (\varphi_{ij}) \in SR^{s\times s}, \quad \varphi_{ij} = \frac{1}{\sigma_i^2 + \sigma_j^2}, \quad 1 \le i, j \le s. \tag{2.4}$$

Proof. We prove only the existence of S_a and (2.3). For any $S = (s_{ij}) \in AR^{s\times s}, G = (g_{ij}) \in R^{s\times s}$, since $s_{ii} = 0, s_{ij} = -s_{ji}$,

$$\begin{aligned}\|\Sigma_0 S - G\|_F^2 &= \sum_{i=1}^{s}\sum_{j=1}^{s}(\sigma_i s_{ij} - g_{ij})^2 \\ &= \sum_{i=1}^{s} g_{ii}^2 + \sum_{1\le i<j\le s}[(\sigma_i^2 + \sigma_j^2)s_{ij}^2 + 2(\sigma_j g_{ji} - \sigma_i g_{ij})s_{ij} + (g_{ij}^2 + g_{ji}^2)].\end{aligned}$$

Hence, there exists a unique solution $S_a = (\hat{s}_{ij}) \in AR^{s\times s}$ for (2.1) such that

$$\hat{s}_{ij} = \frac{\sigma_i g_{ij} - \sigma_j g_{ji}}{\sigma_i^2 + \sigma_j^2}, \quad 1 \le i, j \le s.$$

This is (2.3). ∎

Now suppose the QSVD of the matrix pair $[B, A]$ is (1.3), (1.4), and (1.5) and let

$$Y^{-1}XY^{-T} = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{pmatrix} \begin{matrix} r \\ s \\ k-r-s \\ n-k \end{matrix},$$
$$\begin{matrix} r & s & k-r-s & n-k \end{matrix}$$

$$U^TTV = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} \begin{matrix} r \\ s \\ m-r-s \end{matrix};$$
$$\begin{matrix} p+r-k & s & k-r-s \end{matrix}$$

then

$$\begin{aligned}
&\|BXA^T - T\|_F^2 \\
&= \left\|U\Sigma_B Y^{-1} X Y^{-T} \Sigma_A^T V^T - T\right\|_F^2 \\
&= \left\|\Sigma_B (Y^{-1} X Y^{-T}) \Sigma_A^T - (U^T T V)\right\|_F^2 \\
&= \left\| \begin{pmatrix} 0 & X_{12} & X_{13} \\ 0 & S_{AB}X_{22} & S_{AB}X_{23} \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{pmatrix} \right\|_F^2 \\
&= \|S_{AB}X_{22} - T_{22}\|_F^2 + \|X_{12} - T_{12}\|_F^2 + \|X_{13} - T_{13}\|_F^2 + \|S_{AB}X_{23} \\
&\quad - T_{23}\|_F^2 + \|T_{11}\|_F^2 + \|T_{21}\|_F^2 + \|T_{31}\|_F^2 + \|T_{32}\|_F^2 + \|T_{33}\|_F^2.
\end{aligned} \tag{2.5}$$

About the symmetric solution of (1.1), we have the following theorem.

THEOREM 2.2. (1) *The least squares symmetric solution X_s of* (1.1) *has the general form*

$$X_s = Y \begin{bmatrix} X_{11} & T_{12} & T_{13} & X_{14} \\ T_{12}^T & \hat{X}_{22} & S_{AB}^{-1}T_{23} & X_{24} \\ T_{13}^T & (S_{AB}^{-1}T_{23})^T & X_{33} & X_{34} \\ X_{14}^T & X_{24}^T & X_{34}^T & X_{44} \end{bmatrix} Y^T, \tag{2.6}$$

where $\hat{X}_{22} = \phi * (T_{22}^T S_{AB} + S_{AB}T_{22}), \phi$ *is taken by* (2.4), *and* $X_{11} \in SR^{r\times r}, X_{33} \in SR^{(k-r-s)\times(k-r-s)}, X_{44} \in SR^{(n-k)\times(n-k)}$, $X_{14} \in R^{r\times(n-k)}$, $X_{24} \in R^{s\times(n-k)}$, $X_{34} \in R^{(k-r-s)\times(n-k)}$ *are arbitrary.*

(2) *The system* (1.2) *with* $S = SR^{n\times n}$ *is consistent if and only if*

$$T_{11}, T_{21}, T_{31}, T_{32}, T_{33}$$

are zero submatrices and

$$(S_{AB}^{-1}T_{22})^T = S_{AB}^{-1}T_{22},$$

in which case the general solution is expressed by (2.6), *where* $\widetilde{X}_{22} = S_{AB}^{-1}T_{22}$ *instead of* $\hat{X}_{22}$.

Proof. (1) From (2.5), $X_s \in SR^{n\times n}$ and

$$\|BX_sA^T - T\|_F^2 = \min_{X\in SR^{n\times n}} \|BXA^T - T\|_F^2$$

hold if and only if $X_{ij}^T = X_{ji}$, and

$$\|S_{AB}X_{22} - T_{22}\|_F^2 = \min, \quad \|X_{12} - T_{12}\|_F^2 = \min,$$

$$\|X_{13} - T_{13}\|_F^2 = \min, \quad \|S_{AB}X_{23} - T_{23}\|_F^2 = \min.$$

Therefore $X_{12} = T_{12}, X_{13} = T_{13}, X_{23} = S_{AB}^{-1}T_{23}$, and by (2.1) and (2.2) of Lemma 2.1, $X_{22} = \hat{X}_{22}$.

(2) $BXA^T = T$ with $X^T = X$ if and only if there exists $X \in SR^{n\times n}$ such that $\|BXA^T - T\|_F = 0$. According to (2.5), we obtain that $T_{11}, T_{21}, T_{31}, T_{32}$, and T_{33} are zero matrices, $X_{12} = T_{12}, X_{13} = T_{13}$, $X_{23} = S_{AB}^{-1}T_{23}$, and $X_{22} = \widetilde{X}_{22} = S_{AB}^{-1}T_{22}$ with $\widetilde{X}_{22}^{\ T} = \widetilde{X}_{22}$. This proves Theorem 2.2. □

COROLLARY 2.3. *When B and A are of full column rank, we write*

$$U^TTV = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix} ; \quad \begin{matrix} p-n & n \end{matrix} \tag{2.7}$$

then the following hold:

(1) *There is a unique least squares symmetric solution X_f of* (1.1) *and*

$$X_f = Y[\phi * (T_2^T S_{AB} + S_{AB}T_2)]Y^T.$$

(2) *$BXA^T = T$ with $X^T = X$ is consistent if and only if T_1, T_3, T_4 are zero submatrices and $(S_{AB}^{-1}T_2)^T = S_{AB}^{-1}T_2$, in which case the system has a unique solution X_e and*

$$X_e = YS_{AB}^{-1}T_2Y^T.$$

Proof. Let $\overline{X} = Y^{-1}XY^{-T}$. Using the QSVD (1.3) and (1.6) of matrix pair $[B, A]$, we have

$$\|BXA^T - T\|_F^2 = \|T_1\|_F^2 + \|T_3\|_F^2 + \|T_4\|_F^2 + \|S_{AB}\overline{X} - T_2\|_F^2. \tag{2.8}$$

The remainder of the proof is the same as that of Theorem 2.2. □

In the next step, the skew-symmetric solution of (1.1) is given.

THEOREM 2.4. (1) *The least squares skew-symmetric solution X_a of* (1.1) *has the general form*

$$X_a = Y \begin{pmatrix} X_{11} & T_{12} & T_{13} & X_{14} \\ -T_{12}^T & \hat{X}_0 & S_{AB}^{-1}T_{23} & X_{24} \\ -T_{13}^T & -(S_{AB}^{-1}T_{23})^T & X_{33} & X_{34} \\ -X_{14}^T & -X_{24}^T & -X_{34}^T & X_{44} \end{pmatrix} Y^T, \tag{2.9}$$

where $\hat{X}_0 = \phi * (S_{AB}T_{22} - T_{22}^T S_{AB})$, ϕ is taken by (2.4), *and $X_{11} \in AR^{r\times r}, X_{33} \in AR^{(k-r-s)\times(k-r-s)}, X_{44} \in AR^{(n-k)\times(n-k)}, X_{14} \in R^{r\times(n-k)}, X_{24} \in R^{s\times(n-k)}, X_{34} \in R^{(k-r-s)\times(n-k)}$ are arbitrary.*

(2) *The system* (1.2) *with $S = AR^{n\times n}$ is consistent if and only if*

$$T_{11}, T_{21}, T_{31}, T_{32}, T_{33}$$

are zero submatrices and

$$(S_{AB}^{-1}T_{22})^T = -S_{AB}^{-1}T_{22},$$

in which case the general solution can be expressed by (2.9), *where $\widetilde{X}_0 = S_{AB}^{-1}T_{22}$ instead of $\hat{X}_0$.*

From (2.5), it is seen that the proof of Theorem 2.4 is similar to that of Theorem 2.2.

COROLLARY 2.5. *When B and A are of full column rank, by using the same notation as in Corollary* 2.3*, we have the following:*

(1) *There is a unique least squares skew-symmetric solution W_f of* (1.1) *and*

$$W_f = Y[\phi * (S_{AB}T_2 - T_2^T S_{AB})]Y^T.$$

(2) *$BXA^T = T$ with $X^T = -X$ is consistent if and only if T_1, T_3, T_4 are zero submatrices and $(S_{AB}^{-1}T_2)^T = -S_{AB}^{-1}T_2$, in which case the system has a unique solution W_e and*

$$W_e = YS_{AB}^{-1}T_2Y^T.$$

The proof of Corollary 2.5 is similar to that of Corollary 2.3.

When B and A are of full column rank, the symmetric solution X_f and skew-symmetric solution W_f of (1.1) can be computed in the following way:

1. Find out the QSVD (1.3), (1.6) of $[B, A]$ according to [6] while U, V, Y, and S_{AB} are obtained.
2. Partition U^TTV by (2.7).
3. $\phi := (\varphi_{ij})$, $\varphi_{ij} := \frac{1}{\sigma_i^2 + \sigma_j^2}$, $1 \le i, j \le n$.
4. $X_f := Y[\phi * (T_2^T S_{AB} + S_{AB}T_2)]Y^T$, $W_f := Y[\phi * (S_{AB}T_2 - T_2^T S_{AB})]Y^T$.

3. The solution of (1.1) over $S = R_0^{n\times n}$ when B and A are of full column rank. First we introduce a result about the optimal approximation on the Hilbert space.

Suppose V is a real Hilbert space $\langle ., . \rangle$ denotes the inner product, $\|u\|_V = \sqrt{\langle u, u\rangle}$ is the norm on V, $K \subset V$ is a nonempty closed convex cone with the vertex at the origin.

$$K^\perp = \{u | u \in V, \langle u, k\rangle = 0 \quad \text{for all } k \in K\},$$

$K^{\perp\perp} = (K^\perp)^\perp$, and

$$K^* = \{u \in K^{\perp\perp} | \langle u, k\rangle \ge 0 \quad \text{for all } k \in K\}.$$

It is known that $K^\perp, K^{\perp\perp}$ are closed linear subspaces in V, $K \subset K^{\perp\perp}$; if $K = K^{\perp\perp}$, then $K^* = \{0\}$.

LEMMA 3.1 (see [23]). *For every given $u \in V$, there exist unique u_0, u_1, u_2 with $u_0 \in K^\perp, u_1 \in K$, and $u_2 \in K^*$ such that*

$$u = u_0 + u_1 - u_2, \tag{3.1}$$

$$\langle u_1, u_2\rangle = 0, \quad \langle u_0, u_i\rangle = 0, \quad i = 1, 2, \tag{3.2}$$

$$\|u - u_1\|_V \le \|u - v\|_V \quad \text{for all } v \in K, \tag{3.3}$$

$$\|u + u_2\|_V \le \|u + v\|_V \quad \text{for all } v \in K^*. \tag{3.4}$$

Remark. It can be seen that for $u \in V$, if we know u_0 and find u_2 from (3.4), then we can find u_1 from (3.1).

When $D = \text{diag}(d_1, \ldots, d_n), d_i > 0$ $(i = 1, \ldots, n)$, in $R^{n\times n}$, a new inner product is defined as follows:

$$(A, B)_D = (DA, DB) = \text{tr}(B^T D^2 A),$$

$$\|A\|_D = \sqrt{(A, A)_D} = \sqrt{\text{tr}(A^T D^2 A)}.$$

Then $R^{n\times n}$ with the inner product $(.,.)_D$ is a Hilbert space, which is denoted by $R_D^{n\times n}$.

In $R_D^{n\times n}$, if we take $K = R_0^{n\times n}$, then K is a closed convex cone and the following result holds; the proof is similar to that of Lemma 2.7 in [21] (see also [24]).

LEMMA 3.2. *In $R_D^{n\times n}$, we have*

$$(R_0^{n\times n})^{\perp} = \{0\}, \quad (R_0^{n\times n})^{\perp\perp} = R^{n\times n},$$

$$(R_0^{n\times n})^{*} = \{M \in R^{n\times n} | M = D^{-2}H \ \text{ for all } H \in SR_0^{n\times n}\}.$$

Hence, the following theorem holds from Lemmas 3.1 and 3.2.

THEOREM 3.3. *For every $F \in R_D^{n\times n}$, there exist unique $F_1 \in R_0^{n\times n}$ and $\overline{H} \in SR_0^{n\times n}$ such that*

$$F = F_1 - D^{-2}\overline{H}, \tag{3.5}$$

$$(F_1, D^{-2}\overline{H})_D = 0, \tag{3.6}$$

$$\|F - F_1\|_D = \min_{G\in R_0^{n\times n}} \|F - G\|_D, \tag{3.7}$$

$$\|F + D^{-2}\overline{H}\|_D = \min_{H\in SR_0^{n\times n}} \|F + D^{-2}H\|_D. \tag{3.8}$$

At last we give the positive semidefinite solution of (1.1) and the algorithm when B and A are of full column rank.

THEOREM 3.4. *Suppose B and A are of full column rank, the QSVD of $[B, A]$ is determined by* (1.3) *and* (1.6)*, and U^TTV is partitioned by* (2.7)*; then there exists a unique least squares positive semidefinite solution $X_p \in R_0^{n\times n}$ of the problem* (1.1) *which can be expressed as*

$$X_p = Y\overline{X}_pY^T,$$

where $\overline{X}_p = S_{AB}^{-1}T_2 + S_{AB}^{-2}\overline{H}$, and $\overline{H}$ is the solution of the optimal approximation problem

$$\min_{H\in SR_0^{n\times n}} \|S_{AB}^{-1}H + T_2\|_F. \tag{3.9}$$

In addition, $BXA^T = T$ with $X \in R_0^{n\times n}$ is consistent if and only if T_1, T_3, T_4 are zero submatrices and $S_{AB}^{-1}T_2 \in R_0^{n\times n}$, in which case the solution of (1.2) *is $YS_{AB}^{-1}T_2Y^T$.*

Proof. From (2.8) we know that X_p satisfies

$$\|BX_pA^T - T\|_F = \min_{X\in R_0^{n\times n}} \|BXA^T - T\|_F$$

if and only if $\overline{X}_p$ satisfies

$$\|S_{AB}\overline{X}_p - T_2\|_F = \min_{\overline{X}\in R_0^{n\times n}} \|S_{AB}\overline{X} - T_2\|_F.$$

Let $D = S_{AB}, F = S_{AB}^{-1}T_2$; then $\overline{X}_p$ is the solution of the problem

$$\min_{\overline{X}\in R_0^{n\times n}} \|F - \overline{X}\|_D. \tag{3.10}$$

If we know that $\overline{H}$ is the solution of (3.9), then $\overline{H}$ is also the solution of (3.8). Hence from the remark and Theorem 3.3, the solution of (3.10) is

$$\overline{X}_p = S_{AB}^{-1} T_2 + S_{AB}^{-2} \overline{H}.$$

For the case $BXA^T = T$ with $X \in R_0^{n\times n}$, a similar conclusion can be obtained from (2.8). The theorem is proved. □

From (2.8), we also have the following result.

COROLLARY 3.5. *Suppose the conditions are the same as those in Theorem* 3.2; *then there exists a unique least squares symmetric positive semidefinite solution* $\overline{X}_{sp} \in SR_0^{n\times n}$ *in* (1.1) *and*

$$X_{sp} = Y \overline{X}_{sp} Y^T,$$

where $\overline{X}_{sp}$ *is the solution of the minimum problem*

$$\min_{\overline{X} \in SR_0^{n\times n}} \|S_{AB}\overline{X} - T_2\|_F. \tag{3.11}$$

In addition, $BXA^T = T$ *with* $X \in SR_0^{n\times n}$ *is consistent if and only if* T_1, T_3, T_4 *are zero submatrices and* $S_{AB}^{-1} T_2 \in SR_0^{n\times n}$, *in which case the solution of* (1.2) *is* $Y S_{AB}^{-1} T_2 Y^T$.

The authors of [1, 9, 20, 3] have discussed the algorithms for solving the problem (3.9) or (3.11) in detail. Now we also give an algorithm to compute the positive semidefinite solution X_p and symmetric positive semidefinite solution X_{sp} of (1.1) when B and A are of full column rank.

ALGORITHM 3.1.

1. Find out the $QSVD$ (1.3), (1.6) of $[B, A]$ according to [6] while U, V, Y, and S_{AB} are obtained.
2. Partition $U^T T V$ by (2.7).
3. Determine the solution $\overline{H}$ of (3.9) and the solution $\overline{X}_{sp}$ of (3.11) according to [9] or [20].
4. $X_p := Y(S_{AB}^{-1} T_2 + S_{AB}^{-2} \overline{H}) Y^T$, $\overline{X}_{sp} := Y \overline{X}_{sp} Y^T$.

4. Conclusions. This paper is concerned with a class of Procrustes problems, where the solution is required to be symmetric, skew-symmetric, or positive semidefinite (maybe asymmetrical). The solution is based on the $QSVD$ factorization of the matrix pair $[B, A]$, which is used to reduce the problem to one with a given diagonal matrix.

Acknowledgment. The authors thank the anonymous referees very much for their helpful comments.

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] J. C. ALLWRIGHT AND K. G. WOODGATE, *Erratum and addendum: "Positive semidefinite matrices: Characterization via conical and least-squares solution of a matrix equation,"* SIAM J. Control Optim., 28 (1990), pp. 250–251.

[3] L.-E. ANDERSSON AND T. ELFVING, *A constrained Procrustes problem*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 124–139.

[4] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications*, Wiley, New York, 1974.

[5] K.-W. E. Chu, *Symmetric solutions of linear matrix equations by matrix decompositions*, Linear Algebra Appl., 119 (1989), pp. 35–50.
[6] D. Chu and B. De Moor, *On a variational formulation of the QSVD and the RSVD*, Linear Algebra Appl., 311 (2000), pp. 61–78.
[7] F. J. H. Don, *On the symmetric solution of a linear matrix equation*, Linear Algebra Appl., 93 (1987), pp. 1–7.
[8] R. Escalante and M. Raydan, *Dykstra's algorithm for a constrained least-squares matrix problem*, Numer. Linear Algebra Appl., 3 (1996), pp. 459–472.
[9] R. Escalante and M. Raydan, *Dykstra's algorithm for constrained least-squares rectangular matrix problems*, Comput. Math. Appl., 35 (1998), pp. 73–79.
[10] D. W. Fausett and C. T. Fulton, *Large least squares problems involving Kronecker products*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 219–227.
[11] G. H. Golub and J. H. Welsch, *Calculation of Gauss quadrature rules*, Math. Comput., 23 (1969), pp. 221–230.
[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.
[13] O. Hald, *On Discrete and Numerical Sturm-Liouville Problems*, Ph.D. dissertation, Department of Mathematics, New York University, New York, 1972.
[14] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.
[15] N. J. Higham, *The symmetric Procrustes problem*, BIT, 28 (1988), pp. 133–143.
[16] D. Hua, *On the symmetric solutions of linear matrix equations*, Linear Algebra Appl., 131 (1990), pp. 1–7.
[17] D. Hua and P. Lancaster, *Linear matrix equations from an inverse problem of vibration theory*, Linear Algebra Appl., 246 (1996), pp. 31–47.
[18] J. R. Magnus, *L-structured matrices and linear matrix equations*, Linear and Multilinear Algebra, 14 (1983), pp. 67–88.
[19] J. Sun, *Two kinds of inverse eigenvalue problems for real symmetric matrices*, Math. Numer. Sinica, 3 (1988), pp. 282–290.
[20] K. G. Woodgate, *Least-squares solution of $F = PG$ over positive semidefinite symmetric P*, Linear Algebra Appl., 245 (1996), pp. 171–190.
[21] D.-X. Xie, *Least-squares solution for inverse eigenpair problem of nonnegative definite matrices*, Comput. Math. Appl., 40 (2000), pp. 1241–1251.
[22] H. Y. Zha, *Comments on "Large least squares problems involving Kronecker products,"* SIAM J. Matrix Anal. Appl., 16 (1995), p. 1172.
[23] L. Zhang, *The approximation on the closed convex cone and its numerical application*, Hunan Ann. Math., 6 (1986), pp. 16–22.
[24] L. Zhang, *A class of inverse eigenvalue problem for symmetric nonnegative matrices*, J. Hunan Educ. Inst., 13 (1995), pp. 11–17.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 495–506

LOW RANK PERTURBATION OF JORDAN STRUCTURE*

JULIO MORO† AND FROILÁN M. DOPICO†

Abstract. Let A be a matrix and λ_0 be one of its eigenvalues having g elementary Jordan blocks in the Jordan canonical form of A. We show that for most matrices B satisfying $\operatorname{rank}(B) \leq g$, the Jordan blocks of $A+B$ with eigenvalue λ_0 are just the $g-\operatorname{rank}(B)$ smallest Jordan blocks of A with eigenvalue λ_0. The set of matrices for which this behavior does not happen is explicitly characterized through a scalar determinantal equation involving B and some of the λ_0-eigenvectors of A. Thus, except for a set of zero Lebesgue measure, a low rank perturbation $A+B$ of A destroys for each of its eigenvalues exactly the $\operatorname{rank}(B)$ largest Jordan blocks of A, while the rest remain unchanged.

Key words. Jordan canonical form, matrix spectral perturbation theory

AMS subject classifications. 15A18, 15A21

DOI. 10.1137/S0895479802417118

1. Introduction. It is well known [1, 4] that the multiple eigenvalues of a matrix split typically under perturbation into simple, distinct eigenvalues. If A is the unperturbed matrix, then each Jordan block of dimension k of A gives rise to a so-called *ring* or *cycle* [4, section II.1.2] of k different simple eigenvalues of the perturbed matrix, say $A+B$. This typical behavior takes place for sufficiently small B provided a certain genericity condition is satisfied by the perturbation (see [12, 5, 7] for more details).

In this paper we study a class of perturbations B which are only able to break some, but not all, of the Jordan blocks of A, namely perturbations with low rank. To be more precise, let λ_0 be an eigenvalue of A with geometric multiplicity g, i.e., $g = \dim\ker(A-\lambda_0 I)$, where ker denotes the null space and I is the identity matrix. By "low" rank we will mean in what follows that the rank of B satisfies

$$\operatorname{rank}(B) \leq g. \tag{1.1}$$

It is easy to check that this kind of perturbation cannot break all g Jordan blocks: using the elementary facts that $\operatorname{rank}(A+B-\lambda_0 I) \leq \operatorname{rank}(A-\lambda_0 I) + \operatorname{rank}(B)$ and $\operatorname{rank}(A-\lambda_0 I) = \operatorname{rank}(A+B-\lambda_0 I - B) \leq \operatorname{rank}(A+B-\lambda_0 I) + \operatorname{rank}(B)$, one easily gets

$$g - \operatorname{rank}(B) \leq \dim\ker(A+B-\lambda_0 I) \leq g + \operatorname{rank}(B). \tag{1.2}$$

Since every Jordan block corresponds to one independent eigenvector, the previous inequality implies that the perturbation B can destroy at most $\operatorname{rank}(B)$ of the Jordan blocks of A and can create at most $\operatorname{rank}(B)$ new Jordan blocks associated with each eigenvalue of A. This constraint still allows for a great deal of freedom as to the number and dimensions of the Jordan blocks of $A+B$. The purpose of this paper is to find out which is the most usual behavior in this respect.

*Received by the editors October 31, 2002; accepted for publication (in revised form) by I. C. F. Ipsen March 7, 2003; published electronically September 17, 2003. This research was partially supported by the Ministerio de Ciencia y Tecnología of Spain through grant BFM 2000-0008.
http://www.siam.org/journals/simax/25-2/41711.html
†Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911-Leganés, Spain (jmoro@math.uc3m.es, dopico@math.uc3m.es).

The following naive argument sheds light on the question: for most B's, the equality $\operatorname{rank}(A+B-\lambda_0 I) = \operatorname{rank}(A-\lambda_0 I) + \operatorname{rank}(B)$ holds, and consequently $\dim\ker(A+B-\lambda_0 I) = g - \operatorname{rank}(B)$. Hence, in most cases $A+B$ will have *exactly* $\operatorname{rank}(B)$ fewer Jordan blocks with eigenvalue λ_0 than A. Furthermore, the larger the size of a Jordan block, the more algebraic conditions are needed to ensure its existence, so the largest Jordan blocks should be more sensitive to perturbation than the smaller ones. According to this argument, the *generic* behavior one would expect for most perturbations B is that, *for each eigenvalue λ_0 of A satisfying* (1.1)*, precisely the* $\operatorname{rank}(B)$ *largest Jordan blocks of A corresponding to that eigenvalue are destroyed in the Jordan form of $A+B$, and the other Jordan blocks of A persist as Jordan blocks of $A+B$.*

Of course this hand-waving argument does not always hold true, as shown in the following examples. An appropriately chosen "nontypical" rank one perturbation can increase the size of the Jordan blocks corresponding to $\lambda_0 = 1$ in

$$A+B = \left[\begin{array}{cc|cc|c} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{cccc|c} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

or it may increase the number of Jordan blocks associated with λ_0, as in

$$A+B = \left[\begin{array}{cc|cc|c} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|cc|c} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

However, we will see that, in both cases, very special structures of the perturbation are needed to produce these unusual behaviors.

The main contribution of this paper is to obtain, for any matrix A and each eigenvalue λ_0, a simple, explicit characterization of the set of perturbations B for which the previously described typical behavior occurs. The necessary and sufficient condition for this is simply that a single scalar quantity, denoted by C_0, is not equal to zero. The scalar C_0 is defined through a sum of determinants of matrices involving B and some of the λ_0-eigenvectors of A. As a trivial consequence, the set of perturbations B for which the generic behavior does not happen, i.e., those fulfilling $C_0 = 0$, is an algebraic manifold of zero Lebesgue measure in the set of $n \times n$ complex matrices of given rank. This precise mathematical formulation allows us to term properly the expected behavior described above as *generic*.

The problem we address here was solved when the perturbation B has rank equal to one by Savchenko [9]. In fact, Savchenko conjectured without proof in [9] the generic behavior for perturbations of arbitrary rank. This conjecture motivated our work, leading first to the partial answer given in [8, section 3.2.1] and ultimately to the present paper. Recently, Savchenko [10] has found an independent (and different) proof of the results we present here. Both in [9] and in [10], the proofs rely on functional analytic techniques based on spectral resolvents. Our approach, based only on elementary linear algebra results, is probably better suited for the matrix analysis community. However, the approach in [9, 10] might be more amenable to extend results of this nature to infinite-dimensional operators.

An important point to be made is that all theorems below are valid for perturbations B *of any size*, i.e., they are by no means first-order perturbation results. This makes especially surprising the prominent role of the scalar C_0, a quantity which closely resembles the quantities defining the genericity conditions in first-order eigenvalue perturbation theory [5, 7]. In this respect, the results we present below are related to previous contributions in the context of first-order perturbation theory, dealing with perturbations restricted to some nongeneric manifold. Some preliminary results for nongeneric perturbations may be found in [7, section 3] as an extension of Lidskii's [5] classical results for generic perturbations, but the first systematic description of a class of structured perturbations was obtained by Ma and Edelman [6] for upper k-Hessenberg perturbations of Jordan blocks. More recently, Jeannerod [3] has extended Lidskii's results by obtaining explicit formulas for both the leading exponents and leading coefficients of the Puiseux expansions of the eigenvalues of *analytic* perturbations $J + B(\varepsilon)$ of a Jordan matrix J, provided the powers of ε in the perturbation matrix $B(\varepsilon)$ conform in a certain way to the Jordan structure given by J. However, in both cases [6, 3] the particular structure of the perturbations to the Jordan blocks is not preserved by undoing the change of basis leading to the Jordan form. Hence, not much information is provided for nongeneric perturbations of *arbitrary* matrices. The rank of the perturbation, on the other hand, does not change by undoing the Jordan change of basis. Therefore, to our knowledge, this work is a first contribution in this respect.

Another remarkable feature of the characterization via the scalar C_0 is that, taking into account the properties of the Jordan canonical form (see, for instance, [2, pp. 126–127]), the generic behavior will take place if and only if several equations involving the ranks of different powers of $A + B - \lambda_0 I$ and $A - \lambda_0 I$ are fulfilled. Surprisingly, in the case of low rank perturbations, this set of equations is equivalent to the single condition $C_0 \neq 0$, where C_0 does not involve explicitly any power, either of $A - \lambda_0 I$ or of $A + B - \lambda_0 I$.

Finally, although in this paper we only pay attention to which Jordan blocks are destroyed under a low rank perturbation, and which ones are preserved for each eigenvalue of A, another question which naturally arises is, What happens with the eigenvalues of the destroyed blocks? As stated before, classical first-order eigenvalue perturbation results answer the question for small perturbations: for each destroyed Jordan block of dimension k, a ring of k different simple eigenvalues of $A+B$ appears, and there are explicit formulas for the first-order corrections [5, 7]. For perturbations of arbitrary size, however, the information available is much more limited, and reduces to fairly general (and usually pessimistic) bounds on the variation of the eigenvalues [11].

The paper is organized as follows. In the second section, after setting the appropriate notation, we study in Theorem 2.1 the algebraic multiplicity, as an eigenvalue of $A+B$, of each eigenvalue λ_0 of A for which condition (1.1) holds. This multiplicity turns out to depend crucially on C_0, and $C_0 \neq 0$ is the necessary and sufficient condition for the algebraic multiplicity of λ_0 to be compatible with the predicted generic behavior, i.e., the Jordan blocks of $A+B$ with eigenvalue λ_0 are just the $g - \operatorname{rank}(B)$ smallest Jordan blocks of A with eigenvalue λ_0, where g is the number of λ_0-Jordan blocks of A. However, the algebraic and geometric multiplicity of an eigenvalue do not determine by themselves the corresponding part of the Jordan structure. In the third section, we prove in Theorem 3.1 that $C_0 \neq 0$ ensures the generic behavior by explicitly constructing the corresponding Jordan chains of $A + B$ starting from those

of A. This will show that $C_0 \neq 0$ is a necessary and sufficient condition for the generic behavior, a fact we summarize in a final, concluding theorem.

2. Counting algebraic multiplicities. Throughout this section we follow the notation in [7]: let A be an arbitrary $n \times n$ complex matrix and

$$\left[\begin{array}{c|c} J & \\ \hline & \widehat{J} \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widehat{Q} \end{array}\right] A \left[\begin{array}{c|c} P & \widehat{P} \end{array}\right] \tag{2.1}$$

be a Jordan decomposition of A, so

$$\left[\begin{array}{c} Q \\ \hline \widehat{Q} \end{array}\right] \left[\begin{array}{c|c} P & \widehat{P} \end{array}\right] = I. \tag{2.2}$$

The matrix J contains all Jordan blocks associated with the eigenvalue of interest λ_0, while $\widehat{J}$ is the part of the Jordan form containing the other eigenvalues. Let

$$J = \Gamma_1^1 \oplus \cdots \oplus \Gamma_1^{r_1} \oplus \cdots \oplus \Gamma_q^1 \oplus \cdots \oplus \Gamma_q^{r_q}, \tag{2.3}$$

where, for $j = 1, \ldots, q$,

$$\Gamma_j^1 = \cdots = \Gamma_j^{r_j} = \begin{bmatrix} \lambda_0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0 \end{bmatrix}$$

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_0. The eigenvalue λ_0 is semisimple (nondefective) if $q = n_1 = 1$ and nonderogatory if $q = r_1 = 1$. Set

$$a = \sum_{j=1}^{q} r_j n_j \qquad \text{and} \qquad g = \sum_{j=1}^{q} r_j, \tag{2.4}$$

i.e., we denote by a the *algebraic* multiplicity of λ_0 as an eigenvalue of A, and by g its geometric multiplicity.

We further partition

$$P = \left[\begin{array}{c|c|c|c|c|c|c} P_1^1 & \ldots & P_1^{r_1} & \ldots & P_q^1 & \ldots & P_q^{r_q} \end{array}\right] \tag{2.5}$$

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_0. The lth column of P_j^k is a right Jordan vector of order l. In particular, 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_0. Analogously, we split

$$Q = \begin{bmatrix} \frac{Q_1^1}{\vdots} \\ \hline \frac{Q_1^{r_1}}{\vdots} \\ \hline \frac{Q_q^1}{\vdots} \\ \hline Q_q^{r_q} \end{bmatrix},$$

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_0. The lth row, counting from below, of Q_j^k is a left Jordan vector of order l. 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_0. With these eigenvectors we build up matrices

$$L_j = \begin{bmatrix} y_j^1 \\ \vdots \\ y_j^{r_j} \end{bmatrix}, \qquad R_j = [x_j^1, \ldots, x_j^{r_j}]$$

for $j = 1, \ldots, q$,

$$W_i = \begin{bmatrix} L_1 \\ \vdots \\ L_i \end{bmatrix}, \qquad Z_i = [R_1, \ldots, R_i]$$

for $i = 1, \ldots, q$, and we define square matrices Φ_i of dimension

$$f_i = \sum_{j=1}^{i} r_j$$

by

$$\Phi_i = W_i B Z_i, \qquad i = 1, \ldots, q. \tag{2.6}$$

Note that, due to the cumulative definitions of W_i and Z_i, every $\Phi_{i-1},\ i = 2, \ldots, q$, is the upper left block of Φ_i.

Take, for instance, the unperturbed matrix

$$A = \left[\begin{array}{ccc|cc|cc|c} 0 & 1 & 0 & & & & & \\ 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & & & & & \\ \hline & & & 0 & 1 & & & \\ & & & 0 & 0 & & & \\ \hline & & & & & 0 & 1 & \\ & & & & & 0 & 0 & \\ \hline & & & & & & & 2 \end{array}\right], \tag{2.7}$$

and set $\lambda_0 = 0$, i.e., $g = 3$, $a = 7$, $n_1 = 3$, $n_2 = 2$, $r_1 = 1$, $r_2 = 2$. Then, since the right Jordan vectors of A are columns of the identity matrix, any given perturbation matrix

$$B = \left[\begin{array}{ccc|cc|cc|c} * & * & * & * & * & * & * & * \\ * & * & * & * & * & * & * & * \\ \blacksquare & * & * & \clubsuit & * & \spadesuit & * & * \\ \hline * & * & * & * & * & * & * & * \\ \clubsuit & * & * & \clubsuit & * & \heartsuit & * & * \\ \hline * & * & * & * & * & * & * & * \\ \spadesuit & * & * & \heartsuit & * & \spadesuit & * & * \\ \hline * & * & * & * & * & * & * & * \end{array}\right] \tag{2.8}$$

gives rise to the two matrices

$$\Phi_1 = \left[\begin{array}{c} \blacksquare \\ \end{array}\right], \qquad \Phi_2 = \left[\begin{array}{c|cc} \blacksquare & \clubsuit & \spadesuit \\ \hline \clubsuit & \clubsuit & \heartsuit \\ \spadesuit & \heartsuit & \spadesuit \end{array}\right]$$

with dimensions $f_1 = 1$ and $f_2 = 3$.

As announced in the introduction, we want to determine the most likely Jordan structure for the eigenvalue λ_0 of a low rank perturbation $A + B$ of A, where by low we mean that B and λ_0 satisfy (1.1). Let n_s be the smallest one among the sizes of the rank(B) largest Jordan blocks of A associated with λ_0, i.e., $s \in \{1, \ldots, q\}$ is the index such that

$$\operatorname{rank}(B) \equiv \rho = f_{s-1} + \beta, \quad 0 < \beta \leq r_s, \tag{2.9}$$

where we have set $f_0 = 0$ for convenience. In the 8×8 example above, if we consider perturbations with $\operatorname{rank}(B) = \rho = 2$, then $\rho = f_1 + \beta$ with $\beta = 1 < r_2 = 2$, i.e., $s = 2$ since the two largest Jordan blocks of A are the single 3×3 block, together with either one of the two 2×2 blocks.

We have already seen in formula (1.2) that the geometric multiplicity of λ_0 can decrease at most by ρ under the perturbation B. The following result shows how much the algebraic multiplicity usually decreases. If only the ρ largest Jordan blocks of A with eigenvalue λ_0 disappear, then the algebraic multiplicity of λ_0 in $A + B$ is

$$\widetilde{a} = (r_s - \beta)n_s + r_{s+1}n_{s+1} + \cdots + r_q n_q. \tag{2.10}$$

It is shown in Theorem 2.1 that the algebraic multiplicity of λ_0 in $A + B$ is always larger than or equal to $\widetilde{a}$, and the necessary and sufficient condition for equality is $C_0 \neq 0$.

THEOREM 2.1. *Let A be an $n \times n$ matrix with Jordan form* (2.1)*, i.e., having an eigenvalue λ_0 with Jordan blocks of dimensions $n_1 > n_2 > \cdots > n_q$ repeated $r_1, r_2, \ldots, r_q$ times and algebraic and geometric multiplicities a and g given by* (2.4)*. Let B be an $n \times n$ matrix with rank given by* (2.9)*, and let the matrices Φ_i, $i = 1, \ldots, q$, be given by* (2.6)*. Then the characteristic polynomial of $A + B$ is of the form*

$$p(\lambda) = (\lambda - \lambda_0)^{\widetilde{a}}\, t(\lambda - \lambda_0),$$

where $\widetilde{a}$ is given by (2.10) *and $t(\lambda - \lambda_0)$ is a monic polynomial of degree $n - \widetilde{a}$. Moreover, the constant coefficient of $t(\cdot)$ is*

$$t(0) = (-1)^{\rho+n-a}\, C_0\, \det(\widehat{Q}\, A\, \widehat{P} - \lambda_0 I), \tag{2.11}$$

where $\widehat{Q}$, $\widehat{P}$ *are as in* (2.1) *and* C_0 *is the sum of all principal minors of* Φ_s *corresponding to submatrices of dimension* ρ *containing the upper left block* Φ_{s-1} *of* Φ_s. *(If* $s = 1$, all *principal minors of dimension* ρ *are to be considered.) If, in particular,* $\rho = f_s$ *(i.e., if* $\beta = r_s$*) for some* $s \in \{1, \ldots, q\}$, *then* C_0 *is simply* $\det \Phi_s$.

Proof. We begin by writing the characteristic polynomial of $A + B$ as

$$p(\lambda) = \det((\lambda - \lambda_0)I - \operatorname{diag}(J - \lambda_0 I, \widehat{J} - \lambda_0 I) - \widetilde{B}),$$

where

$$\widetilde{B} = \left[\begin{array}{c} Q \\ \hline \widehat{Q} \end{array}\right] B \left[\begin{array}{c|c} P & \widehat{P} \end{array}\right].$$

For the sake of simplicity we define $\tilde{\lambda} \equiv \lambda - \lambda_0$ and $p_0(\tilde{\lambda}) \equiv p(\lambda)$, so the coefficient of $\tilde{\lambda}^{n-k}$ in $p_0(\tilde{\lambda})$ is $(-1)^k$ times the sum of all k-dimensional principal minors of $\operatorname{diag}(J - \lambda_0 I, \widehat{J} - \lambda_0 I) + \widetilde{B}$ [2, p. 42]. Notice that all principal minors whose corresponding submatrices have more than $\rho = \operatorname{rank}(B)$ rows (equivalently, columns) containing *only* elements of $\widetilde{B}$ are zero, since $\operatorname{rank}(\widetilde{B}) = \operatorname{rank}(B)$. This simple observation is the key to proving the theorem.

To find the lowest power of $\tilde{\lambda}$ in $p_0(\tilde{\lambda})$ we can just look for the largest possible dimension of a principal submatrix of $\operatorname{diag}(J - \lambda_0 I, \widehat{J} - \lambda_0 I) + \widetilde{B}$ containing at most ρ rows with only elements of $\widetilde{B}$. If we denote by $k_{\max}$ the maximal dimension we are looking for, then

$$p_0(\tilde{\lambda}) = \tilde{\lambda}^{n-k_{\max}}\, t(\tilde{\lambda}),$$

with t a monic polynomial of degree $k_{\max}$. Notice first that, since we are looking for the *largest* dimension, we can restrict ourselves to principal submatrices containing *exactly* ρ rows with only elements of $\widetilde{B}$: if the principal submatrix contains less than ρ rows with only elements of $\widetilde{B}$, then one may always construct a new principal submatrix of larger dimension by including a new row with only elements of $\widetilde{B}$ (and the corresponding column). For instance, any row in the position of a bottom row of a Jordan block of $J - \lambda_0 I$ contains only elements of $\widetilde{B}$, and since there are g of them, with $\rho \le g$, at least one of these bottom rows can be used to increase the dimension.

To determine $k_{\max}$, let $\alpha \subset \{1, 2, \ldots, n\}$ be any index set and denote by $(\operatorname{diag}(J - \lambda_0 I, \widehat{J} - \lambda_0 I) + \widetilde{B})(\alpha, \alpha)$ the principal submatrix of $\operatorname{diag}(J - \lambda_0 I, \widehat{J} - \lambda_0 I) + \widetilde{B}$ that lies in the rows and columns indexed by α. By definition, this principal submatrix contains all the diagonal elements in the positions indexed by α. Since the eigenvalues of $\widehat{J}$ are all different from λ_0, the diagonal elements in the positions $a+1, a+2, \ldots, n$ are *not* elements of $\widetilde{B}$, so the corresponding indices can be always included in α without increasing the number of rows with only elements of $\widetilde{B}$. Hence, any set α of the maximal size $k_{\max}$ containing exactly ρ rows with only elements of $\widetilde{B}$ must be of the form

$$\alpha = \{i_1, \ldots, i_j, a+1, a+2, \ldots, n\} \quad \text{with} \quad 1 \le i_1 < i_2 < \cdots < i_j \le a.$$

Furthermore, the rows $i_1, \ldots, i_j$ intersect with a certain number, say l, of the g Jordan blocks in $J - \lambda_0 I$. Take any of these Jordan blocks and denote by i_b the largest index corresponding to a row in α intersecting with that particular Jordan block. Then, the i_bth row contributes to the principal submatrix only with elements of $\widetilde{B}$, either

because it is the bottom row of the Jordan block or because i_b+1 does not belong to α, and thus the element in the position (i_b, i_b+1), where $J-\lambda_0 I$ has a superdiagonal 1, is not in the submatrix. This imposes the restriction $l \leq \rho$ on l. Hence, no choice for α can give rise to a larger dimension than taking $l=\rho$ and choosing the indices $i_1 < \cdots < i_j$ to cover *all* rows of a set of ρ complete *largest* Jordan blocks of $J-\lambda_0 I$. Actually, any of these choices is admissible, since each contains exactly ρ rows with only elements of $\widetilde{B}$, namely one bottom row for each of the ρ Jordan blocks chosen from $J-\lambda_0 I$. The number of possible choices is $r_s!/(\beta!(r_s-\beta)!)$, which is simply one when $\rho = f_s$. Hence, we have shown that

$$k_{\max} = r_1 n_1 + \cdots + r_{s-1} n_{s-1} + \beta n_s + n - a,$$

and consequently $\widetilde{a} = n - k_{\max}$ with $\widetilde{a}$ given by (2.10).

Now we prove (2.11). Recall that $t(0)$ is $(-1)^{k_{\max}}$ times the sum of all $k_{\max}$-dimensional principal minors of $\operatorname{diag}(J-\lambda_0 I, \widehat{J}-\lambda_0 I) + \widetilde{B}$. Moreover, the only nonzero $k_{\max}$-dimensional principal minors correspond to the submatrices described in the previous paragraph. Consider one of these minors and call it M. Set $h = k_{\max} - (n-a) - \rho$ and denote by $1 = j_1 < j_2 < \cdots < j_h$ the indices of rows of the principal submatrix corresponding to M, where $J-\lambda_0 I$ has superdiagonal 1's. The j_kth row of this submatrix is the sum of two rows: one is the (j_k+1)st row e_{j_k+1} of the identity matrix, the other is a piece of a row of $\widetilde{B}$. Using this fact, we can expand M as a sum of 2^h determinants whose j_kth row, with $1 \leq k \leq h$, is either e_{j_k+1} or a row with only elements of $\widetilde{B}$. With the exception of the determinant with all the vectors $e_{j_1+1}, e_{j_2+1}, \ldots, e_{j_h+1}$, the rest of these determinants are zero because each contains more than ρ rows with elements of $\widetilde{B}$. A similar argument on the last $n-a$ rows of the submatrix corresponding to M allows us to replace every element of $\widetilde{B}$ in these rows by zero without changing the value of M. The cofactor expansion of the remaining determinant along the rows $1 = j_1 < j_2 < \cdots < j_h$ leads to a value for M equal to $(-1)^h \det(\widehat{J}-\lambda_0 I)$ times a minor of Φ_s corresponding to a principal submatrix of dimension ρ containing the upper left block Φ_{s-1}. Extending this argument to all nonzero $k_{\max}$-dimensional principal minors of $\operatorname{diag}(J-\lambda_0 I, \widehat{J}-\lambda_0 I) + \widetilde{B}$ leads to (2.11). □

In example (2.7)–(2.8) above, with a perturbation B with

$$\operatorname{rank}(B) = \rho = 2,$$

the quantity C_0 is given by

$$C_0 = \det \left[\begin{array}{c|c} \blacksquare & \clubsuit \\ \hline \clubsuit & \clubsuit \end{array} \right] + \det \left[\begin{array}{c|c} \blacksquare & \spadesuit \\ \hline \spadesuit & \spadesuit \end{array} \right].$$

According to Theorem 2.1, any perturbation with $C_0 \neq 0$ is such that $\lambda_0 = 0$ is an eigenvalue of $A+B$ with algebraic multiplicity two and, according to (1.2), geometric multiplicity at least one. Hence, the Jordan form of $A+B$ can either have just one 2×2, or have two 1×1 Jordan blocks corresponding to λ_0. We shall prove in the next section that $C_0 \neq 0$ actually implies the first possibility.

3. Building Jordan chains. In this section we prove that the genericity condition $C_0 \neq 0$ actually implies that the $\operatorname{rank}(B)$ largest Jordan blocks of A disappear for each eigenvalue, and the rest of the Jordan blocks of A remain as Jordan blocks of $A+B$. If $\operatorname{rank}(B)$ is given by (2.9), we will construct, for the eigenvalue λ_0 of $A+B$,

$r_s - \beta$ Jordan chains of length n_s and r_k chains of length n_k for $k = s+1, \ldots, q$. Due to Theorem 2.1, these are the only Jordan chains of $A+B$ for λ_0, since $C_0 \neq 0$ implies that the algebraic multiplicity of λ_0 is given by (2.10). Although the construction is more involved for perturbations of arbitrary rank, the crucial step in the proof is the recursive formula (3.5), a multidimensional analogue of the one employed by Savchenko [9] for the case of rank one perturbations.

In order to give a concise proof of the results in this section we need to introduce some further notation. Recall that each column of the matrix P in decomposition (2.1) is a Jordan vector of A associated with λ_0. Furthermore, the set of columns of each $P_j^k,\ j = 1, \ldots, q,\ k = 1, \ldots, r_j$, in (2.5) forms a right Jordan chain with length n_j of A associated with λ_0, and the lth column of P_j^k is a right Jordan vector of order l.

For each $l \in \{1, \ldots, n_s\}$ we consider all right Jordan vectors of order l of A associated with λ_0 and denote by X_l (resp., Y_l) the submatrix of P containing all right Jordan vectors of order l corresponding to the ρ largest (resp., the $g-\rho$ smallest) Jordan blocks in J. Both the columns of X_l and of Y_l are assumed to appear in the same relative order as in P. Notice that whenever $\beta < r_s$ in (2.9), the ρ largest Jordan blocks in J are not uniquely determined: we need to further specify which β of the r_s Jordan blocks of size n_s contribute to the X_l, and this fixes which blocks contribute to the Y_l. We do this with the aid of the genericity condition $C_0 \neq 0$: recall that C_0 is the sum of all ρ-dimensional principal minors of Φ_s containing Φ_{s-1}, where $\Phi_s = W_s B Z_s$ and the columns of Z_s are right eigenvectors, i.e., right Jordan vectors of order 1. If $C_0 \neq 0$, then one or more of these principal minors of Φ_s must be different from zero. Let γ be the set of indices corresponding to the ρ rows and columns of Φ_s in any of the nonzero principal minors, and denote, as before, by $\Phi_s(\gamma, \gamma)$ the corresponding principal submatrix of Φ_s. Then γ must be of the form

$$\gamma = \{1, \ldots, f_{s-1}, i_1, i_2, \ldots, i_\beta\}, \qquad f_{s-1} < i_1 < \cdots < i_\beta \leq f_s, \tag{3.1}$$

and we define X_1 as the $n \times \rho$ submatrix of Z_s containing the columns indexed by γ. The $r_s - \beta$ remaining columns of Z_s are assigned to Y_1. Once X_1 (and therefore Y_1) is fixed, the columns of the remaining X_l (resp., Y_l) are chosen from the same set of Jordan blocks as the eigenvectors in X_1 (resp., Y_1). This implies that equations (3.3) below are satisfied.

In the example (2.7)–(2.8), with $\operatorname{rank}(B) = 2$, there are only two principal minors of Φ_2 containing Φ_1, namely

$$\begin{aligned} \Phi_2(\{1,2\},\{1,2\}) &= \det \left[\begin{array}{c|c} \blacksquare & \clubsuit \\ \hline \clubsuit & \clubsuit \end{array}\right], \\ \Phi_2(\{1,3\},\{1,3\}) &= \det \left[\begin{array}{c|c} \blacksquare & \spadesuit \\ \hline \spadesuit & \spadesuit \end{array}\right]. \end{aligned} \tag{3.2}$$

If the first (resp., the second) minor is different from zero, then the two columns of $X_1 \in \mathbb{C}^{8\times 2}$ are the first and second (resp., first and third) columns of $Z_2 \in \mathbb{C}^{8\times 3}$, which are the first and fourth (resp., the first and sixth) columns of $P \in \mathbb{C}^{8\times 7}$. In that case, Y_1 reduces to the third (resp., second) column of Z_2.

Note that all matrices $X_l \in \mathbb{C}^{n\times\rho},\ l = 1, \ldots, n_s$, have the same dimensions, while $Y_l \in \mathbb{C}^{n\times d_l}$, with d_l the number of Jordan blocks of dimension larger than or equal to l among the $g - \rho$ smallest Jordan blocks contributing to Y_1. Hence,

$d_1 = g - \rho \geq d_2 \geq \cdots \geq d_{n_s}$. The fact that both the X_l and the Y_l are constituted by consecutive pieces of Jordan chains is reflected by the conditions

$$(A - \lambda_0 I)X_l = X_{l-1}, \qquad (A - \lambda_0 I)Y_l = Y_{l-1}^{(l)}, \quad l = 1, \ldots, n_s, \tag{3.3}$$

where $Y_{l-1}^{(l)}$ is the leftmost $n \times d_l$ submatrix of Y_{l-1}, and both X_0 and Y_0 are defined to be zero. Notice that if $\beta = r_s$, then $d_{n_s} = 0$ and Y_{n_s} is an empty matrix, so the second equation in (3.3) makes sense only for $l = 1, \ldots, n_{s+1}$.

After all these conventions we are in the position to obtain the main result of this section.

THEOREM 3.1. *Let A, B, λ_0, and C_0 be as in the statement of Theorem* 2.1. *If $C_0 \neq 0$, then the Jordan blocks of $A + B$ with eigenvalue λ_0 are just the $g - \operatorname{rank}(B)$ smallest Jordan blocks of A with eigenvalue λ_0. More precisely, if the rank of B is given by* (2.9)*, then the Jordan structure of $A + B$ with eigenvalue λ_0 consists of $r_s - \beta$ Jordan blocks of dimension n_s and r_k Jordan blocks of dimension n_k for $k = s + 1, \ldots, q$.*

Proof. As commented in the beginning of this section, it suffices to explicitly construct Jordan chains of the appropriate length for $A+B$. This amounts to constructing matrices $\widetilde{Y}_l \in \mathbb{C}^{n \times d_l}$ for $l = 1, \ldots, n_s$ such that

$$(A + B - \lambda_0 I)\widetilde{Y}_l = \widetilde{Y}_{l-1}^{(l)}, \qquad l = 1, \ldots, n_s, \tag{3.4}$$

where $\widetilde{Y}_{l-1}^{(l)}$ is the leftmost $n \times d_l$ submatrix of $\widetilde{Y}_{l-1}$ and $\widetilde{Y}_0 = 0$. We must also prove that the columns of $[\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_{n_s}]$ are linearly independent.

We will construct these matrices recursively through the formula

$$\widetilde{Y}_l = Y_l - \sum_{i=1}^{l} X_i C_{l-i+1}^{(l)}, \qquad l = 1, \ldots, n_s, \tag{3.5}$$

where, at the lth step, the $\rho \times d_l$ matrix $C_l^{(l)}$ is chosen in such a way that

$$B\widetilde{Y}_l = 0, \tag{3.6}$$

and we denote by $C_j^{(l)}$ for $j < l$, the leftmost $\rho \times d_l$ submatrix of the $\rho \times d_j$ matrix $C_j^{(j)}$ already chosen at the jth step. The fact that condition (3.6) uniquely determines the matrix $C_l^{(l)}$ at each step is a consequence of our previous choice of the last β columns of the matrix X_1: since B has rank ρ, one can write $B = \mathcal{U}\mathcal{V}^*$ with $\mathcal{U}$, $\mathcal{V} \in \mathbb{C}^{n \times \rho}$ of full rank and, accordingly, rewrite (3.6) as

$$\mathcal{V}^* X_1 C_l^{(l)} = \mathcal{V}^* \left(Y_l - \sum_{i=2}^{l} X_i C_{l-i+1}^{(l)} \right),$$

where the right-hand side is already known. Hence, the solution $C_l^{(l)}$ is unique provided the square matrix $\mathcal{V}^* X_1$ is nonsingular. Now, recall that the ρ-dimensional principal submatrix $\Phi_s(\gamma, \gamma)$ of Φ_s indexed by the set γ in (3.1) is nonsingular, and $\Phi_s(\gamma, \gamma) = W_s(\gamma) B X_1$, where $W_s(\gamma)$ is the $\rho \times n$ submatrix of W_s containing the rows indexed by γ. Hence, $\Phi_s(\gamma, \gamma)$ is the product of two $\rho \times \rho$ matrices, $W_s(\gamma)\mathcal{U}$ and $\mathcal{V}^* X_1$, each of them nonsingular as well.

We now check (3.4): the definition (3.5) of $\widetilde{Y}_l$, together with (3.6) and (3.3), implies that

$$(A+B-\lambda_0 I)\widetilde{Y}_l = (A-\lambda_0 I)\left(Y_l - \sum_{i=1}^{l} X_i C_{l-i+1}^{(l)}\right) = Y_{l-1}^{(l)} - \sum_{i=2}^{l} X_{i-1} C_{l-i+1}^{(l)}.$$

Shifting the dummy index to $j = i-1$, the previous expression can be rewritten as

$$(A+B-\lambda_0 I)\widetilde{Y}_l = Y_{l-1}^{(l)} - \sum_{j=1}^{l-1} X_j C_{l-j}^{(l)},$$

and, since $C_{l-j}^{(l)}$ is the $\rho \times d_l$ leftmost submatrix of $C_{l-j}^{(l-1)}$, the matrix above is just $\widetilde{Y}_{l-1}^{(l)}$, the leftmost $n \times d_l$ submatrix of $\widetilde{Y}_{l-1}$. This proves that the matrices defined by (3.5) satisfy (3.4). Finally, each $\widetilde{Y}_l$ is just the corresponding Y_l plus some linear combinations of the columns of the matrices $X_1, \ldots, X_l$. Since the columns of all X_l and Y_l are linearly independent (the columns of P are linearly independent), the columns of $\widetilde{Y}_l$ are also linearly independent. ∎

In the example (2.7)–(2.8) with $\operatorname{rank}(B) = 2$, we would need to construct a Jordan chain of length two. If we assume that $C_0 \neq 0$, then one of the two minors in (3.2) is nonzero. Once X_1 is chosen accordingly, the construction of a new Jordan chain of length two for $A+B$ goes as follows: if we write X_i, Y_i, $i = 1, 2$, columnwise as

$$X_i = \left[\, \xi_i^{(1)} \;\; \xi_i^{(2)} \,\right], \quad Y_i = [\, \eta_i \,], \quad i = 1, 2,$$

and denote $C_1^{(1)} = [\, c_{11} \;\; c_{12} \,]^T$, then the first matrix equation $\widetilde{Y}_1 = Y_1 - X_1 C_1^{(1)}$ in (3.5) leads to the eigenvector

$$\widetilde{\eta}_1 = \eta_1 - c_{11}\, \xi_1^{(1)} - c_{12}\, \xi_1^{(2)}$$

of $A+B$, where c_{11} and c_{12} are chosen to ensure that $B\widetilde{\eta}_1 = 0$. The second vector $\widetilde{\eta}_2$ in the new Jordan chain is found through the equation $\widetilde{Y}_2 = Y_2 - X_2 C_1^{(2)} - X_1 C_2^{(2)}$, which, if $C_2^{(2)} = [\, c_{21} \;\; c_{22} \,]^T$, translates into

$$\widetilde{\eta}_2 = \eta_2 - c_{11}\, \xi_2^{(1)} - c_{12}\, \xi_2^{(2)} - c_{21}\, \xi_1^{(1)} - c_{22}\, \xi_1^{(2)}$$

in vector terms. Notice that in this case $C_1^{(2)} = C_1^{(1)}$. Again, the scalars c_{21} and c_{22} are chosen in such a way that $B\widetilde{\eta}_2 = 0$.

We may summarize the discussion throughout the paper by writing the conclusion of both Theorems 2.1 and 3.1 as a final, summarizing theorem.

Concluding Theorem. *Let A be a complex $n \times n$ matrix and λ_0 an eigenvalue of A with geometric multiplicity g. Let B be a complex $n \times n$ matrix with $\operatorname{rank}(B) \leq g$ and C_0 be as in the statement of Theorem* 2.1. *Then the Jordan blocks of $A+B$ with eigenvalue λ_0 are just the $g - \operatorname{rank}(B)$ smallest Jordan blocks of A with eigenvalue λ_0 if and only if $C_0 \neq 0$.*

Acknowledgments. The authors wish to thank Prof. Sergey Savchenko for bringing this interesting problem to their attention and for a fruitful interchange of ideas over e-mail. The authors also thank an anonymous referee whose many detailed suggestions helped to improve the overall presentation of the results.

REFERENCES

[1] H. Baumgärtel, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser, Basel, Boston, Stuttgart, 1985.

[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[3] C.-P. Jeannerod, *On some nongeneric perturbations of an arbitrary Jordan structure*, Linear Algebra Appl., to appear.

[4] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, Heidelberg, New York, 1980.

[5] V. B. Lidskii, *Perturbation theory of non-conjugate operators*, U.S.S.R. Comput. Math. and Math. Phys., 1 (1965), pp. 73–85.

[6] Y. Ma and A. Edelman, *Nongeneric eigenvalue perturbations of Jordan blocks*, Linear Algebra Appl., 273 (1998), pp. 45–63.

[7] J. Moro, J. V. Burke, and M. L. Overton, *On the Lidskii–Vishik–Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 793–817.

[8] J. Moro and F. M. Dopico, *First order eigenvalue perturbation theory and the Newton diagram*, in Applied Mathematics and Scientific Computing, Z. Drmač et al., eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 143–175.

[9] S. Savchenko, *The typical change of the spectral properties under a rank one perturbation*, Mat. Zametki, submitted (in Russian).

[10] S. Savchenko, *On the Change of the Spectral Properties under a Generic Perturbation of Rank r*, preprint (in Russian).

[11] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.

[12] 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 (in English); Uspekhi Mat. Nauk, 15 (1960), pp. 3–80 (in Russian).

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 507–516

THE LOCAL COEFFICIENT OF ERGODICITY OF A NONNEGATIVE MATRIX*

MARC ARTZROUNI†

Abstract. The local coefficient of ergodicity $\tau(T, Y', w)$ of a nonnegative column-allowable matrix T at a fixed positive vector Y is defined as the supremum of $d(X'T, Y'T)/d(X', Y')$ for X not colinear to Y and $d(X', Y') \leq w$ (d is the projective distance in the positive quadrant). A near-closed-form expression is given for $\tau(T, Y', w)$. If T' is scrambling (i.e., no two rows of T' are orthogonal), then for any $Y > 0$, $w < \infty$ we have $\tau(T, Y', w) < 1$. When Y is a positive left eigenvector of T and $X_o > 0$, these results can be used to prove the convergence in direction of $X_o' T^p$ to Y'. Results are illustrated with a numerical example.

Key words. nonnegative matrix, coefficient of ergodicity, eigenvector, dynamical system

AMS subject classifications. 15A18, 15A48, 39A11

DOI. 10.1137/S0895479801391801

1. Definitions and first properties. In the theory of nonnegative matrices the coefficient of ergodicity $\tau(T)$ of a column-allowable $n \times n$ matrix $T = (t_{ij})$ (i.e., a matrix having no zero column) is defined as

$$\tau(T) = \sup_{X,Y>0; X \neq \lambda Y} \frac{d(X'T, Y'T)}{d(X', Y')}, \tag{1.1}$$

where $d(X', Y') = \max_{i,j} \ln(x_i y_j / x_j y_i)$ is the projective distance between the positive vectors $X = (x_i)$ and $Y = (y_i)$ [4, p. 83].

The quantity $\tau(T)$ (which is between 0 and 1) is a contraction coefficient for the linear operator T since $d(X'T, Y'T) \leqslant \tau(T) d(X', Y')$; $\tau(T)$ takes its full usefulness when it is < 1 since T is then a contracting operator.

For an initial vector $X_o > 0$ we may be interested in the dynamical system $X_p' = X_o' T^p$, $p = 0, 1, \ldots$. Suppose Y is a left Perron vector of T. The corresponding eigenvalue is the spectral radius $\rho(T)$ of T [3, p. 493] and $Y'T^p = \rho(T)^p Y'$ for any integer $p \geq 1$. The projective distance between Y (the fixed point of T) and X_p then satisfies

$$d(X_p', Y') = d(X_o' T^p, Y') = d(X_o' T^p, Y'T^p) \leqslant \tau(T)^p d(X_o', Y'), \qquad p = 0, 1, \ldots. \tag{1.2}$$

If $\tau(T) < 1$, then (1.2) shows that the vectors $X_o' T^p$ approach Y in direction when $p \to \infty$.

The problem is that $\tau(T) < 1$ only for T positive, which is a rather strong condition. With a single zero element in the matrix, the coefficient $\tau(T)$ is 1; T is then not a contraction and (1.2) can no longer be used to easily conclude that $d(X_o' T^p, Y') \to 0$.

If T is primitive, then a power of T is positive and the same result holds. However, there are cases where T is imprimitive or even reducible and one would still like to use a simple contraction-type argument to see whether $d(X_o' T^p, Y') \to 0$ for $p \to \infty$.

*Received by the editors July 4, 2001; accepted for publication (in revised form) by H. J. Werner March 4, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/simax/25-2/39180.html
†Department of Applied Mathematics, University of Pau, 64000 Pau, France (marc.artzrouni@univ-pau.fr).

The approach used here will hinge upon the fact that one of the two vectors appearing in the projective distances of (1.2) (i.e., Y' or $Y'T^p$) is always a scalar multiple of Y. Therefore in the definition (1.1) of $\tau(T)$ the vector Y could be fixed. Furthermore, it may not be necessary to find a supremum for arbitrarily large values of $d(X', Y')$ in the denominator. Indeed, in the example of (1.2) it would suffice for $\tau(T)$ to be a supremum over all $d(X', Y')$ bounded by some positive $w > 0$ since the projective distances between the iterates X'_oT^p and Y' are nonincreasing.

In order to address these issues we define $B(Y, w)$ as the ball of center Y and radius $w > 0$ for the projective distance, i.e., $X \in B(Y, w) \iff d(X', Y') \leq w$. We now consider the following definition (in which the vector Y is an arbitrary positive vector; Y is not assumed to be a Perron vector of T).

DEFINITION 1.1. *The local coefficient of ergodicity (LCE) of a nonnegative matrix T in a neighborhood $B(Y, w)$ of a vector $Y > 0$ is defined as*

$$\tau(T, Y', w) \stackrel{\text{def}}{=} \sup_{\substack{X>0; X\neq\lambda Y \\ X\in B(Y,w)}} \frac{d(X'T, Y'T)}{d(X', Y')}. \tag{1.3}$$

If $X \in B(Y, w)$, then

$$d(X'T, Y'T) \leq \tau(T, Y', w)d(X', Y'), \tag{1.4}$$

and if $\tau(T, Y', w) < 1$, we say that T is a local contraction with respect to Y.

The definition and properties of $\tau(T)$ insure that $\tau(T, Y', w)$ is defined and is ≤ 1. The LCE has a submultiplicative property similar to the one that holds for τ (i.e., $\tau(T_1T_2) \leq \tau(T_1)\tau(T_2)$). Indeed, let $\{T_i\}_{i=1,2,\ldots}$ be a sequence of column-allowable matrices and define the forward product $U_p = T_1T_2 \ldots T_p$. The submultiplicative property is then given in the following proposition.

PROPOSITION 1.2. *For positive vectors X, Y we define $w_o \stackrel{\text{def}}{=} d(X', Y')$. With the notation given above, we then have*

$$\tau(T_1T_2, Y', w_o) \leq \tau(T_2, Y'T_1, w_o)\tau(T_1, Y', w_o) \tag{1.5}$$

and more generally

$$\tau(U_k, Y', w_o) \leq \tau(T_k, Y'U_{k-1}, w_o)\tau(T_{k-1}, Y'U_{k-2}, w_o) \ldots \tau(T_1, Y', w_o). \tag{1.6}$$

Proof. Because $d(X'T_1, Y'T_1) \leq d(X', Y')$ we have

$$\tau(T_1T_2, Y', w_o) = \sup_{\substack{X>0; X\neq\lambda Y \\ X\in B(Y,w)}} \frac{d(X'T_1T_2, Y'T_1T_2)}{d(X', Y')} \tag{1.7}$$

$$\leq \sup_{\substack{X'T_1\neq\lambda Y'T_1 \\ X\in B(Y,w)}} \frac{d((X'T_1)T_2, (Y'T_1)T_2)}{d(X'T_1, Y'T_1)} \sup_{\substack{X>0; X\neq\lambda Y \\ X\in B(Y,w)}} \frac{d(X'T_1, Y'T_1)}{d(X', Y')}$$

$$\leq \tau(T_2, Y'T_1, w_o)\tau(T_1, Y', w_o),$$

which is the desired result of (1.5), from which (1.6) follows by induction.

When all the matrices T_i are equal to some T, and $Y > 0$ is a left positive eigenvector of T, then $Y'U_p = Y'T^p$; $Y'T^p$ and Y' are colinear for all p and (1.6) yields

$$d(X_o'T^p, Y') = d(X_o'T^p, Y'T^p) \le \tau(T^p, Y', w_o)w_o \le \tau(T, Y', w_o)^p w_o, \quad p = 0, 1, \ldots, \tag{1.8}$$

which shows that $d(X_o'T^p, Y')$ approaches 0 exponentially fast if $\tau(T, Y', w_o) < 1$.

We will see that $\tau(T, Y', w_o)$ is < 1 under conditions that are much weaker than the positivity assumption needed for $\tau(T) < 1$. In fact we will show that for any $Y > 0$ (not necessarily a positive eigenvector) the LCE $\tau(T, Y', w)$ is < 1 for any finite w as soon as T' is scrambling (i.e., any two rows of T' have at least one positive entry in a coincident position, which means that no two rows are orthogonal). This result is not entirely surprising because when T' is scrambling, then $d(X'T, Y'T) < d(X', Y')$. However, this inequality alone is not sufficient to make a contraction-type argument as in (1.8) when $\tau(T, Y', w_o) < 1$.

The scrambling condition is obviously much weaker than the condition $T > 0$, which must be satisfied in order to have $\tau(T) < 1$. For these reasons the LCE is useful not only here but also in other similar situations when we are interested in the ratio

$$R(X, Y) \stackrel{\text{def}}{=} d(X'T, Y'T)/d(X', Y') \tag{1.9}$$

with X and Y not colinear. The remainder of this paper is devoted to the study of $\tau(T, Y', w)$, with an emphasis on the conditions under which $\tau(T, Y', w) < 1$.

2. Preliminary definitions and results. We first define the subset Ω of the nonnegative quadrant $\mathbb{R}_+^n$ as the set of all vectors with components between 0 and 1, with at least one component equal to 0 and one equal to 1:

$$\Omega = \{E = (e_i) \in \mathbb{R}_+^n : 0 \le e_i \le 1; \text{ at least one } e_i = 0, \text{ one } e_i = 1\}. \tag{2.1}$$

With T column-allowable, we define for a fixed vector $Y = (y_i) > 0$ the row-stochastic matrix $P(T, Y)$ as

$$P(T, Y)_{ij} \stackrel{\text{def}}{=} \frac{y_j t_{ji}}{\sum_{q=1}^{q=n} y_q t_{qi}}, \quad i, j = 1, 2, \ldots, n. \tag{2.2}$$

The matrices $P(T, Y)$ and T have "transposed incidences": the incidence of $P(T, Y)$ is that of T'. If we let $P(T, Y)_i$ denote the ith row of $P(T, Y)$, and if $E = (e_i)$ is a vector of Ω, we note that $P(T, Y)_i E$ is the scalar product $\sum_{k=1}^n P(T, Y)_{ik} e_k$.

We next let $A \circ B$ be the Hadamard (componentwise) product of two matrices or vectors. As in [2] we will express any vector $X > 0$ in a way that will simplify the expressions for $d(X', Y')$ and $d(X'T, Y'T)$.

PROPOSITION 2.1. *For any $X = (x_i) > 0$ not colinear to $Y > 0$ there exists a unique $s > 0$ and a vector $E \in \Omega$ such that $Y + sY \circ E$ and X are colinear. Then*

$$d(X', Y') = d([Y + sY \circ E]', Y') = \ln(1 + s), \tag{2.3}$$

$$d(X'T, Y'T) = \max_{i,j} \ln\left(\frac{1 + s \times P(T, Y)_i E}{1 + s \times P(T, Y)_j E}\right). \tag{2.4}$$

Proof. The components x_i of X can be written as $x_i = y_i(r + \sigma e_i)$, where $r \stackrel{\text{def}}{=} \min_i x_i/y_i$, $\sigma \stackrel{\text{def}}{=} \max_i x_i/y_i - \min_i x_i/y$, and $E = (e_i) = (x_i/y_i - r)/\sigma$ ($\sigma > 0$ because X and Y are not colinear). Then

$$X = Yr + \sigma Y \circ E, \tag{2.5}$$

which after setting $s = \sigma/r$ shows that $Y + sY \circ E$ and X are colinear. This proves (2.3). Also

$$d(X'T, Y'T) = d([Y + sY \circ E]'T, Y'T) = \max_{i,j} \ln \left[\frac{1 + s\sum_k e_k P(T,Y)_{ik}}{1 + s\sum_k e_k P(T,Y)_{jk}} \right], \tag{2.6}$$

which is the desired result of (2.4).

If we define $w^* \stackrel{\text{def}}{=} \exp(w) - 1$, we then have

$$\tau(T, Y, w) = \max_{i,j} \sup_{\substack{0 < s \le w^* \\ E \in \Omega}} \frac{\ln \left[\frac{1+sP(T,Y)_i E}{1+sP(T,Y)_j E} \right]}{\ln(1+s)}. \tag{2.7}$$

We will now proceed in two steps: First we will find for fixed s, i, j the sup over E of the bracketed expression in the numerator. Then we will seek the supremum over $0 < s \le w^*$ of the ratio of the two logarithms.

2.1. Supremum over *E*. We define for two probability-normed vectors $a = (a_i)$, $b = (b_i)$ the function

$$Z(s, a, b, E) = \frac{1 + sa'E}{1 + sb'E}, \tag{2.8}$$

of which we seek the supremum for $E = (e_i) \in \Omega$. ($a'E = \sum_{k=1}^n a_k e_k$, $b'E = \sum_{k=1}^n b_k e_k$, so that a and b represent the ith and jth rows of $P(T, Y)$.)

A supremum is necessarily reached for each e_i equal to either 0 or 1. In this context we define the finite subset Ω' of Ω consisting of vectors having their last k components equal to 1 ($k = 2, 3, \ldots, n$) and the others equal to 0:

$$\Omega' = \{E(k) = (0, 0, \ldots, 0, 1, 1, , , 1), \text{ first "1" in } k\text{th position}, k = 2, 3, \ldots, n\}. \tag{2.9}$$

We now reorder the components (a_i, b_i) in the following way. We first have those components that are both 0. Then we have in increasing order of the ratios $r_i = a_i/b_i$ the components for which $b_i > 0$. Finally we have in increasing order of the a_i's those components for which $b_i = 0$. For example, if $a = (0.3\ 0\ 0\ 0.2\ 0.5)$ and $b = (0\ 0\ 0.2\ 0.1\ 0.7)$, then the vectors with reordered components are $a = (0\ 0\ 0.5\ 0.2\ 0.3)$ and $b = (0\ 0.2\ 0.7\ 0.1\ 0)$. The corresponding vector of increasing ratios is $r = (0/0\ 0/0.2\ 0.5/0.7\ 0.2/0.1\ 0.3/0)$.

In what follows we will assume that the components of a and b have been reordered in this manner, and we say that the pair (a, b) has the *increasing ratio property (IRP).* (Note that in general the reordering of (a, b) and of (b, a) are not the same. In fact the two orderings are mirror images of one another.)

We first dispose of two trivial cases:

i. If $a = b$, then $Z(s, a, b, E)$ is 1 for any E.

ii. If a and b are orthogonal (i.e., do not have a positive term in a coincident position, which means $a_i b_i = 0$ for all i), then there exists an $E(k)$ such that $Z(s, a, b, E(k))$ is equal to its maximum possible value $1 + s$.

We now assume that $a \neq b$ and that a and b are not orthogonal. Then there is necessarily at least one ratio $r_m = a_m/b_m$ that is strictly less than 1 and one that is strictly larger than 1 (and possibly $+\infty$). We thus define

$$m_1 = \{m/\ r_m \leq 1 < r_{m+1}\}, \tag{2.10}$$

$$m_2 = \min\ \{m/\ r_m = r_{m+1} = \cdots = r_n\}. \tag{2.11}$$

In words, r_{m_1} is the largest ratio in the list $\{r_k\}_{k=1,2,\ldots,n}$ to be ≤ 1; r_{m_2} is the first in the list to be equal to the last (and largest) ratio r_n (where r_n may be $+\infty$). With the example $a = (0\ 0\ 0.5\ 0.2\ 0.3)$, $b = (0\ 0.2\ 0.7\ 0.1\ 0)$, $r = (0/0\ 0/0.2\ 0.5/0.7\ 0.2/0.1\ 0.3/0)$, we have $m_1 = 3, m_2 = 5$. If $a = (0.56\ 0.4\ 0.04)$, $b = (0.67\ 0.3\ 0.03)$, then $r = (0.56/0.67\ 0.4/0.3\ 0.04/0.03)$ so that $m_1 = 1, m_2 = 2$ (m_2 is strictly larger than m_1 because $a \neq b$).

We will now use this reordering to partition the set of all positive real numbers into intervals $I(k)$ within which the sup of $Z(s, a, b, E)$ over E is attained for $E(k)$. We define the $m_2 - m_1$ half-open intervals

$$I(k) \stackrel{\text{def}}{=} [S(k-1), S(k)\ ), \quad k = m_1 + 1, m_1 + 2, \ldots, m_2,$$

where the quantities $S(k)$ are given by

$$S(m_1) = 0; \quad S(k) = \frac{a_k - b_k}{b_k \sum_{p=k}^{n} a_p - a_k \sum_{p=k}^{n} b_p}, \qquad k = m_1 + 1, m_1 + 2, \ldots, m_2. \tag{2.12}$$

The $S(k)$'s are nonnegative numbers that satisfy

$$S(m_1) = 0 \leq S(m_1 + 1) \leq \cdots \leq S(m_2 - 1) < S(m_2) = +\infty, \tag{2.13}$$

$$Z(S(k), a, b, E(k)) = r_k, \quad k = m_1 + 1, m_1 + 2, \ldots, m_2. \tag{2.14}$$

In short the $m_2 - m_1$ intervals $I(k)$ constitute a partition of the set of real positive numbers such that

$$s \in I(k) \quad \Longleftrightarrow r_{k-1} \leq Z(s, a, b, E(k)) < r_k. \tag{2.15}$$

We will now show that when (a, b) has the IRP, then for a fixed s in any $I(k)$, the supremum of $Z(s, a, b, E)$ over $E = (e_i) \in \Omega$ is attained at $Z(s, a, b, E(k))$.

PROPOSITION 2.2. *For two probability-normed vectors (a, b) having the IRP, with m_1, m_2 given in* (2.10)–(2.11)*, we have*

$$\begin{aligned} s \in I(k) \Rightarrow & \sup_{E \in \Omega} Z(s, a, b, E) = Z(s, a, b, E(k)) \\ & = \frac{1 + sa'E(k)}{1 + sb'E(k)}, \quad k = m_1 + 1, m_1 + 2, \ldots, m_2, \end{aligned} \tag{2.16}$$

$$b'E(k) \leq a'E(k) \leq 1. \tag{2.17}$$

Proof. The proof will hinge upon the following elementary numerical results concerning four nonnegative numbers u, u', v, v':

$$\frac{u+u'}{v+v'} \lessgtr \frac{u'}{v'} \iff \frac{u}{v} \lessgtr \frac{u'}{v'} \iff \frac{u}{v} \lessgtr \frac{u+u'}{v+v'}. \tag{2.18}$$

We recall that a sup is necessarily reached with each e_k equal to either 0 or 1. Let us assume that the sup is reached for some $E^* = (e_k^*)$ that has a "1" to the left of a "0":

$$\sup_{E\in\Omega} Z(s,a,b,E) = Z(s,a,b,E^*) = \frac{1+s\sum_{i=1}^n a_i e_i^*}{1+s\sum_{i=1}^n b_i e_i^*},\ \exists\ w<q \text{ with } e_w^*=1, e_q^*=0. \tag{2.19}$$

We will show that a contradiction will follow, which will leave only the elements of Ω' as candidates for the optimum E. Indeed if $Z(s,a,b,E^*) \le r_q$, then (2.18) shows that $Z(s,a,b,E^*)$ can be increased by changing e_q^* to 1. If $Z(s,a,b,E^*) > r_q$, then (2.18) shows that $Z(s,a,b,E^*)$ can be increased by changing e_w^* to 0. This contradicts the assumption that E^* is a supremum and shows that the sup is necessarily reached for some element of Ω'. (We note in particular that e_n, the last component of E, is necessarily 1 at the optimal value. This insures that one component is indeed 1, which was an early requirement.)

We now show that $E(k)$ is the element of Ω' at which the supremum in reached. To simplify the writing we define $h(k) = Z(s,a,b,E(k))$. We will show that if $s \in I(k)$, then $h(k)$ is the maximum value of $h(m)$ for all m. First $h(k+1) \le h(k)$ because we obtain $h(k+1)$ by removing (a_k, b_k) from the numerator and the denominator of $h(k)$, which from (2.18) decreases $h(k)$ since $h(k) < r_k \le r_{k+1}$; $h(k+2)$ is obtained by removing (a_{k+1}, b_{k+1}) from $h(k+1)$ but $h(k+1) \le h(k) < r_{k+1}$ so $h(k+2) \le h(k+1)$. Hence with each increase in the index $j, h(k+j)$ becomes smaller because pairs (a_{k+j}, b_{k+j}) with increasing ratios r_{k+j} are removed while $h(k+j)$ decreases. A similar reasoning holds if j decreases: $h(k-1) \le h(k)$ because we obtain $h(k-1)$ by adding (a_{k-1}, b_{k-1}) to the numerator and the denominator of $h(k)$, which from (2.18) implies $r_{k-1} \le h(k-1) \le h(k)$; $h(k-2)$ is obtained by adding (a_{k-2}, b_{k-2}) to $h(k-1)$; however, $r_{k-2} \le r_{k-1} \le h(k-1)$, which insures that $r_{k-2} \le h(k-2) \le h(k-1)$. Hence with each decrease in the index $j, h(k-j)$ becomes smaller while staying larger than r_{k-j}. This shows that for $s \in I(k)$, the maximum value of $h(m)$ is $h(k)$.

2.2. Supremum over s. Now that we have $Z(s,a,b,E(k))$ as the supremum of $Z(s,a,b,E)$ over E in each $I(k)$, we seek the supremum of $\ln[Z(s,a,b,E(k)]/\ln(1+s)$ for $s \in I(k)$. We thus consider the function

$$Q(s,\alpha,\beta) \stackrel{\text{def}}{=} \frac{\ln\frac{1+s\alpha}{1+s\beta}}{\ln(1+s)},\ s>0; \quad 0 \le \beta < \alpha \le 1, \tag{2.20}$$

whose derivative $Q'(s,\alpha,\beta)$ with respect to s is

$$Q'(s,\alpha,\beta) = \frac{\frac{(1+s)(\alpha-\beta)}{1+\beta s}}{1+\alpha s} - \frac{\ln\left(\frac{1+\alpha s}{1+\beta s}\right)}{\ln(1+s)}. \tag{2.21}$$

Two examples of the function $Q(s,\alpha,\beta)$ are given in Figure 2.1, one with $\alpha+\beta = 1.3$, the other with $\alpha+\beta = 0.35$. As we will see below, the function is monotone decreasing when $\alpha+\beta \ge 1$ and has one maximum when $\alpha+\beta < 1$.

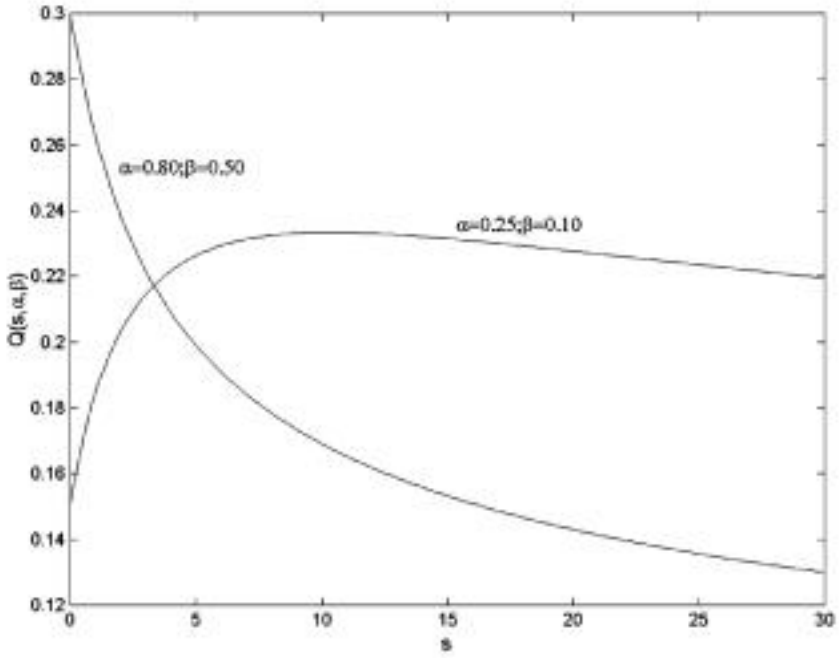


FIG. 2.1. *Two examples of the function* $Q(s, \alpha, \beta)$.

PROPOSITION 2.3. *When* $\alpha > 0$ *the function* $Q(s, \alpha, 0)$ *increases monotonically from* α *to* 1 *as* s *grows from* 0^+ *to* $+\infty$.

i. *If* $\alpha + \beta \geq 1, \beta > 0$, *then* $Q(s, \alpha, \beta)$ *decreases monotonically from* $\alpha - \beta$ *to* 0 *with* s.

ii. *If* $\alpha + \beta < 1, \beta > 0$, *then* $Q(s, \alpha, \beta)$ *first increases from* $\alpha - \beta$ *to a maximum* $Q^*(\alpha, \beta)$, *then decreases to* 0 *as* $s \to +\infty$. *The maximum* $Q^*(\alpha, \beta)$ *is reached at a value* $s^* = s^*(\alpha, \beta)$ *of* s *that is the unique positive root of the equation (in* s*)*

$$\left(\frac{1 + s\alpha}{1 + s\beta}\right) = (1 + s)^{\frac{(1+s)(\alpha-\beta)}{(1+\alpha s)(1+\beta s)}}. \tag{2.22}$$

Then

$$Q^*(\alpha, \beta) = \frac{\ln \frac{1+s^*\alpha}{1+s^*\beta}}{\ln(1 + s^*)} = \frac{(1 + s^*)(\alpha - \beta)}{(1 + \alpha s^*)(1 + \beta s^*)} \leq (\alpha - \beta)/(\alpha + \beta) < 1. \tag{2.23}$$

Proof. By setting $Q'(s, \alpha, \beta)$ of (2.21) equal to 0, one obtains (2.23) and thus (2.22). The upper bound $(\alpha - \beta)/(\alpha + \beta)$ for $Q^*(\alpha, \beta)$ is the maximum (over s^*) of the third term in (2.23). Other elementary details are omitted.

There is no closed-form expression for $s^*(\alpha, \beta)$, the root of (2.22). However, calculating $s^*(\alpha, \beta)$ is an elementary numerical problem because $Q(s, \alpha, \beta)$ is then a simple function that increases then decreases.

Bearing in mind the expression for $\tau(T,Y,w)$ given in (2.7), we now have

$$\tau(T,Y,w) = \max_{i,j} \sup_{\substack{0<s<w^* \\ E\in\Omega}} Q[s, P(T,Y)_i E, P(T,Y)_j E]. \tag{2.24}$$

We will partition $[0,w^*)$ using the intervals $I(k)$ in which we now know that the supremum over E is $E(k)$. If for any $z>0$ we let $k(z)$ denote the index of the interval that contains z, then $w^* \in I(k(w^*)) = [S(k(w^*)-1), S(k(w^*)))$. Then the interval $[0,w^*)$ is the union of the intervals $I(k) = [S(k-1), S(k))$ for k going from m_1+1 to $k(w^*)-1$, to which we add the interval $[S(k(w^*)-1), w^*)$ (which for simplicity of notation we will call $I(k(w^*))$ below, even though strictly speaking it is contained in $I(k(w^*))$ and not equal to $I(k(w^*))$. The next section gives the main result, which hinges upon this particular partitioning of the interval $[0,w^*)$.

3. Main result and applications. With the notation given above, the following theorem provides a near-closed-form expression for the LCE $\tau(T,Y,w)$.

THEOREM 3.1. *For a column-allowable matrix T and a positive vector Y, the LCE $\tau(T,Y,w)$ is equal to*

$$\begin{aligned}&\tau(T,Y,w)\\ &= \max_{i,j} \max_{p=m_1+1,m_1+2,\ldots,k(w^*)-1,k(w^*)} \sup_{s\in I(p)} Q[s, P(T,Y)_i E(p), P(T,Y)_j E(p)],\end{aligned} \tag{3.1}$$

where the vectors $(P(T,Y)_i, P(T,Y)_j)$ have the IRP, i.e., the ratios

$$r_k = P(T,Y)_{ik}/P(T,Y)_{jk} \tag{3.2}$$

are increasing for $k = m_1, m_1+1, \ldots, m_2$, with m_1, m_2 defined in (2.10), (2.11). (m_1, m_2, and $k(w^*)$ depend on the particular pair of indices (i,j).)

If T' is not scrambling (i.e., T' has two orthogonal rows), then two vectors $P(T,Y)_i$ and $P(T,Y)_j$ are orthogonal and $\tau(T,Y,w)=1$. If the columns of T are multiples of a common nonzero vector Z (T of rank 1), then all the rows of $P(T,Y)$ are identical and $\tau(T,Y,w)=0$.

In the general case (T scrambling and of rank >1), we recall the notation

$$S(m_1) = 0, \tag{3.3}$$

$$S(k) = \frac{P(T,Y)_{ik} - P(T,Y)_{jk}}{P(T,Y)_{jk}\sum_{p=k}^{n} P(T,Y)_{ip} - P(T,Y)_{ik}\sum_{p=k}^{n} P(T,Y)_{jp}}; \quad k = m_1+1, m_1+2, \ldots, m_2, \tag{3.4}$$

and $w^* = \exp(w) - 1$. For $p = m_1+1, m_1+2, \ldots, k(w^*)-1$ the intervals $I(p)$ in (3.1) are $I(p) = [S(p-1), S(p))$ as in (2.12). The last interval $I(k(w^*))$ stops at w^* and is $[S(k(w^*)-1), w^*)$ rather than $[S(k(w^*)-1), S(k(w^*)))$.

We next define the values $Q_{le}^{(p)}$ and $Q_{re}^{(p)}$ of the function $Q[s, P(T,Y)_i E(p), P(T,Y)_j E(p)]$ at the left end (le) and right end (re) of the corresponding interval $I(p)$, i.e.,

$$p = m_1+1 \Rightarrow \begin{cases} Q_{le}^{(p)} = P(T,Y)_i E(m_1+1) - P(T,Y)_j E(m_1+1), \\ Q_{re}^{(p)} = \ln(r_{m_1+1})/\ln(1+S(m_1+1)), \end{cases} \tag{3.5}$$

$$p = m_1+2, m_1+3, \ldots, k(w^*)-1 \Rightarrow \quad \begin{cases} Q_{le}^{(p)} = \ln(r_{p-1})/\ln(1+S(p-1)), \\ Q_{re}^{(p)} = \ln(r_p)/\ln(1+S(p)), \end{cases} \tag{3.6}$$

$$p = k(w^*) \Rightarrow \quad \begin{cases} Q_{le}^{(p)} = \ln(r_{k(w*)-1})/\ln(1+S(k(w^*)-1)), \\ Q_{re}^{(p)} = \ln\left(\dfrac{1+w^*P(T,Y)_iE(k(w^*))}{1+w^*P(T,Y)_jE(k(w^*))}\right)/\ln(1+w^*). \end{cases} \tag{3.7}$$

When $P(T,Y)_iE(p) + P(T,Y)_jE(p) < 1$, *we let* s^* *be the value at which* $Q[s, P(T,Y)_iE(p), P(T,Y)_jE(p)]$ *reaches its maximum (see Proposition* 2.3*).*

The suprema of (3.1) *are now as follows:*

i. *If either* $(P(T,Y)_iE(p) + P(T,Y)_jE(p) \geq 1$ *and* $P(T,Y)_jE(p) > 0)$ *or* s^* *is to the* left *of the interval* $I(p)$*, then the supremum is reached at the left end of* $I(p)$*:*

$$\sup_{s\in I(p)} Q[s, P(T,Y)_iE(p), P(T,Y)_jE(p)] = Q_{le}^{(p)}. \tag{3.8}$$

ii. *If either* $(P(T,Y)_iE(p) + P(T,Y)_jE(p) < 1$ *and* $P(T,Y)_jE(p) = 0)$ *or* s^* *is to the* right *of* $I(p)$*, then the supremum is reached at the right end of* $I(p)$*:*

$$\sup_{s\in I(p)} Q[s, P(T,Y)_iE(p), P(T,Y)_jE(p)] = Q_{re}^{(p)}. \tag{3.9}$$

iii. *If* s^* *is inside* $I(p)$*, then*

$$\sup_{s\in I(p)} Q[s, P(T,Y)_iE(p), P(T,Y)_jE(p)] = Q[s^*, P(T,Y)_iE(p), P(T,Y)_jE(p)]. \tag{3.10}$$

Proof. The expressions obtained in this theorem are direct consequences of previous results stemming from Proposition 2.3. Equation (3.1) reflects the partitioning of the interval $[0, w^*)$ into intervals $I(p)$ over which the sup over E is E_p. The supremum over each $I(p)$ is at the left end or right end of $I(p)$, or at the value s^* at which $Q[s, P(T,Y)_iE(p), P(T,Y)_jE(p)]$ reaches a maximum, depending on the value of $P(T,Y)_iE(p)+P(T,Y)_jE(p)$ relative to 1 (see Proposition 2.3). The value $Q_{le}^{(p)}$ given in (3.5) for the left end of the first interval $I(m_1+1) = [S(m_1), S(m_1+1)) = [0, S(m_1+1))$ is obtained by taking the limit of $Q[s, P(T,Y)_iE(m_1+1), P(T,Y)_jE(m_1+1)]$ for $s \to 0$.

We now give a simple condition for $\tau(T,Y,w)$ to be strictly less than 1.

COROLLARY 3.2. *For a column-allowable matrix* T*, a positive vector* Y*, and any* $w > 0$*, the LCE* $\tau(T,Y,w)$ *is strictly less than* 1 *if and only if* T' *is scrambling.*

Proof. We showed above that T' is not scrambling $\Longrightarrow \tau(T,Y,w) = 1$. Conversely, let us assume that $\tau(T,Y,w) = 1$. From Proposition 2.3 a supremum of 1 can be reached only if there is a $P(T,Y)_jE(p) = 0$ and $P(T,Y)_jE(p) = 1$, which means that T' has two orthogonal rows and is not scrambling.

A simple numerical example is provided by the triangular matrix

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix}$$

with positive left eigenvector $Y' = \begin{pmatrix} 2 & 1 & 1 \end{pmatrix}$ and corresponding eigenvalue $\rho(T) = 2$. No conclusion can be drawn on the iterates $X_o'T^p$ by considering the powers of this

reducible matrix which has a coefficient of ergodicity equal to 1 and whose powers remain triangular. However, T' is scrambling, and the results proved here immediately yield the desired conclusion, namely, exponential convergence to 0 of the projective distance $d(X_o'T^p, Y')$. Indeed

$$d(X_o'T^p, Y') \leq \tau(T^p, Y', w_o)w_o \leq \tau(T, Y', w_o)^p w_o \underset{p\longrightarrow\infty}{\longrightarrow} 0. \tag{3.11}$$

With $w_o = 2$ a MATLAB program (available from the author) yields $\tau(T, Y, 2) = 0.88$. As $X_o'T^p$ approaches Y' in direction, the quantity $\tau(T, Y, 0) \stackrel{\text{def}}{=} \lim_{w\to 0} \tau(T, Y, w)$ is an asymptotic rate of convergence equal in the present case to 0.75.

If in order to emphasize the dependence on (i, j) we write $m_1(i, j)$ for m_1 of (3.1), then

$$\tau(T, Y, 0) = \lim_{w\to 0} \tau(T, Y, w) = \max_{i,j} [P(T, Y)_i E(m_1(i, j)+1) - P(T, Y)_j E(m_1(i, j)+1)]. \tag{3.12}$$

It can easily be seen that this asymptotic rate of convergence $\tau(T, Y, 0)$ is equal to $\tau_1(P(T, Y))$, where τ_1 is the classical coefficient of ergodicity defined on row-stochastic matrices [4], i.e., for any row-stochastic matrix $Q = (q_{ij})$,

$$\tau_1(Q) = 0.5 \max_{i,j} \sum_k |q_{ik} - q_{jk}| = \max_{i,j} \sum_{\substack{k\in\Delta(i,j) \\ \Delta(i,j)\stackrel{\text{def}}{=}\{k:q_{ik}-q_{jk}>0\}}} (q_{ik} - q_{jk}). \tag{3.13}$$

We showed in [1] that $\tau(T) = \sup_{Y>0} \tau_1(P(T, Y))$, and we thus come full circle with

$$\sup_{Y>0;\sigma>0} \tau(T, Y, \sigma) = \tau(T) = \sup_{Y>0} \tau_1(P(T, Y)) = \sup_{Y>0} \tau(T, Y, 0). \tag{3.14}$$

Acknowledgment. The author wishes to thank an anonymous referee, whose comments substantially improved the paper.

REFERENCES

[1] M. ARTZROUNI AND X. LI, *A note on the coefficient of ergodicity of a column-allowable nonnegative matrix*, Linear Algebra Appl., 214 (1995), pp. 93–101.

[2] M. ARTZROUNI AND O. GAVART, *Nonlinear matrix iterative processes and generalized coefficients of ergodicity*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1343–1353.

[3] R. A. HORN AND C.R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[4] E. SENETA, *Non-Negative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, Berlin, 1981.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 517–531

MULTIVARIATE FILTER BANKS HAVING MATRIX FACTORIZATIONS*

QIUHUI CHEN†, CHARLES A. MICCHELLI‡, SILONG PENG§, AND YUESHENG XU¶

Abstract. In this paper we design vector-valued multivariate filter banks with a polyphase matrix built by a matrix factorization. These filter banks are suitable for the construction of multivariate multiwavelets with a general dilation matrix. We show that block central symmetric orthogonal matrices provide filter banks having a uniform linear phase. Several examples are included to illustrate our construction.

Key words. multiwavelets, filter banks, matrix factorizations

AMS subject classifications. 42C40, 65T60

DOI. 10.1137/S0895479802412735

1. Introduction.

In 1976, for the purpose of compressing speech signals by subband coding schemes, Croisier, Esteban, and Galand [3] introduced an invertible filter bank, which decomposes a discrete signal into two signals of half its size by using a filtering and subsampling procedure. They showed that the signal can be recovered from these subsampled signals by canceling the aliasing terms with a particular class of filters called conjugate mirror filters (CMFs). This breakthrough motivated an active research effort to build a complete filter bank theory. Necessary and sufficient conditions for decomposing a signal into subsampled components with a filtering scheme, and recovering the same signal with an inverse transform, were established by Smith and Barnwell [20], Vaidyanathan [22], and Vetterli [23].

Filter banks are closely associated with wavelets. The multiresolution theory shows that CMFs and the orthonormal wavelet basis of $L^2(R^d)$ are intimately linked. In fact, a continuous-time wavelet basis can be obtained by iterated filter banks, and filter banks can be considered discrete wavelet transforms. The equivalence between the continuous-time wavelet theory and discrete filter banks leads to a new fruitful interface between digital signal processing and harmonic analysis.

The multiresolution analysis (MRA) theory provides a natural framework for understanding wavelets and filter banks. According to MRA, refinable functions and

*Received by the editors August 5, 2002; accepted for publication (in revised form) by A. H. Sayed February 18, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/simax/25-2/41273.html

†Department of Scientific Computing and Computer Applications, Zhongshan University, Guangzhou, 510275, People's Republic of China. The research of this author was supported in part by NSFC under grants 10201034 and 10226036.

‡Department of Mathematics and Statistics, State University of New York, The University at Albany, Albany, NY 12222 (cam@math.albany.edu). The research of this author was supported in part by the National Science Foundation under grant DMS-9973427.

§Institute of Automation, Chinese Academy of Sciences, Beijing 100080, People's Republic of China.

¶Corresponding author. Department of Mathematics, West Virginia University, Morgantown, WV 26506, and Institute of Mathematics, Academy of Mathematics and System Science, Chinese Academy of Science, Beijing, 100080, People's Republic of China. Current address: Department of Mathematics, Syracuse University, 215 Carnegie Hall, Syracuse, NY 13244 (yxu06@syr.edu). The research of this author was supported in part by the National Science Foundation under grants DMS-9973427 and EPS-0132740, by the program "One Hundred Distinguished Chinese Scientists" of the Chinese Academy of Sciences, and by the WVU Research Corporation under the Research Incentive Competitive Grant Program.

wavelets are completely determined by a low-pass filter and high-pass filters, respectively. In subband code schemes, a low-pass filter and high-pass filters are used as analysis filters and synthesis filters which form perfect reconstruction filter banks. In [7] Herrmann designed maximally flat filters having the finite impulse response (FIR). Daubechies [5] obtained the corresponding univariate FIR two-channel perfect reconstruction filter banks and used them to construct univariate orthonormal wavelets with any prescribed regularity, having compact support and maximally vanishing moments. It is well known that there does not exist a symmetric orthonormal wavelet with a compact support in the univariate dyadic dilation case; that is, two-channel perfect reconstruction FIR banks having a linear phase are not available in the univariate case. Historically, this led to an intense interest in univariate multichannel, high-dimensional, and vector-valued filter banks which correspond to M-band wavelets, multivariate wavelets, and multiwavelets, respectively. For a definitive study of univariate M-band wavelets, see [19].

Our interest here is in multivariate filter banks. Indeed, the study of the two-dimensional case is crucial for digital image processing. A commonly used method builds multivariate filter banks by the tensor products of univariate filters. This construction of filter banks focuses excessively on the coordinate direction. Therefore, nontensor product approaches for construction of multivariate filter banks or wavelets are desirable. Much interest has been given to the study of nonseparable wavelets in $L^2(R^d)$ (see, for example, [2], [10], [11], and also [15], [16] for constructions of multivariate wavelets on invariant sets), as well as to multiwavelets and corresponding vector-valued filter banks [1], [4], [6], [13], [14].

It is not easy to design multivariate filter banks. At present, no general method is available for designing multivariate filter banks and vector-valued filter banks. There are two fundamental difficulties that one encounters in the design of a low-pass filter and high-pass filters which are used for the construction of refinable functions and wavelets, respectively. The first challenge lies in finding trigonometric polynomials that satisfy the perfect reconstruction condition, and the second is met when we extend a block unit vector of trigonometric polynomials to a unitary matrix. These two problems are both difficult; cf. [10], [12], [13]. Most of the current study in multivariate wavelets is given to a dilation matrix with determinant two [11], since in this case only one high-pass filter is needed for the construction and the matrix extension is the same as the univariate two-channel case [2].

The main purpose of this paper is to present a unified approach for the construction of multivariate vector-valued filter banks for arbitrary dilation matrices, using polyphase factorization and block central symmetric orthogonal matrices. In this regard, we were influenced by the papers [17], [18], [21], [11] of Pollen, Vaidyanathan, Nguyen, Kovacevic, and Vetterli. In 1990, Pollen characterized all two-channel univariate scalar FIR wavelet filter banks as a product of a one-parameter family of unitary matrices. Independently, Nguyen and Vaidyanathan [17], [21] used paraunitary transfer matrices to design multichannel CMF banks for particular digital signal processing problems in the one-dimensional case. They referred to them as FIR paraunitary CMF banks and emphasized that in the two-channel case, all real-coefficient FIR paraunitary CMF banks can be represented in this manner. Essentially, their design method coincides with the method of Pollen. Recently, Kovacevic and Vetterli extended this idea to the multidimensional case and constructed some examples of nonseparable wavelet filter banks having a cascade structure in three cases: quincunx lattice, separable lattice, and the face-centered orthorhombic lattice. We shall provide

further insight into the methods presented in [18], [11], [17], and [21] in the case of a vector-valued, multidimensional, and arbitrary lattices.

Often, one seeks filter banks leading to smooth wavelets. However, in the application of filter banks to texture analysis, experiments show that "smooth" filter banks are not suitable because texture images are not smooth. The family of filter banks given in this paper is suitable in this context as it is difficult to achieve smoothness.

We organize this paper into four sections. Section 2 is devoted to a development of multivariate filter banks having matrix factorizations. In section 3, we construct low-pass filters having a uniform linear phase by using block central symmetric matrices. This class of low-pass filters leads to symmetric multiwavelets. We present several examples in section 4 to illustrate the general construction of filter banks.

2. Filter banks having matrix factorizations. In this section, we describe a general construction of vector-valued multivariate filter banks having matrix factorizations. We first present a condition that ensures that the low-pass filter satisfies the *perfect reconstruction* condition. For a low-pass filter having a matrix factorization and satisfying the perfect reconstruction condition, we then develop the corresponding high-pass filters which also have a matrix factorization form.

Let A be a $d \times d$ matrix with integer entries such that all its eigenvalues are greater than 1. Let

$$s := |\det A|,$$

$$Z_s := \{0, 1, \ldots, s-1\},$$

and

$$\Omega(A) := \{\gamma_j : j \in Z_s\}$$

with $\gamma_0 = 0$ being a complete set of representatives of the distinct coset of Z^d/AZ^d. Each $\gamma \in Z^d$ determines the coset

$$\overline{\gamma} := AZ^d + \gamma,$$

and by definition

$$\bigcup_{j \in Z_s} \overline{\gamma}_j$$

forms a partition of Z^d.

Given an $r \times r$ matrix of trigonometric polynomials

$$m_0(\xi) := \sum_{\alpha \in Z^d} c_\alpha e^{-i\alpha \cdot \xi}, \quad \xi \in R^d,$$

with a finite sequence c_α, $\alpha \in Z^d$, of matrix of order r, its polyphase factors are the $r \times r$ matrices of trigonometric polynomials defined for $l \in Z_s$ as

$$m_{0,l}(\xi) = \sum_{\alpha \in Z^d} c_{A\alpha+\gamma_l} e^{-i\alpha \cdot \xi}, \quad \xi \in R^d. \tag{2.1}$$

Reversing the process, we can construct the matrix of trigonometric polynomials m_0 from its polyphase factors $m_{0,j}$, $j \in Z_s$, using the formula

$$m_0(\xi) = \sum_{l \in Z_s} m_{0,l}(A^T\xi)e^{-i\gamma_l \cdot \xi}, \quad \xi \in R^d. \tag{2.2}$$

The construction of multivariate compactly supported orthonormal multiwavelets using MRA leads to the following two problems:

(i) Find an $r \times r$ matrix of trigonometric polynomials m_0 such that its polyphase factors $m_{0,l}$, $l \in Z_s$, satisfy the perfect reconstruction condition

$$W_0(\xi)W_0(\xi)^* = \frac{1}{s}I_r, \qquad \xi \in R^d, \tag{2.3}$$

where W_0 is an $r \times rs$ matrix defined by

$$W_0(\xi) := (m_{0,l}(\xi) : l \in Z_s), \quad \xi \in R^d.$$

(ii) Find $s-1$ $r \times r$ matrices m_j, $j \in Z_s \setminus \{0\}$, of trigonometric polynomials such that the $rs \times rs$ block matrix composed of their polyphase factors given by

$$W(\xi) := (m_{j,l}(\xi) : j, l \in Z_s), \quad \xi \in R^d, \tag{2.4}$$

has the property that $\sqrt{s}W(\xi)$, $\xi \in R^d$, is a unitary matrix.

The purpose of this section is to design a family of filter banks m_0 which have property (i) such that their high-pass filters m_j, $j \in Z_s \setminus \{0\}$, have property (ii) and are easily constructed. Precisely, we choose the low-pass filter m_0 to have the form

$$m_0(\xi) = \frac{1}{\sqrt{s}}X(\xi) \prod_{j \in Z_N} \left(U_j D\left(A^T\xi\right)\right) V, \quad \xi \in R^d, \tag{2.5}$$

where N is an arbitrarily chosen positive integer, $X(\xi)$ is the $r \times rs$ block matrix function defined by

$$X(\xi) := \left(e^{-i\gamma_0 \cdot \xi}I_r, \ldots, e^{-i\gamma_{s-1} \cdot \xi}I_r\right),$$

U_j, $j \in Z_N$, are arbitrary $rs \times rs$ real orthogonal matrices, V is an arbitrary $rs \times r$ matrix satisfying

$$V^TV = I_r, \tag{2.6}$$

and $D(\xi)$ is the block diagonal matrix of order $rs \times rs$ with trigonometric entries defined by

$$D(\xi) := \operatorname{diag}(e^{-i\gamma_0 \cdot \xi}I_r, \ldots, e^{-i\gamma_{s-1} \cdot \xi}I_r).$$

Now we shall show that the filter m_0 having the form (2.5) satisfies the perfect reconstruction condition (i).

THEOREM 2.1. *For any $rs \times rs$ real unitary matrices U_j, $j \in Z_N$, and any $rs \times r$ real constant matrix V satisfying* (2.6)*, the symbol m_0 defined in* (2.5) *satisfies the perfect reconstruction condition* (i).

Proof. We shall confirm that polyphase factors $m_{0,j}$, $j \in Z_s$, of m_0 satisfy (i). To this end, according to (2.2) and (2.5) we observe that the polyphase factors are given by

$$W_0^T(\xi) = \frac{1}{\sqrt{s}} \left(\prod_{j \in Z_N} U_j D(\xi) \right) V, \quad \xi \in R^d,$$

from which we conclude that

$$W_0(\xi) W_0^*(\xi) = \frac{1}{s} V^T \left(\prod_{j \in Z_N} D^T(\xi) U_{N-1-j}^T \right) \left(\prod_{j \in Z_N} U_j D(-\xi) \right) V, \quad \xi \in R^d.$$

Since V satisfies (2.6), U_j, $j \in Z_N$, are orthogonal matrices, and $D(\xi)$ is unitary, we conclude that (2.3) holds, which completes the proof of this theorem. ∎

We need to impose some additional condition for V so that the symbol m_0 is a low-pass filter. In the scalar case, $m_0(0) = 1$ is the necessary and sufficient condition for the refinement equation to have a unique distribution solution. The next theorem proves that V is a specified matrix if $m_0(0) = I_r$.

THEOREM 2.2. *If m_0 is the trigonometric polynomials defined by* (2.5)*, then $m_0(0) = I_r$ if and only if*

$$V = \frac{1}{\sqrt{s}} \left(\prod_{j \in Z_N} U_{N-1-j}^T \right) V_0$$

with

$$V_0 = (I_r, I_r, \ldots, I_r)^T.$$

Proof. We define

$$\tilde{V} := \left(\prod_{j \in Z_N} U_j \right) V$$

and observe that

$$\tilde{V}^T \tilde{V} = I_r.$$

By (2.5), we have that

$$m_0(0) = \frac{1}{\sqrt{s}} V_0^T \tilde{V},$$

and so $m_0(0) = I_r$ is equivalent to the fact that

$$\frac{1}{\sqrt{s}} V_0^T \tilde{V} = I_r.$$

Consequently, it follows that

$$\operatorname{trace}(\tilde{V}^T \tilde{V}) = r, \quad \operatorname{trace}(V_0^T V_0) = rs, \quad \text{and} \quad \operatorname{trace}(V_0^T \tilde{V}) = r\sqrt{s},$$

and so by the Cauchy–Schwarz inequality for the Frobenius norm we conclude that

$$\tilde{V} = \frac{1}{\sqrt{s}} V_0. \qquad \square$$

From the above two theorems, we know that

$$m_0(\xi) = \frac{1}{s} X(\xi) \prod_{j \in Z_N} \left(U_j D\left(A^T \xi\right)\right) \left(\prod_{j \in Z_N} U_{N-1-j}^T \right) V_0 \tag{2.7}$$

is a perfect reconstructional low-pass filter, which is the starting point of our study in the rest of the paper.

The next theorem implies that the low-pass filter defined in (2.7) has accuracy at order 1. To explain this we let

$$\Omega(A^T) := \{\omega_l : l \in Z_s\}$$

be a complete set of the representatives of the coset for $Z^d/A^T Z^d$ with $\omega_0 = 0$.

THEOREM 2.3. *If $l \in Z_s$, then*

$$m_0(2\pi(A^T)^{-1}\omega_l) = \delta_{0l} I_r.$$

Proof. To arrive at this conclusion we observe for $n \in Z_s$ that

$$m_0\left(\pi_n\right) = \frac{1}{s} X(\pi_n) \prod_{j \in Z_N} \left(U_j D(2\pi\omega_n)\right) \left(\prod_{j \in Z_N} U_{N-1-j}^T \right) V_0,$$

where

$$\pi_n := 2\pi(A^T)^{-1}\omega_n.$$

Since $D(\xi)$ is 2π-periodic, we see that $D(2\pi\omega_n)$ is equal to the identity matrix I_{rs} for $n \in Z_s$. Noting the definition of $X(\xi)$, we conclude that

$$m_0\left(\pi_n\right) = \frac{1}{s} \sum_{j \in Z_s} e^{-i\gamma_j \cdot \pi_n} I_r, \quad n \in Z_s.$$

Now, using the identity (see, for example, [9])

$$\frac{1}{s} \sum_{j \in Z_s} e^{2\pi i (A^{-1}\gamma_j)\cdot \omega_n} = \delta_{0n}, \quad n \in Z_s,$$

we prove the theorem. $\square$

Our next task is to construct high-pass filters corresponding to the low-pass filter given by (2.7), which is accomplished by a matrix extension for m_0. Indeed, in this case, the matrix extension for the low-pass filter m_0 is realizable. Specifically, we extend the $rs \times r$ matrix

$$V_0 = (I_r, I_r, \ldots, I_r)^T$$

to an $rs \times rs$ real matrix

$$V := (V_0, V_1, \ldots, V_{s-1})$$

such that $\frac{1}{\sqrt{s}}V$ is an orthogonal matrix, and we define $s-1$ $r\times r$ matrices of trigonometric polynomials m_j, $j\in Z_s\setminus\{0\}$, by the equation

$$m_j(\xi)=\frac{1}{s}X(\xi)\prod_{l\in Z_N}\left(U_lD\left(A^T\xi\right)\right)\left(\prod_{l\in Z_N}U_{N-1-l}^T\right)V_j,\quad \xi\in R^d. \tag{2.8}$$

A matrix extension related to a multiwavelet construction was reformulated in [16] as a matrix equation whose general solution and particular solution were given there. The next theorem shows that the trigonometric polynomials m_j, $j\in Z_s\setminus\{0\}$, form a desired matrix extension.

THEOREM 2.4. *The trigonometric polynomials m_j, $j\in Z_s\setminus\{0\}$, defined by* (2.8) *are high-pass filters corresponding to the low-pass filter m_0 defined by* (2.7).

Proof. It is clear that m_j, $j\in Z_s\setminus\{0\}$, are high-pass filters because

$$m_j(0)=\frac{1}{s}V_0^TV_j=0.$$

It remains to prove that the polyphase matrix W formed from m_j, $j\in Z_s$, satisfies

$$W(\xi)W^*(\xi)=\frac{1}{s}I_{rs},\quad \xi\in R^d.$$

It follows from (2.7) and (2.8) that the polyphase matrix is of the form

$$W^T(\xi)=\frac{1}{s}\prod_{j\in Z_N}\left(U_jD\left(A^T\xi\right)\right)\left(\prod_{j\in Z_N}U_{N-1-j}^T\right)\left(V_0,V_1,\ldots,V_{s-1}\right).$$

Since all the matrices V, U_j, $j\in Z_{N+1}\setminus\{0\}$, and $D(\xi)$ are unitary, we conclude that the matrix $\sqrt{s}W(\xi)$, $\xi\in R^d$, is unitary as well. ∎

To close this section we present an alternative form of the low-pass filter m_0 and the high-pass filters m_j, $j\in Z_s\setminus\{0\}$, defined, respectively, in (2.7) and (2.8).

THEOREM 2.5. *The filters defined by* (2.7) *and* (2.8) *can be expressed in the alternative form*

$$m_j(\xi)=\frac{1}{s}X(\xi)\prod_{k\in Z_N}\left(\tilde{U}_kD\left(A^T\xi\right)\tilde{U}_k^T\right)V_j,\quad j\in Z_s, \tag{2.9}$$

for some $rs\times rs$ unitary matrices $\tilde{U}_k$, $k\in Z_N$.

Proof. Suppose that filters m_j, $j\in Z_s$, have the forms (2.7) and (2.8). We define matrices $\tilde{U}_k$, $k\in Z_N$, by setting

$$\tilde{U}_k:=\prod_{j\in Z_k}U_j.$$

Clearly, the matrices $\tilde{U}_k$, $k\in Z_N$, are unitary as well. This shows that (2.7) and (2.8) can be written in (2.9).

Conversely, suppose that m_j, $j\in Z_s$, have the form of (2.9). We define $U_0=I$ and for $k\in Z_N$ set

$$U_k:=\tilde{U}_{k-1}^T\tilde{U}_k.$$

Noting that

$$\prod_{j\in Z_N}U_{N-1-j}^T=\tilde{U}_N^T$$

and observing that U_j, $j\in Z_N$, are orthogonal matrices, we prove the theorem. ∎

3. The block central symmetric matrix and uniform linear phase. This section focuses on the construction of low-pass filters having a uniform linear phase. We say that the low-pass filter m_0 has a uniform linear phase if there exists a $\mu \in Z^d$ such that for all $\xi \in R^d$

$$\overline{m_0(\xi)} = e^{i\mu\cdot\xi} m_0(\xi).$$

In signal processing, having a linear phase is a central property of filters [5]. Since in this case if the input signal has energy confined to the pass-band of the filter, then the output signal is approximately equal to this input. It is well known that in the univariate case the only two-channel CMF and FIR bank with a linear phase is the Haar filter. In the multivariate, multiple channel, vector-valued case, the situation is very different, and examples of linear phase filter are given in [2], [11], [17].

In this section, we discuss the construction of linear phase filter banks whose polyphase matrix has the matrix factorization (2.7). To this end, we introduce the notion of *block central symmetric* matrices. Let H be the $rs \times rs$ matrix

$$H := \begin{pmatrix} 0 & 0 & \cdots & I_r \\ 0 & \cdots & I_r & 0 \\ \vdots & \vdots & \vdots & \vdots \\ I_r & 0 & \cdots & 0 \end{pmatrix}.$$

Obviously H is a real symmetric orthogonal matrix. For any $rs \times rs$ matrix B, we define

$$B^H := HBH$$

and observe for any two $rs \times rs$ matrices B and C that

$$(B^H)^T = (B^T)^H$$

and

$$(BC)^H = B^H C^H.$$

DEFINITION. *An $rs \times rs$ real matrix U is called $r \times r$ block central symmetric if*

$$U = U^H.$$

The next theorem shows the importance of this notion for the construction of uniform linear phase filters.

THEOREM 3.1. *Suppose that m_0 is the low-pass filter defined in* (2.7) *with*

$$V_0 = (I_r, I_r, \ldots, I_r)^T.$$

If U_j, $j \in Z_N$, are $r \times r$ block central symmetric orthogonal matrices and

$$\gamma_{s-1} - \gamma_j = \gamma_{s-1-j}, \quad j \in Z_s, \tag{3.1}$$

then m_0 has a uniform linear phase.

Proof. By Theorems 2.1 and 2.2, we know that m_0 is a low-pass filter satisfying the perfect reconstruction condition. We only need to verify that m_0 has linear phase; that is, we must find a vector $\mu \in Z^d$ such that for all $\xi \in R^d$

$$\overline{m_0(\xi)} = e^{i\mu\cdot\xi} m_0(\xi).$$

Our choice for μ is that

$$\mu := (NA + I)\gamma_{s-1}.$$

Let us confirm that this is a correct choice. By (2.7) and (2.9), we have that

$$\overline{m_0(\xi)} = \frac{1}{s} X(-\xi) \prod_{j \in Z_N} \left(U_j D\left(-A^T \xi\right) U_j^T\right) V_0,$$

while our hypothesis leads us to conclude that

$$e^{-i\gamma_{s-1}\cdot A^T \xi} D\left(-A^T \xi\right) = H D\left(A^T \xi\right) H$$

and

$$e^{-i\gamma_{s-1}\cdot \xi} X(-\xi) = X(\xi) H.$$

Combining these equations, we get that

$$\overline{m_0(\xi)} = \frac{1}{s} e^{i\mu\cdot\xi} X(\xi) H \prod_{j \in Z_N} \left(U_j H D\left(A^T \xi\right) H U_j^T\right) V_0,$$

from which it follows that

$$\overline{m_0(\xi)} = \frac{1}{s} e^{i\mu\cdot\xi} X(\xi) \prod_{j \in Z_N} \left(H U_j H D\left(A^T \xi\right) H U_j^T H\right) H V_0.$$

Using our hypothesis about the matrices appearing in this product, the result follows. □

To make use of this result we must confirm that the coset representers $\{\gamma_i : i \in Z_s\}$ have the property (3.1). We demonstrate next that this can always be achieved.

LEMMA 3.2. *For any dilation matrix A, there exists a complete set $\{\gamma_j : j \in Z_s\}$ of representatives of the coset of Z^d/AZ^d with $\gamma_0 = 0$ satisfying* (3.1).

To prove this lemma, we need to recall two basic results from group theory. The first result is the Lagrange theorem, which states that the cardinality of any subgroup of a finite group G is a divisor of the cardinality of G. The second result we shall use is the Sylow theorem, which states that if p is a prime number and k is a nonnegative integer such that p^k divides the cardinality of a finite group G, then G contains a subgroup of cardinality p^k; cf. [8].

To facilitate the proof of this lemma, we review some facts of the quotient group Z^d/AZ^d. Recall that any $a \in Z^d$ determines a subset $\overline{a}$ of Z^d given by the formula

$$\overline{a} := AZ^d + a.$$

Note that $\overline{0} = AZ^d$ is also a subgroup of Z^d. Using $\overline{0}$, we can obtain quotient group Z^d/AZ^d, which offers a partition of Z^d,

$$Z^d = \bigcup_{j \in Z_s} \overline{a_j} \quad \text{with} \quad \overline{a_j} \cap \overline{a_k} = \emptyset, \quad j, k \in Z_s, \quad \text{with} \ \ j \neq k,$$

where

$$\Omega := \{a_j : j \in Z_s\}$$

is a complete representative set of the distinct coset of Z^d/AZ^d. A pair a_j and $a_{j'}$ of vectors in Ω is said to be dual relative to $a \in Z^d$ if

$$\overline{a_j} = \overline{a - a_{j'}}.$$

For any elements $\overline{a}, \overline{b} \in Z^d/AZ^d$, the sum of $\overline{a}$ and $\overline{b}$ is defined by

$$\overline{a} + \overline{b} := \overline{a + b}.$$

In particular, we denote

$$2\overline{a} := \overline{a} + \overline{a},$$

and hence

$$2\overline{a} = \overline{2a}.$$

The order of the element $\overline{a} \in Z^d/AZ^d$ is defined as the minimum positive integer n such that $\overline{na} = \overline{0}$.

Proof of Lemma 3.2. Starting from any complete set of representatives of the distinct coset of the quotient group Z^d/AZ^d, we present an algorithm to construct a complete set of representatives of the distinct coset of the quotient group Z^d/AZ^d satisfying (3.1). We consider two cases according to the cardinality s of the group Z^d/AZ^d.

First, we consider the case where s is an odd integer. In this case, $s = 2k+1$, for some positive integer k. We choose any complete set Ω with $a_0 = 0$ of representatives of the distinct coset of Z^d/AZ^d such that

$$Z^d/AZ^d = \{\overline{a_j} : j \in Z_{2k+1}\}.$$

It defines a sequence of length $2k+1$ with a fixed order

$$a_0, a_1, a_2, \ldots, a_{2k-1}, a_{2k}. \tag{3.2}$$

Since the cardinality of Z^d/AZ^d is odd, by the Lagrange theorem, there is no element of order 2 in Z^d/AZ^d. Otherwise, Z^d/AZ^d will contain a cyclic subgroup of cardinality 2, which contradicts the Lagrange theorem. We next prove that

$$2\overline{a_j} \neq 2\overline{a_n} \quad \text{for} \quad j, n \in Z_{2k+1}, \quad \text{with} \quad j \neq n.$$

Assume to the contrary that $2\overline{a_j} = 2\overline{a_n}$ for some $j, n \in Z_{2k+1}$ with $j \neq n$. Hence, $2(a_j - a_n) \in \overline{0}$. In other words, $2\overline{a_j - a_n} = \overline{0}$. This implies that $\overline{a_j - a_n}$ is an element of order 2 and contradicts the nonexistence of elements of order 2 in Z^d/AZ^d. Therefore, we conclude that

$$Z^d/AZ^d = \{2\overline{a_j} : j \in Z_{2k+1}\}.$$

This ensures that there exists a unique positive integer $l \in Z_{2k+1} \setminus \{0, 2k\}$ such that $\overline{a_{2k}} = 2\overline{a_l}$. Hence, in sequence (3.2) replace $\overline{a_{2k}}$ by $2\overline{a_l}$ and obtain a new sequence

$$a_0, a_1, \ldots, a_{2k-1}, 2a_l, \tag{3.3}$$

which satisfies the condition

$$Z^d/AZ^d = \{\overline{a_0}, \overline{a_1}, \ldots, \overline{a_{2k-1}}, 2\overline{a_l}\}.$$

Suppose $l < k$. We interchange $\overline{a_l}$ with $\overline{a_k}$ in sequence (3.3) and obtain a new sequence

$$a_0, \ldots, a_{l-1}, a_k, a_{l+1}, \ldots, a_{k-1}, a_l, a_{k+1}, \ldots, a_{2k-1}, 2a_l. \tag{3.4}$$

Noting that for any $j, j' \in Z_{2k+1}$ with $j \neq j'$, $\overline{2a_l - a_j}$ and $\overline{2a_l - a_{j'}}$ are distinct, we have that

$$Z^d/AZ^d = \left\{\overline{2a_l - a_j} : j \in Z_{2k+1}\right\}.$$

Thus, for any $j \in Z_{2k+1}$ there exists a unique $j' \in Z_{2k+1}$ such that

$$\overline{a_j} = \overline{2a_l - a_{j'}}.$$

It follows that there are $k+1$ dual pairs in sequence (3.4) relative to $2a_l$, including the pair a_0 and $2a_l$, as well as the pair a_l and a_l, which is the only dual pair that is self-dual. We reorder the sequence (3.4) to form the new sequence

$$b_0, b_1, \ldots, b_{2k}$$

with $b_0 = a_0$, $b_{2k} = 2a_l$, and $b_k = a_l$ such that

$$\overline{b_j} = \overline{2a_l - b_{2k-j}}, \quad j \in Z_{2k+1}.$$

Now, we define

$$\begin{gathered}\gamma_0 := b_0, \quad \gamma_j := 2a_l - b_{2k-j}, \quad j = 1, 2, \ldots, k-1, \\ \text{and} \quad \gamma_j := b_j, \quad j = k, k+1, \ldots, 2k.\end{gathered} \tag{3.5}$$

Then we have that

$$Z^d/AZ^d = \{\gamma_j : j \in Z_{2k+1}\}.$$

Hence, the set $\{\gamma_j : j \in Z_{2k+1}\}$ defined by (3.5) is the desired complete set of representatives of the distinct coset of the quotient group Z^d/AZ^d. The cases $l = k$ and $l > k$ can be similarly handled.

Next, we consider the case where s is an even integer. In this case, $s = 2k$ for some positive integer k. We choose a complete set Ω with $a_0 = 0$ of representatives of distinct cosets of Z^d/AZ^d which satisfies

$$Z^d/AZ^d = \{\overline{a_j} : j \in Z_{2k}\}.$$

By the Sylow theorem, in $\{\overline{a_1}, \ldots, \overline{a_{2k-1}}\}$ there is an element of order 2^n for some positive integer n. Without loss of generality, we assume that $\overline{a_{2k-1}}$ is the element having the highest order 2^{n_0}. We then have that

$$\overline{a_{2k-1}} \neq 2\overline{a_j} \quad \text{for} \quad j \in Z_{2k}.$$

Otherwise, there exists an element of order 2^{n_0+1}, contradicting the fact that $\overline{a_{2k-1}}$ is the element of the highest order 2^{n_0} in $\{\overline{a_j} : j \in Z_{2k}\}$. Hence, for any $j, j' \in Z_{2k}$ with $j \neq j'$, $\overline{a_{2k-1} - a_j}$ and $\overline{a_{2k-1} - a_{j'}}$ are distinct. Consequently,

$$Z^d/AZ^d = \{\overline{a_{2k-1} - a_j} : j \in Z_{2k}\}.$$

It follows that for $j \in Z_{2k} \setminus \{0, 2k-1\}$, there exists a unique $j' \in Z_{2k} \setminus \{0, 2k-1, j\}$ such that

$$\overline{a_j} = \overline{a_{2k-1} - a_{j'}}.$$

That is, in the sequence

$$a_0, a_1, \ldots, a_{2k-1} \tag{3.6}$$

there are k dual pairs relative to a_{2k-1}, including the pair a_0 and a_{2k-1}. We reorder the sequence (3.6) and obtain the new sequence

$$b_0, b_1, \ldots, b_{2k-1}$$

with $b_0 = a_0$ and $b_{2k-1} = a_{2k-1}$ such that

$$\overline{b_j} = \overline{b_{2k-1} - b_{2k-1-j}}, \quad j \in Z_{2k}.$$

Define

$$\gamma_j := b_j, \quad j \in Z_k, \quad \gamma_j := b_{2k-1} - b_{2k-1-j}, \quad j = k, k+1, \ldots, 2k-2, \quad \gamma_{2k-1} := b_{2k-1}.$$

This is the desired complete set of representatives of the distinct coset of the group Z^d/AZ^d for the case in which s is even. □

4. Examples. In this section, we list several useful examples of the general constructions presented in the previous sections.

Example 4.1. The case which is most important for application to image processing corresponds to $d = 2$ and the dilation matrix $A = 2I$. Therefore, $s = 4$ and a complete set of coset representers are $\gamma_0 = (0,0)$, $\gamma_1 = (0,1)$, $\gamma_2 = (1,0)$, $\gamma_3 = (1,1)$. For the scalar case, that is, $r = 1$, and also $N = 1$, a convenient choice for a center symmetric orthogonal matrix is

$$U_0 := \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix},$$

and we choose $V = (1,1,1,1)^T$. With this choice, a filter bank with linear phase is determined.

In the case that $r = N = 2$, we can choose the following 2×2 block central symmetric orthogonal matrices of order 8:

$$U_0 := \frac{1}{2\sqrt{2}}\begin{pmatrix} B & C & -B & C \\ -B & C & -B & -C \\ -C & -B & C & -B \\ C & -B & C & B \end{pmatrix},$$

where

$$C^T = B := \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$$

and $U_2 = I$, and we choose the matrix $V := (I, I, I, I)^T$ to generate the low-pass filter having linear phase.

Example 4.2. In this example, we choose the quincunx dilation matrix

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

Thus, we have that $d = s = 2$ and a complete set of representatives of the coset is $\gamma_0 = (0,0)$ and $\gamma_1 = (1,1)$. We can use the four unitary matrices of order 2,

$$U_0 := \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix},$$

$$U_1 := \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix},$$

$$U_j := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad j = 2, \ldots, n+1,$$

and

$$U_{n+2} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad j = n+2, \ldots, 2n+1,$$

to generate low-pass filters.

Example 4.3. Our last example is the univariate $d = 1$, $s = 3$, three-band channel case. A matrix U of order 3 is a central symmetric orthogonal matrix if and only if U has the form

$$\begin{pmatrix} \frac{\cos\beta \pm 1}{2} & \frac{\sqrt{2}}{2}\sin\beta & \frac{\cos\beta \mp 1}{2} \\ \frac{\sqrt{2}}{2}\cos\alpha & \sin\alpha & \frac{\sqrt{2}}{2}\cos\alpha \\ \frac{\cos\beta \mp 1}{2} & \frac{\sqrt{2}}{2}\sin\beta & \frac{\cos\beta \pm 1}{2} \end{pmatrix}$$

with $\alpha - \beta = k\pi + \frac{\pi}{2}$. For example, we can choose the central symmetric orthogonal matrix

$$\frac{1}{4}\begin{pmatrix} \sqrt{2}+2 & -2 & \sqrt{2}-2 \\ 2 & 2\sqrt{2} & 2 \\ \sqrt{2}-2 & -2 & \sqrt{2}+2 \end{pmatrix}$$

with $V_0 = (1,1,1)^T$ to get the low-pass filter. To construct the corresponding two high-pass filters, we extend the unit vector $\frac{1}{\sqrt{3}}V_0$ to a orthonormal basis for the space R^3 with the vectors

$$V_1 = -\frac{\sqrt{2}}{2}(1,-2,1)^T$$

and

$$V_2 = \frac{\sqrt{6}}{2}(1,0,-1)^T.$$

The filter banks of three channels with a linear phase are easier to design with the following three orthogonal matrices of order 3:

$$\begin{pmatrix} \cos\theta_1 & -\sin\theta_1 & 0 \\ \sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} \cos\theta_2 & 0 & -\sin\theta_2 \\ 0 & 1 & 0 \\ \sin\theta_2 & 0 & \cos\theta_2 \end{pmatrix},$$

and

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & -\sin\theta_3 \\ 0 & \sin\theta_3 & \cos\theta_3 \end{pmatrix}.$$

Acknowledgment. The authors would like to thank Dr. Ziqun Lu of Beijing University for a helpful discussion on the proof of Lemma 3.2.

REFERENCES

[1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[2] A. Cohen and I. Daubechies, *Non-separable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[3] A. Croisier, D. Esteban, and C. Galand, *Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques*, in Proceedings of the International Conference on Information Sciences and Systems, Patras, Greece, 1976, pp. 443–446.

[4] W. Dahmen and C. A. Micchelli, *Biorthogonal wavelet expansions*, Constr. Approx., 13 (1997), pp. 293–328.

[5] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[6] C. Heil and D. Colella, *Matrix refinement equation: Existence and uniqueness*, J. Fourier Anal. Appl., 2 (1996), pp. 363–377.

[7] O. Herrmann, *On the approximation problem in nonrecursive digital filter design*, IEEE Trans. Circuit Theory, 18 (1971), pp. 411–413.

[8] N. Jacobson, *Basic Algebra*, W. H. Freeman, New York, 1984.

[9] R. Q. Jia, *Approximation properties of multivariate wavelets*, Math. Comp., 67 (1998), pp. 647–665.

[10] R. Q. Jia and C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets* V: *Extensibility of trigonometric polynomials*, Computing, 48 (1992), pp. 61–72.

[11] J. Kovacevic and M. Vetterli, *Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for R^n*, IEEE Trans. Inform. Theory, 38 (1992), pp. 533–555.

[12] C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets*, Numer. Algorithms, 1 (1991), pp. 75–116.

[13] C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets* VI: *Shift invariant subspaces*, in Approximation Theory, Spline Functions and Applications (Maratea, 1991), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 356, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992, pp. 213–222.

[14] C. A. Micchelli and T. Sauer, *Regularity of multiwavelets*, Adv. Comput. Math., 7 (1997), pp. 455–545.

[15] C. A. Micchelli and Y. Xu, *Using the matrix refinement equation for the construction of wavelets on invariant sets*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 391–401.

[16] C. A. Micchelli and Y. Xu, *Reconstruction and decomposition algorithms for biorthogonal multiwavelets*, Multidimens. Systems Signal Process., 8 (1997), pp. 31–69.

[17] T. Q. Nguyen and P. P. Vaidyanathan, *Two-channel perfect reconstruction FIR QMF structure which yield linear phase FIR analysis and synthesis filters*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 676–690.

[18] D. Pollen, $SU_I(2, F[z, 1/z])$ *for* F *a subfield of* $\mathbf{C}$, J. Amer. Math. Soc., 3 (1990), pp. 611–624.
[19] Q. Sun, N. Bi, and D. Huang, *An Introduction to Multiband Wavelets*, Zhejiang University Press, Hangzhou, 2001 (in Chinese).
[20] M. J. Smith and T. P. Barnwell, *A procedure for designing exact reconstruction filter banks for tree structured sub-band coders*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, San Diego, CA, 1984, pp. 27.1.1–27.1.4.
[21] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[22] P. P. Vaidyanathan, *Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques*, IEEE ASSP Mag., 4 (3) (1987), pp. 4–20.
[23] M. Vetterli, *Filter banks allowing perfect reconstruction*, Signal Process., 10 (1986), pp. 219–244.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 532–550

AN ALGORITHM FOR THE CONSTRUCTION OF SYMMETRIC ORTHOGONAL MULTIWAVELETS*

RADKA TURCAJOVÁ†

Abstract. In this paper, a parameterization is developed of orthogonal multiwavelets that have all scaling and wavelet functions symmetric or antisymmetric about some given point. The parameterization is based on the factorization of the polyphase matrix into the product of an orthogonal matrix and paraunitary linear factors based on complementary orthogonal projectors. The symmetry of scaling and wavelet functions is reflected by the polyphase matrix. It can be enforced by using factors that themselves conform to certain symmetry constraints. Such symmetric factors can be built from smaller orthogonal matrices, which can be parameterized by standard methods. An example is included that uses the proposed parameterization for the construction of symmetric differentiable compactly supported multiwavelets. The factorization presented in this paper can be used also for finding symmetric orthogonal wavelets for an existing set of compactly supported symmetric orthogonal scaling functions.

Key words. orthogonal multiwavelets, symmetry, polyphase matrix, factorization

AMS subject classifications. 15A23, 42C40

DOI. 10.1137/S0895479802411043

1. Introduction.

Wavelet bases have become a very popular tool in various areas of science and engineering. One of their advantages over the traditional methods is the possibility to custom tailor a wavelet basis to a given application. This flexibility, however, has its limits. The classical setting, an orthogonal multiresolution analysis generated by a single scaling and single wavelet function by the means of dilations by the factor of 2 and translations, as developed in [10], may be sometimes too restricting, and desired properties of the basis functions may not be possible to achieve. A typical example of this is the combination of a compact support and the symmetry of the basis functions. Daubechies, when she constructed her famous orthogonal wavelets, proved that the only compactly supported orthogonal wavelet basis consisting of symmetric and antisymmetric functions is the trivial Haar basis; see [2] or [3]. There are various solutions to this problem. One of them is relaxing one of the constraints and choosing wavelets that are orthogonal and compactly supported but with only approximate symmetry; wavelets that are compactly supported, symmetric, but not quite orthogonal; or wavelets that are orthogonal and symmetric but instead of having compact support merely decay fast enough. Daubechies discusses various families of such wavelets in her book [3], and there are also various other sources describing such constructions; see, for example, [1]. Another possible solution to this problem is to use some generalization of the classical multiresolution analysis scheme that is not so restrictive. It is possible, for example, to choose a dilation factor $m > 2$, which results in a scheme with one scaling and $m - 1$ (i.e., more than one) wavelet functions [14]. A more recently explored alternative is multiwavelets, where simultaneously also more than one scaling function is used. A famous example of nontrivial orthogonal, symmetric, compactly supported multiwavelets is due to Geronimo, Hardin, and Massopust [5]. Another construction is due to Strela [13].

*Received by the editors July 16, 2002; accepted for publication by A. H. Sayed March 7, 2003; published electronically October 2, 2003.
http://www.siam.org/journals/simax/25-2/41104.html
†Department of Mathematics, University of St. Thomas, Mail #OSS201, 2115 Summit Ave., St. Paul, MN 55105-1079 (rturcajova@stthomas.edu).

The algorithm presented in [14] for the construction of compactly supported symmetric orthogonal wavelets with dilation factor m larger than 2 is based on the factorization of the so-called *polyphase matrix*, an $m \times m$ matrix trigonometric polynomial constructed from the coefficients of the refinement masks of the scaling and wavelet functions. When all the generating functions are symmetric, the polyphase matrix also displays certain symmetries, and, for some types of symmetry, such symmetric polyphase matrices can be built from symmetric linear factors. The factorization into the product of linear factors based on complementary orthogonal projectors used in [14] is closely related to the products suggested in [12] and was first described (without symmetry) in [8]. It was developed as a generalization of Pollen's factorization of classical orthogonal wavelets [11] and is also closely related to factorizations of paraunitary filter banks presented in [18] and [19]. In comparison with the last mentioned factorizations, however, ours yields a less redundant parameterization (with a smaller number of free parameters) and also a precise control over the order of the polyphase matrix and, consequently, the length of support of scaling and wavelet functions.

In this paper, our aim is to create an algorithm for the construction of symmetric orthogonal compactly supported multiwavelets. In particular, we want to derive a parameterization of symmetric orthogonal multiwavelets similar to the one for the wavelets with the dilation factor larger than 2 given in [14]. It was shown in [14] that symmetry constraints can be accommodated only if the polyphase matrix is larger than 2×2. Otherwise only trivial symmetric factors exist, and the result is the Haar wavelet. Since, for multiwavelets, the polyphase matrix is also larger than 2×2, it also leaves room for additional symmetry constraints. The symmetry pattern of refinement masks of symmetric multiwavelets is, however, different than in the situation when only one scaling function, with dilation factor $m > 2$, is considered. We thus first need to study the refinement masks of symmetric multiwavelets and describe precisely the symmetry of polyphase matrices associated with them. Then we need to come up with a new type of symmetric linear factor that will generate this pattern. The idea is similar to that of [14], but we need to deal here with completely different symmetry constraints.

The paper is organized as follows. First, in section 2, we review some known results we are going to build upon, and we also introduce some terms and notation used throughout the paper. Then, in section 3, we study the symmetry constraints that are imposed upon the polyphase matrix by the requirement that all the generating functions be either symmetric or antisymmetric about some point (common to all the functions). We also prove there that, in such a situation, the numbers of scaling and wavelet functions that are symmetric and antisymmetric cannot be arbitrary, but exactly half of the functions must be symmetric and the other half antisymmetric. In the following section, section 4, we give some motivation for later results and describe the basic building blocks we are going to use, namely, linear factors based on complementary orthogonal projectors and orthogonal matrices showing certain symmetry patterns. These factors and their parameterization are then discussed in more detail in the next two sections. In section 7 we prove that every polyphase matrix that yields orthogonal multiwavelets symmetric or antisymmetric about a given point can be built from these basic building blocks, i.e., that every polyphase matrix associated with such multiwavelets can be factored into the product of these symmetric factors. Finally, in the last section of this paper, we give an example of using the parameterization for the construction of symmetric compactly supported orthogonal multiwavelets. We construct there a new orthogonal multiwavelet that has all the scaling and wavelet functions symmetric or antisymmetric, continuously differentiable,

and supported on the interval $[0, 3]$.

2. Preliminaries and notation. Let us consider an orthogonal multiresolution analysis and an associated orthogonal multiwavelet basis based on the dilation factor 2, r scaling functions φ_j, $j = 1, \dots, r$, and r wavelet functions, ψ_j, $j = 1, \dots, r$. Let $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ denote vectors consisting of all scaling and all wavelet functions, respectively,

$$\boldsymbol{\varphi} = (\, \varphi_1 \quad \varphi_2 \quad \cdots \quad \varphi_r \,)^T, \qquad \boldsymbol{\psi} = (\, \psi_1 \quad \psi_2 \quad \cdots \quad \psi_r \,)^T.$$

The *refinement equations*, two-scale relationships that the scaling and wavelet functions satisfy, can be written in vector form as

$$\boldsymbol{\varphi}(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} \boldsymbol{H}_k \boldsymbol{\varphi}(2x - k), \qquad \boldsymbol{\psi}(x) = \sqrt{2} \sum_{k \in \mathbb{Z}} \boldsymbol{G}_k \boldsymbol{\varphi}(2x - k). \tag{2.1}$$

The coefficients $\boldsymbol{H}_k$ and $\boldsymbol{G}_k$ in these equations are some $r \times r$ matrices. What we are going to do is to parameterize matrices $\boldsymbol{H}_k$ and $\boldsymbol{G}_k$ that satisfy certain necessary conditions. More precisely, we will parameterize a *polyphase matrix* $\boldsymbol{A}(\omega)$, from which the *refinement masks* (the coefficient sequences $\{\boldsymbol{H}_k\}_{k \in \mathbb{Z}}$ and $\{\boldsymbol{G}_k\}_{k \in \mathbb{Z}}$) can be later extracted because it is defined as

$$\boldsymbol{A}(\omega) = \sum_{k \in \mathbb{Z}} \begin{pmatrix} \boldsymbol{H}_{2k} & \boldsymbol{H}_{2k+1} \\ \boldsymbol{G}_{2k} & \boldsymbol{G}_{2k+1} \end{pmatrix} e^{-ik\omega}. \tag{2.2}$$

The scaling and wavelet functions are going to be defined as the solution of the refinement equations. The necessary conditions that we are going to work with on their own do not guarantee the existence of an L^2 solution. However, the sufficient conditions for its existence are known. Existence of the solution and its regularity is governed by the eigenstructure of the so-called *transition operator*, which is constructed from the refinement mask $\{\boldsymbol{H}_k\}_{k \in \mathbb{Z}}$ [7, 17]. It is a generalization of the results for classical wavelets due to Lawton [9] and Eirola [4], which were also successfully generalized to the case of the dilation parameter larger than 2 [6]. These conditions are too complicated to be included directly in the parameterization. However, the lower bound for regularity derived from the eigenvalue structure of the associated transition operator described in [7] can be used as a cost function in numerical optimization exploiting the parameterization. This approach to constructing regular multiwavelets was successfully used in [15] and will be applied also in the example in the last section of this paper. A good parameterization incorporating as many of the desired properties as possible is crucial for this method to work. It is essential to keep the number of free variables small. The parameterization not only reduces the number of free variables but also eliminates constraints (some of them nonlinear) that would otherwise have to be imposed during the optimization.

We are going to construct multiwavelets that are compactly supported. We will achieve this by restricting our attention to refinement masks having only a finite number of coefficients nonzero. We will concentrate on the size of the refinement masks rather than the length of the support of generating functions. It is more convenient for us, but also practical. The discrete wavelet transform utilizes the coefficients of the refinement masks rather than functions themselves. The refinement masks need to be short for the transform to be efficient, and therefore it is highly desirable that only a relatively small number of coefficients $\boldsymbol{H}_k$ and $\boldsymbol{G}_k$ are nonzero. The proposed algorithm allows controlling the number of nonzero coefficients (and consequently also the length of support of the scaling and wavelet functions) by choosing the number

of the nonzero coefficients of the polyphase matrix. Throughout this paper, we thus will assume that only finitely many of the coefficients $\boldsymbol{A}_k$ are nonzero; that is, the polyphase matrix $\boldsymbol{A}(\omega)$ is a matrix trigonometric polynomial. More precisely, we will assume that there is some number $q \geq 0$ such that

$$\boldsymbol{A}(\omega) = \boldsymbol{A}_0 + \boldsymbol{A}_1 e^{-i\omega} + \cdots + \boldsymbol{A}_q e^{-iq\omega}$$

(i.e., $\boldsymbol{A}_k = \boldsymbol{0}$ for all $k < 0$ and $k > q$) and that both $\boldsymbol{A}_0$ and $\boldsymbol{A}_q$ are nonzero. Provided the number of nonzero coefficients of the refinement masks is finite, this can always be achieved by multiplying the polyphase matrix by a suitable factor $e^{-is\omega}$, $s \in \mathbb{Z}$. It is equivalent to simply shifting all the generating functions by $2s$, which is only a cosmetic change—it does not alter the spaces at all and causes only a shift in the wavelet coefficients. We will call the number q the *order* of the polyphase matrix $\boldsymbol{A}(\omega)$.

There are two important sets of necessary conditions that the refinement masks have to satisfy—one relates to the orthogonality properties of the generating functions and the other to the requirement that constants can be reproduced exactly (approximation order 0), which is necessary for the existence of multiresolution analysis, and the consequent requirement that the integrals of scaling functions cannot be simultaneously all zero.

If L^2 functions satisfying (2.1) exist, their integer translates are orthonormal if and only if

$$\sum_{k\in\mathbb{Z}} \boldsymbol{H}_k \boldsymbol{H}_{k+2l}^T = \sum_{k\in\mathbb{Z}} \boldsymbol{G}_k \boldsymbol{G}_{k+2l}^T = \delta_{l0} \boldsymbol{I},$$

$$\sum_{k\in\mathbb{Z}} \boldsymbol{H}_k \boldsymbol{G}_{k+2l}^T = \sum_{k\in\mathbb{Z}} \boldsymbol{G}_k \boldsymbol{H}_{k+2l}^T = \boldsymbol{0}.$$

Here, δ_{lj} is the Kronecker delta (equal to 1 if $l = j$ and to 0 otherwise). This set of equations is equivalent to the equation

$$\sum_{k\in\mathbb{Z}} \boldsymbol{A}_k \boldsymbol{A}_{k+l}^T = \delta_{l0} \boldsymbol{I} \tag{2.3}$$

and, consequently, to the polyphase matrix being *paraunitary*, i.e., unitary for every ω,

$$\boldsymbol{A}(\omega)\boldsymbol{A}(\omega)^* = \boldsymbol{I}. \tag{2.4}$$

Let us now have a look at the other condition we need to incorporate. By integrating both sides of the refinement equation satisfied by the scaling function, one can show that the nonzero vector $\boldsymbol{w} = \int_{\mathbb{R}} \boldsymbol{\varphi}(x)dx$ is an eigenvector of the matrix $\sum_{k\in\mathbb{Z}} \boldsymbol{H}_k$ corresponding to an eigenvalue $\sqrt{2}$. The following lemma reformulates this property of the refinement mask of the scaling function as a restriction on the polyphase matrix.

LEMMA 2.1.

$$\boldsymbol{A}(0) \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w} \end{pmatrix} = \sqrt{2} \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{0} \end{pmatrix}. \tag{2.5}$$

Proof. Since $\boldsymbol{w}$ is an eigenvector of the matrix $\sum_{k\in\mathbb{Z}} \boldsymbol{H}_k$ corresponding to an eigenvalue $\sqrt{2}$, we have

$$\boldsymbol{A}(0) \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{w} \end{pmatrix} = \sqrt{2} \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{x} \end{pmatrix}$$

for some vector $\boldsymbol{x}$. Nevertheless, since $\boldsymbol{A}(0)$ is unitary, we have

$$2\,(\|\boldsymbol{w}\|^2+\|\boldsymbol{x}\|^2) = 2\,\begin{pmatrix}\boldsymbol{w}^T & \boldsymbol{x}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{x}\end{pmatrix} = \begin{pmatrix}\boldsymbol{w}^T & \boldsymbol{w}^T\end{pmatrix}\boldsymbol{A}(0)^T\boldsymbol{A}(0)\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{w}\end{pmatrix}$$
$$= \begin{pmatrix}\boldsymbol{w}^T & \boldsymbol{w}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{w}\end{pmatrix} = 2\,\|\boldsymbol{w}\|^2.$$

Consequently, $\|\boldsymbol{x}\|^2 = 0$ and hence $\boldsymbol{x} = \boldsymbol{0}$. □

Finally, let us recall here also the factorization of paraunitary matrix polynomials into a constant factor and normalized linear factors that we want to adapt to the symmetric case.

THEOREM 2.2. *Let $\boldsymbol{A}(\omega)$ be a polyphase matrix of order q. Then $\boldsymbol{A}(\omega)$ is paraunitary if and only if there exists an orthogonal matrix $\boldsymbol{Q}$ and orthogonal projectors $\boldsymbol{P}_j$ (i.e., $\boldsymbol{P}_j\boldsymbol{P}_j = \boldsymbol{P}_j$, $\boldsymbol{P}_j^T = \boldsymbol{P}_j$), $j = 1,\dots,q$, such that*

$$\boldsymbol{A}(\omega) = \boldsymbol{Q}\ (\boldsymbol{I}-\boldsymbol{P}_1+\boldsymbol{P}_1e^{-i\omega})\cdots(\boldsymbol{I}-\boldsymbol{P}_q+\boldsymbol{P}_qe^{-i\omega}). \tag{2.6}$$

Note that $\boldsymbol{Q} = \boldsymbol{A}(0)$. The condition (2.5) thus represents merely a simple restriction onto one of the factors and therefore can be easily incorporated into the parameterization based on this factorization.

The proof of this theorem can be found in [8, 16]. In section 7 we are going to present a symmetric form of this theorem with a complete proof. The proof is constructive. It gives a practical algorithm for factoring the polyphase matrix. This algorithm can be applied also to rectangular matrices. Besides creating a parameterization of an entire polyphase matrix, as we show it in this paper, the factorization thus may be used also for solving a completion problem—finding the rest of the generating functions when some of the functions are given; for example, for finding wavelets for a given set of scaling functions. This may be required in the case of techniques like the one presented in [13], where the refinement mask for the scaling functions is constructed by the means of the two-scale similarity transform, and the refinement mask for wavelets has to be found afterwards, separately. In the case of classical wavelets, this problem has a unique (up to a multiple by -1 and a shift) trivial solution. In cases when the polyphase matrix is larger than just 2×2, that is, in the case of wavelets with the dilation factor $m > 2$, and also in the case of multiwavelets, a variety of different solutions exists and solving this problem is nontrivial. Factoring the partial polyphase matrix and completing the matrix $\boldsymbol{Q}$ to a square matrix gives an answer. In fact, it allows searching all possible solutions not exceeding the given length.

3. Symmetry. Let us consider the situation when all the generating functions are symmetric or antisymmetric about the same point. More precisely, we will request that there exists some odd integer n such that, for all $j = 1,\dots r$,

$$\varphi_j(x) = \pm\varphi_j(n-x), \qquad \psi_j(x) = \pm\psi_j(n-x).$$

In vector form, we can write this as

$$\boldsymbol{\varphi}(x) = \boldsymbol{\Sigma}\,\boldsymbol{\varphi}(n-x), \qquad \boldsymbol{\psi}(x) = \boldsymbol{\Lambda}\,\boldsymbol{\psi}(n-x), \tag{3.1}$$

where $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ are some diagonal matrices with the diagonal entries equal to ±1. This type of symmetry translates very nicely to polyphase matrices. The following theorem shows the precise pattern.

THEOREM 3.1. *Let φ_j and ψ_j be scaling and wavelet functions generating a multiresolution analysis and a wavelet basis. Then* (3.1) *holds for $n = 2q + 1$ if and only if*

$$\boldsymbol{A}(\omega) = e^{-iq\omega} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda} \end{pmatrix} \boldsymbol{A}(-\omega) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}. \tag{3.2}$$

Proof. First, we need to translate the symmetry of functions into a symmetry pattern for the refinement masks. By combining the symmetry relationships (3.1) with the two-scale equations (2.1) we obtain

$$\begin{aligned} \boldsymbol{\varphi}(x) = \boldsymbol{\Sigma}\, \boldsymbol{\varphi}(n-x) &= \boldsymbol{\Sigma}\, \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{H}_k\, \boldsymbol{\varphi}(2\,(n-x)-k) \\ &= \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{\Sigma}\, \boldsymbol{H}_k\, \boldsymbol{\varphi}(n-(2x-n+k)) = \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{\Sigma}\, \boldsymbol{H}_{n-k} \boldsymbol{\Sigma}\, \boldsymbol{\varphi}(2x-k). \end{aligned}$$

Using the fact that the functions $\varphi_j(x-k)$, $j = 1, \ldots, r$, $k \in \mathbb{Z}$, form a Riesz basis for their span and, therefore, the coefficients in the two-scale relation (2.1) are unique, we observe that the symmetry pattern (3.1) of the generating functions implies the following symmetry pattern of the refinement mask:

$$\boldsymbol{H}_k = \boldsymbol{\Sigma}\, \boldsymbol{H}_{n-k} \boldsymbol{\Sigma} \tag{3.3}$$

for all $k \in \mathbb{Z}$. Similarly, we have

$$\begin{aligned} \boldsymbol{\psi}(x) = \boldsymbol{\Lambda}\, \boldsymbol{\psi}(n-x) &= \boldsymbol{\Lambda}\, \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{G}_k\, \boldsymbol{\varphi}(2\,(n-x)-k) \\ &= \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{\Lambda}\, \boldsymbol{G}_k\, \boldsymbol{\varphi}(n-(2x-n+k)) = \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{\Lambda}\, \boldsymbol{G}_{n-k} \boldsymbol{\Sigma}\, \boldsymbol{\varphi}(2x-k), \end{aligned}$$

and thus

$$\boldsymbol{G}_k = \boldsymbol{\Lambda} \boldsymbol{G}_{n-k} \boldsymbol{\Sigma}. \tag{3.4}$$

To show that the symmetry formula for the polyphase matrix holds, we need to divide the matrix sequences $\{\boldsymbol{H}_k\}_{k\in\mathbb{Z}}$ and $\{\boldsymbol{G}_k\}_{k\in\mathbb{Z}}$ into even and odd entries. We obtain

$$\begin{array}{lcccl} \boldsymbol{H}_{2k} & = & \boldsymbol{\Sigma}\, \boldsymbol{H}_{2q+1-2k} \boldsymbol{\Sigma} & = & \boldsymbol{\Sigma}\, \boldsymbol{H}_{2(q-k)+1} \boldsymbol{\Sigma}, \\ \boldsymbol{H}_{2k+1} & = & \boldsymbol{\Sigma}\, \boldsymbol{H}_{2q+1-(2k+1)} \boldsymbol{\Sigma} & = & \boldsymbol{\Sigma}\, \boldsymbol{H}_{2(q-k)} \boldsymbol{\Sigma}, \\ \boldsymbol{G}_{2k} & = & \boldsymbol{\Lambda}\, \boldsymbol{G}_{2q+1-2k} \boldsymbol{\Sigma} & = & \Lambda\, \boldsymbol{G}_{2(q-k)+1} \boldsymbol{\Sigma}, \\ \boldsymbol{G}_{2k+1} & = & \boldsymbol{\Lambda}\, \boldsymbol{G}_{2q+1-(2k+1)} \boldsymbol{\Sigma} & = & \Lambda\, \boldsymbol{G}_{2(q-k)} \boldsymbol{\Sigma}. \end{array}$$

Substituting this into the definition of the polyphase matrix (2.2) yields (3.2).

Let us now assume that (3.2) holds. The refinement masks then satisfy (3.3) and (3.4) and, consequently,

$$\begin{aligned} \boldsymbol{\Sigma}\boldsymbol{\varphi}(n-x) = \boldsymbol{\Sigma}\, \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{H}_k \boldsymbol{\varphi}(2\,(n-x)-k) &= \boldsymbol{\Sigma} \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{\Sigma}\, \boldsymbol{H}_{n-k} \boldsymbol{\Sigma} \boldsymbol{\varphi}(2\,(n-x)-k) \\ = \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{H}_{n-k} \boldsymbol{\Sigma}\, \boldsymbol{\varphi}(n-(2x-n+k)) &= \sqrt{2} \sum_{k\in\mathbb{Z}} \boldsymbol{H}_k \boldsymbol{\Sigma} \boldsymbol{\varphi}(n-(2x-k)). \end{aligned}$$

Vector function $\boldsymbol{\Sigma}\, \boldsymbol{\varphi}(n-x)$ thus solves the same two-scale equation as $\boldsymbol{\varphi}(x)$ and, because the $\boldsymbol{L}^2$ solution is unique, we have

$$\boldsymbol{\varphi}(x) = \boldsymbol{\Sigma}\, \boldsymbol{\varphi}(n-x).$$

Furthermore, then also

$$\begin{aligned}\boldsymbol{\Lambda}\boldsymbol{\psi}(n-x) &= \boldsymbol{\Lambda}\sqrt{2}\sum_{k\in\mathbb{Z}}\boldsymbol{G}_k\boldsymbol{\varphi}(2(n-x)-k) = \boldsymbol{\Lambda}\sqrt{2}\sum_{k\in\mathbb{Z}}\boldsymbol{\Lambda}\boldsymbol{G}_{n-k}\boldsymbol{\Sigma}\boldsymbol{\varphi}(2\,(n-x)-k)\\ &= \sqrt{2}\sum_{k\in\mathbb{Z}}\boldsymbol{G}_{n-k}\boldsymbol{\Sigma}\boldsymbol{\varphi}(n-(2x-n+k)) = \sqrt{2}\sum_{k\in\mathbb{Z}}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{\varphi}(n-(2x-k))\\ &= \sqrt{2}\sum_{k\in\mathbb{Z}}\boldsymbol{G}_k\boldsymbol{\varphi}(2x-k) = \boldsymbol{\psi}(x). \qquad \square\end{aligned}$$

It turns out that if the symmetry pattern is as described above, the numbers of symmetric and antisymmetric functions cannot be arbitrary.

THEOREM 3.2. *If all the generating functions are symmetric or antisymmetric about some point $q + 1/2$, $q \in \mathbb{Z}$, then exactly r functions must be symmetric and r antisymmetric. In other words, there must be exactly the same number of antisymmetric wavelet functions as symmetric scaling functions, and exactly the same number of symmetric wavelet functions as antisymmetric scaling functions.*

Proof. Substituting $\omega = 0$ into (3.2) yields

$$\boldsymbol{A}(0) = \begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}\end{pmatrix}\boldsymbol{A}(0)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}.$$

Since $\boldsymbol{A}(\omega)$ is paraunitary, $\boldsymbol{A}(0)$ is orthogonal and, therefore, invertible. We thus can write

$$\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}\end{pmatrix} = \boldsymbol{A}(0)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}\boldsymbol{A}(0)^{-1}.$$

The matrices $\left(\begin{smallmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}\end{smallmatrix}\right)$ are hence similar and their eigenvalues the same. Nevertheless, since

$$\frac{1}{\sqrt{2}}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{I}\\ \boldsymbol{I} & -\boldsymbol{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{I}\\ \boldsymbol{I} & -\boldsymbol{I}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{pmatrix}$$

the matrix $\left(\begin{smallmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{smallmatrix}\right)$ is also similar to the matrix $\left(\begin{smallmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{smallmatrix}\right)$. The matrix $\boldsymbol{\Lambda}$ thus must be, up to the permutation of the diagonal elements, equal to $-\boldsymbol{\Sigma}$. From here, the statement of the theorem follows immediately. $\square$

The theorem above implies that, without loss of generality, we can assume that $\boldsymbol{\Lambda} = -\boldsymbol{\Sigma}$. We can always achieve this by simply renumbering the wavelet functions. The resulting symmetry pattern of the polyphase matrix thus is

$$\boldsymbol{A}(\omega) = e^{-iq\omega}\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{pmatrix}\boldsymbol{A}(-\omega)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}. \tag{3.5}$$

4. Basic building blocks. The basic building blocks in the factorization (2.6) are an orthogonal matrix and normalized linear factors. We need to investigate how the symmetry constraint (3.5) can be enforced by the means of allowing only basic building blocks that are themselves symmetric in some way. For the sake of simplicity, let us first have a look at a polyphase matrix that has only two nonzero coefficients; i.e., let us assume for a while that

$$\boldsymbol{A}(\omega) = \boldsymbol{A}_0 + \boldsymbol{A}_1 e^{-i\omega}.$$

According to Theorem 2.2, such a polyphase matrix is paraunitary if and only if

$$\boldsymbol{A}(\omega) = \boldsymbol{Q}\,(\,\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{P}e^{-i\omega}),$$

where $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{P}$ is some orthogonal projector, that is, $\boldsymbol{PP} = \boldsymbol{P}$ and $\boldsymbol{P} = \boldsymbol{P}^T$. Because $\boldsymbol{Q} = \boldsymbol{A}(0)$, it has to satisfy

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Sigma} \end{pmatrix} \boldsymbol{Q} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}. \tag{4.1}$$

We will discuss orthogonal matrices with this type of symmetry in more detail later (see section 5). Now, let us have a close look at how the symmetry condition (3.5) influences the projector. We have

(4.2)

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{P}e^{-i\omega}) = \boldsymbol{A}(0)^{-1}\boldsymbol{A}(\omega) \\
&= \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \boldsymbol{A}(0)^{-1} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Sigma} \end{pmatrix} e^{-i\omega} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Sigma} \end{pmatrix} \boldsymbol{A}(-\omega) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \\
&= e^{-i\omega} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \boldsymbol{A}(0)^{-1}\boldsymbol{A}(-\omega) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \\
&= e^{-i\omega} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{P}e^{i\omega}) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} (\boldsymbol{P} + (\boldsymbol{I} - \boldsymbol{P})e^{-i\omega}) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}.
\end{aligned}$$

The projector $\boldsymbol{P}$ thus must satisfy

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \boldsymbol{P} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} = \boldsymbol{I} - \boldsymbol{P}. \tag{4.3}$$

Note that in the symmetry requirement for the normalized linear factor constructed from a pair of complementary projectors (4.2), the same symmetric orthogonal matrix appears on both sides of the linear factor. Consequently, if we consider a product of several linear factors with this symmetry, the result will have similar symmetry. More precisely, we will have

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{P}_1 + \boldsymbol{P}_1 e^{-i\omega}) \cdots (\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{-i\omega}) \\
&\qquad = e^{-iq\omega} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} (\boldsymbol{I} - \boldsymbol{P}_1 + \boldsymbol{P}_1 e^{i\omega}) \cdots (\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{i\omega}) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}.
\end{aligned}$$

If we, furthermore, add on a square constant matrix conforming to (4.1), then we will have

$$\begin{aligned}
&\boldsymbol{Q}\,(\boldsymbol{I} - \boldsymbol{P}_1 + \boldsymbol{P}_1 e^{-i\omega}) \cdots (\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{-i\omega}) \\
&= e^{-iq\omega} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{\Sigma} \end{pmatrix} \boldsymbol{Q} (\boldsymbol{I} - \boldsymbol{P}_1 + \boldsymbol{P}_1 e^{i\omega}) \cdots (\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{i\omega}) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix},
\end{aligned}$$

which is precisely our desired symmetry (3.5).

We thus now know what our basic building blocks will look like. We will need an orthogonal matrix satisfying (4.1) and some linear factors based on orthogonal projectors satisfying (4.3). We still have to answer two questions, though. First, it is not obvious how to obtain orthogonal matrices satisfying (4.1) and symmetric projectors conforming to (4.3). Second, we should make sure that multiwavelets that can be constructed in this way do not form just some special, very limited small set but, rather, that all possible multiwavelets with our chosen type of symmetry can be obtained in this manner. That is, we should verify that each paraunitary matrix satisfying (3.5) can indeed be expressed as the product of our symmetric factors.

5. Normalizing constant matrix. First, let us take a look at the matrix $\boldsymbol{Q} = \boldsymbol{A}(0)$. This matrix must be orthogonal and has to satisfy two additional conditions, the condition (2.5) involving the vector $\boldsymbol{w} = \int_{\mathbb{R}} \boldsymbol{\varphi}(x)\,dx$ and the symmetry condition (4.1). Let us first concentrate on the symmetry condition. We are going to show that any $2r \times 2r$ orthogonal matrix with this type of symmetry can be constructed from two ordinary $r \times r$ orthogonal matrices.

THEOREM 5.1. *A $2r \times 2r$ matrix $\boldsymbol{Q}$ is orthogonal and satisfies* (4.1) *if and only if there exist some two $r \times r$ orthogonal matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ such that*

$$\boldsymbol{Q} = \frac{1}{2\sqrt{2}} \begin{pmatrix} \boldsymbol{I}+\boldsymbol{\Sigma} & \boldsymbol{I}-\boldsymbol{\Sigma} \\ \boldsymbol{I}-\boldsymbol{\Sigma} & \boldsymbol{I}+\boldsymbol{\Sigma} \end{pmatrix} \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\Sigma} \\ \boldsymbol{I} & -\boldsymbol{\Sigma} \end{pmatrix}. \tag{5.1}$$

Proof. Assume that such orthogonal matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ exist. Since on the right-hand side of (5.1) there is a product of three orthogonal matrices, the matrix $\boldsymbol{Q}$ defined by this expression is also orthogonal. Furthermore,

$$\boldsymbol{Q} = \frac{1}{2\sqrt{2}} \begin{pmatrix} (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_1 + (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_2 & (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_1\boldsymbol{\Sigma} - (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_2\boldsymbol{\Sigma} \\ (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_1 + (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_2 & (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_1\boldsymbol{\Sigma} - (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_2\boldsymbol{\Sigma} \end{pmatrix}$$

and, by simple inspection, (4.1) holds.

Now, let us assume that $\boldsymbol{Q}$ is orthogonal and that it satisfies (4.1). Partitioning it into $r \times r$ blocks we see that, because of (4.1), it must have the form

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{E}_1 & \boldsymbol{\Sigma}\boldsymbol{E}_1\boldsymbol{\Sigma} \\ \boldsymbol{E}_2 & -\boldsymbol{\Sigma}\boldsymbol{E}_2\boldsymbol{\Sigma} \end{pmatrix},$$

where $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are some $r \times r$ matrices. Consequently,

$$\begin{aligned} &\frac{1}{2\sqrt{2}} \begin{pmatrix} \boldsymbol{I}+\boldsymbol{\Sigma} & \boldsymbol{I}-\boldsymbol{\Sigma} \\ \boldsymbol{I}-\boldsymbol{\Sigma} & \boldsymbol{I}+\boldsymbol{\Sigma} \end{pmatrix} \boldsymbol{Q} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{I} \\ \boldsymbol{\Sigma} & -\boldsymbol{\Sigma} \end{pmatrix} \\ &\qquad = \frac{1}{\sqrt{2}} \begin{pmatrix} (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{E}_1 + (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{E}_2 & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{E}_1 + (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{E}_2 \end{pmatrix}, \end{aligned} \tag{5.2}$$

and we can set

$$\boldsymbol{B}_1 = \frac{1}{\sqrt{2}}((\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{E}_1 + (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{E}_2), \qquad \boldsymbol{B}_2 = \frac{1}{\sqrt{2}}((\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{E}_1 + (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{E}_2).$$

Both the matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are orthogonal, because the left-hand side of (5.2) is the product of three orthogonal factors, and thus $\left(\begin{smallmatrix} \boldsymbol{B}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_2 \end{smallmatrix}\right)$, which is on the right-hand side of (5.2), is also orthogonal. □

The theorem above gives the means for parameterizing orthogonal matrices satisfying (4.1). Simply, two $r \times r$ orthogonal matrices are parameterized in a standard

way, by Givens rotations or Householder reflections, and are combined as described by (5.1).

The symmetry (4.1) and orthogonality, though, are not the only properties matrix $\boldsymbol{Q} = \boldsymbol{A}(0)$ has to have. There is also the condition (2.5) concerning the vector $\boldsymbol{w} = \int_{\mathbb{R}} \boldsymbol{\varphi}(x)\, dx$. So as to be able to include this condition in the parameterization, we need to reformulate it as constraints on the matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$.

THEOREM 5.2. *Let the matrix $\boldsymbol{Q}$ be defined by* (5.1). *Then, for any vector $\boldsymbol{w}$,*

$$\boldsymbol{Q}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{w}\end{pmatrix} = \sqrt{2}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{0}\end{pmatrix} \tag{5.3}$$

if and only if

$$\boldsymbol{B}_1\,(\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{w} = (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{w}, \qquad \boldsymbol{B}_2\,(\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{w} = (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{w}. \tag{5.4}$$

Proof. By substituting $\boldsymbol{Q}$ from (5.1) into (5.3) we obtain

$$\frac{1}{2\sqrt{2}}\begin{pmatrix}\boldsymbol{I}+\boldsymbol{\Sigma} & \boldsymbol{I}-\boldsymbol{\Sigma}\\ \boldsymbol{I}-\boldsymbol{\Sigma} & \boldsymbol{I}+\boldsymbol{\Sigma}\end{pmatrix}\begin{pmatrix}\boldsymbol{B}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{B}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{\Sigma}\\ \boldsymbol{I} & -\boldsymbol{\Sigma}\end{pmatrix}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{w}\end{pmatrix} = \sqrt{2}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{0}\end{pmatrix}.$$

Multiplying both sides of this equation by the matrix $\frac{1}{\sqrt{2}}\begin{pmatrix}\boldsymbol{I}+\boldsymbol{\Sigma} & \boldsymbol{I}-\boldsymbol{\Sigma}\\ \boldsymbol{I}-\boldsymbol{\Sigma} & \boldsymbol{I}+\boldsymbol{\Sigma}\end{pmatrix}$ then yields

$$\begin{pmatrix}\boldsymbol{B}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{B}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{\Sigma}\\ \boldsymbol{I} & -\boldsymbol{\Sigma}\end{pmatrix}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{w}\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}+\boldsymbol{\Sigma} & \boldsymbol{I}-\boldsymbol{\Sigma}\\ \boldsymbol{I}-\boldsymbol{\Sigma} & \boldsymbol{I}+\boldsymbol{\Sigma}\end{pmatrix}\begin{pmatrix}\boldsymbol{w}\\ \boldsymbol{0}\end{pmatrix}$$

or

$$\begin{pmatrix}\boldsymbol{B}_1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{B}_2\end{pmatrix}\begin{pmatrix}(\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{w}\\ (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{w}\end{pmatrix} = \begin{pmatrix}(\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{w}\\ (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{w}\end{pmatrix}.$$

From here, the statement of the theorem follows immediately. ∎

The vector $(\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{w}$, respectively, $(\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{w}$, may be zero. If it is not, then it is an eigenvector of the matrix $\boldsymbol{B}_1$, respectively, $\boldsymbol{B}_2$, with eigenvalue 1. How do we obtain an orthogonal matrix that has eigenvalue 1 with a prescribed eigenvector?

Suppose we want to construct an $r \times r$ orthogonal matrix $\boldsymbol{B}$ such that some given vector $\boldsymbol{v}$, $||\boldsymbol{v}|| = 1$, is its eigenvector with eigenvalue 1, i.e.,

$$\boldsymbol{B}\boldsymbol{v} = \boldsymbol{v}. \tag{5.5}$$

If $\boldsymbol{v} = \boldsymbol{e}_1$, the first column of the identity matrix, then (5.5) is equivalent to stating that the first column of $\boldsymbol{B}$ is $\boldsymbol{e}_1$. As the matrix $\boldsymbol{B}$ is orthogonal, the norm of the first row of $\boldsymbol{B}$ is 1. However, because its first column is $\boldsymbol{e}_1$, the element in the upper left corner is 1. This means that all the remaining elements in the first row are necessarily 0. The matrix $\boldsymbol{B}$ thus has the form

$$\boldsymbol{B} = \begin{pmatrix}1 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{K}\end{pmatrix}, \tag{5.6}$$

where $\boldsymbol{K}$ is some $(r-1)\times(r-1)$ orthogonal matrix.

In the case that $\boldsymbol{v} \neq \boldsymbol{e}_1$, we can use a reflection to turn it into $\boldsymbol{e}_1$. If we set

$$\boldsymbol{F} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^T, \qquad \boldsymbol{u} = \frac{\boldsymbol{v}-\boldsymbol{e}_1}{||\boldsymbol{v}-\boldsymbol{e}_1||},$$

then $\boldsymbol{F}$ is a symmetric orthogonal matrix and

$$\boldsymbol{F}\boldsymbol{v} = \boldsymbol{e}_1.$$

By multiplying both sides of (5.5) by $\boldsymbol{F}$ we obtain

$$\boldsymbol{F}\boldsymbol{B}\boldsymbol{v} = \boldsymbol{F}\boldsymbol{v},$$

and from here

$$\boldsymbol{F}\boldsymbol{B}\boldsymbol{F}\boldsymbol{e}_1 = \boldsymbol{e}_1.$$

The matrix $\boldsymbol{F}\boldsymbol{B}\boldsymbol{F}$ thus has the form (5.6), i.e.,

$$\boldsymbol{B} = \boldsymbol{F}\begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{K} \end{pmatrix}\boldsymbol{F}, \tag{5.7}$$

where $\boldsymbol{K}$ is some $(r-1)\times(r-1)$ orthogonal matrix. We thus can create a parameterization of orthogonal matrices satisfying (5.5) by parameterizing an $(r-1)\times(r-1)$ orthogonal matrix $\boldsymbol{K}$ in a standard way and applying (5.7).

6. Projectors. Let us now investigate how a pair of complementary symmetric projectors satisfying (4.3) can be constructed and parameterized. Besides creating the parameterization of polyphase matrices featuring symmetry (3.5), this result will be helpful also in the following section, where we will derive the algorithm for factoring a symmetric paraunitary polyphase matrix into symmetric factors.

THEOREM 6.1. *A $2r\times 2r$ matrix $\boldsymbol{P}$ is an orthogonal projector and it displays symmetry* (4.3) *if and only if there exists a $2r\times 2r$ orthogonal matrix $\boldsymbol{R}$ such that*

$$\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\boldsymbol{R}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} = \boldsymbol{R} \tag{6.1}$$

and

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^T,$$

where the $2r\times r$ matrix $\boldsymbol{U}$ is the left half of $\boldsymbol{R}$.

Proof. Let us first assume that such a matrix $\boldsymbol{R}$ exists. We see immediately that $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^T$ is a symmetric matrix. Moreover, since $\boldsymbol{R}$ is orthogonal, i.e., $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$, we have $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and, consequently, $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$. The matrix $\boldsymbol{P}$ therefore is an orthogonal projector. Let us denote by $\boldsymbol{V}$ the right half of $\boldsymbol{R}$. Since $\boldsymbol{R}$ is orthogonal, $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$ and, consequently, $\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^T$. The condition (6.1) holds if and only if

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\boldsymbol{U}\boldsymbol{\Sigma}. \tag{6.2}$$

Therefore, if it holds, then

$$\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^T = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\boldsymbol{U}\boldsymbol{U}^T\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\boldsymbol{P}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}.$$

Now we need to prove the opposite implication. We are going to assume that $\boldsymbol{P}$ is an orthogonal projector satisfying (4.3), and we will show that such a matrix $\boldsymbol{R}$

exists. Since the matrix $\begin{pmatrix} \mathbf{0} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{0} \end{pmatrix}$ is invertible, it implies that the rank of $\boldsymbol{P}$ is the same as the rank of the complementary orthogonal projector $\boldsymbol{I}-\boldsymbol{P}$. That is, the rank of $\boldsymbol{P}$ is $2r/2 = r$. Let $\boldsymbol{U}$ be any $2r \times r$ matrix such that the columns of $\boldsymbol{U}$ form an orthonormal basis for the range of $\boldsymbol{P}$. We have $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$, and it is not difficult to verify that, because of this, $\boldsymbol{U}\boldsymbol{U}^T$ is an orthogonal projector and its range is the range of $\boldsymbol{U}$. Consequently, since an orthogonal projector is uniquely defined by its range, $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{P}$. Now, let us use (6.2) to define $\boldsymbol{V}$. This definition guarantees that the matrix $\boldsymbol{R} = (\,\boldsymbol{U} \quad \boldsymbol{V}\,)$ satisfies (6.1). So the only thing we need to show to complete the proof is that the matrix $\boldsymbol{R}$ is orthogonal. $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$ when $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{U}^T\boldsymbol{V} = \boldsymbol{V}^T\boldsymbol{U} = \mathbf{0}$. We recall that we have chosen $\boldsymbol{U}$ such that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$, and it is easy to verify that $\boldsymbol{V}$ defined by (6.2) then satisfies $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. We thus need to show that $\boldsymbol{U}^T\boldsymbol{V} = \boldsymbol{V}^T\boldsymbol{U} = \mathbf{0}$; that is, the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are mutually orthogonal. It follows from (4.3) that if $\boldsymbol{u}$ is a vector from the range of $\boldsymbol{P}$, that is, from the nullspace of $\boldsymbol{I}-\boldsymbol{P}$, then the vector $\begin{pmatrix} \mathbf{0} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{0} \end{pmatrix}\boldsymbol{u}$ is from the nullspace of $\boldsymbol{P}$. Since the columns of $\boldsymbol{U}$ form a basis for the range of $\boldsymbol{P}$ and $\boldsymbol{V}$ is defined by (6.2)—that is, each of the columns of $\boldsymbol{V}$ has the form $\begin{pmatrix} \mathbf{0} & \mathbf{\Sigma} \\ \mathbf{\Sigma} & \mathbf{0} \end{pmatrix}\boldsymbol{u}$, where $\boldsymbol{u}$ is some column of $\boldsymbol{U}$—then all the columns of $\boldsymbol{V}$ are from the nullspace of $\boldsymbol{P}$. Because $\boldsymbol{P}$ is an orthogonal projector, its range and nullspace are orthogonal. Consequently, every column of $\boldsymbol{V}$ is orthogonal to every column of $\boldsymbol{U}$, and $\boldsymbol{R}$ is orthogonal. □

Theorem 6.1 transforms the problem of constructing orthogonal projectors satisfying (4.3) into the problem of constructing a $2r \times 2r$ orthogonal matrix satisfying (6.1). Such a matrix can be built from two $r \times r$ orthogonal matrices in a similar fashion as the orthogonal matrix $\boldsymbol{Q}$ satisfying (4.1). Only the transformation matrices are a little different.

THEOREM 6.2. *A $2r \times 2r$ matrix $\boldsymbol{R}$ is orthogonal and satisfies* (6.1) *if and only if there exist some $r \times r$ orthogonal matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ such that*

$$\boldsymbol{R} = \frac{1}{2}\begin{pmatrix} \boldsymbol{I} & \boldsymbol{I} \\ \mathbf{\Sigma} & -\mathbf{\Sigma} \end{pmatrix}\begin{pmatrix} \boldsymbol{C}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_2 \end{pmatrix}\begin{pmatrix} \boldsymbol{I} & \mathbf{\Sigma} \\ \boldsymbol{I} & -\mathbf{\Sigma} \end{pmatrix}. \tag{6.3}$$

Proof. First, let $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ be some orthogonal $r \times r$ matrices. The matrix $\boldsymbol{R}$ defined by (6.3) is then orthogonal because it is the product of three orthogonal matrices. Furthermore,

$$\boldsymbol{R} = \frac{1}{2}\begin{pmatrix} \boldsymbol{C}_1 + \boldsymbol{C}_2 & (\boldsymbol{C}_1 - \boldsymbol{C}_2)\mathbf{\Sigma} \\ \mathbf{\Sigma}(\boldsymbol{C}_1 - \boldsymbol{C}_2) & \mathbf{\Sigma}(\boldsymbol{C}_1 + \boldsymbol{C}_2)\mathbf{\Sigma} \end{pmatrix}, \tag{6.4}$$

and it can be easily verified by inspection that it satisfies (6.1).

To prove the converse, let us now assume that $\boldsymbol{R}$ is some $2r \times 2r$ matrix satisfying (6.1) and let us partition it into $r \times r$ blocks. It satisfies (6.1) if and only if it has the form

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{U}_1 & \mathbf{\Sigma}\boldsymbol{U}_2\mathbf{\Sigma} \\ \boldsymbol{U}_2 & \mathbf{\Sigma}\boldsymbol{U}_1\mathbf{\Sigma} \end{pmatrix},$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are some $r \times r$ matrices. We thus have

$$\begin{aligned} \frac{1}{2}\begin{pmatrix} \boldsymbol{I} & \mathbf{\Sigma} \\ \boldsymbol{I} & -\mathbf{\Sigma} \end{pmatrix}\boldsymbol{R}\begin{pmatrix} \boldsymbol{I} & \boldsymbol{I} \\ \mathbf{\Sigma} & -\mathbf{\Sigma} \end{pmatrix} &= \frac{1}{2}\begin{pmatrix} \boldsymbol{I} & \mathbf{\Sigma} \\ \boldsymbol{I} & -\mathbf{\Sigma} \end{pmatrix}\begin{pmatrix} \boldsymbol{U}_1 & \mathbf{\Sigma}\boldsymbol{U}_2\mathbf{\Sigma} \\ \boldsymbol{U}_2 & \mathbf{\Sigma}\boldsymbol{U}_1\mathbf{\Sigma} \end{pmatrix}\begin{pmatrix} \boldsymbol{I} & \boldsymbol{I} \\ \mathbf{\Sigma} & -\mathbf{\Sigma} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{U}_1 + \mathbf{\Sigma}\boldsymbol{U}_2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{U}_1 - \mathbf{\Sigma}\boldsymbol{U}_2 \end{pmatrix}. \end{aligned}$$

Therefore, if we set

$$\boldsymbol{C}_1 = \boldsymbol{U}_1 + \boldsymbol{\Sigma}\boldsymbol{U}_2, \qquad \boldsymbol{C}_2 = \boldsymbol{U}_1 - \boldsymbol{\Sigma}\boldsymbol{U}_2, \tag{6.5}$$

then $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are orthogonal and (6.3) holds. □

We could create a parameterization for orthogonal projectors $\boldsymbol{P}$ satisfying the symmetry constraint (4.3) by parameterizing $r \times r$ orthogonal matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ in a standard way. Nevertheless, such a parameterization would have twice as many free parameters as is necessary. The map between $\boldsymbol{R}$ and $\boldsymbol{P}$ is many to one. $\boldsymbol{U}$ is not determined by $\boldsymbol{P}$ uniquely. It can be *any* matrix such that its columns form an orthonormal basis for the range of $\boldsymbol{P}$. It turns out that we can eliminate this redundancy simply by setting one of the matrices $\boldsymbol{C}_1$ or $\boldsymbol{C}_2$ equal to the identity matrix. We choose here $\boldsymbol{C}_1 = \boldsymbol{I}$.

THEOREM 6.3. *A $2r \times 2r$ matrix $\boldsymbol{P}$ is an orthogonal projector satisfying* (4.3) *if and only if it has the form*

$$\boldsymbol{P} = \frac{1}{4}\begin{pmatrix} 2\boldsymbol{I} + \boldsymbol{C} + \boldsymbol{C}^T & (\boldsymbol{C} - \boldsymbol{C}^T)\boldsymbol{\Sigma} \\ -\boldsymbol{\Sigma}(\boldsymbol{C} - \boldsymbol{C}^T) & \boldsymbol{\Sigma}(2\boldsymbol{I} - \boldsymbol{C} - \boldsymbol{C}^T)\boldsymbol{\Sigma} \end{pmatrix}, \tag{6.6}$$

where $\boldsymbol{C}$ is an $r \times r$ orthogonal matrix.

Proof. To prove this statement we need to combine the results of the previous two theorems. We have $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is the left half of the matrix $\boldsymbol{R}$ given by (6.4). This means that $\boldsymbol{P}$ can be expressed in terms of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ as follows:

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^T = \frac{1}{4}\begin{pmatrix} (\boldsymbol{C}_1 + \boldsymbol{C}_2)(\boldsymbol{C}_1 + \boldsymbol{C}_2)^T & (\boldsymbol{C}_1 + \boldsymbol{C}_2)(\boldsymbol{C}_1 - \boldsymbol{C}_2)^T\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}(\boldsymbol{C}_1 - \boldsymbol{C}_2)(\boldsymbol{C}_1 + \boldsymbol{C}_2)^T & \boldsymbol{\Sigma}(\boldsymbol{C}_1 - \boldsymbol{C}_2)(\boldsymbol{C}_1 - \boldsymbol{C}_2)^T\boldsymbol{\Sigma} \end{pmatrix}. \tag{6.7}$$

When we substitute $\boldsymbol{C}_1 = \boldsymbol{I}$ and $\boldsymbol{C}_2 = \boldsymbol{C}$ into (6.7), we obtain (6.6). Therefore, by Theorems 6.1 and 6.2, if $\boldsymbol{C}$ is an orthogonal matrix, then $\boldsymbol{P}$ defined by (6.6) is an orthogonal projector that satisfies (6.1).

On the other hand, let us now suppose that $\boldsymbol{P}$ is an orthogonal projector that satisfies (6.1), and let us prove that it must have the form (6.6). Let $\boldsymbol{W}$ be any $r \times r$ orthogonal matrix and let

$$\tilde{\boldsymbol{R}} = \boldsymbol{R}\begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{\Sigma} \end{pmatrix}.$$

This means that

$$\tilde{\boldsymbol{R}} = \frac{1}{2}\begin{pmatrix} \tilde{\boldsymbol{C}}_1 + \tilde{\boldsymbol{C}}_2 & (\tilde{\boldsymbol{C}}_1 - \tilde{\boldsymbol{C}}_2)\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}(\tilde{\boldsymbol{C}}_1 - \tilde{\boldsymbol{C}}_2) & \boldsymbol{\Sigma}(\tilde{\boldsymbol{C}}_1 + \tilde{\boldsymbol{C}}_2)\boldsymbol{\Sigma} \end{pmatrix},$$

where

$$\tilde{\boldsymbol{C}}_1 = \boldsymbol{C}_1\boldsymbol{W}, \qquad \tilde{\boldsymbol{C}}_2 = \boldsymbol{C}_2\boldsymbol{W}.$$

The matrix $\tilde{\boldsymbol{R}}$ is the product of two orthogonal matrices, and therefore it is also orthogonal. Furthermore,

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\tilde{\boldsymbol{R}}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \\ &\quad = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\boldsymbol{R}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{\Sigma} \end{pmatrix}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} \\ &\quad = \boldsymbol{R}\begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}\boldsymbol{W}\boldsymbol{\Sigma} \end{pmatrix} = \tilde{\boldsymbol{R}}. \end{aligned}$$

Also, if we denote by $\tilde{\boldsymbol{U}}$ the left half of $\tilde{\boldsymbol{R}}$, we have

$$\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^T = \boldsymbol{U}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{U}^T = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{P}.$$

For any $r \times r$ orthogonal matrix $\boldsymbol{W}$, orthogonal matrices $\tilde{\boldsymbol{C}}_1 = \boldsymbol{C}_1\boldsymbol{W}$ and $\tilde{\boldsymbol{C}}_2 = \boldsymbol{C}_2\boldsymbol{W}$ thus generate exactly the same projector as $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. If we choose $\boldsymbol{W} = \boldsymbol{C}_1^T$, then $\tilde{\boldsymbol{C}}_1 = \boldsymbol{I}$. We can therefore, without loss of generality, always pick $\boldsymbol{C}_1 = \boldsymbol{I}$. By substituting this into (6.7) we obtain (6.6), with $\boldsymbol{C} = \boldsymbol{C}_2$. □

7. Factorization. We are now going to show that any paraunitary polyphase matrix of order q satisfying the symmetry constraint (3.5) can be factored into the product of an orthogonal matrix satisfying (4.1) and q normalized linear factors that contain complementary orthogonal projectors for which (4.3) holds. That is, we are going to give here a symmetric form of Theorem 2.2. The reason for showing this is to prove that our factorization-based parameterization covers *all* compactly supported multiwavelets with our chosen type of symmetry.

THEOREM 7.1. *A polyphase matrix $\boldsymbol{A}(\omega)$ of order q is paraunitary and satisfies the symmetry constraint* (3.5) *if and only if there exists an orthogonal matrix $\boldsymbol{Q}$ satisfying* (4.1) *and orthogonal projectors $\boldsymbol{P}_j$, $j = 1, \ldots, q$, satisfying* (4.3) *such that*

$$\boldsymbol{A}(\omega) = \boldsymbol{Q}\ (\boldsymbol{I} - \boldsymbol{P}_1 + \boldsymbol{P}_1 e^{-i\omega}) \cdots (\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{-i\omega}). \tag{7.1}$$

Proof. First of all, if $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{P}_j$, $j = 1, \ldots, q$, are orthogonal projectors, then all the factors on the right-hand side of (7.1) are paraunitary, and thus $\boldsymbol{A}(\omega)$ defined by (7.1) is also paraunitary. Also, as we have shown in section 4, if $\boldsymbol{Q}$ satisfies (4.1) and the projectors (4.3), then (3.5) holds. So we only need to show that we can factor each paraunitary polyphase matrix satisfying (3.5) in this way. The proof is constructive and gives a practical algorithm for factoring a polyphase matrix into a constant matrix conforming to (4.1) and symmetric linear factors.

If $q = 0$, then $\boldsymbol{A}(\omega) = \boldsymbol{A}(0) = \boldsymbol{Q}$ and the statement holds. So let us assume that $q > 0$. The strategy here is to find an orthogonal projector $\boldsymbol{P}_q$ satisfying (4.3) such that

$$\boldsymbol{A}(\omega) = \boldsymbol{S}(\omega)(\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{-i\omega}),$$

where $\boldsymbol{S}(\omega)$ is of order $q - 1$. As we will show, $\boldsymbol{S}(\omega)$ then also satisfies (3.5) (with $q - 1$ instead of q), and we thus can repeat the process, reducing in each step the order by 1, until we reach order 0. We then will have an orthogonal matrix satisfying (4.1), and q projectors satisfying (4.3) and (7.1) will hold.

Note that if $\boldsymbol{P}_q$ is an orthogonal projector, then

$$(\boldsymbol{I} - \boldsymbol{P}_q + \boldsymbol{P}_q e^{-i\omega})(\boldsymbol{P}_q e^{i\omega} + \boldsymbol{I} - \boldsymbol{P}_q) = \boldsymbol{I}$$

and hence

$$\boldsymbol{S}(\omega) = \boldsymbol{A}(\omega)(\boldsymbol{P}_q e^{i\omega} + \boldsymbol{I} - \boldsymbol{P}_q). \tag{7.2}$$

If $\boldsymbol{P}_q$ satisfies (4.3), then

$$(\boldsymbol{P}_q e^{i\omega} + \boldsymbol{I} - \boldsymbol{P}_q) = e^{i\omega} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix} (\boldsymbol{P}_q e^{-i\omega} + \boldsymbol{I} - \boldsymbol{P}_q) \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{0} \end{pmatrix}$$

and

$$\begin{aligned}\boldsymbol{S}(\omega) &= \boldsymbol{A}(\omega)(\boldsymbol{P}_q e^{i\omega} + \boldsymbol{I} - \boldsymbol{P}_q)\\ &= e^{-iq\omega}\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{pmatrix}\boldsymbol{A}(-\omega)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}(\boldsymbol{P}_q e^{-i\omega} + \boldsymbol{I} - \boldsymbol{P}_q)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}\\ &= e^{-i(q-1)\omega}\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{pmatrix}\boldsymbol{A}(-\omega)(\boldsymbol{P}_q e^{-i\omega} + \boldsymbol{I} - \boldsymbol{P}_q)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}\\ &= e^{-i(q-1)\omega}\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{\Sigma}\end{pmatrix}\boldsymbol{S}(-\omega)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}.\end{aligned}$$

That is, $\boldsymbol{S}(\omega)$ indeed features the same symmetry as $\boldsymbol{A}(\omega)$, described by (3.5). Only the order is $q-1$ here.

Let us now return to the formula (7.2). If $\boldsymbol{P}_q$ were just a general orthogonal projector, the coefficients $\boldsymbol{S}_{-1}$ and $\boldsymbol{S}_q$ would be nonzero. So as the order of $\boldsymbol{S}(\omega)$ is $q-1$, we need to use an orthogonal projector $\boldsymbol{P}_q$ that will make them zero. That is, we need to find a projector $\boldsymbol{P}_q$ for which (4.3) holds and

$$\boldsymbol{S}_{-1} = \boldsymbol{A}_0\boldsymbol{P}_q = \boldsymbol{0}, \qquad \boldsymbol{S}_q = \boldsymbol{A}_q(\boldsymbol{I} - \boldsymbol{P}_q) = \boldsymbol{0}.$$

To construct such a projector, we need to realize that because $\boldsymbol{A}(\omega)$ is paraunitary and therefore (2.3) holds, we have

$$\boldsymbol{A}_0\boldsymbol{A}_q^T = \boldsymbol{0}. \tag{7.3}$$

Furthermore, the symmetry constraint (3.5) implies that

$$\boldsymbol{A}_0 = \begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}\boldsymbol{A}_q\begin{pmatrix}\boldsymbol{0} & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & \boldsymbol{0}\end{pmatrix}. \tag{7.4}$$

The former equation implies that rank $\boldsymbol{A}_0$ + rank $\boldsymbol{A}_q \le 2r$, while the latter implies that rank $\boldsymbol{A}_0$ = rank $\boldsymbol{A}_q$. Consequently,

$$\text{rank } \boldsymbol{A}_0 = \text{rank } \boldsymbol{A}_q \le r.$$

If the ranks are equal to r, it is sufficient to find any matrix $\boldsymbol{U}$ such that its columns form an orthonormal basis for the row space of $\boldsymbol{A}_q$ and to set $\boldsymbol{P}_q = \boldsymbol{U}\boldsymbol{U}^T$. It follows from (7.4) that if we define a matrix $\boldsymbol{V}$ by (6.2), then the columns of the matrix $\boldsymbol{V}$ form an orthonormal basis for the row space of $\boldsymbol{A}_0$ and, because (7.3) holds, the matrix $\boldsymbol{R} = (\,\boldsymbol{U} \quad \boldsymbol{V}\,)$ is orthogonal. It also satisfies (6.1), and therefore, by Theorem 6.2, $\boldsymbol{P}_q$ is an orthogonal projector satisfying (4.3). The range of this projector is the range of $\boldsymbol{U}$, that is, the row space of $\boldsymbol{A}_q$. We thus have $\boldsymbol{A}_q\boldsymbol{P}_q = \boldsymbol{A}_q$, and therefore $\boldsymbol{A}_q(\boldsymbol{I} - \boldsymbol{P}_q) = \boldsymbol{0}$. The nullspace of this projector is the orthogonal complement of the range of $\boldsymbol{U}$, i.e., of the row space of $\boldsymbol{A}_q$ which is, by (7.3), the row space of $\boldsymbol{A}_0$, and we thus also have $\boldsymbol{A}_0\boldsymbol{P}_q = \boldsymbol{0}$, as required.

The only question that remains to be answered hence is what to do if rank $\boldsymbol{A}_0$ = rank $\boldsymbol{A}_q = s < r$. We need to find a way to complete an orthonormal basis of the row space of $\boldsymbol{A}_q$ to an $2r \times r$ matrix $\boldsymbol{U}$ in such a way that, if we define $\boldsymbol{V}$ by (6.2), $\boldsymbol{R} = (\,\boldsymbol{U} \quad \boldsymbol{V}\,)$ will be an orthogonal matrix. The answer is hidden in the proof of Theorem 6.2. Let $\boldsymbol{T}$ be any matrix such that its columns form an orthonormal basis for the row space of $\boldsymbol{A}_q$. We break the matrix $\boldsymbol{T}$ vertically into two blocks of r rows, $\boldsymbol{T} = \begin{pmatrix}\boldsymbol{T}_1\\ \boldsymbol{T}_2\end{pmatrix}$. Now, following the pattern of (6.5), let $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ be any

$r \times r$ orthogonal matrices such that their first s columns are equal to $\boldsymbol{T}_1 + \boldsymbol{\Sigma}\boldsymbol{T}_2$ and $\boldsymbol{T}_1 - \boldsymbol{\Sigma}\boldsymbol{T}_2$, respectively. Then the matrix $\boldsymbol{R}$ defined by (6.3) is an orthogonal matrix satisfying the symmetry constraint (6.1) and its first s columns are equal to $\boldsymbol{T}$. Hence, if we partition it into halves, $\boldsymbol{R} = (\,\boldsymbol{U} \quad \boldsymbol{V}\,)$ and set $\boldsymbol{P}_q = \boldsymbol{U}\boldsymbol{U}^T$, $\boldsymbol{P}_q$ will be an orthogonal projector satisfying the symmetry constraint (4.3). The range of $\boldsymbol{T}$, i.e., the row space of $\boldsymbol{A}_q$, is a subset of the range of $\boldsymbol{U}$, i.e., of the range of $\boldsymbol{P}_q$, and therefore $\boldsymbol{A}_q(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{0}$. Because $\boldsymbol{R}$ displays the symmetry (6.1) and (7.4) holds, the last s columns of $\boldsymbol{R}$ forms an orthonormal basis for the row space of $\boldsymbol{A}_0$. The row space of $\boldsymbol{A}_0$ thus is a part of the nullspace of $\boldsymbol{P}_q$ and $\boldsymbol{A}_0\boldsymbol{P}_q = \boldsymbol{0}$, as needed. □

8. Example. Let us now demonstrate how the method works. The starting point is the formula (7.1). To keep things simple, let us choose $r = 2$ and $q = 1$. We thus will have two scaling and two wavelet functions, $\boldsymbol{A}(\omega)$ will be a 4×4 linear matrix trigonometric polynomial, and it will have the form

$$\boldsymbol{A}(\omega) = \boldsymbol{Q}\,(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{P}e^{-i\omega}), \tag{8.1}$$

where $\boldsymbol{Q}$ is some orthogonal matrix and $\boldsymbol{P}$ an orthogonal projector. We also need to decide what symmetry pattern we want to achieve. Let us say we want to have one of the scaling functions symmetric and the other antisymmetric. That is, we pick

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Note that this means that we will also have one wavelet symmetric and one antisymmetric (see Theorem 3.2). Finally, we also need to choose the vector $\boldsymbol{w}$. Since $\boldsymbol{w} = \int_{\mathbb{R}} \boldsymbol{\varphi}(x)\,dx$, and $\varphi_2(x)$ is going to be antisymmetric, the second component of $\boldsymbol{w}$ must be zero. We thus set

$$\boldsymbol{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Now we need to use formulas (5.1) and (6.6). Multiplying the three factors on the right-hand side of (5.1) we obtain

$$\boldsymbol{Q} = \frac{1}{2\sqrt{2}} \begin{pmatrix} (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_1 + (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_2 & (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_1\boldsymbol{\Sigma} - (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_2\boldsymbol{\Sigma} \\ (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_1 + (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_2 & (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{B}_1\boldsymbol{\Sigma} - (\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{B}_2\boldsymbol{\Sigma} \end{pmatrix}. \tag{8.2}$$

For the projector $\boldsymbol{P}$ we have

$$\boldsymbol{P} = \frac{1}{4} \begin{pmatrix} 2\boldsymbol{I} + \boldsymbol{C} + \boldsymbol{C}^T & (\boldsymbol{C} - \boldsymbol{C}^T)\boldsymbol{\Sigma} \\ -\boldsymbol{\Sigma}(\boldsymbol{C} - \boldsymbol{C}^T) & \boldsymbol{\Sigma}(2\boldsymbol{I} - \boldsymbol{C} - \boldsymbol{C}^T)\boldsymbol{\Sigma} \end{pmatrix}. \tag{8.3}$$

Hence, there are three unknown 2×2 orthogonal matrices that we need to parameterize: $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, and $\boldsymbol{C}$. We must not forget, though, that (5.4) needs to be satisfied, too. For the matrix $\boldsymbol{\Sigma}$ and the vector $\boldsymbol{w}$ we have chosen, we obtain

$$(\boldsymbol{I}+\boldsymbol{\Sigma})\boldsymbol{w} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \qquad (\boldsymbol{I}-\boldsymbol{\Sigma})\boldsymbol{w} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This means that while $\boldsymbol{B}_2$ can be an arbitrary 2×2 orthogonal matrix, because the second equation of (5.4) does not present any restriction on it, $\boldsymbol{B}_1$ must have eigenvalue 1 with eigenvector $\left(\begin{smallmatrix} 2 \\ 0 \end{smallmatrix}\right)$. This leaves us with only two candidates for $\boldsymbol{B}_1$.

Let us choose $\boldsymbol{B}_1 = \boldsymbol{I}$. Since $\boldsymbol{B}_2$ and $\boldsymbol{C}$ are arbitrary 2×2 orthogonal matrices, each will contribute one free parameter. Let us set

$$\boldsymbol{B}_2 = \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix}, \qquad \boldsymbol{C} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

Substituting this to (8.2) and (8.3) we obtain

$$\boldsymbol{Q} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 & 0 \\ -\sin\beta & \cos\beta & \sin\beta & \cos\beta \\ \cos\beta & \sin\beta & -\cos\beta & \sin\beta \\ 0 & 1 & 0 & -1 \end{pmatrix},$$

$$\boldsymbol{P} = \frac{1}{2} \begin{pmatrix} 1+\cos\theta & 0 & 0 & -\sin\theta \\ 0 & 1+\cos\theta & -\sin\theta & 0 \\ 0 & -\sin\theta & 1-\cos\theta & 0 \\ -\sin\theta & 0 & 0 & 1-\cos\theta \end{pmatrix}.$$

Finally, if we substitute this into (8.1) and extract the refinement masks, we get

$$\boldsymbol{H}_0 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} 1 & 0 \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 1-\cos\theta & 0 \\ 0 & 1-\cos\theta \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 0 & \sin\beta \\ \sin\beta & 0 \end{pmatrix} \right),$$

$$\boldsymbol{H}_1 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} 1 & 0 \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 0 & \sin\beta \\ \sin\beta & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 1+\cos\theta & 0 \\ 0 & 1+\cos\theta \end{pmatrix} \right),$$

$$\boldsymbol{H}_2 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} 1 & 0 \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 1+\cos\theta & 0 \\ 0 & 1+\cos\theta \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 0 & -\sin\beta \\ -\sin\beta & 0 \end{pmatrix} \right),$$

$$\boldsymbol{H}_3 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} 1 & 0 \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 0 & -\sin\beta \\ -\sin\beta & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ \sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} 1-\cos\theta & 0 \\ 0 & 1-\cos\theta \end{pmatrix} \right),$$

$$\boldsymbol{G}_0 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} \cos\beta & \sin\beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1-\cos\theta & 0 \\ 0 & 1-\cos\theta \end{pmatrix} + \begin{pmatrix} -\cos\beta & \sin\beta \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & \sin\beta \\ \sin\beta & 0 \end{pmatrix} \right),$$

$$\boldsymbol{G}_1 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} \cos\beta & \sin\beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & \sin\beta \\ \sin\beta & 0 \end{pmatrix} + \begin{pmatrix} -\cos\beta & \sin\beta \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1+\cos\theta & 0 \\ 0 & 1+\cos\theta \end{pmatrix} \right),$$

$$\boldsymbol{G}_2 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} \cos\beta & \sin\beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1+\cos\theta & 0 \\ 0 & 1+\cos\theta \end{pmatrix} + \begin{pmatrix} -\cos\beta & \sin\beta \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & -\sin\beta \\ -\sin\beta & 0 \end{pmatrix} \right),$$

$$\boldsymbol{G}_3 = \frac{1}{2\sqrt{2}} \left(\begin{pmatrix} \cos\beta & \sin\beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -\sin\beta \\ -\sin\beta & 0 \end{pmatrix} + \begin{pmatrix} -\cos\beta & \sin\beta \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1-\cos\theta & 0 \\ 0 & 1-\cos\theta \end{pmatrix} \right).$$

The result is thus a two-parametric family of refinement masks, each of which has four nonzero coefficients.

For any value of parameters β and θ, the resulting polyphase matrix satisfies the necessary conditions (2.4) and (2.5) and the symmetry constraint (3.5). If the refinement equations (2.1) have a solution in L^2, the resulting multiwavelets are orthogonal, symmetric, and compactly supported. This needs to be accomplished by a suitable choice of the parameter values. By a clever choice of the parameter values, other desirable properties of multiwavelets may also be achieved simultaneously. We used MATLAB to search numerically for the values that yield the maximal smoothness of scaling and wavelet functions by minimizing the value of the largest eigenvalue of the

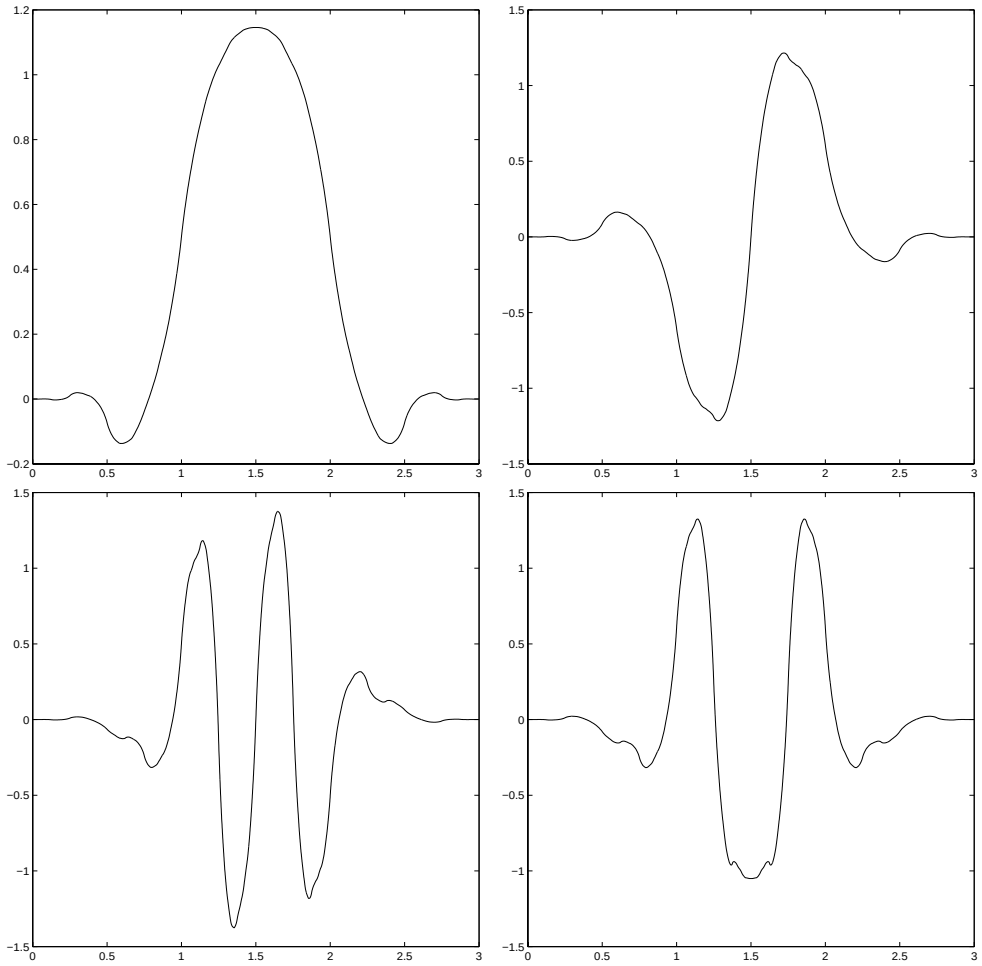


FIG. 8.1. *Scaling (top) and wavelet (bottom) functions for the parameter values* $\theta = 0.2756$, $\beta = 4.6259$.

transition operator that does not fall into the dyadic sequence 1, 1/2, 1/4, ... The values we have found are

$$\theta = 0.2756, \qquad \beta = 4.6259.$$

In this case, the value of the largest nondyadic eigenvalue of the transition operator is $0.0864 \approx 4^{-1.7816}$, which means that the resulting functions belong to the Sobolev space $H^{1.7816}$, and therefore they also belong to C^1 and, in other words, are continuously differentiable. They are depicted in the Figure 8.1. The associated refinement masks are

$$H_0 = \begin{pmatrix} 0.0133 & 0.0962 \\ 0.0050 & -0.0970 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0.6938 & 0.0962 \\ -0.6995 & 0.0359 \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 0.6938 & -0.0962 \\ 0.6995 & 0.0359 \end{pmatrix}, \quad H_3 = \begin{pmatrix} 0.0133 & -0.0962 \\ -0.0050 & -0.0970 \end{pmatrix},$$

$$G_0 = \begin{pmatrix} -0.0970 & -0.0050 \\ -0.0962 & 0.0133 \end{pmatrix}, \quad G_1 = \begin{pmatrix} -0.0359 & -0.6995 \\ 0.0962 & -0.6938 \end{pmatrix},$$

$$G_2 = \begin{pmatrix} 0.0359 & -0.6995 \\ 0.0962 & 0.6938 \end{pmatrix}, \quad G_3 = \begin{pmatrix} 0.0970 & -0.0050 \\ -0.0962 & -0.0133 \end{pmatrix}.$$

Acknowledgments. The author wants to thank the University of St. Thomas for supporting this work through an internal grant, and Jaroslav Kautsky of the Flinders University of South Australia for inviting her to come to Australia to work on this and other related projects and for covering her travel expenses through an ARC grant.

REFERENCES

[1] A. Cohen, *Biorthogonal wavelets*, in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, ed., Academic Press, New York, 1992, pp. 123–152.
[2] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.
[3] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
[4] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.
[5] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Fractal functions and wavelet expansions based on several functions*, J. Approx. Theory, 78 (1994), pp. 373–401.
[6] P. N. Heller and R. O. Wells, Jr., *The spectral theory of multiresolution operators and applications*, in Wavelets: Theory, Algorithms and Applications, C. K. Chui, L. Montefusco, and L. Puccio, eds., Academic Press, New York, 1994, pp. 13–32.
[7] Q. Jiang, *On the regularity of matrix refinable functions*, SIAM J. Math. Anal., 29 (1998), pp. 1157–1176.
[8] J. Kautsky and R. Turcajová, *Pollen product factorization and construction of higher multiplicity wavelets*, Linear Algebra Appl., 222 (1995), pp. 241–260.
[9] W. M. Lawton, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys., 31 (1990), pp. 1898–1901.
[10] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases for $L^2(\mathbb{R})$*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–87.
[11] D. Pollen, *Parametrization of Compactly Supported Wavelets*, Tech. report, AWARE, Inc., 1989.
[12] A. K. Soman, P. P. Vaidyanathan, and T. Q. Nguyen, *Linear phase paraunitary filter banks: Theory, factorizations and designs*, IEEE Trans. Signal Process., 41 (1993), pp. 3480–3496.
[13] V. Strela, *Multiwavelets: Regularity, orthogonality, and symmetry via two-scale similarity transform*, Stud. Appl. Math., 98 (1997), pp. 335–354.
[14] R. Turcajová, *Factorizations and construction of linear phase paraunitary filter banks and higher multiplicity wavelets*, Numer. Algorithms, 8 (1994), pp. 1–25.
[15] R. Turcajová, *Hermite spline multiwavelets for image modeling*, in Wavelet Applications V, Proc. SPIE 3391, SPIE, Bellingham, WA, 1998, pp. 45–56.
[16] R. Turcajová and J. Kautsky, *Shift products and factorizations of wavelet matrices*, Numer. Algorithms, 8 (1994), pp. 27–54.
[17] R. Turcajová and J. Kautsky, *Block Toeplitz-like operators and multiwavelets*, in Wavelet Applications for Dual Use, Proc. SPIE 2491, H. Szu, ed., SPIE, Bellingham, WA, 1995, pp. 957–967.
[18] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
[19] 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.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 551–573

PERTURBATION OF EIGENVALUES FOR MATRIX POLYNOMIALS VIA THE BAUER–FIKE THEOREMS*

ERIC KING-WAH CHU†

Abstract. In earlier papers, the Bauer–Fike technique [F. L. Bauer and C. T. Fike, *Numer. Math.*, 2 (1960), pp. 137–144] was applied to the eigenvalue problem $A\mathbf{x} = \lambda\mathbf{x}$ [E. K.-W. Chu, *Numer. Math.*, 49 (1986), pp. 685–691] and the generalized eigenvalue problem $A\mathbf{x} = \lambda B\mathbf{x}$ [E. K.-W. Chu, *SIAM J. Numer. Anal.*, 24 (1987), pp. 1114–1125]. General multiple eigenvalues were dealt with and perturbation results were obtained for individual as well as clusters of eigenvalues. In this paper, we shall generalize the technique to the eigenvalue problem for matrix polynomials. Multiple eigenvalues for monic as well as regular matrix polynomials will be considered.

Key words. Bauer–Fike theorem, condition number, eigenvalue problem, lambda-matrix, perturbation analysis, matrix polynomial, quadratic eigenvalue problem, second-order system

AMS subject classifications. 15A12, 15A18, 15A22, 65F15, 65F18, 65F35, 65G99

DOI. 10.1137/S0895479802217928

1. Introduction.

The aim of this paper is to apply the Bauer–Fike technique [1] to prove some perturbation results for the matrix polynomial eigenvalue problem (MPEVP)

$$L(\lambda)\mathbf{x} = \mathbf{0}, \quad \mathbf{x} \neq \mathbf{0}, \tag{1.1}$$

with

$$L(\lambda) \equiv \sum_{j=0}^{l} A_j \lambda^j, \quad A_j \in \mathcal{C}^{n\times n}, \quad \lambda \in \mathcal{C}.$$

In this paper, $\mathcal{R}$ and $\mathcal{C}$ denote, respectively, the sets of real and complex numbers.

Let $L(\lambda)$ be *regular*, i.e., $\det L(\lambda)$ is not identically zero. Also, let the perturbed matrix polynomial be denoted by

$$\widetilde{L}(\lambda) \equiv \sum_{j=0}^{l} \widetilde{A}_j \lambda^j = L(\lambda) + \delta L(\lambda), \quad \widetilde{A}_j = A_j + \delta A_j \in \mathcal{C}^{n\times n}.$$

The eigenvalues of $L(\lambda)$ in (1.1) are the roots of the characteristic polynomial $\det L(\lambda)$. When $\det A_l = 0$, $\mu = 0$ is a root of the *modified* matrix polynomial

$$\widehat{L}(\mu) \equiv \sum_{j=0}^{l} A_j \mu^{l-j} = \mu^l L(\mu^{-1}),$$

and the corresponding eigenvalue $\lambda = \mu^{-1}$ of $L(\lambda)$ is considered conventionally to be infinite.

*Received by the editors June 11, 2002; accepted for publication (in revised form) by N. J. Higham June 2, 2003; published electronically November 14, 2003.
http://www.siam.org/journals/simax/25-2/21792.html
†School of Mathematical Sciences, Building 28, Monash University, VIC 3800, Australia (eric.chu@sci.monash.edu.au).

In this paper, $\widehat{L}(\mu)$ and $L(\lambda)$ play equally important roles. If $A_l = I_n$, the matrix polynomial $L(\lambda)$ is described as *monic*. Otherwise, $L(\lambda)$ is a *regular* matrix polynomial.

To make our results accessible to nonexperts as early as possible, the main result is Theorem 4.2 is quoted below, in abbreviated form:

> For a regular matrix pencil $L(\lambda)$ and its perturbation $\tilde{L}(\lambda)$, we have the perturbation result
>
> $$s_{(\alpha,\beta)} \leq \max\{\theta_2, \theta_2^{1/p}\}, \quad \theta_2 \equiv c_1 \mathcal{F} \kappa \Delta,$$
>
> where $s_{(\alpha,\beta)}$ is the spectral variation measuring the distance between a perturbed eigenvalue (α, β) and its nearest neighbor in the spectrum of $L(\lambda)$, c_1 and $\mathcal{F}$ are constants, κ is the product of norms of the left- and right-eigenvector matrices, Δ is the size of the perturbation, and p is the maximum dimension of the Jordan blocks associated with $L(\lambda)$.

The ordinary eigenvalue problem (OEVP)

$$A\mathbf{x} = \lambda \mathbf{x}$$

and the generalized eigenvalue problem (GEVP)

$$\beta A\mathbf{x} = \alpha B\mathbf{x} \tag{1.2}$$

(with $\lambda = \alpha/\beta$ or (α, β)) are special cases of the MPEVP for matrix polynomials of degree one. The OEVP involves a monic matrix polynomial and, in general, the GEVP involves a regular matrix polynomial. The corresponding Bauer–Fike theorems, for individual nondefective and defective as well as clusters of eigenvalues, have been proven in earlier papers by the author [2, 3].

In [3], perturbation results were produced for four different cases, when finite or infinite eigenvalues were perturbed to zero or nonzero eigenvalues. A more refined argument for regular matrix polynomials in this paper eliminates the need to distinguish the four cases and simplifies the results in [3] as well as those in this paper. Perturbation results for two different distance metrics are produced (Corollary 4.3) and the one using the chordal metric [10, 21, 22, 23] is similar to the perturbation results in [3].

The GEVP in (1.2) is written in a form symmetric in A and B. This reflects similar symmetry in the QZ algorithm [12, p. 403] and is important in the development of the theory in [3] and this paper. Ordered pairs (α, β) are equivalent iff they have the same ratios and the equivalence classes of these ordered pairs constitute the spectrum of (A, B). The same generalized representation of classical eigenvalues will be used for regular matrix polynomials later. Thus, (α, β) is an eigenvalue of the matrix polynomial L iff (β, α) is an eigenvalue of the modified matrix polynomial $\widehat{L}$ and classical infinite eigenvalues are represented by $(\alpha, \beta) = (1, 0)$.

For convenience, we may select a representative $(\alpha_\circ, \beta_\circ)$ and its perturbation (α, β) from their corresponding equivalence classes so that

$$|\alpha_\circ|^2 + |\beta_\circ|^2 = 1 = |\alpha|^2 + |\beta|^2, \quad \beta_\circ, \beta \geq 0. \tag{1.3}$$

However, other scaling schemes for (α, β) are sometimes necessary, as in sections 3 and 4. An related important trick is to concentrate on the subspectrum of $L(\lambda)$ inside the unit circle, with the complimentary subspectrum considered via $\widehat{L}(\mu)$. This eliminates the need to consider any eigenvalue outside the unit circle.

The more general result in this paper reduces to those in [1, 2, 3] for OEVPs and GEVPs. For the theory of matrix polynomials and its applications, see [11, 16] for details. For a complete treatment of matrix perturbation theory, consult [24].

The Bauer–Fike technique allows large perturbations, yielding one error bound and its corresponding condition number for the whole spectrum. This is the nature of the technique that it does not distinguish between ill-conditioned and well-conditioned eigenvalues. This is the price we have to pay for allowing perturbations of arbitrary size, with well-conditioned eigenvalues easily perturbed to be ill-conditioned and vice versa. When the perturbations are (asymptotically) small, error bounds and condition numbers for individual or clusters of eigenvalues can be obtained.

1.1. Quadratic eigenvalue problems. The results in this paper first appeared in a technical report [4] in 1992. At the time, we thought the results were of negligible interest, due to the lack of applications. The situation has changed somewhat since then. Tremendous interest for the $l = 2$ case has been shown in the review paper [28] by Tisseur and Meerbergen and in the references therein. Issues related to errors, conditioning, and applications have been investigated in [8, 26, 27], showing consistent results to those here (see Example 2 in section 7.2; see also related results in [18]). The author has also applied the perturbation results to feedback control of second-order systems (modeled by quadratic matrix polynomials) in [5, 7]. In the search of "optimal" controllers [6], "well-conditioned" matrix polynomials have to be constructed from (partially) known spectral information. Consequently, it is important to understand the conditioning of the spectra of matrix polynomials. Note that similar investigation can be performed via linearizations of matrix polynomials, provided that the structures of the perturbations are considered [14]. However, it is more natural to consider matrix polynomials directly, especially when the coefficient matrices (the mass, damping and stiffness matrices [15]) have important engineering interpretation.

1.2. Plan of paper. In section 2, we quote some elementary results on standard triples, resolvent triples, and resolvent forms for matrix polynomials. Section 3 contains some preliminary results, e.g., upper bounds of $\|(\alpha\Lambda_\beta - \beta\Lambda_\alpha)^{-1}\|$ with $(\Lambda_\alpha, \Lambda_\beta)$ in Jordan or Kronecker canonical forms. These results then lead to the Bauer–Fike theorems for monic and regular matrix polynomials in section 4. Section 5 contains some perturbation results in terms of the residual vector $\mathbf{r} \equiv L(\lambda)\mathbf{x}$ ($(\lambda, \mathbf{x})$ is an approximate eigensolution), the perturbation of individual or clusters of eigenvalues, asymptotic perturbation, and condition numbers.

For general matrix polynomials, perturbation results could be obtained by applying known Bauer–Fike theorems [1, 2, 3] to the corresponding linearizations (see [11, pp. 11–15]). For monic matrix polynomials, we show in section 6 that the error bounds calculated via linearizations (with the structure of the perturbation ignored) are no better in general and are worse in some situations, when compared with those in this paper (with the error bounds in terms of the residual $\mathbf{r}$ being sharper).

Section 7 concludes the paper with two numerical examples, one comparing the condition numbers from section 5.4 with those in [8].

We shall denote spectra by $\sigma(\cdot)$ and the Hermitian or complex conjugate transpose by $(\cdot)^*$. Matrix norms are Hölder norms if unspecified. Notation from [11] will be used as much as possible, especially for discussions concerning matrix polynomials.

2. Matrix polynomials. Some elementary results are quoted from [11, chapters 1, 2, 7, and 8] for later use.

For a regular matrix polynomial

$$L(\lambda) = \sum_{j=0}^{l} A_j \lambda^j, \quad A_l \neq 0, \quad \det A_l = 0, \tag{2.1}$$

$(X, T_1 \oplus T_2)$ denotes the corresponding *decomposable pair*. Here we have $X \equiv [X_1, X_2]$, $X_1 \in \mathcal{C}^{n\times m}$, $X_2 \in \mathcal{C}^{n\times(nl-m)}$, $T_1 \in \mathcal{C}^{m\times m}$, $T_2 \in \mathcal{C}^{(nl-m)\times(nl-m)}$. The matrix

$$S_{l-1} \equiv \text{ col } (X_1 T_1^i, X_2 T_2^{l-i-1})_{i=0}^{l-1} = \begin{bmatrix} X_1 & X_2 T_2^{l-1} \\ X_1 T_1 & X_2 T_2^{l-2} \\ \vdots & \vdots \\ X_1 T_1^{l-1} & X_2 \end{bmatrix} \tag{2.2}$$

is nonsingular and

$$\sum_{i=0}^{l} A_i X_1 T_1^i = 0, \qquad \sum_{i=0}^{l} A_i X_2 T_2^{l-i} = 0.$$

The finite and infinite parts of the spectrum $\sigma(L)$ are, respectively, represented in T_1 and T_2. A particularly useful choice will be

$$X_1 = X_F, \quad T_1 = J_F, \quad X_2 = X_\infty, \quad T_2 = J_\infty,$$

with J_F and J_∞ in Jordan form (or with $(\lambda I_m - J_F) \oplus (I_{nl-m} - \lambda J_\infty)$ in Kronecker canonical form). The matrix pair (X_F, J_F) (or (X_∞, J_∞)) is called a finite (or infinite) Jordan pair of $L(\lambda)$.

The matrix polynomial $L(\lambda)$ then has the linearization

$$T(\lambda) = (I_m \lambda - T_1) \oplus (T_2 \lambda - I_{nl-m}) \tag{2.3}$$

and the companion linearization

$$C_L(\lambda) = \begin{bmatrix} I_n & & & \\ & \ddots & & 0 \\ & & \ddots & \\ & 0 & & I_n \\ & & & & A_l \end{bmatrix} \lambda + \begin{bmatrix} 0 & -I_n & 0 & \cdots & 0 \\ \vdots & \ddots & -I_n & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & -I_n \\ A_0 & A_1 & \cdots & \cdots & A_{l-1} \end{bmatrix}. \tag{2.4}$$

For the regular matrix polynomial $L(\lambda)$, we have the following resolvent form [11, Theorem 7.7, pp. 195–197] for $\lambda \notin \sigma(L)$:

$$L^{-1}(\lambda) = X T(\lambda)^{-1} Z, \qquad Z \equiv [I_m \oplus T_2^{l-1}] \begin{bmatrix} S_{l-2} \\ V \end{bmatrix}^{-1} [0, \ldots, 0, I_n]^T, \tag{2.5}$$

with

$$S_{l-2} \equiv \text{ col } (X_1 T_1^i, X_2 T_2^{l-i-1})_{i=0}^{l-2}, \qquad V = \left[A_l X_1 T_1^{l-1}, \ -\sum_{i=0}^{l-1} A_i X_2 T_2^{l-1-i} \right].$$

Jordan pairs or the companion linearization $C_L(\lambda)$ can be used to construct $T(\lambda)$. With the latter, we have

$$L^{-1}(\lambda) = [I_n, 0, \ldots, 0] C_L(\lambda)^{-1} [0, \ldots, 0, I_n]^T. \tag{2.6}$$

The triple $([X_1, X_2], T_1 \oplus T_2, Z)$ is called a *resolvent triple* [11, p. 219]. Here, Z is dependent on and can be constructed from X_i and T_i $(i = 1, 2)$. Note that in general Z is not exactly the same as the left-eigenvector matrix Y, where $Y = [Y_1, Y_2]$ and

$$\sum_{i=0}^{l} A_i^H Y_1 T_1^i = 0, \qquad \sum_{i=0}^{l} A_i^H Y_2 T_2^{l-i} = 0.$$

The matrices Y and Z^T are obviously related, with columns corresponding to the same eigenvalue spanning the same subspace. When the eigenvalues in $T_1 \oplus T_2$ are simple, the vectors $\mathbf{y}_i$ and $\mathbf{z}_j$, corresponding to λ_j in Y and Z^T, respectively, are parallel.

It is also important to note that some of the $\mathbf{z}_j$ may be annihilated by the zeros in T_2 (corresponding to some infinite eigenvalues) in (2.5). In such cases, $\mathbf{z}_j$ can still be retrieved from $\widehat{L}$, interchanging the roles of T_1 and T_2. (See one of such calculations in Example 2 in section 7.2.)

For the monic case, the results are simplified by omitting (X_2, T_2) (or (X_∞, T_∞)). Results for the case with $A_l \neq 0$, $\det A_l \neq 0$ in (2.1) are similar to those for monic matrix polynomials, after minor modifications (see [16, chapter 14]). The resolvent triple is then called the *standard triple*. For $\lambda \notin \sigma(L)$, the resolvent form (2.5) reduces to [11, eqn. (2.16), Theorem 2.4, p. 58]:

$$L(\lambda)^{-1} = XT(\lambda)^{-1}Z, \qquad T(\lambda) \equiv (I_n\lambda - T_1), \tag{2.7}$$

with a simpler structure in Z:

$$Z = \begin{bmatrix} X \\ XT_1 \\ \vdots \\ XT_1^{l-1} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ I_n \end{bmatrix}.$$

As in [9] and for both the regular and monic matrix polynomial cases, it is possible to have other useful rearrangements of $T = T_1 \oplus T_2$, e.g., $T = T_I \oplus T_O$ so that T_I has eigenvalues on or inside a circle of a given radius and T_O represents those outside. The idea of distinguishing eigenvalues on/inside and outside a given circle is important in sections 4 and 5.

3. Preliminary results. We now build up the machinery for the proofs of the main results in section 4.

3.1. Monic matrix polynomials. For monic matrix polynomials, we expand $\widetilde{L}(\lambda)$ such that

$$\widetilde{L}(\lambda) \equiv L(\lambda) + \delta L(\lambda) = L(\lambda)\{I_n + L(\lambda)^{-1}\delta L(\lambda)\}.$$

Using the resolvent form (2.7), we obtain

$$\widetilde{L}(\lambda) = L(\lambda)\{I_n + XT(\lambda)^{-1}Z\ \delta L(\lambda)\}.$$

As in [2, 3], the matrix

$$\widetilde{M_1} \equiv XT(\lambda)^{-1}Z\ \delta L(\lambda) \tag{3.1}$$

has a norm greater than unity when λ is an eigenvalue of $\widetilde{L}(\lambda)$.

We have the following lemma on the upper bound of $\|T(\lambda)^{-1}\|$.

LEMMA 3.1. *Let $T(\lambda)$ be constructed using a Jordan triple, i.e.,*

$$T(\lambda) = \lambda I_{nl} - J, \quad J = J_1 \oplus \cdots \oplus J_r$$

and J_i is the Jordan block for the eigenvalue λ_i. When $\lambda \notin \sigma(T)$, we have

$$\|T(\lambda)^{-1}\| \le c_1 \cdot \max\left\{|z|^{-1}, |z|^{-p}\right\}, \quad c_1 \equiv \min\left\{\frac{2p+1}{p+1}, p\right\}. \tag{3.2}$$

Here

$$z \equiv \min_i |\lambda - \lambda_i| \neq 0$$

with p is the maximum dimension of the Jordan blocks J_i $(i = 1, \ldots, r)$.

Proof. For Hölder norms, we have

$$\|T(\lambda)^{-1}\| \le \max_i \|M_i\|, \quad M_i \equiv (\lambda I - J_i)^{-1}.$$

Let the maximum in the right-hand side in the inequality above occur at $i = k$ and denote $(\lambda - \lambda_k)$ by z_k. The matrix $M_k \in \mathcal{C}^{p_k \times p_k}$ denotes the Toeplitz matrix $(\lambda I_{p_k} - J_k)^{-1}$ or

$$\begin{bmatrix} z_k & -1 & & & \\ & z_k & -1 & & 0 \\ & & \ddots & \ddots & \\ & 0 & & z_k & -1 \\ & & & & z_k \end{bmatrix}^{-1} = \begin{bmatrix} z_k^{-1} & z_k^{-2} & z_k^{-3} & \cdots & z_k^{-p_k} \\ & z_k^{-1} & z_k^{-2} & \cdots & z_k^{-p_k+1} \\ & & z_k^{-1} & \ddots & \vdots \\ & 0 & & \ddots & z_k^{-2} \\ & & & & z_k^{-1} \end{bmatrix}.$$

Note that $\lambda \notin \sigma(T)$, implying that $z_k \neq 0$.

It is easy to prove for the 1-, ∞-, 2-, and F-norms that

$$\|M_k\| \le p_k \cdot \max\left\{|z_k|^{-1}, |z_k|^{-p_k}\right\}. \tag{3.3}$$

The result in (3.2), with $c_1 = p$, is then obtained by replacing p_k and $|z_k|$, respectively, with p and z in (3.3). We shall show that a sharper bound can be obtained when $p > 1$, with the smaller $c_1 = (2p+1)/(p+1) < 2$ (thus $c_1 = 2$ yields a simpler but slightly worse result).

With $M_k^{-1} = z_k(I_{p_k} - z_k^{-1}N)$, $N^{p_k} = 0$, $p_k > 1$, we have

$$M_k = z_k^{-1} \sum_{i=0}^{p_k-1} z_k^{-i} N^i, \quad \|M_k\| \le \eta^{-1} \equiv |z_k|^{-1} \sum_{i=0}^{p_k-1} |z_k|^{-i}.$$

For simplicity, let $x = |z_k|$ and $m = p_k$, the above definition of η leads to the polynomial

$$P_m(x) \equiv x^m - \eta(1 + x + \cdots + x^{m-1}).$$

Descartes' sign rule (*La Géométrie* 1637 [13]) then implies that $P_m(x)$ has at most one positive real root. As $P_m(0) = -\eta < 0$ and $P_m(x) > 0$ as $x \to \infty$, any positive number x^* for which $P_m(x^*) > 0$ is an upper bound of the unique real positive root of P_m. Simple inspection leads to the upper bound $x^* = c_1\eta$ when $\eta > 1$ and $x^* = c_1\eta^{1/m}$ when $\eta \le 1$, with $c_1 = (2m+1)/(m+1)$. With the upper bound of $x = |z_k|$ in terms of η, we can deduce an upper bound of η^{-1} in terms of $|z_k|$, which implies (3.2). ∎

Note. The coefficient $c_1 = 1$ when $p = 1$ and $c_1 = (2p+1)/(p+1) < 2$ when $p > 1$. The possibility of bounding M_k better was raised in [17] (with $c_1 = 2$) without proof. The result in Lemma 3.1 is a slight improvement, with c_1 dependent on p. Sharper bounds with slightly smaller c_1 can easily be found but the complexity of c_1 may not be justified by the slight improvement in the bounds.

3.2. Regular matrix polynomials. For regular matrix polynomials, we shall consider the "symmetrized" polynomials

$$L(\alpha,\beta) \equiv \sum_{j=0}^{l} A_j \alpha^j \beta^{l-j}$$

and, in the modified form,

$$\widehat{L}(\alpha,\beta) \equiv \sum_{j=0}^{l} A_j \alpha^{l-j} \beta^j.$$

Obviously $\widehat{L}(\alpha,\beta) = L(\beta,\alpha)$, but we shall retain the symbol $\widehat{L}(\beta,\alpha)$ to emphasize the use of the modified matrix polynomial $\widehat{L}$.

For $\lambda \equiv \alpha/\beta$ and $\mu \equiv \beta/\alpha$, we have

$$L(\alpha,\beta) = \beta^l L(\lambda), \qquad \widehat{L}(\beta,\alpha) = \alpha^l \widehat{L}(\mu). \tag{3.4}$$

We can expand $\widetilde{L}$ so that

$$\widetilde{L}(\alpha,\beta) = L(\alpha,\beta) + \delta L(\alpha,\beta) = L(\alpha,\beta)\{I_n + L(\alpha,\beta)^{-1}\delta L(\alpha,\beta)\}.$$

Using (3.4), we can rewrite the resolvent form in (2.5) as

$$L(\alpha,\beta)^{-1} = \beta^{-l} L(\lambda)^{-1} = \beta^{-l} X T(\lambda)^{-1} Z = \beta^{1-l} X T(\alpha,\beta)^{-1} Z,$$

with

$$T(\alpha,\beta) \equiv (\alpha I_m - \beta T_1) \oplus (\alpha T_2 - \beta I_{nl-m}) = \beta T(\lambda).$$

As in [3] and analogous to $\widetilde{M_1}$ in (3.1), the matrix

$$\widetilde{M_2} \equiv L(\alpha,\beta)^{-1}\delta L(\alpha,\beta) = \beta^{1-l} X T(\alpha,\beta)^{-1} Z \ \delta L(\alpha,\beta) \tag{3.5}$$

has norm greater than unity when (α,β) is an eigenvalue of $\widetilde{L}(\alpha,\beta)$.

Similar to Lemma 3.1, we have the following lemma on the upper bound of $\|T(\alpha,\beta)^{-1}\|$.

LEMMA 3.2. *Let $T(\alpha,\beta)$ be constructed using a resolvent triple with finite and infinite Jordan pairs,*

$$T(\alpha,\beta) \equiv T_1(\alpha,\beta) \oplus T_2(\alpha,\beta),$$

$$T_1 = \alpha I_m - \beta J_F, \quad J_F = J_1 \oplus \cdots \oplus J_{r_1},$$

$$T_2 = \alpha J_\infty - \beta I_{nl-m}, \quad J_\infty = J_{r_1+1} \oplus \cdots \oplus J_{r_1+r_2},$$

and J_i is the Jordan block for the eigenvalue (α_i,β_i).

Under the assumption that

$$|\beta| \geq \frac{1}{\sqrt{2}}, \tag{3.6}$$

$(\alpha, \beta) \notin \sigma(T)$*, we have*

$$\|T(\alpha,\beta)^{-1}\| \leq c_1 \cdot \max\left\{\widetilde{z}^{-1}, \widetilde{z}^{-p}\right\}, \quad c_1 \equiv \min\left\{\frac{2p+1}{p+1}, p\right\}. \tag{3.7}$$

Here

$$\widetilde{z} \equiv \min_i |\alpha\beta_i - \beta\alpha_i| \neq 0$$

with p *is the maximum dimension of the Jordan blocks in* J_F *or* J_∞*.*

Remark. In the above lemma, we have chosen $\alpha_i \neq 0$, $\beta_i = 1$ for $i = 1, \ldots, r_1$. These eigenvalues do not satisfy the scaling scheme in (1.3) but can be rescaled and will not affect the development of our results. We can modify (I, J_F) and (J_∞, I) to eliminate this violation of the scaling scheme as in [3], but we do not want to introduce any extra notation.

Proof. For Hölder norms, we have

$$\|T(\alpha,\beta)^{-1}\| \leq \max_i \|M_i\|,$$

with

$$M_i \equiv (\alpha I - \beta J_i)^{-1} \quad \text{when } i \in [1, r_1]$$

or

$$M_i \equiv (\alpha J_i - \beta I)^{-1} \quad \text{when } i \in [r_1 + 1, r_2].$$

We shall consider the former case ($\beta_i = 1$), and the proof for the latter case ($\beta_i = 0$) is similar.

Let the maximum in the right-hand side in the above inequality occur at $i = k \in [1, r_1]$ and denote $(\alpha\beta_k - \beta\alpha_k)$ by $\widetilde{z}_k$; the matrix $M_k \in \mathcal{C}^{p_k \times p_k}$ denotes the Toeplitz matrix $(\alpha I - \beta J_k)^{-1}$ or

$$\begin{bmatrix} \widetilde{z}_k & -\beta & & & \\ & \widetilde{z}_k & -\beta & & 0 \\ & & \ddots & \ddots & \\ & 0 & & \widetilde{z}_k & -\beta \\ & & & & \widetilde{z}_k \end{bmatrix}^{-1} = \begin{bmatrix} \widetilde{z}_k^{-1} & \beta\widetilde{z}_k^{-2} & \beta^2\widetilde{z}_k^{-3} & \cdots & \beta^{p_k-1}\widetilde{z}_k^{-p_k} \\ & \widetilde{z}_k^{-1} & \beta\widetilde{z}_k^{-2} & \cdots & \beta^{p_k-2}\widetilde{z}_k^{-p_k+1} \\ & & \widetilde{z}_k^{-1} & \ddots & \vdots \\ & 0 & & \ddots & \beta\widetilde{z}_k^{-2} \\ & & & & \widetilde{z}_k^{-1} \end{bmatrix}.$$

From (1.3) and (3.6), we can deduce that $|\beta| \leq 1$ and λ is on or inside the unit circle. Similar to the proof in Lemma 3.1, we then have

$$\|M_k\| \leq c_1 \cdot \max\left\{|\widetilde{z}_k|^{-1}, |\widetilde{z}_k|^{-p_k}\right\}. \tag{3.8}$$

We obtain (3.7) by replacing p_k and $|\widetilde{z}_k|$, respectively, with p and $\widetilde{z}$ in (3.8). ∎

We also require a third lemma.

LEMMA 3.3. *The function*

$$f(x) = x^2 \sum_{j=0}^{l} \left[\frac{1-x^2}{x^2}\right]^j, \quad x \in \mathcal{I} \equiv \left[\frac{1}{\sqrt{2}}, 1\right], \quad l = 1, 2, \ldots,$$

is positive and monotonically nonincreasing. The maximum of f in $\mathcal{I}$ satisfies

$$\max_{\mathcal{I}} f = f\left(\frac{1}{\sqrt{2}}\right) = \frac{l+1}{2} \leq l.$$

Proof. The function f is a finite sum of $(l+1)$ nonnegative terms with the first term $x^2 \geq \frac{1}{2}$, so f is positive.

The function is obviously differentiable and

$$\frac{df}{dx} = -2 \sum_{j=2}^{l} j x^{-3} (x^{-2} - 1)^{j-1}.$$

The derivative is nonpositive as both x^{-3} and $(x^{-2} - 1)$ are nonnegative. So f is monotonically nonincreasing and the maximum of f in $\mathcal{I}$ is equal to the value of f at the left end-point of $\mathcal{I}$. ∎

4. The Bauer–Fike theorems. We prove the main theorems of the paper in this section.

THEOREM 4.1. *Consider a monic matrix polynomial*

$$L(\lambda) \equiv I_n \lambda^l + \sum_{j=0}^{l-1} A_j \lambda^j$$

and its perturbation

$$\widetilde{L}(\lambda) \equiv I_n \lambda^l + \sum_{j=0}^{l-1} \widetilde{A}_j \lambda^j, \quad \widetilde{A}_j \equiv A_j + \delta A_j \ (j = 0, \ldots, l-1).$$

Let (X, J, Z) be a Jordan triple for L. For $\lambda_i \in \sigma(L)$ and $\lambda \in \sigma(\widetilde{L})$, the spectral variation of $\widetilde{L}$ from L is defined as

$$s_L(\widetilde{L}) \equiv \max_{\lambda} \{s_\lambda\}, \qquad s_\lambda \equiv \min_{i} \{|\lambda - \lambda_i|\}.$$

Let p be the maximum dimension of the Jordan blocks in J.

Then for $\|\cdot\|_\tau$ $(\tau = 1, 2, \infty)$, we have

$$\gamma^{-1} s_\lambda \leq \max\{\theta_1, \theta_1^{1/p}\}, \quad \theta_1 \equiv c_1 \kappa \Delta, \quad c_1 \equiv \min\left\{\frac{2p+1}{p+1}, p\right\}, \tag{4.1}$$

where

$$\gamma \equiv c_2 \sqrt{\sum_{j=0}^{l-1} |\lambda|^{2j}}, \quad \kappa \equiv \|X\| \cdot \|Z\|, \quad \Delta \equiv \|[\delta A_0, \ldots, \delta A_{l-1}]\|,$$

$c_2 = 1$ *for* $\tau = 1, 2$, *and* $c_2 = \sqrt{l}$ *for* $\tau = \infty$. *Also, we have*

$$\widetilde{\gamma}^{-1} s_L(\widetilde{L}) \leq \max\{\widetilde{\theta}_1, \widetilde{\theta}_1^{1/p}\}, \quad \widetilde{\theta}_1 \equiv c_1 \kappa \Delta, \quad \widetilde{\gamma} \equiv \max_{\lambda}\{\gamma\}. \tag{4.2}$$

Proof. From section 3, the matrix $\widetilde{M}_1$ in (3.1) satisfies

$$\|\widetilde{M}_1\| \geq 1 \;\Rightarrow\; \kappa \, \|\delta L(\lambda)\| \, \|T(\lambda)^{-1}\| \geq 1. \tag{4.3}$$

Applying the theory of matrix norms, we obtain

$$\|\delta L(\lambda)\| \leq \|[I_n, \lambda I_n, \ldots, \lambda^{l-1} I_n]\| \Delta \leq \gamma \Delta. \tag{4.4}$$

Substitute the upper bound for $\|T(\lambda)^{-1}\|$ from Lemma 3.1 and substitute (4.4) into (4.3); we have

$$\min\{s_\lambda, s_\lambda^p\} = \min\{z, z^p\} \leq c_1 \kappa \gamma \Delta$$

when $z \neq 0$. When $z = 0$, the left-hand side of (4.1) vanishes and the result is trivial.

The above inequality implies (4.1) after γ or $\gamma^{1/p}$ on the right-hand side has been shifted to the left-hand side, with the latter replaced by the larger γ. Note that $\gamma \geq 1$ from its definition in (4.1). The result in (4.2) then follows from (4.1). ∎

With $n = 1$, the Jordan blocks become trivial and $p = \kappa = 1$. With similar notation, Theorem 4.1 gives rise to the following corollary for scalar polynomials.

COROLLARY 4.1. *Consider a monic scalar polynomial*

$$L_l(x) = x^l + a_{l-1} x^{l-1} + \cdots + a_1 x + a_0$$

and its perturbation

$$\tilde{L}_l(x) = x^l + \tilde{a}_{l-1} x^{l-1} + \cdots + \tilde{a}_1 x + \tilde{a}_0$$

with $\tilde{a}_j = a_j + \delta a_j$ $(j = 0, 1, \ldots, l-1)$. *Let* λ_j *be a root of* L_l *and* λ *be a root of* $\tilde{L}_l$.

Then for $\|\cdot\|_\tau$ $(\tau = 1, 2, \infty)$, *we have*

$$\gamma^{-1} s_\lambda \leq \Delta,$$

where

$$\gamma \equiv c_2 \sqrt{\sum_{j=0}^{l-1} |\lambda|^{2j}}, \quad \Delta \equiv \|[\delta a_0, \ldots, \delta a_{l-1}]\|,$$

$c_2 = 1$ *for* $\tau = 1, 2$, *and* $c_2 = \sqrt{l}$ *for* $\tau = \infty$. *Also, we have*

$$\tilde{\gamma}^{-1} s_L(\tilde{L}) \leq \Delta, \quad \tilde{\gamma} \equiv \max_{\lambda}\{\gamma\}.$$

Notice that for the inequalities in Corollary 4.1, we have the corresponding $c_1 = p = 1$ (in Lemma 3.1) for the scalar case.

The quantities γ and $\widetilde{\gamma}$ are dependent on the perturbed eigenvalue λ, and their appearance in the perturbation results (4.1) and (4.2) has to be interpreted carefully.

We have the following comments:

(i) As $\gamma \geq 1$, we can consider the left-hand side of (4.1) a new distance metric measuring the perturbation error. The quantity $\widetilde{\gamma}^{-1} s_L(\widetilde{L})$ in (4.2) can be considered to be a scaled version of the classical spectral variation.

(ii) If λ is on or inside the unit circle, we have $\gamma \leq \sqrt{l}$. From the proof of Theorem 4.1, the γ^{-1} and $\widetilde{\gamma}^{-1}$ terms on the left-hand sides of (4.1) and (4.2) can be replaced by the constants $\sqrt{l}$ ($\tau = 1, 2$) or l ($\tau = \infty$) in θ_1 on the right-hand sides.

(iii) Another alternative is to replace $\gamma\Delta$ by $\max_{|\lambda|\leq 1} \|\delta L(\lambda)\|$ in the definition of θ_1 in (4.1).

(iv) If λ is outside the unit circle but $\det \widetilde{A}_0 \neq 0$, we can consider instead the monic matrix polynomial

$$\mu^l I_n + \sum_{j=1}^{l} \widetilde{A}_0^{-1} \widetilde{A}_j \mu^{l-j}$$

and its eigenvalue $\mu = \lambda^{-1}$ is inside the unit circle.

(v) We may consider $L(\lambda)$ and its perturbations as general regular matrix polynomials, as in Theorem 4.2.

(vi) As λ and thus γ and $\widetilde{\gamma}$ are normally unknown, only (i), (iii), and (v) above are practical interpretations of the perturbation results concerning γ.

Note that the bounds in (4.1) and (4.2) are not optimal as inequalities for norms have been applied. A balance has to be maintained between the simplicity of the results and the sharpness of the bounds. For instance, the quantity γ could have been defined to be smaller but more complicated expressions for the 1-norm and the ∞-norm would result. Different techniques from those in Lemmas 3.1 and 3.2 can be applied in bounding $T(\lambda)^{-1}$ and $T(\alpha, \beta)^{-1}$ (e.g., the Henrici-type results in [24, Theorems 1.9 and 1.12]). Note also that X in the Jordan triple, and thus κ, are not uniquely defined and can also be minimized.

Expanding on (iv) above, our theory for monic matrix polynomials in Theorem 4.1 is a special case of Theorem 4.2 below. The special case is worthy of consideration because some readers may be interested only in the results for monic matrix polynomials, which are easier to develop and understand. The monic case also shares some common features with the regular case, e.g., the quantity γ and the distinction of eigenvalues on/inside or outside the unit circle, and is a good introduction to the regular case.

THEOREM 4.2. *Consider a regular matrix polynomial*

$$L(\alpha, \beta) \equiv \sum_{j=0}^{l} A_j \alpha^j \beta^{l-j}$$

and its perturbation

$$\widetilde{L}(\alpha, \beta) \equiv \sum_{j=0}^{l} \widetilde{A}_j \alpha^j \beta^{l-j}, \quad \widetilde{A}_j \equiv A_j + \delta A_j \ (j = 0, \ldots, l).$$

Let (X, T, Z) be a resolvent triple for L constructed using some finite and infinite Jordan pairs J_F and J_∞. For $(\alpha_i, \beta_i) \in \sigma(L)$ and $(\alpha, \beta) \in \sigma(\widetilde{L})$, the spectral variation of $\widetilde{L}$ from L is defined as

$$s_L(\widetilde{L}) \equiv \max_{(\alpha,\beta)} \{s_{(\alpha,\beta)}\}, \qquad s_{(\alpha,\beta)} \equiv \min_i \{|\alpha\beta_i - \beta\alpha_i|\}.$$

Let p be the maximum dimension of the Jordan blocks in J_F or J_∞.

Then for $\|\cdot\|_\tau$ $(\tau = 1, 2, \infty)$, we have

$$s_{(\alpha,\beta)} \le \max\{\theta_2, \theta_2^{1/p}\}, \quad \theta_2 \equiv c_1 \mathcal{F} \kappa \Delta, \tag{4.5}$$

where c_1 and c_2 are defined as in Theorem 4.1*,*

$$\mathcal{F} \equiv c_2 \sqrt{\frac{l+1}{2}}, \quad \kappa \equiv \|X\| \cdot \|Z\|, \quad \Delta \equiv \|[\delta A_0, \ldots, \delta A_l]\|.$$

Also, we have

$$s_L(\widetilde{L}) \le \max\{\theta_2, \theta_2^{1/p}\}. \tag{4.6}$$

Proof. From section 3, the matrix $\widetilde{M}_2$ in (3.5) satisfies

$$\|\widetilde{M}_1\| \ge 1 \;\Rightarrow\; |\beta|^{1-l} \kappa \, \|\delta L(\alpha,\beta)\| \, \|T(\alpha,\beta)^{-1}\| \ge 1. \tag{4.7}$$

Apply the theory of matrix norms to obtain

$$\|\delta L(\alpha,\beta)\| \le \|[\beta^l I_n, \alpha\beta^{l-1}\alpha I_n, \ldots, \alpha^{l-1}\beta I_n, \alpha^l I_n]\| \Delta \le \widehat{\gamma}\Delta, \tag{4.8}$$

where

$$\widehat{\gamma} \equiv c_2 \sqrt{\sum_{j=0}^{l} \alpha^{2j} \beta^{2(l-j)}}.$$

With the assumption in (3.6), the scaling schemes in (1.3), and Lemma 3.3, equation (4.8) implies

$$|\beta|^{1-l} \|\delta L(\alpha,\beta)\| \le \widehat{\gamma}\Delta \le c_2 \sqrt{f(\beta)}\Delta \le \mathcal{F}\Delta. \tag{4.9}$$

Substitute the upper bound for $\|T(\alpha,\beta)^{-1}\|$ from Lemma 3.2 and substitute (4.9) into (4.7); we arrive at

$$\min\left\{s_{(\alpha,\beta)}, s^p_{(\alpha,\beta)}\right\} = \min\{\widetilde{z}, \widetilde{z}^p\} \le \theta_2$$

when $\widetilde{z} \equiv \min_i |\alpha\beta_i - \beta\alpha_i| \ne 0$. Otherwise, the left-hand side of (4.5) vanishes and the trivial result follows.

Inequality (4.5), and in turn (4.6), then follows.

It remains to show that the assumption in (3.6) is unnecessary. When (3.6) is violated, we have

$$|\beta| < \frac{1}{\sqrt{2}},$$

implying that $|\alpha| \ge \frac{1}{\sqrt{2}}$. We can then consider, respectively, $\widehat{L}$ and (β, α) instead of L and (α, β). The role of β is now replaced by α and the argument in section 3 and section 4 as well as the earlier part of this proof will be valid. The results in (4.5) and (4.6) still hold as they are symmetric with respect to α and β. ∎

Note that $(\alpha,\beta)=(0,0)$ leads to $s_{(\alpha,\beta)}=0$ and trivial results in Theorem 4.2. However, we always have $(\alpha,\beta)\neq(0,0)$ for regular matrix polynomials, and sensible scaling of (α,β) will avoid the situation where $s_{(\alpha,\beta)}$ is arbitrarily small.

To further illustrate the symmetry between eigenvalues on/inside the unit circle and those outside, we assume from the scaling scheme in (1.3) that

$$(\alpha,\beta)=(\sin\phi,\cos\phi),\qquad (\alpha_\circ,\beta_\circ)=(\sin\phi_\circ,\cos\phi_\circ),$$

where (α,β) is perturbed from $(\alpha_\circ,\beta_\circ)$. We can easily prove that

$$s_{(\alpha,\beta)}=\max_{\phi}|\sin(\phi_\circ-\phi)|. \tag{4.10}$$

The measures $s_{(\alpha,\beta)}$ and $s_L(\widetilde{L})$ are thus invariant when we interchange α ($\alpha_\circ$) and β ($\beta_\circ$), or when ϕ and $\phi_\circ$ are replaced, respectively, by their complementary angles $(\pi/2-\phi)$ and $(\pi/2-\phi_\circ)$.

Finally we also have the following corollary.

COROLLARY 4.3. (i) *Let* $(\alpha,\beta)\neq(0,0)$. *Based on the assumptions and notation in Theorem* 4.2 *and using the chordal metric* [3, 23]

$$\rho\{(\alpha_\circ,\beta_\circ);(\alpha,\beta)\}\equiv\frac{|\alpha\beta_\circ-\beta\alpha_\circ|}{\sqrt{|\alpha_\circ|^2+|\beta_\circ|^2}\sqrt{|\alpha|^2+|\beta|^2}},$$

we have

$$\rho\{(\alpha_\circ,\beta_\circ);(\alpha,\beta)\}\leq\max\{\theta_2,\theta_2^{1/p}\}.$$

(ii) *Let* $(\alpha,\beta)\neq(0,0)$. *Define the new metric*

$$\rho_1\{(\alpha_\circ,\beta_\circ);(\alpha,\beta)\}\equiv\frac{|\alpha\beta_\circ-\beta\alpha_\circ|}{\nu_\circ\nu}$$

with $\nu_\circ\equiv\sqrt{|\alpha_\circ|^2+|\beta_\circ|^2}$ *and*

$$\nu\equiv\begin{cases}\sqrt{\sum_{j=0}^{l}|\alpha|^{2j}|\beta|^{2-2j}} & \text{if}\quad |\beta|\geq|\alpha|,\\[2ex] \sqrt{\sum_{j=0}^{l}|\beta|^{2j}|\alpha|^{2-2j}} & \text{if}\quad |\beta|<|\alpha|;\end{cases}$$

we have

$$\rho_1\{(\alpha_\circ,\beta_\circ);(\alpha,\beta)\}\leq\max\{\theta_3,\theta_3^{1/p}\}\ ,\quad \theta_3\equiv c_1c_2\kappa\Delta\nu_\circ^{-1}.$$

Proof. Result (i) involving ρ is a restatement of (4.5). For result (ii) involving ρ_1, we follow the same argument as in the proof of Theorem 4.2 but without using Lemma 3.3 to bound $f(|\beta|)$. The quantity $f(|\beta|)$ is then replaced by ν from (4.8) on.

We can show that ν is always positive and well defined. When $|\beta|<|\alpha|$, $|\alpha|$ is obviously positive. When $|\beta|\geq|\alpha|$, the fact that $(\alpha,\beta)\neq(0,0)$ implies $|\beta|>0$. ∎

Applying the scaling schemes in (1.3) to result (i) above, we have

$$\rho\{(\alpha_\circ,\beta_\circ);(\alpha,\beta)\}=|\alpha\beta_\circ-\beta\alpha_\circ|. \tag{4.11}$$

Result (ii) can be interpreted, analogous to (4.11), as

$$\rho_1\{(\alpha_\circ,\beta_\circ);(\alpha,\beta)\}=|\alpha\beta_\circ-\beta\alpha_\circ|$$

when the eigenvalues are scaled according to $\nu_\circ=1=\nu$.

5. Perturbation results. Here, some perturbation results are developed utilizing the Bauer–Fike theorems in section 4.

5.1. Residual vectors. Consider the perturbation of the eigenvalues in terms of the residual vector

$$\mathbf{r} \equiv L(\lambda_\circ)\mathbf{x}, \quad \lambda_\circ = \alpha_\circ/\beta_\circ. \tag{5.1}$$

Here $(\alpha_\circ, \beta_\circ)$, with $1 \geq \beta_\circ \neq 0$, denotes an approximate eigenvalue and $\mathbf{x}$ is the corresponding approximate eigenvector. For simplicity, we assume here that both (α, β) and $(\alpha_\circ, \beta_\circ)$ are simple and we shall only consider the regular case.

Let $\mathbf{w}$ be any real unit vector which defines the scaling scheme for $\mathbf{x}$:

$$\mathbf{w}^T\mathbf{x} = 1, \quad \|\mathbf{w}\| = 1. \tag{5.2}$$

From (3.4), the definition of $\mathbf{r}$ in (5.1) and (5.2), we deduce that

$$[L(\alpha_\circ, \beta_\circ) - \delta L]\mathbf{x} = \mathbf{0}, \quad \delta L \equiv \beta_\circ^l \mathbf{r}\mathbf{w}^T. \tag{5.3}$$

Here we are given the eigenvalue $(\alpha_\circ, \beta_\circ)$ of $(L - \delta L)$ and we are interested in the perturbed eigenvalue (α, β) of L, with perturbation δL. (Note that the roles of (α, β) and $(\alpha_\circ, \beta_\circ)$ in section 4 are reversed here.) We can apply Theorem 4.2 to this perturbation problem. However, sharper error bounds can be obtained because δL is independent of $\alpha_\circ$ (with only $\delta A_0 \neq 0$), unlike the general case in section 4. From the proof of Theorem 4.2, it is easy to see, instead of (4.9), that

$$|\beta_\circ|^{1-l}\|\delta L\| = |\beta_\circ|^{1-l}\|\beta_\circ^l \mathbf{r}\mathbf{w}^T\| \leq \|\mathbf{r}\|.$$

We then obtain, instead of (4.5),

$$s_{(\alpha,\beta)} \leq \max\{\theta_4, \theta_4^{1/p}\}, \quad \theta_4 \equiv c_1\kappa\|\mathbf{r}\|. \tag{5.4}$$

Similar to θ_2, we define θ_4 with $\|\mathbf{r}\|$ in place of $\mathcal{F}\Delta$. Here $s_{(\alpha,\beta)}$ represents the distance between the given approximate eigenvalue $(\alpha_\circ, \beta_\circ)$ and a nearest exact eigenvalue (α, β).

We shall show later in section 6 for monic matrix polynomials that the error bound in (5.4) is sharper in general than those calculated via some linearization.

5.2. Clusters of eigenvalues. As in [2, 3], the Bauer–Fike theorems can be generalized to yield perturbation results for clusters of eigenvalues. Note that a cluster can be one eigenvalue (see section 5.4), a group of multiple eigenvalues, or a group of neighboring eigenvalues.

We assume that the resolvent form $L(\alpha, \beta)^{-1}$ is decomposed into

$$L(\alpha, \beta)^{-1} = X_1 T_I(\alpha, \beta)^{-1} Z_1 + X_2 T_O(\alpha, \beta)^{-1} Z_2 \tag{5.5}$$

with T_I and T_O containing two disjoint sets of eigenvalues. The sets are chosen so that the perturbation $\delta L(\alpha, \beta)$ can be neglected in the sense that the resolvent form $L(\alpha, \beta)^{-1}$ is dominated by the first term in (5.5). Equivalently for a negligible positive constant ϵ, we select the cluster in T_I so that

$$\|X_1 T_I(\alpha, \beta)^{-1} Z_1\| \gg \|X_2 T_O(\alpha, \beta)^{-1} Z_2\| \leq \epsilon\|X_1 T_I(\alpha, \beta)^{-1} Z_1\|. \tag{5.6}$$

Consequently, we have

$$\|L(\alpha,\beta)^{-1}\| \le (1+\epsilon)\|X_1 T_I(\alpha,\beta)^{-1} Z_1\|.$$

Similar arguments and techniques to those in sections 3 and 4 can then be applied to $L(\alpha,\beta)+\delta L(\alpha,\beta)$ so that

$$(1+\epsilon)\|X_1 T_I(\alpha,\beta)^{-1} Z_1\|\ \|\delta L(\alpha,\beta)\| \ge \|L(\alpha,\beta)^{-1}\delta L(\alpha,\beta)]\| \ge 1$$

and

$$(1+\epsilon)\kappa_1\|\delta L(\alpha,\beta)\| \ge \|T_I(\alpha,\beta)^{-1}\|^{-1}, \quad \kappa_1 \equiv \|X_1\|\ \|Z_1\|.$$

Replacing $\|T_I(\alpha,\beta)^{-1}\|$ by an upper bound in the above inequality will yield similar results to those in section 4, but for the cluster in T_I rather than the whole $\sigma(L)$. Here, p will be the size of the largest Jordan block associated with the cluster in $T_I(\alpha,\beta)$. Ignoring higher-order terms, the condition numbers will then involve κ_1 instead of κ, similar to the results for clusters in [2, 3]. The price to pay for these condition numbers for clusters is the restriction that the perturbation δL has to be small (in the sense of (5.6)), contrary to the arbitrariness of the size of perturbations in classical Bauer–Fike theorems.

5.3. Asymptotic perturbation. If the perturbation is small in the sense that $\theta_i < 1$ $(i = 1, 2, 3)$ in Theorems 4.1 and 4.2 as well as Corollary 4.3, we have

$$\max\{\theta_i, \theta_i^{1/p}\} = \theta_i^{1/p}.$$

Here, p is the size of the largest Jordan block associated with the cluster in $T_I(\alpha,\beta)$. The pth root in the error bounds is an important and common feature in Bauer–Fike theorems in particular [2, 3] and perturbation results for eigenvalue problems in general (see, e.g., [29, section 23, Chapter 2, p. 81]).

For clusters of eigenvalues as discussed in section 5.2, we usually have the corresponding θ_i less than unity.

When the size of the perturbation is not restricted, several perturbed eigenvalues may correspond to a common unperturbed $\lambda_\circ$. Consequently, perturbation bounds are not available from Bauer–Fike theorems for the whole of $\sigma(L)$. If the perturbation is asymptotic ($\|\delta L\| \to 0$ in some sense), we can be certain that each and every $\lambda \in \sigma(\widetilde{L})$ corresponds to a different $\lambda_\circ \in \sigma(L)$. Perturbation bounds are then available for the whole of $\sigma(L)$.

5.4. Condition numbers for individual eigenvalues. For small perturbations, we can form the clusters in section 5.2 with individual eigenvalues. In place of (4.5), we now have the first-order perturbation result for an individual eigenvalue (α,β):

$$s_{(\alpha,\beta)} \le \{c_1 \mathcal{F} \kappa_1 \Delta\}^{1/p}, \quad \kappa_1 \equiv \|X_1\| \cdot \|Z_1\|, \tag{5.7}$$

with c_1, c_2, and $\mathcal{F}$ defined as in Theorems 4.1 and 4.2. Here p is the size of the largest Jordan block and columns in X_1 (and Z_1) contain the right- and left-eigenvectors and generalized eigenvectors, all associated with (α,β). Note that we have ignored second-order terms, as well as assuming that $\theta = c_1 \mathcal{F} \kappa_1 \Delta < 1$ for small perturbations.

Analogous to the condition number by Rice [20], we can define the condition number for the eigenvalue (α, β) as

$$C_2(\alpha,\beta) \equiv \lim_{\epsilon\to 0} \sup_{\Delta\le\epsilon} \frac{s^p_{(\alpha,\beta)}}{\Delta}.$$

As the perturbation results in (5.7) come from errors bounds from the Bauer–Fike theorems, we can define only the following condition number, which is really an upper bound of the above real condition number C_2:

$$\widetilde{C}_2(\alpha,\beta) \equiv \frac{s^p_{(\alpha,\beta)}}{\Delta}. \tag{5.8}$$

The condition number $\widetilde{C}_2$ picks up the coefficient of Δ in (5.7) while ignoring the index $1/p$. Without the constant $p\mathcal{F}$, $\widetilde{C}_2$ is essentially κ_1. This form of the condition number, as a product of the norms of the left- and right-eigenvectors, is essentially the same for most, if not all, eigenvalue problems. It is easy to see that the above result degenerates back into the more trivial cases when the eigenvalues are simple or when $l = 1$.

When $p = 1$ for simple eigenvalues, $\widetilde{C}_2$ is obviously an upper bound of the condition number C_2 by Dedieu and Tisseur [8]. In the rest of this subsection, we shall compare our perturbation results with those in [8], with $\|\cdot\|$ chosen to be the 2-norm. Note that $\widetilde{C}_2$ is meaningful in terms of the perturbation result in (5.7), when $p \ne 1$ and the perturbation Δ is not asymptotically small. (We still need Δ to be reasonably small so that the clusters of individual multiple eigenvalues are not mixed up, as in section 5.2.)

In [8], first-order variations of the eigenvalue (α, β) and its eigenvector were considered, with condition numbers defined as the norms of the derivatives of the corresponding projections. In terms of the partial differential operators $\mathcal{D}_\alpha$ and $\mathcal{D}_\beta$, we quote [8, Theorem 4.2] (with minor changes in notation).

THEOREM 5.1. *Assume that (α, β) is a simple eigenvalue of $L(\alpha, \beta)$ with corresponding right- and left-eigenvectors $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\mathbf{v} = \overline{\beta}\mathcal{D}_\alpha L\mathbf{x} - \overline{\alpha}\mathcal{D}_\beta L\mathbf{x}$. Then, with $A \equiv [A_0, \ldots, A_l]$ and $\dot{A}$ denoting the derivative of A, the condition operator of the eigenvalue (α, β) is*

$$K_2(A,\mathbf{x},\alpha,\beta) = \frac{y^H L(\dot{A},\alpha,\beta)\mathbf{x}}{\mathbf{y}^H\mathbf{v}}(-\overline{\beta},\overline{\alpha})$$

and the corresponding condition number is

$$C_2(A,\alpha,\beta) = \left(\sum_{k=0}^{l} |\alpha^{2k}|\,|\beta|^{2(l-k)}\right)^{1/2} \frac{\|\mathbf{x}\|\,\|\mathbf{y}\|}{|\mathbf{y}^H\mathbf{v}|}.$$

Under assumptions similar to those in Theorem 5.1 and using (5.7), we can derive the following bounds for $C_2(A, \alpha, \beta)$:

$$C_2(A,\alpha,\beta) \le \widetilde{C}_2 = \sqrt{\frac{l+1}{2}}\,\|\mathbf{x}\|\,\|\mathbf{z}\|. \tag{5.9}$$

Here we have assumed that the perturbation Δ is small enough for higher-order terms to be ignored, with $p = c_2 = 1$. Note that $s_{(\alpha,\beta)}$ is defined in a chordal metric which

equals the sine of the angle $\delta\phi$ between (α, β) and its perturbed value $(\widetilde{\alpha}, \widetilde{\beta})$ (as in (4.10)). When Δ is small,

$$s_{(\alpha,\beta)} = |\sin \delta\phi| \approx |\delta\phi| \approx \|(\widetilde{\alpha} - \alpha, \widetilde{\beta} - \beta)\|$$

with errors being higher-order terms in Δ.

Recall from section 2 that $\mathbf{y}$ and $\mathbf{z}$ are parallel and somehow the scaling factor $(\sum_{k=0}^{l} |\alpha^{2k}|\ |\beta|^{2(l-k)})^{1/2} |\mathbf{y}^H \mathbf{v}|^{-1}$ in Theorem 5.1 is absorbed into $\sqrt{\frac{l+1}{2}}\ \|\mathbf{z}\|$ in (5.9). Note that C_2 is meaningful only for simple eigenvalues and when perturbations are asymptotically small. When the eigenvalue is nearly defective (as in Example 2 in section 7.2, where the validity of the bound in (5.9) is tested), (5.7) implies that

$$C_2 \approx \{c_1 \mathcal{F} \kappa_1\}^{1/p}\, \Delta^{(1-p)/p},$$

which tends to infinity when $\Delta \to 0$. This indicates the ill-conditioning of defective eigenvalues because of the fractional power related to p. In such circumstances, it may be better to consider (5.7) directly when the eigenstructure and p is available.

Finally, it is the characteristic of Bauer–Fike theorems that eigenvectors cannot be treated. For condition operators (K_1) and condition numbers (C_1) for eigenvectors, please consult [8].

5.5. Optimization of condition numbers. We want to reiterate that the matrices X and Z in various κ's are not uniquely defined. Indeed, κ can be arbitrarily large for some choice of X. The condition numbers in the form of $c_1\kappa$ or $c_1\mathcal{F}\kappa$ obtained from Theorems 4.1 and 4.2 can be minimized amongst all possible X. This minimization is time consuming and impractical and is often ignored.

A numerical problem can be considered as a mapping from a data-space to a solution-space. Condition numbers are just upper bounds of the derivatives of such mappings. The nonuniqueness of upper bounds, as well as quantities like X, Z, and κ in these bounds, always makes the comparison of condition numbers from different approaches difficult. However, a comparison of a special case will be considered in sections 6 and 7 below to show that the perturbation results obtained through linearization can be worse than those in section 4.

6. Perturbation through linearization for monic matrix polynomials. We consider only the monic case here, aiming to show that the application of results in [2, 3] to linearizations of L is unlikely to produce superior results.

Let us consider the linearization involving C_L as in (2.6):

$$L^{-1}(\lambda) = [I_n, 0, \ldots, 0] C_L(\lambda)^{-1} [0, \ldots, 0, I_n]^T.$$

Similar to (4.5) for regular matrix polynomials and using the same notation, we can apply the perturbation results in [2, 3] to produce the error bound

$$s_{(\widetilde{\alpha},\widetilde{\beta})} \le \max\{\theta_5, \theta_5^{1/p}\}, \quad \theta_5 \equiv c_1 \kappa_2 \Delta_1. \tag{6.1}$$

Here

$$c_1 \equiv \min\left\{\frac{2p+1}{p+1}, p\right\}, \quad \kappa_2 \equiv \|P\|\ \|P^{-1}\|, \quad \Delta_1 \equiv \sqrt{\|\widetilde{M}_3\|^2 + \|\widetilde{M}_4\|^2}.$$

The linearization C_L is in Jordan canonical form,

$$C_L(\lambda) = P(\widehat{J} - \lambda I_n) P^{-1},$$

and

$$\widetilde{M_3} \equiv \delta A_l, \quad \widetilde{M_4} \equiv [\delta A_0, \delta A_1, \ldots, \delta A_{l-1}].$$

Similar error bounds for clusters of eigenvalues can be obtained, with P (and P^{-1}) in κ_2 replaced by the columns (and rows) of P (and P^{-1}).

Consider a standard triple (X, T, Z) constructed using the Jordan blocks in $\widehat{J}$ and

$$X = [I_n, 0, \ldots, 0]P, \qquad Z = P^{-1}[0, \ldots, 0, I_n]^T. \tag{6.2}$$

With this contrived choice, X and Z are, respectively, the first n rows and last n columns of P and P^{-1}.

It is not easy to compare the results in (6.1) with those in (4.5). The main difficulty lies in the difference between κ and κ_2, as well as their nonuniqueness. Optimization of κ and κ_2 may not be practical, while comparison of suboptimal error bounds is easier but less meaningful. Furthermore, Δ and Δ_1 are slightly different for regular matrix polynomials (with $\Delta = \Delta_1$ for the Frobenius norm, while $\Delta \geq \Delta_1$ for the 1-, 2-, and ∞-norms). There is also the additional factor $\mathcal{F} \equiv c_2\sqrt{\frac{l+1}{2}}$ in (4.5). For the special choice of X and Z in (6.2), however, we have

$$\kappa = \|X\| \, \|Z\| = \|[I_n, 0, \ldots, 0]P\| \, \|P^{-1}[0, \ldots, 0, I_n]^T\| \leq \|P\| \, \|P^{-1}\| = \kappa_2 \tag{6.3}$$

for the 1-, 2-, F-, and ∞-norms. When the perturbation analysis is done in terms of the residual vector $\mathbf{r}$ as in section 5, the additional factor $\mathcal{F}$ disappears, $\Delta_1 = \Delta$, and (6.3) implies that the error bound in (4.5) is sharper than the one calculated via the linearization C_L.

The difficulties in comparing the error bounds from section 4 with those calculated via linearizations are further illustrated in the numerical example in section 7 below. Within the limited numerical experience we have, error bounds from section 4 are usually better than those via linearizations if X and Z are chosen as in (6.2). This difference in sharpness can most likely be accounted for by the lack of consideration in the structure of the perturbation [14].

7. Numerical examples. All calculations were carried out using MATLAB [19]. Working precision is $u = 2^{-53} \approx 1.1 \times 10^{-16}$ and the 2-norm has been used.

7.1. Example 1. To illustrate the results in sections 4–6, we shall calculate some error bounds for a monic matrix polynomial. Essential features of our results can be shown without the more tedious calculations for a regular matrix polynomial. Consider the following 2×2 monic matrix polynomial example [11, p. 55]:

$$L(\lambda) = \begin{bmatrix} \lambda^3 & \sqrt{2}\lambda^2 - \lambda \\ \sqrt{2}\lambda^2 + \lambda & \lambda^3 \end{bmatrix}$$

with

$$\det L(\lambda) = \lambda^2(\lambda + 1)^2(\lambda - 1)^2.$$

The resolvent form involving a Jordan triple is

$$X(I_6\lambda - J)^{-1}Z$$

with

$$X = \left[\begin{array}{cc|cc|cc} 1 & 0 & -\sqrt{2}+1 & \sqrt{2}-2 & \sqrt{2}+1 & \sqrt{2}+2 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{array}\right], \tag{7.1}$$

$$Z^T = \frac{1}{4}\left[\begin{array}{cc|cc|cc} 0 & -4 & \sqrt{2}+2 & -\sqrt{2}-1 & -\sqrt{2}+2 & -\sqrt{2}+1 \\ 4 & 0 & 0 & 1 & 0 & -1 \end{array}\right],$$

and

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

We perturbed the matrix polynomial with randomly chosen δA_j $(j = 0, 1, 2)$, and these matrices are scaled so that $\Delta = 10^{-4}$. The matrices $10^4 \times \delta A_j$ $(j = 0, 1, 2)$ are listed below:

$$\begin{bmatrix} 0.4097 & 0.0264 \\ 0.2205 & 0.1237 \end{bmatrix}, \quad \begin{bmatrix} -0.2450 & 0.0208 \\ 0.5965 & 0.6320 \end{bmatrix}, \quad \begin{bmatrix} 0.0929 & -0.5086 \\ 0.3066 & -0.2466 \end{bmatrix}.$$

The corresponding Δ_1 in θ_5 in (6.1) satisfies

$$\Delta_1 = 1.0939 \times 10^{-4} \approx 10^{-4} = \Delta.$$

The size of the perturbation is adequately small and θ_1 in (4.1) is less than unity. Error bounds were calculated using (4.1) for individual eigenvalues and (4.2) for the whole spectrum. The factor γ is replaced by $\sqrt{l}$ in the definition of θ_1, as suggested in comment (ii) in the discussion following Theorem 4.1. For other clusters, the approach in section 5, i.e., a formula similar to (4.1) but with κ_1 replacing κ, was used. For clusters, the second-order effects of ϵ or T_O were ignored. Similar calculations were also carried out for error bounds using the linearization involving C_L as in (2.4), with $A_l = I_n$.

A right-eigenvector matrix P (with columns normalized to unit length) for C_L in (2.4) was found to be

$$\begin{bmatrix} 1 & 0 & -0.22094 & -0.04576 & -0.53340 & -0.64387 \\ 0 & 1 & 0.53340 & -0.64387 & -0.22094 & 0.04576 \\ 0 & 0 & -0.22094 & -0.26670 & 0.53340 & 0.11047 \\ 0 & 0 & 0.53340 & -0.11047 & 0.22094 & -0.26670 \\ 0 & 0 & -0.22094 & -0.48764 & -0.53340 & 0.42293 \\ 0 & 0 & 0.53340 & 0.42293 & -0.22094 & 0.48764 \end{bmatrix}.$$

The numerical results are summarized in Table 7.1, which has seven columns. The contents in these columns and their abbreviations (in brackets) are listed below:

Column 1. The cluster of eigenvalues under consideration, with $\lambda_1 = 0$, $\lambda_2 = 1$, and $\lambda_3 = -1$.

Column 2 (p). The size of the biggest Jordan block associated with the cluster.

Column 3 (MAE). Maximum of the absolute errors of the eigenvalues in the cluster.

Column 4 (EB-MP). Error bounds calculated using the matrix polynomial formulation in section 4.

Column 5 (R-MP). The ratio EB-MP/MAE.

TABLE 7.1
Comparison of error bounds from the Bauer–Fike theorem and actual errors for various clusters of eigenvalues.

Cluster	p	MAE	EB-MP	R-MP	EB-L	R-L
$\{\lambda_1, \lambda_2, \lambda_3\}$	2	5.9485(-3)	3.5506(-2)	5.97	2.2952(-2)	3.86
$\{\lambda_1, \lambda_2\}$	2	5.9485(-3)	2.0669(-2)	3.48	2.2729(-2)	3.83
$\{\lambda_1, \lambda_3\}$	2	3.0964(-3)	2.9744(-2)	9.61	2.2729(-2)	7.34
$\{\lambda_2, \lambda_3\}$	2	5.9485(-3)	2.9933(-2)	5.03	1.7365(-2)	2.92
$\{\lambda_1\}$	2	2.1176(-5)	1.1785(-4)	5.57	1.6667(-4)	7.87
$\{\lambda_2\}$	2	5.9485(-3)	1.5186(-2)	2.55	1.6843(-2)	2.83
$\{\lambda_3\}$	2	3.0964(-3)	1.5186(-2)	4.91	1.6843(-2)	5.44

TABLE 7.2
Comparison of error bounds from the Bauer–Fike theorem (selecting X and Z from P and P^{-1}) and actual errors for various clusters of eigenvalues.

Cluster	p	MAE	EB-MP	R-MP	EB-L	R-L
$\{\lambda_1, \lambda_2, \lambda_3\}$	2	5.9485(-3)	2.0990(-2)	3.53	2.2952(-2)	3.86
$\{\lambda_1, \lambda_2\}$	2	5.9485(-3)	2.0155(-2)	3.39	2.2729(-2)	3.83
$\{\lambda_1, \lambda_3\}$	2	3.0964(-3)	2.0155(-2)	6.51	2.2729(-2)	7.34
$\{\lambda_2, \lambda_3\}$	2	5.9485(-3)	1.5262(-2)	2.57	1.7365(-2)	2.92
$\{\lambda_1\}$	2	2.1176(-5)	1.1785(-4)	5.57	1.6667(-4)	7.87
$\{\lambda_2\}$	2	5.9485(-3)	1.4242(-2)	2.39	1.6843(-2)	2.83
$\{\lambda_3\}$	2	3.0964(-3)	1.4242(-2)	4.60	1.6843(-2)	5.44

Column 6 (EB-L). Error bounds calculated using the linearization (2.4), as discussed in section 6.

Column 7 (R-L). The ratio EB-L/MAE.

In Tables 7.1 and 7.2, 5.9485×10^{-3} is denoted by $5.9485(-3)$.

From Table 7.1, the error bounds calculated using matrix polynomials are better than those calculated via linearization (2.4), except for the clusters in rows 2, 4, and 5. Recall the difficulties in comparing error bounds discussed earlier.

The ratios in columns five and seven show how much the bounds overestimated the actual errors.

In constructing Table 7.1, (X, Z) and P are not related directly. For X (and Z) chosen from rows (and columns) of P (and P^{-1}) as in (6.2), we summarize the results in Table 7.2. Note that Table 7.2 is the same as Table 7.1 with the exception of the fourth and fifth columns, which contain the new results.

In Table 7.2, the error bounds calculated using matrix polynomials (EB-MP, column 4), with the new choice of X and Z, are better than those in Table 7.1 using the choice of X in (7.1). These bounds are also better than the ones in column 6 calculated via linearization (2.4) for all clusters. The last observation is consistent with the discussion in section 6.

7.2. Example 2. We shall repeat the first example in [8, section 8] to illustrate the bound in (5.9). Many results were quoted from [8], with $\widetilde{C}_2$ calculated using formulae in sections 2 and 5.4.

Consider the quadratic matrix polynomial with

$$A_0 = \begin{bmatrix} 2 & 0 & 9 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix}, \quad A_1 = \begin{bmatrix} -3 & 1 & 0 \\ 0 & -(1+\epsilon) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

TABLE 7.3
Eigenvalues and eigenvectors for $L(\alpha,\beta)$ in Example 2.

k	1	2	3	4	5	6
(α_k,β_k)	$(0,1)$	$(1,1)$	$(1+\epsilon,1)$	$(2,1)$	$(3,1)$	$(1,0)$
α_k/β_k	0	1	$1+\epsilon$	2	3	∞
$\mathbf{x}_k$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\\frac{\epsilon-1}{\epsilon+1}\\0\end{bmatrix}$	$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$
$\mathbf{y}_k$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}\frac{1}{4}\\0\\1\end{bmatrix}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}\frac{1}{5}\\\frac{1}{5(1-\epsilon)}\\1\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$
$\mathbf{z}_k$	$\begin{bmatrix}0\\\frac{-1}{1+\epsilon}\\0\end{bmatrix}$	$\begin{bmatrix}-1\\0\\-4\end{bmatrix}$	$\begin{bmatrix}0\\\frac{1}{\epsilon-1}\\0\end{bmatrix}$	$\begin{bmatrix}-1\\\frac{1}{\epsilon-1}\\-5\end{bmatrix}$	$\begin{bmatrix}0\\0\\-1\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$

(We have interchanged A_0 and A_2 from [8] to be consistent with the convention in this paper. The zero eigenvector is associated with the null-space of A_0, not A_2.)

The parameter ϵ will be used later to change the eigenstructure. The problem is regular as the characteristic polynomial

$$\det P = \alpha^5\beta - (7+\epsilon)\alpha^4\beta^2 + (17+6\epsilon)\alpha_3\beta^3 + (17+11\epsilon)\alpha^2\beta^4 + 6(1+\epsilon)\alpha\beta^5 \not\equiv 0.$$

The eigenvalues and eigenvectors are listed in Table 7.3. We have split the spectrum into two equal groups, with

$$T_1 = \operatorname{diag}\{0,1,1+\epsilon\}, \qquad T_2 = \operatorname{diag}\left\{\frac{1}{2},\frac{1}{3},0\right\}.$$

Here, T_2^{-1} contains the subspectrum $\{2,3,\infty\}$. The left-eigenvectors $\mathbf{z}_k$ are the rows of Z calculated using (2.5):

$$\begin{aligned} Z &\equiv [I_m \oplus T_2^{l-1}] \begin{bmatrix} S_{l-2} \\ V \end{bmatrix}^{-1} [0,\ldots,0,I_n]^T \\ &= \begin{bmatrix} I_3 & 0 \\ 0 & T_2 \end{bmatrix} \begin{bmatrix} X_1 & X_2 \\ A_2X_1T_1 & -A_0X_2T_2 - A_1X_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I_3 \end{bmatrix} \end{aligned} \tag{7.2}$$

as $m=n=3$ and $l=2$. Because of the zeros in T_2, $\mathbf{z}_6$ cannot be retrieved from the above calculation. We retrieved $\mathbf{z}_k$ $(k=1,2,3)$ from (7.2) and the other $\mathbf{z}_k$, using $\widehat{L}$ instead of L, from the first three rows of $\widehat{Z}$ in

$$\widehat{Z} = \begin{bmatrix} I_3 & 0 \\ 0 & T_1 \end{bmatrix} \begin{bmatrix} X_2 & X_1 \\ A_0X_2T_2 & -A_2X_1T_1 - A_1X_1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I_3 \end{bmatrix}.$$

With the addition of our condition number $\widetilde{C}_2$, Table 2 in [8] is transformed into Table 7.4. The condition numbers for eigenvectors in the original table in [8] have been deleted. The condition numbers $\widetilde{C}_2$ are calculated as in (5.9), using the eigenvectors in Table 7.3. The ratios on the last row in Table 7.4 show that $\widetilde{C}_2$ provides tight bounds for C_2 for this example, illustrating the validity of the discussion in section 5.4.

TABLE 7.4
Condition numbers for eigenvalues for $L(\alpha,\beta)$ in Example 2, with $\epsilon=\sqrt{u}$.

α_k/β_k	0	1	$1+\epsilon$	2	3	∞
C_2	1.0	3.6	1.2	4.8	0.9	1.4
$\widetilde{C}_2$	1.2	5.2	1.7	6.4	1.2	1.7
$\widetilde{C}_2/C_2$	1.2	1.4	1.4	1.3	1.4	1.2

TABLE 7.5
Condition numbers for eigenvalues for $L(\alpha,\beta)$ in Example 2, with $\epsilon=-1+\sqrt{u}$.

α_k/β_k	0	$\sqrt{u}$
C_2	8.0(7)	8.0(7)
$\widetilde{C}_2$	1.2(8)	1.2(8)
$\widetilde{C}_2/C_2$	1.5	1.5

When $\epsilon=-1+\sqrt{u}$ and similar to Table 3 in [8], we constructed Table 7.5. Ill-conditioning for the nearly defective pair of eigenvalues ($k=1,3$) is picked up by both condition numbers.

Acknowledgments. The author would like to thank Professors P. Lancaster and L. Rodman for many invaluable comments and suggestions. Comments by the referees also led to several improvements, including the sharpening of c_1 in Lemma 3.1 and the addition of Corollary 4.1 for scalar polynomials.

REFERENCES

[1] F. L. BAUER AND C. T. FIKE, *Norms and exclusion theorems*, Numer. Math., 2 (1960), pp. 137–144.
[2] E. K.-W. CHU, *Generalizations of the Bauer-Fike theorem*, Numer. Math., 49 (1986), pp. 685–691.
[3] E. K.-W. CHU, *Exclusion theorems and the perturbation analysis of the generalized eigenvalue problem*, SIAM J. Numer. Anal., 24 (1987), pp. 1114–1125.
[4] E. K.-W. CHU, *Perturbation of Eigenvalues for Matrix Polynomials*, Appl. Maths. Report and Preprints 92/25, School of Mathematical Sciences, Monash University, Australia, 1992.
[5] E. K.-W. CHU, *Pole assignment for second-ordered systems*, Mech. Syst. Signal Processing, 16 (2002), pp. 39–59.
[6] E. K.-W. CHU, *Optimization and pole assignment in control system*, Internat. J. Appl. Math. Comput. Sci., 11 (2001), pp. 1035–1053.
[7] E. K.-W. CHU AND B. N. DATTA, *Numerically robust pole assignment for second-order systems*, Internat. J. Control, 64 (1996), pp. 1113–1127.
[8] J.-P. DEDIEU AND F. TISSEUR, *Perturbation theory for homogeneous polynomial eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 435–453.
[9] L. ELSNER AND P. LANCASTER, *The spectral variation of pencils of matrices*, J. Comput. Math., 3 (1985), pp. 262–274.
[10] L. ELSNER AND J.-G. SUN, *Perturbation theorems for the generalized eigenvalue problems*, Linear Algebra Appl., 48 (1982), pp. 341–357.
[11] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.
[12] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[13] D. J. GRABINER, *Descartes' rule of signs: Another construction*, Amer. Math. Monthly, 106 (1999), pp. 854–855.
[14] D. J. HIGHAM AND N. J. HIGHAM, *Structured backward error and condition of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 493–512.

[15] P. Lancaster and J. Maroulas, *Selective perturbation of spectral properties of vibrating systems using feedback*, Linear Algebra Appl., 98 (1988), pp. 309–330.
[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.
[17] R.-C. Li, *On Perturbation Bounds for Eigenvalues of a Matrix*, preprint, Computing Center, Academia Sinica, Beijing, P.R. China, 1985; available online from http://www.cs.uky.edu/∼rcli/papers/bafijordan.ps.
[18] A. S. Markus, *Introduction to The Spectral Theory of Polynomial Operator Pencils*, American Mathematical Society, Providence, RI, 1988.
[19] *MATLAB User's Guide*, The Mathworks, Inc., Natick, MA, 2002.
[20] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.
[21] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[22] G. W. Stewart, *Gershgorin theory for the generalized eigenvalue problem* $Ax = \lambda Bx$, Math. Comput., 29 (1975), pp. 600–606.
[23] G. W. Stewart, *Perturbation bounds for the definite generalized eigenvalue problem*, Linear Algebra Appl., 23 (1979), pp. 69–86.
[24] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[25] J.-G. Sun, *The perturbation bounds for eigenspaces of a definite matrix pair*, Numer. Math., 41 (1983), pp. 321–343.
[26] F. Tisseur, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl., 309 (2000), pp. 339–361.
[27] F. Tisseur and N. J. Higham, *Structured pseudospectra for polynomial eigenvalue problems, with applications*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 187–208.
[28] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problems*, SIAM Rev., 43 (2001), pp. 235–286.
[29] J. H. Wilkinson, *Algebraic Eigenvalue Problems*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 574–581

ON RECENT CHEEGER-TYPE BOUNDS FOR NONMAXIMAL EIGENVALUES APPLIED TO POSITIVE MATRICES*

STEPHEN G. WALKER†

Abstract. This paper is concerned with Cheeger-type bounds for nonmaximal eigenvalues of nonnegative irreducible matrices. It is shown that recent upper bounds found by Nabben can be strictly improved when the matrices are positive, stochastic, and reversible, indicating the Nabben bounds are never sharp in this case.

Key words. nonmaximal eigenvalues, reversible Markov chain, stochastic matrix

AMS subject classifications. 15A18, 15A48, 15A51

DOI. 10.1137/S0895479802404684

1. Introduction. Let $P = [p_{i,j}] \in \mathbb{R}^{n,n}$ be an irreducible matrix and the transition probability for an irreducible and reversible Markov chain on the finite state space $X = \{1, \ldots, n\}$. Reversibility of the chain implies

$$\pi_i \, p_{i,j} = \pi_j \, p_{j,i}.$$

Here π is the stationary probability corresponding to the chain, being the normalized eigenvector corresponding to the largest eigenvalue, which is 1. Define $q_{i,j} = \pi_i \, p_{i,j}$. Then $\sum_i \pi_i p_{i,j} = \pi_j$ and $\sum_{i,j} q_{i,j} = 1$.

This paper considers the second largest eigenvalue of P, say β_1, which is both real (a consequence of the reversibility of the chain) and strictly less than 1 (a consequence of the irreducibility of the chain). [4] proved that

$$\beta_1 \leq 1 - l^2/2,$$

where

$$l = \min_{\pi(S) \leq \frac{1}{2}, \, S \neq \emptyset} \frac{Q(S, S')}{\pi(S)}$$

is the Cheeger constant for P; see [1]. Here, for $A, B \subseteq X$, $Q(A, B) = \sum_{i \in A, \, j \in B} q_{i,j}$, $\pi(A) = \sum_{i \in A} \pi_i$ and $A' = X - A$.

This upper bound, which has been studied by [2] and [3], has been improved by [5]:

$$\beta_1 \leq \sqrt{1 - l^2}. \tag{1.1}$$

[5] provides two other upper bounds:

$$\beta_1 \leq \mu + (1 - \mu)\sqrt{1 - i^2} \tag{1.2}$$

*Received by the editors March 25, 2002; accepted for publication (in revised form) by R. Nabben May 16, 2003; published electronically November 14, 2003. This work was supported by an Advanced Research Fellowship from the UK Engineering and Physical Sciences Research Council.
http://www.siam.org/journals/simax/25-2/40468.html
†Department of Mathematical Sciences, University of Bath, BA2 7AY, UK (S.G.Walker@bath.ac.uk).

and

$$\beta_1 \le \xi + \sqrt{(1-\xi)^2 - h^2/\max_i \pi_i^2}. \tag{1.3}$$

Here, $\mu = \max_i p_{i,i}$, $\xi = \min_i p_{i,i}$,

$$i = \min_{S \in C} \frac{Q(S, S')}{\sum_{i \in S}(1 - p_{i,i})\pi_i},$$

$$C = \left\{ S : S, S' \neq \emptyset,\ \sum_{i \in S}(1 - p_{i,i})\pi_i \le \frac{1}{2}\sum_i (1 - p_{i,i})\pi_i \right\},$$

$$h = \min_{S \in D} \frac{Q(S, S')}{|S|},$$

and

$$D = \{S : S, S' \neq \emptyset,\ |S| \le [n/2]\}.$$

These bounds can be found on pages 574 and 575 of [5].

In section 2 it is shown that if $P > 0$, that is, $p_{i,j} > 0$ for all $i, j \in X$, then the upper bounds of [5] are never sharp. This suggests that upper bounds for β_1 need reworking when P is positive. This paper provides a basis for improving the upper bounds in the positive case. Section 3 presents some illustrations and section 4 summarizes.

2. Improved Cheeger bounds. Suppose that $P > 0$ and $P \neq \mathbf{1}\pi'$, and consider the nonnegative stochastic matrix

$$P(\alpha) = \frac{(1+\alpha)\,P - \mathbf{1}\pi'}{\alpha},$$

defined for any α for which

$$1 + \alpha \ge a = \max_{i,j} \frac{\pi_j}{p_{i,j}}.$$

A Poincaré bound of [2] yields

$$\beta_1 \le 1 - \min_{i \neq j} \frac{p_{i,j}}{\pi_j} \le 1 - \min_{i,j} \frac{p_{i,j}}{\pi_j} = \min_{\alpha \ge a - 1} \frac{\alpha}{1+\alpha}.$$

$P(\alpha)$ is irreducible if $\beta_1 < \alpha/(1+\alpha)$, it is reversible with respect to π, and if λ is a nontrivial eigenvalue of P, then $\lambda(1+\alpha)/\alpha$ is a nontrivial eigenvalue of $P(\alpha)$. Consequently, the largest nontrivial eigenvalue of $P(\alpha)$ is $(1+\alpha)\beta_1/\alpha$. Now

$$p_{i,j}(\alpha) = \frac{(1+\alpha)p_{i,j} - \pi_j}{\alpha},$$

so

$$q_{i,j}(\alpha) = \frac{(1+\alpha)q_{i,j} - \pi_i\pi_j}{\alpha}$$

and therefore the corresponding Cheeger constant for $P(\alpha)$ is given by

$$l_\alpha = \min_{\pi(S)\le\frac{1}{2}, S\neq\emptyset}\left\{\frac{Q_\alpha(S,S')}{\pi(S)}\right\}$$

$$= \min_{\pi(S)\le\frac{1}{2}, S\neq\emptyset}\left\{\frac{(1+\alpha)l(S)-\pi(S')}{\alpha}\right\},$$

where $l(S) = Q(S,S')/\pi(S)$ and

$$Q_\alpha(S,S') = \sum_{i\in S, j\in S'} q_{i,j}(\alpha).$$

Let us also define

$$l_\alpha(S) = \frac{(1+\alpha)l(S)-\pi(S')}{\alpha}.$$

It is straightforward to confirm that $l(S) < 1$ for $S \neq \emptyset$, on account of $P > 0$. Hence there exists an α large enough such that $l_\alpha(S) = l(S) + \{l(S) - \pi(S')\}/\alpha \le 1$. We now show that if $P > 0$, then bound (1.1) can be improved.

THEOREM 2.1. *If $P > 0$, there exists a finite α such that*

$$\frac{\alpha}{1+\alpha}\sqrt{1-l_\alpha^2} < \sqrt{1-l^2}.$$

Proof. We prove that

$$\frac{\alpha}{1+\alpha}\sqrt{1-l_\alpha^2(S)} < \sqrt{1-l^2(S)}$$

for all S with $l_\alpha(S) \le 1$ and $S \neq \emptyset$. It is not possible that $l(S) = 1$; otherwise $Q(S,S) = 0$, contradicting the fact that $P > 0$. Squaring and removing the $l^2(S)$ term from both sides, we need to show that

$$\left(\frac{\alpha}{1+\alpha}\right)^2 + \frac{2l(S)\pi(S')}{1+\alpha} - \frac{\pi^2(S')}{(1+\alpha)^2} < 1.$$

Using $l(S) < 1$, this reduces to showing that

$$2\alpha\pi(S') - \pi^2(S') + 2\pi(S') < 1 + 2\alpha,$$

which follows since

$$-\{1-\pi(S')\}^2 < 2\alpha\{1-\pi(S')\},$$

there being a strict inequality as $\pi(S') < 1$. □

We now look at the alternative [5] bounds, starting with (1.2). Let us first define

$$i(S) = \frac{Q(S,S')}{\sum_{i\in S}(1-p_{i,i})\pi_i}$$

for $S \in C$,

$$i_\alpha = \min_{S\in C_\alpha}\frac{Q_\alpha(S,S')}{\sum_{i\in S}(1-p_{i,i}(\alpha))\pi_i}$$

with

$$C_\alpha = \left\{ S : S, S' \neq \emptyset,\ \sum_{i \in S} (1 - p_{i,i}(\alpha))\pi_i \leq \frac{1}{2} \sum_i (1 - p_{i,i}(\alpha))\pi_i \right\},$$

and

$$\begin{aligned} i_\alpha(S) &= \frac{Q_\alpha(S, S')}{\sum_{i \in S}(1 - p_{i,i}(\alpha))\pi_i} \\ &= \frac{\left[(1+\alpha)Q(S, S') - \pi(S)\pi(S')\right]/\alpha}{\sum_{i \in S}\left\{1 - (1+\alpha)p_{i,i}/\alpha + \pi_i/\alpha\right\}\pi_i} \\ &= \frac{(1+\alpha)Q(S, S') - \pi(S)\pi(S')}{\alpha \sum_{i \in S}(1 - p_{i,i})\pi_i + \sum_{i \in S}(\pi_i - p_{i,i})\pi_i}. \end{aligned}$$

THEOREM 2.2. *If $P > 0$, there exists a finite α such that*

$$\frac{\alpha}{1+\alpha}\left\{\mu_\alpha + (1 - \mu_\alpha)\sqrt{1 - i_\alpha^2}\right\} < \mu + (1 - \mu)\sqrt{1 - i^2},$$

where $\mu_\alpha = \max_i p_{i,i}(\alpha)$.

Proof. Now $\alpha\mu_\alpha/(1+\alpha) < \mu$ and so

$$\begin{aligned} \tfrac{\alpha}{1+\alpha}\left\{\mu_\alpha + (1 - \mu_\alpha)\sqrt{1 - i_\alpha^2}\right\} &< \mu\left(1 - \sqrt{1 - i_\alpha^2}\right) + \tfrac{\alpha}{1+\alpha}\sqrt{1 - i_\alpha^2} \\ &= \mu + \tfrac{\alpha}{1+\alpha}\left\{1 - \tfrac{(1+\alpha)\mu}{\alpha}\right\}\sqrt{1 - i_\alpha^2} \\ &< \mu + \tfrac{\alpha}{1+\alpha}(1 - \mu)\sqrt{1 - i_\alpha^2}, \end{aligned}$$

and so we are left with showing that there exists a finite α such that

$$\frac{\alpha}{1+\alpha}\sqrt{1 - i_\alpha^2} \leq \sqrt{1 - i^2}.$$

We achieve this by proving there exists a finite α such that

$$\frac{\alpha}{1+\alpha}\sqrt{1 - i_\alpha^2(S)} \leq \sqrt{1 - i^2(S)} \tag{2.1}$$

for all $S \in C_\alpha$ and then that $C_\alpha \subset C$. Now $i(S) = 1$ iff $\sum_{i \in S}(1 - p_{i,i})\pi_i = Q(S, S')$ iff

$$\sum_{i \in S}(1 - p_{i,i})\pi_i = \sum_{i \in S}\pi_i\left(1 - \sum_{j \in S} p_{i,j}\right)$$

iff

$$\sum_{i \in S} p_{i,i}\pi_i = \sum_{i \in S, j \in S} p_{i,j}\pi_i$$

iff $|S| = 1$, since $p_{i,j} > 0$. Therefore, if $i(S) = 1$, then $|S| = 1$, implying that $i_\alpha(S) = 1$ provided we take α large enough so that $p_{i,j}(\alpha) > 0$. This deals with the case when $i(S) = 1$. Using the expression for $i_\alpha(S)$, we have

$$i_\alpha(S) = (1+\alpha)i(S)/\alpha - \gamma_\alpha(S)/\alpha,$$

where

$$\gamma_\alpha(S) = \frac{\alpha\pi(S)\pi(S')\sum_{i\in S}(1-p_{i,i})\pi_i + (1+\alpha)Q(S,S')\sum_{i\in S}(\pi_i - p_{i,i})\pi_i}{\sum_{i\in S}(1-p_{i,i})\pi_i\left\{\alpha\sum_{i\in S}(1-p_{i,i})\pi_i + \sum_{i\in S}(\pi_i - p_{i,i})\pi_i\right\}}. \tag{2.2}$$

Therefore, from (2.1), we need to show that there exists a finite α such that

$$\left(\frac{\alpha}{1+\alpha}\right)^2 + \frac{2i(S)\gamma_\alpha(S)}{1+\alpha} - \frac{\gamma_\alpha^2(S)}{(1+\alpha)^2} \le 1.$$

Using $i(S) < 1$, we see that this amounts to showing that there exists a finite α for which

$$-(1-\gamma_\alpha(S))^2 \le 2\alpha(1-\gamma_\alpha(S)),$$

which is true if $\gamma_\alpha(S) \le 1$. Recalling (2.2), such an α can be found if

$$\pi(S)\pi(S')\sum_{i\in S}(1-p_{i,i})\pi_i + Q(S,S')\sum_{i\in S}(\pi_i - p_{i,i})\pi_i < \left(\sum_{i\in S}(1-p_{i,i})\pi_i\right)^2.$$

Using $Q(S,S') < \sum_{i\in S}(1-p_{i,i})\pi_i$, that is, $i(S) < 1$, we need to show that

$$\pi(S)\pi(S') + \sum_{i\in S}(\pi_i - p_{i,i})\pi_i \le \sum_{i\in S}(1-p_{i,i})\pi_i,$$

that is,

$$\pi(S)\pi(S') + \sum_{i\in S}\pi_i^2 \le \pi(S).$$

This is seen to be true by noting that

$$\sum_{i\in S}\pi_i^2 \le \left(\sum_{i\in S}\pi_i\right)^2 = \pi^2(S)$$

and that $\pi(S) + \pi(S') = 1$. Finally, we have to show that $S \in C_\alpha$ implies $S \in C$ for some α. So let us assume that $S \in C_\alpha$, that is,

$$\sum_{i\in S}(1-p_{i,i})\pi_i + \frac{1}{\alpha}\sum_{i\in S}(\pi_i - p_{i,i})\pi_i \le \frac{1}{2}\sum_i(1-p_{i,i})\pi_i + \frac{1}{2\alpha}\sum_i(\pi_i - p_{i,i})\pi_i.$$

We can obviously choose α large enough to ensure that this implies

$$\sum_{i\in S}(1-p_{i,i})\pi_i \le \frac{1}{2}\sum_i(1-p_{i,i})\pi_i.$$

This completes the proof. □

Finally, we improve bound (1.3). First define $h(S) = Q(S,S')/|S|$ for $S \in D$ and

$$h_\alpha(S) = \frac{(1+\alpha)h(S) - \tau(S)}{\alpha},$$

where $\tau(S) = \pi(S)\pi(S')/|S|$ for $S \in D_\alpha = D$.

THEOREM 2.3. *If $P > 0$, there exists a finite α such that*

$$\frac{\alpha}{1+\alpha}\left\{\xi_\alpha + \sqrt{(1-\xi_\alpha)^2 - h_\alpha^2/\max_i \pi_i^2}\right\} < \xi + \sqrt{(1-\xi)^2 - h^2/\max_i \pi_i^2}\,,$$

where $\xi_\alpha = \min_i p_{i,i}(\alpha)$.

Proof. We have $\alpha\xi_\alpha < (1+\alpha)\xi$ and so we need to show there exists a finite α for which

$$\left(\frac{\alpha}{1+\alpha}\right)^2 (1-\xi_\alpha)^2 + \frac{2h(S)\tau(S)}{\delta(1+\alpha)} - \frac{\tau^2(S)}{\delta(1+\alpha)^2} \le (1-\xi)^2, \tag{2.3}$$

where $\delta = \max_i \pi_i^2$. Now $p_{i,i}(\alpha) = (1+\alpha)p_{i,i}/\alpha - \pi_i/\alpha$ and so

$$\min_i p_{i,i}(\alpha) \ge (1+\alpha)/\alpha \min_i p_{i,i} - \max_i \pi_i/\alpha,$$

which means that

$$\xi_\alpha \ge (1+\alpha)\xi/\alpha - \sqrt{\delta}/\alpha.$$

Therefore,

$$\begin{aligned}\left(\tfrac{\alpha}{1+\alpha}\right)^2 (1-\xi_\alpha)^2 &\le \left\{(1-\xi) - (1-\sqrt{\delta})/(1+\alpha)\right\}^2 \\ &= (1-\xi)^2 - 2(1-\xi)(1-\sqrt{\delta})/(1+\alpha) + (1-\sqrt{\delta})^2/(1+\alpha)^2.\end{aligned}$$

Using the above inequality in (2.3) and then comparing the $1/(1+\alpha)$ terms in (2.3), if

$$h(S)\tau(S) < \delta(1-\xi)(1-\sqrt{\delta}),$$

that is,

$$\left(\sum_{i\in S, j\in S'} \pi_i p_{i,j}\right) \pi(S)\pi(S') < |S|^2\delta(1-\xi)(1-\sqrt{\delta}),$$

then we can find an α large enough so that (2.3) holds; that is, we can ignore the $1/(1+\alpha)^2$ terms. Now

$$\sum_{i\in S, j\in S'} \pi_i p_{i,j} = \sum_{i\in S} \pi_i \left(1 - \sum_{j\in S} p_{i,j}\right) \le \pi(S)(1-\xi)$$

and so we need to show that

$$\pi^2(S)\pi(S') < |S|^2\delta(1-\sqrt{\delta}),$$

which is equivalent to

$$\pi(S)\pi(S') < |S|\sqrt{\delta}(1-\sqrt{\delta}) \tag{2.4}$$

as $\pi(S) \le |S|\sqrt{\delta}$. Put $\widehat{\pi} = \max\{\pi_1, \ldots, \pi_n\} = \sqrt{\delta}$.

If $\widehat{\pi} \in S$, then $\pi(S') = 1 - \pi(S) \leq 1 - \widehat{\pi}$ and the inequality (2.4) holds, except there is an equality when $S = \{\widehat{\pi}\}$. However, in this case it is easy to show directly that (2.3) holds.

Assume now that $\widehat{\pi} \in S'$. If $\pi(S) \geq \widehat{\pi}$, then $\pi(S') \leq 1 - \widehat{\pi}$ and the inequality (2.4) holds as $\pi(S) < |S|\sqrt{\delta}$. So now look at $\pi(S) < \widehat{\pi}$. If $\pi(S) < \widehat{\pi} \leq \frac{1}{2}$, then $\pi(S)\{1-\pi(S)\} < \widehat{\pi}(1-\widehat{\pi})$ and the inequality (2.4) holds. So, finally, we need to deal with the case when $\widehat{\pi} \in S'$ and $\widehat{\pi} > \frac{1}{2}$. But in this case, as $\widehat{\pi} \in S'$, it follows that $\pi(S') \geq \widehat{\pi} > \frac{1}{2}$ and $\pi(S')\{1-\pi(S')\} \leq \widehat{\pi}(1-\widehat{\pi})$ and the inequality (2.4) holds if $|S| > 1$. If $|S| = 1$, then $\pi(S') > \widehat{\pi}$, so long as $n > 2$, and then $\pi(S')\{1 - \pi(S')\} < \widehat{\pi}(1 - \widehat{\pi})$ and the inequality (2.4) holds. □

3. Illustrations. Let us first consider the 2×2 stochastic matrix

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix},$$

where $0 < a, b < 1$. We will further assume that $a + b < 1$ and $b < a$. The second eigenvalue is $\beta_1 = 1 - a - b$, and

$$\pi = \left(\frac{b}{a+b}, \frac{a}{a+b} \right).$$

It is easy to show that $l = a$, $i = 1$, and $h = ab/(a + b)$ so that the [5] upper bounds are given by $\sqrt{1-a^2}$, $1-b$, and $1-a+\sqrt{(1-a)^2-b^2}$, all of which are strictly greater than $1 - a - b$.

The upper bounds found in this paper are (1.1), (1.2), and (1.3). We use the P_α versions, which are, respectively,

$$\beta_1 \leq \frac{\alpha}{1+\alpha}\sqrt{1-l_\alpha^2},$$

$$\beta_1 \leq \frac{\alpha}{1+\alpha}\left\{\mu_\alpha + (1-\mu_\alpha)\sqrt{1-i_\alpha^2}\right\},$$

and

$$\beta_1 \leq \frac{\alpha}{1+\alpha}\left\{\xi_\alpha + \sqrt{(1-\xi_\alpha)^2 - h_\alpha^2/\max_i \pi_i^2}\right\}.$$

We can take $1 + \alpha = 1/(a+b)$, and then it is easy to see that all the bounds are equal to $1 - a - b$. In fact, $l_\alpha = i_\alpha = h_\alpha = 0$ by virtue of

$$(1+\alpha)Q(S, S') = \pi(S)\pi(S')$$

for $S, S' \neq \emptyset$. That is, $(1 + \alpha)ab/(a + b) = a/(a + b) \times b/(a + b)$.

Now let us consider the reversible $n \times n$ stochastic matrix

$$p_{i,j} = \begin{cases} c, & i \neq j, \\ 1 - c(n-1), & i = j, \end{cases}$$

for $c > 0$ and $c < 1/n$. The best upper bound obtained by [5] is

$$\beta_1 \leq 1 - c(n-1).$$

On the other hand, with $1 + \alpha = 1/(cn)$, $P_\alpha = I_n$ and so the improved upper bound is given by

$$\beta_1 \leq \frac{\alpha}{1+\alpha} = 1 - cn < 1 - c(n-1).$$

4. Summary. If $L(P)$ is an upper Cheeger-type bound of the form (1.1), (1.2), (1.3) for a nonmaximal eigenvalue of the reversible stochastic matrix P, then

$$\tilde{L}(P) = \min_{1+\alpha \geq a} \left\{ \frac{\alpha}{1+\alpha} L\left(\frac{(1+\alpha)P - \mathbf{1}\pi'}{\alpha} \right) \right\},$$

where

$$a = \max_{i,j} \frac{\pi_j}{p_{i,j}}$$

is an improved upper bound.

For a general irreducible positive matrix A with positive left and right eigenvectors u and v, we can define

$$A(\alpha) = \frac{(1+\alpha)A - uv'}{\alpha}$$

and make use of

$$\mathrm{Re}(\beta_1) \leq \tilde{L}(A) = \min_{1+\alpha \geq b} \left\{ \frac{\alpha}{1+\alpha} L(A(\alpha)) \right\},$$

where

$$b = \max_{i,j} \{u_i v_j / a_{i,j}\},$$

to investigate improved upper bounds.

REFERENCES

[1] J. Cheeger, *A lower bound for the lowest eigenvalue of the Laplacian*, in Problems in Analysis: A Symposium in Honor of S. Bochner, R. C. Gunning, ed., Princeton University Press, Princeton, NJ, 1970, pp. 195–199.

[2] P. Diaconis and D. Strook, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–61.

[3] J. Fulman and E. L. Wilmer, *Comparing eigenvalue bounds for Markov chains: When does Poincaré beat Cheeger?*, Ann. Appl. Probab., 8 (1999), pp. 1–13.

[4] M. Jerrum and A. Sinclair, *Approximating the permanent*, SIAM J. Comput., 18 (1989), pp. 1149–1178.

[5] R. Nabben, *Improved upper bounds for the real part of nonmaximal eigenvalues of nonnegative matrices*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 574–579.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 2, pp. 582–600

SOLVING RANK-DEFICIENT AND ILL-POSED PROBLEMS USING UTV AND QR FACTORIZATIONS*

LESLIE V. FOSTER†

Abstract. The algorithm of Mathias and Stewart [*Linear Algebra Appl.*, 182 (1993), pp. 91–100] is examined as a tool for constructing regularized solutions to rank-deficient and ill-posed linear equations. The algorithm is based on a sequence of QR factorizations. If it is stopped after the first step, it produces the same solution as the complete orthogonal decomposition used in LAPACK's xGELSY. However, we show that for low-rank problems a careful implementation can lead to an order of magnitude improvement in speed over xGELSY as implemented in LAPACK. We prove, under assumptions similar to assumptions used by others, that if the numerical rank is chosen at a gap in the singular value spectrum and if the initial factorization is rank-revealing, then, even if the algorithm is stopped after the first step, approximately half the time its solutions are closer to the desired solution than are the singular value decomposition (SVD) solutions. Conversely, the SVD will be closer approximately half the time, and in this case overall the two algorithms are very similar in accuracy. We confirm this with numerical experiments. Although the algorithm works best for problems with a gap in the singular value spectrum, numerical experiments suggest that it may work well for problems with no gap.

Key words. regularization, ill-posed, ill-conditioned, rank-deficient, QR factorization, singular value decomposition, UTV decomposition

AMS subject classifications. 65F25, 65F22, 15A23

DOI. 10.1137/S089547980037785X

1. Introduction.

The solution to ill-posed or nearly rank-deficient linear equations is important in many applications [18]. To solve these systems, some form of regularization is usually used. By regularization we mean the replacement of the original problem with a different, better posed problem. For example, if the original problem is

$$\min \|b - Ax\|, \tag{1}$$

where A is an $m \times n$ ill-conditioned matrix with $m \geq n$ and the norm is the Euclidean norm, it is often recommended to approximate A with an exactly rank-deficient matrix $\widehat{A}$ and solve for the minimum norm solution to (1) with A replaced by $\widehat{A}$. To construct $\widehat{A}$, it is useful to decompose A with a rank-revealing decomposition. The low-rank approximation $\widehat{A}$ to A can be obtained by truncating such decompositions. The singular value decomposition (SVD) is a very good, but expensive, decomposition. We use the complete orthogonal or UTV decomposition $A = UTV^T$ with U orthogonal, V orthogonal, and T triangular. Here the superscript T indicates transpose. Some of our results will apply to any UTV factorization and others to UTV factorizations produced by the algorithm of Mathias and Stewart [21]. This algorithm produces UTV factorizations by using a sequence of QR factorizations. We begin the algorithm with an initial UTV factorization of the form $A = UTV^T = QR\Pi^T$, where Q is an orthogonal matrix, R is upper triangular, and Π is the permutation matrix produced

*Received by the editors September 11, 2000; accepted for publication (in revised form) by P. C. Hansen March 16, 2003; published electronically November 14, 2003. This research was supported in part by the Woodward bequest to the Department of Mathematics, San Jose State University.
http://www.siam.org/journals/simax/25-2/37785.html
†Department of Mathematics and Computer Science, San Jose State University, San Jose, CA 95192 (foster@math.sjsu.edu).

by the standard QR algorithm with pivoting [1, 4, 20]. (We also will discuss a variation where initially A^T is factored in this form.) If the algorithm is stopped after the first step it produces the same solution as the complete orthogonal decomposition used in LAPACK's xGELSY. However, we show that for low-rank problems a careful implementation can lead to an order of magnitude improvement in speed over the two routines xGELSY and xGELSD that LAPACK provides for solving rank-deficient problems. We prove, under assumptions similar to assumptions used by others about the true solution to (1) and the noise in b, that if the numerical rank is chosen at a gap in the singular value spectrum and if the initial factorization is rank-revealing [3, p. 22], then, even if the algorithm is stopped after the first step, approximately half the time its solutions are closer to the desired solution than are the SVD solutions. Conversely, the SVD will be closer approximately half the time, and in this case overall the two algorithms are very similar in accuracy. We confirm this with numerical experiments. Although the algorithm works best for problems with a gap in the singular value spectrum, numerical experiments suggest that it may work well for problems with no gap.

The paper is organized as follows. Following this introduction, section 2 discusses UTV factorizations in general. Section 3 discusses the algorithm in [21]. Section 4 focuses on perturbation errors and section 5 on regularization errors. Section 6 describes implementation of the algorithm and numerical experiments. Section 7 has conclusions.

2. UTV factorizations. Consider any UTV factorization of A, $A = UTV^T$. Let k be the rank of the low-rank approximation to A. It is useful to partition the factorization as follows. If T is lower triangular ($T = L$), we partition UTV^T as

$$A = UTV^T = ULV^T = \begin{pmatrix} \widehat{U} & U_0 \end{pmatrix} \begin{pmatrix} \widehat{L} & 0 \\ H & E \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \widehat{V} & V_0 \end{pmatrix}^T. \tag{2}$$

If T is upper triangular ($T = R$), we partition UTV^T as

$$A = UTV^T = URV^T = \begin{pmatrix} \widehat{U} & U_0 \end{pmatrix} \begin{pmatrix} \widehat{R} & F \\ 0 & G \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \widehat{V} & V_0 \end{pmatrix}^T. \tag{3}$$

In these equations $\widehat{U}$ is $m \times k$, U_0 is $m \times (m-k)$, $\widehat{V}$ is $n \times k$, V_0 is $n \times (n-k)$, $\widehat{L}$ is $k \times k$, H is $(n-k) \times k$, E is $(n-k) \times (n-k)$, $\widehat{R}$ is $k \times k$, F is $k \times (n-k)$, and G is $(n-k) \times (n-k)$. In equations (2) and (3) U_0 corresponds to the last two block rows in the block triangular matrices. If we do not need to distinguish whether T is lower or upper triangular, we will let $\widehat{T}$ represent either $\widehat{L}$ or $\widehat{R}$. In each case we consider two low-rank approximations to A. If T is either lower or upper triangular we will call $\widehat{U}\widehat{T}\widehat{V}^T$ the corner low-rank approximation to A. If T is lower triangular we call $U\,[\widehat{L}^T\ H^T\ 0]^T\,\widehat{V}^T$ the block-column low-rank approximation to A. Similarly, if T is upper triangular, we call $\widehat{U}\,[\widehat{R}\ F]\,V^T$ the block-row low-rank approximation to A.

We will also partition the SVD of A in a similar manner to (2) and (3).

$$A = U_S D V_S^T = \begin{pmatrix} \widehat{U}_S & U_{S0} \end{pmatrix} \begin{pmatrix} \widehat{D} & 0 \\ 0 & D_0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \widehat{V}_S & V_{S0} \end{pmatrix}^T.$$

$\widehat{A}_S = \widehat{U}_S \widehat{D} \widehat{V}_S^T$ is the rank k approximation produced by the SVD. We will use $s_1 \geq s_2 \geq \cdots \geq s_n$ to indicate the singular values of A. We will also use $\sigma_k(A)$, $1 \leq k \leq n$, to indicate the kth singular value of a matrix A. Note that the SVD produces the best rank k approximation to A in the sense that $\|A - \widetilde{A}\|$ is minimized over all rank k matrices $\widetilde{A}$ when $\widetilde{A} = \widehat{A}_S$ [3, p. 12].

When solving (1) we will consider the regularized solution $x_T = \widehat{A}_T^+ b$, where the superscript $^+$ indicates pseudoinverse and $\widehat{A}_T$ is either a corner or block-row/column low-rank approximation to A corresponding to a UTV factorization of A. It will be clear from the context whether x_T refers to a corner or block-row/column low-rank approximation. We call x_T the truncated UTV solution to (1). We assume in the rest of this paper that $\widehat{T}$ and $\widehat{D}$ are nonsingular. In this case the corner low-rank solution has a simple form, $x_T = \widehat{V} \widehat{T}^{-1} \widehat{U}^T b$. The truncated SVD approximate solution to (1) is $x_S = \widehat{V}_S \widehat{D}^{-1} \widehat{U}_S^T b$.

To evaluate the accuracy of x_T we will assume that there is an underlying noiseless solution x_0 to (1) such that $Ax_0 = b_0$ and that in (1) $b = b_0 + \delta b$, where δb is a noise vector in the right-hand-side b. We will prove theorems and carry out numerical experiments that evaluate x_T based on the value of $\|x_T - x_0\|$ and will compare $\|x_T - x_0\|$ with $\|x_S - x_0\|$. We might note that other authors [6, 8, 10] have focused on bounding $\|x_T - x_S\|$. In many cases the goal in solving (1) is to recover an underlying solution x_0 that is different from x_S [18, 22]. In these cases comparison of $\|x_T - x_0\|$ with $\|x_S - x_0\|$ is of interest.

Suppose that C is some regularization operator so that $x = Cb$ is the regularized solution to (1). If x_0 is the underlying noiseless solution, then

$$x - x_0 = Cb - x_0 = (CA - I)x_0 + C(\delta b) \quad \text{and} \tag{4}$$

$$\|x - x_0\| \leq \|(CA - I)x_0\| + \|C(\delta b)\|. \tag{5}$$

The two terms on the right are called, respectively, the regularization error and the perturbation error. In the case that C corresponds to a corner low-rank solution calculated using a truncated UTV factorization, where T is lower triangular, we have a sharper result than (5):

$$\|x - x_0\|^2 = \|(CA - I)x_0\|^2 + \|C(\delta b)\|^2. \tag{6}$$

This result follows since, if $C = \widehat{V} \widehat{L}^{-1} \widehat{U}^T$, then $C^T(CA - I) = 0$ follows easily.

Our first theorem relates $\|x_T - x_0\|$ and $\|x_S - x_0\|$.

THEOREM 1. *Define*

$$\widetilde{U} = U^T U_S = \begin{pmatrix} \widehat{U}^T \widehat{U}_S & \widehat{U}^T U_{S0} \\ U_0^T \widehat{U}_S & U_0^T U_{S0} \end{pmatrix} = \begin{pmatrix} \widetilde{U}_{11} & \widetilde{U}_{12} \\ \widetilde{U}_{21} & \widetilde{U}_{22} \end{pmatrix}, \tag{7}$$

$$\widetilde{V} = V^T V_S = \begin{pmatrix} \widehat{V}^T \widehat{V}_S & \widehat{V}^T V_{S0} \\ V_0^T \widehat{V}_S & V_0^T V_{S0} \end{pmatrix} = \begin{pmatrix} \widetilde{V}_{11} & \widetilde{V}_{12} \\ \widetilde{V}_{21} & \widetilde{V}_{22} \end{pmatrix}, \tag{8}$$

$$M = \begin{pmatrix} -\widehat{D}^{-1} \widetilde{V}_{21}^T \widetilde{V}_{21} \widehat{D}^{-1} & \widetilde{U}_{11}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{12} \\ \widetilde{U}_{12}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{11} & \widetilde{U}_{12}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{12} \end{pmatrix}, \tag{9}$$

and

$$N = \begin{pmatrix} \widetilde{V}_{21}^T \widetilde{V}_{21} & \widetilde{V}_{21}^T \widetilde{V}_{22} \\ \widetilde{V}_{22}^T \widetilde{V}_{21} & -\widetilde{V}_{12}^T \widetilde{V}_{12} \end{pmatrix}. \tag{10}$$

Also let $\widetilde{\delta b} = U_S^T \delta b$ *and* $\widetilde{x}_0 = V_S^T x_0$ *and let* x_T *be the corner low-rank solution to* (1) *calculated from a truncated UTV factorization with* T *lower triangular. Then*

$$\|x_T - x_0\|^2 = \|x_S - x_0\|^2 + \widetilde{\delta b}^T M \, \widetilde{\delta b} + \widetilde{x}_0^T N \, \widetilde{x}_0. \tag{11}$$

Proof. First note that

$$T = U^T U_S D V_S^T V = \widetilde{U} D \widetilde{V}^T = \begin{pmatrix} \widetilde{U}_{11} & \widetilde{U}_{12} \\ \widetilde{U}_{21} & \widetilde{U}_{22} \end{pmatrix} \begin{pmatrix} \widehat{D} & 0 \\ 0 & D_0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \widetilde{V}_{11} & \widetilde{V}_{12} \\ \widetilde{V}_{21} & \widetilde{V}_{22} \end{pmatrix}^T. \tag{12}$$

The perturbation error term for the SVD solution is $\|\widehat{A}_S^+ \, \delta b\|$ where $\widehat{A}_S^+ = \widehat{V}_S \widehat{D}^{-1} \widehat{U}_S^T$ and for the UTV solution it is $\|\widehat{A}_T^+ \, \delta b\|$ where $\widehat{A}_T^+ = \widehat{V} \widehat{T}^{-1} \widehat{U}^T$. Now $\|\widehat{A}_S^+ \, \delta b\|^2 = \|\widehat{D}^{-1} \widehat{U}_S^T U_S \, \widetilde{\delta b}\|^2 = \|\widehat{D}^{-1} (I \; 0) \, \widetilde{\delta b}\|^2$. Note that since $\widetilde{V}$ is orthogonal $I = \widetilde{V}_{11}^T \widetilde{V}_{11} + \widetilde{V}_{21}^T \widetilde{V}_{21}$ and therefore $\widehat{D}^{-2} = \widehat{D}^{-1} \widetilde{V}_{11}^T \widetilde{V}_{11} \widehat{D}^{-1} + \widehat{D}^{-1} \widetilde{V}_{21}^T \widetilde{V}_{21} \widehat{D}^{-1}$. Rewriting (12) as $T\widetilde{V} = \widetilde{U} D$ and since T is lower triangular, it follows that $\widetilde{V}_{11} \widehat{D}^{-1} = \widehat{T}^{-1} \widetilde{U}_{11}$. We may conclude that

$$\|\widehat{A}_S^+ \, \delta b\|^2 = \widetilde{\delta b}^T \begin{pmatrix} \widetilde{U}_{11}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{11} + \widehat{D}^{-1} \widetilde{V}_{21}^T \widetilde{V}_{21} \widehat{D}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \widetilde{\delta b}.$$

Next note that $\|\widehat{A}_T^+ \, \delta b\|^2 = \|(\widehat{V} \widehat{T}^{-1} \widehat{U}^T) \, U_S \, \widetilde{\delta b}\|^2 = \|\widehat{T}^{-1} (\widetilde{U}_{11} \; \widetilde{U}_{12}) \, \widetilde{\delta b}\|^2$. Therefore

$$\|\widehat{A}_T^+ \, \delta b\|^2 = \widetilde{\delta b}^T \begin{pmatrix} \widetilde{U}_{11}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{11} & \widetilde{U}_{11}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{12} \\ \widetilde{U}_{12}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{11} & \widetilde{U}_{12}^T \widehat{T}^{-T} \widehat{T}^{-1} \widetilde{U}_{12} \end{pmatrix} \widetilde{\delta b}.$$

It now follows that

$$\|\widehat{A}_T^+ \, \delta b\|^2 = \|\widehat{A}_S^+ \, \delta b\|^2 + \widetilde{\delta b}^T M \, \widetilde{\delta b}. \tag{13}$$

Using $\widehat{A}_S^+ = \widehat{V}_S \widehat{D}^{-1} \widehat{U}_S^T$ it follows that the regularization error term for the truncated SVD satisfies $\|(\widehat{A}_S^+ A - I) \, x_0\|^2 = \|(\widehat{A}_S^+ A - I) V_S \, \widetilde{x}_0\|^2 = \|(\widehat{V}_S \widehat{V}_S^T - I) V_S \, \widetilde{x}_0\|^2 = \| - (0 \; I) \, \widetilde{x}_0\|^2$. Also $\|(\widehat{A}_T^+ A - I) \, x_0\|^2 = \|(\widehat{V} \widehat{V}^T - I) V_S \, \widetilde{x}_0\|^2 = \|V_0 V_0^T V_S \, \widetilde{x}_0\|^2 = \|V_0^T (\widehat{V}_S \; V_{S0}) \, \widetilde{x}_0\|^2 = \|(\widetilde{V}_{21} \; \widetilde{V}_{22}) \, \widetilde{x}_0\|^2$. Using this result and $I = \widetilde{V}_{22}^T \widetilde{V}_{22} + \widetilde{V}_{12}^T \widetilde{V}_{12}$ (since $\widetilde{V}$ is orthogonal), it follows that

$$\|(\widehat{A}_T^+ A - I) \, x_0\|^2 = \widetilde{x}_0^T \left[\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} + N \right] \widetilde{x}_0 = \|(\widehat{A}_S^+ A - I) \, x_0\|^2 + \widetilde{x}_0^T N \, \widetilde{x}_0.$$

The theorem follows from this equation, (6), and (13). ∎

Note that for a UTV factorization chosen so that $M \neq 0$ and $N \neq 0$, it follows from (9) and (10) that M and N are symmetric indefinite matrices. Therefore, if $M, N \neq 0$ in the UTV factorization of any matrix A, by (11) there exists solution vectors x_0 and noise vectors δb such that the truncated UTV solution is closer to

x_0 than is the truncated SVD solution. We will see in our numerical experiments in section 6 that it is frequently true that x_T is closer to x_0 than x_S is (and, conversely, that x_S is frequently closer to x_0 than x_T is). In sections 4 and 5 we will use Theorem 1 to explore reasons why this is true.

A result from [21] that we will need later relates the singular values of A, $\widehat{T}$, E, and G. If $\|E\| < \sigma_k(\widehat{T})$ and if T is lower triangular, then

$$\sigma_j(\widehat{T}) \le \sigma_j(A) \le \sigma_j(\widehat{T}) \Big/ \left[1 - \frac{\|H\|^2}{\sigma_k^2(\widehat{T}) - \|E\|^2}\right]^{1/2} \quad \text{for } 1 \le j \le k \tag{14}$$

and

$$\sigma_{k+j}(A) \le \sigma_j(E) \le \sigma_{k+j}(A) \Big/ \left[1 - \frac{\|H\|^2}{\sigma_k^2(\widehat{T}) - \|E\|^2}\right]^{1/2} \quad \text{for } 1 \le j \le n-k. \tag{15}$$

If T is upper triangular and if $\|G\| < \sigma_k(\widehat{T})$, then (14) and (15) are also true with H and E replaced, respectively, by F and G.

In the later sections we will also use some of the results [9] which we collect here. These results bound $\sin\theta$, the sine of the angle between the subspaces spanned by $\widehat{V}$ and $\widehat{V}_S$, and $\sin\phi$, the sine of the angle between the subspaces spanned by $\widehat{U}$ and $\widehat{U}_S$. Let $\widetilde{U}$ and $\widetilde{V}$ be defined by (7) and (8). Assume that $\|E\| < \sigma_k(\widehat{T})$ and $\|G\| < \sigma_k(\widehat{T})$. If T is lower triangular, then

$$\sin\phi = \|\widetilde{U}_{12}\| = \|\widetilde{U}_{21}\| \le \frac{\sigma_k(\widehat{T})\|H\|}{\sigma_k^2(\widehat{T}) - \|E\|^2} \qquad \text{and} \tag{16}$$

$$\sin\theta = \|\widetilde{V}_{12}\| = \|\widetilde{V}_{21}\| \le \frac{\|H\|\|E\|}{\sigma_k^2(\widehat{T}) - \|E\|^2}. \tag{17}$$

If T is upper triangular, then

$$\sin\phi = \|\widetilde{U}_{12}\| = \|\widetilde{U}_{21}\| \le \frac{\|F\|\|G\|}{\sigma_k^2(\widehat{T}) - \|G\|^2} \qquad \text{and} \tag{18}$$

$$\sin\theta = \|\widetilde{V}_{12}\| = \|\widetilde{V}_{21}\| \le \frac{\sigma_k(\widehat{T})\|F\|}{\sigma_k^2(\widehat{T}) - \|G\|^2}. \tag{19}$$

Note that in most cases of interest to us, we will have $\|H\| \le \|E\|$ and $\|F\| \le \|G\|$. Assuming this, $\sin\theta$ and $\sin\phi$ can be small for either of two reasons: (1) $\|E\| \ll \sigma_k(\widehat{T})$ and $\|G\| \ll \sigma_k(\widehat{T})$ or (2) $\|H\| \ll \|E\|$ and $\|F\| \ll \|G\|$. The first condition will be true if there is a sufficiently large gap, at singular value k, in the singular values of A and if the UTV factorization is rank-revealing (as defined in the next section). The second condition can be achieved by some of the algorithms for calculating UTV factorizations, even when there is not a gap in the singular values.

3. Calculating UTV factorizations. There are a number of algorithms for calculating UTV factorizations [11, 21, 25]. We will discuss the algorithm in [21] and a variation of this algorithm. One nice feature of this algorithm is that if the

algorithm is stopped after one step, it produces a UTV factorization which uses a single QR factorization and, as the algorithm continues with more steps, it approaches the SVD [21]. The algorithm in [21] does not include interchanges in the columns of A. We will consider a variation that can include column interchanges in the algorithm. At step i the algorithm produces the factorization $A = U_i T_i V_i^T$.

ALGORITHM.

For $i = 1$ let $A = U_1 T_1 V_1^T$, where U_1, T_1, and V_1 are formed either by $A = Q_1 R_1 \Pi_1^T$ with $U_1 = Q_1, T_1 = R_1$, and $V_1 = \Pi_1$ or $A^T = Q_1 R_1 \Pi_1^T$ with $U_1 = \Pi_1, T_1 = R_1^T$, and $V_1 = Q_1$. In this second case we will also use L_1 to indicate T_1 since T_1 is lower triangular.

For $i \geq 2$, if T_{i-1} is upper triangular, form $T_{i-1}^T = Q_i R_i \Pi_i^T$ and let

$$T_i \equiv L_i = R_i^T,\ U_i = U_{i-1}\Pi_i,\ V_i = V_{i-1}Q_i.$$

For $i \geq 2$, if T_{i-1} is lower triangular, form $T_{i-1} = Q_i R_i \Pi_i^T$ and let

$$T_i = R_i,\ U_i = U_{i-1}Q_i,\ V_i = V_{i-1}\Pi_i.$$

To determine when to stop this algorithm one can use the bounds on $\|x_T - x_S\|$ in Theorem 3.3 of [10]. We will select the initial permutation Π_1 using the standard pivoting technique [1, 4, 20] for QR factorizations and let $\Pi_i = I$ for $i \geq 2$. We will also consider a variation where at each step Π_i is chosen by the standard pivoting technique. We will see shortly that often the two alternatives produce identical low-rank solutions. In later section, when we use "the algorithm" or "the algorithm of section 3," we will refer to the first, simpler alternative ($\Pi_i = I,\ i \geq 2$).

We will find it useful to introduce notation to describe the first few steps of the algorithm. When T_1 is upper triangular we use QRP to indicate the first step of the algorithm, QRLP the next step, QRLRP the next step, etc. When T_1 is lower triangular we use QLP for the first step, QLRP for the next step, QLRLP for the next step, etc. Here the Q indicates that we are using QR factorizations, P indicates that we use pivoting at the first step, and the middle letters indicate the history of the steps of the algorithm. When we are calculating x_T using one of these factorizations we will use TQRP, TQRLP, TQLP, TQLRP, etc., to indicate that we are using a truncated factorization—we need only to calculate a portion of U, V, and T.

The above algorithm, without column interchanges, was used in [21] to calculate the SVD. The paper [21] focuses on a block implementation of the above algorithm. Results concerning the convergence of the above algorithm as a tool to estimate singular values and singular vectors is discussed in [7]. Stewart [26, 27] discusses QRLP, with pivoting at both steps, as a tool for estimating singular values and for constructing low-rank approximations. We should note that Stewart uses the designation QLP to refer to what we have called QRLP. In [19] the TQLRP algorithm is described and an example is presented where it works well for regularization. Also we should note that the block-row low-rank approximation produced by TQRP is the same as the approximate solution to (1) produced by LAPACK's xGELSY [1] or by the algorithm HFTI in [20]. Mathematically these algorithms are identical. Our comparison in this paper of TQRP with truncated SVD provides a comparison of the accuracy of xGELSY and xGELSD, the two recommended tools in LAPACK 3.0 for solving rank-deficient problems.

We now show that there is a close relationship between solutions produced by block-row/column low-rank approximations and by corner low-rank approximations.

We also show that often identical low-rank solutions to (1) are produced by the algorithm if $\Pi_i = I,\ i \geq 2$, or if Π_i is chosen by standard column pivoting.

THEOREM 2. *Let $U_i T_i V_i^T$ be the decomposition of A at step $i \geq 1$ of the algorithm. Using the notation of* (2) *and* (3) *with subscripts added to indicate the step number in the algorithm, define $\rho_i = \|E_i\|/\sigma_k(\widehat{T}_i)$ if T_i is lower triangular and $\rho_i = \|G_i\|/\sigma_k(\widehat{T}_i)$ if T_i is upper triangular.*

(a) *Assume $\widehat{T}_{i-1}$ is nonsingular and that Π_i in the algorithm has the form $\Pi_i = \left(\begin{smallmatrix} \Pi_a & 0 \\ 0 & \Pi_b \end{smallmatrix}\right)$, where Π_a and Π_b, respectively, are $k \times k$ and $(n-k)\times(n-k)$ permutation matrices. Then the block-row/column rank k solution x_B calculated from $U_{i-1}T_{i-1}V_{i-1}^T$ is the same as the corner rank k solution x_C calculated from $U_i T_i V_i^T$.*

(b) *If the initial factorization has the property that $\rho_1 < 1$ and if Π_i, $i \geq 2$, is chosen by the standard pivoting algorithm of* [4]*, then for all $i \geq 2$, Π_i is of the form required in part* (a).

(c) *Assume that $\rho_1 < 1$. Then the corner low-rank solution produced at each step of the algorithm using standard column pivoting is identical to the corner low-rank solution produced at the corresponding step of the algorithm where standard column pivoting is used at the first step and no pivoting is used at each following step.*

Proof. Assume that T_{i-1} is lower triangular. We will use the notation of the algorithm, equation (2), and equation (3) except that we will add subscripts to E, $\widehat{V}$, $\widehat{L}$, H, $\widehat{R}$ to indicate the step number of the algorithm.

To prove part (a), note that at step $i-1$ the rank k block-row/column approximation to A is $\widehat{A}_{i-1} = U_{i-1}(\widehat{L}_{i-1}^T\ H_{i-1}^T\ 0)^T \widehat{V}_{i-1}^T$ and by properties of pseudoinverses [3, 20],

$$x_B = \widehat{A}_{i-1}^+ b = \widehat{V}_{i-1} \begin{pmatrix} \widehat{L}_{i-1} \\ H_{i-1} \\ 0 \end{pmatrix}^+ U_{i-1}^T b.$$

However, by our assumption on Π_i and by the constructions of the algorithm,

$$\begin{pmatrix} \widehat{L}_{i-1} \\ H_{i-1} \\ 0 \end{pmatrix} = Q_i \begin{pmatrix} \widehat{R}_i \\ 0 \end{pmatrix} \Pi_a^T \text{ and so } \begin{pmatrix} \widehat{L}_{i-1} \\ H_{i-1} \\ 0 \end{pmatrix}^+ = \Pi_a(\widehat{R}_i^{-1}\ 0)Q_i^T.$$

It now follows that $x_B = \widehat{V}_{i-1}\Pi_a(\widehat{R}_i^{-1}\ 0)Q_i^T U_{i-1}^T b = \widehat{V}_i(\widehat{R}_i^{-1}\ 0)U_i^T b = x_C$. The proof for the case that T_{i-1} is upper triangular is similar.

To show part (b) we again assume that T_{i-1} is lower triangular. We will use induction. Assume that after the first step of the algorithm and prior to step i, the permutation matrices in the algorithm have the form of part (a). It then follows easily from Theorem 2.1 of [21] and its proof that $\sigma_k(\hat{L}_{i-1}) \geq \sigma_k(\widehat{T}_1)$ and, if T_1 is lower triangular, $\|E_1\| \geq \|E_{i-1}\|$ or, if T_1 is upper triangular, $\|G_1\| \geq \|E_{i-1}\|$. Therefore, by the assumption of part (b), $\sigma_k(\widehat{L}_{i-1}) > \|E_{i-1}\|$. If Π_i is of the form of part (a), then for $1 \leq j \leq k$, it follows that $|r_{jj}| \geq \sigma_k(\,(\widehat{L}_{i-1}^T, H_{i-1}^T)\,) \geq \sigma_k(\widehat{L}_{i-1})$. Now suppose, on the other hand, that at step i the column interchanges in the standard pivoted QR factorization move a column of L_{i-1} with column index larger than k into column j, where $1 \leq j \leq k$. The diagonal entry r_{jj} in the QR factorization of L_{i-1} will satisfy $|r_{jj}| \leq \|E_{i-1}\|$. It follows that this last type interchange is not possible since $\sigma_k(\widehat{L}_{i-1}) > \|E_{i-1}\|$ and since standard column pivoting will move to column j, the column of the remaining unprocessed columns that will make $|r_{jj}|$ as large as possible. The proof when T_{i-1} is upper triangular is similar.

Part (c) follows from part (b) and the block structure of Π_i for $i \geq 2$. Our proof is somewhat tedious and we omit it here. Contact the author for the details. ∎

In [27] Stewart notes that in the QRLP factorization if there is a substantial gap in diagonal entries of R_1, "it is unlikely that the pivoting process will interchange columns" across column k in constructing L_2. Theorem 2 proves under a mild condition on R_1 ($\rho_1 < 1$) that Stewart's observation is true. Note that $\rho_1 < 1$ will be true if there is a modest gap in the singular values of A and if the initial QR factorization is rank-revealing [3, p. 22]. Part (c) shows that if $\rho_1 < 1$, pivoting is not necessary after the first step in the algorithm in the sense that the corner solutions are the same with pivoting or without pivoting. This is also true for block-row/column solutions by part (a) of the theorem. Our numerical experience indicates that, even when $\rho_1 > 1$, pivoting at steps after the first step usually makes little difference in the quality of the solution. However, pivoting at the first step is often critical.

The theorem also shows that there is a close connection between solutions produced by rank-revealing QR factorizations and rank-revealing UTV factorizations. A rank-revealing QR factorization of A [3, p. 22] has the properties $\|G\| = O(s_{k+1})$ and $\sigma_k(\widehat{R}) = O(s_k)$. A rank-revealing ULV factorization of A [10, p. 456] has the properties $\|(H\ E)\| = O(s_{k+1})$ and $\sigma_k(\widehat{L}) = O(s_k)$. Assume that, at step 1 of the algorithm, $A = U_1 T_1 V_1^T = QR\Pi^T$ is a rank-revealing QR factorization of A and that Π_2 has the form of part (a) of Theorem 2, for example, if no pivoting is done at step 2. It is not hard to show that the UTV factorization of step 2 of the algorithm will be a rank-revealing ULV factorization of A. It follows from Theorem 2 that the regularized solution produced from a block-row low-rank approximation using a rank-revealing QR factorization is identical to the corner low-rank solution using a related rank-revealing UTV factorization.

Another useful consequence of part (a) is that the results that we develop for corner low-rank solutions to (1) lead directly to results for block-row/column low-rank solutions to (1).

For some of our later results we will need to assume that the UTV factorization is rank-revealing. The first factorization in the algorithm uses the QR factorization with standard column pivoting. There are contrived examples [3, p. 105] where standard column pivoting is not rank-revealing. To overcome this potential problem the algorithm could be started with a QR factorization that guarantees to be reveal rank [3, pp. 22, 108] or one could include pivoting at the second step [27]. Our numerical experience suggests that this is not necessary for examples that are not contrived.

Finally, for later use we would like to present some of the results of [21] that concern the convergence of the algorithm. We define ρ_i as in Theorem 2.

$$\text{If } T_i \text{ is lower triangular, then } \|H_i\| \leq \rho_1 \rho_2 \dots \rho_{i-1} \sigma_k(\widehat{T}_1) \leq \rho_1^{i-1} \sigma_k(\widehat{T}_1). \tag{20}$$

$$\text{If } T_i \text{ is upper triangular, then } \|F_i\| \leq \rho_1 \rho_2 \dots \rho_{i-1} \sigma_k(\widehat{T}_1) \leq \rho_1^{i-1} \sigma_k(\widehat{T}_1). \tag{21}$$

The inequalities (20) and (21) indicate that if $\rho_1 < 1$, then the off-diagonal blocks F_i and G_i are forced to zero as the algorithm proceeds. These results combined with (14) and (15) show that the singular values of $\widehat{T}_i$, E_i, and G_i converge to singular values of A. Inequalities (20) and (21) combined with (16)–(19) show that if $\rho_1 < 1$, then $\sin\phi$ and $\sin\theta$ approach zero as the algorithm proceeds.

4. Perturbation errors. We now compare the perturbation error terms in (4), (5), and (6) for corner low-rank approximations calculated using a truncated UTV

decomposition with the corresponding error terms when using a truncated SVD decomposition. We will assume in (4) that the regularization error term $(CA - I)x_0$ is sufficiently small so that

$$x - x_0 = C(\delta b) \tag{22}$$

is a good approximation. We will also assume initially that the UTV factorization has T lower triangular. As we noted following Theorem 1, equation (11) implies that in some cases $\|x_T - x_0\|$ will be smaller than $\|x_S - x_0\|$. The following theorem concerns the probability that this occurs, if we assume (22).

THEOREM 3. *Let x_T be calculated using a corner low-rank UTV approximation to A with T lower triangular. Assume that* (22) *is true, that $\|E\| < \sigma_k(\widehat{T})$, and that the components of δb come from uncorrelated zero mean Gaussian random variables with common variance (Gaussian white noise). Then as* $\sin\phi$ *approaches* 0 *the probability that $\|x_T - x_0\|$ is less than $\|x_S - x_0\|$ approaches one-half.*

Proof. Due to (22) in (11), we can assume that N is 0. By (9) we can write M as $M = \left(\begin{smallmatrix} -M_{11} & M_{12} \\ M_{12}^T & M_{22} \end{smallmatrix}\right)$ with $M_{11} = \widehat{D}^{-1}\widetilde{V}_{21}^T\widetilde{V}_{21}\widehat{D}^{-1}$, $M_{12} = \widetilde{U}_{11}^T\widehat{T}^{-T}\widehat{T}^{-1}\widetilde{U}_{12}$, and $M_{22} = \widetilde{U}_{12}^T\widehat{T}^{-T}\widehat{T}^{-1}\widetilde{U}_{12}$. Then it follows that

$$\|M_{11}\| \le \frac{s_{k+1}}{s_k}(\tan\theta)\,\|M_{12}\| \le (\tan\phi)\,\|M_{12}\| \text{ and } \|M_{22}\| \le (\tan\phi)\,\|M_{12}\|, \tag{23}$$

where $\tan\phi = \sin\phi/\sqrt{1-\sin^2\phi}$ and $\tan\theta = \sin\theta/\sqrt{1-\sin^2\theta}$. These inequalities follow from (12), $\|E\| < \sigma_k(\widehat{T})$, and the identities $M_{22} = \widetilde{U}_{12}^T\widetilde{U}_{11}^{-T}M_{12}$ and $M_{11} = M_{12}\,(D_0\,0)^T\,\widetilde{V}_{22}^{-1}\widetilde{V}_{21}\widehat{D}^{-1}$, which are consequences of (12) and properties of orthogonal matrices. For $\sin\phi$ small (23) implies that the diagonal blocks of M are small relative to the off-diagonal blocks. Consider the matrix $\widetilde{M} = \left(\begin{smallmatrix} 0 & M_{12} \\ M_{12}^T & 0 \end{smallmatrix}\right)$ formed by the off-diagonal blocks. Note that the eigenvalues of $\widetilde{M}$ come in plus and minus pairs of equal magnitude. Since we are assuming that δb is governed by Gaussian white noise, it follows from Theorem 4.4.8 and Corollary 5.4.2 of [24] that the distribution governing $\widetilde{\delta b}^T\widetilde{M}\widetilde{\delta b}$ is symmetric and that the probability that $\widetilde{\delta b}^T\widetilde{M}\widetilde{\delta b}$ is negative is one-half. Due to (23), the theorem follows from a continuity argument. ∎

By the comments following (19), $\sin\phi$ will be small when $\|H\|$ is sufficiently small or when there is a sufficiently large gap in the singular values of A and the UTV factorization is rank-revealing. It follows under the conditions of the theorem that if $\sin\phi$ is small, then x_T will be closer to x_0 than x_S is approximately half the time and, conversely, x_S will be closer approximately half the time. Our numerical experiments support this. They also suggest that in some cases even when $\sin\phi$ is not small x_T is still frequently as close or closer to x_0 than x_S is.

We assumed in this theorem that the noise is Gaussian white noise. According to [28], "Gaussian white noise is a common occurrence in many signal processing systems."

It is also of interest to look at the expected value of $\|x_T - x_0\|^2$ relative to the expected value of $\|x_S - x_0\|^2$, which we do in Theorem 4. The following lemma is used in the proof of Theorem 4.

LEMMA 1. *Assume that $u \in R^m$ has components that come from uncorrelated zero mean random variables with common variance (white noise) and that the expected value of $\|u\|^2$, indicated by $E(\|u\|^2)$, is Δ^2. Then for an $m \times m$ matrix A, $E(u^T A\,u) = \Delta^2\,\mathrm{trace}(A)/m$. Also for an $n \times m$ matrix A, $E(\|A\,u\|^2) = \Delta^2\,\|A\|_F^2/m$, where $\|A\|_F$ indicates the Frobenius norm.*

Proof. Since $\Delta^2 = E(\Sigma_{i=1}^m u_i^2) = \Sigma_{i=1}^m E(u_i^2)$, then $E(u_i^2) = \Delta^2/m$. Now for an $m \times m$ matrix A, $E(u^T Au) = \Sigma\Sigma a_{ij} E(u_i u_j) = \Sigma a_{ii} E(u_i^2) = \text{trace}(A)\Delta^2/m$. Also for an $n \times m$ matrix A, $E(\|Au\|^2) = E(u^T A^T Au) = \Delta^2 \text{trace}(A^T A)/m = \Delta^2 \|A\|_F^2/m$. ∎

THEOREM 4. *Let x_T be calculated using a corner low-rank UTV approximation to A. Define $\sin\phi$ as in* (16). *Assume that T is lower triangular, that* (22) *is true, and that the components of δb correspond to white noise. Then*

$$0 \leq \frac{E(\|x_T - x_0\|^2) - E(\|x_S - x_0\|^2)}{E(\|x_T - x_0\|^2)} \leq \sin^2\phi. \tag{24}$$

Proof. Let us assume that $E(\|\delta b\|^2) = \Delta^2$. Since $\|x_S - x_0\|^2 = \|\widehat{A}_S^+ \delta b\|^2 = \|\widehat{V}_S \widehat{D}^{-1} \widehat{U}_S^T \delta b\|^2$ and $\|x_T - x_0\|^2 = \|\widehat{A}_T^+ \delta b\|^2 = \|\widehat{V}\widehat{T}^{-1}\widehat{U}^T \delta b\|^2$, it follows from the lemma that $E(\|x_S - x_0\|^2) = \Delta^2 \|\widehat{D}^{-1}\|_F^2/m$ and $E(\|x_T - x_0\|^2) = \Delta^2 \|\widehat{T}^{-1}\|_F^2/m$. The left inequality in (24) is true since $E(\|x_S - x_0\|^2) = \Delta^2 \|\widehat{D}^{-1}\|_F^2/m$, $E(\|x_T - x_0\|^2) = \Delta^2 \|\widehat{T}^{-1}\|_F^2/m$, the Frobenius norm squared is the sum of the square of the singular values, and the left inequality in (14). By (11), (22), and Lemma 1, $E(\|x_T - x_0\|^2) - E(\|x_S - x_0\|^2) = E(\widetilde{\delta b}^T M \widetilde{\delta b}) = \Delta^2 \text{trace}(M)/m = \Delta^2[\text{trace}(M_{22}) - \text{trace}(M_{11})]/m \leq \Delta^2 \text{trace}(M_{22})/m = \Delta^2 \|\widehat{T}^{-1}\widetilde{U}_{12}\|_F^2/m \leq \Delta^2 \|\widetilde{U}_{12}\|^2 \|\widehat{T}^{-1}\|_F^2/m = \Delta^2 (\sin^2\phi) \|\widehat{T}^{-1}\|_F^2/m$. The theorem now follows. ∎

The left-hand inequality in (24) implies under the conditions of the theorem that the truncated SVD solutions will, on average, be better than truncated UTV solutions. However, the right-hand term suggests, as we will see in our numerical experiments, that often the difference, on average, will not be large and the truncated UTV and truncated SVD will be similar in accuracy. Note that by the comments following (19) the size of $\sin\phi$ is related to the size of a gap in the singular values of A and to the size of H. Also note that $\sin^2\phi$ in (24) can be small even for modest $\sin\phi$.

Theorems 3 and 4 are applicable to corner low-rank UTV approximations when T is lower triangular. When T is upper triangular one can prove, although we will not do so here, that (24) is valid except that $\sin^2\phi$ must be replaced by $\sin^2\theta$. Our numerical experiments suggest that there are results similar to Theorem 3 for the case when T is upper triangular.

5. Regularization errors. We now compare the regularization error terms in (4), (5), and (6) for corner low-rank approximations calculated using a truncated UTV decomposition with the corresponding error terms when using a truncated SVD decomposition. We will assume in (4) that the perturbation error term $C(\delta b)$ is sufficiently small so that

$$x - x_0 = (CA - I)x_0 \tag{25}$$

is a good approximation. We will also assume that the UTV factorization has T lower triangular.

Some of our results in this section will involve the values of components of $U_S^T b_0$. The discrete Picard condition [14, p. 507] is that these components decay to zero somewhat faster than the singular values. The condition is required for regularization to produce useful solutions [13, 14, 15]. We will call these components of $U_S^T b_0$ the "Picard coefficients" to indicate their connection to the Picard condition (the term Fourier coefficients is sometimes used). If we let $\widetilde{D}$ be the $n \times n$ diagonal matrix consisting of the first n rows of D in the SVD $A = U_S D V_S^T$, we will model the rate of

decrease of the Picard coefficients by assuming that the first n components of $U_S^T b_0$ are equal to the components of $\widetilde{D}^{p+1}w$ where $\widetilde{D}^{p+1}$ indicates the $(p+1)$st power of $\widetilde{D}$, $p \geq 0$, and w is a vector whose components do not depend on p or the singular values of A. Following Hansen [13, 14, 15], who defines a similar parameter, we will call p the relative decay rate of the Picard or Fourier coefficients.

It will be useful to assume a particular form for the underlying noiseless solution x_0. We assume that

$$x_0 = V_S \widetilde{D}^p w, \tag{26}$$

where $\widetilde{D}$ is first n rows of D. We have two motivations for this choice. First, with this x_0, $b_0 = Ax_0 = U_S(\widetilde{D}^{p+1} \quad 0)^T w$ so that the first n Picard coefficients are $\widetilde{D}^{p+1}w$. Therefore, p is the relative decay rate of the Picard coefficients. If $p > 0$, the Picard condition will be satisfied. Also note that by (26) x_0 is a linear combination of the singular vectors of A. Due to the factor $\widetilde{D}^p$, if the singular values decrease sufficiently rapidly or if p is sufficiently large, the contribution of higher index singular vectors will be small. It is often the case that the lower index singular vectors correspond to smoothly varying functions [17, 18]. If these assumptions are true, as is frequently the case, x_0 will be smoothly varying. We also note that (26) is equivalent to the model [22, p. 640] for characterizing smooth solutions x_0. We conclude that (26) provides a method that has been used by others to generate a class of smoothly varying solutions x_0 that satisfy the Picard condition.

Our results will involve the decay rate p of the Picard coefficients for smaller values of p since these values of p appear to be useful in many practical applications. For example, we looked at 16 examples from Hansen's regularization tools [16]. Most of the problems in [16] come from the literature and all share characteristic features of ill-posed problems. For each example we made a rough estimate of p by estimating the slope of a graph of the log of the Picard coefficients versus the log of the singular values (for values not dominated by errors). In 14 of the 16 cases the rough estimate was 1 or less. Our theorems in this section will assume $0 \leq p \leq 1$ (Theorem 5) and $0 \leq p \leq 2$ (Theorem 6).

As we noted following Theorem 1 and in section 4, equation (11) implies that in some cases $\|x_T - x_0\|$ will be smaller than $\|x_S - x_0\|$. The following theorem concerns the probability that this occurs if we assume (25).

THEOREM 5. *Let x_T be calculated using a corner low-rank UTV approximation to A with T lower triangular. Assume that x_0 satisfies* (26) *with $0 \leq p \leq 1$, that* (25) *is true, that $\|E\| < \sigma_k(\widehat{T})$, and that w follows Gaussian white noise. Then as $\sin\phi$ approaches* 0*, the probability that $\|x_T - x_0\|$ is less than $\|x_S - x_0\|$ approaches one-half.*

Proof. Due to (25) it follows that M is 0 in (11). Due to (26) and (10) we can write $\widetilde{x}_0^T N \widetilde{x}_0 = w^T N_p w$, where

$$N_p = \widetilde{D}^p N \widetilde{D}^p = \begin{pmatrix} \widehat{D}^p \widetilde{V}_{21}^T \widetilde{V}_{21} \widehat{D}^p & \widehat{D}^p \widetilde{V}_{21}^T \widetilde{V}_{22} D_0^p \\ D_0^p \widetilde{V}_{22}^T \widetilde{V}_{21} \widehat{D}^p & -D_0^p \widetilde{V}_{12}^T \widetilde{V}_{12} D_0^p \end{pmatrix} = \begin{pmatrix} N_{11} & N_{12} \\ N_{12}^T & -N_{22} \end{pmatrix}. \tag{27}$$

By (12) and properties of orthogonal matrices it follows that $N_{11} = N_{12}(D_0^{1-p} \quad 0) \times \widetilde{U}_{22}^{-1}\widetilde{U}_{21}\widehat{D}^{p-1}$ and $N_{22} = D_0^p \widetilde{V}_{12}^T \widetilde{V}_{11}^{-T} \widehat{D}^{-p} N_{12}$. From these identities, $0 \leq p \leq 1$, and $\|E\| < \sigma_k(\widehat{T})$, it follows that $\|N_{11}\| \leq (s_{k+1}/s_k)^{1-p}(\tan\phi)\|N_{12}\| \leq (\tan\phi)\,\|N_{12}\|$ and that $\|N_{22}\| \leq (s_{k+1}/s_k)^p\,(\tan\theta)\,\|N_{12}\| \leq (\tan\phi)\,\|N_{12}\|$. The rest of the proof follows in a very similar manner to the proof of Theorem 3. ∎

It follows under the conditions of the theorem that if $\sin\phi$ is small, then $\|x_T - x_0\|$ will be smaller than $\|x_S - x_0\|$ approximately half the time and, conversely, $\|x_S - x_0\|$ will be smaller approximately half the time. One condition of the theorem is that $0 \le p \le 1$. Numerical experiments suggest that the conclusion of the theorem is also true for $1 \le p < 2$. They also suggest that in some cases even when $\sin\phi$ is not small x_T is still frequently as close or closer to x_0 than x_S is.

For regularization errors we can also prove a useful result about the expected value of the errors.

THEOREM 6. *Let x_T be calculated using a corner low-rank UTV approximation to A. Assume that T is lower triangular, that* (25) *is true, that x_0 satisfies* (26) *with $0 \le p \le 2$, and that the components of w correspond to white noise. If*

$$\alpha = \left(\frac{s_k}{s_{k+1}}\right)^p \left[\, \|H\| + (\sin\theta)\,\|E\| \,\right] \|E\|_F \,/\, s_k^2, \quad \textit{then} \tag{28}$$

$$-\sin^2\theta \le \frac{E(\|x_T - x_0\|^2) - E(\|x_S - x_0\|^2)}{E(\|x_S - x_0\|^2)} \le \alpha^2. \tag{29}$$

Proof. Lemma 1 and (25) imply for $p \ge 0$ that $E(\|x_S - x_0\|^2) = \tau^2 \|D_0^p\|_F^2 \,/\, n$, where $\tau^2 \equiv E(\|w\|^2)$. Lemma 1, (11), (25), and (27) imply that $E(\|x_T - x_0\|^2) - E(\|x_S - x_0\|^2) = \tau^2 \mathrm{trace}(N_p)/n$. By (27) it follows that $-(\sin^2\theta)\|D_0^p\|_F^2 \le -\|\widetilde{V}_{12} D_0^p\|_F^2 \le \mathrm{trace}(\mathrm{N_p}) \le \|\widetilde{V}_{21}\widehat{D}^p\|_F^2$. For T lower triangular (12) implies that $\widetilde{V}_{21}\widehat{D}^2 = E^T H \widetilde{V}_{11} + E^T E \widetilde{V}_{21}$. Therefore, for $0 \le p \le 2$ we have $\mathrm{trace}(N_p) \le \|\widetilde{V}_{21}\widehat{D}^2\widehat{D}^{p-2}\|_F^2 \le \|E^T H \widetilde{V}_{11} + E^T E \widetilde{V}_{21}\|_F^2 s_k^{2(p-2)} \le s_k^{2p-4}(\|H\| + (\sin\theta)\,\|E\|)^2 \|E\|_F^2$. Since $s_{k+1}^{2p} \le \|D_0^p\|_F^2$, the theorem follows. ∎

For TQRLP or at any subsequent step of the algorithm with T lower triangular, it follows from (16), (17), (20), and (28) that, for $0 \le p < 2$, $\sin\phi$, $\sin\theta$, and α will be small either if $\|H\|$ is sufficiently small or if the UTV factorization is rank-revealing and there is a sufficiently large gap in the singular values. Note that $\sin^2\theta$ and α^2 may be small even for modest $\sin\theta$ and α.

The right-hand bound in (29) increases in magnitude as p increases. This suggests that the solutions produced by a truncated UTV factorization will be best, relative to those produced by the SVD, for smaller values of p. As mentioned earlier, in practice, values of p one or less appear to be common. For larger values of p, accuracy close to that of the SVD can be achieved by using additional steps in the algorithm. As seen in (20), if $\rho_1 < 1$, these steps will force $\|H_i\|$ and $\sin\theta_i$ to become small. Note that [22, p. 644] discusses the effect of p on classical Tikhonov regularization.

Theorems 5 and 6 assume (26), $x_0 = V_S \widetilde{D}^p w$, applies where w is governed by Gaussian white noise (Theorem 5) or white noise (Theorem 6). These may be only rough models of solutions x_0 as they appear in practical applications. However, note that Neumaier [22, p. 641] comments that a model equivalent to (26) is "a frequently used assumption" and, in addition, the model has been used with similar statistical assumptions about the components of w [2, 22]. A conclusion from Theorem 5 is, under the conditions of the theorem, that $\|x_T - x_0\|$ is frequently smaller than $\|x_S - x_0\|$. This is consistent with our experiments using examples from [16] where x_0 is not chosen randomly (see Table 2).

6. Implementation and numerical experiments. Before discussing our numerical experiments we will discuss some implementation issues for the algorithm of section 3 and the efficiency of the algorithm. For a point of comparison we note

that there are three classical methods for solving least squares problems—the QR factorization without column interchanges, the QR factorizations with column interchanges, and the SVD. The first algorithm is not reliable for solving rank-deficient problems but we include it for comparison of the efficiency of the algorithms. These algorithms are implemented in LAPACK as xGELS, xGELSY, and xGELSD. For large n and $m \geq n$ (but not too much bigger than n) the approximate flop counts are $2mn^2 - 2/3n^3$, $2mn^2 - 2/3n^3$, and $4mn^2 - 4/3n^3$ [23], respectively, for xGELS, xGELSY, and xGELSD for the full-rank case. For the case that $m = n$ these counts predict run times in the ratio 1:1:2 for the three algorithms. However, in practice xGELS makes more effective use of the potential speedup in BLAS-3 calculations and the actual run-time ratios depend on the computer architecture and matrix size. For illustrations of potential actual run-time ratios, note that LAPACK [1, p. 72] reports ratios of 1:1:4 for 900×900 matrices run on an Compaq AlphaServer DS-20, Ren [23, p. 94] reports ratios of 1:1.3:3.4 for 1600×1600 matrices run on an IBM RS 6000/590, and our numerical experiments indicate ratios of 1:2.1:4.7 for 1600×1600 matrices run on a 700 MHz Pentium computer.

We will consider the construction of a block-row/column low-rank approximate solution to (1) using the algorithm in section 3. In the algorithm, if k, the effective numerical rank, is less than n, it is not necessary to do complete QR factorizations. One can start the algorithm with a QR factorization with the usual pivoting scheme [4] and stop the initial factorization when, for example, norms of the columns of E_1 (or G_1) are small. The matrices $E_i, i \geq 1$, and $G_i, i \geq 1$, need not be factored in order to calculate the solutions, x_T, at subsequent steps of the algorithm. An efficient way to implement the algorithm is to begin with the initial partial factorization just described. At subsequent steps one can construct orthogonal factorizations that successively update the block triangular structure of T while keeping the structure (lower or upper triangular) of $\widehat{T}$ fixed. The solution x_T produced by this implementation of the algorithm is identical to the solution produced by the implementation described in section 3. With this implementation the overall flop count for i steps of the algorithm is approximately $4k^2(n - 2/3k) + 2k(n - k)(2n - k)i$, where, for simplicity, the count is for the case where $m = n$. For any $m \geq n$ and for $i = 1$, the flop count for the algorithm is approximately $4mnk - 2k^2m - 2k^3/3$. For $k < n$ this is less than the flop count for xGELSY. Also we can show that for $m \geq n$ and $i \leq 2$, the flop count of the algorithm is less than the theoretical flop count for xGELSD. The advantage of the algorithm is most striking in the low-rank case where $k \ll n$. In this case, for $m \geq n$, the leading-order term in the flop count is $4kmni$, which is substantially less than the theoretical counts for xGELSY and xGELSD. Alternative algorithms for the low-rank case are discussed in [5, 11]. The smallest flop count of the algorithms in [11] is $12mnk$. The flop count for the algorithm of section 3 for $i = 1$ will also be smaller than the count for the algorithm in [5].

The actual time required by an algorithm depends on details of its implementation and the computer architecture as well as flop counts. As discussed earlier the block-row low-rank solution produced by TQRP (or by Theorem 2 the corner low-rank solution produced by TQRLP) will be the same as the solution produced by LAPACK's xGELSY. However, LAPACK does a complete factorization of A and not a partial factorization as discussed in the last paragraph. The routine xGELSY can be modified to incorporate the partial factorization. The tests in xGELSY for the determination of the effective rank can be moved into LAPACK's factorization routine xGEQP3. If these tests are inserted in xGEQP3 immediately after the call

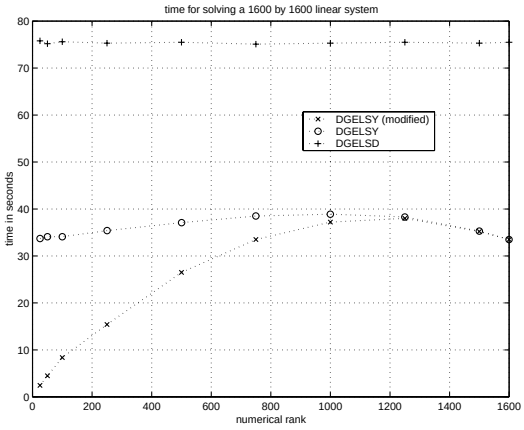


FIG. 1. *Timings for the modification of DGELSY, DGELSY, and DGELSD.*

to LAPACK's xLAQPS, the efficient BLAS-3 calls in xGEQP3 will not be affected. The solution x_T produced by this modification is identical to that of the unmodified LAPACK, but the modified algorithm will run much more quickly for low-rank problems. This is illustrated in Figure 1. The sample problems in Figure 1 were generated by LAPACK's xLATMS and have a gap in the singular values at the indicated numerical rank. They were run on a 700 MHz Pentium computer using BLAS routines supplied by Intel. From the graph it is clear that for low-rank problems the modification of xGELSY is much more efficient than the existing implementations of LAPACK's routines xGELSY and xGELSD. For $k = 25$ the run-time ratios are approximately 1:14:31. Due to this substantial speedup it also clear that in the low-rank case the algorithm of section 3 will remain more efficient than the LAPACK routines if the algorithm is continued with some additional steps.

We should add here that another issue that can be important in choosing an algorithm to solve (1) is the ability to easily do updates and downdates. This is more easily done with a UTV factorization than the SVD [25]. Also note that [26, 27] discuss implementation issues for the QRLP algorithm including the observation that in the low-rank case the savings in stopping the reduction are substantial. Finally, we should note that it is well known [12, p. 250] that truncating the QR factorization reduces the flop count in the factorization to approximately $4mnk$ for small k.

We now present experiments that focus on the accuracy of the algorithm. For our first test results we generated random 64×64 matrices A using REGUTM from [16]. We chose the singular values of A in three manners. To describe the first type of selection let us define quantities which we call the "gap" and the "spread," where the gap is s_{16}/s_{17} and the spread is $s_1/s_{16} = s_{17}/s_{64}$. Singular values s_2 to s_{15} were selected from a log-uniform distribution over s_1 to s_{16} and singular values s_{18} to s_{63} from a log-uniform over s_{17} to s_{64}. For this selection of singular values we fixed the effective numerical rank at 16. In the second selection of singular values we selected 10 singular values from a log-uniform distribution from 1 to .001, 10 singular values equal to .001, and 44 singular values from a log-uniform distribution from .001 to .000001. We again fixed the effective numerical rank at 16. This selection is designed to test the algorithm in an extreme case, forcing the algorithm to select the numerical rank in the middle of a cluster of identical singular values. To describe the third selection

of singular values, we will use a quantity which we call the "mean gap." We let $s_1 = 1$ and $s_{64} = (\text{meangap})^{63}$ and chose s_2 through s_{63} from a log-uniform distribution from s_1 to s_{64}. For this selection of singular values the geometric mean of $s_j/s_{j+1}, j = 1, 2, \ldots, 63$ equals the mean gap. In this case the singular values decay gradually and there is not an obvious gap to help select the effective rank k. A variety of approaches have been suggested in the literature [18] for selecting k. For simplicity and to focus on the approximation scheme, not the technique for selecting k, for the third choice of singular values and for each regularization algorithm, we selected the effective rank by calculating a regularized solution, x, for each numerical rank $k < n$. Among all lower rank approximations, the approximation that minimizes $\|x - x_0\|$ was selected.

For each matrix A we chose the underlying noiseless solution $x_0 = V\widetilde{D}^p w$, $b_0 = Ax_0$ and noise vectors $\delta b = \Delta\|b_0\|v$. We used seven different noise to signal ratios, $\Delta = .3, .1, .01, .001, 10^{-4}, 10^{-6}$, and 10^{-10}. This wide range of noise levels should produce cases where the regularization error dominates (Δ small), cases where the perturbation error dominates (Δ large), and cases in between these extremes. We selected 100 random matrices as described above. For each matrix and for each of the seven noise levels, we selected 100 random values of $x_0 = \widetilde{D}^p w$, with w selected from white noise and for each x_0 we selected a random noise vector δb, with v selected from white noise for a total of 70,000 ($= 100 \times 7 \times 100$) samples. For each sample we calculated x_S as well as five different solutions x_T. To calculate x_T we used block-row/column low-rank approximations for the TQLP, TQLRP, TQRP, TQRLP, and TQRLRP factorizations. To summarize the results in a concise manner, for each low-rank approximation we calculated the mean value of $\|x_T - x_0\|/\|x_S - x_0\| - 1$ over all 70,000 samples. These results are in Table 1. Also in the last column of Table 1 we indicate the percent of the cases where $\|x_T - x_0\|$ is smaller than $\|x_S - x_0\|$ for the block-row low-rank solutions using the TQRP factorization. These solutions can be produced by LAPACK's xGELSY.

Most of the entries in the table are positive, which indicates that on average the truncated SVD solutions are better. However, except for TQLP, if $p < 2$, the truncated UTV solutions are on average not far from the truncated SVD solutions. For example, for $p < 2$, $\|x_T - x_0\|$ was within 15% of $\|x_S - x_0\|$ on average for all the methods except TQLP. Remarkably, this is true for runs with a small gap or no gap in the singular values and when the numerical rank is selected in the middle of a cluster of singular values. As p increases, more steps of the algorithm are required to match the accuracy of the SVD. As mentioned earlier, smaller values of p appear to be more common in practice. Note that we arrive at these same general conclusions by looking at the cases where the rank of the low-rank approximation is fixed at 16 or the "mean gap" cases where the rank is chosen dynamically. Also note that the last column of the table indicates, as suggested by Theorems 3 and 5, that if there is a sufficiently large gap in the singular values and if p is not large, then $\|x_T - x_0\|$ is smaller than $\|x_S - x_0\|$ close to 50% of the time for block-row TQRP solutions. For the problems with a small or no gap and $p = 1$, the percentage of the cases where the block-row TQRP solution is closer to x_0 than is the truncated SVD solution varied between 45% for the cluster example to 32% for the runs with a mean gap of 1.2.

The examples so far have been artificial. To test examples from practice or used elsewhere in the literature, we looked at problems from Hansen's regularization tools [16]. Our sample consists of Hansen's baart, deriv2 (with 3 different solutions), foxgood, heat (with 3 parameter values), ilaplace (with 4 different solutions), phillips, shaw, spikes, and wing for a total of 16 different examples. These are all the ill-

TABLE 1

Mean value of $\frac{\|x_T - x_0\| - \|x_S - x_0\|}{\|x_S - x_0\|}$ *for block-row/column low-rank approximations and, in the last column, the percent of the runs where, for TQRP,* x_T *is closer to* x_0 *than is* x_S*. In the table g stands for gap, s for spread, and m.g. for mean gap. These terms and the term cluster are defined in the text. p is the decay rate in the Picard coefficients. Each entry summarizes* 70,000 *samples.*

Problem properties			Method					%
g	s	p	TQLP	TQLRP	TQRP	TQRLP	TQRLRP	
100	100	.5	.0015	2.1×10^{-7}	$2.8\text{x}10^{-5}$	$7.7\text{x}10^{-8}$	$-7.8\text{x}10^{-10}$	50
100	100	1	.23	$1.2\text{x}10^{-7}$	$6.8\text{x}10^{-5}$	$1.0\text{x}10^{-5}$	$8.4\text{x}10^{-10}$	50
100	100	1.5	5.3	$1.8\text{x}10^{-6}$	$4.1\text{x}10^{-4}$	$3.2\text{x}10^{-4}$	$2.2\text{x}10^{-10}$	49
100	100	2	817	$1.3\text{x}10^{-5}$	.14	.14	$-2.3\text{x}10^{-8}$	44
100	100	3	6882	$3.3\text{x}10^{-6}$	.65	.65	$-2.0\text{x}10^{-8}$	43
100	10^4	1	.14	$3.5\text{x}10^{-7}$	$1.7\text{x}10^{-4}$	$6.5\text{x}10^{-6}$	$2.1\text{x}10^{-9}$	50
100	1	1	.29	$-3.8\text{x}10^{-8}$	$7.3\text{x}10^{-5}$	$3.5\text{x}10^{-5}$	$-3.4\text{x}10^{-10}$	49
10	100	1	.32	$3.1\text{x}10^{-5}$	$5.9\text{x}10^{-3}$	$9.1\text{x}10^{-4}$	$9.9\text{x}10^{-7}$	47
4	100	1	.40	.0016	.040	.0091	$6.5\text{x}10^{-5}$	42
1	100	1	.36	.082	.13	.099	.063	39
cluster		1	.11	.044	.067	.046	.037	45
m.g. 10		1	.28	.0081	.024	.015	.0044	49
m.g. 10		2	81	.019	.18	.16	.0076	44
m.g. 10		3	1163	.11	11	11	.016	23
m.g. 4		1	.29	.015	.038	.024	.0094	48
m.g. 1.2		1	.34	.041	.091	.065	.024	32

conditioned examples in regularization tools, except for parallax and ursell (for which x_0 is not supplied) and blur, which is parameterized differently from the other examples. Most of these examples do not have a clear gap in the singular value spectrum, and so we need a technique to choose the numerical rank k. For simplicity and to focus on the approximation scheme, not the technique for selecting k, for each regularization algorithm we selected the effective rank by calculating a regularized solution, x, for each numerical rank $k < n$ and then among all lower rank approximations, selecting the approximation that minimizes $\|x - x_0\|$.

For each of the 16 examples, we looked at the seven noise levels used in Table 1 for a total of 112 cases. For each case we chose 100 random noise vectors, applied a variety of regularization methods, and calculated the mean values of $(\|x_T - x_0\| - \|x_S - x_0\|)$ and of $\|x_S - x_0\|$. In each of the 112 cases we used the x_0 supplied by regularization tools, not a randomly chosen x_0. Each mean value is a mean over 100 different random noise vectors δb for a fixed x_0. In Table 2 we summarize the results for the block-row/column low-rank solutions produced by TQLP, TQLRP, TQRP, and TQRLRP factorizations. Each entry counts the number of the 112 cases where $\frac{\text{mean}(\|x_T - x_0\| - \|x_S - x_0\|)}{\text{mean}(\|x_S - x_0\|)}$ is in the indicated range.

In this table, the truncated SVD solutions are closer to x_0 in some cases, and in others the truncated UTV solutions are closer. However, overall for this set of problems the truncated UTV algorithm, even when stopped at the first step, appears to work as well as the truncated SVD. The table also indicates that additional steps of

TABLE 2

Counts for examples with characteristic features of ill-posed problems from [16] *of the number or cases, out of* 112 *cases, where* $\frac{\text{mean}(\|x_T-x_0\|-\|x_S-x_0\|)}{\text{mean}(\|x_S-x_0\|)}$ *is in the range indicated in the first row of the table. Rows two through six correspond to block-row/column solutions for different truncated factorizations.*

Range / Method	Less than −50%	−50% to −10%	−10% to −5%	−5% to −1%	−1% to 1%	1% to 5%	5% to 10%	10% to 50%	50% or more
TQRP	0	13	8	17	50	11	8	5	0
TQRLP	0	11	12	16	53	7	8	5	0
TQRLRP	0	7	7	7	77	12	0	2	0
TQLP	9	16	4	10	37	8	5	17	6
TQLRP	0	5	8	8	65	15	5	6	0

the algorithm bring values of mean($\|x_T-x_0\|$) closer to values of mean($\|x_S-x_0\|$). We also kept track of the percent of the time that $\|x_T-x_0\|$ was less that $\|x_S-x_0\|$ over the 11,200 (= 112 × 100) samples. These percents were 51%, 54%, 51%, 57%, and 57%, respectively, for the block-row/column solutions corresponding to the TQRP, TQRLP, TQRLRP, TQLP, and TQLRP factorizations. It is interesting to note that the results for the these test problems, where A and x_0 are not random, seem to favor the truncated UTV solutions more than do the results for test problems involving randomly generated examples. The reason for this merits further investigation.

In order to understand Table 2 better, it is useful to look at a specific case, for example, the Phillips example of [16] when the noise to signal ratio, $\|\delta b\|/\|b\|$, equals 0.1. We can illustrate the results in the table by looking at $\|x_T-x_0\|$ for TQRP and $\|x_S-x_0\|$ for a few typical values of δb. For the Phillips example the underlying true solution x_0 provided by [16] has $\|x_0\| = 2.99$. Six typical values of $\|x_T-x_0\|$ are .25, .29, .17, .30, .25, and .29 and the corresponding values of $\|x_S-x_0\|$ are .21, .34, .14, .34, .23, and .33. Overall, the magnitude of these values is quite similar and the two methods have approximately the same accuracy. The differences between these values are, respectively, .04, −.05, .03, −.04, .02, and −.04 and the corresponding values of $\|x_T-x_S\|$ are .09, .09, .07, .09, .04, and .41, respectively. In the table the sample size was 100, not 6. For these 100 values, $\frac{\text{mean}(\|x_T-x_0\|-\|x_S-x_0\|)}{\text{mean}(\|x_S-x_0\|)}$ was −.042, and this example is one of the 17 entries in the table for the TQRP method with $\frac{\text{mean}(\|x_T-x_0\|-\|x_S-x_0\|)}{\text{mean}(\|x_S-x_0\|)}$ between −5% and −1%.

7. Conclusions. We have discussed the application of the algorithm of [21] to solving ill-posed and rank-deficient problems. The algorithm constructs a UTV factorization of A by using one or more QR factorizations. The following are some of our results.

- The block-row solution produced by a rank-revealing QR factorization is identical to the corner solution produced by a related rank-revealing UTV factorization. (See Theorem 2 and the comments following the theorem.)
- If there is a modest gap in the singular values so that $\rho_1 < 1$, pivoting is not needed after the first step in the algorithm of section 3. (See Theorem 2.)
- We have presented an implementation of LAPACK's xGELSY that, in the low-rank case, is substantially faster than the implementation of xGELSY currently in LAPACK. (See Figure 1 and the discussion prior to Figure 1.)

Some of our results concern the accuracy, relative to truncated SVD solutions, of the solutions to (1) produced by truncated UTV factorizations. The results suggest the following recommendations about the appropriate choice of a method to use to construct regularized solutions to the system (1).

- If one can identify and evaluate the accuracy of typical examples, then we recommend that a variety of methods of regularization be compared for these examples. Our results indicate that although in some examples a relatively expensive method such as the truncated SVD will produce the best solution, in other examples cheaper methods will calculate solutions as close as or closer to the underlying desired solution. (See Theorems 3 and 5 and Tables 1 and 2.)
- If the initial QR factorization is rank-revealing, if the desired regularized solution corresponds to a sufficiently large gap in the singular values, and if p, the decay rate in the Picard coefficients, is not too large, as is often true in practice, then we recommend using the block-row truncated QRP solution. On average this truncated QRP solution will be very close to the accuracy of the truncated SVD solution and it can be calculated more quickly, dramatically so for low-rank problems. (See Theorems 4 and 6, Figure 1, and Tables 1 and 2.)
- If the desired solution does not correspond to a gap in the singular values, our experimental results with random examples suggest that truncated SVD solutions are, on average, somewhat better than truncated UTV solutions, but, for $p < 2$, the difference may be modest (see Table 1). The case where there is not a gap in the singular values merits further investigation. For this case Stewart [27] conjectures for the QRLP decomposition that "the analysis of this decomposition will not be simple."

We also did test runs for the set of problems of [16], which have characteristic features of ill-posed problems. In some cases the truncated SVD solutions were closer to the desired solution and in others the truncated UTV solutions were closer. Overall, for this set of problems the truncated UTV algorithm, even when stopped at the first step, appeared to work as well as the TSVD algorithm (see Table 2).

Acknowledgment. The helpful comments and suggestions from the anonymous referees are gratefully acknowledged.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] M. Bertero, C. De Mol, and G. A. Viano, *The stability of inverse problems*, in Scattering in Optics, H. P. Baltes, ed., Springer-Verlag, New York, 1980, pp. 161–214.

[3] Å. Björck, *Numerical Methods for Least Squared Problems*, SIAM, Philadelphia, 1996.

[4] P. Businger and G. H. Golub, *Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

[5] T. F. Chan and P. C. Hansen, *Low-rank revealing QR factorizations*, Numer. Linear Algebra Appl., 1 (1991), pp. 33–44.

[6] T. F. Chan and P. C. Hansen, *Some applications of the rank revealing QR factorization*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 727–741.

[7] S. Chandrasekaran and I. C. F. Ipsen, *Analysis of a QR algorithm for computing singular values*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 520–535.

[8] R. D. Fierro, *Perturbation analysis for two-sided (or complete) orthogonal decompositions*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 383–400.

[9] R. D. Fierro and J. R. Bunch, *Bounding the subspaces from rank revealing two-sided orthogonal decompositions*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 743–759.
[10] R. D. Fierro and P. C. Hansen, *Accuracy of TSVD solutions computed from rank-revealing decompositions*, Numer. Math., 70 (1995), pp. 453–471.
[11] R. D. Fierro and P. C. Hansen, *Low-rank revealing UTV decompositions*, Numer. Algorithms, 15 (1997), pp. 37–55.
[12] G. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[13] P. C. Hansen, *The discrete Picard condition for discrete ill-posed problems*, BIT, 30 (1990), pp. 658–672.
[14] P. C. Hansen, *Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 503–518.
[15] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.
[16] P. C. Hansen, *Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[17] P. C. Hansen, *Test matrices for regularization methods*, SIAM J. Sci. Comput., 16 (1995), pp. 506–512.
[18] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems. Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.
[19] Y. Hosoda, *Truncated least-squares least-norm solutions by applying the QR decomposition twice*, Trans. Inform. Process. Soc. Japan, 40 (1999), pp. 1051–1055.
[20] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[21] R. Mathias and G. W. Stewart, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1993), pp. 91–100.
[22] A. Neumaier, *Solving ill-conditioned and singular linear systems: A tutorial on regularization*, SIAM Rev., 40 (1998), pp. 636–666.
[23] H. Ren, *On the Error Analysis and Implementation of Some Eigenvalue Decomposition and Singular Value Decomposition Algorithms*, UT-CS-96-336, LAPACK working note 115, 1996, http://www.netlib.org/lapack/lawns/.
[24] V. K. Rohatgi, *An Introduction to Probability Theory and Mathematical Statistics*, John Wiley, New York, 1976.
[25] G. W. Stewart, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Proc., 40 (1992), pp. 1535–1541.
[26] G. W. Stewart, *Matrix Algorithms Volume* 1: *Basic Decompositions*, SIAM, Philadelphia, 1998.
[27] G. W. Stewart, *The QLP approximation to the singular value decomposition*, SIAM J. Sci. Comput., 20 (1999), pp. 1336–1348.
[28] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 601–605

SPECTRAL CHARACTERIZATIONS FOR HERMITIAN CENTROSYMMETRIC K-MATRICES AND HERMITIAN SKEW-CENTROSYMMETRIC K-MATRICES*

MARK YASUDA†

Abstract. Let A and K be real symmetric matrices with $K^2 = I$. In the article "A spectral characterization of generalized real symmetric centrosymmetric and generalized real symmetric skew-centrosymmetric matrices" [D. Tao and M. Yasuda, *SIAM J. Matrix Anal. Appl.*, 23 (2002), pp. 885–895], it was shown that (1) $AK = KA$ if and only if the spectrum of A equals the spectrum of KA up to sign and (2) $AK = -KA$ if and only if the spectrum of A equals the spectrum of KA multiplied by i. This paper extends these spectral characterizations from the case of real symmetric matrices to that of self-adjoint compact linear operators in a complex Hilbert space. Some consequences of these results are mentioned, including an application that describes the correspondence between the spectrum of a real symmetric Toeplitz matrix T and its associated Hankel matrix JT, where J is the so-called exchange matrix.

Key words. centrosymmetric matrix, skew-centrosymmetric matrix, Toeplitz matrix, Hankel matrix, eigenvalues

AMS subject classifications. 15A18, 15A57

DOI. 10.1137/S0895479802418835

1. Introduction.

Let K denote an involutory ($K^2 = I$) matrix. The class of *centrosymmetric K-matrices* consists of those matrices that commute with K. The class of *skew-centrosymmetric K-matrices* consists of those matrices that anticommute with K. These matrices are natural extensions to the classes of centrosymmetric and skew-centrosymmetric matrices that have been studied for many years (see [2] and [3] for a good set of references). Centrosymmetric K-matrices and skew-centrosymmetric K-matrices, however, have not received nearly as much attention. This is somewhat surprising since not much is lost structurally in moving to the more general setting. Some papers that have dealt with centrosymmetric K-matrices, skew-centrosymmetric K-matrices, and other related structures include [1], [5], [7], [11], and [12]. This current paper extends the spectral characterizations of [11] from the real symmetric setting to that of self-adjoint compact linear operators in a complex Hilbert space.

2. Terminology and notation.

Let J represent the *exchange matrix* of order n defined by $J_{i,j} = \delta_{i,n-j+1}$ for $1 \le i, j \le n$, where $\delta_{i,j}$ is the Kronecker delta. *Centrosymmetric matrices* are those matrices which commute with J. *Skew-centrosymmetric matrices* are those matrices which anticommute with J. Papers dealing with centrosymmetric and skew-centrosymmetric matrices often refer to vectors x satisfying $x = Jx$ as *symmetric* and vectors satisfying $x = -Jx$ as *skew-symmetric*. It is well known that real symmetric centrosymmetric matrices have an eigenbasis consisting of $\lceil \frac{n}{2} \rceil$ symmetric eigenvectors and $\lfloor \frac{n}{2} \rfloor$ skew-symmetric eigenvectors [4].

Since the exchange matrix J is involutory, it is natural to consider replacing J with a general involutory matrix K in the above definitions. As stated in the introduction, we refer to matrices that commute with K as *centrosymmetric K-matrices*

*Received by the editors November 29, 2002; accepted for publication (in revised form) by L. Reichel April 15, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/simax/25-3/41883.html
†Raytheon Company, 8680 Balboa Avenue, San Diego, CA 92123 (myasuda@eskimo.com).

and matrices that anticommute with K as *skew-centrosymmetric K-matrices*. When $x = Kx$, we say that the vector x is *K-symmetric*, and, when $x = -Kx$, we say that the vector x is *K-skew-symmetric*.

Let A be a linear operator. We denote the adjoint of A by A^*, and we denote the spectrum of A by $\Lambda(A) = \{\lambda_j(A)\}_{j=1}^{r(A)}$, where $r(A)$ is the spectrum's cardinality. We write $S = \pm\Lambda(A)$ if the elements of the multiset S are the same as those of $\Lambda(A)$ up to sign, and we write $S = z\Lambda(A)$ if $S = \{z\lambda_j(A)\}_{j=1}^{r(A)}$ for a fixed complex number z.

3. The spectral characterizations. An investigation made by McClanahan [10] revealed the strong connection between the spectral characterization theorems of [11] and a result of Gohberg and Krein [6]. By exploiting this connection, one can extend the spectral characterizations to the case of self-adjoint compact linear operators.

LEMMA 3.1 (Gohberg–Krein). *Let A be a compact linear operator in a complex Hilbert space, and let $\lambda_j(A)$ and $\sigma_j(A)$ denote the eigenvalues and singular values of A, arranged in decreasing order of magnitude. Then $|\lambda_j(A)| = \sigma_j(A)$ for $1 \le j \le r(A)$ if and only if A is normal.*

Note. In [6], Gohberg and Krein prove the harder "⇒" direction of the proof. The converse is a straightforward consequence of the diagonalization theorem for compact normal operators.

LEMMA 3.2. *Let A and KA be compact linear operators in a complex Hilbert space, where A and K are self-adjoint and $K^2 = I$. Let $\lambda_j(KA)$ and $\sigma_j(KA)$, respectively, denote the eigenvalues and singular values of KA arranged in decreasing order of magnitude. Then $\Lambda(KA) = \pm z\Lambda(A)$ if and only if $|\lambda_j(KA)| = \sigma_j(KA)$ for $1 \le j \le r(A)$ and $\Lambda(KA) \in z\mathbb{R}$, where z is a complex number of modulus 1.*

Proof. Since $(KA)^*(KA) = A^*A$, we have that $\sigma_j(KA) = \sigma_j(A)$. Since A is normal, Lemma 3.1 implies that $|\lambda_j(A)| = \sigma_j(A)$ for $1 \le j \le r(A)$. Therefore, $|\lambda_j(KA)| = \sigma_j(KA)$ holds if and only if $|\lambda_j(KA)| = |\lambda_j(A)|$. Since the eigenvalues of A are real, $\Lambda(KA) = \pm z\Lambda(A)$ if and only if $|\lambda_j(KA)| = |\lambda_j(A)|$ for $1 \le j \le r(A)$ and $\Lambda(KA) \in z\mathbb{R}$, where z is a complex number of modulus 1. The statement of the lemma follows immediately. ∎

THEOREM 3.3 (McClanahan). *Let A and KA be compact linear operators in a complex Hilbert space, where A and K are self-adjoint and $K^2 = I$. Then $KA = z^2AK$ if and only if $\Lambda(KA) = \pm z\Lambda(A)$, where z is a fixed complex number of modulus 1.*

Proof. Let $\bar{z}$ denote the complex conjugate of z. $KA = z^2AK$ if and only if $\bar{z}KA$ is self-adjoint. Recalling that a compact normal operator is self-adjoint if and only if its spectrum is real (for example, see [9]), we have that $\Lambda(\bar{z}KA) \in \mathbb{R}$. Since $\bar{z}KA$ is normal if and only if KA is normal, Lemma 3.1 implies that $KA = z^2AK$ if and only if $|\lambda_j(KA)| = \sigma_j(KA)$ for $1 \le j \le r(A)$. Application of Lemma 3.2 completes the proof. ∎

Setting z equal to ± 1 and $\pm i$ in the statement of Theorem 3.3 gives the following spectral characterizations for Hermitian centrosymmetric K-matrices and skew-centrosymmetric K-matrices.

COROLLARY 3.4. *Suppose $A \in \mathbb{C}^{n\times n}$ and $K \in \mathbb{C}^{n\times n}$ are Hermitian and $K^2 = I$. Then*

1. *$AK = KA$ if and only if $\Lambda(A) = \pm\Lambda(KA)$,*
2. *$AK = -KA$ if and only if $\Lambda(A) = \pm i\Lambda(KA)$.*

Note 1. The ⇒ directions of the corollary can be proved under weaker conditions (see [11]).

Note 2. It is not hard to show that, under the corollary's hypotheses, substitution of values of z other than ± 1 and $\pm i$ into the statement of Theorem 3.3 forces A to be the zero matrix.

We end this section by describing the eigenbasis of a Hermitian centrosymmetric K-matrix.

PROPOSITION 3.5. *Suppose $A \in \mathbb{C}^{n\times n}$ and $K \in \mathbb{C}^{n\times n}$ are Hermitian and $K^2 = I$. If $AK = KA$, then A has an eigenbasis consisting solely of K-symmetric and K-skew-symmetric eigenvectors.*

Proof. Since A and K commute and are both normal, they can be simultaneously diagonalized by a single unitary matrix (for example, see [8]). Since K's eigenvalues are elements of the set $\{-1, 1\}$, it follows that every eigenvector of K (and therefore A) satisfies $x = Kx$ or $x = -Kx$. □

4. Some consequences. The results in this section can be obtained using Corollary 3.4 and Proposition 3.5. Since the proofs follow along the same lines as those for the corresponding real symmetric results in [11], they are omitted.

PROPOSITION 4.1. *Let $K \in \mathbb{C}^{n\times n}$ be a Hermitian involutory matrix and let $A \in \mathbb{C}^{n\times n}$ be a Hermitian centrosymmetric K-matrix. Assume that K's eigenvalue 1 has multiplicity n_1 and that K's eigenvalue -1 has multiplicity n_2, where $n_1+n_2 = n$.*

If V is a basis for the eigenspace of A consisting entirely of K-symmetric and K-skew-symmetric eigenvectors, then V must contain precisely n_1 K-symmetric eigenvectors and n_2 K-skew-symmetric eigenvectors.

Remark. Proposition 4.1 generalizes the corresponding result in [4] for real symmetric centrosymmetric matrices.

PROPOSITION 4.2. *Let $K \in \mathbb{C}^{n\times n}$ be a Hermitian involutory matrix and let $A \in \mathbb{C}^{n\times n}$ be a Hermitian centrosymmetric K-matrix. Assume that K's eigenvalue -1 has multiplicity n_2. If we let $d(X, Y)$ equal the number of eigenvalues of X which differ from those of Y, then $d(A, KA) \le n_2$.*

If we further stipulate that $|\lambda_i| = |\lambda_j|$ implies $\lambda_i = \lambda_j$ for any $\{\lambda_i, \lambda_j\} \in \Lambda(A)$, then we also have the lower bound $\max\{n_2 - m, 0\} \le d(A, KA)$, where m is the multiplicity of A's zero eigenvalue.

PROPOSITION 4.3. *Suppose $A \in \mathbb{C}^{n\times n}$ and $K \in \mathbb{C}^{n\times n}$ are Hermitian, with $K^2 = I$. If $|\lambda_i| = |\lambda_j|$ implies $\lambda_i = \lambda_j$ for any $\{\lambda_i, \lambda_j\} \in \Lambda(A)$, then $\Lambda(A) = \Lambda(KA)$ if and only if $A = KA$.*

PROPOSITION 4.4. *Suppose $A \in \mathbb{C}^{n\times n}$ and $K \in \mathbb{C}^{n\times n}$ are Hermitian, with $K^2 = I$. If $|\lambda_i| = |\lambda_j|$ implies $\lambda_i = \lambda_j$ for any $\{\lambda_i, \lambda_j\} \in \Lambda(A)$, then $\Lambda(A) = \Lambda(-KA)$ if and only if $A = -KA$.*

5. An application to Toeplitz and Hankel matrices. A matrix $A = \{a_{i,j}\}$ for $1 \le i, j \le n$ is said to be *Toeplitz* if the relationship $a_{r,s} = a_{r+1,s+1}$ holds for all of A's elements. It is said to be *Hankel* if the relationship $a_{r,s} = a_{r-1,s+1}$ holds for all of A's elements. Real symmetric Toeplitz matrices are a frequently studied subclass of the symmetric centrosymmetric matrices.

Clearly, if T is a symmetric Toeplitz matrix, then JT is a centrosymmetric Hankel matrix. Results in the previous sections can be used to show that the spectral decomposition of centrosymmetric Hankel matrices corresponds in a direct way to that of symmetric Toeplitz matrices and vice versa.

Example. The eigenvalues of the real symmetric $n \times n$ tridiagonal Toeplitz matrix

$$\begin{pmatrix} a & b & 0 & \cdots & 0 \\ b & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b \\ 0 & \cdots & 0 & b & a \end{pmatrix}$$

are known to be $\lambda_j = a + 2b\cos\left(\frac{\pi j}{n+1}\right)$ for $1 \leq j \leq n$ with corresponding eigenvectors

$$v_j = \left(\sin\left(\frac{\pi j}{n+1}\right), \sin\left(\frac{2\pi j}{n+1}\right), \ldots, \sin\left(\frac{n\pi j}{n+1}\right)\right)^T.$$

Using Corollary 3.4 and Proposition 3.5, we can immediately write down the spectrum of the corresponding centrosymmetric Hankel matrix

$$\begin{pmatrix} 0 & \cdots & 0 & b & a \\ \vdots & \iddots & \iddots & \iddots & b \\ 0 & \iddots & \iddots & \iddots & 0 \\ b & \iddots & \iddots & \iddots & \vdots \\ a & b & 0 & \cdots & 0 \end{pmatrix}$$

as

$$\lambda_j = a + 2b\cos\left(\frac{\pi j}{n+1}\right) \quad \text{for } \ j \ \text{ odd},$$

$$\lambda_j = -a - 2b\cos\left(\frac{\pi j}{n+1}\right) \quad \text{for } \ j \ \text{ even}.$$

Of course, one needs only to examine the first and last nonzero components of the eigenvectors of a symmetric Toeplitz matrix to determine whether they are symmetric or skew-symmetric and hence which of the corresponding Hankel matrix eigenvalues μ_j has the opposite sign from λ_j.

Acknowledgment. The author would like to thank Kevin McClanahan for providing valuable insights that helped lead to this research.

REFERENCES

[1] A. Andrew, *Eigenvectors of certain matrices*, Linear Algebra Appl., 7 (1973), pp. 151–162.
[2] A. Andrew, *Further comments on "On the eigenvectors of symmetric Toeplitz matrices,"* IEEE Trans. Acoust. Speech Signal Process., 33 (1985), p. 1013.
[3] A. Andrew, *Centrosymmetric matrices*, SIAM Rev., 40 (1998), pp. 697–698.
[4] A. Cantoni and P. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.
[5] H.-C. Chen, *Generalized reflexive matrices: Special properties and applications*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 140–153.
[6] I. Gohberg and M. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.

[7] R. Hill and S. Waters, *On κ-real and κ-Hermitian matrices*, Linear Algebra Appl., 169 (1992), pp. 17–29.
[8] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[9] R. Kadison and J. Ringrose, *Fundamentals of the Theory of Operator Algebras*, AMS, Providence, RI, 1997.
[10] K. McClanahan, *personal correspondence*, 2002.
[11] D. Tao and M. Yasuda, *A spectral characterization of generalized real symmetric centrosymmetric and generalized real symmetric skew-centrosymmetric matrices*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 885–895.
[12] J. Weaver, *Real eigenvalues of nonnegative matrices which commute with a symmetric matrix involution*, Linear Algebra Appl., 100 (1988), pp. 243–253.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 606–626

PERTURBATION THEORY FOR ANALYTIC MATRIX FUNCTIONS: THE SEMISIMPLE CASE*

P. LANCASTER†, A. S. MARKUS‡, AND F. ZHOU†

Abstract. The eigenvalue problem for non-self-adjoint, analytic matrix functions of two variables, $L(\lambda, \alpha)$, is examined with emphasis on the case when, at a fixed α_0, $L(\lambda, \alpha_0)$ has a multiple, semisimple eigenvalue λ_0. New sufficient conditions for analytic dependence of eigenvalue functions, $\lambda(\alpha)$, on α in a neighborhood of α_0 are obtained. An algorithm for generating Taylor coefficients of perturbed eigenvalues and eigenvectors is studied and the existence of positive radii of convergence is established. Connections with known results on self-adjoint problems are made.

Key words. perturbation theory, analytic matrix functions, semisimple eigenvalues, non-self-adjoint functions

AMS subject classifications. 15A18, 47A55, 47A56.

DOI. 10.1137/S0895479803423792

1. Introduction.

Let $L(\lambda, \alpha)$ be a matrix function with values in the $p \times p$ complex matrices. This function is assumed to be analytic in complex variables λ and α on a neighborhood of (λ_0, α_0). It is emphasized that, although our setting is in a finite dimensional space $\mathcal{H}$, there are immediate applications to discrete eigenvalues of more general operator valued functions $L(\lambda, \alpha)$ defined on spaces of infinite dimension (see, for example, Chapter 10 of [B]).

Historically, important contributions have been made to the classical eigenvalue problem $L(\lambda, \alpha) = \lambda I - A(\alpha)$ with $A(\alpha)$ self-adjoint (for real α) by Rellich, Kato, and several others (see, for example, the books of Rellich [R], Kato [K], and Baumgärtel [B]). Special attention has also been paid to the case of quadratic dependence on the eigenvalue parameter (see [B], [GLR2], [LNV], for example):

$$L(\lambda, \alpha) = \lambda^2 I + \lambda B(\alpha) + C(\alpha),$$

but in this work it is found to be convenient (and even helpful) to extend previous analyses to analytic dependence on both λ and α. There are also important results peculiar to *self-adjoint* analytic matrix functions in [GLR1], but the focus here is on *non-self-adjoint* problems.

In this context it is assumed that there is nondegeneracy in the sense that $\det L(\lambda, \alpha_0)$ is not identically equal to zero, and there is said to be an eigenvalue λ_0 of $L(\lambda, \alpha_0)$ if $\det L(\lambda_0, \alpha_0) = 0$. Without loss of generality, it is assumed throughout that $\alpha_0 = 0$.

Suppose that λ_0 is an eigenvalue of $L(\lambda, 0)$ of finite multiplicity. Let this eigenvalue have partial multiplicities $m_1, m_2, \ldots, m_g$. (These partial multiplicities can be determined from a local Smith canonical form for $L(\lambda, 0)$ valid in a neighborhood of λ_0, as in [BGR], for example.)

*Received by the editors February 27, 2003; accepted for publication (in revised form) by M. L. Overton June 4, 2003; published electronically December 17, 2003. This research was supported in part by a Discovery Grant of the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/simax/25-3/42379.html

†Department of Mathematics and Statistics, University of Calgary, Calgary AB T2N 1N4, Canada (lancaste@ucalgary.ca, feiz@math.ucalgary.ca)

‡Department of Mathematics, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (markus@cs.bgu.ac.il).

It will be convenient to introduce the following notation:

$$L_0 = L(\lambda_0, 0), \quad L_{10} = \frac{\partial L}{\partial \lambda}(\lambda_0, 0), \quad L_{01} = \frac{\partial L}{\partial \alpha}(\lambda_0, 0).$$

An eigenvalue λ_0 is said to be *semisimple* when $m_1 = \cdots = m_g = 1$. Geometrically, this is equivalent to the condition that, for every eigenvector x associated with a semisimple eigenvalue λ_0, the singular equation

$$L_0 y = -L_{10} x$$

has no solution y.

The main interest of this paper is the formulation of conditions guaranteeing the existence of analytic eigenvalue functions and corresponding eigenvector functions. The following lemma shows that the second property follows from the first. This and more general results are known (see Theorem 18.2.1 of [GLR3], for example), but a simple proof is provided for the reader's convenience.

LEMMA 1. *If $L(\lambda, \alpha)$ is an analytic matrix function in a neighborhood of $(\lambda_0, 0)$ and has an eigenvalue function $\lambda(\alpha)$ which is analytic at $\alpha = 0$, then there is an eigenvector function $x(\alpha)$ which is analytic at $\alpha = 0$.*

Proof. Suppose that $L(\lambda, \alpha)$ has an eigenvalue $\lambda(\alpha) = \sum_{j=0}^{\infty} a_j \alpha^j$ in a neighborhood of $\alpha = 0$ and that $L(\lambda(\alpha), \alpha)$ has rank r in a deleted neighborhood $\mathcal{N}$ of $\alpha = 0$. Without loss of generality, assume that the minor

$$L \begin{bmatrix} 1 & 2 & \dots & r \\ 1 & 2 & \dots & r \end{bmatrix} \neq 0 \quad \text{in } \mathcal{N}.$$

Consider the minor of order $r+1$:

$$L \begin{bmatrix} 1 & 2 & \dots & r+1 \\ 1 & 2 & \dots & r+1 \end{bmatrix} = \sum_{j=1}^{r+1} l_{r+1,j}(\alpha) x_j(\alpha) = 0,$$

where $x_1(\alpha), \dots, x_{r+1}(\alpha)$ are cofactors of the $r+1$ row in the determinant on the left. Observe also that

$$x_{r+1} = L \begin{bmatrix} 1 & 2 & \dots & r \\ 1 & 2 & \dots & r \end{bmatrix} \neq 0,$$

and complete the construction of a nonzero vector $x(\alpha)$ by setting $x_{r+2} = \cdots = x_n = 0$.

Then, for $r = 1, 2, \dots, n$, consider the jth term of the vector $L(\lambda, \alpha)x(\alpha)$. It has the form

$$\begin{aligned} \sum_{k=1}^{n} l_{jk}(\alpha) x_k(\alpha) &= \sum_{k=1}^{r+1} l_{jk}(\alpha) x_k(\alpha) \\ &= L \begin{bmatrix} 1 & 2 & \dots & r & j \\ 1 & 2 & \dots & r & r+1 \end{bmatrix}. \end{aligned}$$

However, the last expression is zero because, for $j \leq r$, two rows of the minor agree and, for $j \geq r+1$, we have a minor of order $r+1 > r$. Thus there is a vector function $x(\alpha)$, nonvanishing in $\mathcal{N}$ and analytic in $\mathcal{N} \cup \{0\}$, such that $L(\lambda(\alpha), \alpha)x(\alpha) \equiv 0$. If this vector has a zero of order k at $\alpha = 0$, then the vector function $x(\alpha)\alpha^{-k}$ has all the required properties. ∎

Note that the Weierstrass preparation theorem (Theorem 3.10 of [M], for example) underlies much of our analysis and, locally, allows us to treat $\det L(\lambda(\alpha), \alpha) = 0$ as an algebraic equation.

2. Preliminaries. For the reader's convenience some known results are presented here concerning perturbations of eigenvalues and eigenvectors.

Let λ_0 be an eigenvalue of the unperturbed matrix function $L(\lambda, 0)$ with partial multiplicities $m_1, m_2, \dots, m_g$. Then there are numbers $\varepsilon > 0$ and $\delta > 0$ such that, for $|\alpha| < \varepsilon$, the spectrum of $L(\lambda, \alpha)$ in $|\lambda - \lambda_0| < \delta$ consists of $m = \sum_1^g m_j$ eigenvalues $\lambda_j(\alpha)$, which can be represented by branches of several Puiseux series

$$\mu_\nu(\alpha) = \lambda_0 + \sum_{k=1}^{\infty} c_{\nu k} \alpha^{k/q_\nu}, \qquad \nu = 1, 2, \dots, r \tag{2.1}$$

(and the q_ν are positive integers). In general, the only connection between the numbers $\{m_j\}_1^g$ and $\{q_\nu\}_1^r$ is the equality $\sum_{j=1}^g m_j = \sum_{\nu=1}^r q_\nu$. The function $\mu_\nu(\alpha)$ is an algebraic q_ν-valued function on a cut neighborhood of $\alpha = 0$ (say $|\alpha| < \varepsilon$, $-\pi < \arg\alpha \le \pi$). It determines q_ν values of $\lambda_j(\alpha)$ which correspond to q_ν values of $\alpha^{1/q}$ in this neighborhood, namely

$$|\alpha|^{1/q} \exp(i(\arg\alpha + 2\pi j)/q), \qquad j = 0, 1, \dots, q-1.$$

Also, in expansion (2.1), $\alpha^{k/q} = (\alpha^{1/q})^k$.

As defined by Langer, Najman, and Veselić (see [LNV]), a general eigenvalue λ_0 with partial multiplicities $m_1, m_2, \dots, m_g$ has the *regular splitting property* if, for each m_i, there emerge from λ_0 (in the complex plane) m_i eigenvalues $\lambda_{ij}(\alpha)$ with Puiseux expansions for which

$$\lambda_{ij}(\alpha) = \lambda_0 + \lambda'_{ij} \alpha^{1/m_i} + o(|\alpha|^{1/m_i}) \tag{2.2}$$

holds as $\alpha \to 0$, $i = 1, 2, \dots, g$, $j = 1, 2, \dots, m_i$, and $\lambda'_{ij} \neq 0$ whenever $m_i > 1$. If, in addition, $\lambda'_{ij} \neq 0$ for all i, j in (2.2), then there is a *complete regular splitting* at λ_0. We say that λ_0 has the complete regular splitting (CRS) property.

As we are particularly interested in the case in which λ_0 is a semisimple eigenvalue, note that regular splitting corresponds to the existence of asymptotic relations

$$\lambda_j(\alpha) = \lambda_0 + \lambda'_j \alpha + o(|\alpha|) \tag{2.3}$$

as $\alpha \to 0$ for $j = 1, 2, \dots, g$, and CRS means that $\lambda'_j \neq 0$ for each j. Returning to the Puiseux series (2.1), observe that, in this case, regular splitting means that for every $q_\nu > 1$, $c_{\nu k} = 0$ for $k < q_\nu$. CRS means that, in addition, $c_{\nu q_\nu} \neq 0$ for all ν.

A statement concerning Puiseux series for eigenvectors will also be useful.

LEMMA 2. *For every eigenvalue*

$$\lambda(\alpha) = \lambda_0 + \sum_{j=1}^{\infty} c_j \alpha^{j/q} \tag{2.4}$$

of $L(\lambda, \alpha)$ defined on a cut neighborhood of $\alpha = 0$, there exists an associated eigenvector on this same neighborhood of the form

$$x(\alpha) = \sum_{k=0}^{\infty} \xi_k \alpha^{k/q}, \qquad \xi_0 \neq 0. \tag{2.5}$$

In these two expansions, $\alpha^{1/q}$ denotes the same branch of the corresponding q-valued function and $\alpha^{k/q} = (\alpha^{1/q})^k$.

Proof. This follows immediately from Lemma 1 on replacing α by $\alpha^{1/q}$. ∎

3. A useful construction. Some useful notations are introduced in this section. They play an important part in some basic definitions of section 4. Let $\mathcal{E}$ and $\mathcal{F}$ be g dimensional subspaces of $\mathcal{H}$ and $M \in \mathcal{L}(\mathcal{H})$. Choose bases $\{e_k\}_1^g$ for $\mathcal{E}$ and $\{f_j\}_1^g$ for $\mathcal{F}$, and define a matrix

$$[M]_{\mathcal{E},\mathcal{F}} := [(Me_k, f_j)]_{j,k=1}^g. \tag{3.1}$$

Of course, this matrix depends on the choice of basis vectors, but this is not important in what follows and is suppressed. Another representation for this matrix is required.

Denote by $\hat{M}$ the linear operator mapping $\mathcal{E}$ into $\mathcal{F}$ defined as follows: If $y = [M]_{\mathcal{E},\mathcal{F}}\, x$ $(x, y \in \mathcal{C}^g)$, then

$$\hat{M} \sum_1^g x_j e_j = \sum_1^g y_j f_j. \tag{3.2}$$

There are two equivalent forms for this definition of $\hat{M}$:

(a) If S is the isomorphism of $\mathcal{E}$ onto $\mathbb{C}^g$ defined by

$$S\left(\sum_{j=1}^g x_j e_j\right) = [x_j]_1^g$$

and T is the isomorphism of $\mathbb{C}^g$ onto $\mathcal{F}$ defined by

$$T([y_j]_1^g) = \sum_{j=1}^g y_j f_j,$$

then $\hat{M} = T[M]_{\mathcal{E},\mathcal{F}} S$.

(b) If R is the linear transformation from $\mathcal{H}$ to $\mathcal{F}$ defined by $R = \sum_{j=1}^g (\,.\,, f_j) f_j$, then

$$\hat{M} = RM|\mathcal{E}.$$

(Note that if $\{f_j\}_1^g$ is an orthonormal basis, then R is the orthogonal projector onto $\mathcal{F}$.) To verify (b), observe that if $h = \sum_1^g x_k e_k \in \mathcal{E}$, then

$$RMh = R \sum_1^g x_k M e_k = \sum_{j=1}^g \left(\sum_{k=1}^g (Me_k, f_j) x_j \right) f_j,$$

which coincides with the definition (3.2).

4. Semisimple eigenvalues. A third characterization of the semisimple property will be helpful. Also, the semisimple property with respect to the second variable, α, can play a role. Thus, the eigenvalue λ_0 at $\alpha = 0$ is said to be *α-semisimple* if, for all nonzero $x \in \mathrm{Ker} L_0$, the singular equation $L_0 y = -L_{01} x$ has no solution y. Write $\mathcal{K} := \mathrm{Ker} L_0$ and $\mathcal{K}' := \mathrm{Ker}(L_0^*) = (\mathrm{Im} L_0)^\perp$. In what follows, the construction of section 3 is used with $\mathcal{E} = \mathcal{K}$ and $\mathcal{F} = \mathcal{K}'$.

LEMMA 3. (a) *An eigenvalue λ_0 at $\alpha = 0$ is semisimple if and only if $[L_{10}]_{\mathcal{K},\mathcal{K}'}$ is nonsingular.* (b) *An eigenvalue λ_0 at $\alpha = 0$ is α-semisimple if and only if $[L_{01}]_{\mathcal{K},\mathcal{K}'}$ is nonsingular.*

Proof. (a) Let $\xi_0 \in \mathcal{K}$, $\xi_0 \neq 0$, and consider the equation

$$L_0\xi_1 = -L_{10}\xi_0 \tag{4.1}$$

for ξ_1. There exists a solution if and only if $L_{10}\xi_0$ is orthogonal to $\mathcal{K}'$. But this is equivalent to $RL_{10}\xi_0 = 0$, i.e., $\hat{L}_{10}\xi_0 = 0$. Thus, there is no nonzero solution for (4.1) if and only if $\hat{L}_{10}$, and hence $[L_{10}]_{\mathcal{K},\mathcal{K}'}$ is nonsingular. The proof for (b) is similar. □

Remark. The validity of statement (a) is easily verified directly for the classical eigenvalue problem $L(\lambda, \alpha) = \lambda I - A$.

We now extend Lemma 3.10 of [HL] as follows.

LEMMA 4. *A semisimple eigenvalue has the regular splitting property and, if the eigenvalue is also α-semisimple, then it has the CRS property.*

Proof. Let λ_0 be semisimple with multiplicity g. Consider an eigenvalue function of the form (2.4) where $1 \leq q \leq g$. By Lemma 2 there is an associated eigenvector function of the form (2.5). Denote the first nonzero coefficient in (2.4) by $c_{k'}$, and assume that $k' < q$. Now compare coefficients of $\alpha^{k/q}$ for $k = 0$ and $k = k'$ in $L(\lambda(\alpha), \alpha)x(\alpha) \equiv 0$, i.e., in

$$\{L_0+(c_{k'}\alpha^{k'/q}+c_{k'+1}\alpha^{(k'+1)/q}+\cdots)L_{10}+\alpha L_{01}+\cdots\}\{\xi_0+\xi_1\alpha^{1/q}+\cdots+\xi_q\alpha+\cdots\} = 0.$$

It is found that $L_0\xi_j = 0$ and $c_{k'}L_{10}\xi_0 + L_0\xi_{k'} = 0$. It follows that ξ_0, $\xi_{k'}/c_{k'}$ form a Jordan chain for λ_0 and the assumption that λ_0 is semisimple is contradicted. Consequently, $k' \geq q$ and there must be a regular splitting.

Suppose now that the splitting is not complete. Then there is an integer $r > 0$ and an eigenvalue

$$\lambda(\alpha) - \lambda_0 = c_{q+r}\alpha^{(q+r)/q} + \cdots$$

with $c_{q+r} \neq 0$, and

$$\{L_0 + (c_{q+r}\alpha^{(q+r)/q} + \cdots)L_{10} + \alpha L_{01} + \cdots\}\{\xi_0 + \xi_1\alpha^{1/q} + \cdots\} = 0.$$

The coefficients of α^0 and α^1 yield

$$L_0\xi_0 = 0 \quad \text{and} \quad L_0\xi_q + L_{01}\xi_0 = 0.$$

This contradicts the definition of an α-semisimple eigenvalue and concludes the proof. □

Thus, for any eigenvalue function $\lambda(\alpha)$ (emanating from a semisimple λ_0),

$$\lambda(\alpha) = \lambda_0 + \lambda'\alpha + o(|\alpha|) \quad \text{as } \alpha \to 0,$$

where $\lambda' = c_q$ of (2.4), and $\lambda(\alpha)$ is said to be *real differentiable* at $\alpha = 0$. Then the equation $L(\lambda(\alpha), \alpha)x(\alpha) = 0$ implies

$$\{L_0 + \alpha(\lambda' L_{10} + L_{01}) + \cdots\}\{\xi_0 + \xi_1\alpha^{1/q} + \cdots\} = 0 \quad (1 \leq q \leq g),$$

whence

$$L_0\xi_q + (\lambda' L_{10} + L_{01})\xi_0 = 0.$$

Using a basis $\{e_1, e_2, \ldots, e_g\}$ for $\mathcal{K}$, write $E = [e_1 \cdots e_g]$ and $\xi_0 = E\phi$, $0 \neq \phi \in \mathbb{C}^g$. Now introduce a basis $\{f_1, \ldots, f_g\}$ for $\mathcal{K}'$ so that $L_0^* f_j = 0$ and

$$((\lambda' L_{10} + L_{01})E\phi, f_j) = 0, \qquad j = 1, 2, \ldots, g,$$

i.e.,

$$(\lambda'[L_{10}]_{\mathcal{K},\mathcal{K}'} + [L_{01}]_{\mathcal{K},\mathcal{K}'})\phi = 0, \quad \phi \neq 0,$$

and λ' is an eigenvalue of the pencil

$$\mathcal{P}(\mu) := \mu[L_{10}]_{\mathcal{K},\mathcal{K}'} + [L_{01}]_{\mathcal{K},\mathcal{K}'}. \tag{4.2}$$

Concerning the effect of the choice of bases for $\mathcal{K}$ and $\mathcal{K}'$ on the pencil $\mathcal{P}(\mu)$, it is easily seen that a second pair of bases generates a strictly equivalent pencil $\tilde{\mathcal{P}}(\mu)$, i.e., there are nonsingular U and V such that $\tilde{\mathcal{P}}(\mu) = U\mathcal{P}(\mu)V$, and all essential eigenvalue properties are invariant.

Now a converse statement is to be established: that each eigenvalue of $\mathcal{P}(\mu)$ is one of the coefficients λ_j' of (2.3).

LEMMA 5. *Let λ_0 be a semisimple eigenvalue of $L(\lambda, \alpha)$ at $\alpha = 0$ of multiplicity g. Then each eigenvalue of $\mathcal{P}(\mu)$ determines one of the coefficients $\lambda_1', \lambda_2', \ldots, \lambda_g'$ of the expansions* (2.3).

Proof. From Lemma 4, the relations (2.3) hold for g eigenvalue functions $\lambda_j(\alpha)$ of

$$L(\lambda, \alpha) = \sum_{i,j=0}^{\infty} (\lambda - \lambda_0)^i \alpha^j L_{ij},$$

and L_{00} is identified with L_0. Writing $L(\lambda, 0) = \sum_{i=0}^{\infty}(\lambda - \lambda_0)^i L_{i0}$, put

$$L(\lambda, \alpha) = L(\lambda, 0) + \alpha L_{01} + \alpha B(\lambda, \alpha),$$

where $B(\lambda, \alpha)$ is analytic near $(\lambda_0, 0)$ and $B(\lambda_0, 0) = 0$.

To find the possible values of λ_j', substitute (2.3) into the characteristic equation $\det L(\lambda, \alpha) = 0$, collect powers of α, and then equate the coefficient of the lowest power of α to zero. In the terminology associated with Newton's diagram, this equation in λ_j' is known as the "determining equation" (see [Bl] or [VT]), and all of its solutions determine coefficients λ_j' in (2.3) (section 15 of [Bl], for example). Although we are considering the lowest power of α, it is important to recognize that zero roots of the determining equation, if any, are admissible. Note also that the characteristic equation can be multiplied by an arbitrary analytic function of λ, provided it does not vanish at λ_0, without affecting the determining equation.

Since λ_0 is semisimple with multiplicity g, $L(\lambda, 0)$ has local Smith normal form

$$D(\lambda) = \begin{bmatrix} (\lambda - \lambda_0)I_g & 0 \\ 0 & I_{n-g} \end{bmatrix},$$

where I_k denotes the identity matrix of size k (see [BGR], for example). Thus, there are $n \times n$ matrix functions $E(\lambda)$, $F(\lambda)$ which are analytic and invertible near λ_0 such that

$$F(\lambda)L(\lambda, 0)E(\lambda) = D(\lambda). \tag{4.3}$$

Also, the first g columns of $E(\lambda_0)$ and $F(\lambda_0)^*$ form bases for $\mathcal{K}$ and $\mathcal{K}'$, respectively. It follows that the determining equation for $\det L(\lambda,\alpha)=0$ is the same as that for

$$\det\left[D(\lambda)+\alpha F(\lambda)L_{01}E(\lambda)+\alpha F(\lambda)B(\lambda,\alpha)E(\lambda)\right]=0. \tag{4.4}$$

Considering the coefficient of λ in $D(\lambda)$, it is seen that (with these bases) $[L_{10}]_{\mathcal{K},\mathcal{K}'}=I_g$. Now substitute $\lambda(\alpha)=\lambda_0+\mu\alpha+o(|\alpha|)$ so that the leading term of $D(\lambda(\alpha))$ has the form

$$\begin{bmatrix} \mu[L_{10}]_{\mathcal{K},\mathcal{K}'}\alpha & 0 \\ 0 & I_{n-g} \end{bmatrix}.$$

Also, the leading term of $\alpha F(\lambda)L_{01}E(\lambda)$ has the form $\alpha F(\lambda_0)L_{01}E(\lambda_0)$ and all other terms have a factor α^2. On examining the block structure of (4.4), it follows that the left-hand side of the determining equation is the coefficient of α^g and, furthermore, this coefficient is just

$$\det(\mu[L_{10}]_{\mathcal{K},\mathcal{K}'}+[L_{01}]_{\mathcal{K},\mathcal{K}'})=\det\mathcal{P}(\mu),$$

a polynomial of degree g, as required. □

The following theorem is now established and its first part generalizes a result of [LN1] and [LN2]—after reduction to the semisimple case. There, the unperturbed eigenvalue is also required to be α-semisimple (see Lemma 3) and, here, this assumption is not made.

THEOREM 6. *Let $L(\lambda,\alpha)$ be an analytic matrix function of λ and α with a semisimple eigenvalue λ_0 at $\alpha=0$ of multiplicity g. Then there are exactly g eigenvalues $\lambda_j(\alpha)$, $j=1,2,\dots,g$, of $L(\lambda,\alpha)$ for which $\lambda_j(\alpha)\to\lambda_0$ as $\alpha\to 0$. These eigenvalues have Puiseux expansions for which (2.3) holds and there is a one-to-one correspondence between the coefficients λ_j' and the eigenvalues of the pencil $\mathcal{P}(\mu)$.*

For every eigenvalue $\lambda_j(\alpha)$ of $L(\lambda,\alpha)$ there is a corresponding eigenvector $x_j(\alpha)$ which also has a Puiseux expansion about $\alpha=0$, and, if $x_j(0)=\sum_{k=1}^g\phi_{jk}e_k$, then the vector $\phi_j=[\phi_{jk}]_{k=1}^g\in\mathbb{C}^g$ is an eigenvector of the pencil $\mathcal{P}(\mu)$ corresponding to λ_j'.

Notice also that if, in addition, λ_0 is α-semisimple, then $\lambda_j'\neq 0$ for each j and the CRS property holds.

The following example illustrates the techniques discussed.

EXAMPLE 1.

$$L(\lambda,\alpha)=\begin{bmatrix} -1+\lambda-2\alpha & \alpha & \lambda^3 \\ \alpha & -\lambda+\lambda^2 & \alpha+\alpha^2 \\ 0 & \lambda\alpha & \lambda^2 \end{bmatrix}.$$

There is an eigenvalue $\lambda_0=1$ at $\alpha=0$. This case is constructed so that the reduction of $L(\lambda,0)$ to Smith form is easy (although this is not necessary in general). It is found that, with

$$E(\lambda)=\begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda^{-1} & 0 \\ 0 & 0 & \lambda^{-2} \end{bmatrix},\quad F(\lambda)=\begin{bmatrix} 1 & 0 & -\lambda \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$F(\lambda)L(\lambda,0)E(\lambda)=\begin{bmatrix} \lambda-1 & 0 & 0 \\ 0 & \lambda-1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

the Smith normal form. Bases for $\mathcal{K}$ *and* $\mathcal{K}'$ *can then be formed from* $E(1)$ *and* $F(1)$*:*

$$\mathcal{K} = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\right\}, \qquad \mathcal{K}' = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\right\}.$$

Then

$$[L_{10}]_{\mathcal{K},\mathcal{K}'} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad [L_{01}]_{\mathcal{K},\mathcal{K}'} = \begin{bmatrix} -2 & 0 \\ 1 & 0 \end{bmatrix},$$

and

$$\mathcal{P}(\mu) = \begin{bmatrix} \mu - 2 & 0 \\ 1 & \mu \end{bmatrix}.$$

To find the determining equation directly from the definition, write

$$L(\lambda, \alpha) = L_0 + (\lambda - 1)L_{10} + \alpha L_{01} + \cdots$$

and substitute $\lambda(\alpha) = 1 + \lambda'\alpha + o(\alpha)$ *to obtain*

$$L(\lambda, \alpha) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} (\lambda' - 2) & 1 & 3\lambda' \\ 1 & \lambda' & 1 \\ 0 & 1 & 2\lambda' \end{bmatrix} \alpha + (\text{terms of } o(|\alpha|)).$$

Then it is found that

$$\det L(\lambda, \alpha) = \lambda'(\lambda' - 2)\alpha^2 + o(|\alpha|^2).$$

Thus, the determining equation is $\lambda'(\lambda' - 2) = 0$ *and agrees with the characteristic equation of* $\mathcal{P}(\mu)$. □

5. Analytic eigenvalues. The main result of this section depends on techniques and results developed in [HL]. Some preparations are needed and are contained in the following definition and lemmas. This definition comes from [HL].

DEFINITION. *Let* $\lambda(\alpha)$, $x(\alpha)$ *be an eigenvalue-eigenvector pair of the form* (2.4), (2.5), *respectively. Then* $x(0)$ $(= \xi_0)$ *is called a generating eigenvector of* $L(\lambda, \alpha)$ *(at the point* $(\lambda_0, 0)$ *and associated with* $\lambda(\alpha)$*).*

Consider also the adjoint matrix function L_* defined by

$$L_*(\lambda, \alpha) = (L(\bar{\lambda}, \bar{\alpha}))^*. \tag{5.1}$$

Using $L_*(\lambda, \alpha)$ instead of $L(\lambda, \alpha)^*$ requires reformulation of some results from [HL]. Thus, the next two lemmas are equivalent to Lemmas 3.1 and 3.2 of [HL], and the proofs are the same. Note, in particular, that $\lambda(\alpha)$ is an eigenvalue of $L(\lambda, \alpha)$ if and only if $\overline{\lambda(\bar{\alpha})}$ is an eigenvalue of $L_*(\lambda, \alpha)$.

LEMMA 7. *Let* $\lambda_1(\alpha)$ *and* $\lambda_2(\alpha)$ *be different eigenvalue functions of the form* (2.4) *on the same cut neighborhood of* $\alpha = 0$. *Let* $x_1(\alpha)$ *be an eigenvector of the form* (2.5) *corresponding to the eigenvalue* $\lambda_1(\alpha)$ *of* $L(\lambda, \alpha)$, *and let* $y_2(\alpha)$ *be an eigenvector of the form* (2.5) *corresponding to the eigenvalue* $\overline{\lambda_2(\bar{\alpha})}$ *of* $L_*(\lambda, \alpha)$. *Then* $(L_{10}x_1(0),\, y_2(0)) = 0$.

LEMMA 8. *Let* $\lambda_1(\alpha), \ldots, \lambda_q(\alpha)$ *be eigenvalues of* $L(\lambda, \alpha)$ *which tend to* λ_0 *as* $\alpha \to 0$ *and which constitute all the different branches of the same Puiseux series* (2.1).

Then there are corresponding eigenvectors of the form (2.5)*, say* $x_1(\alpha), \dots, x_q(\alpha)$*, for which* $x_1(0) = \cdots = x_q(0)$

In particular, it follows from this lemma that there is at least one generating eigenvector of $L(\lambda, \alpha)$ at the point $(\lambda_0, 0)$.

Theorem 3.6 of [HL] can be refined in the following way. (The proof is practically unaltered.)

THEOREM 9. *Let* $\lambda_1(\alpha)$ *be an eigenvalue function of* $L(\lambda, \alpha)$ *of the form* (2.4)*. Assume that, for every generating eigenvector* x *of* $L(\lambda, \alpha)$ *associated with* $\lambda_1(\alpha)$*, there exists a generating eigenvector* y *of* $L_*(\lambda, \alpha)$ *associated with* $\overline{\lambda_1(\bar{\alpha})}$ *such that*

$$(L_{10}x,\ y) \neq 0.$$

Then $\lambda_1(\alpha)$ *depends on* α *analytically, and there is a corresponding eigenvector* $x(\alpha)$ *which is analytic in* α *for* α *sufficiently close to zero.*

Proof. Assume that $\lambda_1(\alpha)$ is nonanalytic. Then $\lambda_1(\alpha)$ is a branch of a Puiseux series at λ_0. Let $\lambda_2(\alpha)$ be a different branch of the same algebraic function. By Lemma 8, there are corresponding continuous eigenvectors $x_1(\alpha)$ and $x_2(\alpha)$ of the form (2.5) such that $x_1(0) = x_2(0) := x_0$.

Now let $\lambda(\alpha)$ and $y(\alpha)$ be *any* eigenvalue-eigenvector pair of $L_*(\lambda, \alpha)$. Then $\overline{\lambda(\bar{\alpha})}$ is an eigenvalue of $L(\lambda, \alpha)$, and, since $\overline{\lambda(\bar{\alpha})}$ cannot coincide identically with both $\lambda_1(\alpha)$ and $\lambda_2(\alpha)$, it follows from Lemma 7 that $(L_{10}x_0,\ y(0)) = 0$, and this contradicts our hypothesis. Hence $\lambda_1(\alpha)$ depends on α analytically. Also, it follows from Lemma 1 that, in some neighborhood of $\alpha = 0$, there is an associated analytic eigenvector function $x(\alpha)$. ∎

In the following statement, matrix pencil $\mathcal{P}(\mu)$ is defined in (4.2), and a *simple* eigenvalue of $\mathcal{P}(\mu)$ is just a simple zero of $\det\mathcal{P}(\mu)$.

LEMMA 10. *If* λ_0 *is a semisimple eigenvalue of* $L(\lambda, \alpha)$ *and* λ' *is a simple eigenvalue of* $\mathcal{P}(\mu)$*, then, for each generating eigenvector* x_0 *of* $L(\lambda, \alpha)$ *associated with the eigenvalue* $\lambda(\alpha) = \lambda_0 + \alpha\lambda' + o(|\alpha|)$*, there exists a generating eigenvector* y_0 *of* $L_*(\lambda, \alpha)$ *at* $(\bar{\lambda}_0, 0)$ *such that*

$$(L_{10}x_0,\ y_0) \neq 0.$$

Proof. By the definition of generating eigenvector, there is an eigenvector $x(\alpha)$ of the form (2.5) corresponding to $\lambda(\alpha)$ for which $x(0) = x_0$. Consider the eigenvalue $\overline{\lambda(\bar{\alpha})}$ of $L_*(\lambda, \alpha)$, and let $y(\alpha)$ be a corresponding eigenvector of the form (2.5). Set $y_0 = y(0) \neq 0$.

Now it follows from Theorem 6 that

$$\mathcal{P}(\lambda')u = 0, \tag{5.2}$$

where vector $u = [u_j]_{j=1}^g$ $(\in \mathbb{C}^g)$ is defined via the decomposition of x_0 with respect to the basis $\{e_j\}$ for $\mathcal{K}$, i.e., $x_0 = \sum_{j=1}^g u_j e_j$.

Now consider $L_*(\lambda, \alpha)$, $\overline{\lambda(\bar{\alpha})}$, and y_0. Decompose y_0 with respect to the basis $\{f_k\}_{k=1}^g$; $y_0 = \sum_{k=1}^g v_k f_k$, and $v = [v_k] \in \mathbb{C}^g$. Clearly $\overline{\lambda(\bar{\alpha})} = \overline{\lambda_0} + \overline{\lambda'}\alpha + o(|\alpha|)$ as $\alpha \to 0$.

Applying the arguments above to the matrix function $L_*(\lambda, \alpha)$ leads to the definition of the pencil

$$\mathcal{P}_*(\mu) = \mu[L_{10}^*]_{\mathcal{K}',\mathcal{K}} + [L_{01}^*]_{\mathcal{K}',\mathcal{K}}, \tag{5.3}$$

and, by Theorem 6,

$$\mathcal{P}_*(\overline{\lambda}')v = 0. \tag{5.4}$$

However, it follows from definition (3.1) that $[M^*]_{\mathcal{K}',\mathcal{K}} = ([M]_{\mathcal{K},\mathcal{K}'})^*$, and hence

$$\mathcal{P}_*(\bar{\mu}) = (\mathcal{P}(\mu))^*. \tag{5.5}$$

Since λ' is a simple eigenvalue of $\mathcal{P}(\mu)$, there are no Jordan chains at λ'. Hence, with u from (5.2) the equation

$$\mathcal{P}(\lambda')w = -[L_{10}]_{\mathcal{K},\mathcal{K}'}u \tag{5.6}$$

has no solution w. By (5.5), $\dim(\mathrm{Ker}\mathcal{P}_*(\overline{\lambda}')) = \dim(\mathrm{Ker}\mathcal{P}(\lambda')) = 1$. Therefore, $\mathrm{Ker}\mathcal{P}_*(\overline{\lambda}') = \mathrm{span}\{v\}$, and (using the well-known criterion for the solvability of inhomogeneous equations) the fact that (5.6) has no solution implies

$$([L_{10}]_{\mathcal{K},\mathcal{K}'}u, v) \neq 0. \tag{5.7}$$

Finally, using (3.1),

$$\begin{aligned}([L_{10}]_{\mathcal{K},\mathcal{K}'}u,\ v) &= \sum_{j=1}^{g}\sum_{k=1}^{g}(L_{10}e_k, f_j)u_k\bar{v_j} \\ &= \left(L_{10}\sum_{k=1}^{g}u_ke_k, \sum_{j=1}^{g}v_jf_j\right) \\ &= (L_{10}x_0, y_0).\end{aligned}$$

Thus, the lemma follows from (5.7). □

Lemma 10 now admits a direct application of Theorem 9 to obtain the main result.

THEOREM 11. *Let λ_0 be a semisimple eigenvalue of $L(\lambda, 0)$. Suppose also that λ' is a simple eigenvalue of $\mathcal{P}(\mu)$ with corresponding eigenvector u. Then for some $\varepsilon > 0$ there exists a simple eigenvalue function $\lambda(\alpha)$ of $L(\lambda, \alpha)$ which is analytic in $|\alpha| < \varepsilon$ and satisfies $\lambda(0) = \lambda_0$, $\lambda'(0) = \lambda'$. A corresponding eigenvector $x(\alpha)$ can be chosen analytic in $|\alpha| < \varepsilon$ and such that $x(0) = \xi$, where $\xi = \sum_{i=1}^{g} u_ie_i$ and $[u_i]_1^g = u$.*

Proof. By Theorem 6 there exists an eigenvalue $\lambda(\alpha)$ with the representation

$$\lambda(\alpha) = \lambda_0 + \lambda'\alpha + o(|\alpha|) \quad \text{as} \quad \alpha \to 0.$$

It follows from Lemma 10 that, for every generating eigenvector x_0 of $L(\lambda, \alpha)$ at $(\lambda_0, 0)$ corresponding to $\lambda(\alpha)$, there is a generating eigenvector y_0 of $L_*(\lambda, \alpha)$ at $(\overline{\lambda_0}, 0)$ such that $(L_{10}x_0,\ y_0) \neq 0$. By Theorem 9, this implies that the eigenvalue $\lambda(\alpha)$ is analytic in a neighborhood of $\alpha = 0$, and the corresponding eigenvector $x(\alpha)$ can be chosen analytic there. The statement connecting the vectors $x(0)$ and u follows from Theorem 6.

It remains to prove that, for $\alpha \neq 0$, $\lambda(\alpha)$ is a simple eigenvalue of $L(\lambda, \alpha)$. Suppose that $\hat{\lambda}(\alpha)$ is another eigenvalue of $L(\lambda, \alpha)$ defined on a neighborhood of $\alpha = 0$ with an asymptotic representation

$$\hat{\lambda}(\alpha) = \lambda_0 + \lambda''\alpha + o(|\alpha|).$$

Since λ' is a simple eigenvalue of $\mathcal{P}(\mu)$, it follows from Theorem 6 that $\lambda'' \neq \lambda'$. Since

$$\lambda(\alpha) - \hat{\lambda}(\alpha) = (\lambda' - \lambda'')\alpha + o(|\alpha|),$$

it follows that $\lambda(\alpha) \neq \hat{\lambda}(\alpha)$ when $0 < |\alpha| < \varepsilon$. □

Note that the conclusions of this theorem (with, additionally, $\lambda' \neq 0$) were deduced in [LN1] under the additional assumption that λ_0 is also α-semisimple. For the classical eigenvalue problem $L(\lambda, \alpha) = \lambda I - A(\alpha)$, the theorem is well known (see, for example, p. 269 of [B]).

The next example shows that the assumption that λ' is a simple eigenvalue for $\mathcal{P}(\mu)$ cannot be relaxed to admit a multiple semisimple eigenvalue.

EXAMPLE 2. *Consider the classical eigenvalue problem*

$$L(\lambda, \alpha) = \begin{bmatrix} \lambda - \alpha & \alpha^2 \\ \alpha^3 & \lambda - \alpha \end{bmatrix},$$

with a double semisimple eigenvalue $\lambda_0 = 0$ at $\alpha = 0$. It is easily seen that $\mathcal{P}(\mu) = \mu I_2 - I_2$ so that $\mathcal{P}(\mu)$ has a semisimple eigenvalue $\mu = 1$. However, the eigenvalue functions are $\lambda_1(\alpha) = \alpha + \alpha^{5/2}$, $\lambda_2(\alpha) = \alpha - \alpha^{5/2}$ and are not analytic at $\alpha = 0$. □

Theorem 11 immediately implies the following.

COROLLARY 12. *Let λ_0 be a semisimple eigenvalue of $L(\lambda, 0)$. Suppose also that the eigenvalues $\{\mu_j\}_{j=1}^g$ of the pencil $\mathcal{P}(\mu)$ are distinct, and let $\{u_j\}_1^g$ be a corresponding set of eigenvectors. Then there are numbers $\epsilon > 0$ and $\delta > 0$ such that, for all α satisfying $0 < |\alpha| < \epsilon$, the spectrum of $L(\lambda, \alpha)$ in $|\lambda - \lambda_0| < \delta$ consists of g distinct eigenvalues $\lambda_1(\alpha), \ldots, \lambda_g(\alpha)$ which are analytic in $|\alpha| < \epsilon$ and, for $j = 1, 2, \ldots, g$,*

$$\lambda_j(0) = \lambda_0, \qquad \lambda_j'(0) = \mu_j.$$

Corresponding eigenvectors $x_j(\alpha)$ of $L(\lambda, \alpha)$ can be chosen analytic in $|\alpha| < \epsilon$ and such that, for $j = 1, 2, \ldots, g$,

$$x_j(0) = \xi_j,$$

where $\xi_j = \sum_{i=1}^g u_{ji} e_i$ and $[u_{ji}]_{i=1}^g = u_j$.

A convenient way to determine those eigenvectors in $\mathcal{K}$ which are generating is not immediately obvious and would be useful in applications. Note, in particular, that linear combinations of generating eigenvectors are not necessarily generating. The last theorem provides a way to find the generating eigenvectors associated with a semisimple eigenvalue.

COROLLARY 13. *Let λ_0 be a semisimple eigenvalue of $L(\lambda, 0)$. Suppose also that all eigenvalues of $\hat{\mathcal{P}}(\mu)$ are distinct. Then ξ_o is a generating eigenvector at λ_0 if and only if it is an eigenvector of $\hat{\mathcal{P}}(\mu)$.*

(Note that eigenvalues of $\hat{\mathcal{P}}(\mu)$ are, by definition, eigenvalues of any matrix representation $\mathcal{P}(\mu)$ and eigenvectors of $\hat{\mathcal{P}}(\mu)$ are necessarily eigenvectors of $L(\lambda, 0)$ at λ_0.)

Proof. Corollary 12 shows that each eigenvector of $\hat{\mathcal{P}}(\mu)$ can be extended analytically into a neighborhood of $\alpha = 0$ as an eigenvector of $L(\lambda, \alpha)$, and, therefore, each of these g linearly independent eigenvectors is generating.

Conversely, given a generating eigenvector ξ_0 of $L(\lambda, \alpha)$ at λ_0, there is an eigenvector function $x(\alpha)$ with a Puiseux expansion (2.5), and, as in Theorem 6, it follows that ξ_0 is an eigenvector of $\hat{\mathcal{P}}(\mu)$. □

6. Taylor coefficients. Consider the Taylor decomposition of $L(\lambda,\alpha)$, valid in some neighborhood of $(\lambda_0,0)$ where λ_0 is a semisimple eigenvalue of $L(\lambda,0)$:

$$L(\lambda,\alpha)=\sum_{j,k=0}^{\infty}(\lambda-\lambda_0)^j\alpha^k L_{jk},$$

and $L_{00}=L_0$. As above, write $\operatorname{Ker}L_0=\mathcal{K}$, $\operatorname{Ker}L_0^*=\mathcal{K}'$, and

$$\hat{\mathcal{P}}(\mu)=\mu\hat{L}_{10}+\hat{L}_{01}. \tag{6.1}$$

Let this pencil have a simple eigenvalue b_1 with associated eigenvector ξ_0. By Theorem 11, there is an analytic eigenvalue function of $L(\lambda,\alpha)$,

$$\lambda(\alpha)=\lambda_0+\sum_{k=1}^{\infty}b_k\alpha^k, \tag{6.2}$$

and a corresponding analytic eigenvector function,

$$x(\alpha)=\sum_{k=0}^{\infty}\xi_k\alpha^k, \tag{6.3}$$

both valid in a neighborhood of $\alpha=0$, and b_1, ξ_0 are the eigenvalue and eigenvector of $\hat{\mathcal{P}}(\mu)$ introduced above. The series of (6.2) and (6.3) can be substituted in the identity

$$L(\lambda(\alpha),\alpha)\,x(\alpha)=0 \tag{6.4}$$

to obtain

$$\sum_{i,j=0}^{\infty}\left(\sum_{k=1}^{\infty}b_k\alpha^k\right)^i\alpha^j L_{ij}\left(\sum_{m=0}^{\infty}\xi_m\alpha^m\right)=0. \tag{6.5}$$

The constant term on the left is $L_{00}\xi_0=0$. By equating coefficients of α^j to zero for $j=1,2,\ldots$, an infinite system of equations is obtained for the numbers b_2, $b_3,\ldots$ and vectors $\xi_1,\xi_2,\ldots$. Thus,

$$L_0\xi_1+(b_1L_{10}+L_{01})\xi_0=0, \tag{6.6}$$

$$L_0\xi_2+(b_1L_{10}+L_{01})\xi_1+(b_2L_{10}+b_1^2L_{20}+b_1L_{11}+L_{02})\xi_0=0, \tag{6.7}$$

$$\begin{aligned}&L_0\xi_3+(b_1L_{10}+L_{01})\xi_2+(b_2L_{10}+b_1^2L_{20}+b_1L_{11}+L_{02})\xi_1\\&+(b_3L_{10}+2b_1b_2L_{20}+b_2L_{11}+b_1^3L_{30}+b_1^2L_{21}+b_1L_{12}+L_{03})\xi_0=0,\end{aligned} \tag{6.8}$$

$$\vdots$$

$$\begin{aligned}&L_0\xi_n+(b_1L_{10}+L_{01})\xi_{n-1}+\cdots\\&+(b_{n-1}L_{10}+\cdots+L_{0,n-1})\xi_1+(b_nL_{10}+\cdots+L_{on})\xi_0=0,\end{aligned} \tag{6.9}$$

and so on.

THEOREM 14. *Let λ_0 be a semisimple eigenvalue of $L(\lambda,0)$, and let b_1 be a simple eigenvalue of $\hat{\mathcal{P}}(\mu)$ with associated eigenvector ξ_0. Then the infinite system* (6.6), (6.7), ... *has a solution $\{b_j\}_2^\infty$, $\{\xi_j\}_1^\infty$ such that the series* (6.2) *and* (6.3) *converge in a neighborhood of $\alpha=0$ and represent there an eigenvalue-eigenvector pair of $L(\lambda,\alpha)$.*

The numbers $\{b_j\}_2^\infty$ *are uniquely determined by this system. The solution* $\{b_j\}_2^\infty$, $\{\xi_j\}_1^\infty$, *which gives an analytic eigenvalue-eigenvector pair* $\lambda(\alpha)$, $x(\alpha)$, *can be found by successive computation of the unknowns* $\xi_1, b_2, \xi_2, b_3, \ldots$.

Proof. The first statement follows immediately from the existence of an analytic eigenvalue-eigenvector pair (Theorem 11).

Before proving the uniqueness statement, we need a preliminary remark. Since b_1 is a simple eigenvalue of $\mathcal{P}(\mu)$, $\operatorname{Ker}(b_1\hat{L}_{10} + \hat{L}_{01})^*$ has dimension one. So write $\operatorname{Ker}(b_1\hat{L}_{10}+\hat{L}_{01})^* = \operatorname{span}\{\eta_0\}$, where $\eta_0 \neq 0$. Also, there is no Jordan chain associated with the eigenvalue b_1, so the equation

$$(b_1\hat{L}_{10} + \hat{L}_{01})x = -\hat{L}_{10}\xi_0$$

has no solution. Hence the right-hand side is not orthogonal to η_0:

$$(\hat{L}_{10}\xi_0,\ \eta_0) \neq 0. \tag{6.10}$$

Now suppose that b_1 and ξ_0 are given (as above), and there are two solutions: $\{b_k\}_2^\infty$, $\{\xi_k\}_1^\infty$ and $\{b'_k\}_2^\infty$, $\{\xi'_k\}_1^\infty$ of the system (6.6), (6.7), ... It is to be proved that

$$b_k = b'_k, \qquad k = 2, 3, \ldots. \tag{6.11}$$

Together with the system of equations (6.6), (6.7), ... obtained by substituting the first solution $\{b_k\}_1^\infty$, $\{\xi_k\}_1^\infty$ in the system, consider the system obtained by substituting the second solution $\{b'_k\}_1^\infty$, $\{\xi'_k\}_1^\infty$. Denote this new set of equations by (6.6′), (6.7′), ...

It follows from (6.6) and (6.6′) that $\xi_1 - \xi'_1 \in \mathcal{K}$ and, from (6.7) and (6.7′), we have

$$L_0(\xi_2 - \xi'_2) + (b_1L_{10} + L_{01})(\xi_1 - \xi'_1) + (b_2 - b'_2)L_{10}\xi_0 = 0. \tag{6.12}$$

As in (b) of section 3, define $R = \sum_{j=1}^g (., f_j)f_j$, where $\{f_j\}_1^g$ is a basis for $\mathcal{K}'$. Applying this transformation we obtain

$$(b_1\hat{L}_{10} + \hat{L}_{01})(\xi_1 - \xi'_1) + (b_2 - b'_2)\hat{L}_{10}\xi_0 = 0. \tag{6.13}$$

Taking the inner product with $\eta_0 \in \mathcal{K}'$ gives

$$(b_2 - b'_2)(\hat{L}_{10}\xi_0, \eta_0) = 0,$$

and then (6.10) implies

$$b_2 = b'_2. \tag{6.14}$$

Now (6.13) can be written in the form

$$(b_1\hat{L}_{10} + \hat{L}_{01})(\xi_1 - \xi'_1) = 0,$$

which means that $\xi_1 - \xi'_1 \in \operatorname{Ker}(b_1\hat{L}_{10} + \hat{L}_{01}) = \operatorname{span}\{\xi_0\}$, i.e.,

$$\xi_1 - \xi'_1 = \alpha_1\xi_0 \tag{6.15}$$

for some $\alpha_1 \in \mathbb{C}$. It follows from (6.12), (6.14), and (6.15) that

$$L_0(\xi_2 - \xi'_2) + (b_1L_{10} + L_{01})(\alpha_1\xi_0) = 0.$$

Equation (6.6) gives $L_0(\alpha_1\xi_1)+(b_1L_{10}+L_{01})(\alpha_1\xi_0)=0$ so that $L_0(\xi_2-\xi_2'-\alpha_1\xi_1)=0$, or

$$\xi_2-\xi_2'=\alpha_1\xi_1+\xi,\qquad \xi\in\mathcal{K}. \tag{6.16}$$

Progress now to (6.8) and (6.8′). Their difference gives

$$\begin{aligned}&L_0(\xi_3-\xi_3')+(b_1L_{10}+L_{01})(\xi_2-\xi_2')\\ &+(b_2L_{10}+b_1^2L_{20}+b_1L_{11}+L_{02})(\xi_1-\xi_1')+(b_3-b_3')L_{10}\xi_0=0.\end{aligned} \tag{6.17}$$

From this equation subtract (6.7) multiplied by α_1 and use (6.16) to obtain

$$L_0(\xi_3-\xi_3'-\alpha_1\xi_2)+(b_1L_{10}+L_{01})\xi+(b_3-b_3')L_{10}\xi_0=0. \tag{6.18}$$

Apply the transformation R to obtain

$$(b_1\hat{L}_{10}+\hat{L}_{01})\xi+(b_3-b_3')\hat{L}_{10}\xi_0=0,$$

and taking the inner product with η_0, $((b_3-b_3')\hat{L}_{10}\xi_0,\eta_0)=0$, whence

$$b_3=b_3'. \tag{6.19}$$

Now rewrite (6.18) in the form

$$L_0(\xi_3-\xi_3'-\alpha_1\xi_2)+(b_1L_{10}+L_{01})\xi=0, \tag{6.20}$$

and use R to get $(b_1\hat{L}_{10}+\hat{L}_{01})\xi=0$. This means that $\xi=\alpha_2\xi_0$ for some $\alpha_2\in\mathbb{C}$. Hence it follows from (6.20) and (6.6) that

$$L_0(\xi_3-\xi_3'-\alpha_1\xi_2-\alpha_2\xi_1)=0,$$

i.e.,

$$\xi_3-\xi_3'=\alpha_1\xi_2+\alpha_2\xi_1+\xi,\qquad \xi\in\mathcal{K}. \tag{6.21}$$

Now suppose that, inductively,

$$\begin{aligned}b_2=b_2',\ b_3=b_3',\dots,b_{n-1}&=b_{n-1}',\\ \xi_1-\xi_1'&=\alpha_1\xi_0,\\ \xi_2-\xi_2'&=\alpha_1\xi_1+\alpha_2\xi_0,\\ &\vdots\\ \xi_{n-2}-\xi_{n-2}'&=\alpha_1\xi_{n-3}+\alpha_2\xi_{n-4}+\cdots+\alpha_{n-2}\xi_0,\\ \xi_{n-1}-\xi_{n-1}'&=\alpha_1\xi_{n-2}+\alpha_2\xi_{n-3}+\cdots+\alpha_{n-2}\xi_1+\xi,\end{aligned}$$

where $\xi\in\mathcal{K}$.

Consider the difference of equations (6.9) and (6.9′):

$$\begin{aligned}&L_0(\xi_n-\xi_n')+(b_1L_{10}+L_{01})(\xi_{n-1}-\xi_{n-1}')+\cdots\\ &+(b_{n-1}L_{10}+\cdots+L_{0,n-1})(\xi_1-\xi_1')+(b_n-b_n')L_{10}\xi_0=0.\end{aligned}$$

Subtract from this equation the $(n-1)$st equation of the system (6.6), (6.7), ... multiplied by α_1, then subtract the preceding equation multiplied by α_2, and so on, ending with (6.7) multiplied by α_{n-2}. Using the induction hypotheses, we obtain

$$L_0(\xi_n-\xi_n'-\alpha_1\xi_{n-1}-\cdots-\alpha_{n-2}\xi_2)+(b_1L_{10}+L_{01})\xi+(b_n-b_n')L_{10}\xi_0=0, \tag{6.22}$$

where $\xi \in \mathcal{K}$ is the last term in the induction hypotheses. Act on (6.22) with R to obtain

$$(b_1\hat{L}_{10} + \hat{L}_{01})\xi + (b_n - b_n')\hat{L}_{10}\xi_0 = 0.$$

Take the inner product with η_0 and divide by $(\hat{L}_{10}\xi_0, \eta_0)$ to obtain

$$b_n = b_n'. \tag{6.23}$$

Now (6.22) becomes

$$L_0(\xi_n - \xi_n' - \alpha_1\xi_{n-1} - \cdots - \alpha_{n-2}\xi_2) + (b_1L_{10} + L_{01})\xi = 0. \tag{6.24}$$

Apply R to obtain $(b_1\hat{L}_{10} + \hat{L}_{01})\xi = 0$. Hence $\xi = \alpha_{n-1}\xi_0$ for some $\alpha_{n-1} \in \mathbb{C}$, and it follows from (6.24) and (6.6) that

$$L_0(\xi_n - \xi_n' - \alpha_1\xi_{n-1} - \cdots - \alpha_{n-2}\xi_2 - \alpha_{n-1}\xi_1) = 0,$$

or

$$\xi_n - \xi_n' = \alpha_1\xi_{n-1} + \alpha_2\xi_{n-2} + \cdots + \alpha_{n-2}\xi_2 + \alpha_{n-1}\xi_1 + \xi$$

for some $\xi \in \mathcal{K}$. All the induction hypotheses are satisfied for the number n, so $b_n = b_n'$ holds for any n.

Now an algorithm is formulated for the computation of the coefficients $\{b_j\}_{j=2}^{\infty}$ and $\{\xi_j\}_1^{\infty}$ of the expansions (6.2) and (6.3). The equation $L_0u = v$ has a solution if and only if v is orthogonal to the subspace $\mathcal{K}' = \mathrm{Ker}(L_0^*)$, which is then equivalent to $Rv = 0$. Thus (6.6) has a solution ξ_1 if and only if

$$R(b_1L_{10} + L_{01})\xi_0 = 0,$$

or

$$(b_1\hat{L}_{10} + \hat{L}_{01})\xi_0 = 0.$$

But this holds because ξ_0 is an eigenvector of $\hat{\mathcal{P}}$ at $\mu = b_1$.

Now let ξ_1^0 be a particular solution of (6.6) and write the general solution in the form $\xi_1 = \xi_1^0 + \xi$ where $\xi \in \mathcal{K}$. The vector ξ will be chosen to ensure the existence of a solution of the next equation, (6.7). This equation has a solution ξ_2 if and only if

$$R\{(b_1L_{10} + L_{01})\xi_1 + (b_2L_{10} + b_1^2L_{20} + b_1L_{11} + L_{02})\xi_0\} = 0. \tag{6.25}$$

Substitute $\xi_1 = \xi_1^0 + \xi$ and rewrite (6.25) in the form

$$(b_1\hat{L}_{10} + \hat{L}_{01})\xi = -R(b_1L_{10} + L_{01})\xi_1^0 - (b_2\hat{L}_{10} + b_1^2\hat{L}_{20} + b_1\hat{L}_{11} + \hat{L}_{02})\xi_0. \tag{6.26}$$

Now the number b_2 is determined in such a way that (6.26) has a solution ξ, i.e., the right-hand side of (6.26) will be orthogonal to $\mathrm{Ker}(b_1\hat{L}_{10} + \hat{L}_{01})^*$. The latter subspace has the same dimension as $\mathrm{Ker}(b_1\hat{L}_{10} + \hat{L}_{01})$, namely one. Thus, $\mathrm{Ker}(b_1\hat{L}_{10} + \hat{L}_{01})^* = \mathrm{span}\{\eta_0\}$ for some $\eta_0 \in \mathcal{K}'$, and the right-hand side of (6.26) has to be orthogonal to η_0. Given (6.10), we may define

$$b_2 = -\frac{1}{(\hat{L}_{10}\xi_0, \eta_0)}\left(R(b_1L_{10} + L_{01})\xi_1^0 + (b_1^2\hat{L}_{20} + b_1\hat{L}_{11} + \hat{L}_{02})\xi_0,\ \eta_0\right), \tag{6.27}$$

and then the orthogonality condition holds.

Thus, if b_2 is defined by (6.27), then (6.26) has a solution $\xi = \xi_1' \in \mathcal{K}$, and for $\xi_1 = \xi_1^0 + \xi_1'$ (6.25) is satisfied. Hence (6.7) has a solution ξ_2^0 and its general solution is $\xi_2 = \xi_2^0 + \xi$, $\xi \in \mathcal{K}$. Then the vector ξ is to be chosen when considering the next equation, (6.8).

The general induction step can now be formulated. Suppose that numbers $b_2, \dots,$ b_{n-1} and vectors $\xi_1, \dots, \xi_{n-2}$ have been obtained which satisfy the first $n-1$ equations of the infinite system together with vector $\xi_{n-1} = \xi_{n-1}^0 + \xi$, where ξ is a vector from $\mathcal{K}$ to be determined via the nth equation of the system, (6.9).

This equation is solvable for ξ_n if and only if

$$R\left((b_1 L_{10} + L_{01})\xi_{n-1} + \cdots + (b_n L_{10} + \cdots + L_{0n})\xi_0\right) = 0. \tag{6.28}$$

Substitute $\xi_{n-1} = \xi_{n-1}^0 + \xi$ ($\xi \in \mathcal{K}$) and rewrite (6.28) in the form

$$\begin{aligned}(b_1 \hat{L}_{10} + \hat{L}_{01})\xi = -R\{&(b_1 L_{10} + L_{01})\xi_{n-1}^0 + (b_2 L_{10} + \cdots + L_{02})\xi_{n-2} + \cdots \\ &+ (b_n L_{10} + 2b_1 b_{n-1} L_{20} \cdots + L_{0n})\xi_0\}.\end{aligned} \tag{6.29}$$

Using (6.10), we define

$$\begin{aligned}b_n = -\frac{1}{(\hat{L}_{10}\xi_0,\ \eta_0)}(R\{&(b_1 L_{10} + L_{01})\xi_{n-1}^0 + (b_2 L_{10} + \cdots + L_{02})\xi_{n-2} + \cdots \\ &+ (2b_1 b_{n-1} L_{20} + \cdots L_{0n})\xi_0\},\ \eta_0),\end{aligned}$$

and then the right-hand side of (6.29) will be orthogonal to η_0, and so (6.29) will have a solution $\xi = \xi_{n-1}' \in \mathcal{K}$, and, for $\xi_{n-1} = \xi_{n-1}^0 + \xi_{n-1}'$, equation (6.28) holds. Hence (6.9) has a solution ξ_n^0. The general solution of this equation is $\xi_n = \xi_n^0 + \xi$ with $\xi \in \mathcal{K}$.

This completes the induction and the proof that the proposed algorithm admits the successive calculation of the coefficients $\xi_1, b_2, \xi_2, b_3 \dots$.

The following arguments will show that the series of (6.2) and (6.3) has positive radii of convergence. For (6.2) this is clear, because $\lambda(\alpha)$ is known to be analytic and the coefficients $\{b_j\}_2^\infty$ are uniquely determined. But we do not have this assurance for (6.3). A more precise choice of the vectors $\{\xi_j\}_1^\infty$ will now be made so that estimates of their norms can be made.

Since $L(\lambda, \alpha)$ and $\lambda(\alpha)$ are analytic functions, the Cauchy inequalities can be applied and yield

$$\|L_{ij}\| \le \frac{M}{r^{i+j}} \qquad (i, j = 0, 1, \dots), \qquad |b_j| \le \frac{m}{\rho^j} \qquad (j = 1, 2, \dots)$$

for some positive $M,\ r,\ m,\ \rho$. It will be convenient (and sufficient) to use less precise inequalities:

$$\|L_{ij}\| \le C^{i+j+1} \qquad (i, j = 0, 1, \dots), \qquad |b_j| \le C^{j+1} \qquad (j = 1, 2, \dots), \tag{6.30}$$

where $C = \max(M, r^{-1}, m, \rho^{-1})$.

Let ν_1 be a positive number with the following property: if the equation $L_0x = y$ holds $(x, y \in \mathcal{H})$, then there is a solution x such that $\|x\| \le \nu_1\|y\|$. To achieve this it is sufficient to take, for example, the solution orthogonal to $\operatorname{Ker} L_0$. Similarly, let ν_2 be a positive number with the property that: if the equation $(b_1\hat{L}_{10} + \hat{L}_{01})\xi = \eta$ holds $(\xi \in \mathcal{K},\ \eta \in \mathcal{K}')$, then there is a solution ξ such that $\|\xi\| \le \nu_2\|\eta\|$.

By construction, $\xi_n = \xi_n^0 + \xi_n'$, and $\|\xi_n^0\|$, $\|\xi_n'\|$ will be estimated separately. The solution ξ_n^0 of (6.9) is chosen in such a way that

$$\|\xi_n^0\| \le \nu_1\|(b_1L_{10} + L_{01})\xi_{n-1} + \cdots + (b_nL_{10} + \cdots L_{0n})\xi_0\|. \tag{6.31}$$

Vector ξ_{n-1}' is a solution of (6.29) and is chosen in such a way that

$$\begin{aligned}\|\xi_{n-1}'\| \le \nu_2\|R\{&(b_1L_{10} + L_{01})\xi_{n-1}^0 \\ &+ (b_2L_{10} + \cdots + L_{02})\xi_{n-2} + \cdots + (b_nL_{10} + \cdots + L_{0n})\xi_0\}\|.\end{aligned} \tag{6.32}$$

Consider a typical term from (6.31) or (6.32):

$$(b_kL_{10} + \cdots + L_{0k})\xi_{n-k}. \tag{6.33}$$

To proceed it is necessary to estimate the number of summands in the parenthesis of this expression. Call the expression in parentheses S_k. Rewriting the analytic matrix function $L(\lambda(\alpha), \alpha)$ as a power series in α, we obtain

$$L(\lambda(\alpha), \alpha) = \sum_{i,j=0}^{\infty}\left(\sum_{k=1}^{\infty} b_k\alpha^k\right)^i \alpha^j L_{ij} = L_0 + \sum_{k=1}^{\infty}\alpha^k S_k.$$

In the simplest case when $L_{ij} \equiv I$ and $b_k \equiv 1$, we obtain $L(\lambda(\alpha), \alpha) = f(\alpha)I$, where

$$\begin{aligned}f(\alpha) &= \sum_{i,j=0}^{\infty}\left(\sum_{k=1}^{\infty}\alpha^k\right)^i \alpha^j = \sum_{j=0}^{\infty}\alpha^j\sum_{i=0}^{\infty}\left(\frac{\alpha}{1-\alpha}\right)^i \\ &= (1-\alpha)^{-1}\left(1 - \frac{\alpha}{1-\alpha}\right)^{-1} = \frac{1}{1-2\alpha} = \sum_{k=0}^{\infty}2^k\alpha^k.\end{aligned}$$

Consequently, if every summand in S_k is replaced by 1, we obtain 2^k. This means that the number of summands is 2^k (if we do not collect similar terms; e.g., $b_1b_2L_{20} + b_2b_1L_{20}$ counts as two terms in (6.8)).

It is easy to see that all the summands in S_k are of the form

$$b_{p_1}b_{p_2}\ldots b_{p_i}L_{ij}, \tag{6.34}$$

where $p_1, \ldots, p_i \ge 1,\ i, j \ge 0,\ i + j \ge 1$, and $p_1 + p_2 + \cdots + p_i + j = k$.

Now estimate the term (6.34) using (6.30) to obtain

$$|b_{p_1}b_{p_2}\ldots b_{p_i}|\,\|L_{ij}\| \le C_1^{p_1+p_2+\cdots+p_i+i}C^{i+j+1} = C^{k+2i+1} \le C^{4k},$$

since $i \le k$ and $k \ge 1$. Finally, $\|S_k\| \le 2^kC^{4k}$, and from (6.31), (6.32),

$$\|\xi_n^0\| \le \nu_1(C_1\|\xi_{n-1}\| + C_1^2\|\xi_{n-2}\| + \cdots + C_1^n\|\xi_0\|), \tag{6.35}$$

$$\|\xi_{n-1}'\| \le \nu_2\|R\|(C_1\|\xi_{n-1}^0\| + C_1^2\|\xi_{n-2}\| + \cdots + C_1^n\|\xi_0\|),$$

where $C_1 = 2C^4$. Rewrite the last inequality replacing $n-1$ by n to obtain

$$\|\xi_n'\| \le \nu_2\|R\|(C_1\|\xi_n^0\| + C_1^2\|\xi_{n-1}\| + \cdots + C_1^{n+1}\|\xi_0\|). \tag{6.36}$$

From (6.35) and (6.36), a recursive inequality is obtained for the norm of $\xi_n = \xi_n^0 + \xi_n'$:

$$\begin{aligned}\|\xi_n\| &\le (\nu_2\|R\|C_1 + 1)\|\xi_n^0\| + \nu_2\|R\|(C_1^2\|\xi_{n-1}\| + \cdots + C_1^{n+1}\|\xi_0\|)\\ &\le \nu(C_1\|\xi_{n-1}\| + \cdots + C_1^n\|\xi_0\|),\end{aligned} \tag{6.37}$$

where $\nu := (\nu_2\|R\|C_1 + 1)\nu_1 + \nu_2\|R\|C_1$.

Now use the following elementary statement:

(a) If a sequence of positive numbers $\{t_n\}_0^\infty$ satisfies

$$t_n \le \nu(C_1 t_{n-1} + C_1^2 t_{n-2} + \cdots + C_1^n t_0) \quad (n \ge 1)$$

for some positive ν and C_1, then $t_n \le C_2^{n+1},\ n = 0, 1, \ldots$, where $C_2 = \max(2\nu C_1, t_0)$.

This can be proved by a direct induction argument assuming initially that the result holds for all $k < n$.

Now statement (a) and inequality (6.37) imply that $\|\xi_n\| \le C_2^{n+1}$ $(n = 0, 1, \ldots)$ for some $C_2 > 0$. Hence the series (6.3) converges for $|\alpha| < C_2^{-1}$, and its sum represents an analytic eigenvector function corresponding to the eigenvalue $\lambda(\alpha)$. □

This section is concluded with a discussion of the connection between the solutions of the infinite system and its finite subsystems.

THEOREM 15. *Let the hypotheses of Theorem* 14 *hold. Consider the subsystem of* n *equations* (6.6)–(6.9) *from the infinite system. Every solution of this subsystem has the form* $\{b_j\}_2^n,\ \{\xi_j^{(n)}\}_1^n$, *where the numbers* $\{b_j\}_2^n$ *are uniquely determined and coincide with the corresponding Taylor coefficients of the analytic eigenvalue* $\lambda(\alpha)$ *satisfying* $\lambda(0) = \lambda_0,\ \lambda'(0) = b_1$.

There is a corresponding eigenvector function $x(\alpha)$ *which is analytic in a neighborhood of* $\alpha = 0$ *and such that the polynomial*

$$\xi_0 + \sum_{j=1}^{n-1} \xi_j^{(n)} \alpha^j$$

is the $(n-1)$*st Taylor polynomial of* $x(\alpha)$ *at* $\alpha = 0$.

In other words, the theorem asserts that the polynomials

$$\lambda_0 + \sum_{j=1}^{n} b_j \alpha^j, \qquad \xi_0 + \sum_{j=1}^{n} \xi_j^{(n)} \alpha^j$$

formed from a solution of the subsystem (6.6)–(6.9) can be extended to an analytic eigenvalue-eigenvector pair for $L(\lambda, \alpha)$ at the point $(\lambda_0, 0)$—with the possible exception that the last coefficient of the second polynomial, $\xi_n^{(n)}$, may have to be replaced.

Proof. The first statement of the theorem has already been established in proving Theorem 14. Using the vectors $\xi_1^{(n)}, \ldots, \xi_{n-1}^{(n)}$ instead of $\xi_1, \ldots, \xi_{n-1}$ and $\xi_n^{(n)}$ instead of ξ_n^0, and repeating the inductive construction from the proof of Theorem 14, a vector $\xi_n' \in \mathcal{K}$ can be found such that, with $\xi_n = \xi_n^0 + \xi_n'$, the $(n+1)$st equation of the infinite system can be solved for ξ_{n+1}, and so on. A solution $\{b_j\}_2^\infty,\ \{\xi_j\}_1^\infty$ of the infinite system is obtained such that $\xi_j = \xi_j^{(n)}$ for $j = 1, 2, \ldots, n-1$. Now it only has to be shown that the radius of convergence of the series $\sum_{j=0}^\infty \xi_j \alpha^j$ is positive.

With this in mind, return to the estimates of $\|\xi_k\|$. As in the proof of the last part of Theorem 14, vectors ξ_n, $\xi_{n+1}, \dots$ can be chosen satisfying

$$\|\xi_k\| \le \nu(C_1\|\xi_{k-1}\| + \cdots + C_1^k\|\xi_0\|)$$

for $k \ge n$. Now use a slightly modified form of statement (a) (from the proof of Theorem 14):

(a′) If a sequence of positive numbers $\{t_k\}_0^\infty$ satisfies

$$t_k \le \nu(C_1 t_{k-1} + C_1^2 t_{k-2} + \cdots + C_1^k t_0)$$

for some positive ν and C_1 and for any $k \ge n$, then

$$t_k \le C_2^{k+1} \quad \text{for} \quad k = 1, 2, \dots,$$

where $C_2 = \max(2\nu C_1,\ t_0,\ t_1^{1/2},\ t_2^{1/3}, \dots, t_{n-1}^{1/n})$.

Using this statement, we find that there is a number $C_2 > 0$ such that $\|\xi_k\| \le C_2^{k+1}$ for all k. This proves the analyticity of $x(\alpha) = \sum_{j=0}^\infty \xi_j \alpha^j$ in a neighborhood of $\alpha = 0$. □

7. Self-adjoint functions. An analytic matrix function $L(\lambda, \alpha)$ is said to be *self-adjoint* if $L_*(\lambda, \alpha) = L(\lambda, \alpha)$ for all λ and α. From the definition (5.1) of L_*, it is clear that, if λ_0 is an eigenvalue of L at α_0, then $\overline{\lambda}_0$ is an eigenvalue of L_* at $\overline{\alpha}_0$. Furthermore, it follows readily from the Smith canonical form that these two eigenvalues have the same partial multiplicities. In particular, if α is confined to the real numbers, then the nonreal eigenvalues of self-adjoint functions arise in complex conjugate pairs having the same partial multiplicity structure.

Now consider the case of a real eigenvalue λ_0 at $\alpha = 0$, and recall that, when $L(\lambda, \alpha)$ is self-adjoint and $\lambda \in \mathbb{R}$, there is a unitary decomposition as follows (see Theorem 1.1 of [GLR1]; the semisimple hypothesis is made here for convenience).

PROPOSITION 16. *Let λ_0 be a semisimple real eigenvalue of multiplicity g of the self-adjoint function $L(\lambda, \alpha)$ at $\alpha = 0$. Then, on a neighborhood $\mathcal{N}$ of λ_0, there is an analytic matrix function $U(\lambda)$ which is unitary for $\lambda \in \mathcal{N} \cap \mathbb{R}$ and such that*

$$L(\lambda, 0) = U(\lambda)K(\lambda)U(\lambda)^{-1} \quad (\lambda \in \mathcal{N}), \tag{7.1}$$

$$K(\lambda) = \operatorname{diag}[(\lambda - \lambda_0)r_1(\lambda), \dots, (\lambda - \lambda_0)r_g(\lambda), 1, \dots, 1], \tag{7.2}$$

and $r_1(\lambda), \dots, r_g(\lambda)$ are real analytic functions for which $\rho_j := r_j(\lambda_0) \ne 0$ for $j = 1, 2, \dots, g$.

This proposition is now used to determine a corresponding pencil $\mathcal{P}(\mu)$ (of (4.2)). Let $x_1, \dots, x_n$ be the standard basis for $\mathcal{H} = \mathbb{C}^n$ (i.e., x_j has a one in position j and zeros elsewhere), and let $h_j = U(\lambda_0)x_j$. By (7.1) and (7.2), $\{h_j\}_1^g$ is an orthonormal basis for subspace $\mathcal{K} = \mathcal{K}'$, and it can be used in the role of both bases $\{e_j\}$ and $\{f_j\}$ of section 4.

From (7.2),

$$K(\lambda_0) = \operatorname{diag}[0, \dots, 0, 1, \dots, 1],$$

$$K'(\lambda_0) = \operatorname{diag}[\rho_1, \dots, \rho_g, 0, \dots, 0],$$

and hence

$$K(\lambda_0)x_j = 0, \quad K'(\lambda_0)x_j = \rho_j x_j \quad (j = 1, \ldots, g). \tag{7.3}$$

From (7.1),

$$L_{10} = U'(\lambda_0)K(\lambda_0)U^{-1}(\lambda_0) + U(\lambda_0)K'(\lambda_0)U^{-1}(\lambda_0) + U(\lambda_0)K(\lambda_0)\frac{d}{d\lambda}U^{-1}(\lambda)|_{\lambda=\lambda_0},$$

and

$$\begin{aligned}(L_{10}h_k, h_j) &= (U'(\lambda_0)K(\lambda_0)U^{-1}(\lambda_0)h_k, h_j) + (K'(\lambda_0)U^{-1}(\lambda_0)h_k,\ U^{-1}(\lambda_0)h_j) \\ &\quad + \left(\frac{d}{d\lambda}U^{-1}(\lambda)|_{\lambda=\lambda_0}h_k,\ K(\lambda_0)U^{-1}(\lambda_0)h_j\right) \\ &= (U'(\lambda_0)K(\lambda_0)x_k, h_j) + (K'(\lambda_0)x_k, x_j) + \left(\frac{d}{d\lambda}U^{-1}(\lambda)|_{\lambda=\lambda_0}h_k, K(\lambda_0)x_j\right).\end{aligned}$$

Using (7.3), we obtain

$$(L_{10}h_k, h_j) = (K'(\lambda_0)x_k, x_j) = \rho_k\delta_{jk}$$

for $j, k = 1, \ldots, g$. Finally,

$$\mathcal{P}(\mu) = \mu R_0 + [(L_{01}h_k, h_j)]_{j,k=1}^{g}, \tag{7.4}$$

where $R_0 = \operatorname{diag}[\rho_1, \ldots, \rho_g]$. Now the results of section 4 yield the following statement.

PROPOSITION 17. *If λ_0 is a real semisimple eigenvalue of the self-adjoint analytic matrix function $L(\lambda, \alpha)$, then the eigenvalue derivatives λ_j' of* (2.3) *are the eigenvalues of the self-adjoint pencil* (7.4).

Since both coefficient matrices of the pencil (7.4) may be indefinite, the eigenvalue derivatives can be nonreal and, under real variations in α, λ_0 may split into nonreal eigenvalue functions (as in Example 3 below). In the terminology of [GLR1], the numbers $\operatorname{sgn}(\rho_j)$, $j = 1, 2, \ldots, g$, determine the *sign characteristic* of λ_0. When R_0 is positive (negative) definite, λ_0 is said to have positive (negative) type.

Note that there are cases in which the requirement of Theorem 11 that λ' be a *simple* eigenvalue of $\mathcal{P}(\mu)$ can be relaxed. For example, when R_0 is definite (of either type), then *all* the eigenvalue functions $\lambda_j(\alpha)$ emanating from λ_0 can be chosen real for real α and analytic at $\alpha = 0$ (Corollary 3.8 of [HL]). Furthermore, the same conclusion holds if the matrix $[(L_{01}h_k, h_j)]$ of (7.4) is definite (Corollary 4.4 of [HL]).

EXAMPLE 3. *The following simple example is instructive. Consider the self-adjoint function*

$$L(\lambda, \alpha) = \begin{bmatrix} \lambda & \alpha \\ \alpha & -\lambda \end{bmatrix}.$$

There is a semisimple eigenvalue $\lambda_0 = 0$ at $\alpha = 0$. Eigenvalue functions emanating from λ_0 have the form $\pm i\alpha$ and are analytic but not real. Here,

$$\mathcal{P}(\mu) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\mu + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

with eigenvalues $\pm i$.

All nonzero vectors in $\mathbb{C}^2$ *are eigenvectors at* λ_0*, but (by Corollary* 13*) generating eigenvectors are confined to the nonzero scalar multiples of*

$$\begin{bmatrix} 1 \\ i \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ -i \end{bmatrix}. \qquad \square$$

See Example 1.2 of [HL] for a self-adjoint function with a real eigenvalue which is both semisimple and α-semisimple and, nevertheless, has nonanalytic behavior under perturbations of α.

Acknowledgments. The work of the first author began during his tenure of a Research Award of the Humboldt Foundation at the Technical University of Darmstadt. The second author acknowledges the hospitality of the University of Calgary during the summer of 2002.

REFERENCES

[B] H. Baumgärtel, *Analytic Perturbation Theory*, Oper. Theory Adv. Appl. 15, Birkhäuser, Basel, 1985.

[BGR] J. A. Ball, I. Gohberg, and L. Rodman, *Interpolation of Rational Matrix Functions*, Oper. Theory Adv. Appl. 45, Birkhäuser, Basel, 1990.

[Bl] G. A. Bliss, *Algebraic Functions*, Dover Publications, New York, 1966.

[GLR1] I. Gohberg, P. Lancaster, and L. Rodman, *Perturbations of analytic hermitian matrix functions*, Appl. Anal., 20 (1985), pp. 23–48.

[GLR2] I. Gohberg, P. Lancaster, and L. Rodman, *Quadratic matrix polynomials with a parameter*, Adv. Appl. Math., 7 (1986), pp. 253–281.

[GLR3] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.

[HL] R. O. Hryniv and P. Lancaster, *On the perturbation of analytic matrix functions*, Integral Equations Operator Theory, 34 (1999), pp. 325–338.

[K] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1976.

[LN1] H. Langer and B. Najman, *Remarks on the perturbation of analytic matrix functions*, II, Integral Equations Operator Theory, 12 (1989), pp. 392–407.

[LN2] H. Langer and B. Najman, *Leading coefficients of the eigenvalues of perturbed analytic matrix functions*, Integral Equations Operator Theory, 16 (1993), pp. 600–604.

[LNV] H. Langer, B. Najman, and K. Veselić, *Perturbation of the eigenvalues of quadratic matrix polynomials*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 474–489.

[M] A. I. Markushevich, *Theory of Functions of a Complex Variable*, Vol. 2, Prentice–Hall, Englewood Cliffs, NJ, 1965.

[R] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.

[VT] M. M. Vainberg and V. A. Trenogin, *Theory of Branching of Solutions of Non-linear Equations*, Noordhoff, Leyden, 1974.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 627–629

A STABILITY PROPERTY OF T. CHAN'S PRECONDITIONER*

XIAO-QING JIN†, YI-MIN WEI‡, AND WEI XU§

Abstract. In this short note, we prove that T. Chan's preconditioner proposed in [SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771] is stable for matrices that are normal and stable.

Key words. T. Chan's circulant preconditioner, stable

AMS subject classifications. 65F10, 65N22, 65L05, 65F15

DOI. 10.1137/S0895479803422701

In 1988, T. Chan in [3] proposed a circulant preconditioner for Toeplitz systems. The use of circulant preconditioners for solving Toeplitz systems has been studied extensively since 1986; see [2, 7, 9]. For any Toeplitz matrix T_n, T. Chan's circulant preconditioner $c_F(T_n)$ proposed in [3] is defined to be the minimizer of the Frobenius norm

$$\|T_n - W_n\|_F,$$

where W_n runs over all circulant matrices. The $c_F(T_n)$ is called the optimal circulant preconditioner in [3].

Since T. Chan's preconditioner is defined not just for Toeplitz matrices but for general matrices as well, we then begin with the general case. Given a unitary matrix $U \in C^{n\times n}$, let

$$\mathcal{M}_U \equiv \{U^*\Lambda_n U \mid \Lambda_n \text{ is any } n \times n \text{ diagonal matrix}\}. \tag{1}$$

We note that in (1), when $U = F$, the Fourier matrix $\mathcal{M}_F$ is the set of all circulant matrices; see [5].

Let $\delta(E_n)$ denote the diagonal matrix whose diagonal is equal to the diagonal of the matrix E_n. The following lemma can be found in [1, 6].

LEMMA 1. *For any arbitrary $A_n = [a_{pq}] \in C^{n\times n}$, let $c_U(A_n)$ be the minimizer of*

$$\|W_n - A_n\|_F$$

over all $W_n \in \mathcal{M}_U$. Then $c_U(A_n)$ is uniquely determined by A_n and is given by

$$c_U(A_n) \equiv U^*\delta(UA_nU^*)U.$$

Proof. Since the Frobenius norm is unitary invariant, we have

$$\|W_n - A_n\|_F = \|U^*\Lambda_nU - A_n\|_F = \|\Lambda_n - UA_nU^*\|_F.$$

*Received by the editors February 11, 2003; accepted for publication (in revised form) by D. P. O'Leary June 23, 2003; published electronically December 17, 2003.

http://www.siam.org/journals/simax/25-3/42270.html

†Department of Mathematics, University of Macau, Macau, China (xqjin@umac.mo). The research of this author was supported by research grant RG031/02-03S/JXQ/FST from the University of Macau.

‡Departmant of Mathematics, Fudan University, Shanghai, 200433, China (ymwei@fudan.edu.cn). The research of this author was supported by the National Natural Science Foundation of China under grant 19901006.

§Institute of Mathematics, Fudan University, Shanghai, 200433, China (xu_wei8@hotmail.com).

Thus the problem of minimizing $\|W_n - A_n\|_F$ over $\mathcal{M}_U$ is equivalent to the problem of minimizing $\|\Lambda_n - UA_nU^*\|_F$ over all diagonal matrices. Since Λ_n can only affect the diagonal entries of UA_nU^*, we see that the solution for the latter problem is $\Lambda_n = \delta(UA_nU^*)$. Hence

$$c_U(A_n) \equiv U^*\delta(UA_nU^*)U$$

is the minimizer of $\|W_n - A_n\|_F$. □

It is natural to ask if a general matrix A_n is stable, i.e., the real part of the eigenvalues of A_n are negative, how about T. Chan's preconditioner $c_U(A_n)$? We should emphasize that the stable property of a matrix is very important in control theory and dynamic systems; see [4, 8].

Let us first consider the following example:

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

We immediately have

$$c_F(A) = \begin{bmatrix} -1 & 2 \\ 2 & -1 \end{bmatrix}.$$

It is easy to check that the eigenvalues of A are all -1, but the eigenvalues of $c_F(A)$ are 1 and -3, i.e., T. Chan's preconditioner cannot keep the stable property in general.

We want to investigate when T. Chan's preconditioner will be stable.

THEOREM 2. *Let A_n be normal and stable. Then T. Chan's preconditioner $c_U(A_n)$ is also stable.*

Proof. Since A_n is normal and stable, A_n can be written as

$$A_n = Q^*D_nQ,$$

where $Q \in C^{n\times n}$ is a unitary matrix and

$$D_n = \mathrm{diag}(d_1, d_2, \ldots, d_n).$$

It is obvious that the eigenvalues of

$$c_U(A_n) = U^*\delta(UA_nU^*)U$$

are equal to the diagonal elements of UA_nU^*. For simplicity, we denote UQ^* as

$$UQ^* = [b_1, b_2, \ldots, b_n],$$

where b_i is the ith column of matrix UQ^* and

$$b_i = (\beta_{1i}, \beta_{2i}, \ldots, \beta_{ni})^T$$

for $i = 1, 2, \ldots, n$.

By direct computation, we get

$$\begin{aligned} UA_nU^* &= UQ^*D_nQU^* \\ &= [b_1, b_2, \ldots, b_n]\mathrm{diag}(d_1, d_2, \ldots, d_n) \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_n^* \end{bmatrix} \\ &= d_1b_1b_1^* + d_2b_2b_2^* + \cdots + d_nb_nb_n^*, \end{aligned}$$

where $b_i b_i^* \in C^{n\times n}$ for $i = 1, 2, \ldots, n$. Moreover,

$$\begin{aligned}\delta(UA_nU^*) &= \delta(d_1 b_1 b_1^* + d_2 b_2 b_2^* + \cdots + d_n b_n b_n^*)\\ &= d_1\delta(b_1 b_1^*) + d_2\delta(b_2 b_2^*) + \cdots + d_n\delta(b_n b_n^*).\end{aligned}$$

It is easy to see that

$$\delta(b_i b_i^*) = \begin{bmatrix} \beta_{1i}\overline{\beta_{1i}} & & 0 \\ & \ddots & \\ 0 & & \beta_{ni}\overline{\beta_{ni}} \end{bmatrix}, \qquad i = 1, 2, \ldots, n,$$

i.e., the diagonal elements are nonnegative real numbers.

Since the real part of d_i is negative, for $i = 1, 2, \ldots, n$, and the jth diagonal element of matrices,

$$\delta(b_1 b_1^*), \quad \delta(b_2 b_2^*), \quad \ldots, \quad \delta(b_n b_n^*),$$

cannot be zero at the same time due to nonsingularity of UQ^*, we conclude that the real part of the eigenvalues of $\delta(UA_nU^*)$ is also negative and therefore T. Chan's preconditioner $c_U(A_n)$ is stable. □

We remark that if the real part of the eigenvalues of A_n is positive, then the real part of the eigenvalues of $c_U(A_n)$ is also positive.

Acknowledgment. X. Jin and Y. Wei would like to thank Prof. Raymond Chan for his hospitality when they visited the Chinese University of Hong Kong in January, 2003. Part of this work was finished during their stay there.

REFERENCES

[1] R. Chan, X. Jin, and M. Yeung, *The circulant operator in the Banach algebra of matrices*, Linear Algebra Appl., 149 (1991), pp. 41–53.

[2] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[3] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[4] B. N. Datta, *Applied and Computational Control, Signals and Circuits: Recent Developments*, Kluwer Academic Publishers, Boston, 2001.

[5] P. Davis, *Circulant Matrices*, John Wiley & Sons, New York, Chichester, Brisbane, 1979.

[6] T. Huckle, *Circulant and skewcirculant matrices for solving Toeplitz matrix problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 767–777.

[7] X.-Q. Jin, *Developments and Applications of Block Toeplitz Iterative Solvers*, Comb. Comput. Sci. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[8] D. D. Siljak, *Large-Scale Dynamic Systems. Stability and Structure*, North–Holland Ser. System Sci. Engrg. 3, North–Holland, New York, Amsterdam, 1978.

[9] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 630–641

DOUBLE ORDERING AND FILL-IN FOR THE *LU* FACTORIZATION*

MARKUS BAUMANN†, PETER FLEISCHMANN†, AND OTTO MUTZBAUER†

Abstract. We present a new method, called *reversed double ordering*, for reordering arbitrary matrices prior to *LU* factorization. This reordering creates a variable-band matrix. We compare the fill-in of the *LU* factorization for sparse matrices with respect to reversed double ordering, column minimum degree ordering, and the reversed Cuthill–McKee algorithm. Moreover, we combine the first two reorderings with good success.

Key words. double ordering, minimum degree ordering, Cuthill–McKee algorithm, *LU* factorization, sparse matrices

AMS subject classifications. 65F50, 65F05

DOI. 10.1137/S0895479802392989

1. Introduction. We introduce a new reordering, called *reversed double ordering*, for arbitrary matrices and apply it prior to *LU* factorization. We show that this reordering reduces fill-in. In particular, for sparse systems of linear equations this proves to be advantageous. Moreover, we compare the reduction in fill-in using reversed double ordering, column minimum degree ordering, and the reversed Cuthill–McKee algorithm for some application-related matrices from Davis [2], and for larger random matrices. Note that column minimum degree ordering is the same as a symmetric minimum degree ordering for the matrix A^TA (cf. [4]). For *LU* factorization, column minimum degree ordering, and the reversed Cuthill–McKee algorithm, we use MATLAB [6]. Our code for reversed double ordering is a naive implementation in C (cf. section 6). Finally we show that the combination of reversed double ordering and column minimum degree ordering often reduces fill-in more than column minimum degree ordering alone or the reversed Cuthill–McKee algorithm. The choice of this combination was motivated by the fact that the Cuthill–McKee algorithm can be treated as a special case of double ordering. Moreover, we apply double ordering first, because double ordering may destroy the original intention of the minimum degree algorithm, namely, to reduce fill-in.

Double ordering was introduced by Mader and Mutzbauer [5]. The nonzero pattern of a matrix can be considered as a $(0,1)$-matrix if we agree that entries not equal to zero are represented by 1. In [5, Theorem 2] it is shown that $(0,1)$-matrices can be reordered such that the rows and columns, considered as binary representations on natural numbers, are ordered simultaneously downward and from left to right. This is called *double ordering*. A general matrix is called *doubly ordered* if its nonzero pattern is doubly ordered. Algebraically, an $(m \times n)$-matrix A is *permutation equivalent* to a doubly ordered matrix $B = PAQ$, where P and Q are suitable permutation matrices.

Double ordering results in a variable-band nonzero pattern. This was our motivation for applying double ordering as a reordering prior to *LU* factorization, since a variable-band pattern promises less fill-in. Clearly, double ordering destroys the symmetry of the nonzero pattern of the matrix. For matrices which are already in

*Received by the editors May 16, 2002; accepted for publication (in revised form) by Z. Strakoš April 10, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/simax/25-3/39298.html
†Mathematisches Institut, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany (msbaumann@web.de, fleischmann@mathematik.uni-wuerzburg.de, mutzbauer@mathematik.uni-wuerzburg.de).

variable-band form, the bandwidth after double ordering may not be smaller. Thus the application of double ordering makes sense mostly for matrices without a specific nonzero pattern. We visualize the effect of double ordering with random matrices (cf. Figures 2 and 3).

There is an obvious link to the Cuthill–McKee algorithm [3], since the Cuthill–McKee algorithm is a permutation similarity, i.e., $B = PAP^{-1}$. In particular, it maintains the symmetry of matrices. This is obviously not the case for double ordering. In this sense the Cuthill–McKee algorithm is a natural restriction of double ordering to matrices with a symmetric nonzero pattern. Both these reorderings lead to variable-band matrices. Usually the Cuthill–McKee algorithm is applied together with reversion (cf. "rcm" in MATLAB). This reversion significantly reduces fill-in. For the same reason we use only reversed double ordering prior to LU factorization.

The new Davis package UMFPACK for MATLAB comes with an extraordinarily good column permutation; it is an asymmetric multifrontal method, which seems to be better than column minimum degree ordering and reversed double ordering in their present form.

2. Double ordering. First we describe the double ordering of $(0,1)$-matrices. A *line* of a matrix denotes either a row or a column. The set of $(0,1)$-rows allows a lexicographic ordering induced by $1 > 0$ from the left. Analogously $(0,1)$-columns allow a lexicographic ordering induced by $1 > 0$ from the top. This is equivalent to considering each of the rows of a $(0,1)$-matrix as the dyadic representation of a natural number. Then the lexicographic order of rows is just the usual linear order on natural numbers, and an analogous statement is true for the lexicographic order of the columns. With this in mind we can talk about greater and smaller lines. A matrix is said to be *doubly ordered* if the set of the rows from top to bottom and the set of the columns from left to right simultaneously form descending sequences.

Every $(0,1)$-matrix allows line permutations such that a double ordering is obtained. The proof of this fact is given in the form of an explicit and properly working algorithm. Define the *degree of order* $\mathrm{dgo}(M)$ of a $(0,1)$-matrix $M = [m_{ij}]$ of size m by n by setting

$$\mathrm{dgo}(M) = \sum_{(i,j)} m_{ij} 2^{m+n-i-j}.$$

The degree of order is a weight function such that entries of a matrix more to the top and/or more to the left get higher weight. Thus, moving a greater row to the top or moving a greater column to the left increases the degree of order. This idea proves the following statement.

LEMMA 2.1 (see [5, Lemma 1]). *The degree of order of a $(0,1)$-matrix increases under transpositions of lines if either a greater row is permuted toward the top or a greater column is permuted toward the left.*

Now it is straightforward to show that all $(0,1)$-matrices can be doubly ordered.

THEOREM 2.2 (see [5, Theorem 2]). *Every $(0,1)$-matrix is permutation equivalent to a doubly ordered matrix. The doubly ordered matrix is obtained by interchangeably sorting rows and columns.*

Proof. By interchangeably sorting rows and columns, the degree of order increases by Lemma 2.1 until the matrix is doubly ordered. This must happen in finitely many steps since the degree of order is bounded. □

Remark. The above proof with interchangeably sorting rows and columns is a simple and properly working algorithm. There is another algorithm [5, Theorem 6]

doubly ordering a $(0,1)$-matrix that works line by line. Double ordering of a matrix does not have a unique result; in other words, there is a certain freedom in double ordering a matrix.

In [5, Theorem 8] it is shown that double ordering of a matrix always displays the finest block diagonal structure that is obtainable for this matrix by permutations of lines. However, double ordering does more than merely determine the diagonal block structure.

3. Structure of doubly ordered matrices: Numerical aspects. In this section we want to analyze and visualize the structure of doubly ordered matrices. Instead of using the term "dyadic representation," it is preferable to use the lexicographic ordering of the rows and the columns as $(0,1)$-vectors, since this simplifies the discussion of the pattern of matrices after double ordering. For instance, for a doubly ordered $(0,1)$-matrix, the columns decrease to the right; hence the 1's retreat from the top. More precisely, the first row must be of the form $(1,\dots,\overset{i}{1},0,\dots,0)$, where the last 1 is in the ith column, and $(*,\dots,*,\overset{i+1}{1},\dots,1,0,\dots,0)$ must be the form of the second row, where "$*$" denotes any entry and the connected 1's start in the $(i+1)$th column, and so on. In particular, the 1's form 1-bars at the margin of the nonzero pattern. Since the rows also decrease downward, the same arguments apply for the columns; the 1's retreat from the left and form 1-bars at the margin. Thus the consequence of double ordering is a variable-band matrix, or as we say, the pattern of a doubly ordered matrix has "leaf" form, with a concentration of entries at the margin of the leaf (cf. the typical doubly ordered matrix below),

$$\begin{bmatrix} 1&1&1&0&0&0&0&0&0&0\\ 1&0&0&1&1&0&0&0&0&0\\ 0&1&0&1&0&1&1&0&0&0\\ 0&1&0&0&0&0&0&1&0&0\\ 0&0&1&0&1&1&0&0&1&1\\ 0&0&0&1&0&1&0&0&1&0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \bullet&\bullet&\bullet&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\ \bullet&\cdot&\cdot&\bullet&\bullet&\cdot&\cdot&\cdot&\cdot&\cdot\\ \cdot&\bullet&\cdot&\bullet&\cdot&\bullet&\bullet&\cdot&\cdot&\cdot\\ \cdot&\bullet&\cdot&\cdot&\cdot&\cdot&\cdot&\bullet&\cdot&\cdot\\ \cdot&\cdot&\bullet&\cdot&\bullet&\bullet&\cdot&\cdot&\bullet&\bullet\\ \cdot&\cdot&\cdot&\bullet&\cdot&\bullet&\cdot&\cdot&\bullet&\cdot \end{bmatrix},$$

which is more suggestively displayed by replacing the 1's by bullets. For the visualization of larger random matrices, doubly ordered, compare Figures 2 and 3.

We find the *occupation degree*, i.e., the average number of entries on a row, most suitable to characterize random matrices in the context of fill-in. For random matrices the shape and the width of the leaf depend on this occupation degree. There is another special property of the shape of doubly ordered matrices. Since double ordering displays the finest possible diagonal block structure of a matrix that can be achieved by line permutations, and since double ordering is more likely to move longer rows, i.e., those with more nonzero entries, to the top, a doubly ordered random matrix of low occupation degree always has a leaf form ending with a tail. This tail is formed by small and very small diagonal blocks, and, in particular, most of the row singletons, which form (1×1)-blocks, are sorted toward the lower right corner (cf. Figure 2). Note that doubly ordered matrices have all zero columns at the right and all zero rows at the bottom.

To clarify heuristically the effect of double ordering, we picture how the nonzero pattern of a random matrix is transformed by double ordering. Recall that a matrix has a nonzero pattern, which, written as a $(0,1)$-matrix, can be viewed as a black and white picture; this is the so-called *density map*. Figure 1 displays the density

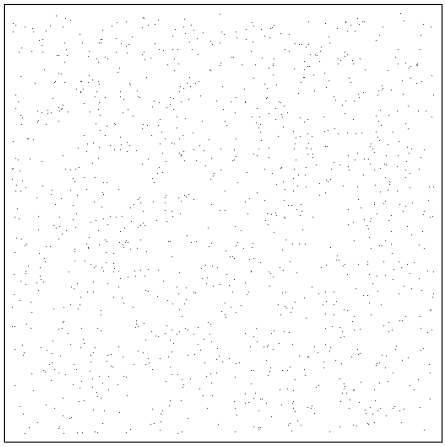


FIG. 1. 500×500 *random matrix with* 1,250 *entries, i.e., of occupation degree* 2.5.

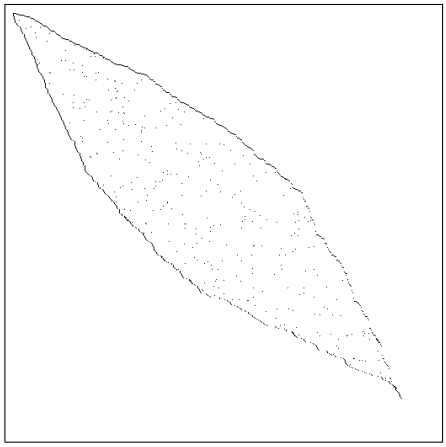


FIG. 2. *This random matrix doubly ordered.*

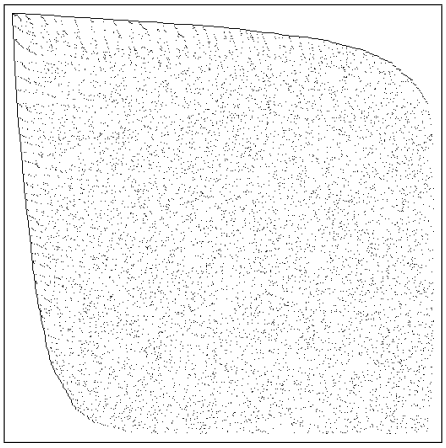


FIG. 3. 500×500 *random matrix with* 10,000 *entries, i.e., of occupation degree* 20, *doubly ordered.*

map of a square random matrix of size 500 with 1,250 entries, which consequently has occupation degree 2.5. Figures 2 and 3 show the nonzero patterns, similar to leaves, that are typical for doubly ordered matrices. Observe that the density inside of a leaf is about the same as the original density and that the nonzero entries are concentrated only at the margin of the leaf. This effect is a straightforward consequence of double ordering.

If the occupation degree of a random matrix is higher, for instance 20, as in Figure 3, double ordering will not result in a small bandwidth, and reversed double ordering prior to LU factorization may not be advantageous.

Double ordering of random matrices has many features, all of which are numerically relevant.

(1) Arbitrary matrices can be doubly ordered.

(2) The resulting matrices have variable-band form.
(3) Double ordering destroys the symmetry of the nonzero pattern.
(4) Longer rows (with more nonzero entries) are more likely to be sorted to the top and longer columns to the left; in other words, nodes of higher degree are likely to be on top (cf. Figure 3).
(5) Double ordering of a matrix does not lead to a unique matrix.
(6) Doubly ordered matrices display the finest possible block diagonal decomposition [5, Theorem 8].
(7) The occupation degree of a random matrix controls the bandwidth of its doubly ordered forms (cf. section 6).
(8) Double ordering does not relate to the values of the nonzero entries.

(1) and (2) indicate that double ordering promises advantages for LU factorization. Double ordering is a way to equip matrices with some structure.

(3) makes clear that double ordering is not necessarily advantageous for symmetric matrices or, more precisely, for matrices with a symmetric nonzero pattern. The reversed Cuthill–McKee algorithm [3], for instance, may be better and cheaper for symmetric matrices.

(4) indicates that reversing the matrix after double ordering is advisable prior to LU factorization.

(5) shows the internal freedom in obtaining a doubly ordered matrix. This can be used to modify double ordering in order to get a nonzero pattern which is numerically better. For instance we can think of some line permutations prior to double ordering: for example, first permute the longest rows to the top; then use column permutations to increase these longer rows lexicographically, i.e., move nonzero elements to the left. If, after this treatment, an algorithm is started which doubly orders the matrix, these longer rows will likely stay on top of the matrix. Now, if the matrix is reversed, the shortest rows are on top. This reduces the fill-in of a later LU factorization. Moreover, instead of strictly double ordering a matrix, some "almost double ordering" is interesting in view, for instance, of getting a smaller bandwidth, minimizing the fill-in, or simply, producing faster codes.

A block diagonal decomposition (6), if there is one, is always helpful. However, there are cheaper algorithms to decompose a matrix. Random matrices of higher occupation degree are in general not decomposable (cf. Figure 3 and section 6).

The occupation degree of a random matrix (cf. (7)) allows an estimate of the bandwidth of the doubly ordered nonzero pattern. Thus, for random matrices with higher occupation degree, the effect of double ordering is not so significant (cf. Figure 3) and possibly the advantages for the LU factorization are not that remarkable.

By (8), pivoting strategies, which do not permute severely the lines of a matrix, allow a reasonable combination with double ordering.

4. LU factorization and fill-in. We show that reversed double ordering applied prior to LU factorization diminishes fill-in. The effect of the LU factorization is shown by the density maps of the generated matrices. We use partial pivoting, i.e., the maximal pivot in a column is always chosen. Clearly, threshold pivoting strategies to maintain sparsity are preferable and may be sometimes unavoidable (cf. section 6).

The original matrix of size 500 has 1,250 entries, i.e., its occupation degree is 2.5. Triangularizing this matrix with partial pivoting leads to about 7,300 entries for U. If the matrix is first doubly ordered and then triangularized, the final number of entries in U is even greater than for the original matrix, namely, about 13,800. This

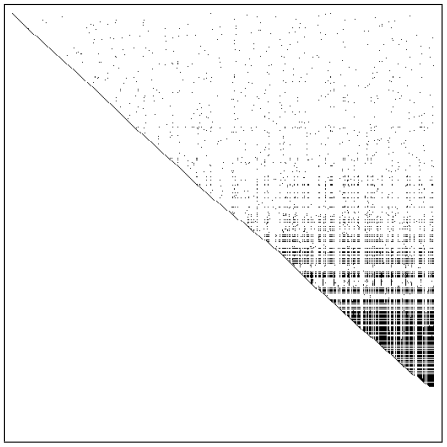


FIG. 4. 500×500 *random matrix with* 1,250 *entries, i.e., of occupation degree* 2.5*, triangularized with partial pivoting.*

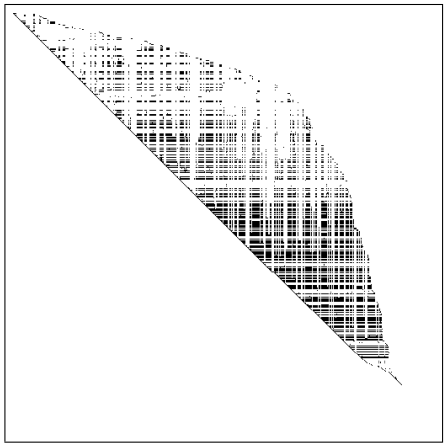


FIG. 5. 500×500 *random matrix with* 1,250 *entries, i.e., of occupation degree* 2.5*, triangularized with partial pivoting after double ordering.*

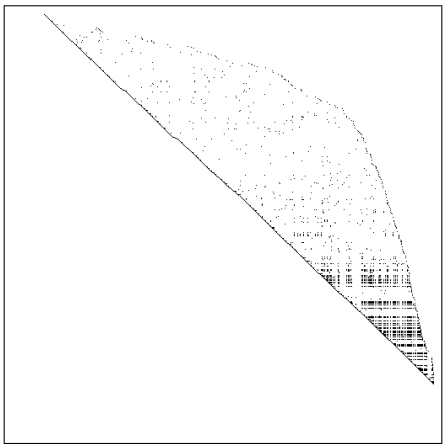


FIG. 6. 500×500 *random matrix with* 1,250 *entries, i.e., of occupation degree* 2.5*, triangularized with partial pivoting after* reversed *double ordering.*

disadvantage is caused by point (4) in section 3. The LU factorization is done row by row; thus, in spite of some row permutations due to partial pivoting, the leaf form is roughly kept. The number of entries in U after reversed double ordering is only 2,650; Figures 4 to 6 clarify the situation. Fill-in is significantly diminished by reversed double ordering.

5. Results. We consider only some of the larger matrices in the UF Sparse Matrix Collection [2], and we use MATLAB to compare fill-in after the reversed Cuthill–McKee algorithm, column minimum degree ordering, and reversed double ordering. Moreover, we combine reversed double ordering and column minimum degree ordering in that order. We provide two tables, Table 1 for matrices with a symmetric pattern,

TABLE 1
Symmetric matrices, occupation degree of L, U, and L+U after LU factorization under different reorderings and partial pivoting.

	Reordering method	Occupation degree			Time
		L	U	L+U	
`bcsstk14`	RCM	101.19	175.55	276.74	2.59
1806×1806	CMD	87.90	138.91	226.82	2.09
Occupation:	RDO	103.59	172.96	276.55	2.75
35.13	RDO+CMD	86.60	134.21	220.81	1.95
`bcsstk26`	RCM	76.06	110.79	186.86	1.58
1922×1922	CMD	44.28	72.90	117.19	0.80
Occupation:	RDO	80.94	116.62	197.57	1.81
15.78	RDO+CMD	47.37	76.68	124.06	0.92

and Table 2 for matrices with a definitely asymmetric pattern. By the arguments in the introduction, we expect significantly different results comparing the Cuthill–McKee algorithm and the reversed double ordering for matrices with an asymmetric pattern (cf. Table 2 for matrices with an asymmetric pattern).

We compare the occupation degrees of L, U, and $L+U$ after LU factorization using the reversed Cuthill–McKee algorithm (RCM), column minimum degree ordering (CMD), reversed double ordering (RDO), and a combination of column minimum degree ordering after reversed double ordering (RDO+CMD). Column minimum degree ordering and reversed double ordering are quite comparable; the combination of both is sometimes better. Moreover, we list the CPU time (cf. "cputime" in MATLAB) for the LU factorization, not including the time for the respective reordering.

6. Implementation and experiments. We use MATLAB [6] for LU factorization, the reversed Cuthill–McKee algorithm "rcm," and column minimum degree ordering "colmmd" [4].

The proof of Theorem 2.2 is an algorithm which doubly orders matrices. Namely, first the rows, considered as $(0,1)$-vectors, are ordered lexicographically downward; second, the columns are ordered to the right; third, the rows are ordered again, and so on, interchangeably. The output of this algorithmic double ordering is two permutations, one for the rows and one for the columns.

We implemented this algorithm for sparse matrices as a C program using only standard libraries. Peter Fleischmann wrote the first code for double ordering of $(0,1)$-matrices in 1999. All our implementations are on a PC under Linux. It is easy to check the code, since the density map of a doubly ordered matrix is characteristic. Our naive implementation of double ordering is not optimized. We used the algorithm from the proof of Theorem 2.2 and sorted the rows and columns with "quicksort." However, this is expensive. For instance, it is enough to sort only some of the greater rows and after that some of the greater columns and so on. This may be an improvement by the factor $\ln n$, where n is the size of the matrix. The proof of Theorem 2.2 still works, since it depends only on the increase of the degree of order. Moreover, we used the double pointer technique to store the sparse matrix. This is probably not the optimal data structure, since, for permutations of lines, the administration of the pointers is expensive. Our double ordering of a random matrix of size 50,000 of occupation degree 2 takes about 40 seconds. The time for the reversed Cuthill–McKee algorithm and for the column minimum degree algorithm is less than 4 seconds. In particular, reversed double ordering with our code very often takes more time than LU factorization with MATLAB. A fast algorithm for double ordering will be the

TABLE 2
Matrices with asymmetric patterns, occupation degree of L, U, and $L+U$ after LU factorization under different reorderings and partial pivoting.

	Reordering method	Occupation degree			Time
		L	U	L+U	
`gemat11`	RCM	152.40	170.09	322.49	23.49
4929×4929	CMD	7.33	11.67	19.01	0.14
Occupation:	RDO	22.93	23.80	46.74	0.46
5.87	RDO+CMD	6.99	11.37	18.37	0.14
`lhr04c`	RCM	143.83	148.38	292.21	10.48
4101×4101	CMD	33.69	49.88	83.57	0.71
Occupation:	RDO	33.86	46.76	80.62	0.55
20.16	RDO+CMD	33.94	50.69	84.63	0.70
`lhr07c`	RCM	201.65	212.44	414.09	42.05
7337×7337	CMD	34.75	53.35	88.10	1.34
Occupation:	RDO	31.98	51.04	83.02	1.01
21.33	RDO+CMD	36.18	56.53	92.72	1.48
`lhr10c`	RCM	210.07	191.47	401.55	49.20
10672×10672	CMD	35.01	53.98	89.00	1.97
Occupation:	RDO	30.32	44.29	74.62	1.29
21.79	RDO+CMD	35.17	56.29	91.47	2.02
`lhr11c`	RCM	238.82	258.48	497.31	99.74
10964×10964	CMD	40.44	57.83	98.28	2.72
Occupation:	RDO	52.23	66.51	118.74	3.21
21.31	RDO+CMD	39.36	58.95	98.32	2.77
`lhr14c`	RCM	256.36	253.41	509.78	117.65
14270×14270	CMD	41.75	58.66	100.41	3.58
Occupation:	RDO	50.75	62.49	113.25	3.43
21.57	RDO+CMD	42.77	60.95	103.72	3.89
`lhr17c`	RCM	392.00	435.55	827.55	444.69
17576×17576	CMD	41.61	55.90	97.52	4.23
Occupation:	RDO	49.91	61.33	111.25	4.33
21.73	RDO+CMD	42.44	58.91	101.36	4.79
`lhr34`	RCM	161.73	180.82	342.55	243.73
35152×35152	CMD	27.97	51.00	78.97	8.25
Occupation:	RDO	41.98	55.50	97.49	8.81
21.24	RDO+CMD	28.47	50.95	79.43	8.86

subject of future work. However, we doubly ordered a random matrix of size $320,000$ with $960,000$ entries, i.e., of occupation degree 3, in less than 16 minutes [7] (cf. also [1]). With the choice of this larger matrix we want to show that double ordering is still possible even if the LU factorization for a random matrix of this size is beyond the scope of our hardware.

In section 5 we work mostly with application-related matrices. However, we find the comparison of reorderings interesting, as an experiment; thus we used random matrices, as an objective input. We take square random matrices with nonzero diagonal and entries between 0 and 1 of size 50,000. The nonzero diagonal is needed for MATLAB's LU factorization to work properly. The occupation degree of the random matrices is between 2.0 and 2.4. As above, we list the occupation degrees of L and U and the time in seconds for the LU factorization using the different reorderings and different threshold pivoting. In particular, the time does not include the time for the respective reordering.

The results for random matrices (cf. Table 3) show that the Cuthill–McKee algorithm gets significantly worse than the other reorderings, whereas column minimum

TABLE 3
Occupation degree of L, U, and $L+U$ after LU factorization of square random matrices of size 50,000.

	Reordering method	Occupation degree			Time
		L	U	L+U	
50RND20	RCM	25.70	21.18	46.89	27.91
50000 × 50000	CMD	4.21	5.19	9.40	3.19
Occupation:	RDO	8.28	6.99	15.28	4.00
2.00	RDO+CMD	3.77	4.76	8.54	2.58
50RND21	RCM	85.20	74.23	159.43	193.12
50000 × 50000	CMD	16.73	19.62	36.36	34.17
Occupation:	RDO	33.37	27.21	60.59	39.29
2.10	RDO+CMD	13.97	17.76	31.74	26.73
50RND22	RCM	199.04	184.18	383.22	747.98
50000 × 50000	CMD	43.44	50.37	93.81	156.92
Occupation:	RDO	90.17	77.02	167.20	209.60
2.20	RDO+CMD	43.87	51.61	95.49	165.12
50RND23	RCM	350.47	342.33	692.80	1906.18
50000 × 50000	CMD	89.56	105.41	194.97	431.18
Occupation:	RDO	170.53	152.17	322.71	551.05
2.30	RDO+CMD	92.02	109.68	201.70	480.72
50RND24	RCM	515.46	500.73	1016.19	3711.22
50000 × 50000	CMD	172.76	199.17	371.93	1391.70
Occupation:	RDO	309.54	287.85	597.39	1468.50
2.40	RDO+CMD	165.46	194.22	359.69	1294.26

degree ordering is slightly better than reversed double ordering. Column minimum degree ordering is designed for reducing fill-in step by step following the pattern of the *LU* factorization. Reversed double ordering has the decrease of the fill-in as a side effect. Thus it is not surprising that reversed double ordering is worse. The combination of both of these reorderings is sometimes better.

The results for random matrices allow a kind of a global comparison between these different reorderings. However, the importance of this should not be overestimated, since these experiments are clearly irrelevant for application purposes.

We tried some random matrices and found that the bandwidth after double ordering is significantly smaller than the bandwidth obtained by the Cuthill–McKee algorithm. This is not surprising, since double ordering is the more general reordering and it also displays the finest diagonal block decomposition (cf. section 3).

We found that square random matrices of occupation degree 2.5 have, after double ordering, an indecomposable first block of about 80% of the rows and several very small blocks, which form a "tail." For occupation degree higher than 3.5, random matrices seem to be indecomposable in general. These observations were independent of the size of the matrices.

For demonstration purposes we include the density pattern of the matrix "lhr01" (Light hydrocarbon recovery problems, from J. Mallya and Mark Stadtherr, Univ. of Illinois) under different reorderings (see Figures 7–10).

We used threshold pivoting in MATLAB, but we found no acceleration by this pivoting strategy. The MATLAB manual specifies, "The sparse LU factorization does not pivot for sparsity, but it does pivot for numerical stability." Nevertheless we did experiments with threshold pivoting for sparsity, and we found that this is unavoidable for the performance of *LU* factorization. However, we decided not to include this topic, since our implementation is naive and threshold pivoting without a discussion

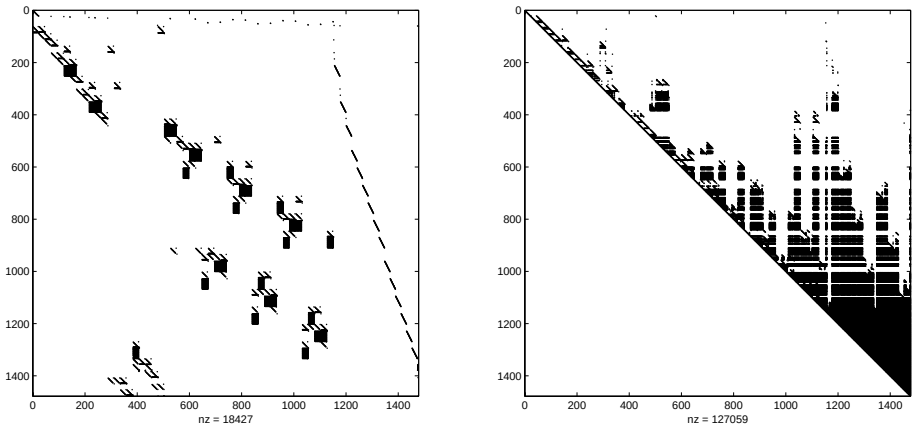


FIG. 7. *lhr*01 *and U without reordering.*

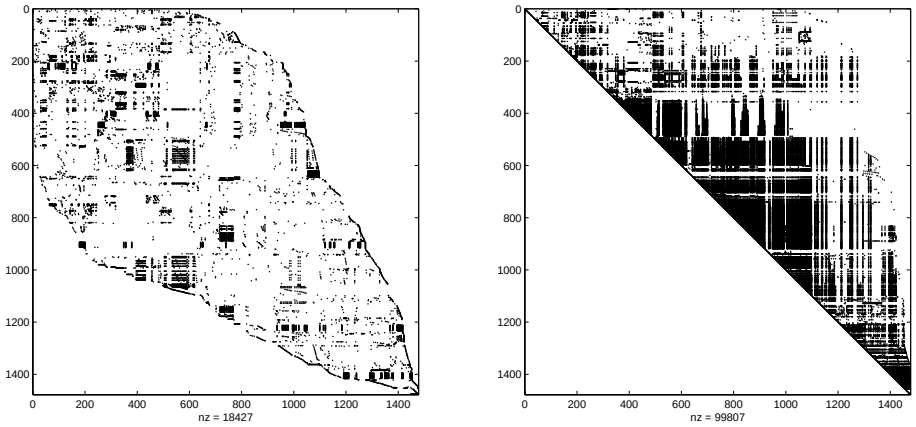


FIG. 8. *lhr*01 *and U after reversed Cuthill–McKee.*

on stability does not make sense. The new Davis package UMFPACK for MATLAB comes with an extraordinarily good column permutation, an asymmetric multifrontal method, and a new threshold strategy pivoting for sparsity. Using this new feature makes all reorderings discussed here obsolete, at least in the present setting.

7. Discussion and conclusion. For larger random matrices, the fill-in caused by *LU* factorization is very sensitive to the occupation degree, as is shown in Table 3. For a fixed size there is always a kind of critical value of the occupation degree beyond which fill-in explodes. This critical value depends on the reordering algorithm. We use partial pivoting. Pivoting strategies, which do not severely destroy the effect of double ordering, promise success in combination with reversed double ordering.

Double ordering creates a variable-band matrix; thus, if applied to a matrix which is already in variable-band form, the effect of double ordering may even be negative. Moreover, double ordering destroys the symmetry of the nonzero pattern of a matrix; thus it is best applied to matrices without a specific structure.

The decrease of the fill-in by reversed double ordering is comparable to that for column minimum degree ordering and, for matrices with a symmetric pattern, with

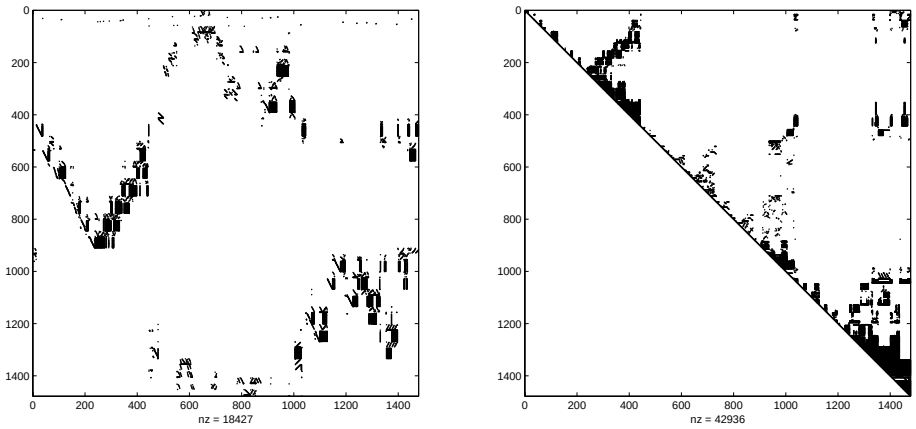


FIG. 9. *lhr*01 *and U after column minimum degree ordering.*

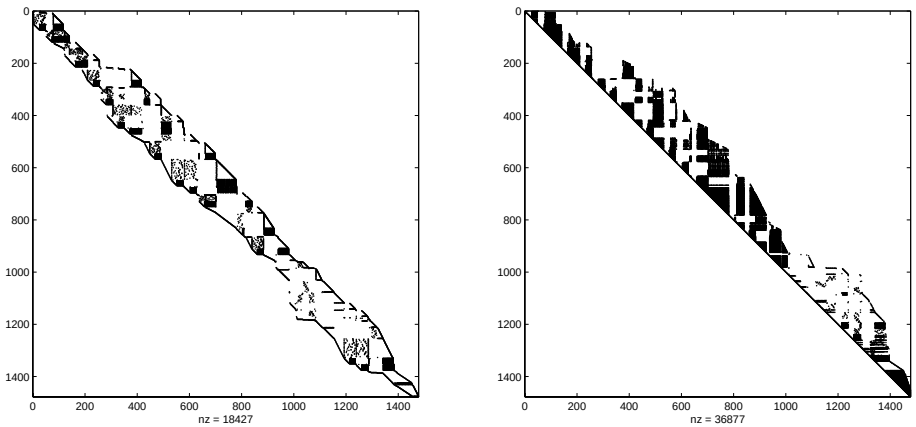


FIG. 10. *lhr*01 *and U after reversed double ordering.*

the Cuthill–McKee algorithm, all in the case of the indicated Davis matrices. Also, for random matrices of size 50,000, double ordering and column minimum degree are comparable. However, the Cuthill–McKee algorithm is obviously worse. The effect of applying both techniques to these random matrices, namely, first our implementation of reversed double ordering and then column minimum degree ordering "colmmd" in MATLAB, seems to be an improvement.

Double ordering is not a fixed algorithm. Not even the doubly ordered final form of a matrix is unique. There is a degree of freedom that can be used to obtain numerically better results. It might even be desirable to get only an almost doubly ordered matrix which behaves numerically better or allows a faster code. Another idea is a combination of both algorithms, reversed double ordering and column minimum degree, in one code.

As a final remark we would like to emphasize that the asymmetric multifrontal method implemented in the new package UMFPACK by Davis has an excellent column ordering algorithm which is significantly better than all reorderings considered in this paper, at least in the present setting.

Acknowledgment. We cordially thank the referees for their intensive help.

REFERENCES

[1] M. Baumann, *Lineare Gleichungssysteme mit dünn besetzten Matrizen*, Diplomarbeit, Math. Institut, Universität Würzburg, Würzburg, Germany, 2000.
[2] T. Davis, *University of Florida Sparse Matrix Collection*, http://www.cise.ufl.edu/research/sparse/matrices, ftp://ftp.cise.ufl.edu/pub/faculty/davis/matrices (1994, 1996, 1997).
[3] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Clarendon Press, Oxford, UK, 1986.
[4] 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.
[5] A. Mader and O. Mutzbauer, *Double ordering of* $(0,1)$*-matrices*, Ars Combin., 61 (2001), pp. 81–95.
[6] MATLAB, Version 6.5.0.180913a, Release 13.
[7] F. Schmitt, *Doppelordnung mit Doppelzeiger-Technik und Anwendungen*, Diplomarbeit, Math. Institut, Universität Würzburg, Würzburg, Germany, 2001.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 642–661

AN IMPROVED DISCRETE WAVELET TRANSFORM PRECONDITIONER FOR DENSE MATRIX PROBLEMS*

JUDITH M. FORD†

Abstract. We present a new preconditioning method for a class of dense matrices, based on sparse approximation in a discrete wavelet basis. We first prove theoretical results that enable us to reduce the cost of an existing wavelet-based preconditioner and then incorporate those ideas into the design of our new preconditioner. We demonstrate the effectiveness of our methods with numerical examples drawn from one-dimensional physical problems and indicate how the method could be incorporated into a Kronecker product–based strategy for solving problems in higher dimensions. A notable feature of our new method is that it enables an optimal wavelet level to be chosen automatically, making it more robust than previous wavelet preconditioners that depend on the user choosing an appropriate level. Thus we are now closer to the ultimate goal of a black-box preconditioner for dense matrices.

Key words. wavelet-based preconditioner, discrete wavelet transform, dense matrix with diagonal singularity

AMS subject classifications. 65F10, 65N22, 65T60

DOI. 10.1137/S0895479802416526

1. Introduction.

We are concerned with the fast solution of large, dense linear systems

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

where the (often nonsymmetric) $n \times n$ matrix A is smooth apart from a diagonal singularity. By "smooth" we mean that the rows and columns of the matrix, viewed as the values of a function on a uniform grid, have small divided differences. In other words, the values of adjacent entries differ by only relatively small amounts compared with their magnitudes. For each element, $a_{i,j}$, of A we define the level 1 *difference vector*, $d^1_{i,j}$, whose two components are the differences $a_{\langle i+1\rangle_n,j} - a_{i,j}$ and $a_{i,\langle j+1\rangle_n} - a_{i,j}$. Here the notation $\langle p\rangle_q$, $p, q \in \mathbb{N}$, denotes the remainder when p is divided by q. Higher order difference vectors are defined recursively in the obvious way: $d^{k+1}_{i,j} = (d^k_{\langle i+1\rangle_n,j}(1) - d^k_{i,j}(1), d^k_{i,\langle j+1\rangle_n}(2) - d^k_{i,j}(2))$. A matrix S is considered to have level ℓ smoothness if $\|d^\ell_{i,j}\|_\infty < \epsilon_\ell$ for some constant ϵ_ℓ, which is small compared with the magnitude of a typical entry of S. Here we will be concerned with matrices with level ℓ smoothness for small values of ℓ, and we then say that the matrix is "smooth." If a matrix A can be expressed as the sum of a smooth matrix S and a band matrix B of bandwidth w, then we say that A is smooth apart from a diagonal singularity of width w.

Dense matrices with diagonal singularities arise during numerical solution of a variety of problems, including Cauchy singular integral equations (see, e.g., [22, 26]), which arise in many physical models including elasticity and aerodynamics; boundary integral equations that result when a boundary element method is used to solve

*Received by the editors October 18, 2002; accepted for publication (in revised form) by M. Hauke June 17, 2003; published electronically December 17, 2003. This work was supported by EPSRC Postdoctoral Research Fellowship GR/R95982/01.

http://www.siam.org/journals/simax/25-3/41652.html

†Department of Mathematics, UMIST, P.O. Box 88, Manchester M60 1QD, UK (j.ford@umist.ac.uk).

PDE boundary value problems such as the Helmholtz equation (see, e.g., [2, 7]); and integro-differential equations such as those arising in elastohydrodynamic lubrication problems (see, e.g., [28]).

Solution of such large, dense linear systems by a direct method is prohibitively expensive, so an iterative technique, such as one of the Krylov subspace methods, is usually adopted, and effective preconditioning of the matrix A is required in order to keep the number of iterations small. Our ultimate goal is to produce "black-box" software capable of providing effective preconditioners for dense matrices based on the structure of the matrices themselves rather than relying on detailed knowledge of the underlying problem from which they arise.

One well-established approach is to form a preconditioner $M \approx A$. The (right) preconditioned version of (1.1) is then

$$AM^{-1}y = b, \quad Mx = y. \tag{1.2}$$

Provided that M is a close approximation to A, we expect that our preconditioned matrix $AM^{-1} \approx I$ will give fast convergence of our iterative method, but for the preconditioner to be cost-effective we also require that the matrix-vector multiplication $M^{-1}v$, which is performed at each iteration, be cheap to compute. One way of achieving this is to choose M to be sparse with a sparsity pattern that does not cause much fill-in under LU factorization. When matrix A is "smooth," application of a discrete wavelet transform (DWT) produces a transformed matrix $\tilde{A}$ with a large proportion of very small entries. By setting to zero entries that are deemed to be insignificant, a sparse approximation $\tilde{M}$ to $\tilde{A}$ can be formed. This "wavelet-compression" technique is often used successfully for image processing (see, e.g., [12]), but in the context of preconditioner design not only the *number* of nonzero entries in $\tilde{M}$ is important, but also the sparsity pattern is important, since fill-in during LU factorization may result in a high cost preconditioner even if it is very sparse. Several ways of producing sparse wavelet-based approximations that are suitable for preconditioning have been proposed (see, e.g., [3, 4, 5, 10, 11, 17]), and wavelet transforms have also been used to derive sparse approximate inverse preconditioners for sparse matrices with dense inverses (see, e.g., [6, 8, 9]).

Here we start from the "DWTPer" preconditioner first proposed by Chen [11] (see (3.1) for a formal definition) and further developed in [16, 17, 18, 20] and show how it can be modified to enable a larger proportion of the most significant entries in the transformed matrix to be retained in the preconditioner without a large increase in cost. The resultant preconditioner is more effective than DWTPer for many matrices that respond well to DWTPer preconditioning and is also competitive with other methods for some matrices for which DWTPer gives poor results. This means that our preconditioner is more robust than its competitors in the sense that it performs reliably for a wider range of problems. It also requires less user input (as we shall show in section 3), making it less likely that an inappropriate choice will be made. For example, the performance of a DWTPer band preconditioner depends on the "level" of wavelet transform used, which must be decided in advance by the user, while our new method automatically determines an optimal choice of transform level. Our technique is a purely algebraic one, based on consideration of the properties of the dense matrix A without reference to its origins. For dense matrices arising from operator equations, continuous wavelet theory has been used in the construction of preconditioners (see, for example, [13]). Our approach is distinctively different in that it attempts to use wavelet compression to enable preconditioning of a discrete linear

system without using knowledge derived from any underlying continuous operator. Such knowledge may not always be available, for example, in the case of Jacobian matrices in a Newton iteration (see section 4) or Kronecker factors approximating a multidimensional operator matrix (see section 5). The hierarchical matrices of Hackbusch [25] and the mosaic skeletons of Tyrtyshnikov [30] offer alternative ways of approximating dense matrices, and it would be interesting in future work to compare these methods with wavelet-based approaches.

The structure of the paper is as follows: in section 2 we summarize the ideas behind DWT-based preconditioning and explain the principle upon which our new algorithm is based. We then present, in section 3, some new theoretical results relating to the DWTPer transform and give a detailed presentation of the new algorithm. Section 4 contains numerical examples of the use of our preconditioner to solve both test problems and linear systems arising from real applications. We illustrate the potential of our new algorithm for extension to matrices arising from discretization of multidimensional problems by an example in section 5. Finally, in section 6 we summarize our conclusions and outline plans for future developments.

2. DWT-based preconditioning. Let $v = (s_0^{(0)}, s_1^{(0)}, \dots, s_{n-1}^{(0)})^T$ be a vector of length n. The (periodized) level k DWT of v is then defined by the following recurrence relations:

$$s_j^{(i+1)} = \sum_{\ell=0}^{D-1} h_\ell s^{(i)}_{\langle \ell+2j \rangle_{n/2^i}}, \quad d_j^{(i+1)} = \sum_{\ell=0}^{D-1} g_\ell s^{(i)}_{\langle \ell+2j \rangle_{n/2^i}}. \tag{2.1}$$

The $s_j^{(i+1)}$ represent weighted averages of the elements $s^{(i)}_{\langle \ell+2j \rangle_{n/2^i}}, \ell = 0, \dots, D-1$, with the weights being defined by the "filter coefficients" $h_0, h_1, \dots, h_{D-1}$. The $d_j^{(i+1)}$ are weighted differences of the same elements. Often (for example, in the case of the Daubechies wavelet family) the g_j are related to the h_j by $g_j = (-1)^j h_{D-1-j}$. The filter length, D, is also known as the *order* of the DWT. For a smooth vector, we expect the $d_j^{(i+1)}$ to be small compared with the $s_j^{(i+1)}$. At each level the number of elements to be transformed is halved. The level i transformed vector, $v^{(i)}$, is of the form

$$\left(s_0^{(i)}, \dots, s^{(i)}_{n/2^i-1}, d_0^{(i)}, \dots, d^{(i)}_{n/2^i-1}, d_0^{(i-1)}, \dots, d^{(i-1)}_{n/2^{i-1}-1}, \dots, d_0^{(1)}, \dots, d^{(1)}_{n/2-1}\right)^T.$$

The level $i+1$ transform is obtained by transforming the components

$$s_0^{(i)}, s_1^{(i)}, \dots, s^{(i)}_{n/2^i-1}$$

to give

$$s_0^{(i+1)}, s_1^{(i+1)}, \dots, s^{(i+1)}_{n/2^{(i+1)}-1}, d_0^{(i+1)}, d_1^{(i+1)}, \dots, d^{(i+1)}_{n/2^{(i+1)}-1}.$$

An equivalent definition of the (level k) DWT is

$$\tilde{v} = Wv = W_k W_{k-1} \dots W_1 v, \tag{2.2}$$

where each $n \times n$ matrix W_i, $i = 1, 2, \dots, k$, is such that $v^{(i)} = W_i v^{(i-1)}$.

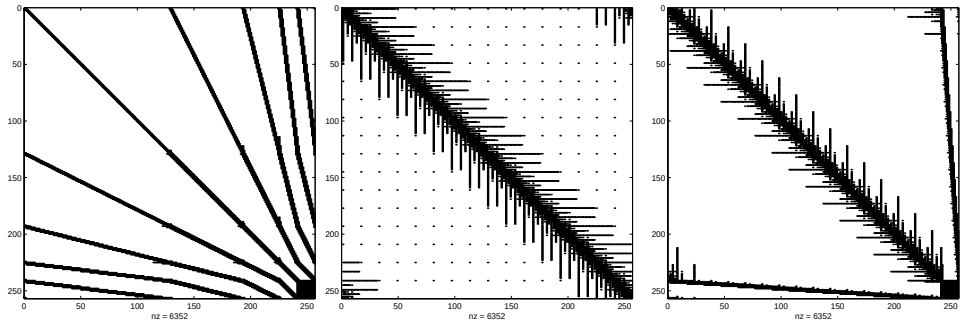


FIG. 2.1. *Sparsity pattern of smooth matrix with nonsmooth diagonal under DWT. Left: standard DWT; center: DWTPer; right: modified DWTPer.*

A matrix can be transformed by applying a DWT to both the rows and the columns:

$$\tilde{A} = WAW^T = W_k W_{k-1} \dots W_1 A W_1^T \dots W_{k-1}^T W_k^T. \tag{2.3}$$

The magnitudes of the $d_j^{(i)}$ depend on the smoothness of the vector v and on the number of *vanishing moments* of the DWT. A DWT with r vanishing moments will give zero $d_j^{(i)}$ entries when applied to a vector v whose components are the values of a polynomial of degree $r-1$ at uniformly spaced points on the real line (equivalently, v's divided differences of level r and above are zero). Thus, we would expect to obtain a matrix with many very small entries when a DWT with r vanishing moments is applied to a matrix with level r smoothness. A Daubechies DWT of order $2r$ has r vanishing moments. In our numerical results we make extensive use of the Daubechies 4 DWT, whose filter coefficients are

$$\frac{1+\sqrt{3}}{4\sqrt{2}}, \quad \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad \frac{1-\sqrt{3}}{4\sqrt{2}}.$$

If A is smooth, the transformed matrix $\tilde{A}$ will have a large proportion of small entries, corresponding to weighted differences between entries that are near-neighbors within the matrix. A sparse approximation can then be formed by setting to zero the smallest of these entries. Nonsmooth features within A give rise to additional large entries in $\tilde{A}$; for a matrix that is smooth apart from along the diagonal, the large entries form a "finger" pattern, such as that shown in the left-hand diagram of Figure 2.1. The dense square at the bottom right of the transformed matrix consists of entries corresponding to weighted average of both rows and columns; the "fingers" correspond to the diagonal singularity. This sparsity pattern is not convenient for preconditioning purposes, because of the large amount of fill-in that occurs under LU factorization. One way of avoiding fill-in is to include in the preconditioner only entries on the diagonal band (i.e., neglecting entries in the outlying "fingers"), and this is sometimes effective, but usually too many significant entries are ignored by this approach, and the resulting preconditioner gives poor convergence.

Another way of avoiding the "finger" pattern is to permute the rows and columns of $\tilde{A}$ so as to bring the large entries associated with the diagonal singularity into a diagonal band. The sparsity pattern of this "DWTPer" transform is shown in the center diagram of Figure 2.1. Here the entries corresponding to weighted differences of entries near the diagonal singularity are contained within a wrap-round diagonal

band, but note that the weighted average entries are now dispersed at regular intervals across the whole matrix rather than being confined to a small square. A typical DWTPer-based preconditioner would be formed by setting to zero all elements outside of a diagonal band. This preconditioner undergoes very little fill-in under LU factorization and so is cheap to apply; particularly for matrices with a pronounced diagonal singularity, this approach has been shown to produce effective preconditioners (see [11, 17, 20]). A bound for the width of the diagonal band of large entries is proved in [11], and this enables a suitable bandwidth for the preconditioner to be chosen, based on the width of the nonsmooth band in A and the order (number of filter coefficients) and level of the DWTPer transform. In section 3 we will show that this bound is not tight and prove results that enable a narrower band to be used, thus reducing the cost of each application of the preconditioner.

Our new preconditioner is based on the idea that, in order to approximate $\tilde{A}$ well, we need to include *both* the large entries corresponding to weighted differences near to the singularity *and* those corresponding to weighted averages of all the entries in A. Band approximation of standard DWT allows all the weighted average entries to be included, but neglects a large proportion of weighted difference entries, and is effective when the diagonal singularity is not very marked; band approximation of DWTPer includes almost all the large weighted difference entries, but neglects most of the weighted average entries, and works well for matrices with a pronounced diagonal singularity. By taking into account both types of large entry, we aim to design a preconditioner that will outperform both methods and will be applicable to smooth matrices with both strong and weak singular features. We cannot include all the large entries from the standard DWT sparsity pattern without an unacceptable increase in cost due to fill-in. The DWTPer sparsity pattern produces much less fill-in, and to improve efficiency further we plan to apply additional permutations to move the large entries into a more convenient pattern. We can do this very simply starting from the DWTPer sparsity pattern, by moving all the rows and columns that contain weighted average entries to the bottom and right-hand edges, respectively. The result is the sparsity pattern shown in the right-hand diagram of Figure 2.1. We can now form a "bordered block" preconditioner by selecting entries from bands along the diagonal and the bottom and right-hand edges. This structure allows us to apply the preconditioner at low cost (see [15, Appendix B]).

3. A modification of the DWTPer preconditioner.

3.1. Determining the bandwidth for a DWTPer-based preconditioner.

Suppose that we wish to form a DWTPer-based band preconditioner for a matrix A that is smooth apart from a wrap-round diagonal band having lower bandwidth α and upper bandwidth β. We can think of A as the sum of a smooth matrix S and a band matrix B. When A is transformed using DWTPer, we obtain $\tilde{A} = W(S + B)W^T = WSW^T + WBW^T$. DWTPer band preconditioning is based on forming an approximation to $\tilde{A}$ by retaining all the entries corresponding to nonzero entries in $\tilde{B} = WBW^T$. In [11] it is proved that for a level k, order D transform, these nonzero entries must all lie within a wrap-round band of lower bandwidth $\leq \alpha + D(2^k - 1)$ and upper bandwidth $\leq \beta + D(2^k - 1)$. We now present some tighter bounds for these bandwidths, which will enable us to reduce the number of nonzero entries in DWTPer band preconditioners with a consequent reduction in both CPU and storage costs. To improve clarity, we ignore the wrap-round of the DWTPer matrices in the proofs of our theoretical results. It is trivial to establish that this does not affect the validity of the results.

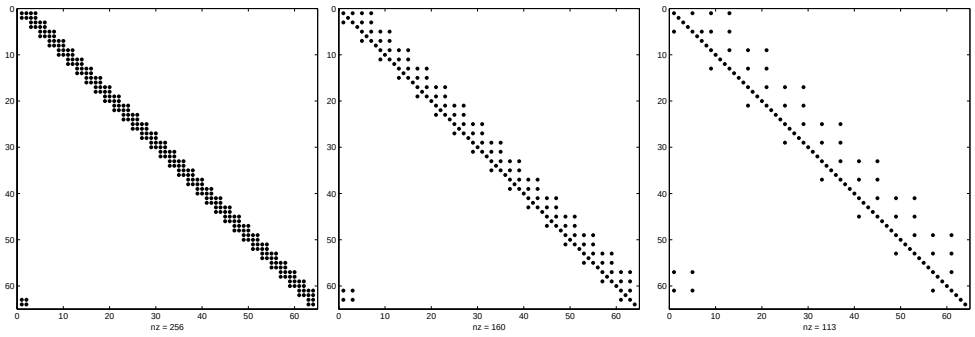


FIG. 3.1. *Sparsity patterns of* $\hat{W}_1, \hat{W}_2, \hat{W}_3$ *for the Daubechies* 4 *DWTPer transform.*

Application of DWTPer (see [11]) is equivalent to replacing the matrices W_i of the standard transform with

$$\hat{W}_i = \begin{pmatrix}
h_0 & 0 & h_1 & 0 & h_2 & 0 & \cdots & & & & & h_{D-1} & 0 & \cdots & & \\
0 & I & 0 & 0 & 0 & 0 & \cdots & & & & & 0 & 0 & \cdots & & \\
g_0 & 0 & g_1 & 0 & g_2 & 0 & \cdots & & & & & g_{D-1} & 0 & \cdots & & \\
0 & 0 & 0 & I & 0 & 0 & \cdots & & & & & & & & & \\
0 & 0 & 0 & 0 & h_0 & 0 & h_1 & 0 & h_2 & 0 & \cdots & & & & h_{D-1} & \cdots \\
\vdots & \vdots & & & 0 & I & 0 & 0 & 0 & 0 & \cdots & & & & & \\
 & & & & g_0 & 0 & g_1 & 0 & g_2 & 0 & \cdots & & & & g_{D-1} & \cdots \\
 & & & & 0 & 0 & 0 & I & 0 & 0 & \cdots & & & & & \\
 & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & & & &
\end{pmatrix}. \tag{3.1}$$

Here the h_j, g_j are the filter coefficients defining the DWT; I is an identity matrix of dimension $2^{i-1}-1$; the 0's are block zero matrices of the appropriate sizes. (For $i=1$, both I and all the 0's are of dimension 0.) So the level k, order D DWTPer of a matrix A is given by

$$\tilde{A} = \mathcal{W}_k A \mathcal{W}_k^T = \hat{W}_k \hat{W}_{k-2} \dots \hat{W}_1 A \hat{W}_1^T \dots \hat{W}_{k-1}^t \hat{W}_k^T. \tag{3.2}$$

We now use this definition of DWTPer to establish bounds on the bandwidth of the DWTPer transform of a band matrix. Before stating our results formally, we will illustrate them by means of examples. Figure 3.1 shows the sparsity patterns of $\hat{W}_1, \hat{W}_2, \hat{W}_3$ for the Daubechies 4 DWTPer. When these matrices are combined to give $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$, the resulting sparsity patterns are those shown in Figure 3.2. Our previous estimate for the bandwidth of $\tilde{B}$ was based on considering $\hat{W}_i$ to be a band matrix with lower bandwidth 2^{i-1} and upper bandwidth $(D-1)2^{i-1}$. This means that $\mathcal{W}_k$ is a band matrix with lower bandwidth 2^k-1 and upper bandwidth $(D-1)(2^k-1)$. However, as we now show, the block structure of the W_i ensures that the lower bandwidth of $\mathcal{W}_k$ is actually only 2^{k-1}.

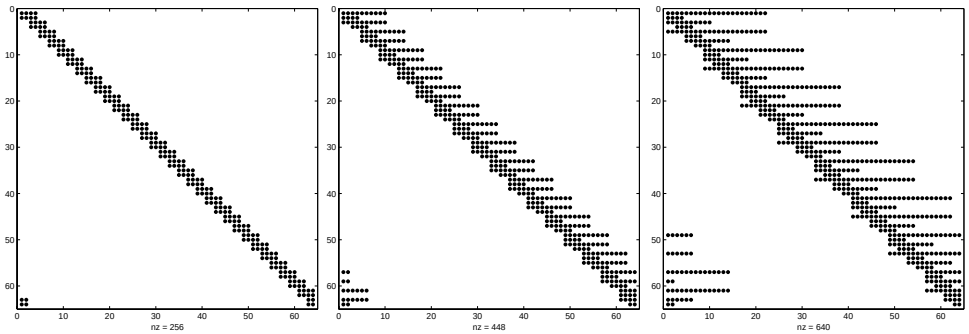


FIG. 3.2. *Sparsity patterns of* $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ *for the Daubechies* 4 *DWTPer transform.*

LEMMA 3.1. *The order* D*, level* k *DWTPer matrix* $\mathcal{W}_k = \hat{W}_k \hat{W}_{k-2} \dots \hat{W}_1 A \hat{W}_1^T \dots \hat{W}_{k-1}^t \hat{W}_k^T$ *is of the form*

$$\begin{pmatrix} B_k & 0 & 0 & \cdots \\ 0 & B_k & 0 & \cdots \\ & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3.3}$$

where each B_k *is a* $2^k \times (D-1)(2^k-1)+1$ *block of the form*

$$\begin{array}{r} \text{row number} \\ 1 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 2^{k-1} \\ 2^{k-1}+1 \\ 2^{k-1}+2 \\ \vdots \\ \vdots \\ 2^{k-1}+2^{k-2}+1 \\ 2^{k-1}+2^{k-2}+2 \\ \vdots \\ 2^k-1 \\ 2^k \end{array} \begin{pmatrix} * & * & \cdots & & & & & & & & & & \cdots & * \\ * & * & \cdots & & & & & & & & & & \cdots & * \\ \vdots & & \ddots & & & & & & & & & & & * \\ \vdots & & & \ddots & & & & & & & & & & * \\ \vdots & & & & \ddots & & & & & & & & & \vdots \\ \vdots & & & & & \ddots & & & & & & & & \vdots \\ * & \cdots & & & & \cdots & * & \cdots & & & & & \cdots & * \\ * & \cdots & & & & & \cdots & * & \cdots & & & & \cdots & * \\ & & & & & & & * & * & \cdots & * & * & \cdots & * \\ & & & & & & & \vdots & & \ddots & & & & \vdots \\ & & & & & & & \vdots & & & \ddots & & & \vdots \\ & & & & & & & * & \cdots & & \cdots & * & \cdots & * \\ & & & & & & & & & & & * & * & * \\ & & & & & & & & & & & \vdots & \ddots & \vdots \\ & & & & & & & & & & & & * & * \\ & & & & & & & & & & & & * & * \end{pmatrix},$$

and the $(1,1)$ *entry of each* B_k *lies on the leading diagonal of* $\mathcal{W}_k$*.*

Proof.

$$\mathcal{W}_1 = \begin{pmatrix} h_0 & h_1 & h_2 & \cdots & & & & h_{D-1} & 0 & \cdots & & \\ g_0 & g_1 & g_2 & \cdots & & & & g_{D-1} & 0 & \cdots & & \\ 0 & 0 & h_0 & h_1 & h_2 & \cdots & & & & & h_{D-1} & \cdots \\ 0 & 0 & g_0 & g_1 & g_2 & \cdots & & & & & g_{D-1} & \cdots \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \end{pmatrix}$$

is clearly of the form (3.3) (with $k=1$). Hence the result is true for $k=1$. Now

suppose that it is true for $k = i - 1$.

$$\mathcal{W}_i = \hat{W}_i \mathcal{W}_{i-1} = \hat{W}_i \begin{pmatrix} B_{i-1} & 0 & 0 & \cdots \\ 0 & B_{i-1} & 0 & \cdots \\ & \ddots & \ddots & \ddots \\ & & & \end{pmatrix}. \tag{3.4}$$

Because of the block structure of the matrices, it is only necessary to consider the first B_i block in $\mathcal{W}_i$. Since columns $1, \ldots, 2^{i-1}$ of $\mathcal{W}_{i-1}$ are zero below row 2^{i-1} (row 2^{i-1} is the bottom row in the first B_{i-1} block in $\mathcal{W}_{i-1}$), columns $1, \ldots, 2^{i-1}$ of $\mathcal{W}_i$ must be linear combinations of columns $1, \ldots, 2^{i-1}$ of $\hat{W}_i$. Hence, columns $1, \ldots, 2^{i-1}$ of $\mathcal{W}_i$ are zero below row $2^{i-1} + 1$ (see (3.1)). Similarly, rows $1, \ldots, 2^{i-1} + 1$ of $\hat{W}_i$ are zero to the right of column $(D-1)(2^{i-1} - 1) + 2$ so that rows $1, \ldots, 2^{i-1} + 1$ of $\mathcal{W}_i$ are linear combinations of rows $1, \ldots, (D-1)(2^{i-1} - 1) + 2$ of $\mathcal{W}_{i-1}$ and hence are zero to the right of column $(D-1)(2^{i-1}-1)+2+(D-1)(2^{i-1}-1) = (D-1)(2^i - 1) - D + 3 \leq (D-1)(2^i - 1) + 1$, since $D \in \{2, 4, \ldots\}$. This gives the first subblock of B_i in $\mathcal{W}_i$. The remaining subblocks of B_i are the result of the identity matrix in rows $2^{i-1} + 2, \ldots, 2^i$ of $\hat{W}_i$ multiplying the second B_{i-1} block of $\mathcal{W}_{i-1}$. We have shown that if the lemma is true for $k = i - 1$, it is also true for $k = i$, hence it is true for $k = 1, 2, \ldots$. □

COROLLARY 3.2. *$\mathcal{W}_k$ is a (wrap-round) band matrix with lower bandwidth $\leq 2^{k-1}$ and upper bandwidth $\leq (D-1)(2^k - 1)$.*

We can now deduce the following theorem.

THEOREM 3.3. *When an order D, level k DWTPer is applied to a band matrix B with lower bandwidth α and upper bandwidth β, the resultant matrix $\tilde{B}$ has lower bandwidth $\leq \alpha + (D-1)(2^k - 1) + 2^{k-1}$ and upper bandwidth $\leq \beta + (D-1)(2^k - 1) + 2^{k-1}$.*

Proof. $\tilde{B} = \mathcal{W}_k B \mathcal{W}_k^T$ is the product of band matrices. Hence $\tilde{B}$ is a band matrix whose bandwidths are the sums of the bandwidths of $\mathcal{W}_k$, B, and $\mathcal{W}^T$. □

The difference between the estimate for the bandwidths given in [11] and that given by Theorem 3.3 is $D(2^k - 1) - (D-1)(2^k - 1) - 2^{k-1} = 2^k - 2^{k-1} + 1$. When k is small this is not very great, but as k increases the savings achievable by using our new estimate become significant. Previous experience suggests that higher transform levels are required for larger matrices, so this improvement will be particularly valuable in solving very large (and hence very costly) dense linear systems.

3.1.1. Tightening the bound further. It is not possible to improve upon the bound given by Theorem 3.3 for the application of DWTPer to a *general* band matrix, since there exist examples for which these bounds are achieved. However, depending on the order of the DWTPer and the bandwidths, α and β, of the matrix B, the *actual* bandwidth of $\tilde{B}$ may be considerably less than that given by Theorem 3.3. Figure 3.3 shows examples of Daubechies 4 DWTPer applied to band matrices. When B is a diagonal matrix ($\alpha = \beta = 0$), $\tilde{B}$ has lower and upper bandwidths $(D-1)(2^k - 1) + 1$; when B has $\alpha = \beta = 3$, $\tilde{B}$ has bandwidths $(D-1)(2^k - 1) + 2^{k-1}$. We now present tighter bounds for the bandwidth of $\tilde{B}$ in the special (but commonly occurring) case where B is a diagonal matrix.

Example 3.1 (application of order 4 DWTPer to a diagonal matrix). The order 4, level k DWTPer matrix is of the form

(3.5)

$$\mathcal{W}_k = \begin{pmatrix}
* & \cdots & & \cdots & * & & & & & & & & & & \\
\vdots & \ddots & & & * & & & & & & & & & & \\
\vdots & & \ddots & & * & & & & & & & & & & \\
* & \cdots & & \cdots & * & & \cdots & * & & & & & & & \\
 & & & & \ddots & & & & & & & & & & \\
 & & & & & \ddots & & & & & & & & & \\
 & & * & & & \ddots & & & & & & & & & \\
 & & * & * & * & * & * & \cdots & & \cdots & * & & & & \\
 & & & & & & * & * & \cdots & \cdots & * & & & & \\
 & & & & & & * & * & * & \cdots & & \cdots & * & & \\
 & & & & & & & & * & * & \cdots & \cdots & * & & \\
 & & & & & & & & & & & \ddots & \ddots & \ddots &
\end{pmatrix},$$

in which the upper block bandwidth is $3(2^k-1)$. Applying this transform to a diagonal matrix B gives $\tilde{B} = \mathcal{W}_k B \mathcal{W}_k^T$. $B\mathcal{W}_k^T$ has the same sparsity pattern as $\mathcal{W}_k^T$. We now consider the sparsity pattern of $\tilde{B}$ when it is formed from $\mathcal{W}_k$ and $B\mathcal{W}_k^T$.

Columns $1 \ldots 2^k$ of $B\mathcal{W}_k^T$ are zero below row $3(2^k-1)+1$, so columns $1 \ldots 2^k$ of $\tilde{B}$ are linear combinations of columns $1, \ldots, 3(2^k-1)+1$ of $\mathcal{W}_k$. $3(2^k-1)+1 = 3 \cdot 2^k - 2$, so we are concerned with the columns up to 2 columns before the final column in a 2^k block. It is easy to see that these columns are zero below row $3(2^k-1)+1+1$, giving a bandwidth of $3(2^k-1)+1$.

Columns $2^k+1 \ldots 2^{k+1}$ of $B\mathcal{W}_k^T$ are zero below row $3(2^k-1)+2^k+1$, so columns $2^k+1 \ldots 2^{k+1}$ of $\tilde{B}$ are linear combinations of columns $1, \ldots, 3(2^k-1)+2^k+1$ of $\mathcal{W}_k$. Again this gives a bandwidth of $3(2^k-1)+1$.

Continuing in this way, we see that the overall bandwidth of $\tilde{B}$ is $3(2^k-1)+1$, which for large values of k is considerably less than the $3(2^k-1)+2^{k-1}$ predicted by Theorem 3.3.

THEOREM 3.4. *When an order D, level k DWTPer transform is applied to a diagonal matrix B, the bandwidths of the resultant $\tilde{B}$ are $(D-1)(2^k-1)+\ell$, where ℓ is a constant dependent only on D. Moreover, $\ell \leq D-2$, and for all values of D sufficiently small to be useful for preconditioning, $\ell \leq 3$.*

Proof. As in the above example, the sparsity pattern of $\tilde{B}$ is that of the product of matrices with the sparsity patterns of $\mathcal{W}_k$ and $\mathcal{W}_k^T$. The first 2^k columns of $\mathcal{W}_k^T$ are zero below row $(D-1)(2^k-1)+1 = (D-1)2^k+2-D$, which implies that the first 2^k columns of $\tilde{B}$ are linear combinations of the first $(D-1)2^k+2-D$ columns of $\mathcal{W}_k$. The bandwidth of $\tilde{B}$ is thus given by $(D-1)(2^k-1)+\ell$, where ℓ is the number of nonzero entries below the diagonal in column 2^k+2-D of $\mathcal{W}_k$ and so must satisfy $\ell \leq D-2$. ∎

The precise value of ℓ for any given D can readily be found by inspection of the sparsity pattern of $\mathcal{W}_k$. The table below gives the values of ℓ for $D = 2, 4, \ldots, 20$. In practice, the large bandwidths associated with the use of high order DWTPer means that, for preconditioning purposes, $2 \leq D \leq 8$, giving $\ell \leq 3$.

D	2	4	6	8	10	12	14	16	18	20
ℓ	0	1	1	3	1	3	5	7	1	3

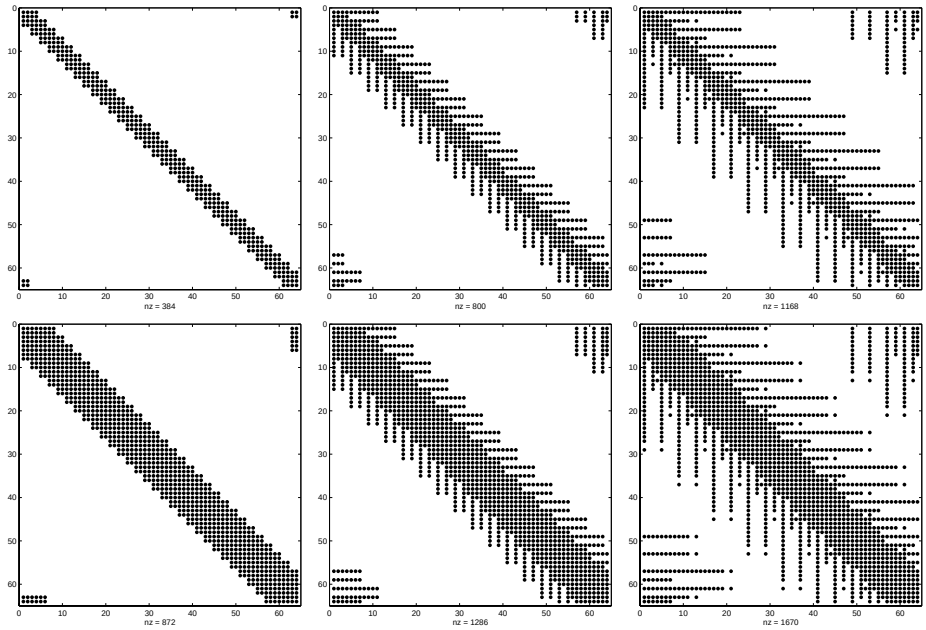


FIG. 3.3. *Daubechies* 4 *DWTPer, levels* 1–3*, applied to band matrices. Top: diagonal matrix; below: upper and lower bandwidth* 3.

3.2. Our new preconditioner. When a level k DWTPer is applied to an $N \times N$ matrix A, the weighted average entries in $\tilde{A}$ are $A_{ij},\ i,j = 2^k, 2\cdot 2^k, 3\cdot 2^k, \ldots$. We can permute the rows and columns so that these entries are confined to bands of width $N/2^k$ at the bottom and right-hand edges of $\tilde{A}$. We can then form a preconditioner by setting to zero all entries that fall outside the diagonal band and these two edge bands. As we have seen in the previous section, the width of the diagonal band depends on the transform level k. In previous work we have used lower and upper bandwidths $(2^k-1)D+\alpha$, $(2^k-1)D+\beta$ (see [11]), so a band preconditioner based on "standard" DWTPer would have $\approx N(2(2^k-1)D+\alpha+\beta+1)$ entries. We can reduce the bandwidth, while still retaining all the entries corresponding to a diagonal band singularity, by using the tighter bound given in Theorem 3.3, or, in the case of a singularity confined to the leading diagonal, by using the smaller bandwidths given in Theorem 3.4.

As k increases, the cost of the standard DWTPer band preconditioner increases (since the bandwidth increases), but the proportion of large entries that are neglected falls, so that we expect the convergence rate of our iterative method to improve. Successful DWTPer preconditioning depends on selecting k to be sufficiently large to give fast convergence, without the preconditioner becoming too expensive. This choice has usually been made by making an "educated guess" based on the dimension and structure of A; for example, a higher level DWTPer is needed for larger matrices and for those with a less pronounced diagonal singularity. Although this may appear to be a rather "hit and miss" procedure, it often works well, because, although it is relatively unlikely that the "optimal" wavelet level will be chosen, it is usually quite easy to choose a level that gives satisfactory performance. (In particular, it is considerably easier to choose an appropriate k for DWTPer preconditioning than

to choose a good bandwidth for preconditioning the untransformed A with a band approximation.) We now show that, with our new preconditioner, which we shall call DWTPerMod, it is possible to choose an optimal value of k given N and D.

Once we have chosen a suitable method for determining the diagonal bandwidth (using Theorem 3.3 or Theorem 3.4 as appropriate), we can estimate the cost of applying the factorized bordered block preconditioner by forward and backward substitution, based on the widths of the diagonal, horizontal, and vertical bands. Since the preconditioner is designed to include both the weighted average and the largest weighted difference entries in $\tilde{A}$, we do not expect the convergence rate to change much as k changes, so we can expect to minimize the overall cost by minimizing the cost of the preconditioning step at each iteration. The cost of forward and backward substitution is proportional to the number $N_z(k)$ of nonzero entries in the LU factors of $\tilde{M}$. Hence we choose k such that $N_z(k)$ is a minimum. A matrix of dimension N with borders of width r and a diagonal band with upper and lower bandwidth p can be factored into LU factors such that $N_z \approx N(3p + 2r)$ (a trivial extension of Theorem 4.3.2 in [24]). We summarize our new preconditioning method in Algorithm 3.1 below.

ALGORITHM 3.1 (DWTPerMod preconditioner). *Given a dense matrix A of dimension N and a DWT of order D, compute DWTPerMod preconditioner as follows:*

1. **for** $i = 1, 2, \ldots, \log_2(N/(D-1))$, *compute*
   $p(i)$ *using Theorem* 3.3 *or Theorem* 3.4,
   $r(i) = N/(2^i)$,
   $N_z(i) = N(3p(i) + 2r(i))$.
2. *Choose k such that $N_z(k) = \min_i[N_z(i)]$.*
3. *Apply a level k DWTPer to A to obtain $\tilde{A}$.*
4. *Permute the rows and columns of $\tilde{A}$ so that the weighted average entries lie in bands at the right-hand and bottom edges.*
5. *Form a bordered block preconditioner $\tilde{M}$ by setting to zero entries in $\tilde{A}$ outside of a diagonal band of width $p(k)$ and borders of width $r(k)$.*

Remark 3.1. In the above analysis we have assumed, for simplicity, that the upper and lower bandwidths of the diagonal singularity in A are equal. However, this preconditioning strategy is equally applicable when the bandwidths are different and this requires only trivial modifications to take into account the two bandwidths.

4. Numerical results. In this section we give illustrative results comparing the performance of our new modified DWTPer preconditioner with other preconditioning methods commonly used for solving linear systems defined by a smooth matrix with diagonal singularity. In each case the results we give record the performance of our new preconditioner using Daubechies 4 DWT with the optimal level described in section 3.2, and we use right preconditioned GMRES and iterate until the relative residual falls below 10^{-6}.

Bearing in mind our aim to develop algorithms that can be used for solving dense systems without the need for the user to bring to bear detailed knowledge of the continuous problem from which they arise, we have chosen to use the Daubechies family of wavelets whose filter coefficients are readily available and which have the convenient property of orthogonality. It is likely that, for individual examples, better compression could be achieved by using custom-built wavelets tuned to the properties of the particular matrix involved, but here we are concerned primarily with the effect of applying different permutations in order to produce the most effective preconditioner for a given choice of wavelet. The choice of order 4 Daubechies DWT (with 2 vanishing

moments) was arrived at in the light of previous experience [17, 20]. Higher order Daubechies wavelets would give better compression of the smooth part of the matrix (because of the larger number of vanishing moments), but the larger number of filter coefficients results in more large entries corresponding to the diagonal singularity (see Theorems 3.3 and 3.4). The optimal wavelet order for a given matrix depends on the width of the diagonal singularity, the "smoothness" of the smooth part and the size of the matrix; we have not attempted to find this optimal order here.

We appreciate that the stopping criterion is somewhat arbitrary, but our desire here is simply to compare the ability of several preconditioners to speed up convergence of the GMRES iteration for various dense systems, and so we have not attempted to choose a tolerance dependent on the discretization.

The other preconditioners that we use are

1. a band preconditioner formed by setting to zero entries of the matrix outside a chosen bandwidth;
2. a standard DWT "finger pattern" preconditioner using Daubechies 4 DWT;
3. a standard DWTPer band preconditioner (see [11]) using Daubechies 4 DWT.

It is not easy to decide a priori what will be the optimal choice of bandwidth for band preconditioners or wavelet level for DWT and DWTPer preconditioners. In our experiments we solved each linear system using band preconditioners with bandwidths varying up to $N/4$, where N is the dimension of the linear system, and recorded the best performance. Similarly, we used DWT levels of 1 to 5 for testing the DWT and DWTPer preconditioners and recorded the *best* performance. In practice, the optimal bandwidth and wavelet level would frequently not be chosen, so these results compare the *best possible* performance of the band, DWT, and DWTPer preconditioners with the *standard* performance of our new preconditioner. This illustrates one of the important features of our new method, namely that, by providing an optimal choice of wavelet level (and hence of sparsity pattern for the preconditioner) based only on the order of the DWT and the dimension of the linear system, it removes the difficulty over the rather arbitrary choice of bandwidth or wavelet level needed for band, DWT, or DWTPer preconditioner design.

We would expect that the standard DWT finger pattern preconditioner will give faster convergence than the equivalent DWTPer band preconditioner, since both the weighted difference entries associated with the diagonal singularity and all of the weighted average entries are included in the finger pattern, while many of the weighted average entries are neglected in the band. However, the additional cost of applying the finger pattern preconditioner can be expected to make it less cost-effective. Our new preconditioner should enable us to retain the fast convergence of the finger pattern preconditioner while reducing the cost of preconditioner application. Our experimental results confirm this.

4.1. Example 1: Calderón–Zygmund matrices. We consider matrices of the form

$$A_{ij} = \begin{cases} 1/|i-j|^{\alpha}, & i \neq j, \\ 1, & i = j, \end{cases} \tag{4.1}$$

where α is a constant which controls the "steepness" of the singularity at the diagonal. DWTPer-based preconditioners are known to be effective for such matrices when $\alpha \geq 1$ but are less satisfactory for smaller values of α (corresponding to less pronounced diagonal singularities). A comparison of the performance of DWTPer and standard DWT preconditioners for such matrices can be found in [19]. Here, in Table 4.1,

TABLE 4.1
Comparison of preconditioners for Example 1 *using GMRES iterated to a tolerance of* 10^{-6}.

		Preconditioned GMRES							
	Direct	Band		Std. DWT		DWTPer		DWTPerMod	
N	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops
128	1.5	123	16.5	16	2.5	19	3.0	7	1.9
256	12	211	113	20	13	20	14	9	9
512	91	402	839	34	94	47	83	10	46
1024	723	888	7285	592	3920	145	762	12	218

we compare the performance of different preconditioners for $\alpha = 0.2$, corresponding to a rather weak diagonal singularity and resulting poor performance of DWTPer. Although this matrix is symmetric, we have chosen to use GMRES, rather than CG or some other symmetric solver, for consistency with our other examples and for ease of comparison with the results in [19].

For this matrix, band preconditioners are ineffective. The wavelet-based preconditioners perform better, but none of them fully succeed in keeping the number of iterations constant as the problem size increases. The DWTPerMod preconditioner comes closest to this objective with only modest increases in iteration counts and with the overall solution cost being approximately $\mathcal{O}(N^2)$, while the cost of the other preconditioners is almost $\mathcal{O}(N^3)$.

4.2. Example 2: A matrix used previously to test wavelet-based solution techniques. We consider the matrix

$$A_{ij} = \begin{cases} C/\tan(\pi(i-j)/N), & i \neq j, \\ 1, & i = j, \end{cases} \tag{4.2}$$

where $i, j = 1, 2, \ldots, N$ and $C = 1/N$. This skew-symmetric matrix is used in [23] to compare the performance of direct solution methods based on standard DWT and NS-forms with a direct solution method by LU factorization of the original dense matrix. Table 4.2 shows our results using DWTPerMod, DWTPer, and band preconditioners. For this matrix, we have previously found that better results for DWTPer band preconditioners are obtained using the Haar DWT rather than Daubechies 4, so we used this wavelet basis for each of the DWT-based preconditioners.

In this case *all* of the preconditioners tested give $\mathcal{O}(N^2)$ solution costs and fairly steady iteration counts. Our new preconditioner gives modest savings compared with the others, but it must be remembered that a less than optimal choice of transform level or bandwidth could severely impair performance of all but the DWTPerMod preconditioner so that, even if performance were *equal*, DWTPerMod could be said to offer a more reliable alternative.

4.3. Example 3: Solution of the boundary integral equation associated with the solution of Helmholtz equation in an elliptic domain. The Helmholtz equation models the propagation of acoustic waves through a medium. The exterior Helmholtz problem can be formulated as an integral equation with the following formulation due to Burton and Miller [7]:

$$\left(-\frac{1}{2}\mathcal{I} + \mathcal{M}_k + i\eta\mathcal{N}_k\right)\varphi = \left[\mathcal{L}_k + i\eta\left(\frac{1}{2}\mathcal{I} + \mathcal{M}_k^T\right)\right]\frac{\partial\varphi}{\partial n}. \tag{4.3}$$

TABLE 4.2
Comparison of preconditioners for Example 2 using GMRES iterated to a tolerance of 10^{-6}.

		Preconditioned GMRES							
	Direct	Band		Std. DWT		DWTPer		DWTPerMod	
N	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops
128	1.5	8	0.8	9	1.0	16	1.4	4	0.8
256	12	9	4	16	3	16	5	5	3
512	91	10	9	16	14	16	21	5	14
1024	723	11	59	10	63	16	82	5	52
2048	5759	12	247	10	251	16	325	5	212

TABLE 4.3
Comparison of preconditioners for Example 3 using GMRES iterated to a tolerance of 10^{-6}.

		Preconditioned GMRES							
	Direct	Band		Std. DWT		DWTPer		DWTPerMod	
N	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops
128	6	24	8	7	6	24	10	6	6
256	47	25	31	6	28	25	37	7	23
512	366	26	127	5	158	26	157	6	92
1024	2894	26	515	5	1003	26	636	7	369

Here η is an arbitrary nonzero coupling parameter and the operators $\mathcal{L}_k$, $\mathcal{M}_k$, and $\mathcal{N}_k$ are defined by

$$\mathcal{L}_k\varphi(p) = \int_S G_k(p,q)\varphi(q)dS, \tag{4.4}$$

$$\mathcal{M}_k\varphi(p) = \int_S \frac{\partial G_k(p,q)}{\partial n_q}\varphi(q)dS, \tag{4.5}$$

$$\mathcal{N}_k\varphi(p) = \frac{\partial}{\partial n_p}\mathcal{M}_k\varphi(p) = \frac{\partial}{\partial n_p}\int_S \varphi(q)\frac{\partial G_k(p,q)}{\partial n_q}dS_q. \tag{4.6}$$

Here $G_k(p,q)$ is the free-space Green's function for the Helmholtz equation.

This boundary integral equation (BIE) can be solved numerically using a collocation method (see, e.g., [1]). This solution method requires that a non-Hermitian, complex, dense linear system be solved, and previous work (see, e.g., [11, 17]) has shown that DWT-based preconditioners can be effective in speeding up solution by iterative solvers. Table 4.3 gives a comparison between our new preconditioner and the others that we have been considering here for the case where S is a circle.

Although the standard DWT preconditioner usually gives the fewest iterations, the DWTPerMod preconditioner clearly outperforms all the others in terms of computational cost. We can see why this should be by looking at the magnitudes of the entries in the original matrix and its (level 3) DWT, DWTPer, and DWTPerMod transforms as shown in Figure 4.1. The entries in the original matrix do not decay very fast away from the diagonal, which means that a large bandwidth must be used in order to get good convergence with a band preconditioner; fill-in in a preconditioner based on the finger pattern of the standard DWT transform makes it expensive to apply; many more of the largest entries can be included in the bordered block preconditioner based on DWTPerMod than in a band preconditioner based on DWTPer.

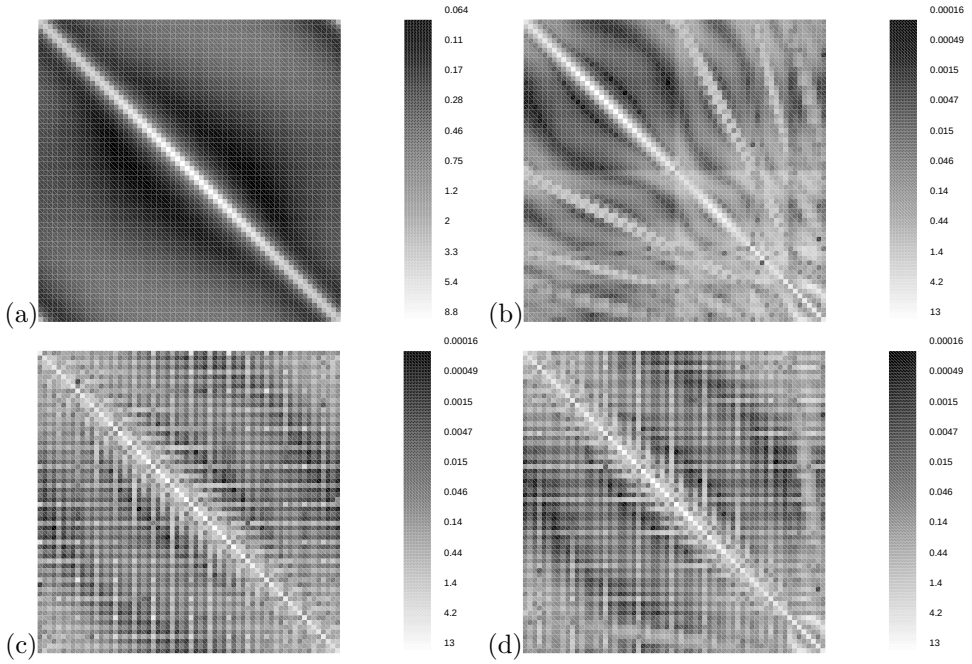


FIG. 4.1. *Modulus of (complex) matrix from Example* 3. (a) *original matrix;* (b) *standard DWT;* (c) *DWTPer;* (d) *modified DWTPer.*

4.4. Example 4: Elastohydrodynamic lubrication Jacobian matrix. Elastohydrodynamic lubrication (EHL) theory models the behavior of lubricating fluids in circumstances where the elastic deformation of the surrounding surfaces plays a significant role in the hydrodynamic process. For example, the EHL line contact problem models the flow of lubricating fluid between two cylinders rotating under an applied load. This physical situation can be modeled by a coupling of Reynolds equation for flow of the lubricating fluid with the elastic deformation equation for the cylinders [14]. Under the assumption that the length of the cylinders is large compared with the width of the contact area and can therefore be considered to be infinite, the problem is essentially one-dimensional; we are interested in finding the film thickness and pressure at each point along the contact width, which can be viewed as an interval in $\mathbb{R}^1$.

A standard solution method for the resulting integro-differential equation is to use a finite difference discretization [20, 27, 29], which yields a system of nonlinear equations, which are solved using a Newton-based search method. Here we are concerned with the fast solution of the system of linear equations, corresponding to a Jacobian matrix, which must be solved at each step of the Newton iteration. This matrix is smooth, dense, and highly nonsymmetric, with a nonsmooth diagonal band (see, for example, [20]).

We have previously found (see [20]) that DWTPer preconditioning gives much improved performance compared with band or standard DWT preconditioning, particularly for the more difficult high-load problems. We now compare these approaches with our new DWTPerMod preconditioner. The results for a typical Jacobian matrix from the first Newton iteration can be found in Table 4.4. For all matrix sizes, the two DWTPer-based methods are clearly superior to the others considered. DWTPerMod

TABLE 4.4
Comparison of preconditioners for Example 4 *using GMRES iterated to a tolerance of* 10^{-6}.

		Preconditioned GMRES							
	Direct	Band		Std. DWT		DWTPer		DWTPerMod	
N	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops	Its.	Mflops
129	1.6	8	1.8	20	2.9	15	2.0	7	1.9
257	12	9	11	17	11	18	8	10	9
513	92	10	60	17	55	22	34	9	32
1025	726	11	366	19	330	28	170	12	139

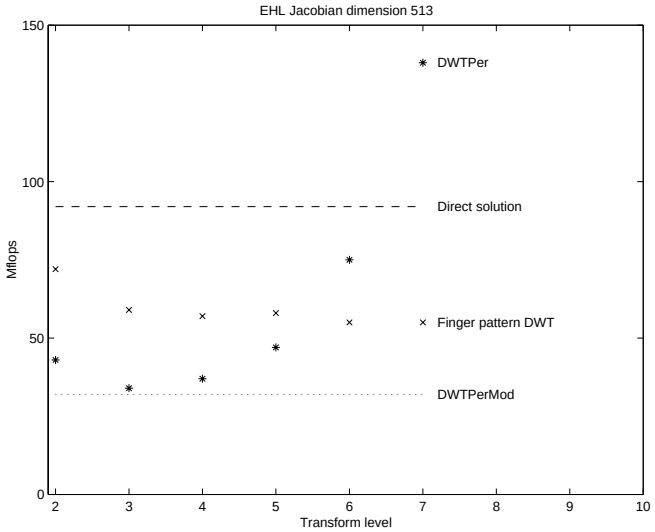


FIG. 4.2. *Cost of solving linear systems of Example* 4 *using DWT and DWTPer preconditioners of different levels.*

is competitive with DWTPer for all sizes and considerably more cost-effective for the largest matrix size. Taking into account the fact that the results recorded here for DWTPer are those using the "best possible" transform level, we can conclude that the DWTPerMod preconditioner provides a reliable method for preconditioning this problem.

Figure 4.2 shows how the costs of DWTPer and standard DWT preconditioners vary with the transform level with $N = 513$. Notice that, although the "best" DWTPer preconditioner is almost as cost-effective as the DWTPerMod preconditioner, if the "wrong" level is chosen, solution using a DWTPer preconditioner may cost even more than using a direct solver.

Remark 4.1. Although, with the exception of section 4.3, the matrices in our example problems are *not* periodic, our methods can be seen to have been effective in preconditioning them using *periodized* Daubechies wavelets.

5. Potential for application to higher-dimensional problems. A limitation of the theory and numerical results presented in the previous sections is that they relate only to one-dimensional problems. For matrices that correspond to discretizations of higher-dimensional problems, it is often possible to use a Kronecker product approximation (see, e.g., [34]) to reduce the storage and computational costs of solving the linear system. Recent developments by Tyrtyshnikov [31, 32] allow

dense matrices related to functions on a two-dimensional grid to be approximated by a sum of Kronecker products in an extremely efficient way. When the Kronecker factors are themselves smooth (as is frequently the case), further savings can be made by applying a DWT to the factors and setting to zero negligibly small entries. Details of this strategy for solving large (having perhaps 1 million or more unknowns) linear systems of this type, using standard (unpermuted) DWT, can be found in [21]; we summarize it briefly here.

5.1. Wavelet-enhanced Kronecker-product approximation. Suppose that the matrix A is associated with a function of two variables in the following way:

$$A = (a_{ij}) = (f(z_i, z_j)), \qquad i, j = 1, 2, \ldots, n = pq, \tag{5.1}$$

where $\{z^i\}$ and $\{z^j\}$ are the nodes of some grids logically equivalent to the Cartesian product of two one-dimensional grids with p and q grid points, respectively. Then it is possible (see [21, 31, 32]) to approximate A, to any required accuracy, by a sum of Kronecker products:

$$A \approx U_1 \otimes V_1 + U_2 \otimes V_2 + \cdots + U_r \otimes V_r, \tag{5.2}$$

where the U_i and V_j matrices are of dimensions $p \times p$ and $q \times q$, respectively. Provided that r is small compared with pq, this offers significant savings in storage. If the underlying function f is smooth, then the U_i and V_j can be expected to be smooth matrices (like those corresponding to discretizations of functions on a one-dimensional grid), and further compression of the data can be achieved by applying a DWT and setting to zero entries that fall below a chosen threshold. This gives a new approximation $P_1 \otimes Q_1 + P_2 \otimes Q_2 + \cdots + P_r \otimes Q_r$, where the P_i and Q_i are sparse matrices expressed in a wavelet basis. The linear system $Ax = b$ can be solved approximately by solving the approximate system. Inverting $\hat{A} = P_1 \otimes Q_1 + P_2 \otimes Q_2 + \cdots + P_r \otimes Q_r$ would be costly, but this can be overcome by using an iterative method such as GMRES. The matrix-vector multiplication at each iteration can be achieved at low cost using the identity (see, e.g., [33])

$$(U \otimes V)\, x = \operatorname{vec}\left(VXU^T\right), \tag{5.3}$$

where X is the $p \times q$ matrix obtained by listing the entries of x in the columns of X, and $\operatorname{vec}(Y)$ is the vector obtained by columnwise listing of the entries of Y.

To speed up convergence of the iteration, a preconditioner may be needed. Two options that have been tried [21] are

- an inverse Kronecker product (IKP) preconditioner,
- an incomplete LU preconditioner with threshold (ILUT).

The first of these uses the single Kronecker product $P_1 \otimes Q_1$, whose inverse $P_1^{-1} \otimes Q_1^{-1}$ can be applied by LU factorization of P_1 and Q_1. This essentially reduces the task to that of preconditioning the dense matrices U_1 and V_1, each of which is smooth with a diagonal singularity. The DWTPerMod approach can, therefore, be expected to be an improvement on standard DWT, and preliminary tests confirm this: for some examples, savings of as much as one quarter in the number of nonzero entries in the factorized preconditioner are possible without affecting the number of iterations required for convergence of the iterative method.

The second option relies on applying a higher threshold to obtain very sparse approximations $\bar{P}_i$, $\bar{Q}_i$ of the Kronecker factors P_i, Q_i so that $\bar{A} = \bar{P}_1 \otimes \bar{Q}_1 + \bar{P}_2 \otimes \bar{Q}_2 + \cdots + \bar{P}_r \otimes \bar{Q}_r$ is sparse enough to be held in memory explicitly. An incomplete LU

factorization of $\bar{A}$ then provides a preconditioner. In experiments this preconditioner gave similar performance to that of the IKP preconditioner for "smaller" matrices (up to dimension 260000 on an AMD-1000 computer with 1 Gbyte operative memory) but was infeasible for larger matrices because of the amount of memory required to store the preconditioner. The advantages of DWTPerMod over standard DWT are less clear-cut here, because fill-in under *incomplete* LU factorization is less dependent on the sparsity structure of the matrix than is the full LU factorization algorithm. We have not so far found any significant benefit from using the new transform in this context.

6. Conclusion and future work. We have designed a new DWTPer-based preconditioning method for dense matrices with nonsmooth diagonal bands. This improves on previous preconditioners in three main ways:

1. tighter bounds on the bandwidth are required for DWTPer band preconditioning, enabling such preconditioning to be done at lower cost;
2. there is inclusion of more of the significant entries in the preconditioner, giving a better approximation and hence faster convergence;
3. there is removal of uncertainty about choosing an appropriate bandwidth or wavelet level, giving a more robust method.

We have tested the method using several example problems and have found that, in every case, the new method performs substantially better than diagonal, band, and standard DWT preconditioners. In a majority of cases it also outperforms the "best possible" DWTPer band preconditioner, and in every case the new method is competitive with it and is significantly more effective than DWTPer band preconditioners for which the "wrong" transform level has been chosen. This means that, for the examples that we have tried, the best preconditioner may be a DWTPer band approximation, but the most reliable preconditioner is our new DWTPerMod approach since no user intervention is required to choose an appropriate transform level for each problem.

Work is currently ongoing to extend the approach to tackle higher dimension problems using a combination of Kronecker product and one-dimensional DWT compression, and we expect that DWTPerMod will be a useful tool in this context as well as in the one-dimensional case.

Using our new preconditioner, given the order of the DWT, we can determine the optimal transform level, but there still remains the question of how to choose the most effective wavelet basis for compressing a given matrix. We plan in the future to develop ways of using measured smoothness properties of a matrix (such as those defined in [15, Chap. 4]) to determine the compression effect of a discrete wavelet transform with a given number of vanishing moments and hence to choose an appropriate order from a family of discrete wavelet transforms, such as the Daubechies family. This would represent a significant step towards the development of a purely algebraic wavelet-based preconditioning strategy for dense matrices.

Acknowledgments. I would like to thank the anonymous referees and my colleagues Dr. David Silvester and Prof. Neville Ford for their careful reading of the draft of this paper and for their helpful suggestions. I am also grateful to Dr. Ke Chen of Liverpool University, whose computer program produced the matrices used in section 4.3, Example 3, to Dr. Laurence Scales of Shell Global Solutions, who introduced me to elastohydrodynamic lubrication problems, and to Prof. Eugene Tyrtyshnikov, who developed the Kronecker product approximation method for dense matrices relating to two-dimensional problems.

REFERENCES

[1] S. Amini, P. J. Harris, and D. T. Wilton, *Coupled Boundary and Finite Element Methods for the Solution of the Dynamic Fluid-Structure Interaction Problem*, Lecture Notes in Engrg. 77, Springer-Verlag, Berlin, 1992.

[2] S. Amini and N. Maines, *Iterative solutions of boundary integral equations*, in Boundary Element Methods XVI, Vol. 1, C. A. Brebbia, ed., Computational Mechanics Publications, Southampton, UK, 1994, pp. 193–200.

[3] G. Beylkin, R. R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[4] G. Beylkin, R. R. Coifman, and V. Rokhlin, *Wavelets in numerical analysis*, in Wavelets and Their Applications, Jones and Bartlett, Boston, 1992, pp. 181–210.

[5] G. Beylkin and J. M. Keiser, *On the adaptive numerical solution of nonlinear partial differential equations in wavelet bases*, J. Comput. Phys., 132 (1997), pp. 233–259.

[6] R. Bridson and W.-P. Tang, *Multiresolution approximate inverse preconditioners*, SIAM J. Sci. Comput., 23 (2001), pp. 463–479.

[7] A. J. Burton and G. F. Miller, *The application of integral equation methods to the numerical solution of some exterior boundary-value problems*, Proc. Roy. Soc. London Ser. A, 323 (1971), pp. 201–210.

[8] T. F. Chan and K. Chen, *Two-Stage Preconditioners Using Wavelet Band Splitting and Sparse Approximation*, Report CAM 00-26, University of California Los Angeles, Los Angeles, 2000.

[9] T. F. Chan, W. P. Tang, and W. L. Wan, *Wavelet sparse approximate inverse preconditioners*, BIT, 37 (1997), pp. 644–660.

[10] K. Chen, *On a class of preconditioning methods for dense linear systems from boundary elements*, SIAM J. Sci. Comput., 20 (1998), pp. 684–698.

[11] K. Chen, *Discrete wavelet transforms accelerated sparse preconditioners for dense boundary element systems*, Electron. Trans. Numer. Anal., 8 (1999), pp. 138–153.

[12] C. K. Chui, *Wavelets: A Mathematical Tool for Signal Analysis*, SIAM Monogr. Math. Model. Comput., SIAM, Philadelphia, 1997.

[13] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.

[14] D. Dowson and G. R. Higginson, *Elastohydrodynamic Lubrication*, Pergamon Press, Oxford, 1977.

[15] J. Ford, *Wavelet-based Preconditioning of Dense Linear systems*, Ph.D. thesis, University of Liverpool, Liverpool, 2001; available online from http://www.ma.umist.ac.uk/jf/judythes.ps.gz.

[16] J. Ford and K. Chen, *An algorithm for accelerated computation of DWTPer-based band preconditioners*, Numer. Algorithms, 26 (2001), pp. 167–172.

[17] J. Ford and K. Chen, *Wavelet-based preconditioners for dense matrices with non-smooth local features*, BIT, 41 (2001), pp. 282–307.

[18] J. Ford and K. Chen, *Speeding up the solution of thermal EHL problems*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 2305–2310.

[19] J. Ford, K. Chen, and D. Evans, *On a recursive Schur preconditioner for iterative solution of a class of dense matrix problems*, Int. J. Comput. Math., 80 (2003), pp. 105–122.

[20] J. Ford, K. Chen, and L. Scales, *A new wavelet transform preconditioner for iterative solution of elastohydrodynamic lubrication problems*, Int. J. Comput. Math., 75 (2000), pp. 497–513.

[21] J. M. Ford and E. E. Tyrtyshnikov, *Combining Kronecker product approximation with discrete wavelet transforms to solve dense, function-related linear systems*, SIAM J. Sci. Comput., 25 (2003), pp. 961–981.

[22] A. Gerasoulis, *Nystrom's iterative variant methods for the solution of Cauchy singular integral equations*, SIAM J. Numer. Anal., 26 (1989), pp. 430–441.

[23] D. Gines, G. Beylkin, and J. Dunn, *LU factorization of non-standard forms and direct multiresolution solvers*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 156–201.

[24] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, London, 1996.

[25] W. Hackbusch, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part* I: *Introduction to $\mathcal{H}$-matrices*, Computing, 62 (1999), pp. 89–108.

[26] N. I. Ioakimidis and P. S. Theocaris, *A comparison between the direct and the classical numerical methods for the solution of Cauchy type singular integral equations*, SIAM J. Numer. Anal., 17 (1980), pp. 115–118.

[27] G. E. Morales-Espejel and A. Felix-Quinonez, *Kinematics of two-sided surface features in elastohydrodynamic lubrication*, Proc. Inst. Mech. Engrs., 213 (1999), pp. 95–108.
[28] E. Nurgat, M. Berzins, and L. Scales, *Solving EHL problems using iterative, multigrid, and homotopy methods*, Trans. ASME, 121 (1999), pp. 28–34.
[29] T.-J. Park and K.-W. Kim, *Elastohydrodynamic lubrication of a finite line contact*, Wear, 223 (1998), pp. 102–109.
[30] E. E. Tyrtyshnikov, *Mosaic-skeleton approximations*, Calcolo, 33 (1996), pp. 47–57.
[31] E. E. Tyrtyshnikov, *Kronecker-product approximations for some function-related matrices*, Linear Algebra Appl., to appear.
[32] E. E. Tyrtyshnikov, *Tensor approximations of matrices associated with asymptotically smooth functions*, Sb. Math., to appear.
[33] C. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100.
[34] C. Van Loan and N. P. Pitsianis, *Approximation with Kronecker products*, NATO Sci. Ser. E Appl. Sci., 232 (1993), pp. 293–314.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 662–668

HIGHER ORDER LOGARITHMIC DERIVATIVES OF MATRICES IN THE SPECTRAL NORM*

RAJENDRA BHATIA† AND LUDWIG ELSNER‡

Abstract. For the spectral norm ||.|| on $n \times n$ complex matrices, we derive the first three right-hand derivatives of $\phi(t) = ||e^{tA}||$ at $t = 0$. The first one is the well-known logarithmic derivative. This study was inspired by a recent result by Kohaupt, where the second derivative is studied for the l_p norms, $p = 1, \infty$.

Key words. logarithmic derivative, exponential function, higher order derivatives

AMS subject classifications. 65F35, 65F05

DOI. 10.1137/S0895479802413662

1. Introduction.

Let A be an $n \times n$ complex matrix, and let $||A||$ be its norm as a linear operator on the Euclidean space $\mathbb{C}^n$; i.e.,

$$||A|| = \sup \{||Ax|| : x \in \mathbb{C}^n,\ ||x|| = 1\}, \tag{1}$$

where $||x||$ is the Euclidean norm of the vector x. The initial-value problem

$$\dot{x}(t) = Ax(t),\ x(0) = x_0 \tag{2}$$

has the solution

$$x(t) = e^{tA}\ x_0. \tag{3}$$

For many purposes—such as error bounds—one needs upper bounds for the quantity $||e^{tA}||$. A very useful bound is given in terms of the *logarithmic derivative* of A defined as

$$\mu(A) = \lim_{h \to 0^+} \frac{||e^{hA}|| - 1}{h}. \tag{4}$$

We have

$$||e^{tA}|| \ \leq\ e^{\mu(A)t} \quad \text{for all} \quad t \geq 0, \tag{5}$$

and $\mu(A)$ is the smallest number for which such an inequality holds. We know that

$$\mu(A) = \lambda_1 \left(\frac{A + A^*}{2} \right), \tag{6}$$

where $\lambda_1(H)$ denotes the maximum eigenvalue of a Hermitian matrix H. See [1, 5].

In a recent paper [4], Kohaupt studied the problem of finding the second logarithmic derivative and solved it when the operator norm is induced not by the Euclidean

*Received by the editors May 8, 2002; accepted for publication (in revised form) by U. Helmke April 2, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/simax/25-3/41366.html
†Indian Statistical Institute, New Delhi 110016, India (rbh@isid.ac.in).
‡Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, D-33501 Bielefeld, Germany (elsner@mathematik.uni-bielefeld.de).

norm as in our definition (1) but by the p-norm where $p = 1$ or ∞. In this note we resolve the problem for $p = 2$. The somewhat unexpected answer led us to investigate the third derivative as well. We prove the following theorem.

THEOREM 1. *Let $\varphi(t) = ||e^{tA}||,\ t \geq 0$, and let $\dot{\varphi}(0),\ \ddot{\varphi}(0),\ \dddot{\varphi}(0)$ denote the first three right derivatives of φ at 0. Then*

$$\ddot{\varphi}(0) = \dot{\varphi}(0)^2. \tag{7}$$

Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of $A + A^*$ and $x_1, \ldots, x_n$ the corresponding eigenvectors.

Then

$$\dddot{\varphi}(0) = \dot{\varphi}(0)^3 - \frac{1}{4} \sum_{j=2}^{n} (\lambda_1 - \lambda_j)\ |\langle x_j, Ax_1\rangle|^2. \tag{8}$$

Note that $\dot{\varphi}(0)$ is just $\mu(A)$. The equality (7) is a little surprising and does not persist when we go to the third derivative. Our proof of (8) requires the assumption that the eigenvalue λ_1 is simple. It might be possible to drop this requirement.

2. Proofs. To handle higher order terms we need an extension of a standard perturbation result. The discussion in the next paragraph is modeled on that in [6, p. 69]. Series expansions of the kind we use are also derived in [3, p. 120].

Consider the eigenequation

$$(A + \epsilon B + \epsilon^2 C)\ x_1(\epsilon) = \lambda_1(\epsilon)\ x_1(\epsilon), \tag{9}$$

where A, B, C are Hermitian, and $\lambda_1(0) = \lambda_1$ is a simple eigenvalue of A. Then we have a series expansion

$$\lambda_1(\epsilon) = \lambda_1 + \epsilon k_1 + \epsilon^2 k_2 + \cdots. \tag{10}$$

Let $x_1,\ x_2, \ldots, x_n$ be the eigenvectors of A corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. The vector $x_1(\epsilon)$ has a series expansion

$$x_1(\epsilon) = x_1 + (\epsilon t_{21} + \epsilon^2 t_{22} + \cdots)\ x_2 + \cdots + (\epsilon t_{n1} + \epsilon^2 t_{n2} + \cdots)\ x_n. \tag{11}$$

The coefficients k_1 and k_2 are found as follows. Combine (9), (10), and (11) and equate the first order terms in ϵ to get

$$A(t_{21}x_2 + t_{31}x_3 + \cdots + t_{n1}x_n) + Bx_1 = \lambda_1(t_{21}x_2 + t_{31}x_3 + \cdots + t_{n1}x_n) + k_1x_1.$$

Taking the inner product of both sides with x_1, we get

$$\langle x_1, Bx_1\rangle = k_1, \tag{12}$$

while taking inner products with $x_j,\ j \geq 2$, we get (using $Ax_j = \lambda_j x_j$)

$$t_{j1} = \frac{\langle x_j, Bx_1\rangle}{\lambda_1 - \lambda_j},\ j \geq 2. \tag{13}$$

Again, using (9), (10), and (11) and equating second order terms in ϵ, we get

$$A\left(\sum_{j=2}^{n} t_{j2}x_j\right) + B\left(\sum_{j=2}^{n} t_{j1}x_j\right) + C\ x_1 = \lambda_1\left(\sum_{j=2}^{n} t_{j2}x_j\right) + k_1\left(\sum_{j=2}^{n} t_{j1}x_j\right) + k_2x_1.$$

Taking the inner product of both sides with x_1, and then substituting for t_{j1} from (13), we get

$$\sum_{j=2}^{n} \frac{|\langle x_j, Bx_1\rangle|^2}{\lambda_1 - \lambda_j} \; + \; \langle x_1, Cx_1\rangle \; = k_2. \tag{14}$$

The information contained in (10) and (12) is often written as

$$\lambda_1(A + \epsilon B) = \lambda_1 \; + \; \epsilon\langle x_1, Bx_1\rangle \; + \; o(\epsilon), \tag{15}$$

where x_1 is the normalized eigenvector corresponding to the simple eigenvalue λ_1 of A. More generally, when λ_1 is not a simple eigenvalue, we have for small ϵ

$$\lambda_1(A + \epsilon B) = \lambda_1 + \epsilon \max_{x\in M,\ ||x||=1} \langle x, Bx\rangle \; + \; o(\epsilon), \tag{16}$$

where M is the eigenspace corresponding to the eigenvalue λ_1 of A. See, e.g., (3.8) in [2].

Now, for any matrix A, consider the function

$$g(t) = \varphi(t)^2 = ||e^{tA}||^2 = \lambda_1(e^{tA}\; e^{tA^*}).$$

Then

$$\begin{aligned}
\dot g(t) &= \lim_{h\to 0^+} \frac{1}{h} \; \Big[\lambda_1(e^{tA}(I + hA)\; (I + hA^*)\; e^{tA^*}) - \lambda_1(e^{tA}\; e^{tA^*})\Big] \\
&= \lim_{h\to 0^+} \frac{1}{h} \; \Big[\lambda_1(e^{tA}\; e^{tA^*} + h\; e^{tA}(A + A^*)\; e^{tA^*}) - \lambda_1(e^{tA}\; e^{tA^*})\Big] \\
&= \max_{x\in M(t),\ ||x||=1} \langle x, e^{tA}(A + A^*)e^{tA^*}\; x\rangle, \qquad (17)
\end{aligned}$$

where $M(t)$ is the eigenspace of $e^{tA}\; e^{tA^*}$ corresponding to the largest eigenvalue.

When $t = 0$, this is the entire space $\mathbb{C}^n$, and hence

$$\dot g(0) = \max_{||x||=1} \langle x, (A + A^*)x\rangle \; = \lambda_1(A + A^*). \tag{18}$$

Since

$$\dot\varphi(t) = \frac{\dot g(t)}{2\varphi(t)}, \tag{19}$$

this gives the known result $\mu(A) = \dot\varphi(0) = \lambda_1(\frac{A+A^*}{2})$.

To calculate the second and the third derivatives we need the following lemma.

LEMMA 2. *Let x be a (normalized) eigenvector of $A + A^*$. Then*

$$\langle x, A^*\; f(A + A^*)Ax\rangle = \langle x, A\; f(A + A^*)A^*x\rangle \tag{20}$$

for every function f. In particular,

$$\langle x, A^*Ax\rangle = \langle x, AA^*x\rangle, \tag{21}$$

$$\langle x, A^*(A + A^*)Ax\rangle = \langle x, A(A + A^*)A^*x\rangle. \tag{22}$$

Proof. Choose an orthonormal basis consisting of eigenvectors of $A + A^*$ starting with x. Let $A = B + iC$, where $B = \frac{1}{2}(A + A^*)$, $C = \frac{1}{2i}(A - A^*)$. In the basis we have chosen, let $b_1, \ldots, b_n$ be the diagonal entries of B, and let a_{ij} be the entries of A. The two sides of (20) are the $(1,1)$ entries of the matrices $A^* f(A + A^*)A$ and $A f(A + A^*)A^*$, respectively. A simple calculation shows that each of them is equal to $\sum_{j=1}^n f(2b_j)\,|a_{1j}|^2$. □

Proof of Theorem 1. We have

$$g(h) = ||e^{hA}||^2 = \lambda_1(e^{hA}\, e^{hA^*}) = \lambda_1\left[\left(I + hA + \frac{h^2}{2}A^2\right)\left(I + hA^* + \frac{h^2}{2}A^{*2}\right)\right] + o(h^2)$$
$$= 1 + h\,\lambda_1\left[(A + A^*) + \frac{h}{2}(2AA^* + A^2 + A^{*2})\right] + o(h^2).$$

Using (16), we get from this

$$g(h) = 1 + h\,\lambda_1(A + A^*) + \frac{h^2}{2}\max_{x\in M,\ ||x||=1}\langle x, (2AA^* + A^2 + A^{*2})x\rangle + o(h^2),$$

where M is the eigenspace of $A + A^*$ corresponding to its largest eigenvalue. Now using (21), we see that for every $x \in M$

$$\langle x, (2AA^* + A^2 + A^{*2})x\rangle = \langle x, (A + A^*)^2 x\rangle = \lambda_1^2(A + A^*).$$

This shows that

$$\ddot{g}(0) = \lambda_1^2(A + A^*). \tag{23}$$

Since $\ddot{g}(t) = 2\,\dot{\varphi}(t)^2 + 2\varphi(t)\,\ddot{\varphi}(t)$, we have $\ddot{\varphi}(0) = \frac{\ddot{g}(0) - 2\dot{\varphi}(0)^2}{2\varphi(0)}$. Substituting the values of $\ddot{g}(0)$ and $\dot{\varphi}(0)$ from (23) and (6), we get (7).

To study the third derivative, write out the expansion of $g(h)$ as $g(h) = 1 + h\,\lambda_1(\widetilde{A} + h\widetilde{B} + h^2\,\widetilde{C}) + o(h^3)$, where

(24)
$$\widetilde{A} = A + A^*,\ \widetilde{B} = \frac{1}{2}(A^2 + A^{*2} + 2AA^*),\ \widetilde{C} = \frac{1}{6}\{A^3 + A^{*3} + 3A(A + A^*)A^*\}.$$

From (9), (10), (12), and (14), we know that if $\lambda_1(\widetilde{A})$ is simple, then $\lambda_1(\widetilde{A} + h\widetilde{B} + h^2\widetilde{C}) = \lambda_1(\widetilde{A}) + h\,k_1 + h^2\,k_2 + o(h^2)$, where

$$k_2 = \langle x_1, \widetilde{C}x_1\rangle + \sum_{j=2}^{n}\frac{|\langle x_j, \widetilde{B}x_1\rangle|^2}{\lambda_1 - \lambda_j}, \tag{25}$$

λ_j being the eigenvalues of $\widetilde{A} = A + A^*$ and x_j the corresponding eigenvectors. To calculate the second term in (25), note that

$$2\langle x_i, \widetilde{B}x_j\rangle = \langle x_i, (A^2 + A^{*2} + 2AA^*)x_j\rangle$$
$$= \langle x_i, [(A + A^*)^2 + A(A + A^*) - (A + A^*)A]x_j\rangle$$
$$= \{\lambda_i^2\,\delta_{ij} + (\lambda_j - \lambda_i)\,\langle x_i, Ax_j\rangle\}.$$

Hence,

$$\sum_{j=2}^{n} \frac{|\langle x_j, \widetilde{B}x_1\rangle|^2}{\lambda_1 - \lambda_j} = \frac{1}{4} \sum_{j=2}^{n} (\lambda_1 - \lambda_j)\ |\langle x_j, Ax_1\rangle|^2. \tag{26}$$

To calculate the first term in (25), note that

$$\begin{aligned} 6\ \widetilde{C} &= (A + A^*)^3 + A(AA^* - A^*A) + (AA^* - A^*A)A^* \\ &\quad + [A(A + A^*)A^* - A^*(A + A^*)A]. \end{aligned} \tag{27}$$

If W is the term inside the square brackets in (27), then by (22)

$$\langle x_1, Wx_1\rangle = 0. \tag{28}$$

Furthermore, note that

$$\begin{aligned} \langle x_1, A(AA^* - A^*A)x_1\rangle &= \langle A^*x_1, [(A + A^*)A^* - A^*(A + A^*)]x_1\rangle \\ &= \langle A^*x_1, (A + A^* - \lambda_1\ I)A^*x_1\rangle \\ &= \left\langle A^*x_1, \left(\sum_{j=2}^{n} (\lambda_j - \lambda_1)x_jx_j^*\right) A^*x_1 \right\rangle \\ &= \sum_{j=2}^{n} (\lambda_j - \lambda_1)\ |\langle x_j, A^*x_1\rangle|^2 \\ &= \sum_{j=2}^{n} (\lambda_j - \lambda_1)\ |\langle x_j, Ax_1\rangle|^2. \end{aligned} \tag{29}$$

(In the last step we used the fact that x_j are eigenvectors of $A + A^*$).

This shows also that

$$\langle x_1, (AA^* - A^*A)A^*x_1\rangle = \sum_{j=2}^{n} (\lambda_j - \lambda_1)\ |\langle x_j, Ax_1\rangle|^2. \tag{30}$$

Equations (26)–(30) show that

$$6\langle x_1, \widetilde{C}x_1\rangle = \lambda_1^3 + 2 \sum_{j=2}^{n} (\lambda_j - \lambda_1)\ |\langle x_j, Ax_1\rangle|^2. \tag{31}$$

From (25), (26), and (31) we obtain

$$6\ k_2 = \lambda_1^3 - \frac{1}{2} \sum_{j=2}^{n} (\lambda_1 - \lambda_j)\ |\langle x_j, Ax_1\rangle|^2. \tag{32}$$

This is then the value of $\ddot{g}(0)$. Since

$$\dddot{\varphi}(0) = \frac{\dddot{g}(0) - 6\ \dot{\varphi}(0)\ddot{\varphi}(0)}{2\varphi(0)},$$

we obtain equality (8) from the expressions already derived for $\dot{\varphi}(0)$ and $\ddot{\varphi}(0)$. ∎

3. Remarks.

1. We have proved (7) without the assumption that λ_1 is a simple eigenvalue of $A + A^*$. The proof is facilitated by the first order expansion (16). We do not know of an analogous second order expansion when λ_1 is a multiple eigenvalue. This compels us to assume λ_1 is simple while proving (8). We believe this assumption is not necessary.

2. Inequality (5) says $\varphi(t) \le e^{\dot\varphi(0)t}$. Because of (6) and (7), we know that $e^{\dot\varphi(0)t} - \varphi(t) = O(t^3)$, and (8) tells us that no further improvement is possible in general.

3. When the maximum eigenvalue of $A + A^*$ is simple, we can get conditions for equality in (5) using our result (8).

PROPOSITION 3. *Suppose $\lambda_1(A+A^*)$ is a simple eigenvalue of $A+A^*$. Then the following conditions are equivalent:*

(i) $||e^{tA}|| = e^{\mu(A)t}$ *for all* $t \ge 0$.

(ii) $||e^{hA}|| = e^{\mu(A)h}$ *for some* $h > 0$.

(iii) *The eigenvector x_1 of $A + A^*$ corresponding to λ_1 is also an eigenvector of A.*

Proof. Clearly (i) $\Rightarrow$ (ii). If (ii) holds for some $h > 0$, then for all natural numbers m

$$||e^{h/m\ A}|| = e^{\mu(A)h/m}$$

because of submultiplicativity of the norm. Since $\dot\varphi(0) = \mu(A)$, $\ddot\varphi(0) = \mu(A)^2$, we have from (8)

$$\sum_{j=2}^{n} (\lambda_1 - \lambda_j)\ |\langle x_j, Ax_1\rangle|^2 = 0.$$

Since $\lambda_j \ne \lambda_1$ for $j \ge 2$, this implies $\langle x_j, Ax_1\rangle = 0$. Hence $A\ x_1$ is a multiple of x_1. Thus statement (iii) is true if (ii) is.

Now suppose (iii) holds. If $Ax_1 = \lambda x_1$, then $A^*x_1 = \bar\lambda x_1$ and $\lambda_1 = \lambda + \bar\lambda$. In the orthonormal basis $x_1, \ldots, x_n$, we can write

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & A_1 \end{bmatrix}.$$

Note that $\mu(A_1) \le \mu(A) = \lambda_1/2 = \text{Re } \lambda$. Hence

$$||e^{tA}|| = \max\ (|e^{t\lambda}|,\ ||e^{tA_1}||) = e^{\mu(A)t}.$$

Thus (i) is true if (iii) is. □

Acknowledgments. The second author thanks the National Board for Higher Mathematics (India) and the Indian Statistical Institute for sponsoring a visit in February, 2001 when this work was done. He thanks the ISI, New Delhi, for its hospitality. Discussions with Tirthankar Bhattacharyya, Bangalore, which led to the last proposition, are gratefully acknowledged.

REFERENCES

[1] G. DAHLQUIST, *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations*, Kungl. Tekn. Högsk. Handl. Stockholm 130, Stockholm, 1959.

[2] J.-B. Hiriart-Urruty and D. Ye, *Sensitivity analysis of all eigenvalues of a symmetric matrix*, Numer. Math., 70 (1995), pp. 45–72.
[3] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.
[4] L. Kohaupt, *Second logarithmic derivative of a complex matrix in the Chebyshev norm*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 382–389.
[5] S. M. Lozinskii, *Error estimates for the numerical integration of ordinary differential equations* I, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958), pp. 52–90 (in Russian).
[6] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 669–693

A SUPERFAST TOEPLITZ SOLVER WITH IMPROVED NUMERICAL STABILITY*

MICHAEL STEWART†

Abstract. This paper describes a new $O(n\log^3(n))$ solver for the positive definite Toeplitz system $Tx = b$. Instead of computing generators for the inverse of T, the new algorithm adjoins b to T and applies a superfast Schur algorithm to the resulting augmented matrix. The generators of this augmented matrix and its Schur complements are used by a divide-and-conquer block back-substitution routine to complete the solution of the system. The goal is to avoid the well-known numerical instability inherent in explicit inversion. Experiments suggest that the algorithm is backward stable in most cases.

Key words. Toeplitz matrix, Schur algorithm

AMS subject classification. 65F05

DOI. 10.1137/S089547980241791X

1. Background.

We start with the positive definite Toeplitz matrix

$$T = \begin{bmatrix} t_0 & t_1 & \cdots & \cdots & t_{m-1} \\ \bar{t}_1 & t_0 & t_1 & & \vdots \\ \vdots & \bar{t}_1 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & t_1 \\ \bar{t}_{m-1} & \cdots & \cdots & \bar{t}_1 & t_0 \end{bmatrix} \in \mathbb{C}^{m\times m}$$

and the system of equations $Tx = b$. There are several classes of algorithms for solving such systems: these include slow algorithms requiring $O(n^3)$ unstructured matrix computations, fast $O(n^2)$ algorithms that exploit the Toeplitz structure, and superfast algorithms that achieve a complexity strictly less than $O(n^2)$. Examples of fast algorithms include the Schur and Levinson algorithms. Superfast algorithms have been developed in [3, 5, 10, 1, 7, 2].

One way to view the related approaches of [10, 1, 7, 2] is as a divide-and-conquer variant of the $O(n^2)$ Schur algorithm with fast polynomial multiplication via the FFT used to extend computations from submatrices and Schur complements to the full matrix T. The underlying Schur algorithm, along with several generalizations, is numerically stable [4, 15, 6], but it has not been shown that this stability extends to the superfast Schur algorithm. In fact, the proposed application of the algorithm to linear systems involves computing generators of T^{-1} and then forming $T^{-1}b$ using the FFT. Numerical methods based on explicit inversion are usually unstable [9]. Experiments presented in section 5 show that the superfast Schur algorithm is no exception; it is not a backward stable algorithm.

We will propose an alternative method that parallels the conventional and stable method of triangular factorization and back-substitution. Instead of inverting T we

*Received by the editors November 13, 2002; accepted for publication (in revised form) by L. Reichel May 6, 2003; published electronically December 17, 2003.
http://www.siam.org/journals/simax/25-3/41791.html
†Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303 (mstewart@mathstat.gsu.edu).

will transform the system $Tx = b$ to

$$\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} - T_{12}^{\mathrm{H}} T_{11}^{-1} T_{12} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 - T_{12}^{\mathrm{H}} T_{11}^{-1} b_1 \end{bmatrix} \tag{1}$$

and successively solve the two smaller systems

$$(T_{22} - T_{12}^{\mathrm{H}} T_{11}^{-1} T_{12}) x_2 = b_2 - T_{12}^{\mathrm{H}} T_{11}^{-1} b_1, \qquad T_{11} x_1 = b_1 - T_{12} x_2. \tag{2}$$

When dividing the system $Tx = b$ into two systems, it will simplify the presentation to assume that we have divided them in half. Thus T_{11} is $m/2 \times m/2$. We will assume that T so partitioned.

A superfast Schur algorithm applied to the augmented matrix formed from T and b can be used to compute the triangular system (1). The block back-substitution (2) is nothing more than the solution of two smaller Toeplitz-like systems that can be combined to solve the full system using a divide-and-conquer procedure. The multiplication $T_{12}x_2$ can be performed using the FFT.

The resulting algorithm avoids the suspect step of multiplication by T^{-1} but at the cost of increasing the complexity from $O(n \log^2(n))$ to $O(n \log^3(n))$ floating point operations. The increase in complexity occurs because the second system of (2) involves a modified right-hand side $b_1 - T_{12}x_2$ and consequently a modified augmented matrix. The new augmented matrix requires its own $O(n \log^2(n))$ superfast Schur factorization so that the overall procedure is $O(n \log^3(n))$.

The algorithm of [2] is the model for the derivation of the new algorithm as well as the benchmark for evaluating stability and efficiency. In the remainder of this section we will describe both the algorithm of [2] and the generalized Schur algorithm for a matrix with arbitrary displacement rank. In section 2 we show how the same divide-and-conquer idea can be applied to an augmented system that incorporates the right-hand side vector b. In section 3 we show how the information computed by a superfast block triangularization of the augmented matrix can be used to solve the system $Tx = b$ without the need for explicit matrix inversion. In section 4 we evaluate the computational complexity of the algorithm. In section 5 we present the results of numerical experiments that demonstrate the improved stability of the algorithm. Finally, in section 6 we make some observations on the possibility of a proof of numerical stability and compare the new method to another stabilized superfast algorithm.

1.1. The generalized Schur algorithm. A Toeplitz matrix T has an indefinite rank 2 *displacement*

$$T - ZTZ^{\mathrm{H}} = Y \Sigma Y^{\mathrm{H}}, \tag{3}$$

where Z is the downshift matrix, $[Z]_{ij} = 1$ if $i - j = 1$ and $[Z]_{ij} = 0$ otherwise, $\Sigma = 1 \oplus -1$, and

$$Y^{\mathrm{H}} = \begin{bmatrix} \sqrt{t_0} & t_1/\sqrt{t_0} & t_2/\sqrt{t_0} & \cdots & t_{n-1}/\sqrt{t_0} \\ 0 & t_1/\sqrt{t_0} & t_2/\sqrt{t_0} & \cdots & t_{n-1}/\sqrt{t_0} \end{bmatrix}.$$

Equation (3) is called a *displacement equation.* The matrix Σ is the *signature* matrix and Y is the *generator* matrix for T. Any matrix for which the displacement has rank significantly lower than n is *Toeplitz-like.*

The generators of a Toeplitz-like matrix are not unique. Given a generator matrix Y and a matrix H satisfying $H\Sigma H^{\mathrm{H}} = \Sigma$, we have

$$(YH)\Sigma(YH)^{\mathrm{H}} = Y(H\Sigma H^{\mathrm{H}})Y^{\mathrm{H}} = Y\Sigma Y^{\mathrm{H}}$$

so that YH is also a generator matrix for T. For general $\Sigma = I_p \oplus -I_q$, matrices H satisfying $H\Sigma H^{\mathrm{H}} = \Sigma$ are known as Σ-unitary. In the particular case $\Sigma = 1 \oplus -1$, all Σ-unitary matrices have the form

$$H = \frac{1}{\sqrt{1-|\rho|^2}} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 1 & \overline{\rho} \\ \rho & 1 \end{bmatrix},$$

where $|a| = |b| = 1$, i.e., the Σ-unitary matrices are just the product of hyperbolic rotations and unitary diagonal matrices.

In section 2 we will need to consider rank $p+q$ displacements (3) with $\Sigma = I_p \oplus -I_q$. Among the more useful general Σ-unitary transformations are the block diagonal unitary matrices $U \oplus V$ and hyperbolic rotations,

$$\begin{bmatrix} I & & & \\ & \frac{1}{\sqrt{1-|\rho|^2}} & & \frac{\rho}{\sqrt{1-|\rho|^2}} \\ & & I & \\ & \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} & & \frac{1}{\sqrt{1-|\rho|^2}} \\ & & & & I \end{bmatrix},$$

in which the rotation acts on one index in the positive part of the signature and one in the negative. There is also a hyperbolic version of a Householder transformation [11, 14].

The product of Σ-unitary matrices can be shown to be Σ-unitary. Thus, in applying general Σ-unitary transformations, it is natural to decompose them into a product of hyperbolic rotations and block unitary transformations. In this paper we will make use of products of hyperbolic rotations and block diagonal plane rotations. However, we will use a somewhat nonstandard signature matrix $\Sigma = 1 \oplus -1 \oplus 1 \oplus -1$ so that the block diagonal unitary rotation matrices become

$$\begin{bmatrix} c_1 & 0 & -\overline{s_1} & 0 \\ 0 & c_2 & 0 & -\overline{s_2} \\ s_1 & 0 & c_1 & 0 \\ 0 & s_2 & 0 & c_2 \end{bmatrix},$$

where c_1 and c_2 are real and nonnegative and $c_1^2 + |s_1|^2 = c_2^2 + |s_2|^2 = 1$. The hyperbolic rotations have the form

$$\begin{bmatrix} \frac{1}{\sqrt{1-|\rho_1|^2}} & 0 & \frac{\rho_1}{\sqrt{1-|\rho_1|^2}} & 0 \\ 0 & \frac{1}{\sqrt{1-|\rho_2|^2}} & 0 & \frac{\rho_2}{\sqrt{1-|\rho_2|^2}} \\ \frac{\overline{\rho_1}}{\sqrt{1-|\rho_1|^2}} & 0 & \frac{1}{\sqrt{1-|\rho_1|^2}} & 0 \\ 0 & \frac{\overline{\rho_2}}{\sqrt{1-|\rho_2|^2}} & 0 & \frac{1}{\sqrt{1-|\rho_2|^2}} \end{bmatrix}.$$

The Schur algorithm is a fast ($O(n^2)$) algorithm for the Cholesky factorization of T. It achieves the reduction in computation by working with the generator matrix

instead of on the entire matrix T. Since we will need the generality in section 2, we will describe the generalized Schur algorithm for factorization of a displacement rank $p+q$ Toeplitz-like matrix.[1]

We start with a matrix

$$T = \begin{bmatrix} t_0 & t_{21}^{\mathrm{H}} \\ t_{21} & T_{22} \end{bmatrix}$$

satisfying (3), where $\Sigma = I_p \oplus -I_q$. The first step of the generalized Schur algorithm is to transform the matrix Y. We partition Y as

$$Y = \left[\begin{array}{cc|c} y_{11} & y_{12}^{\mathrm{H}} & y_{13}^{\mathrm{H}} \\ y_{21} & Y_{22} & Y_{23} \end{array}\right],$$

where y_{11} is a scalar and the vertical line marks the boundary between the first p columns and the last q. We then compute a Σ-unitary H as a product of plane rotations and hyperbolic rotations so that

$$\hat{Y} = YH = \left[\begin{array}{cc|c} \hat{y}_{11} & 0 & 0 \\ \hat{y}_{21} & \hat{Y}_{22} & \hat{Y}_{23} \end{array}\right].$$

Thus $\hat{Y}$ is a generator matrix in which only the leading element of the first row is nonzero. Such generators are said to be in *proper form*. The displacement equation (3) implies that

$$\begin{bmatrix} t_0 & t_{21}^{\mathrm{H}} \end{bmatrix} = \hat{y}_{11} \begin{bmatrix} \hat{y}_{11} & \hat{y}_{21}^{\mathrm{H}} \end{bmatrix}$$

so that the first row of $\hat{Y}^{\mathrm{H}}$ is the first row of the Cholesky factor of T. Furthermore, if we define

$$T_S = \begin{bmatrix} 0 & 0 \\ 0 & T_{22} - t_{21} t_0^{-1} t_{21}^{\mathrm{H}} \end{bmatrix}, \qquad Y_S = \begin{bmatrix} Z\hat{Y}(:,1) & \hat{Y}(:, 2:p+q) \end{bmatrix},$$

then

$$T_S - Z T_S Z^{\mathrm{H}} = Y_S \Sigma Y_S^{\mathrm{H}}.$$

Thus the zero-bordered Schur complement of T inherits the displacement structure of T and its generators are easily determined by the proper form generators for T. The generalized Schur algorithm repeats this process recursively on T_S with generator matrix Y_S to compute successively the rows of the Cholesky factor.

1.2. The superfast Schur algorithm. We will now give a description of the superfast Schur algorithm. The presentation here summarizes material from [1, 2]. The main idea behind speeding up the Schur algorithm is to represent the generators as polynomials and then to use fast polynomial multiplication via the FFT to implement

[1] In reference to Schur algorithms, the term "generalized" has been used with two distinct meanings. In the sense we are using it here, it refers to a fast algorithm that factors any matrix, not necessarily Toeplitz, satisfying a displacement equation with $\Sigma = I_p \oplus -I_q$. In [2], however, it refers to a superfast algorithm for solving ordinary Toeplitz systems—what we refer to here as the superfast Schur algorithm.

the generator transformations of the Schur algorithm. Suppose T is a Toeplitz-like matrix of displacement rank 2 with generators

$$Y = \begin{bmatrix} v_0 & w_0 \\ v_1 & w_1 \\ \vdots & \vdots \\ v_{n-1} & w_{n-1} \end{bmatrix}.$$

We define the polynomial generators

$$Y_0(z) = \begin{bmatrix} v_0(z) & w_0(z) \end{bmatrix},$$

where

$$v_0(z) = v_0 + v_1 z + v_2 z^2 + \cdots + v_{n-1} z^{n-1}$$

and

$$w_0(z) = w_0 + w_1 z + w_2 z^2 + \cdots + w_{n-1} z^{n-1}.$$

Multiplication by z replaces the shift of the first column of Y so that the first step of the Schur algorithm becomes

$$\begin{bmatrix} v_1(z) & w_1(z) \end{bmatrix} = \frac{1}{\sqrt{1-|\rho_1|^2}} \begin{bmatrix} v_0(z) & w_0(z) \end{bmatrix} \begin{bmatrix} 1 & \rho_1 \\ \overline{\rho_1} & 1 \end{bmatrix} \begin{bmatrix} z & 0 \\ 0 & 1 \end{bmatrix}.$$

At step k of the Schur algorithm we have

$$\begin{bmatrix} v_k(z) & w_k(z) \end{bmatrix} = \frac{1}{\sqrt{1-|\rho_k|^2}} \begin{bmatrix} v_{k-1}(z) & w_{k-1}(z) \end{bmatrix} \begin{bmatrix} 1 & \rho_k \\ \overline{\rho_k} & 1 \end{bmatrix} \begin{bmatrix} z & 0 \\ 0 & 1 \end{bmatrix}$$

so that

$$\begin{bmatrix} v_k(z) & w_k(z) \end{bmatrix} = \begin{bmatrix} v_0(z) & w_0(z) \end{bmatrix} \begin{bmatrix} a_k^{(0)}(z) & b_k^{(0)}(z) \\ \tilde{b}_k^{(0)}(z) & \tilde{a}_k^{(0)}(z) \end{bmatrix},$$

where

$$\begin{bmatrix} a_k^{(0)}(z) & b_k^{(0)}(z) \\ \tilde{b}_k^{(0)}(z) & \tilde{a}_k^{(0)}(z) \end{bmatrix} = \left(\prod_{j=1}^{k} \frac{1}{\sqrt{1-|\rho_j|^2}} \right) \begin{bmatrix} z & \rho_1 \\ z\overline{\rho_1} & 1 \end{bmatrix} \begin{bmatrix} z & \rho_2 \\ z\overline{\rho_2} & 1 \end{bmatrix} \cdots \begin{bmatrix} z & \rho_k \\ z\overline{\rho_k} & 1 \end{bmatrix}.$$

It can be shown inductively that

$$\tilde{a}_k^{(0)}(z) = z^k \overline{a}_k^{(0)}(1/z), \qquad \tilde{b}_k^{(0)}(z) = z^k \overline{b}_k^{(0)}(1/z).$$

Hence the product resulting from k steps of the Schur algorithm can be represented by just the *Schur polynomials* $a_k^{(0)}(z)$ and $b_k^{(0)}(z)$.

To represent an arbitrary sequence of k consecutive steps of the Schur algorithm we define

$$\begin{bmatrix} a_k^{(l)}(z) & b_k^{(l)}(z) \\ \tilde{b}_k^{(l)}(z) & \tilde{a}_k^{(l)}(z) \end{bmatrix} = \left(\prod_{j=l+1}^{l+k} \frac{1}{\sqrt{1-|\rho_j|^2}} \right) \begin{bmatrix} z & \rho_{l+1} \\ z\overline{\rho_{l+1}} & 1 \end{bmatrix} \begin{bmatrix} z & \rho_{l+2} \\ z\overline{\rho_2} & 1 \end{bmatrix} \cdots \begin{bmatrix} z & \rho_{l+k} \\ z\overline{\rho_{l+k}} & 1 \end{bmatrix}$$

so that

$$\begin{bmatrix} v_{l+k}(z) & w_{l+k}(z) \end{bmatrix} = \begin{bmatrix} v_l(z) & w_l(z) \end{bmatrix} \begin{bmatrix} a_k^{(l)}(z) & b_k^{(l)}(z) \\ \tilde{b}_k^{(l)}(z) & \tilde{a}_k^{(l)}(z) \end{bmatrix}. \tag{4}$$

Thus $a_k^{(l)}(z)$ and $b_k^{(l)}(z)$ are Schur polynomials that apply k steps of the Schur algorithm, transforming the generator polynomials $v_l(z)$ and $w_l(z)$ into $v_{l+k}(z)$ and $w_{l+k}(z)$. Since they are formed from products of elementary hyperbolic rotations in exactly the same way as $a_k(z)$ and $b_k(z)$, they also satisfy

$$\tilde{a}_k^{(l)}(z) = z^k \overline{a}_k^{(l)}(1/z), \qquad \tilde{b}_k^{(l)}(z) = z^k \overline{b}_k^{(l)}(1/z).$$

Given the Schur polynomials, we can perform k steps of the Schur algorithm via the polynomial multiplication (4). If we use the FFT, the computational cost of the multiplication will be $O(n \log(n))$. The Schur polynomials can be computed using a divide-and-conquer procedure based on the doubling step

$$\begin{bmatrix} a_{2k}^{(0)}(z) & b_{2k}^{(0)}(z) \\ \tilde{b}_{2k}^{(0)}(z) & \tilde{a}_{2k}^{(0)}(z) \end{bmatrix} = \begin{bmatrix} a_k^{(0)}(z) & b_k^{(0)}(z) \\ \tilde{b}_k^{(0)}(z) & \tilde{a}_k^{(0)}(z) \end{bmatrix} \begin{bmatrix} a_k^{(k)}(z) & b_k^{(k)}(z) \\ \tilde{b}_k^{(k)}(z) & \tilde{a}_k^{(k)}(z) \end{bmatrix}. \tag{5}$$

This equation represents the multiplication of the polynomials for the first k steps of the Schur algorithm with those for the next k to get the polynomials for carrying out $2k$ steps. Again, (5) is just polynomial multiplication which can be carried out with the FFT in $O(n \log(n))$ operations.

Multiplication by the Schur polynomials in (4) increases the degree of the generator polynomials. Since the length of the generator vectors does not increase in the course of applying the Schur algorithm, the higher powers of z are not necessary for computing a factorization. Thus to save memory and computation we should truncate the generator polynomials. For

$$v(z) = v_0 + v_1 z + \cdots + v_{n-1} z^{n-1},$$

we let a superscript (k) for $k < n$ denote the truncation

$$v^{(k)}(z) = v_0 + v_1 z + \cdots + v_{k-1} z^{(k-1)}.$$

Note that this meaning for a superscript applies only to generator polynomials $v(z)$ and $w(z)$ but not to the Schur polynomials $a(z)$ and $b(z)$ for which the superscript has a completely different meaning.

Combining (5) with (4), we get a divide-and-conquer algorithm for computing the Schur polynomials $a_k^{(0)}(z)$ and $b_k^{(0)}(z)$.

`function` $[a(z),\, b(z)]$=`sfschur`$(v(z),\, w(z), n)$
  `if` $n > 1$ `then`
    $[a_{n/2}^{(0)}(z),\, b_{n/2}^{(0)}(z)]$ = `sfschur`$(v^{(n/2)}(z),\, w^{(n/2)}(z), n/2)$
    $v_{n/2}(z) = v(z) a_{n/2}^{(0)}(z) + w(z) \tilde{b}_{n/2}^{(0)}(z)$
    $w_{n/2}(z) = v(z) b_{n/2}^{(0)}(z) + w(z) \tilde{a}_{n/2}^{(0)}(z)$
    $[a_{n/2}^{(n/2)}(z),\, b_{n/2}^{(n/2)}(z)\,]$ = `sfschur`$(v_{n/2}^{(n/2)}(z),\, w_{n/2}^{(n/2)}(z), n/2)$
    $a(z) = a_{n/2}^{(0)}(z) a_{n/2}^{(n/2)}(z) + b_{n/2}^{(0)}(z) \tilde{b}_{n/2}^{(n/2)}(z)$
    $b(z) = a_{n/2}^{(0)}(z) b_{n/2}^{(n/2)}(z) + b_{n/2}^{(0)}(z) \tilde{a}_{n/2}^{(n/2)}(z)$

```
else
    ρ = −w(z)/v(z)
    a(z) = z/√(1 − |ρ|²)
    b(z) = ρ/√(1 − |ρ|²)
endif
```

The function `sfschur`() takes two generator polynomials of degree $n-1$ representing two generator vectors of length n. The length is passed as a separate parameter. The output polynomials $a(z)$ and $b(z)$ are the Schur polynomials $a_n^{(0)}(z)$ and $b_n^{(0)}(z)$ for applying n steps of the Schur algorithm. The computation is $O(n\log^2(n))$.

Since the recursive calls to `sfschur`() use the truncated generators $v^{(n/2)}(z)$ and $w^{(n/2)}(z)$, the problem size is halved with each level of depth in the recursion. In the termination case $n = 1$, only one easily computed hyperbolic rotation needs to be applied: if $n = 1$, then $v(z)$ and $w(z)$ are constants and the Schur algorithm reduces to the proper form transformation

$$\begin{bmatrix}\tilde{v}(z) & 0\end{bmatrix} = \frac{1}{\sqrt{1-|\rho|^2}}\begin{bmatrix}v(z) & w(z)\end{bmatrix}\begin{bmatrix}1 & \rho\\ \overline{\rho} & 1\end{bmatrix}$$

with $\rho = -w(z)/v(z)$ and

$$\begin{bmatrix}a(z) & b(z)\\ \tilde{b}(z) & \tilde{a}(z)\end{bmatrix} = \frac{1}{\sqrt{1-|\rho|^2}}\begin{bmatrix}z & \rho\\ z\overline{\rho} & 1\end{bmatrix}.$$

To solve a Toeplitz system, we pass the generator polynomials $v(z)$ and $w(z)$ to `sfschur`() to compute the Schur polynomials

$$\begin{bmatrix}a_n^{(0)}(z) & b_n^{(0)}(z)\end{bmatrix} = \texttt{sfschur}(v(z), w(z), n).$$

The inverse matrix T^{-1} is known to be Toeplitz-like. If

$$\phi(z) = \tilde{a}_n^{(0)}(z) + b_n^{(0)}(z), \qquad \tilde{\phi}(z) = a_n^{(0)}(z) + \tilde{b}_n^{(0)}(z),$$

then the pair of polynomials $\phi(z)$ and $\tilde{\phi}(z)$ are polynomial generators for T^{-1}. This fact is expressed by the well-known Gohberg–Semencul formula. Since `sfschur`() gives generators for T^{-1}, we can use the FFT to apply T^{-1} to the right-hand side vector b to solve the system $Tx = b$ using $O(n\log^2(n))$ operations. Since the multiplication by T^{-1} is $O(n\log(n))$, the additional cost of solving the system with a different right-hand side is $O(n\log(n))$. More details on the use of this algorithm for solving systems can be found in [1, 2].

2. The augmented system. Instead of factoring just T we will generalize the superfast Schur algorithm to the augmented system

$$M = \begin{bmatrix}T & b\\ b^{\mathrm{H}} & 1\end{bmatrix}.$$

Suppose T is positive definite of displacement rank 2 and satisfies the displacement equation (3). We extend the displacement equation to

$$M - \begin{bmatrix}Z & 0\\ 0 & 0\end{bmatrix} M \begin{bmatrix}Z^{\mathrm{H}} & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}Y\Sigma Y^{\mathrm{H}} & b\\ b^{\mathrm{H}} & 1\end{bmatrix}.$$

This displacement has a factorization

$$M - \begin{bmatrix} Z & 0 \\ 0 & 0 \end{bmatrix} M \begin{bmatrix} Z^{\mathrm{H}} & 0 \\ 0 & 0 \end{bmatrix} = \hat{Y} \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma \end{bmatrix} \hat{Y}^{\mathrm{H}},$$

where

$$\hat{Y} = \begin{bmatrix} Y & b & b \\ 0 & 1 & 0 \end{bmatrix}. \tag{6}$$

The generalized Schur algorithm can use any Σ-unitary transformation to put $\hat{Y}$ into proper form. However, in generalizing the superfast Schur algorithm, it is convenient to use a special transformation that preserves the structure within the generator matrix. In particular, we will use transformations that keep the generator matrices of T and its Schur complements as submatrices of the generator matrices of M and its Schur complements. The resulting algorithm extends but does not otherwise alter the superfast Schur algorithm; it computes every polynomial computed by the superfast Schur algorithm in exactly the same manner in which it is computed by the superfast Schur algorithm. The right-hand side part of the augmented matrix is handled through the addition of two new Schur polynomials and one new generator polynomial. These additional polynomials depend on the factorization of T, but the computations relating to T, its Schur complements, and their generators do not in any way depend on the new polynomials. The use of structured generator transformations reduces the total amount of computation, while also making available generators of both T and M.

We assume that at some point in the application of the generalized Schur algorithm to M we have generators of the form

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ v_1 & w_1 & b_1 & b_1 \\ v_2 & w_2 & b_2 & b_2 \\ 0 & 0 & \delta & \delta - 1/\delta \end{bmatrix},$$

where v_1, w_1, b_1, and δ are scalars and v_1 is real and positive. Note that the initial generators (6) are of this form but with no leading zero rows and with $\delta = 1$. Note also that the matrix

$$\begin{bmatrix} v_1 & w_1 \\ v_2 & w_2 \end{bmatrix}$$

is a generator matrix for the Toeplitz-like leading block of M. Thus positive definiteness of T guarantees that $v_1 \neq 0$ and if $\rho = -w_1/v_1$, then $|\rho| < 1$.

We propose a structured transformation to proper form

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ v_1 & w_1 & b_1 & b_1 \\ v_2 & w_2 & b_2 & b_2 \\ 0 & 0 & \delta & \delta - 1/\delta \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1-|\rho|^2}} & \frac{\rho}{\sqrt{1-|\rho|^2}} & 0 & 0 \\ \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{1}{\sqrt{1-|\rho|^2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c & 0 & -\overline{s} & 0 \\ 0 & 1 & 0 & 0 \\ s & 0 & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\cdot \begin{bmatrix} \frac{1}{c} & 0 & 0 & \frac{-\overline{s}}{c} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{-s}{c} & 0 & 0 & \frac{1}{c} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \tilde{v}_1 & 0 & 0 & 0 \\ \tilde{v}_2 & \tilde{w}_2 & \tilde{b}_2 & \tilde{b}_2 \\ \tilde{\delta}_1 & 0 & \tilde{\delta} & \tilde{\delta} - 1/\tilde{\delta} \end{bmatrix}.$$

Clearly if $c = \sqrt{1-|s|^2}$, then each of these transformations is Σ-unitary. Multiplying them together gives

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ v_1 & w_1 & b_1 & b_1 \\ v_2 & w_2 & b_2 & b_2 \\ 0 & 0 & \delta & \delta - 1/\delta \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{1-|\rho|^2}} & \frac{\rho}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} \\ \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{1}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s\rho}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s\rho}}{\sqrt{1-|\rho|^2}} \\ \frac{s}{c} & 0 & c & \frac{-|s|^2}{c} \\ \frac{-s}{c} & 0 & 0 & \frac{1}{c} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ \tilde{v}_1 & 0 & 0 & 0 \\ \tilde{v}_2 & \tilde{w}_2 & \tilde{b}_2 & \tilde{b}_2 \\ \tilde{\delta}_1 & 0 & \tilde{\delta} & \tilde{\delta} - 1/\tilde{\delta} \end{bmatrix}. \tag{7}$$

We will show that s and ρ can be chosen so that the transformed generators have the form shown in (7). Let

$$\rho = -\frac{w_1}{v_1},$$

where $|\rho| < 1$ and

$$s = \frac{\overline{b_1}}{\sqrt{v_1^2(1-|\rho|^2) + |b_1|^2}}, \qquad c = \sqrt{1-|s|^2}.$$

Since $|\rho| \neq 1$ and $v_1 \neq 0$, $|s| < 1$ so that $0 < c \leq 1$.

The value of ρ has been chosen to introduce a zero into the w_1 element of the generator matrix. In addition to this,

$$\tilde{v}_1 = v_1 \frac{1}{\sqrt{1-|\rho|^2}} + w_1 \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} = v_1 \sqrt{1-|\rho|^2} > 0.$$

The cosine c is

$$c = \sqrt{1-|s|^2} = \sqrt{\frac{v_1^2(1-|\rho|^2)}{v_1^2(1-|\rho|^2) + |b_1|^2}} = \frac{\tilde{v}_1}{\sqrt{v_1^2(1-|\rho|^2) + |b_1|^2}}.$$

The first b_1 transforms to

$$cb_1 - \overline{s}\left(v_1 \frac{1}{\sqrt{1-|\rho|^2}} + w_1 \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} \right) = cb_1 - \overline{s}\tilde{v}_1 = 0,$$

and the second transforms to

$$\frac{1}{c}b_1 - \frac{|s|^2}{c}b_1 - \overline{s}\left(v_1 \frac{1}{\sqrt{1-|\rho|^2}} + w_1 \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} \right) = cb_1 - \overline{s}\tilde{v}_1 = 0.$$

Similarly the first and second b_2 element are transformed to the same vector

$$\tilde{b}_2 = cb_2 - \overline{s}\left(v_2 \frac{1}{\sqrt{1-|\rho|^2}} + w_2 \frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} \right).$$

The presence of the zero element on the bottom row is obvious. The element corresponding to δ is

$$\tilde{\delta} = c\delta. \tag{8}$$

Finally, the element corresponding to $\delta - 1/\delta$ is

$$-\frac{|s|^2}{c}\delta + (\delta - 1/\delta)\frac{1}{c} = \frac{1-c^2}{c}\delta + -\frac{1}{\delta c} = \delta c - \frac{1}{\delta c} = \tilde{\delta} - \frac{1}{\tilde{\delta}}.$$

This verifies that all elements of the transformed generator matrix are as shown. In particular, we have shown that if the generators are

$$\begin{bmatrix} v & w & b & b \\ 0 & 0 & \delta & \delta - 1/\delta \end{bmatrix}, \tag{9}$$

then (7) puts the transformed generators into the form

$$\begin{bmatrix} \tilde{v} & \tilde{w} & \tilde{b} & \tilde{b} \\ \tilde{\delta}_1 & 0 & \tilde{\delta} & \tilde{\delta} - 1/\tilde{\delta} \end{bmatrix}. \tag{10}$$

To show that at every stage of the Schur algorithm the generators have the form (9), we first note that the initial generators (6) are of this form. As shown in (7), the generator transformations preserve the structure except for the addition of a nonzero δ_1. However, in computing the generators of a Schur complement, the first column of the transformed generator matrix (10) is multiplied by

$$\begin{bmatrix} Z & 0 \\ 0 & 0 \end{bmatrix},$$

which zeros the element δ_1.

In addition to the preservation of the pattern of repeated vectors in the generators, (7) implies that, for the first two columns of the transformed generator matrix,

$$\begin{bmatrix} 0 & 0 \\ \tilde{v}_1 & 0 \\ \tilde{v}_2 & \tilde{w}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ v_1 & w_1 \\ v_2 & w_2 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \overline{\rho} & 1 \end{bmatrix} / \sqrt{1 - |\rho|^2}.$$

These two columns are no different than if we had simply applied the 2×2 hyperbolic rotation from the ordinary Schur algorithm to the first two columns of the generator matrix. The result is that, in applying this form of the generalized Schur algorithm, the first two columns of the generator matrix will be generators of T and its Schur complements.

As in the superfast Schur algorithm, the generators and generator transformations can be represented by polynomials. We represent all but the last row of the generator matrix by three polynomials

$$\begin{bmatrix} v_k(z) & w_k(z) & \beta_k(z) & \beta_k(z) \end{bmatrix} = \begin{bmatrix} 1 & z & \cdots & z^{n-1} \end{bmatrix} \begin{bmatrix} v & w & b & b \end{bmatrix}$$

so that

$$\begin{bmatrix} v_k(z) & w_k(z) & \beta_k(z) & \beta_k(z) \end{bmatrix} = \begin{bmatrix} v_{k-1}(z) & w_{k-1}(z) & \beta_{k-1}(z) & \beta_{k-1}(z) \end{bmatrix} \cdot \begin{bmatrix} z\frac{1}{\sqrt{1-|\rho|^2}} & \frac{\rho}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} \\ z\frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{1}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}\rho}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}\rho}{\sqrt{1-|\rho|^2}} \\ z\frac{s}{c} & 0 & c & \frac{-|s|^2}{c} \\ z\frac{-s}{c} & 0 & 0 & \frac{1}{c} \end{bmatrix}.$$

The scalar δ will not be incorporated into any polynomial; it will be stored and kept track of separately from $v(z)$, $w(z)$, and $\beta(z)$.

As in the displacement rank 2 case, accumulating the product of such transformations results in a special structure. We will show that k steps of the polynomial version of the generalized Schur algorithm have the form

$$\begin{bmatrix} v_{l+k}(z) & w_{l+k}(z) & \beta_{l+k}(z) & \beta_{l+k}(z) \end{bmatrix} = \begin{bmatrix} v_l(z) & w_l(z) & \beta_l(z) & \beta_l(z) \end{bmatrix} \cdot \begin{bmatrix} a_k^{(l)}(z) & b_k^{(l)}(z) & c_k^{(l)}(z) & c_k^{(l)}(z) \\ \tilde{b}_k^{(l)}(z) & \tilde{a}_k^{(l)}(z) & d_k^{(l)}(z) & d_k^{(l)}(z) \\ e_k^{(l)}(z) & f_k^{(l)}(z) & g_k^{(l)}(z) & g_k^{(l)}(z) - 1/g_k^{(l)}(0) \\ -e_k^{(l)}(z) & -f_k^{(l)}(z) & -g_k^{(l)}(z) + g_k^{(l)}(0) & -g_k^{(l)}(z) + 1/g_k^{(l)}(0) + g_k^{(l)}(0) \end{bmatrix} \tag{11}$$

This is verified inductively in the following theorem.

THEOREM 1. *The product of k matrices of the form*

$$\begin{bmatrix} z\frac{1}{\sqrt{1-|\rho|^2}} & \frac{\rho}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} \\ z\frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{1}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}\overline{\rho}}{\sqrt{1-|\rho|^2}} \\ z\frac{s}{c} & 0 & c & \frac{-|s|^2}{c} \\ z\frac{-s}{c} & 0 & 0 & \frac{1}{c} \end{bmatrix} \tag{12}$$

has the form

$$\begin{bmatrix} a_k(z) & b_k(z) & c_k(z) & c_k(z) \\ \tilde{b}_k(z) & \tilde{a}_k(z) & d_k(z) & d_k(z) \\ e_k(z) & f_k(z) & g_k(z) & g_k(z) - 1/g_k(0) \\ -e_k(z) & -f_k(z) & -g_k(z) + g_k(0) & -g_k(z) + 1/g_k(0) + g_k(0) \end{bmatrix},$$

where $e(0) = f(0) = 0$ and where $\tilde{a}(z) = z^k\overline{a}(1/z)$ and $\tilde{b}(z) = z^k\overline{b}(1/z)$.

Proof. We note that the single transformation (12) has the specified form. We assume that the theorem holds for a product of $k-1$ transformations and write

$$\begin{bmatrix} a_k(z) & b_k(z) & c_k(z) & c_k(z) \\ \tilde{b}_k(z) & \tilde{a}_k(z) & d_k(z) & d_k(z) \\ e_k(z) & f_k(z) & g_k(z) & g_k(z) - 1/g_k(0) \\ -e_k(z) & -f_k(z) & -g_k(z) + g_k(0) & -g_k(z) + 1/g_k(0) + g_k(0) \end{bmatrix} = \begin{bmatrix} a_{k-1}(z) & b_{k-1}(z) & c_{k-1}(z) & c_{k-1}(z) \\ \tilde{b}_{k-1}(z) & \tilde{a}_{k-1}(z) & d_{k-1}(z) & d_{k-1}(z) \\ e_{k-1}(z) & f_{k-1}(z) & g_{k-1}(z) & g_{k-1}(z) - 1/g_{k-1}(0) \\ -e_{k-1}(z) & -f_{k-1}(z) & -g_{k-1}(z) + g_{k-1}(0) & -g_{k-1}(z) + 1/g_{k-1}(0) + g_{k-1}(0) \end{bmatrix} \cdot \begin{bmatrix} z\frac{1}{\sqrt{1-|\rho|^2}} & \frac{\rho}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} \\ z\frac{\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{1}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}\overline{\rho}}{\sqrt{1-|\rho|^2}} & \frac{-\overline{s}\overline{\rho}}{\sqrt{1-|\rho|^2}} \\ z\frac{s}{c} & 0 & c & \frac{-|s|^2}{c} \\ z\frac{-s}{c} & 0 & 0 & \frac{1}{c} \end{bmatrix}.$$

This gives two relations for each of the polynomials $e_k(z)$, $f_k(z)$, $c_k(z)$, and $d_k(z)$. The relations for $e_k(z)$ and $-e_k(z)$ (obtained by computing the $(3,1)$ and $(4,1)$ elements

of the left-hand side) are

$$e_k(z) = \frac{z}{\sqrt{1-|\rho|^2}} e_{k-1}(z) + \frac{z\overline{\rho}}{\sqrt{1-|\rho|^2}} f_{k-1}(z) + \frac{zs}{c} g_{k-1}(z) - \frac{zs}{c} g_{k-1}(z) + \frac{zs}{c} \frac{1}{g_{k-1}(0)}$$

and

$$-e_k(z) = -\frac{z}{\sqrt{1-|\rho|^2}} e_{k-1}(z) - \frac{z\overline{\rho}}{\sqrt{1-|\rho|^2}} f_{k-1}(z) - \frac{zs}{c} g_{k-1}(z) + \frac{zs}{c} g_{k-1}(0) + \frac{zs}{c} g_{k-1}(z) - \frac{zs}{c} \frac{1}{g_{k-1}(0)} - \frac{zs}{c} g_{k-1}(0).$$

These two relations clearly define the same polynomial $e_k(z)$. Verification that the two relations for $f_k(z)$ give the same polynomial is similar, as is the verification for $c_k(z)$ and $d_k(z)$.

Every term in the expression for $e_k(z)$ has z as a factor. Thus $e_k(0) = 0$. Since

$$f_k(z) = e_{k-1}(z) \frac{\rho}{\sqrt{1-|\rho|^2}} + f_{k-1}(z) \frac{1}{\sqrt{1-|\rho|^2}},$$

we see that $f_k(0) = 0$ follows from $e_{k-1}(0) = 0$ and $f_{k-1}(0) = 0$.

The relations $\tilde{a}_k(z) = z^k \overline{a}_k(1/z)$ and $\tilde{b}_k(z) = z^k \overline{a}_k(1/z)$ follow from

$$\begin{bmatrix} a_k(z) & b_k(z) \\ \tilde{b}_k(z) & \tilde{a}_k(z) \end{bmatrix} = \begin{bmatrix} a_{k-1}(z) & b_{k-1}(z) \\ \tilde{b}_{k-1}(z) & \tilde{a}_{k-1}(z) \end{bmatrix} \begin{bmatrix} z & \rho \\ z\overline{\rho} & 1 \end{bmatrix} / \sqrt{1-|\rho|^2}$$

in exactly the same way as this result follows for the displacement rank 2 case.

To verify the identities for the polynomial $g_k(z)$, we note that if we define

$$h_k(z) = e_{k-1}(z) \frac{-\overline{s}}{\sqrt{1-|\rho|^2}} + f_{k-1}(z) \frac{-\overline{s\rho}}{\sqrt{1-|\rho|^2}},$$

then $h_k(0) = 0$ since $e_{k-1}(0) = f_{k-1}(0) = 0$. In terms of $h_k(z)$, the lower right 2×2 block is

$$\begin{bmatrix} g_k(z) & g_k(z) - 1/g_k(0) \\ -g_k(z) + g_k(0) & -g_k(z) + 1/g_k(0) + g_k(0) \end{bmatrix}$$
$$= \begin{bmatrix} h_k(z) + cg_{k-1}(z) & h_k(z) + cg_{k-1}(z) - 1/(cg_{k-1}(0)) \\ -h_k(z) - cg_{k-1}(z) + cg_{k-1}(0) & -h_k(z) - cg_{k-1}(z) + cg_{k-1}(0) + 1/(cg_{k-1}(0)) \end{bmatrix},$$

where

$$g_k(z) = h_k(z) + cg_{k-1}(z)$$

and

$$g_k(0) = cg_{k-1}(0) \tag{13}$$

since $h_k(0) = 0$. □

Having established the form of the generator transformations, we construct a superfast version of the Schur algorithm for the augmented matrix using a doubling relation analogous to (5):

$$
\begin{aligned}
&\begin{bmatrix} a_{2k}^{(0)}(z) & b_{2k}^{(0)}(z) & c_{2k}^{(0)}(z) & c_{2k}^{(0)}(z) \\ \tilde{b}_{2k}^{(0)}(z) & \tilde{a}_{2k}^{(0)}(z) & d_{2k}^{(0)}(z) & d_{2k}^{(0)}(z) \\ e_{2k}^{(0)}(z) & f_{2k}^{(0)}(z) & g_{2k}^{(0)}(z) & g_{2k}^{(0)}(z) - 1/g_{2k}^{(0)}(0) \\ -e_{2k}^{(0)}(z) & -f_{2k}^{(0)}(z) & -g_{2k}^{(0)}(z) + g_{2k}^{(0)}(0) & -g_{2k}^{(0)}(z) + 1/g_{2k}^{(0)}(0) + g_{2k}^{(0)}(0) \end{bmatrix} \\
&= \begin{bmatrix} a_{k}^{(0)}(z) & b_{k}^{(0)}(z) & c_{k}^{(0)}(z) & c_{k}^{(0)}(z) \\ \tilde{b}_{k}^{(0)}(z) & \tilde{a}_{k}^{(0)}(z) & d_{k}^{(0)}(z) & d_{k}^{(0)}(z) \\ e_{k}^{(0)}(z) & f_{k}^{(0)}(z) & g_{k}^{(0)}(z) & g_{k}^{(0)}(z) - 1/g_{k}^{(0)}(0) \\ -e_{k}^{(0)}(z) & -f_{k}^{(0)}(z) & -g_{k}^{(0)}(z) + g_{k}^{(0)}(0) & -g_{k}^{(0)}(z) + 1/g_{k}^{(0)}(0) + g_{k}^{(0)}(0) \end{bmatrix} \\
&\cdot \begin{bmatrix} a_{k}^{(k)}(z) & b_{k}^{(k)}(z) & c_{k}^{(k)}(z) & c_{k}^{(k)}(z) \\ \tilde{b}_{k}^{(k)}(z) & \tilde{a}_{k}^{(k)}(z) & d_{k}^{(k)}(z) & d_{k}^{(k)}(z) \\ e_{k}^{(k)}(z) & f_{k}^{(k)}(z) & g_{k}^{(k)}(z) & g_{k}^{(k)}(z) - 1/g_{k}^{(k)}(0) \\ -e_{k}^{(k)}(z) & -f_{k}^{(k)}(z) & -g_{k}^{(k)}(z) + g_{k}^{(k)}(0) & -g_{k}^{(k)}(z) + 1/g_{k}^{(k)}(0) + g_{k}^{(k)}(0) \end{bmatrix}. \qquad (14)
\end{aligned}
$$

It turns out that a complete algorithm for the solution of $Tx = b$ can be formulated without computing $e_k(z)$, $f_k(z)$, and $g_k(z)$. We will, however, need $a_k(z)$, $b_k(z)$, $c_k(z)$, $d_k(z)$, and the scalar $g_k(0)$. The updates from (14) that we will use include the Schur polynomial computation

$$
\begin{bmatrix} a_{2k}^{(0)}(z) & b_{2k}^{(0)}(z) \\ \tilde{b}_{2k}^{(0)}(z) & \tilde{a}_{2k}^{(0)}(z) \end{bmatrix} = \begin{bmatrix} a_{k}^{(0)}(z) & b_{k}^{(0)}(z) \\ \tilde{b}_{k}^{(0)}(z) & \tilde{a}_{k}^{(0)}(z) \end{bmatrix} \begin{bmatrix} a_{k}^{(k)}(z) & b_{k}^{(k)}(z) \\ \tilde{b}_{k}^{(k)}(z) & \tilde{a}_{k}^{(k)}(z) \end{bmatrix}, \qquad (15)
$$

and the updates for $c(z)$ and $d(z)$,

$$
c_{2k}^{(0)}(z) = a_{k}^{(0)}(z) c_{k}^{(k)}(z) + b_{k}^{(0)}(z) d_{k}^{(k)}(z) + g_{k}^{(k)}(0) c_{k}^{(0)}(z), \qquad (16)
$$

$$
d_{2k}^{(0)}(z) = \tilde{b}_{k}^{(0)}(z) c_{k}^{(k)}(z) + \tilde{a}_{k}^{(0)}(z) d_{k}^{(k)}(z) + g_{k}^{(k)}(0) d_{k}^{(0)}(z). \qquad (17)
$$

Since $e(0) = f(0) = 0$,

$$
\begin{aligned}
g_{2k}^{(0)}(z) = {} & e_{k}^{(0)}(z) c_{k}^{(k)}(z) + f_{k}^{(0)}(z) d_{k}^{(k)}(z) + g_{k}^{(0)}(z) g_{k}^{(k)}(z) \\
& + (g_{k}^{(0)}(z) - 1/g_{k}^{(0)}(0))(-g_{k}^{(k)}(z) + g_{k}^{(k)}(0))
\end{aligned}
$$

gives the scalar update

$$
g_{2k}^{(0)}(0) = g_{k}^{(0)}(0) g_{k}^{(k)}(0). \qquad (18)
$$

From (11) we use the generator transformations

$$
v_{k/2}(z) = v_0(z) a_{k/2}^{(0)}(z) + w_0(z) \tilde{b}_{k/2}^{(0)}(z)
$$

and

$$
w_{k/2}(z) = v_0(z) b_{k/2}^{(0)}(z) + w_0(z) \tilde{a}_{k/2}^{(0)}(z).
$$

For $\beta(z)$ we use

$$\beta_{k/2}(z) = v_0(z)c^{(0)}_{k/2}(z) + w_0(z)d^{(0)}_{k/2}(z) + \beta_0(z)g^{(0)}_{k/2}(0).$$

We keep track of the quantity δ in the augmented matrix generators by noting that if $\delta_0 = 1$ and δ_k represents the quantity δ after k steps of the Schur algorithm, then, by comparing (8) and (13), we see that both δ_k and $g^{(0)}_k(0)$ are products of the cosines c used in the generalized Schur factorization of the augmented matrix. It follows that

$$\delta_k = g^{(0)}_k(0).$$

In the algorithm we will use the relation

$$\delta_{l+k/2} = g^{(l)}_{k/2}(0)\delta_l.$$

Since we will need only $g^{(l)}_k(0)$ and not the full polynomial $g^{(l)}_k(z)$, we set $g^{(l)}_k = g^{(l)}_k(0)$. Putting everything together, we get the superfast generalized Schur algorithm for the augmented matrix.

`function` $[a(z),\, b(z),\, c(z),\, d(z),\, g,\, S(z),\, P(z)]$= `bsfschur`$(v(z),\, w(z), \beta(z),\, \delta,\, n)$

`if` $n > 1$ `then`

$[a^{(0)}_{n/2}(z),\, b^{(0)}_{n/2}(z),\, c^{(0)}_{n/2}(z),\, d^{(0)}_{n/2}(z),\, g^{(0)}_{n/2},\, S^{(0)}_{n/2}(z),\, P^{(n/2)}_{n/2}\ ] =$
`bsfschur`$(v^{(n/2)}(z),\, w^{(n/2)}(z),\, \beta^{(n/2)}(z),\, \delta,\, n/2)$

$v_{n/2}(z) = v(z)a^{(0)}_{n/2}(z) + w(z)\tilde{b}^{(0)}_{n/2}(z)$

$w_{n/2}(z) = v(z)b^{(0)}_{n/2}(z) + w(z)\tilde{a}^{(0)}_{n/2}(z)$

$\beta_{n/2}(z) = v(z)c^{(0)}_{n/2}(z) + w(z)d^{(0)}_{n/2}(z) + \beta(z)g^{(0)}_{n/2}$

$[a^{(n/2)}_{n/2}(z),\, b^{(n/2)}_{n/2}(z),\, c^{(n/2)}_{n/2}(z),\, d^{(n/2)}_{n/2}(z),\, g^{(n/2)}_{n/2},\, S^{(n/2)}_{n/2}(z),\, P^{(n/2)}_{n/2}(z)\ ] =$
`bsfschur`$(v^{(n/2)}_{n/2}(z),\, w^{(n/2)}_{n/2}(z),\, \beta^{(n/2)}_{n/2}(z),\, \delta g^{(0)}_{n/2},\, n/2)$

$a(z) = a^{(0)}_{n/2}(z)a^{(n/2)}_{n/2}(z) + b^{(0)}_{n/2}(z)\tilde{b}^{(n/2)}_{n/2}(z)$

$b(z) = a^{(0)}_{n/2}(z)b^{(n/2)}_{n/2}(z) + b^{(0)}_{n/2}(z)\tilde{a}^{(n/2)}_{n/2}(z)$

$c(z) = a^{(0)}_{n/2}(z)c^{(n/2)}_{n/2}(z) + b^{(0)}_{n/2}(z)d^{(n/2)}_{n/2}(z) + g^{(n/2)}_{n/2}c^{(0)}_{n/2}(z)$

$d(z) = \tilde{b}^{(0)}_{n/2}(z)c^{(n/2)}_{n/2}(z) + \tilde{a}^{(0)}_{n/2}(z)d^{(n/2)}_{n/2}(z) + g^{(n/2)}_{n/2}d^{(0)}_{n/2}(z)$

$g = g^{(0)}_{n/2}g^{(n/2)}_{n/2}$

$S(z) = \begin{bmatrix} v(z) & w(z) & \beta(z)/\delta \end{bmatrix}$

$S(z) = \begin{bmatrix} S(z) \\ S^{(0)}_{n/2}(z) \\ S^{(n/2)}_{n/2}(z) \end{bmatrix}$

$P(z) = \begin{bmatrix} a(z) & b(z) \end{bmatrix}$

$P(z) = \begin{bmatrix} P(z) \\ P^{(0)}_{n/2}(z) \\ P^{(n/2)}_{n/2}(z) \end{bmatrix}$

`else`

$\rho = -w(z)/v(z)$

$a(z) = z/\sqrt{1 - |\rho|^2}$

$b(z) = \rho/\sqrt{1 - |\rho|^2}$

$s = \overline{\beta(z)}/\sqrt{(1-|\rho|^2)|v(z)|^2 + |\beta(z)|^2}$
$c(z) = -\overline{s}/\sqrt{1-|\rho|^2}$
$d(z) = -\overline{s\rho}/\sqrt{1-|\rho|^2}$
$g = \sqrt{1-|s|^2}$
$S(z) = \begin{bmatrix} v(z) & w(z) & \beta(z)/\delta \end{bmatrix}$
$P(z) = \begin{bmatrix} a(z) & b(z) \end{bmatrix}$
`endif`

Several features of the algorithm need to be explained. The function `bsfschur`() takes generator polynomials $v(z)$, $w(z)$, and $\beta(z)$; the scalar parameter δ; and the integer parameter n, where $n-1$ is the degree of the polynomials $v(z)$, $w(z)$, and $\beta(z)$. The parameter n is also the number of Schur steps to be performed. The algorithm is recursive and uses the doubling recurrence for the polynomials $a(z)$, $b(z)$, $c(z)$, and $d(z)$. The recursion terminates when $n = 1$. This corresponds to a 1×1 Toeplitz-like matrix or a 2×2 augmented matrix. When $n = 1$ the polynomials are just the elements of the matrix in (7).

The function `bsfschur`() incorporates code to compute and store generators for the augmented matrix and its Schur complements in $S(z)$. The corresponding Schur polynomials are stored in $P(z)$. To understand the storage scheme, partition an $(n+1) \times (n+1)$ augmented matrix as

$$M = \begin{bmatrix} T & b \\ b^{\mathrm{H}} & 2 - \frac{1}{\delta^2} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & b_1 \\ T_{12}^{\mathrm{H}} & T_{22} & b_2 \\ b_1^{\mathrm{H}} & b_2 & 2 - \frac{1}{\delta^2} \end{bmatrix}.$$

Suppose M has generators

$$\begin{bmatrix} v & w & b\delta & b\delta \\ 0 & 0 & \delta & \delta - 1/\delta \end{bmatrix}$$

with

$$\begin{bmatrix} v(z) & w(z) & \beta(z) \end{bmatrix} = \begin{bmatrix} 1 & z & \cdots & z^{n-1} \end{bmatrix} \begin{bmatrix} v & w & b\delta \end{bmatrix}.$$

Thus $v(z)$ and $w(z)$ are generators of T and the coefficients of $\beta(z)/\delta$ are the elements of the vector b. If `sfschur`() is run on generators for M, then the matrix $S(z)$ is constructed recursively as

$$S(z) = S_0^{(0)}(z) = \begin{bmatrix} v(z) & w(z) & \beta(z)/\delta \\ & S_{n/2}^{(0)}(z) & \\ & S_{n/2}^{(n/2)}(z) & \end{bmatrix}. \tag{19}$$

Matrices $S_{n/2}^{(0)}(z)$ and $S_{n/2}^{(n/2)}$ are defined in the same way but for the submatrix

$$M_1 = \begin{bmatrix} T_{11} & b_1 \\ b_1^{\mathrm{H}} & 2 - \frac{1}{\delta^2} \end{bmatrix}$$

and for the Schur complement

$$M_S = \begin{bmatrix} T_{22} - T_{12}^{\mathrm{H}} T_{11}^{-1} T_{12} & b_2 - T_{12}^{\mathrm{H}} T_{11}^{-1} b_1 \\ b_2^{\mathrm{H}} - b_1^{\mathrm{H}} T_{11}^{-1} T_{12} & 2 - \frac{1}{\delta^2} - b_1^{\mathrm{H}} T_{11}^{-1} b_1 \end{bmatrix}.$$

This recursive definition of $S(z)$ terminates when the relevant augmented matrix is 2×2, in which case

$$S(z) = \begin{bmatrix} v(z) & w(z) & \beta(z)/\delta \end{bmatrix}.$$

The structure of $P(z)$ is defined in a similar manner:

$$P(z) = P_0^{(0)}(z) = \begin{bmatrix} a_0^{(0)}(z) & b_0^{(0)}(z) \\ P_{n/2}^{(0)}(z) & \\ P_{n/2}^{(n/2)}(z) & \end{bmatrix}. \tag{20}$$

The information contained in $S(z)$ and $P(z)$ will be used by a function that solves the system $Tx = b$ or by a function to recompute the polynomials associated with a different right-hand side.

Given the Schur polynomials in $P(z)$ and the generators in $S(z)$, it is possible to recompute $c(z)$, $d(z)$, g, and $\beta(z)$ for a different right-hand side without repeating the computation of $v(z)$, $w(z)$, $a(z)$, and $b(z)$. In the following we assume that the inputs $S_0^{(0)}(z)$ and $P_0^{(0)}(z)$ are partitioned as in (19) and (20). The elements of the new right-hand side vector are the coefficients of the input polynomial $\beta(z)/\delta$. The outputs are the new polynomials $c_n^{(0)}(z)$ and $d_n^{(0)}(z)$, the scalar $g_n^{(0)}$ and $S_n^{(0)}(z)$ updated with the new right-hand side.

`function` $[c(z),\, d(z),\, g,\, S(z)] =$ `rhs`$(S_0^{(0)}(z),\, P_0^{(0)}(z),\, \beta(z),\, \delta,\, n)$

`if` $n > 1$ `then`

$[c_{n/2}^{(0)}(z)\ ,\ d_{n/2}^{(0)}(z),\, g_{n/2}^{(0)},\, S_{n/2}^{(0)}(z)\ ] =$

`rhs`$(S_{n/2}^{(0)},\, P_{n/2}^{(0)},\, \beta^{(n/2)}(z),\, \delta,\, n/2)$

$\beta_{n/2}(z) = v_0(z)c_{n/2}^{(0)}(z) + w_0(z)d_{n/2}^{(0)}(z) + \beta(z)g_{n/2}^{(0)}$

$[c_{n/2}^{(n/2)}(z),\, d_{n/2}^{(n/2)}(z),\, g_{n/2}^{(n/2)},\, S_{n/2}^{(n/2)}(z)\ ] =$

`rhs`$(S_{n/2}^{(n/2)},\, P_{n/2}^{(n/2)}\ ,\, \beta_{n/2}^{(n/2)}(z),\, \delta g_{n/2}^{(0)},\, n/2)$

$c(z) = a_{n/2}^{(0)}(z)c_{n/2}^{(n/2)}(z) + b_{n/2}^{(0)}(z)d_{n/2}^{(n/2)}(z) + g_{n/2}^{(n/2)}c_{n/2}^{(0)}(z)$

$d(z) = \tilde{b}_{n/2}^{(0)}(z)c_{n/2}^{(n/2)}(z) + \tilde{a}_{n/2}^{(0)}(z)d_{n/2}^{(n/2)}(z) + g_{n/2}^{(n/2)}d_{n/2}^{(0)}(z)$

$g = g_{n/2}^{(0)}g_{n/2}^{(n/2)}$

$S(z) = \begin{bmatrix} v(z) & w(z) & \beta(z)/\delta \end{bmatrix}$

$S(z) = \begin{bmatrix} S(z) \\ S_{n/2}^{(0)}(z) \\ S_{n/2}^{(n/2)}(z) \end{bmatrix}$

`else`

$\rho = -w_0(z)/v_0(z)$

$s = \overline{\beta(z)}/\sqrt{(1-|\rho|^2)|v(z)|^2 + |\beta(z)|^2}$

$c(z) = -\overline{s}/\sqrt{1-|\rho|^2}$

$d(z) = -\overline{s\rho}/\sqrt{1-|\rho|^2}$

$g = \sqrt{1-|s|^2}$

$S(z) = \begin{bmatrix} v(z) & w(z) & \beta(z)/\delta \end{bmatrix}$

`endif`

Note that there is no $P(z)$ as output for `rhs`(). This is because $P(z)$ does not depend on the right-hand side. The use of function `rhs`() substantially reduces the computation for problems that involve multiple right-hand sides. More significantly,

as we will see in the next section, it allows us to efficiently deal with transformations of the right-hand side when solving the system $Tx = b$.

3. Divide-and-conquer back-substitution. The function `bsfschur`() is a divide-and-conquer procedure for factoring the $(n+1) \times (n+1)$ augmented matrix

$$M = \begin{bmatrix} T & b \\ b^{\mathrm{H}} & 1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & b_1 \\ T_{12}^{\mathrm{H}} & T_{22} & b_2 \\ b_1^{\mathrm{H}} & b_2^{\mathrm{H}} & 1 \end{bmatrix}. \tag{21}$$

After $n/2$ steps of elimination on this matrix, we have the factorization

$$M = \begin{bmatrix} I & 0 & 0 \\ T_{12}^{\mathrm{H}} T_{11}^{-1} & I & 0 \\ b_1^{\mathrm{H}} T_{11}^{-1} & 0 & 1 \end{bmatrix} \begin{bmatrix} T_{11} & 0 & 0 \\ 0 & T_{22} - T_{12}^{\mathrm{H}} T_{11}^{-1} T_{12} & b_2 - T_{12}^{\mathrm{H}} T_{11}^{-1} b_1 \\ 0 & b_2^{\mathrm{H}} - b_1^{\mathrm{H}} T_{11}^{-1} T_{12} & 1 - b_1^{\mathrm{H}} T_{11}^{-1} b_1 \end{bmatrix} \cdot \begin{bmatrix} I & T_{11}^{-1} T_{12} & T_{11}^{-1} b_1 \\ 0 & I & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Both the vector

$$b_S = b_2 - T_{12}^{\mathrm{H}} T_{11}^{-1} b_1$$

and the Schur complement

$$T_S = T_{22} - T_{12}^{\mathrm{H}} T_{11}^{-1} T_{12}$$

can be found from the matrix $S(z)$ returned by `bsfschur`(). In fact, the generators of the matrix

$$\begin{bmatrix} T_S & b_S \\ b_S^{\mathrm{H}} & 2 - \frac{1}{\delta_{n/2}} \end{bmatrix}$$

are available in polynomial form as the first row of $S_{n/2}^{(n/2)}(z)$. The Schur complements of the augmented matrix, stored in $S(z)$, are the data that will be used to solve $Tx = b$.

Given a linear system partitioned as

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{12}^{\mathrm{H}} & T_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

elimination gives

$$\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_S \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_S \end{bmatrix}.$$

Block back-substitution gives two smaller linear systems

$$T_S x_2 = b_S, \qquad T_{11} x_1 = b_1 - T_{12} x_2. \tag{22}$$

Using `bsfschur`(), the computation of generators for the Schur complement system and its right-hand side is $O(n \log^2(n))$.

To form the right-hand side for $T_{11}x_1 = b_1 - T_{12}x_2$, we note that T_{12} is a block of the Toeplitz-like matrix T, so it is also Toeplitz-like. More precisely, if we partition the displacement equation for T, $T - ZTZ^{\mathrm{H}} = vv^{\mathrm{H}} - ww^{\mathrm{H}}$, as

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{12}^{\mathrm{H}} & T_{12} \end{bmatrix} - \begin{bmatrix} Z_{11} & 0 \\ e_1 e_{n/2}^{\mathrm{H}} & Z_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{12}^{\mathrm{H}} & T_{12} \end{bmatrix} \begin{bmatrix} Z_{11}^{\mathrm{H}} & e_{n/2} e_1^{\mathrm{H}} \\ 0 & Z_{22}^{\mathrm{H}} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \begin{bmatrix} v_1^{\mathrm{H}} & v_2^{\mathrm{H}} \end{bmatrix} - \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \begin{bmatrix} w_1^{\mathrm{H}} & w_2^{\mathrm{H}} \end{bmatrix},$$

then

$$T_{12} - Z_{11} T_{12} Z_{22}^{\mathrm{H}} = v_1 v_2^{\mathrm{H}} - w_1 w_2^{\mathrm{H}} + Z_{11} T_{11} e_{n/2} e_1^{\mathrm{H}}. \tag{23}$$

Thus T_{12} is in general a displacement rank 3 Toeplitz-like matrix. Multiplication by T_{12} is $O(n \log(n))$ using the FFT. Thus both half-size systems (22) can be computed efficiently.

This motivates the following divide-and-conquer algorithm. We assume that the inputs $S(z)$ and $P(z)$ are partitioned as in (19) and (20).

`function` x=`solve`($S(z)$,$P(z)$,n)
  `if` $n > 1$ `then`
    x_2=`solve` ($S_{n/2}^{(n/2)}(z)$,$P_{n/2}^{(n/2)}(z)$,$n/2$)
    b_1 = `vec`($\beta(z)/\delta$)
    $b_1 = b_1(1 : n/2)$
    T = `toeplitz`($v(z), w(z)$)
    $T_{12} = T(1 : n/2, n/2 + 1 : n)$
    $b_1 = b_1 - T_{12} x_2$
    $\beta(z) = \begin{bmatrix} 1 & z & \cdots z^{n/2-1} \end{bmatrix} b_1$
    [$c(z)$, $d(z)$, g, $S_{n/2}^{(0)}(z)$]=
      `rhs`($S_{n/2}^{(0)}(z)$, $P_{n/2}^{(0)}(z)$,$\beta^{(n/2)}(z)$, 1, $n/2$)
    x_1=`solve` ($S_{n/2}^{(0)}(z)$,$P_{n/2}^{(0)}(z)$,$n/2$)
    $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
  `else`
    $x = (v(z)\overline{v(z)} - w(z)\overline{w(z)})^{-1} \beta(z)/\delta$
  `endif`

The function `toeplitz`() constructs a Toeplitz-like matrix from the generators $v_0(z)$ and $w_0(z)$. The function `vec`() forms a vector from the coefficients of a polynomial. These functions make possible the matrix notation

$$b_1 - T_{12} x_2.$$

In practice, this would not be done explicitly; instead the multiplication by T_{12} would be carried out with the generators and the FFT using only $O(n \log(n))$ operations. Unfortunately the call to `rhs`() is not so efficient; it is $O(n \log^2(n))$. As we will see, this makes the procedure `solve`() an $O(n \log^3(n))$ algorithm.

In the case that the recursion terminates with $n = 1$, the matrix T is 1×1, $v(z)$, $w(z)$, and $\beta(z)$ are constants and

$$T = v(z)\overline{v(z)} - w(z)\overline{w(z)}, \qquad b = \beta(z)/\delta$$

so that the solution to $Tx = b$ is

$$x = (v(z)\overline{v(z)} - w(z)\overline{w(z)})^{-1}\beta(z)/\delta.$$

The function assumes that $S(z)$ and $P(z)$ are available. Thus `solve()` would be used as follows:
$[a(z), b(z), c(z), d(z), g, S(z), P(z)]$= `bsfschur`$(v(z), w(z), \beta(z), 1, k)$
x=`solve`$(S(z), P(z), n)$
This naturally assumes that the function `rhs()` is also available.

4. Computational complexity. The complexity of solving a Toeplitz system using `solve()` is $O(n\log^3(n))$. In this section we will provide a more detailed count of arithmetic operations. We will assume that T and b are real, and we will use the assumptions from [2] on the complexity of convolutions, in particular that convolutions are implemented using a split-radix FFT so that a real cyclic convolution of two length n vectors takes

$$6n\log_2(n) - 9n + 14$$

real floating point operations.

The cyclic convolution is

$$w_j = \sum_{k=0}^{n-1} x_k y_{j-k},$$

where y_k and x_k are defined for $k = 0, 1, \ldots, n-1$, $y_{-k} = y_{n-k}$ and $j = 0, 1, \ldots, n-1$. Multiplication of polynomials is a linear rather than a cyclic convolution: It is assumed that $y_k = 0$ for $k < 0$ and $j = 0, 1, \ldots, 2n-2$. Thus to multiply two length n polynomials with coefficients given as elements of the vectors $x = [x_k]$ and $y = [y_k]$, we can pad the vectors with n zeros and compute the cyclic convolution

$$w = \begin{bmatrix} x \\ 0 \end{bmatrix} * \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

In analyzing the complexity of `bsfschur()`, it is important to take note of the length of each of the convolutions. The equations

$$v_{n/2}(z) = v(z)a^{(0)}_{n/2}(z) + w(z)\tilde{b}^{(0)}_{n/2}(z), \qquad w_{n/2}(z) = v(z)b^{(0)}_{n/2}(z) + w(z)\tilde{a}^{(0)}_{n/2}(z)$$

involve four convolutions, each with one length n vector and one length $n/2$ vector. This can be implemented using a length $3n/2$ cyclic convolution. However, because $v_{n/2}(z)$ and $w_{n/2}(z)$ will be truncated and will have leading zeros, only the middle $n/2$ elements of the convolutions will be needed. It follows that these can be done using four length n convolutions [2]. This also applies to the convolutions in

$$\beta_{n/2}(z) = v(z)c^{(0)}_{n/2}(z) + w(z)d^{(0)}_{n/2}(z) + \beta(z)g^{(0)}_{n/2}.$$

The convolutions in the equations for $a(z)$, $b(z)$, $c(z)$, and $d(z)$ can also be implemented using length n cyclic convolutions.

Let $B(n)$ be the complexity of `bsfschur()`. The function calls itself twice on half-size problems and performs 14 length n cyclic convolutions. The complexity satisfies

$$B(n) = 2B(n/2) + 14\,(6n\log_2(n) - 9n + 14) + \frac{19}{2}n + 2$$

or

$$B(n) = 2B(n/2) + 84n\log_2(n) - \frac{233}{2}n + 198. \tag{24}$$

The term $19n/2 + 2$ in the first expression is the cost of the vector additions, scalar-vector multiplications, and two scalar-scalar multiplications. The scalar multiplications are the products $g^{(0)}_{n/2}g^{(n/2)}_{n/2}$ and $\delta g^{(0)}_{n/2}$. In assessing the cost of the vector additions, we have assumed that elements of vectors that are to be truncated are not computed.

The general solution to (24) is

$$B(n) = 42n\log_2^2(n) - \frac{149}{2}n\log_2(n) + Cn - 198.$$

To determine the constant C, we note that, if $n = 1$, then `bsfschur`() performs 18 real operations. Thus

$$B(1) = C - 198 = 18$$

so that

$$B(n) = 42n\log_2^2(n) - \frac{149}{2}n\log_2(n) + 216n - 198.$$

Now let the complexity of `rhs`() be $R(n)$. Only six convolutions are required, so

$$R(n) = 2R(n/2) + 6\left(6n\log_2(n) - 9n + 14\right) + \frac{13}{2}n + 2,$$

or

$$R(n) = 2R(n/2) + 36n\log_2(n) - \frac{95}{2}n + 86.$$

The general solution to this is

$$R(n) = 18n\log_2^2(n) - \frac{59}{2}n\log_2(n) + Cn - 86.$$

Since $R(1) = C - 86 = 16$,

$$R(n) = 18n\log_2^2(n) - \frac{59}{2}n\log_2(n) + 102n - 86.$$

For `solve`() we assume that the multiplication by the $n/2 \times n/2$ Toeplitz-like matrix T_{12} involves

$$4\left(6(2n)\log_2(2n) - 9(2n) + 14\right) + n/2 = 48n\log_2(n) - \frac{47}{2}n + 56$$

operations. The justification is as follows. We note that

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{12}^{\mathrm{H}} & T_{22} \end{bmatrix} \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = \begin{bmatrix} T_{12}x_2 \\ T_{22}x_2 \end{bmatrix}$$

so that to multiply a vector by T_{12} efficiently it suffices to have an efficient means of multiplying a vector by T. Since T is Toeplitz-like, it can be represented through the Gohberg–Semencul formula

$$T = L_+L_+^{\mathrm{H}} - L_-L_-^{\mathrm{H}},$$

where $L_\pm$ are lower triangular and Toeplitz. Each matrix $L\pm$ can then be embedded in a circulant matrix of size $2n \times 2n$. Multiplication by these circulants is simply a cyclic convolution. Thus multiplication by T can be reduced to four convolutions of size $2n$ with cost

$$4\left(6(2n)\log_2(2n) - 9(2n) + 14\right).$$

The additional $n/2$ is the complexity of the length $n/2$ addition that finally computes $T_{12}x_2$. Alternately it is possible to use (23) directly and in so doing avoid expanding the size of the convolutions by a factor of 4. However, this would involve a greater number of convolutions and the additional computation of $T_{11}e_n$.

Under the above assumption, if $S(n)$ is the cost of `solve`(), then

$$S(n) = 2S(n/2) + R(n/2) + \left(48n\log_2(n) - \frac{47}{2}n + 56\right) + n/2$$

or

$$S(n) = 2S(n/2) + 9n\log_2^2(n) + \frac{61}{4}n\log_2(n) + \frac{207}{4}n - 30.$$

The general solution is

$$S(n) = 3n\log_2^3(n) + \frac{97}{8}n\log_2^2(n) + \frac{487}{8}n\log_2(n) + Cn + 30.$$

For the case $n = 1$, $S(1) = C + 30 = 4$ so that

$$S(n) = 3n\log_2^3(n) + \frac{97}{8}n\log_2^2(n) + \frac{487}{8}n\log_2(n) - 26n + 30$$

real floating point operations.

To fully solve a Toeplitz system requires an initial call to `bsfschur`() so that the complexity of solving $Tx = b$ is

(25)
$$T(n) = S(n) + B(n) = 3n\log_2^3(n) + \frac{433}{8}n\log_2^2(n) - \frac{109}{8}n\log_2(n) + 190n - 168.$$

To compare this superfast algorithm to the most comparable fast methods, we note that Schur's algorithm requires $3n^2$ operations to compute the Cholesky factor of T. Another $2n^2$ is required for back-substitution, so the solution of $Tx = b$ requires roughly $5n^2$ operations. The smallest value of n for which $T(n)$ is smaller than $5n^2$ is $n = 2148$. Of course `solve`() assumes that n is a power of 2 and the smallest power of 2 for which `solve`() has a smaller operation count than the Schur algorithm is $n = 4096$. Nevertheless the algorithm is very close to breaking even at $n = 2048$.

In contrast, since the superfast Schur algorithm of [2] computes generators of T^{-1}, it is perhaps most naturally compared to the Levinson algorithm, which also computes these generators. In [2] it was shown that the superfast Schur algorithm breaks even in comparison to the Levinson algorithm for $n = 256$.

Finally, we note that the overall storage required by the recursive algorithm is $O(n\log(n))$. Instead of making assumptions about how computer memory is used, we will analyze the storage required by $S(z)$. All the other polynomials computed by the algorithm could be stored in a similar array so that is sufficient to show

that $S(z)$ requires $O(n\log(n))$ storage. Note that this analysis assumes the recursive formulation given here; there is redundancy in $S(z)$ and it might be possible to develop a nonrecursive algorithm that uses only $O(n)$ storage.

Given an $n \times n$ Toeplitz system, let $M(n)$ be the storage required for $S_0^{(0)}(z)$. Then

$$M(n) = 2M(n/2) + 3n,$$

which has solution

$$M(n) = 3n\log_2(n) + Cn.$$

If $n = 1$, then $S(z)$ stores only three constants so that

$$M(1) = C = 3$$

so that

$$M(n) = 3n\log_2(n) + 3n.$$

5. Numerical experiments. The reason for formulating a superfast algorithm in terms of factorization and divide-and-conquer back-substitution was in the hope of achieving some of the stability inherent in unstructured triangularization and back-substitution. Unfortunately the algorithm is quite complicated; a rigorous error analysis has not been performed and might well be extremely difficult. Instead we will attempt to assess stability through numerical experiments. All experiments were conducted using code written in Matlab and run on a Pentium III PC with machine precision approximately $\epsilon = 1 \times 10^{-16}$. The FFT routines used were those built into Matlab.

For the first experiment, we generated a 128×128 positive definite Toeplitz matrix from random Schur parameters ρ_k distributed uniformly over the interval $[-.5, .5]$. These parameters resulted in ill-conditioned but numerically nonsingular matrices. For the right-hand side vector b, we randomly generated a vector $\hat{x}$ and then formed the product $T\hat{x} = b$. For solutions x obtained by `solve`() and by the superfast inversion algorithm of [2] combined with the Gohberg–Semencul formula for multiplication by T^{-1}, the relative residuals

$$r(T, b, x) = \frac{\|Tx - b\|}{\|T\|\|x\| + \|b\|}$$

are shown in the first three lines of Table 1. (Note that each line in the table corresponds to a different matrix.) As expected for a method based on inversion, the residuals for the Gohberg–Semencul approach are large. The residuals for `solve`() are what might be expected for a backward stable algorithm. These results were typical for random problems generated in this way.

Next we generated ill-conditioned Toeplitz matrices for which $|\rho_k|$ was close to 1 for some k. In particular, we generated random 128×128 Toeplitz matrices with ρ_k uniformly distributed over $[-.3, .3]$ with two of the ρ_k changed to

$$\rho_{10} = .9999999, \qquad \rho_{15} = -.99.$$

The right-hand side vectors were generated in the same manner as before. The results are shown in lines 4–6 of Table 1. Note that these matrices are almost numerically

TABLE 1
Relative residuals.

Experiment	$\kappa(T)$	$\|T_{11}^{-1}T_{12}\|$	`solve()`	Inversion
1	5.14×10^8	25.3	1.2×10^{-14}	6.6×10^{-10}
	3.5×10^8	12.0	6.7×10^{-16}	2.2×10^{-10}
	2.4×10^9	25.7	9.1×10^{-15}	1.02×10^{-8}
2	1.1×10^{14}	37.8	9.4×10^{-15}	1×10^{-4}
	2.3×10^{13}	11.6	5.0×10^{-15}	7.5×10^{-4}
	1.5×10^{14}	83.6	6.42×10^{-15}	8.7×10^{-5}
3	7.8×10^{11}	1.1×10^4	2.0×10^{-10}	2.5×10^{-7}
	3.15×10^{13}	1.7×10^3	2.7×10^{-13}	3.6×10^{-6}

singular. The errors for `solve()` remain on the order of the machine precision, while those for the other algorithm have increased with the increasing condition number.

Finally, we devise an experiment to highlight a notable weakness of the new algorithm: `solve()` can lose accuracy when the quantity $\|T_{11}^{-1}T_{12}\|$ (or the equivalent quantity for any of the Schur complements of T) becomes large. It was shown in [13] that for a positive definite Toeplitz matrix, or for the Schur complement of a positive definite Toeplitz matrix, the quantity $\|T_{11}^{-1}T_{12}\|_2$ can be bounded by an expression that depends only on the sizes of the matrices and not on $\|T_{11}^{-1}\|$. The reason is that if

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{12}^{\mathrm{H}} & T_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} L_{11}^{\mathrm{H}} & L_{21}^{\mathrm{H}} \\ 0 & L_{22}^{\mathrm{H}} \end{bmatrix}$$

is an LDL^{H} factorization of T, then

$$T_{11}^{-1}T_{12} = (L_{11}D_1L_{11}^{\mathrm{H}})^{-1}L_{11}D_1L_{21}^{\mathrm{H}} = L_{11}^{-1}L_{12}.$$

However, it can be shown that the rows of L^{-1} are the coefficients of optimal filters solving a linear prediction problem [8]. Consequently they have the following well-known minimum phase property: the polynomials with coefficients taken from the rows of L^{-1} have zeros only in the unit circle. This implies that

$$|L^{-1}| < \begin{bmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & 2 & 1 & & & \\ 1 & 3 & 3 & 1 & & \\ 1 & 4 & 6 & 4 & 1 & \\ \vdots & & & & & \ddots \end{bmatrix}.$$

Equivalently

$$|L_{ij}| \leq \begin{pmatrix} i-1 \\ j-1 \end{pmatrix}.$$

It is shown in [13] that the same bounds hold for $|L|$, although this fact follows for different reasons; the polynomials formed from the rows of L do not have the minimum phase property and can have zeros outside the unit circle. The result of the two inequalities is that $L_{11}^{-1}L_{12}$, and hence $T_{11}^{-1}T_{12}$, must be bounded by a function of n independent of the size of $\|T_{11}^{-1}\|$. Unfortunately the bounds are not completely

satisfactory; binomial coefficients grow quickly with increasing n. Furthermore, there is an example of a sequence of positive definite Toeplitz matrices for which both L and L^{-1} approach these bounds in the limit. (In this limit the Toeplitz matrix also becomes singular.)

Nevertheless, the bounds are in practice very pessimistic. It is extremely difficult to generate Toeplitz matrices of any reasonable size that come close to achieving these bounds and are still numerically positive definite. For the next experiment, we will apply `solve()` to positive definite Toeplitz matrices for which $\|T_{11}^{-1}T_{12}\|$ is as large as we are able to make it. The examples are small because larger examples that are numerically nonsingular could not be generated.

We started with a Toeplitz matrix for which

$$\rho_1 = \rho_2 = \cdots = \rho_7 = .99$$

and

$$\rho_8 = \rho_9 = \cdots = \rho_{32} = .2.$$

The right-hand side was generated in the same way as before. The results are on the seventh line of Table 1. Clearly $\|T_{11}^{-1}T_{12}\|$ is larger than before, and there has been a proportional increase in the errors.

Nevertheless, it seems that the algorithm is stable in most circumstances. This example is very extreme, and even seemingly minor changes in the parameters considerably reduce $\|T_{11}^{-1}T_{12}\|$. Suppose we keep the previous set of Schur parameters, changing only $\rho_4 = .1$ and $\rho_7 = -.8$. The results are in the final line of Table 1. The quantity $\|T_{11}^{-1}T_{12}\|$ has dropped an order of magnitude, and the error has improved for `solve()` but not for the inversion. Note that the condition number has become worse; growth in errors is apparently not linked to ill-conditioning in a simple or direct way.

6. Observations. The algorithm proposed in this paper is a divide-and-conquer $O(n\log^3(n))$ method for the solution of positive definite Toeplitz systems. It achieves a crossover point at which it beats the Schur algorithm for $n = 4096$. Its strength over previous superfast methods is that it is observed to be relatively numerically stable. Experiments suggest that this stability is connected with the tendency of the block eliminators,

$$\begin{bmatrix} I & 0 \\ -T_{12}^{\mathrm{H}}T_{11}^{-1} & I \end{bmatrix},$$

to be of modest size when T is positive definite and Toeplitz. It was shown in [13] that the Schur complements of a Toeplitz matrix are insensitive to perturbations when $T_{11}^{-1}T_{12}$ is not large. This might make possible an error analysis based on forward accuracy in computed Schur complements. This is a possible direction for further research.

A stabilized superfast algorithm for nonsymmetric Toeplitz systems was published in [16]. However, that algorithm depended in part on iterative refinement for its stability. Iterative refinement is well known to stabilize algorithms that are not too unstable applied to problems that are not too ill-conditioned [12, 9]. The algorithm of [16] appeared to be stable in most cases, but it displayed growth in relative residuals, despite iterative refinement, when tested on some very large problems. In contrast,

the algorithm presented here is stable on at least some extremely ill-conditioned problems. Furthermore, the mild degree of instability exhibited by the algorithm is not so extreme as to prevent iterative refinement from restoring backward stability. Unfortunately it is not clear how the new algorithm might be extended to the nonsymmetric Toeplitz matrices considered in [16] or to any broader class of structured matrices. Furthermore, it seems likely that the stability of the algorithm depends on $\|T_{11}^{-1}T_{12}\|$ not being too large. Bounds of this sort have been established only for positive definite Toeplitz matrices.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *The implementation and use of the generalized Schur algorithm*, in Computational and Combinatorial Methods in Systems Theory, C. I. Byrnes and A. Linquist, eds., North–Holland, Amsterdam, 1986, pp. 265–279.

[2] G. S. Ammar and W. B. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[3] R. R. Bitmead and B. D. O. Anderson, *Asymptotically fast solution of Toeplitz and related systems of linear equations*, Linear Algebra Appl., 34 (1980), pp. 103–116.

[4] 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.

[5] R. P. Brent, F. Gustavson, and D. Yun, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[6] S. Chandrasekaran and A. H. Sayed, *Stabilizing the generalized Schur algorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 950–983.

[7] F. de Hoog, *A new algorithm for solving Toeplitz systems of equations*, Linear Algebra Appl., 88/89 (1987), pp. 123–138.

[8] S. Haykin, *Adaptive Filter Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1991.

[9] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[10] B. R. Musicus, *Levinson and Fast Choleski Algorithms for Toeplitz and Almost Toeplitz Matrices*, Research report, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[11] C. M. Rader and A. O. Steinhardt, *Hyperbolic Householder transforms*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 269–290.

[12] R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

[13] M. Stewart, *Cholesky factorization of semidefinite Toeplitz matrices*, Linear Algebra Appl., 254 (1997), pp. 497–525.

[14] M. Stewart and G. W. Stewart, *On hyperbolic triangularization: Stability and pivoting*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 847–860.

[15] M. Stewart and P. Van Dooren, *Stability issues in the factorization of structured matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 104–118.

[16] M. Van Barel, G. Heinig, and P. Kravanja, *A stabilized superfast solver for nonsymmetric Toeplitz systems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 494–510.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 694–717

SUPPORT THEORY FOR PRECONDITIONING*

ERIK G. BOMAN† AND BRUCE HENDRICKSON†

This paper is dedicated to the memory of Fred Howes

Abstract. We present *support theory*, a set of techniques for bounding extreme eigenvalues and condition numbers for matrix pencils. Our intended application of support theory is to enable proving condition number bounds for preconditioners for symmetric, positive definite systems. One key feature sets our approach apart from most other works: We use *support numbers* instead of generalized eigenvalues. Although closely related, we believe support numbers are more convenient to work with algebraically.

This paper provides the theoretical foundation of support theory and describes a set of analytical tools and techniques. For example, we present a new theorem for bounding support numbers (generalized eigenvalues) where the matrices have a known factorization (not necessarily square or triangular). This result generalizes earlier results based on graph theory. We demonstrate the utility of this approach by a simple example: block Jacobi preconditioning on a model problem. Also, our analysis of a new class of preconditioners, *maximum-weight basis preconditioners*, in [E. G. Boman, D. Chen, B. Hendrickson, and S. Toledo, *Numer. Linear Algebra Appl.*, to appear] is based on results contained in this paper.

Key words. preconditioning, eigenvalue bounds, condition number, support theory

AMS subject classifications. 65F10, 65F15, 65F35

DOI. 10.1137/S0895479801390637

1. Introduction.

The solution of linear systems of equations is at the heart of many computations in science, engineering, and other disciplines. Iterative methods are often the most efficient means to solve such systems. In many cases, the matrix describing the system is symmetric, positive definite, in which case the preconditioned conjugate gradients method is the algorithm of choice. The cost of using an iterative method like preconditioned conjugate gradients is the cost of a single iteration (involving the operation of the matrix and of the preconditioner on a vector) multiplied by the number of iterations. Preconditioning is important to keep the number of iterations small. For (preconditioned) conjugate gradients or Chebyshev iteration, the number of iterations is known to be bounded by a constant times the square root of the condition number (after preconditioning). This analysis is based on Chebyshev polynomials and represents a worst-case scenario, so in practice the number of iterations may be much smaller, for instance, when the eigenvalues are clustered. Still, the spectral condition number is a useful indicator of the quality of a preconditioner.

The dual goals of finding a preconditioner that is both of good quality and inexpensive to compute and apply often conflict, and the design of effective preconditioners continues to be a very active area of research. Many of the best preconditioners are

*Received by the editors June 8, 2001; accepted for publication (in revised form) by M. Hanke May 26, 2003; published electronically December 17, 2003. This work was funded by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research, and was performed at Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC-94AL85000.

http://www.siam.org/journals/simax/25-3/39063.html

†Discrete Algorithms and Math Department, Sandia National Laboratories, Albuquerque, NM 87185-1111 (eboman@cs.sandia.gov, bah@cs.sandia.gov).

specialized to individual problems. Some general-purpose preconditioning techniques include variants of incomplete factorizations, approximate inverses, algebraic multilevel methods, or domain decomposition. None of these approaches is a panacea, and preconditioning remains as much an art as a science. One of the biggest problems with preconditioning is that convergence analysis is generally limited to simple model problems. For problems with irregular numerical or topological structure, condition number bounds are generally difficult to obtain.

Much work has been done in the field of bounding eigenvalues and condition numbers. In this paper we introduce *support theory* as a mathematical framework to analyze condition numbers of preconditioned systems. Our focus will be on symmetric positive definite (spd) and symmetric positive semidefinite (spsd) systems. We provide a set of tools with which one can bound *support numbers* (to be defined in the next section). Support numbers are closely related to generalized eigenvalues. Several authors have earlier derived eigenvalue bound techniques for certain families of preconditioners, in particular incomplete factorizations; see, for example, work by Axelsson and Barker [3], Axelsson [1], Beauwens [4, 5], Magolu and Notay [20], Magolu [19], and Notay [21, 22]. Although some of the basic tools in the present paper have implicitly been used earlier by others, we believe that our main support theory results (section 4) are new and different. Also, these results apply to all spsd matrices, not just M-matrices.

Many of our *support theory* techniques can be viewed as an algebraic generalization of recent work on a little-known technique called *support-graph* preconditioning; hence the name. Several core ideas in support-graph theory can be traced back to Beauwens [5] and were rediscovered by Vaidya, who used them to study spanning tree preconditioners [28]. The techniques were extended and applied to multilevel methods by Gremban [11], Gremban, Miller, and Zagha [12], Reif [24], and Bern et al. [6]. The resulting methods have been applied to the analysis of incomplete Cholesky factorization by Guattery [13] and by Bern et al. [6] and to multilevel diagonal scaling [6]. Unfortunately, support-graph theory is fairly limited in its applicability. It applies only to spsd diagonally dominant M-matrices (a subset of Stieltjes matrices) and, in some cases, to all spsd diagonally dominant matrices. In contrast, our algebraic support theory applies to all spsd matrices. Furthermore, as we discuss in section 9, support-graph theory is a special case of our methodology.

In this paper we present a collection of propositions and theorems, some of which are quite elementary and correspond to well-known facts in linear algebra. We show that the *support number* used in our analysis is the largest generalized eigenvalue in a certain subspace. More specifically, support numbers are well-defined under rank-deficiency and in that sense more robust than generalized eigenvalues. The support number definition is often easier to work with than that of eigenvalues. Our hope is that by reformulating results in terms of support numbers and gathering them into a single paper, this will become a useful resource for future work. This paper forms the foundation for several forthcoming papers by the present authors and collaborators.

In section 2 we review the concept of support number and describe how it can be used to bound condition numbers. In section 3 we provide a collection of fundamental algebraic properties of support numbers. This is followed in section 4 with our most important set of tools and techniques for analyzing preconditioners. In section 5 we expand our tool kit to address diagonal matrices (preconditioners). A few basic results about Schur complements are stated in section 6. We then present some fairly

specialized techniques for analyzing Hadamard products and negative semidefinite matrices in sections 7 and 8, respectively. We discuss the relationship between this paper and previous work on support-graph theory in section 9. In section 10 we demonstrate how our support tools can be used to analyze a simple, well-known preconditioner, namely, block Jacobi preconditioning. In section 11 we propose a generalization of support numbers that may be useful for analyzing nonsymmetric or indefinite systems.

2. Support theory definitions and concepts. The main goal of the *support theory* in this paper is to provide techniques to bound the generalized eigenvalues and condition number for a matrix pencil (A, B). Think of B as being a preconditioner for A. We study only real matrices in this paper, but most of the results carry over to the complex case (substitute Hermitian for symmetric). If both A and B are spd, then the convergence of many preconditioned iterative methods (and, specifically, preconditioned conjugate gradients) depends on the condition number of the preconditioned operator $B^{-1/2}AB^{-1/2}$. We define the generalized (spectral) condition number by

$$\kappa(A, B) \equiv \kappa(B^{-1/2}AB^{-1/2}) = \frac{\lambda_{\max}(B^{-1/2}AB^{-1/2})}{\lambda_{\min}(B^{-1/2}AB^{-1/2})} = \frac{\lambda_{\max}(A, B)}{\lambda_{\min}(A, B)},$$

where $\lambda(A)$ denotes an eigenvalue of A while $\lambda(A, B)$ denotes a generalized eigenvalue for (A, B).

The central concept in support theory is the *support number* of a matrix pair (A, B), sometimes simply called the *support*. We remark that the definition we use is slightly different from the one in [6] and [11] but only when A or B is indefinite.

DEFINITION 2.1. *The support number of* (A, B)*, where* $A \in \mathbb{R}^{n\times n}$*,* $B \in \mathbb{R}^{n\times n}$*, is defined by*

$$\sigma(A, B) = \min\left\{t \in \mathbb{R} \,|\, x^T(\tau B - A)x \geq 0 \text{ for all } x \in \mathbb{R}^n \text{ and for all } \tau \geq t\right\}.$$

For some pencils (A, B), there is no such t and we define the support number $\sigma(A, B)$ to be ∞. Similarly, if $\tau B - A$ is positive semidefinite (psd) for all τ we define the support number to be $-\infty$. (This cannot happen if B is psd.) In this paper, we say that a matrix C is psd if $y^TCy \geq 0$ for all real vectors y, even if C is not symmetric (cf. [10, section 4.2]).

The definition above does not require A and B to be symmetric. However, symmetric matrices will be the main focus of this paper. We remark that by choosing $B = I$, the techniques in this paper can be used to bound the largest eigenvalue and spectral condition number of A. For symmetric matrices, the support is closely related to a generalized eigenvalue. Axelsson [1, Corollary 2.1] showed the following result.

LEMMA 2.2. *Suppose* A *is spsd and* B *is spd. For any* τ *such that* $\lambda_{\min}(\tau B - A) \geq 0$ *we have*

$$\lambda_{\max}(B^{-1}A) \leq \tau.$$

In other words, an upper bound on the support number $\sigma(A, B)$ is also a bound on the generalized eigenvalue $\lambda_{\max}(A, B) \equiv \max\{\lambda \,|\, Ax = \lambda Bx, x \neq 0\}$. (More general versions of this lemma can be found as Theorem 3.16 and Theorem 10.1 in [2].) Next, we elaborate on this important result and include the case where B is spsd and may be singular. The theorem below is an extension of Gremban's support lemma [11, Lemma 4.4] and similar lemmas in [6].

THEOREM 2.3. *Let A and B be symmetric matrices.*

1. *If B is spd, then $\sigma(A,B) = \lambda_{\max}(A,B)$.*
2. *If B is spsd and $Null(B) \subseteq Null(A)$, then*

$$\sigma(A,B) = \max\{\lambda \mid Ax = \lambda Bx,\ Bx \neq 0\},$$

or, equivalently,

$$\sigma(A,B) = \lambda_{\max}(Z^T AZ, Z^T BZ),$$

where Z is such that the columns of Z span the range of B.

3. *If B is not spsd, then $\sigma(A,B)$ is infinite.*

Proof. The first part follows from the variational characterization

$$\lambda_{\max}(A,B) = \tau \max_{x \neq 0} \frac{x^T Ax}{x^T(\tau B)x},$$

where B is assumed to be spd. For any τ such that $x^T(\tau B - A)x \geq 0$ the condition above implies that $\lambda_{\max}(A,B) \leq \tau$. Equality holds when τ is the largest generalized eigenvalue and x is the corresponding eigenvector. To show the second part, use the same argument but restrict x to the space where $Bx \neq 0$. For the third part, let x be a vector such that $x^T Bx < 0$. Then $x^T(\tau B - A)x < 0$ for any sufficiently large τ, so the support is unbounded (infinite). ∎

The support number can therefore be interpreted as an extension of generalized eigenvalues that is robust under rank-deficiency. When both matrices are spd, then the (generalized) condition number is the ratio of the largest to smallest generalized eigenvalues.

PROPOSITION 2.4. *When A and B are both spd, the generalized condition number $\kappa(A,B)$ satisfies $\kappa(A,B) = \sigma(A,B)\sigma(B,A)$.*

Proof. By Theorem 2.3, $\sigma(A,B) = \lambda_{\max}(A,B)$, and therefore $\sigma(B,A) = 1/\lambda_{\min}(A,B)$. ∎

The condition number is unbounded (infinite) if either A or B is rank deficient, but $\sigma(A,B)\sigma(B,A)$ may still be finite and can therefore be viewed as a more robust generalization of the condition number. In practice one should be cautious about using a singular matrix as a preconditioner.

Our technique to bound the support of (A,B) is to break the matrices up into pieces which are in some sense *simpler*. In the sections that follow, *simple* can mean different things, for example, sparse and of low rank. We will rely heavily upon the following splitting principle, a slight variation of Lemma 4.7 in [11].

PROPOSITION 2.5 (splitting). *Split A and B into $A = A_1 + A_2 + \cdots + A_q$ and $B = B_1 + B_2 + \cdots + B_q$. If all B_i are psd, then $\sigma(A,B) \leq \max_i \sigma(A_i, B_i)$.*

The key to proving good support bounds is to find good splittings of A and B. (We remark that "multisplitting" might be a more appropriate term since the matrices can be split into several parts.) In our framework, each B_i must be psd, while there is no restriction of the definiteness of A_i. However, in practice we usually employ splittings where all the A_i are also spsd.

An important observation for using support theory is that one may use different splittings of A and B when proving bounds on $\sigma(A,B)$ and $\sigma(B,A)$. Different splittings may give quite different bounds on the condition number, so identifying good splittings is crucial.

In some applications, there is a natural splitting of the form $A = \Sigma_i A_i$. For example, in finite element analysis, A could correspond to the global mass or stiffness

matrix, while each A_i corresponds to an element matrix. Analysis by splitting into element matrices is a technique used by several authors and goes back at least to the early 1970s. Irons and Treharne [17] described the splitting theorem in the context of finite elements as "a familiar but undervalued theorem" and advocated that it should be taught in finite element courses. More recently, Wathen [30] and Lee and Wathen [18] used the splitting property to prove upper and lower eigenvalue bounds for element-by-element preconditioners. Similar splittings are also used in domain decomposition [26]. We do not discuss finite elements any further here because it is outside the scope of the present paper.

3. Fundamental properties of support numbers. We state some fundamental properties of the support number and skip the simplest proofs.

PROPOSITION 3.1. *When A is psd and $\alpha \neq 0$, then $\sigma(\alpha A, A) = \alpha$.*

PROPOSITION 3.2. *Let B be psd and $\alpha > 0$. Then $\sigma(\alpha A, B) = \alpha\sigma(A, B)$ and $\sigma(A, \alpha B) = \alpha^{-1}\sigma(A, B)$.*

PROPOSITION 3.3. *If B is psd, then*

$$\sigma(A + C, B) \leq \sigma(A, B) + \sigma(C, B).$$

PROPOSITION 3.4. *If B and C are psd, then*

$$\sigma(A, B + C) \leq \frac{\sigma(A, B)\sigma(A, C)}{\sigma(A, B) + \sigma(A, C)} \leq \frac{1}{2} \max\left\{\sigma(A, B), \sigma(A, C)\right\}.$$

Proof. Using Propositions 3.2 and 2.5, we have that $\sigma(A, B + C) = \sigma(\frac{1}{2}A + \frac{1}{2}A, B + C) \leq \frac{1}{2}\max\{\sigma(A, B), \sigma(A, C)\}$, which proves the weaker bound. The stronger bound is derived similarly by a splitting $A = \alpha A + (1-\alpha)A$ for α such that $\alpha\sigma(A, B) = (1 - \alpha)\sigma(A, C)$. ∎

PROPOSITION 3.5. *If B and C are psd, then*

$$\sigma(A, B) \leq \sigma(A + C, B).$$

When A and $B - C$ are also psd, then

$$\sigma(A, B) \leq \sigma(A, B - C).$$

The triangle inequality holds for support numbers.

PROPOSITION 3.6. *Suppose that B and C are psd. Then*

$$\sigma(A, C) \leq \sigma(A, B)\sigma(B, C).$$

Note that none of the propositions in this section so far require symmetry. The support number essentially ignores the nonsymmetric part of the matrices, as shown below.

PROPOSITION 3.7. *Suppose that B is psd. Then $\sigma(A, B) = \sigma(A^T, B) = \sigma(A, B^T)$, and hence*

$$\sigma(A, B) = \sigma(Sym(A), Sym(B)),$$

where $Sym(X) \equiv \frac{1}{2}(X + X^T)$ denotes the symmetric part of X.

Proof. The result follows from Definition 2.1 and the fact that $x^T Ax = x^T A^T x$ for any square (not necessarily symmetric) matrix. ∎

COROLLARY 3.8. *Suppose that A, B, and C are spsd. Then*

$$\sigma(AC, B) = \sigma(CA, B) \quad \textit{and} \quad \sigma(A, BC) = \sigma(A, CB).$$

Proof. By using Proposition 3.7 and the symmetry of A and C, we have that $\sigma(AC, B) = \sigma((AC)^T, B) = \sigma(C^T A^T, B) = \sigma(CA, B)$. Similarly for the second part. ∎

We will use a well-known eigenvalue result; see, for example, Corollary 3.14 in [2].

LEMMA 3.9. *Let A and B be spsd matrices of the same order. Then*

$$\lambda_{\max}(AB) \leq \lambda_{\max}(A)\lambda_{\max}(B).$$

Using this lemma and Theorem 2.3, we get the following results for symmetric matrices.

PROPOSITION 3.10. *When A, B, and C are all spsd, then*

$$\sigma(AC, B) \leq \lambda_{\max}(C)\sigma(A, B).$$

Proof. Suppose that B is nonsingular. Then $\sigma(AC, B) = \lambda_{\max}(B^{-1}AC) \leq \lambda_{\max}(B^{-1}A)\lambda_{\max}(C) \leq \lambda_{\max}(C)\sigma(A, B)$. If B is singular, the same argument holds in a subspace (the range of B). ∎

The next proposition extends lemmas that were used by Gremban [11] and by Bern et al. [6] to partially factor a matrix and preconditioner while maintaining a bound on the support number.

PROPOSITION 3.11. *Let $B \in \mathbb{R}^{n\times n}$ be spsd. Then for any $G \in \mathbb{R}^{n\times p}$,*

$$\sigma(G^T AG, G^T BG) \leq \sigma(A, B),$$

and if $Null(G^T) \subseteq Null(A)$ and $Null(G^T) \subseteq Null(B)$, then

$$\sigma(G^T AG, G^T BG) = \sigma(A, B).$$

Proof. Let $\tau = \sigma(A, B)$. Then $x^T(\tau B - A)x \geq 0$ for all x. For any $y \in \mathbb{R}^p$, let $x = Gy$. Then $y^T G^T(\tau B - A)Gy \geq 0$, and it follows that $\sigma(G^T AG, G^T BG) \leq \tau$. This proves the first part of the proposition. For the second part, note that $Null(G^T) = \text{Range}(G)^{\perp}$. Any vector $x \in \mathbb{R}^n$ can be split into two parts, $x = \hat{x} + \tilde{x}$, where $\hat{x} \in \text{Range}(G)$ and $\tilde{x} \in Null(G^T)$. Suppose $Null(G^T) \subseteq Null(A)$ and $Null(G^T) \subseteq Null(B)$. It follows that $x^T(\tau B - A)x = \hat{x}^T(\tau B - A)\hat{x}$, and since $\hat{x} \in \text{Range}(G)$ there exists y such that $\hat{x} = Gy$. ∎

PROPOSITION 3.12. *Suppose that A and B are spd. Then $\sigma(A, B) = \sigma(B^{-1}, A^{-1})$.*

Proof. First consider the case where $B = I$. Let $C = A^{1/2}$ be a symmetric square root of A, that is, $A = CC^T = C^2$. From Proposition 3.11 (with $G = C^{-1}$) it follows that

$$\sigma(A, I) = \sigma(C^{-T}AC^{-1}, C^{-T}C^{-1}) = \sigma(I, A^{-1}).$$

The general case where $B \neq I$ can be reduced to the case where $B = I$. Let $B^{1/2}$ denote a symmetric square root of B. Then $\sigma(A, B) = \sigma(B^{-1/2}AB^{-1/2}, I)$ and $\sigma(B^{-1}, A^{-1}) = \sigma(I, B^{1/2}A^{-1}B^{1/2})$, and the desired reduction is complete. ∎

The next result is a slight generalization of Lemma 3.3 in [6], which was used to prove a bound on modified incomplete Cholesky preconditioners.

PROPOSITION 3.13. *When A and B are psd, then*

$$\sigma(A,B) \le \frac{1}{1-\sigma(A-B,A)}.$$

Proof. Let $\tau' = \sigma(A-B,A)$. Observe that $\tau' \le 1$ because $A-(A-B) = B$ is psd. Also, $\tau' A-(A-B)$ is psd by Definition 2.1. We have that $\tau' A-(A-B) = B-(1-\tau')A$; hence $\frac{1}{1-\tau'}B - A$ is also psd because $\tau' \le 1$. Consequently, $\sigma(A,B) \le \frac{1}{1-\tau'}$. ∎

The following proposition may be useful when A and B are spsd but not diagonally dominant since there are more efficient algorithms for solving diagonally dominant systems. By choosing C to be diagonal with sufficiently large positive elements, $A+C$ and $B+C$ can be made diagonally dominant.

PROPOSITION 3.14. *Suppose A and B are psd. Then for any psd C and $\alpha > 0$ such that $\alpha\sigma(A+C, B+\alpha C) \le 1$, then*

$$\sigma(A,B) \le \sigma(A+C, B+\alpha C).$$

Proof. For any $\alpha > 0$ there exists a τ such that $\tau(B+\alpha C) - (A+C)$ is spsd. Consequently, $\tau B - A$ is spsd when $(1-\tau\alpha)C$ is spsd. By assumption, $\tau\alpha \le 1$, so the desired result follows. ∎

When A and B have block diagonal structure, the support number can be computed by looking at the blocks independently and taking the maximum. This is a special case of splitting where equality holds.

PROPOSITION 3.15. *Suppose B is psd and A, B are of the form*

$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}.$$

Then $\sigma(A,B) = \max\{\sigma(A_{11},B_{11}), \sigma(A_{22},B_{22})\}$.

In some situations it is helpful to obtain a support bound by expanding the matrices into a higher dimension. The following proposition explains how.

PROPOSITION 3.16. *Let A_{11}, B_{11} denote principal submatrices of A and B, respectively. Then $\sigma(A_{11},B_{11}) \le \sigma(A,B)$.*

Proof. Let $\tau = \sigma(A,B)$. Then $\tau B - A$ is psd. Any principal submatrix of $\tau B - A$ is also psd; in particular, $\tau B_{11} - A_{11}$. ∎

4. Main support results. This section contains our main results. Recall from Proposition 2.5 that we want to break A and B into sums of *simple* pieces. A key kind of simplicity that we will exploit is to have the pieces be of low rank. We can exploit the fact that symmetric rank-1 and rank-2 matrices have spectra that are simple to express.

LEMMA 4.1. *Let $A = uu^T$. Then all eigenvalues of A are zero except $\lambda_1(A) = u^T u$. Furthermore, if B is invertible (nonsingular), then all generalized eigenvalues of (A,B) are zero except $\lambda = u^T B^{-1} u$.*

LEMMA 4.2. *Let $A = uv^T + vu^T$. Then all the eigenvalues of A are zero except $\lambda_{1,2}(A) = \pm\|u\|_2\|v\|_2 + u^T v$.*

Lemma 4.1 gives us a formula for the support for a symmetric rank-1 matrix A.

PROPOSITION 4.3. *Let $A = uu^T$ and let B be spd. Then*

$$\sigma(A,B) = u^T B^{-1} u.$$

Proof. From Theorem 2.3 we have that $\sigma(A,B) = \lambda_{\max}(A,B)$. By Lemma 4.1, all the eigenvalues $\lambda(A,B)$ are zero except one, which is $u^T B^{-1} u$. Since B is spd, $u^T B^{-1} u > 0$ for any u, so $\lambda_{\max}(A,B) = u^T B^{-1} u$. ∎

Next we show a more general result that includes the case where B is semidefinite and does not have full rank.

THEOREM 4.4 (rank-1 support theorem). *Suppose $u \in \mathbb{R}^n$ is in the range of $V \in \mathbb{R}^{n\times k}$. Then*

$$\sigma(uu^T, VV^T) = \min_w w^T w \quad \textit{subject to } Vw = u.$$

Proof. Let w be a vector that satisfies $Vw = u$. By applying Proposition 3.11, we get

$$\sigma(uu^T, VV^T) = \sigma(Vww^TV^T, VV^T) \le \sigma(ww^T, I) = w^T w.$$

Next we prove that there exists a w such that equality holds. The smallest norm solution to $Vw = u$ is given by $w = V^+u$, where V^+ is the Moore–Penrose pseudoinverse of V [10, p. 243]. We have that $\sigma(uu^T, VV^T) = \lambda_{\max}(V^+uu^T(V^+)^T) = \|V^+u\|_2^2$. ∎

We remark that any w satisfying $Vw = u$ gives an upper bound on $\sigma(uu^T, VV^T)$. Further observe that when V has full column rank, then there is a unique w such that $Vw = u$. The theorem above can also be restated in terms of the pseudoinverse, that is, $\sigma(uu^T, VV^T) = \|V^+u\|_2^2$.

Note that all spsd matrices can be constructed as a sum of symmetric outer products like those in the theorem. For instance, the Cholesky decomposition (in outer-product form) provides such a splitting. However, there are many alternatives, and the Cholesky decomposition may not be the best choice for proving bounds or building preconditioners.

In the special case where each column of U and V has only two nonzero entries and these entries have the same magnitude, this proposition reduces to the congestion-dilation lemma discussed in section 9. The congestion-dilation lemma is based on a specific graph interpretation that we will examine in section 9 and is the cornerstone of support-graph theory [11, 6]. In support-graph theory, the vector u with its two nonzeros in locations i and j represents an edge between vertices i and j, and the set of columns of V corresponds to a path (a sequence of edges) between the same vertices. Unfortunately, only a very limited class of matrices can be represented as sums of outer products of these specialized vectors. Specifically, as discussed in section 9, if the two values are of the opposite sign, then all symmetric, diagonally dominant, psd M-matrices can be generated. And if values of the same sign are included, then the class grows to be all symmetric, diagonally dominant, psd matrices. Support-graph theory is limited to these classes of matrices. But with a general u, the much more important class of spsd matrices can be addressed.

We next state the higher-rank generalization of Theorem 4.4.

THEOREM 4.5 (symmetric product support). *Suppose $U \in \mathbb{R}^{n\times k}$ is in the range of $V \in \mathbb{R}^{n\times p}$. Then*

$$\sigma(UU^T, VV^T) = \min_W \|W\|_2^2 \quad \textit{subject to } VW = U.$$

Proof. Let W satisfy $VW = U$. Then

$$\begin{aligned}\sigma(UU^T, VV^T) &= \sigma(VWW^TV^T, VV^T) \le \sigma(WW^T, I)\\ &= \lambda_{\max}(WW^T) = \|W\|_2^2.\end{aligned}$$

As in the proof of Theorem 4.4, one can show that equality is achieved for $W = V^+U$. ∎

We will often use this theorem as a tool for obtaining an upper bound on $\sigma(UU^T, VV^T)$. Note that any W for which $VW = U$ provides an upper bound on the support number. One special case of interest is when the columns of U are a subset of the columns of V (or vice versa).

COROLLARY 4.6. *Suppose the columns of U are a subset of the columns of V. Then $\sigma(UU^T, VV^T) \leq 1$.*

The result above follows by letting W be an appropriate subset of the identity matrix, so $\|W\|_2^2 \leq 1$. Alternatively, it is easy to show that $VV^T - UU^T$ is spsd, which also gives a bound of one for the support number.

The following theorem is a slight generalization of Theorem 4.5.

THEOREM 4.7. *Suppose $U \in \mathbb{R}^{n\times k}$ is in the range of $V \in \mathbb{R}^{n\times p}$ and let $D \in \mathbb{R}^{k\times k}$ be symmetric. Then*

$$\sigma(UDU^T, VV^T) \leq \lambda_{\max}(WDW^T) \leq \lambda_{\max}(D)\|W\|_2^2$$

for all W such that $VW = U$.

Proof. Let W satisfy $VW = U$. Then

$$\sigma(UDU^T, VV^T) = \sigma(VWDW^TV^T, VV^T) \leq \sigma(WDW^T, I) = \lambda_{\max}(WDW^T),$$

which proves the first part. The second follows from $\lambda_{\max}(WDW^T) = \lambda_{\max}(DW^TW)$ $\leq \lambda_{\max}(D)\lambda_{\max}(W^TW) = \lambda_{\max}(D)\|W\|_2^2$. ∎

Recall that the support number may be negative.

COROLLARY 4.8. *Suppose $U \in \mathbb{R}^{n\times k}$ is in the range of $V \in \mathbb{R}^{n\times p}$ and let D be a block diagonal matrix in $\mathbb{R}^{k\times k}$, where the blocks are either of the type ± 1 or $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. Then $\sigma(UDU^T, VV^T) \leq \|W\|_2^2$ for all W such that $VW = U$.*

Proof. The eigenvalues of D can only take on two different values: 1 or -1. Hence $\lambda_{\max}(D) \leq 1$, and the result follows from Theorem 4.7. ∎

We remark that any symmetric matrix (possibly indefinite) has a decomposition of the type UDU^T, where U is square and lower triangular and D is as described in the corollary above. However, this may not be the best way to apply the corollary.

Further note that $\|W\|_2^2$ may be expensive to compute. Nonetheless, as is well known, the 2-norm can be bounded by easy-to-compute quantities.

LEMMA 4.9. *For any matrix W, we have that*

(i) $\|W\|_2^2 \leq \|W\|_1\|W\|_\infty = (\max_j \sum_i |W_{ij}|)\,(\max_i \sum_j |W_{ij}|)$,

(ii) $\|W\|_2^2 \leq \|W\|_F^2 = \sum_{i,j} W_{ij}^2$.

Most of the preceding set of results have involved symmetric outer products to construct low rank matrices. We now extend the rank-1 support theorem to the rank-2 case.

THEOREM 4.10. *Suppose $u, v \in \mathbb{R}^n$ are in the range of $Y \in \mathbb{R}^{n\times k}$. Then*

$$\sigma(uv^T + vu^T, YY^T) \leq \|w\|_2\|\hat{w}\|_2 + w^T\hat{w}$$

for any w and $\hat{w}$ such that $Yw = u$ and $Y\hat{w} = v$.

Proof.

$$\begin{aligned}
\sigma(uv^T + vu^T, YY^T) &= \sigma(Y(w\hat{w}^T + \hat{w}w^T)Y^T, YY^T) \\
&\leq \sigma(w\hat{w}^T + \hat{w}w^T, I) \\
&= \lambda_{\max}(w\hat{w}^T + \hat{w}w^T) \\
&= \|w\|_2\|\hat{w}\|_2 + w^T\hat{w} \quad \text{by Lemma 4.2.}
\end{aligned}$$

∎

COROLLARY 4.11. *Suppose $u, v \in \mathbb{R}^n$ are in the range of $Y \in \mathbb{R}^{n\times k}$. Then*

$$\sigma(uv^T + vu^T, YY^T) \le 2\|w\|_2\|\hat{w}\|_2 \le \|w\|_2^2 + \|\hat{w}\|_2^2$$

for all $w, \hat{w}$ such that $Yw = u$, $Y\hat{w} = v$.

Proof. The result follows from Theorem 4.10 and the Cauchy–Schwarz inequality. ∎

We can extend Theorem 4.10 to the case where U and V are matrices.

THEOREM 4.12. *Suppose $U, V \in \mathbb{R}^{n\times p}$ are in the range of $Y \in \mathbb{R}^{n\times k}$. Then*

$$\sigma(UV^T + VU^T, YY^T) \le \lambda_{\max}(W\hat{W}^T + \hat{W}W^T) \le 2\|W\|_2\|\hat{W}\|_2$$

for any W and $\hat{W}$ such that $YW = U$ and $Y\hat{W} = V$.

We omit the proof because it is essentially a combination of the proofs of Theorem 4.10 and of Corollary 4.11.

5. Diagonal support. In section 4 we described tools for bounding support numbers when the pieces involved have low rank. Another kind of simple structure we can exploit occurs when one of the matrices is diagonal. Any matrix can be supported by a positive diagonal matrix. We remark that computing the exact support $\sigma(A, B)$ when B is diagonal is not much easier than for a general spd B and requires the computation of an extremal eigenvalue.

Fortunately, we will see that it is easy to obtain a bound. We need the following well-known fact, which is easily derived from Gerschgorin's theorem.

LEMMA 5.1. *If A is symmetric, weakly (strictly) diagonally dominant, and has nonnegative diagonal entries, then A is spsd (spd).*

Using the above lemma, one way to bound $\sigma(A, B)$ is to find τ such that $\tau B - A$ is diagonally dominant with positive diagonal entries. Unfortunately, this strategy only works for certain B and, further, computing the optimal value of τ may require the solution of a linear program. However, when B is diagonal we can obtain a bound as follows.

THEOREM 5.2. *Suppose A is symmetric (not necessarily spd) and B is diagonal with $b_{ii} \ge 0$ for all i. Assume that $W = \{w_{ij}\}$ satisfies $w_{ij} > 0$ and $w_{ij} = 1/w_{ji}$ for all i and j, and that $b_{ii} = 0$ only if $a_{ii} + \sum_{j\neq i} w_{ij}|a_{ij}| \le 0$. Then*

$$\sigma(A, B) \le \max_i \left\{ \frac{a_{ii} + \sum_{j\neq i} w_{ij}|a_{ij}|}{b_{ii}} \right\}, \; b_{ii} \neq 0.$$

Proof. We will describe how to find an spsd matrix $\hat{A}$ such that $D \equiv A + \hat{A}$ is diagonal. From Proposition 3.5 it follows that $\sigma(A, B) \le \sigma(A + \hat{A}, B) = \sigma(D, B)$. Let $\hat{A} = \sum_{ij} \hat{A}_{ij}$, where $\hat{A}_{ij}$ is chosen to cancel out the off-diagonal element a_{ij}. Specifically, $\hat{A}_{ij}$ is zero except in rows and columns i and j, where it is

$$\begin{pmatrix} |a_{ij}|/w_{ij} & -a_{ij} \\ -a_{ij} & |a_{ij}|w_{ij} \end{pmatrix} = \begin{pmatrix} |a_{ij}|w_{ji} & -a_{ij} \\ -a_{ij} & |a_{ij}|w_{ij} \end{pmatrix}.$$

Consequently, $D = A + \hat{A}$ is diagonal. By simple algebra, $d_{ii} = a_{ii} + \sum_{j\neq i} w_{ij}|a_{ij}|$, and the desired result follows. ∎

By setting $B = I$, we obtain an interesting eigenvalue bound.

COROLLARY 5.3. *Let A be a symmetric matrix (not necessarily spd). Then for any positive matrix W such that $w_{ij} = 1/w_{ji}$ for all i and j,*

$$\lambda_{\max}(A) \le \max_i \left\{ a_{ii} + \sum_{j \neq i} w_{ij} |a_{ij}| \right\}.$$

By setting all the w_{ij} values to be 1, we get a different special case.

COROLLARY 5.4. *Suppose A is symmetric (not necessarily spd), $B \ge 0$ is diagonal, and $b_{ii} = 0$ only if $a_{ii} + \sum_{j \neq i} |a_{ij}| \le 0$. Then*

$$\sigma(A, B) \le \max_i \left\{ \frac{a_{ii} + \sum_{j \neq i} |a_{ij}|}{b_{ii}} \right\}, \; b_{ii} \neq 0.$$

When $B = I$ and all the w_{ij} values are 1, then each of these corollaries reduces to Gerschgorin's well-known bound on the maximal eigenvalue. Furthermore, Theorem 5.2 contains as a special case the scaled Gerschgorin bound obtained by diagonal scaling of A, that is, the Gerschgorin eigenvalue bound for SAS^{-1} where S is diagonal.

How can we choose W to improve the bound? Computing the optimal W is difficult and could even be more expensive than computing $\lambda_{\max}(A)$ directly. Intuitively, we want to choose w_{ij} small when row i has a large (absolute) row sum, i.e., when $a_{ii} + \sum_{k \neq i} |a_{ik}|$ is large. One possible such strategy is to let

$$w_{ij} = \frac{a_{jj} + \sum_{k \neq j} |a_{jk}| - a_0}{a_{ii} + \sum_{k \neq i} |a_{ik}| - a_0},$$

where $a_0 = \min_i a_{ii}$. (Because we subtract a_0, the bound is invariant under shifting of the eigenvalues.) We remark that the proposed bound is often, but not always, better than the Gerschgorin bound. For example, for

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 6 & 3 \\ 1 & 3 & 9 \end{pmatrix},$$

the Gerschgorin bound is 13 but our new bound is 11.7. The largest eigenvalue is 11.3.

An alternative approach is to start out with $w_{ij} \equiv 1$ and then iteratively pick an entry w_{ij} to adjust. Keeping all other coefficients fixed, one can compute a new value for w_{ij} that tightens the eigenvalue bound.

We note that tighter bounds may be obtained by using matrices with nonzeros in more than two rows (columns) to cancel out positive off-diagonals. Such a strategy requires finding cliques in the graph of the matrix. We do not examine this option any further here.

A technique used by several previous authors for preconditioning diagonally dominant matrices is to first subtract a diagonal matrix such that the remaining part is semidefinite and rank deficient. Then one preconditions the semidefinite part using support theory and adds back the diagonal part. The following lemma is used. (Note that in this and the subsequent lemmas, D is a general spsd matrix, but for current purposes we are interested in the case where D is diagonal.)

LEMMA 5.5. *If A is symmetric and B and D are spsd, then $\sigma(A + D, B + D) \le \max\{\sigma(A, B), 1\}$.*

Clearly, the diagonal elements are not fully exploited in this approach. Basically, B supports A while D supports only itself. Going to the other extreme, we could let

D support both A and D, which yields $\sigma(A+D, B+D) \le \sigma(A+D, D) \le \sigma(A, D)+1$. This method is also unsatisfactory because B is not utilized at all. A better approach is to let parts of D support A and parts of it support itself. From this idea we obtain the following result.

PROPOSITION 5.6. *If A is symmetric and B and D are spsd, then*

$$\sigma(A+D, B+D) \le \frac{1+\sigma(A,D)}{1+\sigma(A,D)/\sigma(A,B)}.$$

Proof. We use the splitting $B + D = (B + \alpha D) + (1 - \alpha)D$, and by applying Propositions 2.5 and 3.4 we find that

$$\begin{aligned}\sigma(A+D, B+D) &\le \max\{\sigma(A, B+\alpha D), \sigma(D, (1-\alpha)D\} \\ &\le \max\left\{\frac{\sigma(A,B)+\sigma(A,D)}{\alpha\sigma(A,B)+\sigma(A,D)}, \frac{1}{1-\alpha}\right\}\end{aligned}$$

for any α such that $0 < \alpha < 1$. We want the tightest possible bound, which occurs when the two arguments in max are equal. Hence we solve, for α, the equation

$$(1-\alpha)(\sigma(A,B)+\sigma(A,D)) = \alpha\sigma(A,B)+\sigma(A,D),$$

which has the solution

$$\alpha = \frac{\sigma(A,D)(1+\sigma(A,B)}{\sigma(A,B)(1+\sigma(A,D))}.$$

The desired support bound is $1/(1-\alpha)$, which after some algebra is shown to equal

$$\frac{1}{1-\alpha} = \frac{\sigma(A,B)(1+\sigma(A,D))}{\sigma(A,B)+\sigma(A,D)} = \frac{1+\sigma(A,D)}{1+\sigma(A,D)/\sigma(A,B)}. \qquad \square$$

6. Schur complement support. Another special matrix structure that commonly arises in practice is the Schur complement—the remaining portion of a matrix after a subset of rows and columns has been factored (by Gaussian elimination). This section contains tools to address this special matrix structure.

A matrix can be supported in a "higher-dimensional space" using the Schur complement.

PROPOSITION 6.1. *Let A and B be spsd and of the form*

$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix},$$

where B_{22} is nonsingular. Then $\sigma(A,B) = \sigma(A_{11}, B_{11} - B_{12}B_{22}^{-1}B_{12}^T)$.

Proof. Let $G^T = \left(\begin{smallmatrix} I & -B_{12}B_{22}^{-1} \\ 0 & I \end{smallmatrix}\right)$, which is always nonsingular. Let S denote the Schur complement $B_{11} - B_{12}B_{22}^{-1}B_{12}^T$. It is easy to verify that $G^TAG = A$ and $G^TBG = \left(\begin{smallmatrix} S & 0 \\ 0 & B_{22} \end{smallmatrix}\right)$. By Proposition 3.11, $\sigma(G^TAG, G^TBG) = \sigma(A,B)$. Since the lower right block of A is zero, the support number is determined by the upper left blocks of the block diagonal matrix pencil (G^TAG, G^TBG), and we have that $\sigma(G^TAG, G^TBG) = \sigma(A_{11}, S)$. $\square$

A useful special case of the preceding result is as follows.

COROLLARY 6.2. *Suppose A and B are spsd and of the form*

$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \alpha A_{11} + VV^T & \beta V \\ \beta V^T & \beta^2 I \end{pmatrix},$$

where $\alpha > 0$, $\beta \neq 0$, and V is any matrix of appropriate dimensions. Then

$$\sigma(A, B) = 1/\alpha.$$

Proof. Proposition 6.1 yields

$$\begin{aligned}\sigma(A, B) &= \sigma(A_{11}, \alpha A_{11} + VV^T - \beta V \beta^{-2} \beta V^T) \\ &= \sigma(A_{11}, \alpha A_{11}) = 1/\alpha. \qquad \square\end{aligned}$$

This corollary contains the clique-star lemma from [11, 6] as a special case, where $\alpha = 1/k$, $\beta = 1$, $A_{11} = kI - ee^T$, $V = e$, and e is a vector of all ones. The clique-star lemma was used by Gremban [11] in the analysis of multilevel support-graph preconditioners (see also [6]).

7. Hadamard product support. In this section we restate some known results about eigenvalues and Hadamard products in terms of support numbers. The *Hadamard* product is the elementwise matrix product; that is, if $C = A \circ B$, then $c_{ij} = a_{ij} b_{ij}$ for all i, j. Schur [25] proved several properties of the Hadamard product, including the important results below.

LEMMA 7.1. *If A and C are both spsd, then*

$$\lambda_{\min}(A)\lambda_{\min}(C) \leq \lambda_i(A \circ C) \leq \lambda_{\max}(A)\lambda_{\max}(C) \text{ for all } i.$$

COROLLARY 7.2. *If A and C are both spsd, then $A \circ C$ is also spsd.*

The next proposition follows directly from Schur's results.

PROPOSITION 7.3. *If A, B, and C are spsd, then*

$$\sigma(A \circ C, B \circ C) \leq \sigma(A, B).$$

Proof. Let $\tau = \sigma(A, B)$, so $\tau B - A$ is spsd. By Corollary 7.2, $(\tau B - A) \circ C = \tau(B \circ C) - A \circ C$ is also spsd for any spsd C. $\square$

Restating a variation of Schur's result [23, Lemma 2.1] in support theory notation, we get the proposition below.

PROPOSITION 7.4. *Suppose A is spsd and C is symmetric. Let D_A denote the diagonal matrix with the same diagonal as A. Then*

$$\sigma(A \circ C, D_A) \leq \lambda_{\max}(C).$$

If C is spd, then we also have

$$\sigma(D_A, A \circ C) \leq \frac{1}{\lambda_{\min}(C)}.$$

Fiedler and Markham [9] proved the following result.

PROPOSITION 7.5. *Suppose A is spsd and C is spd. Then*

$$\sigma(A, A \circ C) \leq e^T C^{-1} e,$$

where e is the all-ones vector.

This result may be useful in our context when, for example, the preconditioner B has a sparsity pattern that is a subset of the nonzeros of A, so there exists a C such that $A \circ C = B$. As a simple example, consider the case when B (and hence also C) is diagonal. Then $\sigma(A, B) \leq \sum_i (a_{ii}/b_{ii})$. Observe that when $B = I$ this bound reduces to the well-known trace bound, $\lambda_{\max}(A) \leq \operatorname{tr}(A) = \sum_i a_{ii}$.

Recently, several extensions to the Fiedler–Markham result (Proposition 7.5) have been developed [23, 15]. These extensions hold when C is either positive definite or *conditionally* positive definite, that is, positive definite in a subspace.

8. Supporting negative semidefinite parts. It is trivial to support a negative semidefinite matrix.

PROPOSITION 8.1. *If A is negative semidefinite, then $\sigma(A,0) = -\infty$. Furthermore, $\sigma(A,B) \le 0$ for any psd B.*

This proposition gives us two preconditioning strategies when applied to a part A_i of a matrix A. First, any negative semidefinite part of A can be ignored (preconditioned by 0). We remark that a better condition number bound may possibly be obtained by utilizing the negative semidefinite part. Second, we can add any psd matrix B_i to a preconditioner B and the support number $\sigma(A,B)$ will not increase. Implicitly, there is a corresponding term $A_i = 0$, so $\sigma(A_i, B_i) \le 0$ for any psd B_i. It may seem strange to make the preconditioner B more complicated than necessary, but in fact B can often be made "simpler" (for example, sparser) by adding additional psd terms. This strategy is particularly well suited for canceling out off-diagonal elements that make the preconditioner hard to factor.

Recall that when we split a preconditioner B into parts, $B = \sum_i B_i$, we normally require that all B_i be psd. There is one exception to this rule. A matrix B_i may be indefinite or negative definite if it is supported by a set of psd matrices $\sum_{j\in S} B_j$ with support at most one. The combined matrix $B_i + \sum_{j\in S} B_j$ is then psd. In the expression

$$\tau B - A = \tau \sum_{i=1}^{k'} B_i - \sum_{i=1}^{k} A_i,$$

A and B are not necessarily decomposed into the same number of terms; that is, $k' \neq k$ is allowed. Hence some terms in B can be used to support non-psd terms in B. A special case of this technique was used by Bern et al. [6, section 3.2].

9. Laplacian matrices and support graphs. As mentioned in the introduction, several previous authors have analyzed preconditioners using a closely related technique called *support-graph theory*. In this section we review the essentials of support-graph theory and show that they are a special case (albeit a very useful one) of our basis support results from section 4. Specifically, in Theorem 4.4 we showed how to support a rank-1 matrix uu^T with a larger symmetric matrix VV^T. In support-graph theory the vectors u and the columns of v are generally limited to have two nonzeros each. And the two nonzeros are of equal magnitude. Recall that a basic tool in support theory is to split a general matrix into simpler parts. What classes of matrices can be split into sums of such restricted outer products?

Consider first the case where the two nonzeros in u are of opposite sign, so $u_i = \sqrt{\alpha}$ and so $u_j = -\sqrt{\alpha}$. Then the nonzero portion uu^T (in rows/columns i and j) is

$$\begin{pmatrix} \alpha & -\alpha \\ -\alpha & \alpha \end{pmatrix}.$$

A positive linear combination of such matrices can produce any matrix that is spsd, diagonally dominant, has nonpositive off-diagonal elements, and has zero row sums. We call this class of matrices *Laplacians* (Gremban called them *generalized Laplacians* [11]). This class of matrices includes many standard discretizations of Laplace's or Poisson's equation and other elliptic equations and so is quite important in practice. By also including u vectors with a single nonzero, one can augment the diagonal values, thus allowing matrices with positive row sums. This corresponds to different (e.g., Dirichlet) boundary conditions in the differential equation.

If we also allow the two nonzeros in the u vector to be of equal sign, then the nonzero contribution from uu^T is

$$\begin{pmatrix} \alpha & \alpha \\ \alpha & \alpha \end{pmatrix}.$$

Any positive linear combination of such matrices is spsd and diagonally dominant, but now the off-diagonal values are nonnegative. Combining all these observations, it is easy to show the following.

PROPOSITION 9.1. *A symmetric matrix A with nonnegative diagonal entries is diagonally dominant if and only if there exists a decomposition of the form $A = UU^T$, where each column of U has either one nonzero or exactly two nonzero entries and these two entries have the same magnitude. Furthermore, if all off-diagonal entries of A are nonpositive, then A is also an M-matrix, and any column of U with two nonzeros has entries of opposite signs.*

The columns of U are easy to construct in linear time. Each symmetric pair of off-diagonal nonzeros in A corresponds to a single column of U. Additional columns of U can be added to augment the diagonals. This correspondence between nonzeros of A and simple columns of U can be expressed in terms of graphs. Specifically, consider the rows of the symmetric matrix A to be vertices of a graph, and for each nonzero off-diagonal a_{ij} add an edge between vertices i and j with weight equal to a_{ij}. Note that each such edge corresponds to a column of U. This relationship between Laplacian matrices, and more generally, diagonally dominant matrices, and graphs is at the heart of support-graph theory.

Key tools in support-graph theory are various forms of what are called congestion-dilation lemmas. Here we show that they follow directly from Theorem 4.4. A *path* between vertices i and j is a series of edges which leads from i to j. Let e^{ij} be a vector corresponding to the edge between i and j in which all elements are zero except for $e^{ij}_i = 1$ and $e^{ij}_j = -1$. Define $E^{ij} = e^{ij}(e^{ij})^T$. Consider the set of vectors comprising a path from i to j. By adding or subtracting these vectors as appropriate, all the intermediate values will cancel and the result will be equal to e^{ij}. In this way, a path can be used to support an edge. In particular, as we state more formally below, the support number is equal to the *dilation*, the number of edges in the path. A preconditioner containing a set of such paths can be built which supports any symmetric, diagonally dominant matrix with nonpositive off-diagonals. This was Vaidya's key observation and is a principal idea in support-graph theory.

Note that a single edge in the preconditioner might be on many such support paths. In this case, the support number also depends on the number of paths it must support—its *congestion*. These observations are made more rigorous in the following results.

PROPOSITION 9.2 (path congestion-dilation). *Suppose $A = aE^{1,k+1}$ for some k and that $B = \sum_{i=1}^{k} b_i E^{i,i+1}$, where $a, b_i > 0$ and E^{ij} is as defined above. Then*

$$\sigma(A, B) = \sum_{i=1}^{k} \frac{a}{b_i}.$$

Proof. From Theorem 4.4 with $u = \sqrt{a}e^{1,k+1}$ and $V = (\sqrt{b_1}e^{1,2}, \sqrt{b_2}e^{2,3}, \ldots, \sqrt{b_k}e^{k,k+1})$ we find that $w = (\sqrt{\frac{a}{b_1}}, \ldots, \sqrt{\frac{a}{b_k}})^T$, and the result follows. ∎

This proposition says that the support is bounded by the sum of the edge congestions along a path. In the simpler case where all edge weights in B are constant

(i.e., $b_i = b$ for all i), the support number is just $\sigma(A,B) = k\,(a/b)$, where k is the length of the path. (This was proven in [6].) The path congestion-dilation proposition is not new; variations have been stated by Gremban [11, Lemma 4.6] and by Guattery [13]. The proposition above was also (implicitly) used by Guattery, Leighton, and Miller [14] in their *path resistance method* to bound the Fiedler eigenvalue of Laplacians.

The preceding proposition considers only the support for a single edge by a single path. More interesting is the case for a set of edges being supported by a set of paths; that is, we have a graph embedding. The set of edges will correspond to a matrix A and the set of paths to a preconditioner B, where both A and B are Laplacians. Represent A and B by graphs G_A and G_B, respectively, and each edge $e \in G_A$ is mapped to a path in G_B that connects the endpoints of e. (Note that a path may be a single edge.) One strategy is to use the splitting proposition and break A into a sum of edges and B into a sum of paths, and apply Proposition 9.2 to each of these pairs. The following result ensues.

PROPOSITION 9.3 (basic graph congestion-dilation). *Given Laplacian matrices A and B, choose a mapping of the edges in the graph G_A onto paths in G_B. For each $e \in E(G_A)$, let $path(e)$ denote the corresponding path in G_B, and let $c(f)$ denote the number of supporting paths an edge f participates in, where $f \in E(G_B)$. Then*

$$\sigma(A,B) \le \max_{e \in E(G_A)} \sum_{f \in path(e)} \frac{a_e\, c(f)}{b_f}.$$

This result is a slight extension of the "worst congestion times worst dilation" bound used in [11, 6]. With our symmetric product theorem (Theorem 4.5), we can show the following stronger result, which to the best of our knowledge is new.

THEOREM 9.4 (graph congestion-dilation). *Given Laplacian matrices A and B, choose a mapping of the edges in the graph G_A onto paths in G_B. For each $e \in E(G_A)$, let $path(e)$ denote the corresponding path in G_B. Then*

$$\sigma(A,B) \le \left(\max_{e \in E(G_A)} \sum_{f \in path(e)} \sqrt{\frac{a_e}{b_f}} \right) \left(\max_{f \in E(G_B)} \sum_{e \mid f \in path(e)} \sqrt{\frac{a_e}{b_f}} \right),$$

and also

$$\sigma(A,B) \le \sum_{e \in E(G_A)} \sum_{f \in path(e)} \frac{a_e}{b_f} = \sum_{f \in E(G_B)} \sum_{e \mid f \in path(e)} \frac{a_e}{b_f}.$$

Proof. Let U, V have the structure described in Proposition 9.1 and $UU^T = A$ and $VV^T = B$. Let $w_{ef} = \sqrt{a_e}/\sqrt{b_f}$, where $e \in E(G_A)$ and $f \in E(G_B)$ if f belongs to $path(e)$. It is straightforward to verify that for appropriately chosen signs (the signs do not affect the norms of W), $W = \{\pm w_{ef}\}$ satisfies $VW = U$. By Theorem 4.5 and Lemma 4.9, $\sigma(A,B) \le \|W\|_1 \|W\|_\infty$ and also $\sigma(A,B) \le \|W\|_F^2$. ∎

In the unweighted case (a_e, b_f, and w_{ef} are 0 or 1), the first bound has a simple interpretation: The first term, $\max_e \sum_f w_{ef}$, is the maximum number of support paths that include any particular edge—that is, the maximum congestion. The second term, $\max_f \sum_e w_{ef}$, is the length of the longest path, or the maximum dilation. Thus the support number is bounded by the product of the maximum congestion and the maximum dilation. In the weighted case, the square roots in the definition of w_{ef} are significant and our result is different from previously used bounds.

The second bound, based on the Frobenius norm, shows that the support number is bounded by the sum of all congestions, or, equivalently, the sum of all dilations in the graph embedding. This bound is tighter than the bound in Proposition 9.3. In the weighted case, the two bounds given in Theorem 9.4 are not comparable.

Theorem 9.4 assumes that each edge in G_A is supported by a unique path in G_B. More generally we can support an edge by a (finite) set of paths. This corresponds to a fractional mapping where each edge weight may be split up into several parts and mapped to different paths in G_B. It is straightforward to extend the theorem to fractional mappings.

Vaidya [28] used the above graph interpretation to construct preconditioners for Laplacian matrices based on *maximum-weight spanning trees*. A spanning tree is a tree that spans all vertices of a given graph, and in which the weight of a tree is the sum of the weights of the edges in the tree. There are efficient algorithms to find spanning trees of maximum weight. One advantage of using a tree is that the corresponding matrix can be factored in linear time with no fill. It is easy to show that the edges of a spanning tree constitute a basis for a graph and hence also for a Laplacian.

Vaidya showed [28, 6] that when A is Laplacian and B is the matrix that corresponds to the maximum-weight spanning tree for the graph of A, then $\sigma(B, A) \leq 1$ and $\sigma(A, B) \leq mn$, where n is the number of vertices and m is the number of edges in the graph. (m is about half the number of nonzeros in A.) This implies that the condition number of the preconditioned system $B^{-1}A$ is at most of order mn, independent of the matrix coefficients. The (upper) bound mn can be reduced by adding additional edges (nonzeros) to the preconditioner, which lowers the condition number but increases the work per iteration in an iterative solver. The optimal trade-off depends on the graph type (e.g., planar).

Vaidya claimed but did not prove that his techniques could be extended to all diagonally dominant matrices (that is, graphs with both positive and negative edge weights). We finally prove this claim in recent work with Chen and Toledo [7] using techniques from the present paper. One key idea is to factor A into $A = UU^T$, where each column of U has at most two nonzeros, but these two elements may have the same sign (cf. Proposition 9.1). The preconditioner $B = VV^T$ is chosen such that the columns of V are a subset of the columns of U, and V is a basis for the range of U.

10. Example: Block Jacobi. In this section, we show how support theory can be used to analyze the well-known block Jacobi preconditioner for a model problem. The analysis is purely algebraic. We reproduce known bounds in a different and perhaps simpler way.

10.1. The one-dimensional model problem. We start with the one-dimensional (higher dimensions will be considered later) Laplace equation with Dirichlet boundary conditions,

$$-u_{xx} = f(x), \quad x \in \Omega = [0, 1].$$

Suppose that Ω has been uniformly discretized using n points, and let $h = 1/n$. We need to solve a system $Au = f$, where A is a tridiagonal matrix with all 2's on the diagonal and -1 on the sub- and superdiagonals, and u and f are discretizations of $u(x)$ and $f(x)$, respectively.

We wish to analyze the block Jacobi method, which corresponds to a simple domain decomposition method without overlap. Let B be the block Jacobi operator for a certain decomposition of A. Note that we do not assume that the blocks have

the same sizes, or, in other words, the subdomains may vary in size. Let q denote the number of subdomains, or, equivalently, the number of diagonal blocks in B.

Consider the following example, where $n = 7$ and $q = 3$:

$$A = \begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & & & \\ & & -1 & 2 & -1 & & \\ & & & -1 & 2 & & \end{pmatrix}, \quad B = \begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & & & & \\ & & & 2 & -1 & & \\ & & & -1 & 2 & & \\ & & & & & 2 & -1 \\ & & & & & -1 & 2 \end{pmatrix}.$$

We now bound the eigenvalues of the preconditioned operator $B^{-1/2}AB^{-1/2}$ using support theory. Recall (see Definition 2.1) that the support number $\sigma(A, B)$ is roughly given by $\sigma(A, B) = \min\{t \mid tB - A \text{ is psd}\}$ and that $\kappa(B^{-1}A) \leq \sigma(A, B)\sigma(B, A)$ (Proposition 2.4). It is easy to bound $\sigma(A, B)$, so the bound that is harder to prove is $\sigma(B, A)$.

LEMMA 10.1. *Let A be the discrete Laplace operator as defined above, and let B be a block diagonal approximation for A formed by dropping some of the off-diagonal entries. Then $\sigma(A, B) \leq 2$.*

Proof. We observe that $2B - A$ is diagonally dominant with positive diagonal and hence psd (by Lemma 5.1). Thus, $\sigma(A, B) \leq 2$ because $t = 2$ in Definition 2.1 ensures that $tB - A$ is psd. ∎

In order to bound $\sigma(B, A)$ we will use the symmetric product support theorem (Theorem 4.5). We factorize $A = VV^T$ and $B = UU^T$, where V is n by $(n + 1)$ and U is n by $(n + q)$. For our example, we obtain

$$V = \begin{pmatrix} 1 & 1 & & & & & & \\ & -1 & 1 & & & & & \\ & & -1 & 1 & & & & \\ & & & \ddots & \ddots & & & \\ & & & & -1 & 1 & & \\ & & & & & -1 & 1 \end{pmatrix}, \; U = \begin{pmatrix} 1 & 1 & & & & & & & \\ & -1 & 1 & & & & & & \\ & & -1 & 1 & & & & & \\ & & & & 1 & 1 & & & \\ & & & & & -1 & 1 & & \\ & & & & & & & 1 & 1 \\ & & & & & & & & -1 & 1 \end{pmatrix}.$$

We seek a matrix W such that $VW = U$. Clearly, there are many choices for W. We would like W to have small norm(s). The following short algorithm constructs a suitable W:

Input: V, U, n, q
Output: W such that $VW = U$

```
w_ij := 0 for all i, j
p := 0
for j := 1 to n + q
    if U_j = V_k for some k, then w_kj := 1
    else        // U_j must contain a single nonzero
        p := p + 1
        k := the index for which u_kj = 1
        if p < q, then
            w_1j := 1
            for i := 2 to k, w_ij := −1, end
        else
            for i := k to n, w_ij := 1, end
        endif
    endif
end
```

Since U and V have many columns in common, W by construction has mostly columns with only one nonzero element, which is 1. A few columns of W will have all nonzero entries either above or below a certain index k. One can verify that $VW = U$ using only elementary algebra. For our specific example, we get

$$W = \begin{pmatrix} 1 & & & 1 & 1 & & & & \\ & 1 & & -1 & -1 & & & & \\ & & 1 & -1 & -1 & & & & \\ & & & & -1 & & & & \\ & & & & & 1 & 1 & & \\ & & & & & & 1 & 1 & 1 & \\ & & & & & & 1 & 1 & & 1 \end{pmatrix}.$$

By inspection, the largest (absolute) row sum in W for the example above is 3 and the largest (absolute) column sum is 4, so $\|W\|_2^2 \le \|W\|_1 \|W\|_\infty \le 12$. Columns in W with more than one nonzero correspond to boundaries between subdomains. The corresponding columns in U have to be "supported" from the external boundary, $\partial\Omega$. In general, W has at most $q+1$ nonzeros in each row (one nonzero for each boundary of $q/2$ subdomains plus one additional "diagonal" nonzero) and at most $(n+1)/2$ nonzeros in each column. Since each nonzero in W is ± 1, it follows that $\sigma(B, A) \le \|W\|_1 \|W\|_\infty \le (q+1)(n+1)/2$.

LEMMA 10.2. *Let A be the discrete Laplace operator as defined above, and let B be a block diagonal approximation for A with $q > 1$ blocks formed by dropping some of the off-diagonal entries. Then $\sigma(B, A) \le (q+1)(n+1)/2$.*

Another way to obtain this result is to use the congestion-dilation proposition (Proposition 9.3) for graphs. In our case, we need support paths from the boundary nodes to each interior node that is on the boundary of a subdomain. Consequently, the dilation is $O(n)$ while the congestion is $O(q)$, which also gives the support bound $O(nq)$. (The factor $1/2$ comes from routing half the support paths from each boundary.)

By combining the two bounds on the support numbers, we get the following bound on the condition number.

THEOREM 10.3. *Let A be the discrete Laplace operator as defined above, and let B be a block diagonal approximation for A with $q > 1$ blocks formed by dropping some of the off-diagonal entries. Then the condition number κ satisfies $\kappa(B^{-1}A) \le (q+1)(n+1)$.*

A more detailed analysis in [8] showed that the condition number is bounded by $qn+q+1$. Our bound agrees with that bound up to a lower order term and is simpler to derive. Since the Chang–Schultz bound is known to be tight [8], our bound is also tight asymptotically.

For the special case where uniform blocks are used, let $H = hn/q$ such that H denotes the subdomain size. This gives us the well-known result from domain decomposition that the condition number is bounded by $O(1/(hH))$.

10.2. Higher dimensions. We will show that the following result holds for block Jacobi preconditioning in dimensions higher than one.

THEOREM 10.4. *Consider a regular $n_1 \times n_2 \times \cdots \times n_d$ grid in d dimensions. Let A be the finite difference discretization of the Laplace equation. Suppose the domain is partitioned into subdomains, possibly in an unstructured fashion. Let B be the block Jacobi preconditioner corresponding to this partitioning (domain decomposition).*

Then

$$\kappa(B^{-1}A) = O(\max_{1\le i\le d} n_i q_i),$$

where q_i is the maximum number of subdomains along any line in the ith dimension.

Proof. Split $A = A_1 + A_2 + \cdots + A_d$ and, similarly, $B = B_1 + \cdots + B_d$, where A_i corresponds to the Laplace finite difference operator along the lines in the ith dimension. Similarly, let B_i correspond to the block Jacobi approximation in the ith dimension. By the splitting proposition (Proposition 2.5), we have that

$$\sigma(B, A) \le \max_i \{\sigma(B_i, A_i)\}.$$

Consider the algebraic equations along one line of gridpoints. Such a subset of equations corresponds precisely to the one-dimensional problem we analyzed in the previous section. Hence, $\sigma(B_i, A_i) = O(n_i q_i)$, and it follows that

$$\sigma(B, A) = O(\max_i n_i q_i).$$

The desired condition number bound follows by noting that $\sigma(A, B) \le 2$ as in the one-dimensional case. ∎

For a regular grid on the unit cube with $n^{1/d}$ gridpoints in each dimension and a uniform partitioning ($H = 1/q$) we obtain the expected bound $\sigma(B, A) = O(1/(hH))$.

10.3. Block Jacobi summary. We have rederived known bounds for block Jacobi using support theory. While a traditional analysis is based on calculating the eigenvectors (eigenfunctions) of the Laplacian, the support theory analysis is purely algebraic and does not require analytic expressions for the eigenvectors. Our analysis is a bit similar to the one in [8] but simpler in several ways. One advantage of our analysis is that it is easy to analyze nonuniform (irregular) decompositions of a domain. In this example, we examined only the Laplace equation on a structured grid. Our analysis tools also apply to more complicated equations and unstructured grids, though it is harder to obtain any general (a priori) bound.

11. Extensions to general matrices. Support theory was developed with spd systems in mind. Nevertheless, much of the theory developed in the preceding sections can be extended to general (including indefinite and nonsymmetric) matrices through a small change in the definition of support number.

DEFINITION 11.1. *For matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$ with the same number of columns, the* generalized *support number of (A, B) is defined by*

$$\hat{\sigma}(A, B) = \min\left\{t \,|\, x^T(\tau^2 B^T B - A^T A)x \ge 0 \text{ for all } x \in \mathbb{R}^n \text{ and for all } \tau \ge t\right\}.$$

Note that generalized support numbers cannot be negative.

Since both $B^T B$ and $A^T A$ are spsd, all of the techniques introduced in the previous sections can be used to analyze $\hat{\sigma}(A, B)$. When $B^T B$ has full rank, then by Theorem 2.3 $\hat{\sigma}(A, B) = \sqrt{\lambda_{\max}(A^T A, B^T B)} = \mu_{\max}(A, B)$, where $\mu_{\max}(A, B)$ is the largest generalized singular value of the matrix pencil (A, B). For a brief description of generalized singular values, see section 8.7.3 of [10]. (We use μ to denote singular values since the symbol σ has been reserved for support numbers in this paper.)

The spectral condition number $\kappa_2(C)$ is defined as $\kappa_2(C) = \|C\|_2\|C^{-1}\|_2 = \mu_{\max}(C)/\mu_{\min}(C)$. For nonsingular B, it follows that $\kappa_2(B^{-1}A) \le \hat{\sigma}(A, B)\hat{\sigma}(B, A)$.

When A and B are singular but share the same nullspace, then $\hat{\sigma}(A,B)\hat{\sigma}(B,A)$ bounds the effective condition number of the pencil (A,B) outside the nullspace. In short, the generalized support number can be used to bound the condition number in much the same way as the standard support number.

The quadratic form in the definition of generalized support can be factored in a useful manner. Specifically, for $A, B \in \mathbb{R}^{m\times n}$,

$$x^T(\tau^2 B^T B - A^T A)x = x^T(\tau B^T - A^T)(\tau B + A)x.$$

If A and B have different sizes, one can pad the smaller matrix with zeros. When A and B are both symmetric, this factorization reveals a close relationship between the generalized support number and the (standard) support number. Since the product of two psd matrices that commute is also psd, the quadratic form on the right will be nonnegative when both the matrix terms are psd. These terms have the form used in standard support numbers, which leads to the following.

PROPOSITION 11.2. *If A and B are symmetric, then $\hat{\sigma}(A,B) \le \max\{\sigma(A,B), \sigma(-A,B)\}$. Equality holds when B is spsd.*

If B is not psd, then $\sigma(A,B)$ is infinite and the bound becomes useless. In the case where both A and B are spsd a further reduction is possible. In this case, $\sigma(-A,B)$ is nonpositive, so Proposition 11.2 reduces to the following.

COROLLARY 11.3. *When A and B are both spsd, then $\hat{\sigma}(A,B) = \sigma(A,B)$.*

Thus, generalized support numbers are strict generalizations of the support numbers we defined in section 2. Note, however, that there is a discrepancy in definitions if either A or B is not psd. For example, if A is symmetric but negative definite, then the standard support number σ will be negative and corresponds to the largest (rightmost) generalized eigenvalue of (A,B). In contrast, the generalized support number $\hat{\sigma}$ is always nonnegative and corresponds to the largest magnitude of a generalized eigenvalue of (A,B).

Some of the propositions presented in this paper hold for generalized support numbers as well as the standard support number, but not all. In particular, the splitting proposition (Proposition 2.5) needs to be modified, as shown below.

PROPOSITION 11.4. *For splittings $A = A_1 + A_2$ and $B = B_1 + B_2$, where $B_1^T B_2$ is psd (possibly zero),*

$$\hat{\sigma}(A_1 + A_2, B_1 + B_2) \le \max\left\{\hat{\sigma}(A_1,B_1), \hat{\sigma}(A_2,B_2), \sqrt{\max\{0, \sigma(A_1^T A_2, B_1^T B_2)\}}\right\}.$$

Proof. We have that $A^T A = (A_1+A_2)^T(A_1+A_2) = A_1^T A_1 + A_1^T A_2 + A_2^T A_1 + A_2^T A_2$, and similarly for $B^T B$. Hence

$$\begin{aligned} &x^T(\tau^2 B^T B - A^T A)x \\ &\quad \le x^T\left((\tau^2 B_1^T B_1 - A_1^T A_1) + (\tau^2 B_2^T B_2 - A_2^T A_2) + 2(\tau^2 B_1^T B_2 - A_1^T A_2)\right)x. \end{aligned}$$

Now choose τ by the right-hand side bound in the proposition. Since each of the three terms in the quadratic form above is then nonnegative, the total quadratic form must also be nonnegative. The desired result follows from Definitions 2.1 and 11.1. ∎

In the special case when $A_1^T A_2$ and $B_1^T B_2$ are both zero, the proposition reduces to the standard splitting property.

Finally, it is possible that the standard support number may provide an indication about convergence even for non-spd systems. An analysis by Starke [27] shows that the residual of the GMRES method can be bounded by a simple function of the

support number (although he did not use that terminology). We have not tried to determine which approach gives better bounds.

12. Summary and future work. All the results in this paper that hold for real symmetric matrices generalize to complex Hermitian matrices. This feature complements the work of Howle and Vavasis [16], who considered complex symmetric matrices. It is more difficult to go from symmetric to nonsymmetric systems. A major difficulty is that the correspondence between the support number and the largest generalized eigenvalue (Theorem 2.3) breaks down. In section 11 we proposed to use the *generalized* support number, which is closely related to the generalized singular values, to bound the condition number in the non-spd case. The convergence analysis for iterative methods for nonsymmetric problems is quite complicated and further work is needed.

In the symmetric case, the Chebyshev (semi)iterative method [29, 31] can benefit from support analysis because good bounds on the extreme eigenvalues are required. We remark that Chebyshev iteration has the same worst-case complexity as conjugate gradients but requires no inner products. This may give Chebyshev iteration an advantage for large-scale problems on parallel computers. Also note that in general the convergence of iterative methods depends not only on the extreme eigenvalues but also on the distribution of all the eigenvalues. The support theory presented here bounds only the extreme eigenvalues. It is more difficult to obtain bounds for interior eigenvalues. See [1] for some such results.

The present paper extends the existing support-graph theory [6] from spsd, diagonally dominant M-matrices to a much wider class of matrices, namely, all spsd matrices. Our framework is purely algebraic and no longer relies on graph theory (though graphs may still be useful in an analysis). The work presented here has enabled us to generalize Vaidya's preconditioners to all spd diagonally dominant matrices [7]. Using vectors with two nonzeros but possibly different magnitudes, we conjecture that the max-weight-basis preconditioners can be extended to all H-matrices.

The authors believe that the tools presented in the present paper are well suited both to analyze existing preconditioners and to develop new types of preconditioners. Promising candidates for analysis include incomplete factorizations and algebraic multilevel methods. The earlier support-graph theory has already been successfully applied to a multilevel preconditioner by Gremban [11], and to incomplete factorization preconditioners by Guattery [13] and Bern et al. [6]. However, the results are restricted to fairly specific problem instances and matrix classes. We hope that the techniques presented in the present paper can be used to extend some of these methods and results to all spd matrices.

The support preconditioners we and others have developed all rely on using the rank-1 support theorem (Theorem 4.4) or the symmetric product support theorem (Theorem 4.5) where columns of U and V correspond to edges in a graph (that is, they have only two nonzeros and these have the same magnitude). An open question is whether efficient preconditioners can be constructed that employ column vectors with three or more nonzeros. Although the theory in the present paper can handle this situation, a major obstacle in practice is that the resulting preconditioner may be difficult to solve for (i.e., factorize).

Acknowledgments. We would like to thank John Gilbert for introducing us to support-graph preconditioners, Roger Horn for pointing us to the results in section 7, and David Keyes for suggesting the example in section 10. The remarks of an anonymous referee greatly improved the paper. We also thank Michele Benzi, Gene Golub, Steve Guattery, and Sivan Toledo for helpful comments.

REFERENCES

[1] O. Axelsson, *Bounds of eigenvalues of preconditioned matrices*, SIAM J. Matrix Anal., 13 (1992), pp. 847–862.

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.

[3] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems. Theory and Computation*, Academic Press, Orlando, FL, 1984.

[4] R. Beauwens, *Upper eigenvalue bounds for pencils of matrices*, Linear Algebra Appl., 62 (1984), pp. 87–104.

[5] R. Beauwens, *Lower eigenvalue bounds for pencils of matrices*, Linear Algebra Appl., 85 (1987), pp. 101–119.

[6] M. Bern, J. R. Gilbert, B. Hendrickson, N. Nguyen, and S. Toledo, *Support-Graph Preconditioners*, tech. report, School of Computer Science, Tel-Aviv University, 2001; SIAM J. Matrix Anal. Appl., submitted.

[7] E. G. Boman, D. Chen, B. Hendrickson, and S. Toledo, *Maximum-weight-basis preconditioners*, Numer. Linear Algebra Appl., to appear.

[8] M. Y. Chang and M. H. Schultz, *Bounds on block diagonal preconditioning*, Parallel Algorithms Appl., 1 (1993), pp. 141–164.

[9] M. Fiedler and T. L. Markham, *An observation on the Hadamard product of Hermitian matrices*, Linear Algebra Appl., 215 (1995), pp. 179–182.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[11] K. Gremban, *Combinatorial Preconditioners for Sparse, Symmetric, Diagonally Dominant Linear Systems*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1996; available as Tech. Report CMU-CS-96-123.

[12] K. Gremban, G.L. Miller, and M. Zagha, *Performance evaluation of a new parallel preconditioner*, in Proceedings of the 9th International Parallel Processing Symposium, IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 65–69; a longer version is available as Tech. Report CMU-CS-94-205, Carnegie Mellon University, Pittsburgh, PA, 1994.

[13] S. Guattery, *Graph Embedding Techniques for Bounding Condition Numbers of Incomplete Factor Preconditioners*, Tech. Report 97-47, ICASE, NASA Langley Research Center, Hampton, VA, 1997.

[14] S. Guattery, F. T. Leighton, and G. Miller, *The path resistance method for bounding λ_2 of a Laplacian*, in Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 1997, pp. 201–210.

[15] R. A. Horn, *Personal communication*, 2001.

[16] V. Howle and S. Vavasis, *An Iterative Method for Solving Complex-Symmetric Systems Arising in Electric Power Modeling*, tech. report, Cornell University, Ithaca, NY, 2000; SIAM J. Matrix Anal. Appl., submitted.

[17] B. M. Irons and G. Treharne, *A bound theorem in eigenvalues and its practical applications*, in Proceedings of the Third Conference on Matrix Methods in Structural Mechanics, Wright–Patterson AFB, Dayton, Ohio, 1971.

[18] H.-C. Lee and A. J. Wathen, *On element-by-element preconditioning for general elliptic problems*, Comput. Methods Appl. Mech. Engrg., 92 (1991), pp. 215–229.

[19] M. M. Magolu, *Analytical bounds for block approximate factorizations*, Linear Algebra Appl., 179 (1993), pp. 33–57.

[20] M. M. Magolu and Y. Notay, *On the conditioning analysis of block approximate factorization methods*, Linear Algebra Appl., 154/156 (1991), pp. 583–599.

[21] Y. Notay, *Conditioning analysis of modified block incomplete factorizations*, Linear Algebra Appl., 154/156 (1991), pp. 711–722.

[22] Y. Notay, *Upper eigenvalue bounds and related modified incomplete factorizations*, in Iterative Methods in Linear Algebra, North–Holland, Amsterdam, 1992, pp. 551–562.

[23] R. Reams, *Hadamard inverses, square roots and products of almost semidefinite matrices*, Linear Algebra Appl., 288 (1999), pp. 35–43.

[24] J. H. Reif, *Efficient approximate solution of sparse linear systems*, Comput. Math. Appl., 36 (1998) pp. 37–58; see also errata in Comput. Math. Appl., 38 (1999), p. 141.

[25] I. Schur, *Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen*, J. Reine Angew. Math., 140 (1911), pp. 1–28.

[26] B. Smith, P. E. Bjørstad, and W. Gropp, *Domain Decomposition. Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[27] G. Starke, *Field-of-values analysis of preconditioned iterative methods for nonsymmetric el-*

liptic problems, Numer. Math., 78 (1997), pp. 103–117.
[28] P. M. Vaidya, *Solving Linear Equations with Symmetric Diagonally Dominant Matrices by Constructing Good Preconditioners*, manuscript.
[29] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[30] A. J. Wathen, *An analysis of some element-by-element preconditioning techniques*, Comput. Methods Appl. Mech. Engrg., 74 (1989), pp. 271–287.
[31] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 718–738

OPTIMAL PRECONDITIONING FOR RAVIART–THOMAS MIXED FORMULATION OF SECOND-ORDER ELLIPTIC PROBLEMS*

CATHERINE ELIZABETH POWELL† AND DAVID SILVESTER†

Abstract. We evaluate two preconditioning strategies for the indefinite linear system obtained from Raviart–Thomas mixed finite element formulation of a second-order elliptic problem with variable diffusion coefficients. It is known that the underlying saddle-point problem is well-posed in two function spaces, $H(\mathrm{div}) \times L^2$ and $L^2 \times H^1$, leading to the possibility of two distinct types of preconditioner. For homogeneous Dirichlet boundary conditions, the discrete problems are identical. This motivates our use of Raviart–Thomas approximation in both frameworks, yielding a nonconforming method in the second case. The focus is on linear algebra; we establish the optimality of two parameter-free block-diagonal preconditioners using basic properties of the finite element matrices. Uniform eigenvalue bounds are established and the impact of the PDE coefficients is explored in numerical experiments. A practical scheme is discussed, the key building block for which is a fast solver for a scalar diffusion operator based on algebraic multigrid. Trials of preconditioned MINRES illustrate that both preconditioning schemes are optimal with respect to the discretization parameter and robust with respect to the PDE coefficients.

Key words. second-order elliptic problems, saddle-point problems, variable coefficients, mixed finite elements, Raviart–Thomas, MINRES, preconditioning

AMS subject classifications. 65F10, 65N22, 65M60, 65N12

DOI. 10.1137/S0895479802404428

1. Introduction.

Let Ω be a bounded domain in $\mathbb{R}^2$. We consider scalar second-order elliptic problems of the form

$$\begin{aligned} -\nabla \cdot \mathcal{A}\nabla p &= f \quad \text{in } \Omega, \\ p &= g \quad \text{on } \partial\Omega_D, \\ \mathcal{A}\nabla p \cdot \vec{n} &= 0 \quad \text{on } \partial\Omega_N, \end{aligned} \tag{1.1}$$

where $\partial\Omega_D \neq \emptyset$ and $\mathcal{A} = \mathcal{A}(\vec{x})$ is a 2×2 bounded, symmetric, and uniformly positive-definite matrix-valued function. This implies that there exist positive constants γ and Γ with $0 < \gamma \leq \Gamma$ such that

$$\gamma(\vec{v}, \vec{v}) \leq (\mathcal{A}^{-1}\vec{v}, \vec{v}) \leq \Gamma(\vec{v}, \vec{v}) \tag{1.2}$$

for every $\vec{v} : \Omega \to \mathbb{R}^2$. Boundary-value problems of this type occur in mathematical models of important physical processes such as fluid flow in porous media. To fix ideas, we call p and $\vec{u} = \mathcal{A}\nabla p$ the pressure and velocity solutions, respectively. Mixed finite element methods are favored when $\vec{u}$ is the variable of interest since postprocessing primal pressure solutions leads to loss of accuracy. Mixed velocity solutions are insensitive to the variation in the coefficient term (see [15, pp. 240–241]). In addition, mixed methods conserve mass locally, a crucial feature in the modelling of groundwater flow. Other advantages of a mixed approximation are discussed in [6].

*Received by the editors March 21, 2002; accepted for publication (in revised form) by S. A. Vavasis June 19, 2003; published electronically January 30, 2004.
http://www.siam.org/journals/simax/25-3/40442.html
†Mathematics Department, UMIST, Manchester, M60 1QD, United Kingdom (cp@fire.ma.umist.ac.uk, djs@fire.ma.umist.ac.uk).

1.1. Preconditioning strategies. It is known (see [1], [5], [6], [16], [20], [24]) that mixed finite element formulation of (1.1) yields an indefinite linear system of the form

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{C} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ -\underline{f} \end{pmatrix}, \tag{1.3}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive-definite and $\underline{u}$ and $\underline{p}$ are the discrete velocity and pressure solutions, respectively. The coefficient matrix C is ill-conditioned with respect to the discretization parameter (see, for example, [20]) and the coefficient term $\mathcal{A}$. Krylov solution methods are therefore not robust; preconditioners are essential.

Two approaches are possible. The majority of preconditioners that have already been suggested require the transformation of the underlying saddle-point problem to a *positive-definite* one. Indeed, there is a rich literature concerning domain-decomposition and multigrid techniques for such reduced problems (see, for example, [7], [8], [9], [10], [23], [27]). This is not our approach. We follow [1], [20], and [28] and solve the full *indefinite* system (1.3) using preconditioned MINRES (see [14]). In [20], [21], and [22], Rusten et al. propose preconditioners of the form

$$\mathcal{P} = \begin{pmatrix} \mathcal{I} & 0 \\ 0 & \mathcal{S} \end{pmatrix}, \tag{1.4}$$

where $\mathcal{S}$ is an approximation to a Schur complement operator on the pressure space and $\mathcal{I}$ is the identity operator. Vassilevski and Lazarov [28] construct a preconditioner for a transformed indefinite problem. Convergence rates are shown to be independent of the mesh parameter, but, critically, iteration counts are affected by two artificial parameters, and it is not clear what the optimal choices are. Arnold, Falk, and Winther observe in [1] that C also has the same mapping properties as the matrix operator,

$$\mathcal{P} = \begin{pmatrix} \mathcal{H} & 0 \\ 0 & \mathcal{I} \end{pmatrix}, \tag{1.5}$$

where $\mathcal{H} : H(\mathrm{div}) \times H(\mathrm{div}) \to \mathbb{R}$ is the $H(\mathrm{div})$ operator defined, for vector functions $\vec{u}$ and $\vec{v}$, by $(\mathcal{H}\vec{u}, \vec{v}) = (\vec{u}, \vec{v}) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v})$. $\mathcal{H}$ is approximated using domain decomposition and multigrid and it is shown (see [1], [2]) that the resulting preconditioner is optimal for quasi-uniform meshes and "trivial" coefficients of the form $\mathcal{A} = \delta\mathcal{I}$, where δ is a constant.

1.2. Overview. An *optimal* preconditioner is a matrix operator that accelerates the convergence rate of MINRES so that convergence to a fixed tolerance is independent of the discretization parameter. The aim here is to illustrate the robustness of two optimal parameter-free preconditioners for the indefinite problem when the ratio $\frac{\Gamma}{\gamma}$ of constants in (1.2) is large. In this section, we review Raviart–Thomas approximation and convergence properties of MINRES. In section 2 we look for a solution to (1.1) in the space $H(\mathrm{div}) \times L^2$, derive the discrete problem, and construct an optimal preconditioner associated with that space. A new uniform eigenvalue bound is established and shown to be tight. Numerical experiments assess the impact of a range of different coefficients on that bound. In section 3, the solution is sought in the space $L^2 \times H^1$. A second optimal preconditioning strategy is described. We propose a

novel practical implementation, the key building block for which is a fast solver for a scalar diffusion operator based on black-box algebraic multigrid (AMG). Performance of both preconditioners is assessed numerically in MINRES trials.

1.3. Notation. Let $\Omega \subseteq \mathbb{R}^2$ be a convex polygon with boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$. $L^2(\Omega)$ is the usual space of square-integrable scalar functions with inner-product $(\cdot,\cdot)$. For the space of vector functions $L^2(\Omega)^2$, the definition is understood to hold componentwise. We write $\|\cdot\|_0$ to denote the Lebesgue measure for both spaces. The subspace,

$$H(\mathrm{div};\Omega) = \left\{ \vec{v} \in L^2(\Omega)^2 \mid \nabla\cdot\vec{v} \in L^2(\Omega) \right\},$$

contains vectors with square-integrable divergence. The associated inner-product is $(\vec{u},\vec{v})_{div} = (\vec{u},\vec{v}) + (\nabla\cdot\vec{u}, \nabla\cdot\vec{v})$, and we denote the induced norm $\|\cdot\|_{div}$. The Sobolev space,

$$H^1(\Omega) = \left\{ w \in L^2(\Omega) \mid \nabla w \in L^2(\Omega)^2 \right\},$$

is equipped with the usual inner-product $(w,v)_1 = (w,v) + (\nabla w, \nabla v)$ and the induced norm $\|\cdot\|_1$.

1.4. Raviart–Thomas finite elements. Let T_h denote a partition of Ω into a mesh of triangles or rectangles $\{K_1,\ldots,K_s\}$ with maximum edge size h. The mesh is refined such that $\{T_{h_1}, T_{h_2}, \ldots\}$ is a family of shape-regular quasi-uniform partitions of Ω. Our numerical experiments are performed on rectangular grids. For given h and T_h, the associated Raviart–Thomas spaces $V_h \subset H(\mathrm{div};\Omega)$ and $W_h \subset L^2(\Omega)$ of integer index k (see [16]) are

$$V_h = \left\{ \vec{v_h} \in H(\mathrm{div};\Omega) \mid \vec{v_h}\,|_K \in Q_{k+1,k}(K) \times Q_{k,k+1}(K) \quad \forall\, K \in T_h \right\} \tag{1.6}$$

and $W_h = \{ w_h \in L^2(\Omega) \mid w_h\,|_K \in P_k(K)\ \forall\, K \in T_h \}$. Here, $P_k(K)$ denotes the space of polynomials of degree $\le k$ on rectangle K and $Q_{r,s}(K)$ is the space of polynomials of degree $\le r$ in x and $\le s$ in y. We use the lowest order[1] ($k = 0$) spaces. Thus, velocity and pressure test functions have the special piecewise forms,

$$\vec{v_h}\,|_K = \begin{pmatrix} v_1 + v_2 x \\ v_3 + v_4 y \end{pmatrix}, \quad w_h\,|_K = w_1, \tag{1.7}$$

respectively, where v_1, v_2, v_3, v_4, and w_1 are constants. For an $H(\mathrm{div};\Omega)$ conforming velocity approximation, the degrees of freedom are normal components at edge midsides; the pressure solution is sampled at element centroids (see Figure 1.1).

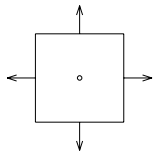


FIG. 1.1. *Degrees of freedom for the lowest order, rectangular, Raviart–Thomas element.*

[1] For highly discontinuous and anisotropic diffusion coefficients, solution regularity is low.

1.5. MINRES. Let us denote the symmetric and indefinite system (1.3) by $C\underline{x} = \underline{b}$ and suppose that a preconditioner with symmetric and positive-definite blocks,

$$P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}, \tag{1.8}$$

has been chosen. Given a zero initial guess $\underline{x}_0$, the iterative solver MINRES (see [14]) applied to the symmetrically preconditioned system

$$\begin{pmatrix} P_1^{-\frac{1}{2}} A P_1^{-\frac{1}{2}} & P_1^{-\frac{1}{2}} B^T P_2^{-\frac{1}{2}} \\ P_2^{-\frac{1}{2}} B P_1^{-\frac{1}{2}} & 0 \end{pmatrix} \begin{pmatrix} \underline{y} \\ \underline{z} \end{pmatrix} = \begin{pmatrix} P_1^{-\frac{1}{2}} \underline{g} \\ -P_2^{-\frac{1}{2}} \underline{f} \end{pmatrix}, \quad \begin{pmatrix} \underline{y} \\ \underline{z} \end{pmatrix} = \begin{pmatrix} P_1^{\frac{1}{2}} \underline{u} \\ P_2^{\frac{1}{2}} \underline{p} \end{pmatrix}, \tag{1.9}$$

generates a sequence of iterates $\underline{x}_k$, belonging to the Krylov space

$$K_k = \underline{x}_0 + \mathtt{span}\left(P^{-1}\underline{r}_0, P^{-1}CP^{-1}\underline{r}_0, \ldots, \left(P^{-1}C\right)^{k-1} P^{-1}\underline{r}_0 \right),$$

with the minimization property

$$\| \underline{b} - C\underline{x}_k \|_{P^{-1}} = \min_{\underline{x} \in K_k} \| \underline{b} - C\underline{x} \|_{P^{-1}},$$

where $\underline{r}_k = \underline{b} - C\underline{x}_k$ is the kth residual and $\| \underline{v} \|_{P^{-1}}^2 = \underline{v}^T P^{-1} \underline{v}$. Since P is symmetric, it can be shown that a sharp upper bound for the residual error after k iterations (see [12, p. 51]) is given by

$$\frac{\| \underline{r}^{(k)} \|_{P^{-1}}}{\| \underline{r}^{(0)} \|_{P^{-1}}} \leq \min_{p_k} \max_{i=1:(n+m)} | p_k(\lambda_i) |, \tag{1.10}$$

where p_k is a polynomial of degree k satisfying $p_k(0) = 1$, and $\{\lambda_i\}_{i=1}^{n+m}$ denotes the eigenvalues of $P^{-1}C$.

Theoretical eigenvalue bounds for the system matrix (1.3) are derived by Rusten and Winther in [20] and Silvester and Wathen in [24]. For the lowest order Raviart–Thomas discretization of (1.1) on quasi-uniform meshes, it can be shown (see section 2.3 in [23]) that there exist constants C_1 and C_2, depending on the coefficient $\mathcal{A}$ but independent of h, such that

$$C_1 h^2 \leq \frac{\underline{u}^t A \underline{u}}{\underline{u}^t \underline{u}} \leq C_2 h^2 \quad \forall\, \underline{u} \in \mathbb{R}^n \backslash \{\underline{0}\}. \tag{1.11}$$

Further, it can be shown that the minimum and maximum singular values of B are bounded by constants that depend on h^2 and h, respectively. The theory of [20] consequently yields an h-dependent eigenvalue bound of the form $[-ah, -bh^2] \cup [ch^2, dh]$ for positive constants a, b, c, d. The solution process becomes prohibitively expensive as the mesh is refined, and this is a serious flaw in practical applications. When $\mathcal{A}$ is ill-conditioned, convergence is observed to deteriorate further. In sections 2 and 3 we discuss two optimal block-diagonal preconditioners of the form (1.8), choosing matrices P_1 and P_2 to be discrete representations of suitable norms on V_h and W_h so that inclusion intervals for the eigenvalues of $P^{-1}C$ are independent of h.

2. $H_{div} \times L^2$ formulation. Substituting $\vec{u} = \mathcal{A}\nabla p$ in (1.1) yields the first-order system

$$\begin{aligned} \mathcal{A}^{-1}\vec{u} - \nabla p &= 0, \\ \nabla \cdot \vec{u} &= -f \quad \text{in } \Omega, \\ p &= g \quad \text{on } \partial\Omega_D, \\ \vec{u} \cdot \vec{n} &= 0 \quad \text{on } \partial\Omega_N. \end{aligned} \tag{2.1}$$

The data requirements are $f \in L^2(\Omega)$ and $g \in H^{\frac{1}{2}}(\partial\Omega_D)$, the set of traces of $H^1(\Omega)$ functions on $\partial\Omega_D$. It is well known (see [1], [3]) that (2.1) may be formulated as a saddle-point problem in two distinct function spaces. First, we choose

$$V = H_{0,N}(\mathrm{div};\Omega) = \left\{ \vec{v} \in H(\mathrm{div};\Omega) \mid \vec{v} \cdot \vec{n} = 0 \text{ on } \partial\Omega_N \right\}$$

and $W = L^2(\Omega)$. Multiplying by arbitrary $\vec{v} \in V$ and $w \in W$ in (2.1) and integrating the *first* equation by parts, we look for a solution $(\vec{u}, p) \in V \times W$ satisfying

$$\begin{aligned} (\mathcal{A}^{-1}\vec{u}, \vec{v}) + (p, \nabla \cdot \vec{v}) &= \langle g, \vec{v} \cdot \vec{n} \rangle \qquad \forall\, \vec{v} \in V, \\ (w, \nabla \cdot \vec{u}) &= -(f, w) \qquad \forall\, w \in W, \end{aligned} \tag{2.2}$$

where $\langle g, \vec{v} \cdot \vec{n} \rangle = \int_{\partial\Omega_D} g\, \vec{v} \cdot \vec{n}\, ds$. Condition (1.2) ensures that $(\mathcal{A}^{-1}\vec{u}, \vec{v})$ is bounded.

The stability theory of Brezzi and Babuška (see [6], [17]) shows that a unique solution $(\vec{u}, p)$ exists if and only if there are constants $\alpha_0 > 0$ and $\beta_0 > 0$ satisfying

$$(\mathcal{A}^{-1}\vec{v}, \vec{v}) \geq \alpha_0 \,\|\, \vec{v} \,\|_{div}^2 \qquad \forall\, \vec{v} \in Z, \tag{2.3}$$

$$\sup_{\vec{v} \in V \setminus \{\vec{0}\}} \frac{(w, \nabla \cdot \vec{v})}{\|\, \vec{v} \,\|_{div}} \geq \beta_0 \,\|\, w \,\|_0 \qquad \forall\, w \in W, \tag{2.4}$$

where $Z = \{\vec{v} \in V \mid (w, \nabla \cdot \vec{v}) = 0 \;\forall\, w \in W\}$. Since $\nabla \cdot V \subset W$, the constraint space Z contains only divergence-free velocities and $\|\cdot\|_{div}$ is equivalent to $\|\cdot\|_0$ on this space. Thus, $\alpha_0 = \gamma$ in (1.2). The constant β_0 depends only on the shape of the domain Ω.

To implement the conforming Raviart–Thomas discretization, we choose finite-dimensional subspaces $V_h \subset V$ and $W_h \subset W$ as defined in section 1.4. We then look for $(\vec{u_h}, p_h) \in V_h \times W_h$ satisfying

$$\begin{aligned} (\mathcal{A}^{-1}\vec{u_h}, \vec{v_h}) + (p_h, \nabla \cdot \vec{v_h}) &= \langle g, \vec{v_h} \cdot \vec{n} \rangle \qquad \forall\, \vec{v_h} \in V_h, \\ (w_h, \nabla \cdot \vec{u_h}) &= -(f, w_h) \qquad \forall\, w_h \in W_h. \end{aligned} \tag{2.5}$$

Stability of the approximation is established in [16]. A crucial property of this scheme is that $\nabla \cdot V_h \subset W_h$ and hence the discrete version of condition (2.3) is trivially satisfied if (1.2) holds.

2.1. Finite element matrices. In the usual way, we choose bases $V_h = \mathtt{span}\{\vec{\varphi}_i\}_{i=1}^n$ and[2] $W_h = \mathtt{span}\{\phi_j\}_{j=1}^m$. Defining the finite element matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{m \times n}$ by $A_{i,j} = (\mathcal{A}^{-1}\vec{\varphi}_i, \vec{\varphi}_j)$ for $1 \leq i, j \leq n$ and $B_{k,j} = (\phi_k, \nabla \cdot \vec{\varphi}_j)$ for $1 \leq k \leq m$, and the vectors $\underline{g} \in \mathbb{R}^n$ and $\underline{f} \in \mathbb{R}^m$ by $g_i = \langle g, \vec{\varphi}_i \cdot \vec{n} \rangle$ and $f_k = -(f, \phi_k)$, leads to the linear algebra problem (1.3) with $\vec{u_h} = \sum_{i=1}^n u_i \vec{\varphi}_i$ and $p_h = \sum_{j=1}^m p_j \phi_j$.

[2] n = number of edges − number of edges on $\partial\Omega_N$; m = number of elements.

To construct our first preconditioner, we also define the velocity mass matrix $M \in \mathbb{R}^{n\times n}$, the velocity divergence matrix $D \in \mathbb{R}^{n\times n}$, and the pressure mass matrix $N \in \mathbb{R}^{m\times m}$ by

$$\begin{aligned} M_{i,j} &= (\vec{\varphi}_i, \vec{\varphi}_j), && 1 \le i,j \le n,\\ D_{i,j} &= (\nabla\cdot\vec{\varphi}_i, \nabla\cdot\vec{\varphi}_j), && 1 \le i,j \le n,\\ N_{r,s} &= (\phi_r, \phi_s), && 1 \le r,s \le m. \end{aligned}$$

These give useful discrete representations of the norms $\|\cdot\|_0$ and $\|\cdot\|_{div}$ on V_h and $\|\cdot\|_0$ on W_h since for any $\vec{v_h}$ in V_h and w_h in W_h we can now write

$$\|\vec{v_h}\|_0^2 = \underline{v}^T M \underline{v}, \quad \|\vec{v_h}\|_{div}^2 = \underline{v}^T (M+D)\underline{v}, \quad \|w_h\|_0^2 = \underline{w}^T N \underline{w}, \tag{2.6}$$

where $\underline{v}$ and $\underline{w}$ are the vectors of coefficients corresponding to the expansions of the functions $\vec{v_h}$ and w_h in the chosen basis sets. Now, in matrix form, Brezzi's discrete inf-sup stability condition (2.4) is

$$\beta^2 \le \frac{\underline{w}^T B\,(M+D)^{-1}\,B^T \underline{w}}{\underline{w}^T N \underline{w}} \qquad \forall\, \underline{w} \in \mathbb{R}^m \backslash \{\underline{0}\}, \tag{2.7}$$

and so a computable upper bound for the discrete inf-sup constant β is $\sqrt{\lambda_{min}}$, where λ_{min} is the minimum eigenvalue of

$$B\,(M+D)^{-1}\,B^T \underline{w} = \lambda N \underline{w}. \tag{2.8}$$

Note that β does not depend on the coefficient $\mathcal{A}$ and so the method is inf-sup stable even when $\frac{\Gamma}{\gamma}$ is large. Using the Cauchy–Schwarz inequality we have

$$\sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{|\,(w_h, \nabla\cdot\vec{v}_h)\,|^2}{\|\,\vec{v}_h\,\|_{div}^2\,\|\,w_h\,\|_0^2} \le \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{\|\,w_h\,\|_0^2\,\|\,\nabla\cdot\vec{v}_h\,\|_0^2}{\|\,\vec{v}_h\,\|_{div}^2\,\|\,w_h\,\|_0^2} \le 1,$$

and it can be shown that

$$\frac{\underline{w}^T B\,(M+D)^{-1}\,B^T \underline{w}}{\underline{w}^T N \underline{w}} \le 1 \qquad \forall\, \underline{w} \in \mathbb{R}^m \backslash \{\underline{0}\}. \tag{2.9}$$

2.2. Preconditioning. Consider now preconditioning (1.3) with the matrix

$$P_{div} = \begin{pmatrix} A+D & 0 \\ 0 & N \end{pmatrix}. \tag{2.10}$$

From a practical point of view, D and N are cheap to assemble. The entries are constants and no extra integration is required to construct P_{div}. N is a diagonal matrix with entries $|\,K_1\,|, \ldots, |\,K_m\,|$, where $|\,K_i\,|$ denotes the area of the ith finite element. We call P_{div} an "H(div)" preconditioner since for any $\vec{v_h}$ in V_h,

$$\underline{v}^T \left(\frac{1}{\Gamma}A + D\right)\underline{v} \le \|\,\vec{v_h}\,\|_{div}^2 \le \underline{v}^T \left(\frac{1}{\gamma}A + D\right)\underline{v}, \tag{2.11}$$

where γ and Γ are the constants associated with the diffusion coefficients in (1.2). That is, $A+D$ engenders a discrete representation of a norm that is equivalent to $\|\cdot\|_{div}$ on V_h. In the trivial case $\mathcal{A} = \mathcal{I}$, the matrices M and A are identical. Thus, the leading block of P_{div} represents precisely the H_{div} operator $\mathcal{H}$ defined in (1.5). Below, we report numerical results that highlight interesting phenomena in the eigenvalue spectrum of the preconditioned matrix when other coefficients are applied. A parameterized version of (2.10) is applied to a modified saddle-point problem in [28]. Here, we are interested in parameter-free preconditioning for problems with general coefficients. We now derive an eigenvalue bound for this case.

2.3. Eigenvalues. First, it is instructive to establish the algebraic relationship between the matrices B, D, and N. This is nicely summarized in the following lemma.

LEMMA 2.1. *If $\nabla \cdot V_h \subset W_h$, then $D = B^T N^{-1} B$.*

Proof. Consider the matrices B, N, and D in operator form. That is,

$$\begin{aligned}
(\mathcal{B}\vec{x_h}, z_h) &= (z_h, \nabla \cdot \vec{x_h}) = (\mathcal{B}^T z_h,\, \vec{x_h}) && \forall\, \vec{x_h} \in V_h,\, \forall\, z_h \in W_h,\\
(\mathcal{D}\vec{x_h}, \vec{y_h}) &= (\nabla \cdot \vec{x_h}, \nabla \cdot \vec{y_h}) && \forall\, \vec{x_h}, \vec{y_h} \in V_h,\\
(\mathcal{N} z_h, w_h) &= (z_h, w_h) = (z_h,\, \mathcal{N}^{-1} w_h) && \forall\, z_h, w_h \in W_h,
\end{aligned}$$

where $\mathcal{N}$ is the identity operator on W_h. If by construction we have $\nabla \cdot V_h \subset W_h$, then

$$\begin{aligned}
(\mathcal{D}\vec{x_h}, \vec{x_h}) &= (\nabla \cdot \vec{x_h}, \nabla \cdot \vec{x_h})\\
&= (\nabla \cdot \vec{x_h},\, \mathcal{N}\nabla \cdot \vec{x_h})\\
&= (\mathcal{B}\vec{x_h},\, \mathcal{N}\nabla \cdot \vec{x_h})\\
&= (\mathcal{N}^{-1}\mathcal{B}\vec{x_h}, \nabla \cdot \vec{x_h})\\
&= (\mathcal{B}^T\mathcal{N}^{-1}\mathcal{B}\vec{x_h},\, \vec{x_h}),
\end{aligned}$$

which proves the result. ∎

It is also important to establish the dependence on h of the minimum eigenvalue of the Schur complement matrix $BA^{-1}B^T$.

LEMMA 2.2. *Let μ_{min} denote the minimum eigenvalues of $BA^{-1}B^T$. There exists a constant C, independent of h, such that $\mu_{min} \geq Ch^2$.*

Proof. This follows directly from Brezzi and Babuška's stability criterion and condition (1.2). For quasi-uniform meshes, it can also be shown that for any $w_h \in W_h$ there exist constants c_1 and c_2, independent of h, such that

$$c_1 h^2 \underline{w}^T \underline{w} \leq \| w_h \|_0^2 \leq c_2 h^2 \underline{w}^T \underline{w}, \tag{2.12}$$

where $\underline{w}$ is the vector of coefficients corresponding to the expansion of w_h in the standard basis for W_h. Combining these results, we obtain

$$\begin{aligned}
\beta^2 \| w_h \|_0^2 &\leq \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{|\,(w_h, \nabla \cdot \vec{v}_h)\,|^2}{\| \vec{v}_h \|_{div}^2} \leq \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{|\,(w_h, \nabla \cdot \vec{v}_h)\,|^2}{\| \vec{v}_h \|_0^2}\\
&\leq \Gamma \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{|\,(w_h, \nabla \cdot \vec{v}_h)\,|^2}{(\mathcal{A}^{-1}\vec{v}_h, \vec{v}_h)}.
\end{aligned}$$

Applying (2.12) and translating into matrix notation yield

$$\frac{c_1 \beta^2 h^2}{\Gamma} \leq \frac{\underline{w}^T B A^{-1} B^T \underline{w}}{\underline{w}^T \underline{w}} \quad \forall\, \underline{w} \in \mathbb{R}^m \backslash \{\underline{0}\}. \tag{2.13}$$ ∎

The following result extends the bound[3] established by Vassilevski and Lazarov in [28]. The main point is that here D supplies scaling with respect to N (which, by (2.12) is an h-dependent scaling).

LEMMA 2.3. *If T_h is a quasi-uniform mesh, the $n + m$ eigenvalues of the generalized eigenvalue problem,*

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \begin{pmatrix} A + D & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{2.14}$$

[3]The analysis here is for a different A and X and no penalty parameters.

lie in the union of the intervals,

$$\left(-1,\, -\frac{c\mu_{min}}{|\,K\,|_{min}+\mu_{min}}\right]\cup[1]\,, \tag{2.15}$$

where μ_{min} is the minimum eigenvalue of the Schur complement $BA^{-1}B^T$, $|\,K\,|_{min}$ is the area of the smallest element in T_h, and c is a constant independent of h.

Proof. The eigenvalues $\{\lambda_i\}_{i=1}^{m+n}$ satisfy

$$\begin{aligned} A\underline{u}+B^T\underline{p} &= \lambda\,(A+D)\,\underline{u},\\ B\underline{u} &= \lambda N\underline{p}. \end{aligned}$$

Suppose $\lambda = 1$; then $B^TN^{-1}B\underline{u} = D\underline{u}$, which is true, by Lemma 2.1, for every $\underline{u}\in\mathbb{R}^n$. Since $D\in\mathbb{R}^{n\times n}$ is symmetric, there are n linearly independent eigenvectors corresponding to $\lambda = 1$ and hence n distinct eigenvalues equal to 1. In addition, $\underline{u}$ and $\underline{p}$ satisfy

$$\begin{aligned} B^T\underline{p} &= (\lambda-1)\,(A+D)\,\underline{u}+D\underline{u},\\ B\underline{u} &= \lambda N\underline{p}. \end{aligned}$$

Applying Lemma 2.1 yields

$$\begin{aligned} B\,(A+D)^{-1}\,B^T\underline{p} &= (\lambda-1)\,B\underline{u}+B\,(A+D)^{-1}\,D\underline{u},\\ &= \lambda\,(\lambda-1)\,N\underline{p}+B\,(A+D)^{-1}\,B^TN^{-1}B\underline{u}\\ &= \lambda\,(\lambda-1)\,N\underline{p}+\lambda B\,(A+D)^{-1}\,B^T\underline{p}. \end{aligned}$$

Thus, the remaining m eigenvalues $\{\lambda_i\}_{i=1}^{m}$ satisfy

$$B\,(A+D)^{-1}\,B^T\underline{p} = -\lambda N\underline{p}. \tag{2.16}$$

Since $D = B^TN^{-1}B$, these are the eigenvalues of the matrix,

$$-N^{-\frac{1}{2}}B\,(A+B^TN^{-1}B)^{-1}\,B^TN^{-\frac{1}{2}}.$$

Rearranging gives

$$\begin{aligned} &N^{-\frac{1}{2}}B\,(A+B^TN^{-1}B)^{-1}\,B^TN^{-\frac{1}{2}}\\ &= N^{-\frac{1}{2}}BA^{-\frac{1}{2}}\left(I+A^{-\frac{1}{2}}B^TN^{-\frac{1}{2}}N^{-\frac{1}{2}}BA^{-\frac{1}{2}}\right)^{-1}A^{-\frac{1}{2}}B^TN^{-\frac{1}{2}}\\ &= X\,(I+X^TX)^{-1}\,X^T, \end{aligned} \tag{2.17}$$

where $X = N^{-\frac{1}{2}}BA^{-\frac{1}{2}}$. We then apply the Sherman–Morrison–Woodbury formula to obtain

$$\left(I+X^TX\right)^{-1} = I-X^T\left(I+XX^T\right)^{-1}X,$$

and so $X\,(I+XX^T)^{-1}\,X^T = X\,(I-X^T\,(I+XX^T)^{-1}\,X)\,X^T$. Now, we can apply Lemma 3.1 of [28] with $X = N^{-\frac{1}{2}}BA^{-\frac{1}{2}}$ to relate the eigenvalues of (2.16) to those of $BA^{-1}B^T$. For completeness, we reproduce this argument.

Let $\underline{x}_i$ be an eigenfunction of XX^T and σ_i denote the corresponding eigenvalue. Then,

$$\begin{aligned} X\left(I+XX^T\right)^{-1}X^T\underline{x}_i &= XX^T\underline{x}_i - XX^T\left(I+XX^T\right)^{-1}XX^T\underline{x}_i \\ &= \sigma_i\underline{x}_i - \left(\frac{\sigma_i^2}{1+\sigma_i}\right)\underline{x}_i = \left(\frac{\sigma_i}{1+\sigma_i}\right)\underline{x}_i. \end{aligned}$$

Hence, the eigenvalues of $X\left(I+XX^T\right)^{-1}X^T$ are the set of values $\{\frac{\sigma_i}{1+\sigma_i}\}_{i=1}^m$, where $\{\sigma_i\}_{i=1}^m$ are the eigenvalues of $XX^T = N^{-\frac{1}{2}}BA^{-1}B^TN^{-\frac{1}{2}}$.

Since $N^{-1}BA^{-1}B^T$ has the same eigenvalue spectrum as $N^{-\frac{1}{2}}BA^{-1}B^TN^{-\frac{1}{2}}$, the negative eigenvalues of our generalized eigenvalue problem (2.14) lie in the interval,

$$\left[-\max_i \frac{\sigma_i}{1+\sigma_i}, -\min_i \frac{\sigma_i}{1+\sigma_i}\right].$$

Since A is positive-definite and $\sigma_i > 0$ for all i, we have

$$\{\lambda_i\}_{i=1}^m \in \left(-1, -\frac{\sigma_{min}}{1+\sigma_{min}}\right],$$

where σ_{min} is the minimum eigenvalue of $N^{-1}BA^{-1}B^T$. Finally, recall that in the formulation considered here, the eigenvalues of N are the values $|K_1|, \ldots, |K_n|$. Clearly then,

$$\frac{\mu_{min}}{|K|_{max}} \leq \sigma_{min} \leq \frac{\mu_{min}}{|K|_{min}},$$

where $|K|_{min}$ and $|K|_{max}$ denote the smallest and largest eigenvalues of N, or, equivalently, the smallest and largest areas of the finite elements in T_h. Finally, we obtain

$$-\frac{\sigma_{min}}{1+\sigma_{min}} \leq -\frac{\frac{\mu_{min}}{|K|_{max}}}{1+\frac{\mu_{min}}{|K|_{min}}},$$

and so for quasi-uniform T_h, there exists a constant c, independent of h, such that

$$\{\lambda_i\}_{i=1}^m \in \left(-1, -\frac{c\mu_{min}}{|K|_{min}+\mu_{min}}\right]. \qquad \square$$

By Lemma 2.2, μ_{min} exhibits the same dependence on the discretization parameter as $|K|_{min}$. We note that this is also true in three dimensions. The eigenvalue bound established in Lemma 2.3 is optimal with respect to h. The robustness of the preconditioning is then completely determined by the dependence of μ_{min} on $\mathcal{A}$. Lemma 2.2 suggests that small coefficients can lead to small μ_{min} and thus poor MINRES convergence. For the trivial case $\mathcal{A} = \mathcal{I}$, Lemma 2.3 can be simplified to give a bound that depends only on the discrete inf-sup stability constant β.

COROLLARY 2.4. *If $\mathcal{A} = \mathcal{I}$, the $n+m$ eigenvalues of the generalized eigenvalue problem,*

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}\begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda\begin{pmatrix} A+D & 0 \\ 0 & N \end{pmatrix}\begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{2.18}$$

lie in the union of the intervals $[-1, -\beta^2] \cup [1]$*, where* β *is the discrete inf-sup constant.*

Proof. If $\mathcal{A} = \mathcal{I}$ we have $A = M$. The proof is the same as that of Lemma 2.3 except that in (2.16) we can apply (2.7) and (2.9). ∎

To obtain a practical implementation of P_{div}, $A+D$ also requires a preconditioner. It can be inverted approximately by applying a V-cycle of the specialized multigrid algorithms proposed in [1], [2], or [13]. To be more specific, the authors of [1] construct an operator $\mathcal{V}$ consisting of a V-cycle of standard geometric multigrid with additive Schwarz smoothing and prove the following result.

THEOREM 2.5. *Let* $\delta > 0$ *and let* $\mathcal{H}_\delta : H(\mathrm{div}) \times H(\mathrm{div}) \to \mathbb{R}$ *be the operator defined via* $(\mathcal{H}_\delta \vec{u}, \vec{v}) = \delta(\vec{u}, \vec{v}) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v})$*. The eigenvalues of* $\mathcal{V}\mathcal{H}$ *lie in the interval* $[1 - c, 1]$*, where* c *is a constant independent of* δ *and* h*.*

Proof. See Theorem 3.1 in [1]. ∎

Thus, $\mathcal{V}$ is an optimal preconditioner for $A + D$. Further, the condition number of $\mathcal{V}\mathcal{H}$ is independent of the coefficients if $\mathcal{A} = \delta\mathcal{I}$. As far as we are aware, there are no theoretical results concerning the robustness of the approximation when more general coefficients are present. In the rest of this section, we consider properties of the exact version of the preconditioner.

2.3.1. Numerical examples. First, we study (1.1) on $\Omega = [0, 1] \times [0, 1]$ with uniform meshes, $\mathcal{A} = I$ and $\partial\Omega = \partial\Omega_D$. The observed eigenvalues of the preconditioned system $\{\lambda_1, \ldots, \lambda_{n+m}\}$ are listed in Table 2.1; they confirm that the bound in Corollary 2.4 is tight.

TABLE 2.1
Eigenvalues of preconditioned matrix; unit coefficients.

h	$-\beta^2$	μ_{min}	$-\frac{\mu_{min}}{\lvert K\rvert_{min}+\mu_{min}}$	λ_1	λ_m	λ_{m+1}	λ_{m+n}
$\frac{1}{8}$	−0.9524	0.3124	−0.9524	−0.9993	−0.9524	1	1
$\frac{1}{16}$	−0.9519	0.0774	−0.9519	−0.9998	−0.9519	1	1
$\frac{1}{32}$	−0.9518	0.0193	−0.9518	−0.9999	−0.9518	1	1

To illustrate the role of the mesh, consider a class of anisotropic test problems. Take $\mathcal{A}(\vec{x}) = \mathtt{diag}(\epsilon, 1)$ for all $\vec{x} \in \Omega$, with $\epsilon \to \infty$ or $\epsilon \to 0$, so that $\mathcal{A}$ is ill-conditioned. The eigenvalues of the Schur complement matrix are listed in Table 2.2. The use of rectangular finite elements is crucial to the success of the preconditioning. A and consequently $BA^{-1}B^T$ have a special block structure (see [11]) that can be exploited. In this case, for $\epsilon \to 0$ and fixed h, μ_{min} does not decay to zero and MINRES convergence is insensitive to ϵ. Solving the same problem using triangular Raviart–Thomas elements, $\mu_{min} \to 0$ as $\epsilon \to 0$. To see this, compare the eigenvalues of the preconditioned systems in Figure 2.1. Note that for anisotropic coefficients and homogeneous Dirichlet boundary conditions, the analytical pressure solution has boundary layers. Anisotropic meshes should be used to resolve them. For the calculations here, uniform meshes were used simply to illustrate the asymptotic properties of the eigenvalue bound.

In practical applications, both entries on the diagonal of $\mathcal{A}$ may be small. Consider a class of discontinuous test problems. Take

$$\mathcal{A}(\vec{x}) = \begin{cases} \epsilon I & \forall\, \vec{x} \in \Omega_*, \\ I & \forall\, \vec{x} \in \Omega \backslash \Omega_*, \end{cases} \tag{2.19}$$

TABLE 2.2
Minimum eigenvalue of Schur complement, anisotropic coefficients; rectangles.

h ϵ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\frac{1}{8}$	0.1718	0.1578	0.1564	0.1562	0.1562	0.1562
$\frac{1}{16}$	0.0425	0.0391	0.0387	0.0387	0.0387	0.0387
h ϵ	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
$\frac{1}{8}$	1.7182	15.777	1.564e2	1.562e3	1.562e4	1.562e5
$\frac{1}{16}$	0.4254	3.9064	3.872e1	3.868e2	3.868e3	3.868e4

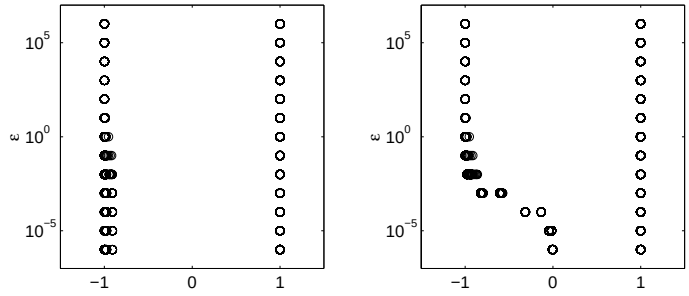


FIG. 2.1. *Eigenvalues of preconditioned system, $h = \frac{1}{8}$, rectangles (left), triangles (right).*

TABLE 2.3
Minimum eigenvalue of Schur complement, jumping coefficients.

h ϵ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\frac{1}{8}$	0.2074	0.0308	0.0036	3.648e−4	3.648e−5	3.649e−6
$\frac{1}{16}$	0.0506	0.0078	0.0008	8.051e−5	8.053e−6	8.054e−7
h ϵ	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
$\frac{1}{8}$	0.3409	0.3452	0.3456	0.3457	0.3547	0.3547
$\frac{1}{16}$	0.0846	0.0857	0.0858	0.0858	0.0858	0.0858

where I is the identity matrix and $\Omega_* \subset \Omega$. Setting $0 < \epsilon \ll 1$ describes a zone of low permeability, a feature common to groundwater flow problems. We choose $\Omega_* = [0.25, 0.5] \times [0.25, 0.75]$ and $\epsilon \in [10^{-6}, 10^{6}]$. The corresponding values of μ_{min} are listed in Table 2.3. If ϵ is large, the right-hand bound for the negative eigenvalues in Lemma 2.3 converges rapidly to -1 as $h \to 0$. MINRES convergence is fast. On the other hand, if $\epsilon \ll 1$, then μ_{min} decays to zero.

Alternatively, take $\epsilon < 1$ and

$$\mathcal{A}(\vec{x}) = \begin{cases} I & \forall\, \vec{x} \in \Omega_*, \\ \frac{1}{\epsilon} I & \forall\, \vec{x} \in \Omega \backslash \Omega_*. \end{cases} \tag{2.20}$$

The eigenvalues of the preconditioned systems associated with coefficients (2.19) and (2.20), for fixed h and $\epsilon \in [10^{-6}, 1]$, are shown in Figure 2.2. The plot on the left illustrates the decay of a subset of the negative eigenvalues of problem (2.19) as $\epsilon \to 0$. MINRES convergence is more efficient in the second case. Now, if the source term f is rescaled, solving problem (2.20) instead of problem (2.19) corresponds to multiplying the underlying PDE (1.1) by a constant and has the effect of multiplying all the eigenvalues of $BA^{-1}B^T$ by $\frac{1}{\epsilon}$. The result is that for $\epsilon \ll 1$, μ_{min} is large

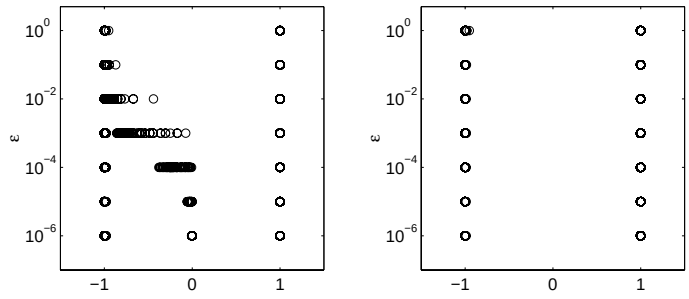


FIG. 2.2. *Eigenvalues of preconditioned system;* $h = \frac{1}{16}$*, unscaled (left) and scaled (right).*

and by Lemma 2.3, large eigenvalues of $BA^{-1}B^T$ produce a tight cluster of negative eigenvalues.

2.3.2. Preconditioned MINRES. We now report on MINRES convergence for a range of coefficients. The iteration counts listed are for exact preconditioning. Consider problem (1.1), discretized on $\Omega = [0,1] \times [0,1]$ using rectangular meshes. We apply preconditioned MINRES to the assembled system with a stopping tolerance of 10^{-6} on the relative residual error in the $\|\cdot\|_{P^{-1}}$ norm with $P = P_{div}$. The symbol $*$ signifies that more than 1,000 iterations were needed.

Example 2.1. We begin with a trivial case. Choose $\mathcal{A} = I$, $f = 1$, and homogeneous Dirichlet boundary conditions. Iteration counts for the preconditioned ($P = P_{div}$) and unpreconditioned ($P = I$) problems are given in Table 2.4; they confirm that the preconditioner is optimal. Alternatively, choose mixed boundary conditions: $p = 1$ on $\{0\} \times [0,1]$, $p = 0$ on $\{1\} \times [0,1]$, and $\vec{u} \cdot \vec{n} = 0$ on $(0,1) \times \{0,1\}$. Iteration counts are given in Table 2.5. The preconditioning is hardly affected by the boundary condition. Unless stated otherwise, a homogeneous Dirichlet condition is applied in the following examples.

TABLE 2.4
MINRES *iterations; homogeneous Dirichlet boundary condition.*

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$P = P_{div}$	5	5	5	5
$P = I$	25	75	165	311

TABLE 2.5
MINRES *iterations; mixed boundary conditions.*

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$P = P_{div}$	6	6	6	6
$P = I$	18	34	66	130

Example 2.2. The convergence is completely determined by the eigenvalues of $\mathcal{A}$. To demonstrate this, we solve (1.1) with nondiagonal coefficients,

$$\mathcal{A}(\vec{x}) = \begin{pmatrix} 1 + 4(x^2 + y^2) & 3xy \\ 3xy & 1 + 11(x^2 + y^2) \end{pmatrix}. \tag{2.21}$$

Iteration counts are reported in Table 2.6. Next we set $\mathcal{A}$ to be the diagonal matrix of the eigenvalues of $\mathcal{A}$ in (2.21). (They satisfy $1 \leq \lambda(\mathcal{A}) \leq 25$.) Iteration counts are given in Table 2.7. Convergence does not deteriorate in the nondiagonal case.

TABLE 2.6
MINRES *iterations; nondiagonal coefficients.*

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$P = P_{div}$	5	5	5	5
$P = I$	191	426	849	$*$

TABLE 2.7
MINRES *iterations; diagonal coefficients.*

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$P = P_{div}$	5	5	5	5
$P = I$	146	343	706	$*$

Example 2.3. Next, consider a problem with a coefficient that varies by three orders of magnitude across the domain. Take

$$\mathcal{A}(\vec{x}) = \begin{pmatrix} \frac{1}{1+1000(x^2+y^2)} & 0 \\ 0 & \frac{1}{1+1000(x^2+y^2)} \end{pmatrix}. \tag{2.22}$$

Iteration counts are reported in Table 2.8. Analysis of the negative eigenvalues shows that this problem is more challenging due to the small magnitude of the coefficient in some parts of the domain. The iteration count rises because μ_{min} is smaller than in the other examples. However, the preconditioner is optimal with respect to the discretization parameter.

TABLE 2.8
MINRES *iterations; variable coefficients.*

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$P = P_{div}$	24	27	27	27
$P = I$	442	$*$	$*$	$*$

Example 2.4. Finally, we test the performance of the preconditioner when discontinuous coefficients are present. Choose $\mathcal{A}$ as in (2.19) with $\epsilon \in [10^{-6}, 10^{6}]$. We take $\Omega_* = [0.25, 0.75] \times [0.25, 1]$, $f = 0$, and mixed boundary conditions $\vec{u} \cdot \vec{n} = 0$ on $\partial\Omega_N$ and $p = 1 - x$ on $\partial\Omega_D$ with $\partial\Omega_N = [0,1] \times 0 \cup \{0,1\} \times [0, 0.75]$ and $\partial\Omega_D = \Omega \backslash \Omega_N$. Without scaling, it is no surprise (recall Figure 2.2) that the iteration counts listed in Table 2.9 deteriorate as $\epsilon \to 0$. Again, this behavior is due to the small magnitude of μ_{min}.

However, if for $\epsilon < 1$, we solve the rescaled problem as discussed in section 2.3.1 by applying the coefficients (2.20), we obtain the iteration counts listed in Table 2.10. The accuracy of the solution is not unaffected by the rescaling but we can compensate for this cheaply by iterating to a lower tolerance. We apply the same stopping criterion as in the above examples with a tolerance of 10^{-9}. This is sufficient to ensure that the velocity solution to the rescaled problem is the same as that of the original problem to eight decimal places (measured in the discrete l^2 norm) for the smallest value of ϵ considered.

TABLE 2.9
MINRES *iterations; tol* $= 10^{-9}$*; unscaled.*

h	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
$\frac{1}{16}$	105	89	64	30	15	9	7	7	7	7	7	7
$\frac{1}{32}$	187	143	77	30	15	9	7	8	7	7	7	7
$\frac{1}{64}$	308	191	82	32	15	9	7	8	7	7	7	7

TABLE 2.10
MINRES *iterations; tol* $= 10^{-9}$*; scaled.*

h	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
$\frac{1}{16}$	6	6	6	6	6	7	7	7	7	7	7	7
$\frac{1}{32}$	6	6	6	7	6	7	7	8	7	7	7	7
$\frac{1}{64}$	6	6	6	7	7	7	7	8	7	7	7	7

In all of these cases, optimal rates of convergence are recovered without the use of artificial parameters. Examples 2.3 and 2.4 show that the impact of the coefficients is case dependent and not always trivial. Scaling with respect to the smallest coefficient can improve convergence. The message is that the magnitude of the coefficients and not the structure determines the robustness of this exact preconditioning scheme. Note also that when the coefficients are such that μ_{min} is small, convergence cannot be improved upon by adding multigrid. Numerical experiments not reported here confirm this fact. An alternative exact preconditioner is discussed in the next section and extended to a practical scheme.

3. $L^2 \times H^1$ formulation. It is known that the variational problem (2.2) is well-posed in a second pair of function spaces. To see this, set $V = L^2(\Omega)^2$ and

$$W = H^1_D(\Omega) = \left\{ w \in H^1(\Omega) \mid w = g \text{ on } \partial\Omega_D \right\}. \tag{3.1}$$

Multiplying by arbitrary $\vec{v} \in V$ and $w \in W$ in (2.1) and integrating the *second* equation by parts, we now look for $(\vec{u}, p) \in V \times W$ satisfying

$$\begin{aligned} (\mathcal{A}^{-1}\vec{u}, \vec{v}) - (\vec{v}, \nabla p) &= 0 && \forall\, \vec{v} \in V, \\ -(\vec{u}, \nabla w) &= -(f, w) - \langle g, \vec{u} \cdot \vec{n} \rangle && \forall\, w \in W. \end{aligned} \tag{3.2}$$

Observe that problem (3.2) is equivalent to (2.2) if $g = 0$ and $\partial\Omega = \partial\Omega_D$. Condition (2.3) certainly holds in the norm $\|\cdot\|_0$ if (1.2) holds, and it is a trivial exercise (see [3]) to show that the problem is inf-sup stable in the new norms with $\beta = \frac{1}{c}$, where c is a constant arising in Friedrich's inequality. Suppose we formulate the discrete problem using the lowest order Raviart–Thomas spaces V_h and W_h. Since $W_h \nsubseteq W$ this corresponds to a nonconforming approach and the discrete variational problem consists of finding $(\vec{u_h}, p_h) \in V_h \times W_h$ satisfying

$$\begin{aligned} (\mathcal{A}^{-1}\vec{u_h}, \vec{v_h}) - (\vec{v_h}, \nabla p_h) &= 0 && \forall\, \vec{v_h} \in V_h, \\ -(\vec{u_h}, \nabla w_h) &= -(f, w_h) && \forall\, w_h \in W_h. \end{aligned} \tag{3.3}$$

Now, since $-(\vec{v_h}, \nabla w_h) = (w_h, \nabla \cdot \vec{v_h})$ for w_h in H^1_0, the coefficient matrix of the associated linear algebra problem is identical to the one derived in section 2.

3.1. Preconditioning. It is known (see [1]) that the matrix operator associated with (3.2) defines an isomorphism from $L^2 \times H_0^1$ to $L^2 \times H^{-1}$. Block-elimination yields

$$\begin{pmatrix} \mathcal{A}^{-1}\mathcal{I} & \nabla \\ 0 & \nabla \cdot (\mathcal{A}\nabla) \end{pmatrix} \begin{pmatrix} \vec{u_h} \\ p_h \end{pmatrix} = \begin{pmatrix} 0 \\ -f_h \end{pmatrix}. \tag{3.4}$$

Thus, it is natural to approximate the coefficient operator in (3.4) by a block-diagonal matrix whose blocks are discrete representations of the operators $\mathcal{A}^{-1}\mathcal{I}$ and $\nabla \cdot (\mathcal{A}\nabla)$ acting on V_h and W_h, respectively. The Schur complement matrix $S = BA^{-1}B^T$ is the obvious choice for the scalar diffusion operator $\nabla \cdot (\mathcal{A}\nabla)$. Indeed, it can be shown that

$$P = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \tag{3.5}$$

is an optimal preconditioner for (1.3). To make the approach (3.5) feasible in practice, Rusten and Winther [20] replace S with the Cholesky factors of $BB^T \approx \nabla \cdot (\mathcal{I}\nabla)$. Unfortunately, this scheme is not robust with respect to $\mathcal{A}$. We need an approximation to $\nabla \cdot (\mathcal{A}\nabla)$ that is simple to construct and cheap to implement. However, $W_h \nsubseteq H^1$ and so it is not obvious how to achieve this.

On rectangular meshes, a simple idea is to construct a 5-point finite difference approximation to $\nabla \cdot (\mathcal{A}\nabla)$. Given $\mathcal{A}(\vec{x}) = \texttt{diag}(a_1(\vec{x}), a_2(\vec{x}))$, it is straightforward to construct a finite difference matrix on the set of element centroids, scaling x and y connections by $a_1(\vec{x})$ and $a_2(\vec{x})$, respectively. The boundary condition is imposed by extending the domain by a ficticious layer of rectangles and applying a standard centered difference. The key observation is that given *any* available P_S satisfying

$$\theta^2 \leq \frac{\underline{p}^t S \underline{p}}{\underline{p}^t P_S \underline{p}} \leq \Theta^2 \quad \forall\, \underline{p} \in \mathbb{R}^n \backslash \{\underline{0}\}, \tag{3.6}$$

with constants θ^2 and Θ^2 independent of h, an optimal preconditioner for (1.3) is

$$P_{schur} = \begin{pmatrix} A & 0 \\ 0 & P_S \end{pmatrix}. \tag{3.7}$$

This statement is made more precise in the following theorem.

THEOREM 3.1. *The eigenvalues of the generalized problem,*

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \begin{pmatrix} A & 0 \\ 0 & P_S \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{3.8}$$

lie in the intervals $[-\hat{b}, -\hat{a}] \cup [1, 1] \cup [1 + \hat{a}, 1 + \hat{b}]$, *where*

$$\hat{a} = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\theta^2}, \quad \hat{b} = -\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\Theta^2}.$$

Proof. See Theorem 2.3 in [24]. ∎

Any black-box solver for Poisson problems is a potential candidate for P_S. Thus, a range of practical and feasible preconditioners is possible. We outline *one* approach based on black-box algebraic multigrid. (See [18] and [19] for basic AMG algorithms and convergence theory and [26] for a review of current trends and related literature.)

Consider preconditioning (1.3) with the block-diagonal matrix,

$$P_{amg} = \begin{pmatrix} D_A & 0 \\ 0 & P_S \end{pmatrix}, \tag{3.9}$$

where $D_A = \mathtt{diag}(A)$ and P_S is one V-cycle of black-box AMG applied to the approximate Schur complement $S_D = BD_A^{-1}B^T$. Our choice is motivated by the observation that for diagonal coefficients, $\mathcal{A} = \mathtt{diag}(a_1(\vec{x}), a_2(\vec{x}))$, D_A is an optimal preconditioner for A (see Lemma 3.5). The matrix S_D is much sparser than S and AMG is known (see [18]) to be an efficient *solver* for sparse linear systems arising from discretizations of the operator $\nabla \cdot (\mathcal{A}\nabla)$. Moreover, AMG is more generally applicable than traditional multigrid methods to problems with anisotropic and discontinuous coefficients.

3.2. Eigenvalues. Our starting point is the eigenvalue bound established by Rusten and Winther in [20].

THEOREM 3.2. *Let $0 < \mu_1 \leq \cdots \leq \mu_n$ be the eigenvalues of A and let $0 < \rho_1 \leq \cdots \leq \rho_m$ be the singular values of B. The eigenvalues of the system matrix* (1.3) *lie in the union of the intervals,*

$$\left[\frac{1}{2}\left(\mu_1 - \sqrt{\mu_1^2 + 4\rho_m^2}\right), \frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\rho_1^2}\right)\right] \cup \left[\mu_1, \frac{1}{2}\left(\mu_n + \sqrt{\mu_n^2 + 4\rho_m^2}\right)\right]. \tag{3.10}$$

Proof. See [20]. ∎

Suppose that C in (1.3) is preconditioned with the exact (no AMG) version of the preconditioner (3.9). We obtain

$$P^{-\frac{1}{2}} C P^{-\frac{1}{2}} = \begin{pmatrix} D_A^{-\frac{1}{2}} A D_A^{-\frac{1}{2}} & D_A^{-\frac{1}{2}} B^T \left(BD_A^{-1}B^T\right)^{-\frac{1}{2}} \\ \left(BD_A^{-1}B^T\right)^{-\frac{1}{2}} BD_A^{-\frac{1}{2}} & 0 \end{pmatrix} = \begin{pmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & 0 \end{pmatrix}.$$

Applying Theorem 3.2 to the preconditioned saddle-point system leads to the following result.

COROLLARY 3.3. *Let $0 < \tilde{\mu}_1 \leq \cdots \leq \tilde{\mu}_n$ be the eigenvalues of $D_A^{-1}A$. The eigenvalues of*

$$P^{-\frac{1}{2}} \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} P^{-\frac{1}{2}}, \text{ with } P = \begin{pmatrix} D_A & 0 \\ 0 & BD_A^{-1}B^T \end{pmatrix},$$

lie in the union of the intervals,

$$\left[\frac{1}{2}\left(\tilde{\mu}_1 - \sqrt{\tilde{\mu}_1^2 + 4}\right), \frac{1}{2}\left(\tilde{\mu}_n - \sqrt{\tilde{\mu}_n^2 + 4}\right)\right] \cup \left[\tilde{\mu}_1, \frac{1}{2}\left(\tilde{\mu}_n + \sqrt{\tilde{\mu}_n^2 + 4}\right)\right].$$

Proof. Observe that

$$\tilde{B}\tilde{B}^T = \left(BD_A^{-1}B^T\right)^{-\frac{1}{2}} BD_A^{-\frac{1}{2}} D_A^{-\frac{1}{2}} B^T \left(BD_A^{-1}B^T\right)^{-\frac{1}{2}} = I,$$

where I is the identity matrix. The result follows immediately from Theorem 3.2 since the singular values of $\tilde{B}$ satisfy $\tilde{\rho}_1 = \cdots = \tilde{\rho}_m = 1$ and $\tilde{A}$ has the same eigenvalue spectrum as $D_A^{-1}A$. ∎

Now, given any preconditioner P_S for $BD_A^{-1}B^T$ satisfying

$$\tilde{\theta}^2 \leq \frac{\underline{p}^t BD_A^{-1}B^T\underline{p}}{\underline{p}^t P_S\underline{p}} \leq \tilde{\Theta}^2 \quad \forall\, \underline{p} \in \mathbb{R}^m\backslash\{\underline{0}\},$$

for some constants $\tilde{\theta}^2$ and $\tilde{\Theta}^2$, Theorem 3.2 can be extended to obtain the following theoretical eigenvalue bound.

COROLLARY 3.4. *Let $0 < \tilde{\mu}_1 \leq \cdots \leq \tilde{\mu}_n$ be the eigenvalues of $D_A^{-1}A$. The eigenvalues of*

$$P^{-\frac{1}{2}}\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}P^{-\frac{1}{2}}, \text{ with } P = \begin{pmatrix} D_A & 0 \\ 0 & P_S \end{pmatrix},$$

lie in the union of the intervals,

$$\left[\frac{1}{2}\left(\tilde{\mu}_1 - \sqrt{\tilde{\mu}_1^2 + 4\tilde{\Theta}^2}\right), \frac{1}{2}\left(\tilde{\mu}_n - \sqrt{\tilde{\mu}_n^2 + 4\tilde{\theta}^2}\right)\right] \cup \left[\tilde{\mu}_1, \frac{1}{2}\left(\tilde{\mu}_n + \sqrt{\tilde{\mu}_n^2 + 4\tilde{\Theta}^2}\right)\right].$$

Proof. The result follows from Theorem 3.2 and Corollary 3.3 with $\tilde{B} = P_S^{-\frac{1}{2}}BD_A^{-\frac{1}{2}}$. Observe that

$$\begin{aligned}\underline{p}^T\tilde{B}\tilde{B}^T\underline{p} &= \underline{p}^T P_S^{-\frac{1}{2}}BD_A^{-1}B^T P_S^{-\frac{1}{2}}\underline{p} \\ &\leq \tilde{\Theta}^2\underline{p}^T P_S^{-\frac{1}{2}}P_S P_S^{-\frac{1}{2}}\underline{p} \\ &= \tilde{\Theta}^2\underline{p}^T\underline{p}.\end{aligned}$$

That is, the maximum singular value $\tilde{\rho}_m$ of $\tilde{B}$ satisfies $\tilde{\rho}_m^2 \leq \tilde{\Theta}^2$. Similarly, it can be shown that $\tilde{\rho}_1^2 \geq \tilde{\theta}^2$. ∎

In section 3.3, we show, numerically, that black-box AMG is an optimal choice for P_S yielding constants $\tilde{\theta}^2$ and $\tilde{\Theta}^2$ that are independent of h and insensitive to $\mathcal{A}$. By Corollary 3.4, an optimal eigenvalue bound is obtained if and only if the constants $\tilde{\mu}_1$ and $\tilde{\mu}_n$ are independent of h and $\mathcal{A}$. By a result of Wathen [29] it is sufficient to consider the *element* matrices. Denoting by λ_{min}^k and λ_{max}^k the minimum and maximum eigenvalues of the diagonally preconditioned element matrix, $(D_A^k)^{-1}A^k$, associated with element k, we have

$$\min_k\{\lambda_{min}^k\} \leq \tilde{\mu}_1, \quad \tilde{\mu}_n \leq \max_k\{\lambda_{max}^k\}.$$

For simple geometries and coefficients, we can compute explicit bounds for these values. As an illustration, consider $\mathcal{A}(\vec{x}) = \texttt{diag}(a_1(\vec{x}), a_2(\vec{x}))$ and square meshes.

LEMMA 3.5. *Consider a square domain tiled with squares of edge length h. Then,*

$$\frac{1}{2} \leq \min_k\{\lambda_{min}^k\}, \quad \max_k\{\lambda_{max}^k\} \leq \frac{3}{2}. \tag{3.11}$$

Proof. For element k, label the Raviart–Thomas velocity degrees of freedom on the vertical edges (x_1, y_1) and (x_2, y_2) and those on the horizontal edges (x_3, y_3) and (x_4, y_4). We can fix normal vectors at each edge so that the element basis functions are

$$\vec{\varphi}_1 = \begin{pmatrix} \frac{x_2}{h} - \frac{x}{h} \\ 0 \end{pmatrix}, \quad \vec{\varphi}_2 = \begin{pmatrix} -\frac{x_1}{h} + \frac{x}{h} \\ 0 \end{pmatrix}, \quad \vec{\varphi}_3 = \begin{pmatrix} 0 \\ \frac{y_4}{h} - \frac{y}{h} \end{pmatrix}, \quad \vec{\varphi}_4 = \begin{pmatrix} 0 \\ -\frac{y_3}{h} + \frac{y}{h} \end{pmatrix}.$$

Now, let

$$\mathcal{A}_c^k = \begin{pmatrix} a_1(x_c^k, y_c^k) & 0 \\ 0 & a_2(x_c^k, y_c^k) \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix},$$

denote $\mathcal{A}$ evaluated at the centroid (x_c^k, y_c^k). The element contribution to A is then defined by

$$A^k(i,j) = \iint_{\square} (\mathcal{A}_c^k)^{-1} \vec{\varphi}_i \cdot \vec{\varphi}_j \, dxdy, \quad i,j = 1:4.$$

Integrating yields

$$A^k = h^2 \begin{pmatrix} \frac{1}{3a_1} & \frac{1}{6a_1} & 0 & 0 \\ \frac{1}{6a_1} & \frac{1}{3a_1} & 0 & 0 \\ 0 & 0 & \frac{1}{3a_2} & \frac{1}{6a_2} \\ 0 & 0 & \frac{1}{6a_2} & \frac{1}{3a_2} \end{pmatrix}, \quad (D_A^k)^{-1} A^k = \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 1 \end{pmatrix}. \tag{3.12}$$

Hence, $\lambda_1^k = \frac{1}{2}$, $\lambda_2^k = \frac{1}{2}$, $\lambda_3^k = \frac{3}{2}$, $\lambda_4^k = \frac{3}{2}$. ∎

A similar result can be proved for triangles. Note, however, that for nondiagonal $\mathcal{A}$, the values $\tilde{\mu}_{min}$ and $\tilde{\mu}_{max}$ will depend on the coefficients. The exact nature of this is case dependent but can easily be evaluated. Other diagonal approximations to the matrix A may have to be considered.

3.3. Preconditioned MINRES. The following MINRES trials illustrate, numerically, the robustness of the practical preconditioner (3.9) for a range of diagonal coefficients. We apply one V-cycle ($P_S = V$) of the algorithm `amg1r5` (see [18]) to the sparse matrix $S_D = BD_A^{-1}B^T$. We implement it as a black-box with symmetric smoothing; no parameters are estimated a priori. The experiments are performed using a `mex` Fortran interface in Matlab 6.0 on a SUN ultraSPARC workstation. In each example, problem (1.1) is discretized on $[0,1] \times [0,1]$. We apply the same stopping criteria as in section 2.3.2 with a tolerance of 10^{-6}.

Example 3.1. $\mathcal{A} = I$, $f = 1$, $g = 0$. Iteration counts are reported in Table 3.1. The second column corresponds to the unpreconditioned case. Results for the exact version of the preconditioner ($P_S = S_D$) are also listed. The preconditioned iteration counts decrease slightly with mesh refinement. The observed eigenvalues are listed in Table 3.2. We obtain $\tilde{\theta}^2 \approx 0.9453$ and $\tilde{\Theta}^2 = 1$. Substituting these values in Corollary 3.4 gives the theoretical bound $[-0.7808, -0.4778] \cup [0.5, 2]$.

TABLE 3.1
MINRES *iterations.*

h	$P = I$	$P_S = S_D$	$P_S = V$
$\frac{1}{16}$	75	24	25
$\frac{1}{32}$	165	23	24
$\frac{1}{64}$	311	20	22
$\frac{1}{128}$	574	17	19

Example 3.2. Now choose $\mathcal{A}$ to be the variable coefficient matrix (2.22). Here, the eigenvalues are very close to those of Example 3.1 despite the variation in $\mathcal{A}$. This is reflected in the iteration counts shown in Table 3.3. The eigenvalues of the preconditioned system are listed in Table 3.4. We obtain $\tilde{\theta}^2 \approx 0.9460$ and $\tilde{\Theta}^2 = 1$,

TABLE 3.2
Eigenvalues of indefinite preconditioned system; $P_S = V$.

h	$\tilde{\mu}_1$	$\tilde{\mu}_n$	$\tilde{\theta}^2$	$\tilde{\Theta}^2$	Observed eigenvalues
$\frac{1}{8}$	0.5	1.5	0.9564	1	$[-0.7772, -0.4931] \cup [0.5, 1.9553]$
$\frac{1}{16}$	0.5	1.5	0.9501	1	$[-0.7787, -0.4833] \cup [0.5, 1.9739]$
$\frac{1}{32}$	0.5	1.5	0.9453	1	$[-0.7803, -0.4794] \cup [0.5, 1.9781]$

TABLE 3.3
MINRES *iterations.*

h	$P = I$	$P_S = S_D$	$P_S = V$
$\frac{1}{16}$	*	26	26
$\frac{1}{32}$	*	23	25
$\frac{1}{64}$	*	20	22
$\frac{1}{128}$	*	19	21

TABLE 3.4
Eigenvalues of indefinite preconditioned system; $P_S = V$.

h	$\tilde{\mu}_1$	$\tilde{\mu}_n$	$\tilde{\theta}^2$	$\tilde{\Theta}^2$	Observed eigenvalues
$\frac{1}{8}$	0.5	1.5	0.9573	1	$[-0.7772, -0.4947] \cup [0.5, 1.9514]$
$\frac{1}{16}$	0.5	1.5	0.9507	1	$[-0.7786, -0.4844] \cup [0.5, 1.9735]$
$\frac{1}{32}$	0.5	1.5	0.9460	1	$[-0.7802, -0.4803] \cup [0.5, 1.9783]$

yielding the theoretical bound $[-0.7808, -0.4782] \cup [0.5, 2]$. Since we are dealing with diagonal coefficients, D_A is an optimal preconditioner for A. The robustness of P_{amg} is completely determined by the constants $\tilde{\theta}^2$ and $\tilde{\Theta}^2$ in this case. The numerical experiments show that these values are not sensitive to the magnitude of the coefficients.

Example 3.3. Finally, we consider a discontinuous coefficient example with mixed boundary conditions (Example 2.4 with $\epsilon < 1$). Iteration counts for the AMG preconditioner are reported in Table 3.5. Eigenvalues for the case $\epsilon = 10^{-3}$ are listed in Table 3.6. We obtain $\tilde{\theta}^2 \approx 0.8952$ and $\tilde{\Theta}^2 = 1$, yielding the theoretical bound $[-0.7808, -0.4574] \cup [0.5, 2]$. The AMG preconditioner is completely insensitive to the coefficient ϵ.

TABLE 3.5
MINRES *iterations.*

h	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\frac{1}{16}$	32	30	30	30	30	30
$\frac{1}{32}$	32	32	32	32	32	32
$\frac{1}{64}$	33	33	33	33	33	33
$\frac{1}{128}$	33	33	32	32	32	32

A key observation in all of these examples is that solve times grow only linearly with respect to the problem size, making this a cheap and feasible solution scheme. For diagonal $\mathcal{A}$, the practical preconditioner P_{amg} is insensitive to the magnitude of the coefficients.

TABLE 3.6
Eigenvalues of indefinite preconditioned system; $P_S = V$, $\epsilon = 10^{-3}$.

h	$\tilde{\mu}_1$	$\tilde{\mu}_n$	$\tilde{\theta}^2$	$\tilde{\Theta}^2$	Observed eigenvalues
$\frac{1}{16}$	0.5	1.5	0.9262	1	$[-0.7772, -0.4820] \cup [0.5, 1.9737]$
$\frac{1}{32}$	0.5	1.5	0.8952	1	$[-0.7798, -0.4595] \cup [0.5, 1.9784]$

4. Concluding remarks. In this paper we have explained—without penalty parameters or reduction techniques—why the exact preconditioners P_{div} and P_{schur} are optimal with respect to the discretization parameter. We have explored the impact of the coefficient $\mathcal{A}$ and identified the key parameters that determine the efficiency of MINRES convergence. The results apply to the lowest order Raviart–Thomas spaces in $\mathbb{R}^2$ and $\mathbb{R}^3$. A practical preconditioner P_{amg} based on black-box algebraic multigrid was proposed for diagonal coefficients. There are no parameters to estimate. Numerical experiments show that the resulting solution scheme is insensitive to the discretization parameter and is robust with respect to jumps and variations in the coefficients.

Acknowledgment. We would like to thank the Fraunhofer Institute SCAI (a former GMD institute) for permission to publish numerical results obtained using the Fortran 77 code `amg1r5`.

REFERENCES

[1] D.N. ARNOLD, R.S. FALK, AND R. WINTHER, *Multigrid in $H(div)$ and $H(curl)$*, Numer. Math., 85 (2000), pp. 197–218.
[2] D.N. ARNOLD, R.S. FALK, AND R. WINTHER, *Preconditioning in $H(div)$ and applications*, Math. Comp., 66 (1997), pp. 957–984.
[3] D. BRAESS, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.
[4] J.H. BRAMBLE, *Multigrid Methods*, Pitman Res. Notes Math. Ser. 294, Longman, Harlow, UK, 1993.
[5] F. BREZZI AND K.J. BATHE, *A discourse on the stability conditions for mixed finite element formulations*, Comput. Methods Appl. Mech. Engrg., 82 (1990), pp. 27–57.
[6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[7] Z. CHEN, R.E. EWING, AND R. LAZAROV, *Domain decomposition algorithms for mixed methods for second-order elliptic problems*, Math. Comp., 65 (1996), pp. 467–490.
[8] K.A. CLIFFE, I.G. GRAHAM, R. SCHEICHL, AND L. STALS, *Parallel computation of flow in heterogeneous media modelled by mixed finite elements*, J. Comput. Phys., 164 (2000), pp. 258–282.
[9] R.E. EWING AND J. WANG, *Analysis of the Schwartz algorithm for mixed finite elements methods*, M2AN Math. Model. Numer. Anal., 26 (1992), pp. 739–756.
[10] R.E. EWING AND J. WANG, *Analysis of multilevel decomposition iterative methods for mixed finite element methods*, M2AN Math. Model. Numer. Anal., 28 (1994), pp. 377–398.
[11] R.E. EWING, R.D. LAZAROV, P. LU, AND P.S. VASSILEVSKI, *Preconditioning indefinite systems arising from mixed finite element discretization of second-order elliptic problems*, in Preconditioned Conjugate Gradient Methods, Lecture Notes in Math. 1457, Springer, Berlin, 1990, pp. 28–43.
[12] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.
[13] R. HIPTMAIR, *Multigrid methods for $H(div)$ in three dimensions*, Electron. Trans. Numer. Anal., 6 (1997), pp. 133–152.
[14] C.C. PAIGE AND M.A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[15] A. QUARTERONI AND A. VALLI, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, New York, 1994.

[16] P.A. Raviart and J.M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977.
[17] J.E. Roberts and J.M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II: Finite Element Methods (Part 1), Handb. Numer. Anal., II, P.G. Ciarlet and J.L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 523–639.
[18] J.W. Ruge and K. Stüben, *Efficient solution of finite difference and finite element equations by algebraic multigrid (AMG)*, in Multigrid Methods for Integral and Differential Equations, Inst. Math. Appl. Conf. Ser. New Ser. 3, D.J. Paddon and H. Holstein, eds., Oxford, UK, 1985, pp. 169–212.
[19] J.W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, Frontiers Appl. Math. 3, S.F. McCormick, ed., SIAM, Philadelphia, 1987, pp. 73–130.
[20] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.
[21] T. Rusten and R. Winther, *Substructure preconditioners for elliptic saddle point problems*, Math. Comp., 60 (1993), pp. 23–48.
[22] T. Rusten, P.S. Vassilevski, and R. Winther, *Interior preconditioners for mixed finite element approximations of elliptic problems*, Math. Comp., 65 (1996), pp. 447–466.
[23] R. Scheichl, *Iterative Solution of Saddle-Point Problems Using Divergence-Free Finite Elements with Applications to Groundwater Flow*, Ph.D. thesis, University of Bath, Bath, UK, 2000.
[24] D. Silvester and A. Wathen, *Fast iterative solution of stabilised stokes systems part* II: *Using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1367.
[25] D. Silvester and A. Wathen, *Fast and robust solvers for time-discretised incompressible Navier–Stokes equations*, in Numerical Analysis, Dundee, 1995, Pitman Res. Notes Math. Ser. 344, D.F. Griffiths and G.A. Watson, eds., Longman, Harlow, UK, 1996.
[26] K. Stüben, *A review of algebraic multigrid*, J. Comput. Appl. Math., 128 (2001), pp. 281–309.
[27] P.S. Vassilevski and J. Wang, *Multilevel iterative methods for mixed finite element discretizations of elliptic problems*, Numer. Math., 63 (1992), pp. 503–520.
[28] P.S. Vassilevski and R.D. Lazarov, *Preconditioning mixed finite element saddle-point elliptic problems*, Numer. Linear Algebra Appl., 3 (1996), pp. 1–20.
[29] A.J. Wathen, *Realistic eigenvalue bounds for the Galerkin mass matrix*, IMA J. Numer. Anal., 7 (1987), pp. 449–457.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 739–751

ON THE CONVERGENCE TO ZERO OF INFINITE PRODUCTS OF INTERVAL MATRICES*

SY-MING GUU† AND CHIN-TZONG PANG‡

Abstract. A necessary and sufficient condition for the consecutive powers of an interval matrix to converge to the null matrix was established by Mayer. Motivated by the issue of globally asymptotic stability of Takagi–Sugeno free fuzzy systems with time-varying uncertainty, we study the conditions for the infinite products of a finite number of interval matrices to converge to the null matrix. As an application, convergence to null matrix of infinite products of the associated finite interval matrices implies the globally asymptotic stability of Takagi–Sugeno free fuzzy systems with time-varying uncertainty.

Key words. norms of matrices, infinite products of interval matrices, free fuzzy systems

AMS subject classifications. 15A60, 40A20, 93C42

DOI. 10.1137/S0895479802408840

1. Introduction.

In 1984, Mayer [9] established a remarkable result that the sequence of consecutive powers of an interval matrix converges to the zero matrix if and only if a constructed real matrix has its spectral radius less than 1. The way to construct the associated real matrix is based on the *property* $(*)$ (see Definition 2.1). Pang, Lur, and Guu [12] provided a new proof for Mayer's convergence theorem. Pang [10] extended Mayer's conditions to study the stability of linear interval systems. Shih, Lur, and Pang [14] applied the extended Mayer conditions to study the asymptotic stability of discrete-time linear interval systems.

In the current paper, we study the conditions for the infinite products of a finite number of interval matrices to converge to the null matrix. The motivation for the case with a finite number of interval matrices comes from our stability study of Takagi–Sugeno free fuzzy systems[1] with time-varying uncertainty. Our proof for the convergence of infinite products of a finite number of interval matrices is based on our new proof for Mayer's convergence theorem [12]. Characterizations for the convergence of the infinite products of a finite number of interval matrices are established as well.

As an application, we apply the extended Mayer convergence theorem to the study of globally asymptotic stability of Takagi–Sugeno free fuzzy systems with time-

*Received by the editors May 29, 2002; accepted for publication (in revised form) by U. Helmke July 1, 2003; published electronically January 30, 2004.
http://www.siam.org/journals/simax/25-3/40884.html

†Department of Business Administration, Yuan Ze University, Taoyuan, Taiwan, 320, R.O.C. (iesmguu@saturn.yzu.edu.tw). The research of this author was supported by NSC 91-2213-E-155-037.

‡Department of Information Management, Yuan Ze University, Taoyuan, Taiwan, 320, R.O.C. (imctpang@saturn.yzu.edu.tw). The research of this author was supported by NSC 91-2213-E-155-055.

[1]Fuzzy systems have been employed in the control processes in many practical applications. Two major types of fuzzy control systems are prevailing: the Mamdani fuzzy systems and the Takagi–Sugeno fuzzy systems [16]. The major difference between these two systems lies in their consequent parts of fuzzy rules [18]. The consequent parts of the rules of Mamdani systems are fuzzy sets, while the Takagi–Sugeno systems employ linear functions of system states and inputs. Matrix representations of these linear functions are called the characteristic matrices. For the stability of deterministic linear fuzzy systems, we refer to [6, 17].

varying uncertainty. The linkage between two topics lies in the representation[2] of the time-varying uncertainty incurred in characteristic matrices, where these matrices employed in each iteration come from certain interval matrices (Pang and Guu [11]).

The rest of the paper is as follows. Notations and preliminary materials are given in section 2. Section 3 contains the main results for the convergence of infinite products of a finite number of interval matrices. In section 4, we shall review the background of Takagi–Sugeno free fuzzy systems with time-varying uncertainty. Then an application of the extended Mayer's convergence theorem to the globally asymptotic stability of Takagi–Sugeno free fuzzy systems with time-varying uncertainty is mentioned.

2. Preliminaries. Let Σ be a bounded set of $n \times n$ complex matrices. For $m \geq 1$, Σ^m is the set of all products of matrices in Σ of length m, that is,

$$\Sigma^m = \{A_1 A_2 \cdots A_m : A_i \in \Sigma,\ i = 1, 2, \ldots, m\}.$$

The set $\Sigma' := \cup_{m \geq 1} \Sigma^m$ denotes the multiplicative semigroup generated by Σ. $\rho(A)$ and $||A||$ denote the spectral radius and an operator norm of matrix A, respectively. The *joint spectral radius* of Σ (see Rota and Strang [13]), $\hat{\rho}(\Sigma)$, is defined by

$$\hat{\rho}(\Sigma) = \limsup_{m \to \infty} \left[\sup_{A \in \Sigma^m} \|A\| \right]^{\frac{1}{m}}.$$

The *generalized spectral radius* of Σ (see Daubechies and Lagarias [4]), $\rho(\Sigma)$, is defined by

$$\rho(\Sigma) = \limsup_{m \to \infty} \left[\sup_{A \in \Sigma^m} \rho(A) \right]^{\frac{1}{m}}.$$

It has been proved that $\rho(\Sigma) = \hat{\rho}(\Sigma)$ (see [2, 5, 15]).

We refer to Alefeld and Herzberger [1] for the background materials of interval matrices. Real numbers are denoted by lowercase letters $a, b, \ldots$. The $\bar{a}$ and $\underline{a}$ denote the upper and lower bounds, respectively, of a real closed interval $[\underline{a}, \bar{a}]$. The set of all these closed intervals is denoted by $I(\mathbb{R})$. Interval matrices with entries belonging to $I(\mathbb{R})$ are denoted by $\mathcal{A}, \mathcal{B}, \ldots$. It is convenient to denote $\mathcal{A} = (A_{ij})$, $\mathcal{B} = (B_{ij}), \ldots$. Two interval matrices (A_{ij}) and (B_{ij}) are equal if and only if $A_{ij} = B_{ij}$ for all i and j. Let $* \in \{+, -, \cdot, /\}$ be one of the usual binary operations on the set of real numbers $\mathbb{R}$. For $X, Y \in I(\mathbb{R})$ the binary operation

$$X * Y := \{x * y : x \in X, y \in Y\},$$

assuming that $0 \notin Y$ in the case of division. Let $A = (a_{ij})$ be a real matrix. A is *generated* from interval matrix $\mathcal{A}$ (denoted by $A \in \mathcal{A}$) if $a_{ij} \in A_{ij}$ for each i and j. The set $M_n(\mathbb{R})$ denotes all $n \times n$ real compact interval matrices.

Real matrices and interval matrices in this paper are of size $n \times n$. For interval matrices (A_{ij}) and (B_{ij}), and an interval $X \in I(\mathbb{R})$, the matrix operations $+, -, \cdot$ are formally defined as

$$\begin{aligned} \mathcal{A} \pm \mathcal{B} &:= (A_{ij} \pm B_{ij}), \\ \mathcal{A} \cdot \mathcal{B} &:= \left(\sum_s A_{is} \cdot B_{sj} \right), \\ X \cdot \mathcal{A} &:= (X \cdot A_{ij}). \end{aligned}$$

[2] Joh, Chen, and Langari [7] considered a similar representation yet with the issue of quadratic stability.

Let I be an $n \times n$ identity matrix. The powers of interval matrix $\mathcal{A}$ are defined as

$$\begin{aligned} &\mathcal{A}^0 := I, \\ &\mathcal{A}^k := \mathcal{A}^{k-1} \cdot \mathcal{A}, \quad k = 1, 2, \ldots. \end{aligned}$$

As noted by Mayer [9], the product of interval matrices is not associative in general. Therefore, $(\mathcal{A} \cdot \mathcal{B}) \cdot \mathcal{C}$ may not be equal to $\mathcal{A} \cdot (\mathcal{B} \cdot \mathcal{C})$.

For a compact interval $[\underline{a}, \overline{a}]$, the *width* $\mathrm{d}\,([\underline{a}, \overline{a}])$ and the *absolute value* $|[\underline{a}, \overline{a}]|$ are defined by

$$\mathrm{d}\,[\underline{a}, \overline{a}] = \overline{a} - \underline{a}, \qquad |[\underline{a}, \overline{a}]| = \max\{|\underline{a}|, |\overline{a}|\},$$

respectively. For an $n \times n$ real interval matrix $\mathcal{A} = (A_{ij})$, we define two nonnegative real matrices $\mathrm{d}\,(\mathcal{A}) = (\mathrm{d}\,(A_{ij}))$ and $|\mathcal{A}| = (|A_{ij}|)$, which are called the *width* and *absolute value* of the interval matrix $\mathcal{A} = (A_{ij})$, respectively.

On the set $M_n(\mathbb{R})$ of real $n \times n$ matrices we introduce the usual partial ordering $\leq$ by defining $(a_{ij}) \leq (b_{ij})$ if and only if $a_{ij} \leq b_{ij}$ for all $1 \leq i, j \leq n$. Then the following important properties of $\mathrm{d}\,(\mathcal{A})$ and $|\mathcal{A}|$ can be found in Mayer [9]:

$$0 \leq \mathrm{d}\,(\mathcal{A}),\ 0 \leq |\mathcal{A}|,$$

$$\mathrm{d}\,(\mathcal{A}) = 0 \Leftrightarrow \mathcal{A} \text{ is a real matrix},$$

$$|\mathcal{A}| = 0 \Leftrightarrow \mathcal{A} = 0,$$

$$\mathrm{d}\,(\mathcal{A} \pm \mathcal{B}) = \mathrm{d}\,(\mathcal{A}) + \mathrm{d}\,(\mathcal{B}),$$

$$|\mathcal{A} \pm \mathcal{B}| \leq |\mathcal{A}| + |\mathcal{B}|,$$

$$\mathrm{d}\,(\mathcal{A})|\mathcal{B}|,\ |\mathcal{A}|\mathrm{d}\,(\mathcal{B}) \leq \mathrm{d}\,(\mathcal{A} \cdot \mathcal{B}) \leq \mathrm{d}\,(\mathcal{A})|\mathcal{B}| + |\mathcal{A}|\mathrm{d}\,(\mathcal{B}),$$

$$|\mathcal{A} \cdot \mathcal{B}| \leq |\mathcal{A}||\mathcal{B}|.$$

Definition 2.1 (Mayer [9]). *Let $\mathcal{A}$ be an $n \times n$ interval matrix. We say that the jth column of $\mathcal{A}$ has the property $(*)$ if there exists a power $\mathcal{A}^m$ containing in the same jth column at least one interval not degenerated to a point interval. Furthermore, a real matrix $A = (a_{ij})$ can be constructed as*

$$a_{ij} = \begin{cases} |A_{ij}| & \text{if the } j\text{th column of } \mathcal{A} \text{ has the property } (*), \\ A_{ij} & \text{otherwise.} \end{cases}$$

Let Ψ be a finite set in $M_n(\mathbb{R})$. Denote by Ψ^m the set of all products [3] *of interval matrices in Ψ of length m, that is,*

$$\Psi^m = \{\mathcal{A}_1 \mathcal{A}_2 \cdots \mathcal{A}_m : \mathcal{A}_i \in \Psi, i = 1, 2, \ldots, m\}.$$

[3] To ease our notation, we omit the $\cdot$ and parentheses in products of interval matrices. Namely, $\mathcal{A}\mathcal{B}\mathcal{C}$, $\mathcal{A}\mathcal{B}\mathcal{C}\mathcal{D}$, etc., instead of $(\mathcal{A} \cdot \mathcal{B}) \cdot \mathcal{C}$, $((\mathcal{A} \cdot \mathcal{B}) \cdot \mathcal{C}) \cdot \mathcal{D}$, etc.

The set $\Psi' = \bigcup_{m\geq 1} \Psi^m$ *denotes the multiplicative semigroup generated by* Ψ.

We shall now turn to extend Mayer's property $(*)$ *to the case with multiple interval matrices. For* $j = 1, 2, \ldots, n$, $\Psi_*(j)$ *denotes the set of any interval matrix* $\mathcal{A} \in \Psi'$ *with its* j*th column containing at least one nondegenerated interval element.*

DEFINITION 2.2. *Let* Ψ *be a set in* $M_n(\mathbb{IR})$ *and* $j \in \{1, 2, \ldots, n\}$. *The set* Ψ *has the* generalized property $(*)$ in j *if* $\Psi_*(j)$ *is nonempty. Furthermore, for each interval matrix* $\mathcal{A}$ *in* Ψ, *we construct a real matrix* $\tilde{\mathcal{A}} = (a_{ij})$ *by*

$$a_{ij} = \begin{cases} |A_{ij}| & \text{if } \Psi \text{ has the generalized property } (*) \text{ in } j, \\ A_{ij} & \text{otherwise.} \end{cases}$$

We denote $\tilde{\Psi} = \{\tilde{\mathcal{A}} : \mathcal{A} = (A_{ij}) \in \Psi\}$.

When Ψ *consists of a single interval matrix* $\mathcal{A}$, *the generalized property* $(*)$ *becomes Mayer's property* $(*)$.

Example 2.1. Consider the following set:

$$\Psi = \left\{ \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ [0,1] & 0 \end{pmatrix} \right\}.$$

The set Ψ has the generalized property $(*)$ in $j = 1$ and $j = 2$, respectively. Moreover,

$$\tilde{\Psi} = \left\{ \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right\}.$$

3. Main results. Motivated by the study of globally asymptotic stability of Takagi–Sugeno free fuzzy systems with time-varying uncertainty, we shall consider a finite set Ψ in this paper. Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ be a set in $M_n(\mathbb{IR})$. The main result in this section is to establish the convergence of infinite products of interval matrices in Ψ. Precisely, we shall establish conditions for any sequence $\{\mathcal{A}_k \in \Psi : k \in \mathbb{N}\}$ to have $\lim_{k\to\infty} \mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_k = 0$. As a notation, $|\Psi| = \{|\mathcal{A}^{(1)}|, |\mathcal{A}^{(2)}|, \ldots, |\mathcal{A}^{(N)}|\}$.

In this section, we shall present eight lemmas and three theorems. Among those results, Theorem 3.11 represents the most important consequence of this paper. The following three lemmas are needed to establish Lemma 3.4, which will play a role in the proof of Theorem 3.6.

LEMMA 3.1 (König [8]). *If* G *is an infinite graph such that* G *is connected and locally finite (i.e., each vertex of* G *has finite degree), then for any vertex* ν *of* G, *there exists an infinite path with initial vertex* ν.

LEMMA 3.2. *Let* $\Sigma = \{A^{(1)}, A^{(2)}, \ldots, A^{(N)}\}$ *be a set in* $M_n(\mathbb{C})$. *Then* $\hat{\rho}(\Sigma) < 1$ *if and only if there exists a norm* $||\cdot||$ *on* $\mathbb{C}^n$ *such that* $||A^{(i)}|| < 1$ *for* $i = 1, 2, \ldots, N$.

Proof. The direction "$\Leftarrow$" is trivial. On the other hand, the direction "$\Rightarrow$" follows from Rota and Strang's theorem [13].

LEMMA 3.3 (Elsner [5]). *Let* $||\cdot||$ *denote a vector norm on* $\mathbb{C}^n$ *and its operator norm in the space of* $n \times n$ *matrices. There exists a constant* c depending on $||\cdot||$ *such that for any* $z \in \mathbb{C}^n$ *with* $||z|| = 1$, *any* $n \times n$ *matrix* A *with* $||A|| \leq 1$, *and eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_n$, *the following inequality holds:*

$$\min_i |1 - \lambda_i| \leq c||Az - z||^{\frac{1}{n}}.$$

LEMMA 3.4. *Let* $\Sigma = \{A^{(1)}, A^{(2)}, \ldots, A^{(N)}\}$ *be a set in* $M_n(\mathbb{C})$ *and* Σ' *be bounded. If* $\lim_{k\to\infty}[\max_{A\in\Sigma^k} \rho(A)] = 0$, *then there exists a norm* $||\cdot||$ *on* $\mathbb{C}^n$ *such that* $||A^{(i)}|| < 1$ *for* $i = 1, 2, \ldots, N$.

Proof.[4] Since Σ' is bounded, we can construct a norm by

$$||x|| := \sup\{||Mx||_2 : M \in \Sigma' \cup \{I\}\}, \quad x \in \mathbb{C}^n.$$

Then for each i the inequality $||A^{(i)}x|| \leq ||x||$ holds for $x \in \mathbb{C}^n$. This in turn implies that $||A^{(i)}|| \leq 1$ for $i = 1, 2, \ldots, N$ (see Bonsall and Duncan [3, p. 21].) For a finite sequence $w = (n_1, n_2, \ldots, n_s)$ of integers in $\{1, 2, \ldots, N\}$, we define

$$A_w^* := A^{(n_s)} \cdots A^{(n_1)} \in \Sigma^s.$$

Let $W = \{w : ||A_w^*|| = 1\}$. We shall show that the set W is finite. Suppose to the contrary that W is infinite; by König's lemma there exists a sequence $\{d_i\}_{i=1}^{\infty}$ with $1 \leq d_i \leq N$ such that $||T_k|| = 1$ for all $k \in \mathbb{N}$, where $T_k = A^{(d_{n_k})} \cdots A^{(d_{n_1})}$. Choose a sequence $\{x_n\}_{n=1}^{\infty}$ with $||x_n|| = ||T_n x_n|| = 1$ for $n \in \mathbb{N}$. We can select an appropriate subsequence $\{n_i\}_{i=1}^{\infty}$ such that

$$x_{n_i} \to \zeta,\ T_{n_i}\zeta \to \eta \text{ as } i \to \infty.$$

From $T_n\zeta = T_n x_n + T_n(\zeta - x_n)$, we get $||\eta|| = 1$. Hence for any given $\epsilon > 0$ we can choose $r = n_i$, $s = n_{i+1}$ such that

$$||T_s\zeta - T_r\zeta|| \leq \epsilon \text{ and } ||T_r\zeta|| \geq \frac{1}{2}.$$

Set $T_s = BT_r$ for some $B \in \Sigma^{s-r}$ and let z denote $\frac{T_r\zeta}{||T_r\zeta||}$. As $||B|| \leq 1$, we have $||Bz - z|| \leq 2\epsilon$. By Lemma 3.3, we obtain a sequence $\{A_k\}$ in Σ' such that $\lim_{k\to\infty} \rho(A_k) = 1$. Then there exists a subsequence $\{A_{n_k}\}$ with $A_{n_k} \in \Sigma^{l(n_k)}$, $l(n_1) < l(n_2) < \cdots$, so that $\lim_{k\to\infty}[\max_{A_{n_k}\in\Sigma^{l(n_k)}} \rho(A_{n_k})] \neq 0$, a contradiction. Therefore, W is finite. Now choose k so that $\max_{A\in\Sigma^k} ||A|| < 1$. Define

$$|||x||| = ||x|| + \max_{B\in\Sigma} ||Bx|| + \cdots + \max_{B\in\Sigma^{k-1}} ||Bx||, \quad x \in \mathbb{C}^n.$$

Then for any $A^{(i)} \in \Sigma$ we have

$$|||A^{(i)}x||| < |||x||| \text{ for all } x \in \mathbb{C}^n.$$

Hence $|||A^{(i)}||| < 1$ for $i = 1, 2, \ldots, N$.

The following important lemma by Elsner plays the key role in establishing Theorem 3.6.

LEMMA 3.5 (Elsner [5]). *Let Σ be a bounded set in $M_n(\mathbb{C})$ and $\hat{\rho}(\Sigma) = 1$. If Σ' is unbounded, then there is a nonsingular matrix S and $1 \leq n_1 < n$ such that for all $A \in \Sigma$,*

$$S^{-1}AS = \begin{pmatrix} A_{(2)} & * \\ 0 & A_{(1)} \end{pmatrix},$$

where $A_{(1)} \in M_{n_1}(\mathbb{C})$.

THEOREM 3.6. *Let $\Sigma = \{A^{(1)}, A^{(2)}, \ldots, A^{(N)}\}$ be a set in $M_n(\mathbb{C})$. Then $\lim_{k\to\infty}$ $[\max_{A\in\Sigma^k} \rho(A)] = 0$ if and only if there exists a norm $||\cdot||$ on $\mathbb{C}^n$ such that $||A^{(i)}|| < 1$ for $i = 1, 2, \ldots, N$.*

[4]Part of this proof adopts the approach employed in the proof of Lemma 3 in [5], where Professor Elsner established conditions for $\rho(\Sigma) = \hat{\rho}(\Sigma)$ (see [4] for the issue of $\rho(\Sigma) = \hat{\rho}(\Sigma)$).

Proof. The direction "⇐" holds because of

$$\max_{A\in\Sigma^k}\rho(A)\le\left(\max_{1\le i\le N}||A_i||\right)^k \quad \text{for } k=1,2,\ldots.$$

"⇒" By Lemma 3.2, it suffices to show that $\hat{\rho}(\Sigma)<1$. We proceed with the proof by induction on the dimension n. The assertion is true for $n=1$. Assume that the theorem holds for dimensions $1,2,\ldots,n-1$. Let $\mathcal{B}=\Sigma/\hat{\rho}(\Sigma)$. Then $\hat{\rho}(\mathcal{B})=1$. We shall show that this theorem holds for dimension n. Suppose to the contrary that $\hat{\rho}(\Sigma)\ge 1$. Then

$$\lim_{k\to\infty}\left[\max_{A\in\mathcal{B}^k}\rho(A)\right]=0.$$

If $\mathcal{B}'$ is bounded, by Lemmas 3.2 and 3.4, $\hat{\rho}(\mathcal{B})<1$, which violates $\hat{\rho}(\mathcal{B})=1$. Therefore, $\mathcal{B}'$ is unbounded. By Lemma 3.5, there exists a nonsingular S and $1\le n_1<n$ such that

$$S^{-1}BS=\begin{pmatrix} B_{(2)} & * \\ 0 & B_{(1)}\end{pmatrix}$$

for all $B\in\mathcal{B}$, where $B_{(1)}\in M_{n_1}(\mathbb{C})$. Set

$$\mathcal{B}_{(i)}=\{B_{(i)}: B\in\mathcal{B}\},\quad i=1,2.$$

Then

$$\lim_{k\to\infty}\left[\max_{A\in\mathcal{B}_{(i)}^k}\rho(A)\right]=0,\quad i=1,2.$$

As the dimension associated with $\mathcal{B}_{(i)}$ is less than n, by an induction assumption,

$$\hat{\rho}(\mathcal{B}_{(i)})<1 \text{ for } i=1,2.$$

Since

$$\hat{\rho}(\mathcal{B})=\max\{\hat{\rho}(\mathcal{B}_{(1)}),\hat{\rho}(\mathcal{B}_{(2)})\},$$

we have $\hat{\rho}(\mathcal{B})\ne 1$, a contradiction. This completes the proof of $\hat{\rho}(\Sigma)<1$.

The following two lemmas are needed to establish Theorem 3.9.

LEMMA 3.7. *Let* $\Psi=\{\mathcal{A}^{(1)},\mathcal{A}^{(2)},\ldots,\mathcal{A}^{(N)}\}$ *be a set in* $M_n(\mathbb{IR})$. *If* $\lim_{k\to\infty}[\max_{A\in|\Psi|^k}\rho(A)]=0$, then for any sequence $\{\mathcal{A}_k\in\Psi: k\in\mathbb{N}\}$, $\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k=0$.

Proof. Since $\lim_{k\to\infty}[\max_{A\in|\Psi|^k}\rho(A)]=0$, by Theorem 3.6 there exists a norm $\|\cdot\|$ on $\mathbb{R}^n$ such that

$$\|\,|\mathcal{A}^{(i)}|\,\|<1 \quad \text{for } i=1,2,\ldots,N.$$

Let $\{\mathcal{A}_k\in\Psi: k\in\mathbb{N}\}$ be a sequence. Then

$$\begin{aligned}\|\,|\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k|\,\|_F &\le \|\,|\mathcal{A}_1||\mathcal{A}_2|\cdots|\mathcal{A}_k|\,\|_F\\ &\le \alpha\|\,|\mathcal{A}_1||\mathcal{A}_2|\cdots|\mathcal{A}_k|\,\|\\ &\le \alpha\left(\max_{1\le i\le N}\|\,|\mathcal{A}^{(i)}|\,\|\right)^k\to 0 \text{ as } k\to\infty,\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm and the α occurs because of a changing norm. Therefore $\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = 0$. This completes the proof.

LEMMA 3.8. *Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ be a set in $M_n(\mathbb{IR})$ with the generalized property $(*)$ in some j. For any sequence $\{\mathcal{A}_k \in \Psi : k \in \mathbb{N}\}$, if*

$$\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = 0,$$

then for the sequence $\{|\mathcal{A}_k| \in |\Psi| : k \in \mathbb{N}\}$, the jth row of $|\mathcal{A}_1||\mathcal{A}_2|\cdots|\mathcal{A}_k|$ converges to 0 as $k \to \infty$.

Proof. Consider the sequence $\{|\mathcal{A}_k| \in |\Psi| : k \in \mathbb{N}\}$. Since Ψ has the generalized property $(*)$ in j, there exists a $\mathcal{B} \in \Psi'$ with $\mathrm{d}(\mathcal{B}_{ij}) > 0$ for some i. Then for each $k = 1, 2, \ldots,$

$$\begin{aligned}\mathrm{d}[(\mathcal{B}\mathcal{A}_1\cdots\mathcal{A}_k)_{il}] &\geq [\mathrm{d}(\mathcal{B})|\mathcal{A}_1|\cdots|\mathcal{A}_k|]_{il}\\ &\geq \mathrm{d}(\mathcal{B})_{ij}[|\mathcal{A}_1|\cdots|\mathcal{A}_k|]_{jl}\\ &\geq 0 \quad \text{for all } l = 1, \ldots, n.\end{aligned}$$

Since $\mathcal{B}\mathcal{A}_1\cdots\mathcal{A}_k \to 0$ as $k \to \infty$, the jth row of $|\mathcal{A}_1|\cdots|\mathcal{A}_k|$ converges to 0 as $k \to \infty$. This completes the proof.

THEOREM 3.9. *Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ be a set in $M_n(\mathbb{IR})$ with the generalized property $(*)$ for all $1 \leq j \leq n$. Then $\lim_{k\to\infty}[\max_{A\in|\Psi|^k}\rho(A)] = 0$ if and only if for any sequence $\{\mathcal{A}_k \in \Psi : k \in \mathbb{N}\}$, $\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = 0$.*

Proof. The proof for the direction "⇒" follows from Lemma 3.7. The direction "⇐" follows from Lemma 3.2, Lemma 3.7, and Berger and Wang [2, Theorem I].

The following lemma is crucial in the proof of our main theorem (Theorem 3.11).

LEMMA 3.10. *Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ be a set in $M_n(\mathbb{IR})$. If there exists an index subset Λ of $\{1, 2, \ldots, n\}$ with $1 \leq \mathrm{card}(\Lambda) = k < n$ such that Ψ does not possess the generalized property $(*)$ in j for $j \in \Lambda$ and Ψ has the generalized property $(*)$ in j for $j \notin \Lambda$, then there exists a (common) permutation matrix P such that for each $i = 1, 2, \ldots, N$,*

$$P^{-1}\mathcal{A}^{(i)}P = \begin{pmatrix}\mathcal{A}_{11}^{(i)} & *\\ 0 & \mathcal{A}_{22}^{(i)}\end{pmatrix},$$

where $\mathcal{A}_{11}^{(i)}$ is a $k \times k$ real matrix. *Moreover, $P^{-1}\Psi P$ has the generalized property $(*)$ in j for $j = k+1, \ldots, n$.*

Proof. For all $\mathcal{A} \in \Psi$, we choose a common permutation matrix P such that $P^{-1}\Psi P$ has no generalized property $(*)$ in $j = 1, 2, \ldots, k$ and $P^{-1}\Psi P$ has the generalized property $(*)$ in $j = k+1, \ldots, n$.

Claim. $(P^{-1}\mathcal{A}P)_{rs} = 0$ for $k+1 \leq r \leq n$ and $1 \leq s \leq k$.

Assume $(P^{-1}\mathcal{A}P)_{rs} \neq 0$ for some $k+1 \leq r \leq n$ and some $1 \leq s \leq k$. Since $P^{-1}\Psi P$ has the generalized property $(*)$ in r, there exists a $\mathcal{B} \in \Psi'$ and $1 \leq t \leq n$ such that $\mathrm{d}(P^{-1}\mathcal{B}P)_{tr} > 0$. Hence

$$\mathrm{d}(P^{-1}\mathcal{B}\mathcal{A}P)_{ts} = \mathrm{d}[(P^{-1}\mathcal{B}P)(P^{-1}\mathcal{A}P)]_{ts} \geq \mathrm{d}(P^{-1}\mathcal{B}P)_{tr}|P^{-1}\mathcal{A}P|_{rs} > 0.$$

It follows that $P^{-1}\Psi P$ has the generalized property $(*)$ in s, a contradiction. Therefore, $(P^{-1}\mathcal{A}P)_{rs} = 0$ for $k+1 \leq r \leq n$ and $1 \leq s \leq k$. This completes the proof.

We are ready to present the main theorem in this paper, to which we now turn.

THEOREM 3.11. *Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ be a set in $M_n(\mathbb{IR})$. Then the following statements are mutually equivalent:*

(i) $\lim_{k\to\infty}[\max_{A\in\tilde{\Psi}^k}\rho(A)]=0$.
(ii) $\hat{\rho}(\tilde{\Psi})<1$.
(iii) *There exists a norm* $\|\cdot\|$ *on* $\mathbb{C}^n$ *such that* $\|\widetilde{\mathcal{A}^{(i)}}\|<1$ *for* $i=1,2,\ldots,N$.
(iv) *For any sequence* $\{\mathcal{A}_k\in\Psi : k\in\mathbb{N}\}$, $\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k=0$.

Proof. The proof of "(i) $\Leftrightarrow$ (ii)" follows from Lemma 3.2 and Theorem 3.6. The proof of "(ii) $\Leftrightarrow$ (iii)" follows from Theorem 3.6.

"(iii) $\Rightarrow$ (iv)" Consider the sequence $\{\mathcal{A}_k\in\Psi : k\in\mathbb{N}\}$. We proceed with the proof in the following three cases.

Case 1. We consider that Ψ has the generalized property $(*)$ in j for each $j=1,2,\ldots,n$. Then for each $i=1,2,\ldots,N$,

$$\widetilde{\mathcal{A}^{(i)}}=|\mathcal{A}^{(i)}|.$$

It follows that for each $k=1,2,\ldots,$

$$\begin{aligned}\|\,|\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k|\,\|_F &\le \|\,|\mathcal{A}_1||\mathcal{A}_2|\cdots|\mathcal{A}_k|\,\|_F\\ &\le \alpha\|\,|\mathcal{A}_1||\mathcal{A}_2|\cdots|\mathcal{A}_k|\,\|\\ &\le \alpha\|\tilde{\mathcal{A}}_1\|\|\tilde{\mathcal{A}}_2\|\cdots\|\tilde{\mathcal{A}}_k\|\\ &\le \alpha\left(\max_{1\le i\le N}\|\widetilde{\mathcal{A}^{(i)}}\|\right)^k\to 0 \text{ as } k\to\infty,\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm and the constant α occurs because of changing norm. Therefore $\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k=0$.

Case 2. We consider that Ψ has no generalized property $(*)$ in j for each $j=1,2,\ldots,n$. Then for each $i=1,2,\ldots,N$,

$$\widetilde{\mathcal{A}^{(i)}}=\mathcal{A}^{(i)}.$$

It follows that for each $k=1,2,\ldots,$

$$\|\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k\|\le\left(\max_{1\le i\le N}\|\widetilde{\mathcal{A}^{(i)}}\|\right)^k\to 0 \text{ as } k\to\infty.$$

Therefore $\lim_{k\to\infty}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k=0$.

Case 3. We assume the existence of index set $\Lambda\subset\{1,2,\ldots,n\}$ with $1\le\text{card}(\Lambda)<n$ such that Ψ has no generalized property $(*)$ in j for $j\in\Lambda$ and Ψ has the generalized property $(*)$ in j for $j\notin\Lambda$. By Lemma 3.10 there exists a permutation matrix P such that for each $i=1,2,\ldots,N$,

$$P^{-1}\mathcal{A}^{(i)}P=\mathcal{D}_{\mathcal{A}^{(i)}}+\mathcal{N}_{\mathcal{A}^{(i)}},$$

where

$$\mathcal{D}_{\mathcal{A}^{(i)}}=\begin{pmatrix}\mathcal{A}_{11}^{(i)} & 0\\ 0 & \mathcal{A}_{22}^{(i)}\end{pmatrix},\qquad \mathcal{N}_{\mathcal{A}^{(i)}}=\begin{pmatrix}0 & \mathcal{A}_{12}^{(i)}\\ 0 & 0\end{pmatrix},$$

$\mathcal{A}_{11}^{(i)}$ is a $k\times k$ real matrix, and $P^{-1}\Psi P$ has the generalized property $(*)$ in j for $j=k+1,\ldots,n$. Thus for each $i=1,2,\ldots,N$ we have

$$\begin{aligned}(\widetilde{P^{-1}\mathcal{A}^{(i)}P}) &= P^{-1}\widetilde{\mathcal{A}^{(i)}}P\\ &= \mathcal{D}_{\widetilde{\mathcal{A}^{(i)}}}+\mathcal{N}_{\widetilde{\mathcal{A}^{(i)}}},\end{aligned}$$

where

$$\mathcal{D}_{\widetilde{\mathcal{A}^{(i)}}} = \begin{pmatrix} \mathcal{A}_{11}^{(i)} & 0 \\ 0 & |\mathcal{A}_{22}^{(i)}| \end{pmatrix} \text{ and } \mathcal{N}_{\widetilde{\mathcal{A}^{(i)}}} = \begin{pmatrix} 0 & |\mathcal{A}_{12}^{(i)}| \\ 0 & 0 \end{pmatrix}.$$

We can define a new norm $\|\cdot\|_\alpha$ on $\mathbb{R}^n$ by $\|x\|_\alpha = \|Px\|$. Then $\|P^{-1}\widetilde{\mathcal{A}^{(i)}}P\|_\alpha < 1$ for $i = 1, 2, \ldots, N$. For $k = 1, 2, \ldots$, we have

$$\begin{aligned} P^{-1}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_kP &= P^{-1}\mathcal{A}_1PP^{-1}\mathcal{A}_2P\cdots P^{-1}\mathcal{A}_kP \\ &= [\mathcal{D}_{\mathcal{A}_1} + \mathcal{N}_{\mathcal{A}_1}][\mathcal{D}_{\mathcal{A}_2} + \mathcal{N}_{\mathcal{A}_2}]\cdots[\mathcal{D}_{\mathcal{A}_k} + \mathcal{N}_{\mathcal{A}_k}] \\ &= \mathcal{D}_{\mathcal{A}_1}\mathcal{D}_{\mathcal{A}_2}\cdots\mathcal{D}_{\mathcal{A}_k} + \cdots + \mathcal{N}_{\mathcal{A}_1}\mathcal{N}_{\mathcal{A}_2}\cdots\mathcal{N}_{\mathcal{A}_k}. \end{aligned}$$

Note that the terms which contain at least two nilpotent factors must be zero. Thus

$$P^{-1}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_kP = \mathcal{D}_{\mathcal{A}_1}\mathcal{D}_{\mathcal{A}_2}\cdots\mathcal{D}_{\mathcal{A}_k} + \Omega^1 + \cdots + \Omega^k,$$

where Ω^i denotes $\mathcal{D}_{\mathcal{A}_1}\cdots\mathcal{D}_{\mathcal{A}_{i-1}}\mathcal{N}_{\mathcal{A}_i}\mathcal{D}_{\mathcal{A}_{i+1}}\cdots\mathcal{D}_{\mathcal{A}_k}$ with $\mathcal{N}_{\mathcal{A}_i}$ appearing in the ith place of the product. Note that

$$\begin{aligned} &\| \, |P^{-1}\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_kP| \, \|_F \\ &\le \| \, |\mathcal{D}_{\mathcal{A}_1}\mathcal{D}_{\mathcal{A}_2}\cdots\mathcal{D}_{\mathcal{A}_k}| \, \|_F + \sum \| \, |\mathcal{D}_{\mathcal{A}_1}\cdots\mathcal{D}_{\mathcal{A}_{i-1}}| \, \|_F \| \, |\mathcal{N}_{\mathcal{A}_i}| \, \|_F \| \, |\mathcal{D}_{\mathcal{A}_{i+1}}\cdots\mathcal{D}_{\mathcal{A}_k}| \, \|_F \\ &\le \| \, |\mathcal{D}_{\tilde{\mathcal{A}}_1}\mathcal{D}_{\tilde{\mathcal{A}}_2}\cdots\mathcal{D}_{\tilde{\mathcal{A}}_k}| \, \|_F + \sum \| \, |\mathcal{D}_{\tilde{\mathcal{A}}_1}\cdots\mathcal{D}_{\tilde{\mathcal{A}}_{i-1}}| \, \|_F \| \, |\mathcal{N}_{\tilde{\mathcal{A}}_i}| \, \|_F \| \, |\mathcal{D}_{\tilde{\mathcal{A}}_{i+1}}\cdots\mathcal{D}_{\tilde{\mathcal{A}}_k}| \, \|_F \\ &\le \|(P^{-1}\tilde{\mathcal{A}}_1P)\cdots(P^{-1}\tilde{\mathcal{A}}_kP)\|_F \\ &\quad + \sum \|(P^{-1}\tilde{\mathcal{A}}_1P)\cdots(P^{-1}\tilde{\mathcal{A}}_{i-1}P)\|_F\|\mathcal{N}_{\tilde{\mathcal{A}}_i}\|_F\|(P^{-1}\tilde{\mathcal{A}}_{i+1}P)\cdots(P^{-1}\tilde{\mathcal{A}}_kP)\|_F \\ &\le c\|(P^{-1}\tilde{\mathcal{A}}_1P)\cdots(P^{-1}\tilde{\mathcal{A}}_kP)\|_\alpha \\ &\quad + c\sum \|(P^{-1}\tilde{\mathcal{A}}_1P)\cdots(P^{-1}\tilde{\mathcal{A}}_{i-1}P)\|_\alpha\|\mathcal{N}_{\tilde{\mathcal{A}}_i}\|_F\|(P^{-1}\tilde{\mathcal{A}}_{i+1}P)\cdots(P^{-1}\tilde{\mathcal{A}}_kP)\|_\alpha \\ &\le c\alpha^k + cMk\alpha^{k-1}, \end{aligned}$$

where $\alpha = \max_{1\le i\le N} ||P^{-1}\widetilde{\mathcal{A}^{(i)}}P||_\alpha < 1$, $M = \max_{1\le i\le N} \|\mathcal{N}_{\widetilde{\mathcal{A}^{(i)}}}\|_F$, and the constant c occurs because of a changing norm. Hence $\lim_{k\to\infty} \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = 0$.

The proof of the direction "(iv) ⇒ (iii)" involves three cases as well.

Case 1. Ψ has the generalized property $(*)$ in every $j = 1, 2, \ldots, n$. Then $\widetilde{\mathcal{A}^{(i)}} = |\mathcal{A}^{(i)}|$ for $i = 1, 2, \ldots, N$. By Theorems 3.6 and 3.9 there exists a norm $\|\cdot\|$ on $\mathbb{R}^n$ such that $\|\widetilde{\mathcal{A}^{(i)}}\| < 1$ for $i = 1, 2, \ldots, N$.

Case 2. Ψ has no generalized property $(*)$ in every $j = 1, 2, \ldots, n$. Then $\widetilde{\mathcal{A}^{(i)}} = \mathcal{A}^{(i)}$ for $i = 1, 2, \ldots, N$. By Berger and Wang [2, Theorem I] and Lemma 3.2 we have a norm $\|\cdot\|$ on $\mathbb{R}^n$ such that $\|\widetilde{\mathcal{A}^{(i)}}\| < 1$ for $i = 1, 2, \ldots, N$.

Case 3. We assume the existence of index set $\Lambda \subset \{1, 2, \ldots, n\}$ with $1 \le \text{card}(\Lambda) < n$ such that Ψ has no generalized property $(*)$ for $j \in \Lambda$ and Ψ has the generalized property $(*)$ for $j \notin \Lambda$. By Lemma 3.10 there exists a permutation matrix P such that for each $i = 1, 2, \ldots, N$,

$$P^{-1}\mathcal{A}^{(i)}P = \mathcal{D}_{\mathcal{A}^{(i)}} + \mathcal{N}_{\mathcal{A}^{(i)}},$$

where

$$\mathcal{D}_{\mathcal{A}^{(i)}} = \begin{pmatrix} \mathcal{A}_{11}^{(i)} & 0 \\ 0 & \mathcal{A}_{22}^{(i)} \end{pmatrix}, \qquad \mathcal{N}_{\mathcal{A}^{(i)}} = \begin{pmatrix} 0 & \mathcal{A}_{12}^{(i)} \\ 0 & 0 \end{pmatrix},$$

$\mathcal{A}_{11}^{(i)}$ is a $k \times k$ real matrix, and $P^{-1}\Psi P$ has the generalized property $(*)$ for $j = k+1, \ldots, n$. Thus for each $i = 1, 2, \ldots, N$ we have

$$\begin{aligned} \widetilde{(P^{-1}\mathcal{A}^{(i)}P)} &= P^{-1}\widetilde{\mathcal{A}^{(i)}}P \\ &= \mathcal{D}_{\widetilde{\mathcal{A}^{(i)}}} + \mathcal{N}_{\widetilde{\mathcal{A}^{(i)}}}, \end{aligned}$$

where

$$\mathcal{D}_{\widetilde{\mathcal{A}^{(i)}}} = \begin{pmatrix} \mathcal{A}_{11}^{(i)} & 0 \\ 0 & |\mathcal{A}_{22}^{(i)}| \end{pmatrix} \text{ and } \mathcal{N}_{\widetilde{\mathcal{A}^{(i)}}} = \begin{pmatrix} 0 & |\mathcal{A}_{12}^{(i)}| \\ 0 & 0 \end{pmatrix}.$$

Since

$$\lim_{k\to\infty} \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = \lim_{k\to\infty} P^{-1}\mathcal{A}_1PP^{-1}\mathcal{A}_2P\cdots P^{-1}\mathcal{A}_kP = 0,$$

we have

$$\lim_{k\to\infty} \mathcal{D}_{\mathcal{A}_1}\mathcal{D}_{\mathcal{A}_2}\cdots\mathcal{D}_{\mathcal{A}_k} = 0.$$

By Case 1 and Case 2 there exist norms $\|\cdot\|_1$ on $\mathbb{R}^{k\times k}$ and $\|\cdot\|_2$ on $\mathbb{R}^{(n-k)\times(n-k)}$, respectively, such that $\|\mathcal{A}_{11}^{(i)}\|_1 < 1$ and $\||\mathcal{A}_{22}^{(i)}|\|_2 < 1$ for $i = 1, 2, \ldots, N$. It follows that $\lim_{k\to\infty}[\max_{A\in\tilde{\Psi}^k}\rho(A)] = 0$. By Theorem 3.6 we have a norm $\|\cdot\|$ on $\mathbb{R}^n$ such that $\|\widetilde{\mathcal{A}^{(i)}}\| < 1$ for $i = 1, 2, \ldots, N$. This completes the proof.

4. Stability of Takagi–Sugeno free fuzzy systems with time-varying uncertainty. In this section, we first review the (deterministic) Takagi–Sugeno free fuzzy system [16]. Consider the following free fuzzy system. Let the system state vector at time instant k be $\bar{x}(k) = [x_1(k), \ldots, x_n(k)]^T$, where $x_1(k), \ldots, x_n(k)$ are state variables of the system at time instant k. Then the free fuzzy system is defined by the implications below: For $i = 1, 2, \ldots, N$ and $A_i \in R^{n\times n}$ we have the rule

RULE^i : IF ($x_1(k)$ is S_1^i, AND,..., AND $x_n(k)$ is S_n^i), THEN $\bar{x}(k+1) = A_i\bar{x}(k)$.

Note that S_j^i is the fuzzy set corresponding to x_j and the implication RULE^i. The A_i's are the system characteristic matrices. Let Σ denote the set of characteristic matrices, that is, $\Sigma = \{A_1, A_2, \ldots, A_N\}$. The truth value of RULE^i at time instant k is defined as

$$w_i(k) = \wedge\{\mu_{S_1^i}(x_1(k)), \ldots, \mu_{S_n^i}(x_n(k))\},$$

where $\wedge$ usually stands for the minimum operator and $\mu_S(x)$ is the value of membership function of the fuzzy set S at position x. Then the state vector at time instant $k+1$ is updated by

$$\bar{x}(k+1) = \left[\sum_{i=1}^{N} \alpha_i(k)A_i\right]\bar{x}(k), \tag{1}$$

where $\alpha_i(k) = w_i(k)/\sum_{i=1}^{N} w_i(k)$.

Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$. The following mathematical model of the Takagi–Sugeno free fuzzy system with time-varying uncertainty was employed by Pang and Guu [11]. They assumed that the consequent parameters are subjected to time-varying uncertainty. In other words, the characteristic matrices employed in each rule are varying due to uncertainty for each time epoch k. Precisely, RULE^i becomes

(2) IF ($x_1(k)$ is S_1^i, AND,..., AND $x_n(k)$ is S_n^i), THEN $\bar{x}(k+1) = [A_i(k)]\bar{x}(k)$,

where $A_i(k) \in \mathcal{A}^{(i)}$. The states of the system are updated by

$$\bar{x}(k+1) = \left[\sum_{i=1}^{N} \alpha_i(k) A_i(k)\right] \bar{x}(k), \tag{3}$$

where $\alpha_i(k)$s are defined as in the deterministic case.

DEFINITION 4.1. *The Takagi–Sugeno free fuzzy system with time-varying uncertainty is* globally asymptotically stable *if*

$$\bar{x}(k) \to 0 \text{ as } k \to \infty \tag{4}$$

for any initial values $\bar{x}(0) \in R^n$.

THEOREM 4.2. *If there exists a norm* $||\cdot||$ *on* $\mathbb{R}^n$ *such that* $\|\widetilde{\mathcal{A}^{(i)}}\| < 1$ *for all* $i = 1, 2, \ldots, N$, *then the Takagi–Sugeno free fuzzy system with time-varying uncertainty is globally asymptotically stable.*

Proof. For $k = 0, 1, 2, \ldots$, since $A_i(k) \in \mathcal{A}^{(i)}$ for $i = 1, 2, \ldots, N$, we have $A_i(k) \leq \widetilde{\mathcal{A}^{(i)}}$ for all $i = 1, 2, \ldots, N$. Then for any initial $\bar{x}(0) \in R^n$,

$$\begin{aligned}
||\bar{x}(k+1)||_F &= \left|\left|\left(\sum_{i=1}^{N} \alpha_i(k) A_i(k)\right) \bar{x}(0)\right|\right|_F \\
&\leq \left|\left|\left(\sum_{i=1}^{N} \alpha_i(k) A_i(k)\right)\right|\right|_F ||\bar{x}(0)||_F \\
&\leq \left|\left|\left(\sum_{i=1}^{N} \alpha_i(k) \widetilde{\mathcal{A}^{(i)}}\right)\right|\right|_F ||\bar{x}(0)||_F \\
&\leq c\left|\left|\left(\sum_{i=1}^{N} \alpha_i(k) \widetilde{\mathcal{A}^{(i)}}\right)\right|\right| ||\bar{x}(0)||_F \\
&\leq c\left(\max_{1\leq i\leq N} ||\widetilde{\mathcal{A}^{(i)}}||\right)^{k+1} ||\bar{x}(0)||_F \to 0 \text{ as } k \to \infty.
\end{aligned}$$

This shows that the Takagi–Sugeno free fuzzy system with time-varying uncertainty is globally asymptotically stable.

COROLLARY 4.3. *The Takagi–Sugeno free fuzzy system with time-varying uncertainty is globally asymptotically stable if* $\lim_{k\to\infty}[\max_{A\in\tilde{\Psi}^k} \rho(A)] = 0$.

Proof. Corollary 4.3 follows from Theorems 3.11 and 4.2.

COROLLARY 4.4. *The Takagi–Sugeno free fuzzy system with time-varying uncertainty is globally asymptotically stable if for any sequence* $\{\mathcal{A}_k \in \Psi : k \in \mathbb{N}\}$, $\lim_{k\to\infty} \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = 0$.

Proof. Corollary 4.4 follows from Theorems 3.11 and 4.2.

COROLLARY 4.5. *The Takagi–Sugeno free fuzzy system with time-varying uncertainty is globally asymptotically stable if* $\hat{\rho}(\tilde{\Psi}) < 1$.

Proof. Corollary 4.5 follows from Theorems 3.11 and 4.2.

5. Conclusion. In the literature, a condition for the consecutive powers of an interval matrix to converge to the zero matrix has been characterized by Mayer. In the current paper, we extended Mayer's convergence theorem to the case with a finite number of interval matrices. Precisely, we studied the conditions for the infinite products of a finite number of interval matrices to converge to the null matrix. Let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ be a finite set of interval matrices in $M_n(\mathbb{IR})$. We proved that the following statements are mutually equivalent:

(i) $\lim_{k\to\infty}[\max_{A\in\tilde{\Psi}^k} \rho(A)] = 0$.

(ii) $\hat{\rho}(\tilde{\Psi}) < 1$.

(iii) There exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ such that $\|\widetilde{\mathcal{A}^{(i)}}\| < 1$ for $i = 1, 2, \ldots, N$.

(iv) For any sequence $\{\mathcal{A}_k \in \Psi : k \in \mathbb{N}\}$, $\lim_{k\to\infty} \mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_k = 0$.

The motivation for the case with a finite number of interval matrices came from our stability study of Takagi–Sugeno free fuzzy systems with time-varying uncertainty. As an application, let $\Psi = \{\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}\}$ denote the set of interval matrices from which at each iteration the characteristic matrices are generated. We then showed that the Takagi–Sugeno free fuzzy system with time-varying uncertainty is globally asymptotically stable if there exists a norm $||\cdot||$ on $\mathbb{R}^n$ such that for $i = 1, 2, \ldots, N$,

$$\|\widetilde{\mathcal{A}^{(i)}}\| < 1,$$

where the $\widetilde{\mathcal{A}^{(i)}}$s are real matrices constructed from Ψ by the extended property $(*)$.

REFERENCES

[1] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] M. A. BERGER AND Y. WANG, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[3] F. F. BONSALL AND J. DUNCAN, *Numerical Ranges of Operators on Normed Spaces and Elements of Normed Algebras*, London Mathematical Society Lecture Note Series 2, Cambridge University Press, London, New York, 1971.

[4] I. DAUBECHIES AND J. C. LAGARIAS, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[5] L. ELSNER, *The generalized spectral-radius theorem: An analytic-geometric proof*, Linear Algebra Appl., 220 (1995), pp. 151–159.

[6] S.-M. GUU AND C.-T. PANG, *On the asymptotic stability of free fuzzy systems*, IEEE Trans. Fuzzy Systems, 7 (1999), pp. 467–468.

[7] J. JOH, Y.-H. CHEN, AND R. LANGARI, *On the stability issues of linear Takagi-Sugeno fuzzy models*, IEEE Trans. Systems, 6 (1998), pp. 402–410.

[8] D. KÖNIG, *Theory of Finite and Infinite Graphs*, Birkhäuser Boston, Boston, 1990.

[9] G. MAYER, *On the convergence of powers of interval matrices*, Linear Algebra Appl., 58 (1984), pp. 201–216.

[10] C.-T. PANG, *Asymptotic Stability of Interval Matrices and Interval Systems*, Ph.D. thesis, Department of Mathematics, National Central University, Taiwan, R.O.C., 1997.

[11] C.-T. PANG AND S.-M. GUU, *Sufficient conditions for the stability of linear Takagi-Sugeno free fuzzy systems*, IEEE Trans. Fuzzy Systems, 11 (2003), pp. 695–700.

[12] C.-T. PANG, Y.-Y. LUR, AND S.-M. GUU, *A new proof of Mayer's theorem*, Linear Algebra Appl., 350 (2002), pp. 273–278.

[13] G. C. Rota and W. G. Strang, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.

[14] M.-H. Shih, Y.-Y. Lur, and C.-T. Pang, *Simultaneous Schur Stability of Interval Matrices*, working paper, Department of Information Management, Yuan Ze University, Taiwan, R.O.C., 2002.

[15] M.-H. Shih, J.-W. Wu, and C.-T. Pang, *Asymptotic stability and generalized Gelfand spectral formula*, Linear Algebra Appl., 252 (1997), pp. 61–70.

[16] T. Takagi and M. Sugeno, *Fuzzy identification of systems and its applications to modelling and control*, IEEE Trans. Systems, Man. Cybernet., 15 (1985), pp. 116–132.

[17] M. A. L. Thathachar and P. Viswanath, *On the stability of fuzzy systems*, IEEE Trans. Fuzzy Systems, 5 (1997), pp. 145–151.

[18] H. Ying, *General SISO Takagi-Sugeno fuzzy systems with linear rule consequent are universal approximators*, IEEE Trans. Fuzzy Systems, 6 (1998), pp. 582–587.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 752–765

THE EQUALITY CASES FOR THE INEQUALITIES OF FISCHER, OPPENHEIM, AND ANDO FOR GENERAL M-MATRICES*

XIAO-DONG ZHANG†

Abstract. In this paper, we study the zero pattern structure of two general M-matrices whose Fan product is singular. Then, for general M-matrices, necessary and sufficient conditions are obtained for equality in the inequalities of Fischer, Oppenheim, and Ando.

Key words. Fischer's inequality, Oppenheim's inequality, Ando's inequality, general M-matrix, Fan product

AMS subject classifications. 15A45, 15A57

DOI. 10.1137/S0895479802410086

1. Introduction.

A square $n \times n$ matrix A is called a *general M-matrix* if A can be expressed in the form $A = sI - P$ with $P \geq 0$, where $s \geq \rho(P)$, the spectral radius of nonnegative matrix P. Thus general M-matrices consist of nonsingular M-matrices and singular M-matrices. General M-matrices arise in investigations concerning the convergence of iterative processes for systems of linear or nonlinear equations and in the study of nonnegative solutions to such systems. These investigations have a variety of applications to problems in economics and linear programming. An extensive list of references to studies of general M-matrices may be found in [2] and [3]. In a series of papers [5], [6], [7], etc., Fan established remarkable determinant inequalities as well as some matrix inequalities for nonsingular M-matrices. In particular, Fan in [5] and [6] (see also Ando [1]) proved Fischer's inequality for nonsingular M-matrices and Oppenheim's inequality for the Fan product of two nonsingular M-matrices (in their papers, "M-matrices" means "nonsingular M-matrices"). For two general M-matrices $A = (a_{ij})$ and $B = (b_{ij})$, the *Fan product* of A and B, denoted by $A \circ B$, is the matrix $C = (c_{ij})$, where $c_{ii} = a_{ii}b_{ii}$ for all i and $c_{ij} = -a_{ij}b_{ij}$ for $i \neq j$. Moreover, throughout this paper, if an $n \times n$ matrix A is partitioned into the form $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$, then we always assume that A_{11} is square. Hence we get the following theorem.

THEOREM 1.1. *Let $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ and $B = (b_{ij})$ be two $n \times n$ nonsingular M-matrices. Then the inequalities of Fischer and Oppenheim hold. That is,*

$$\det A \leq \det A_{11} \det A_{22} \tag{1}$$

and

$$\left(\prod_{i=1}^{n} b_{ii}\right) \det A \leq \det(A \circ B). \tag{2}$$

Further, Ando in [1] proved the following inequality, which improved Oppenheim's inequality and is called *Ando's inequality.*

*Received by the editors June 24, 2002; accepted for publication (in revised form) by R. Nabben July 15, 2003; published electronically January 30, 2004. This research was supported by the National Natural Science Foundation of China (10371075) and sponsored by SRF for ROCS, SEM.
http://www.siam.org/journals/simax/25-3/41008.html

†Department of Mathematics, Shanghai Jiao Tong University, 1954 Huashan road, Shanghai, 200030, People's Republic of China (xiaodong@sjtu.edu.cn).

THEOREM 1.2. *Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $n \times n$ nonsingular M-matrices. Then*

$$\left(\prod_{i=1}^{n} b_{ii}\right) \det A + \left(\prod_{i=1}^{n} a_{ii}\right) \det B - \det A \det B \leq \det(A \circ B). \tag{3}$$

Recently, Lee in [10] and [11] extended Oppenheim's inequality for irreducible general M-matrices. Later, Smith in [14] showed that Oppenheim's inequality holds for the Fan product of any two general M-matrices. Further, he also proved Ando's inequality for any two general M-matrices. For studies of inequalities and related determinant inequalities for general M-matrices, the reader is referred to [1], [4], [8], [12], and [9].

This paper is organized as follows: In section 2, we investigate the zero pattern structure of two general M-matrices whose Fan product is singular. In section 3, for nonsingular M-matrices, necessary and sufficient conditions are obtained for equality in Fischer's inequality. These results, in sections 4 and 5, are applied to describe, for general M-matrices, necessary and sufficient conditions for the equality in the inequalities of Oppenheim and Ando, respectively.

2. Fan product of two general M-matrices. An $n \times n$ square matrix $A = (a_{ij})$ is called *row diagonally dominant* if

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}| \quad \text{for} \quad i = 1, \ldots, n.$$

An $n \times n$ general M-matrix A is called *cyclic* if

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & \cdots & 0 & 0 \\ 0 & a_{22} & a_{23} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ a_{n1} & 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}, \tag{4}$$

where $a_{ii}a_{i,i+1} \neq 0$ for $i = 1, \ldots, n, (n+1 \equiv 1)$. It is easy to see that $\det A = a_{11} \cdots a_{nn} - (-1)^n a_{12} \cdots a_{n-1,n} a_{n1}$ for $n > 1$. Hence, a cyclic matrix A with $n > 1$ is singular if and only if $a_{11} \cdots a_{nn} = (-1)^n a_{12} \cdots a_{n-1,n} a_{n1}$. In particular, a 1×1 cyclic matrix is nonsingular. Moreover, a cyclic matrix is irreducible. In order to obtain the zero pattern structure of two general M-matrices whose Fan product is singular, we need some lemmas.

LEMMA 2.1. *Let A and B be two $n \times n$ row diagonally dominant general M-matrices. Then $A \circ B$ is an irreducible singular M-matrix if and only if there exists a permutation matrix P such that PAP^T and PBP^T are singular and cyclic.*

Proof. Assume that $A \circ B$ is an irreducible singular M-matrix. Since A and B are row diagonally dominant general M-matrices, we have

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|, \quad b_{ii} \geq \sum_{j \neq i} |b_{ij}|, \quad i = 1, \ldots, n. \tag{5}$$

Then

$$a_{ii}b_{ii} \geq \sum_{j \neq i} |a_{ij}| \sum_{j \neq i} |b_{ij}| \geq \sum_{j \neq i} |a_{ij}b_{ij}|, \quad i = 1, \ldots, n. \tag{6}$$

Hence $A \circ B$ is an irreducible, row diagonally dominant general M-matrix. Since $A \circ B$ is irreducible and singular, by Taussky's theorem of [15], equality holds in (6) for $i = 1, \ldots, n$ and $a_{ii}b_{ii} > 0$. Therefore

$$a_{ii} = \sum_{j \neq i} |a_{ij}|, \quad b_{ii} = \sum_{j \neq i} |b_{ij}|, \quad i = 1, \ldots, n, \tag{7}$$

$$|a_{ij}| \sum_{k \neq i,j} |b_{ik}| = 0, \quad |b_{ij}| \sum_{k \neq i,j} |a_{ik}| = 0, \quad j \neq i. \tag{8}$$

It follows from (7) and (8) that each row of A and B has only one nonzero off-diagonal entry, respectively. Moreover, $A \circ B$ is irreducible implies that A and B are irreducible. Hence every row and column of A and B must contain only one nonzero off-diagonal entry, respectively. Further, for $i \neq j$, the (i, j) entry of A is nonzero if and only if the (i, j) entry of B is nonzero. Therefore, there exists a permutation matrix P such that PAP^T and PBP^T are singular and cyclic.

Conversely, without loss of generality, we may assume that $A = (a_{ij})$ and $B = (b_{ij})$ are singular and cyclic, since $P(A \circ B)P^T = (PAP^T) \circ (PBP^T)$ by Lemma 3.1 of [14]. Hence $A \circ B$ is cyclic. Moreover, $\prod_{i=1}^n a_{ii} = (-1)^n \prod_{i=1}^n a_{i,i+1}$ and $\prod_{i=1}^n b_{ii} = (-1)^n \prod_{i=1}^n b_{i,i+1}$. Then $\det(A \circ B) = \prod_{i=1}^n (a_{ii}b_{ii}) - \prod_{i=1}^n (a_{i,i+1}b_{i,i+1}) = 0$. We conclude that $A \circ B$ is an irreducible singular M-matrix. □

LEMMA 2.2. *Let A and B be two $n \times n$ general M-matrices. Then $A \circ B$ is an irreducible singular M-matrix if and only if there exists a permutation matrix P such that both PAP^T and PBP^T are singular and cyclic.*

Proof. We assume that $A \circ B$ is an irreducible singular M-matrix. Then A and B are irreducible M-matrices. By Theorem 6.4.16 of [3], there exist two positive diagonal matrices D and E such that AD and BE are row diagonally dominant general M-matrices. Hence $(AD) \circ (BE) = (A \circ B)DE$ is an irreducible singular M-matrix. By Lemma 2.1, there exists a permutation matrix P such that $PADP^T$ and $PBEP^T$ are singular and cyclic. Then $PAP^T = (PADP^T)(PD^{-1}P^T)$ and $PBP^T = (PBEP^T)(PE^{-1}P^T)$ are singular and cyclic, since $PD^{-1}P^T$ and $PE^{-1}P^T$ are positive diagonal matrices.

The converse follows as in the proof of Lemma 2.1. □

COROLLARY 2.3. *Let A and B be two $n \times n$ general M-matrices. If $A \circ B$ is an irreducible general M-matrix, then $a_{ii}b_{ii} > 0$ for $i = 1, \ldots, n$.*

Proof. If $A \circ B$ is singular, it follows from Lemma 2.2 that $a_{ii}b_{ii} > 0$ for $i = 1, \ldots, n$. If $A \circ B$ is nonsingular, then by Theorem 6.2.3 of [3], $a_{ii}b_{ii} > 0$ for $i = 1, \ldots, n$. □

We are ready to present the main result in this section after recalling the following notation. If $A = (a_{ij})$ is an $n \times n$ matrix and σ is a permutation on n objects, then the n-tuple $(a_{1,\sigma(1)}, a_{2,\sigma(2)}, \ldots, a_{n,\sigma(n)})$ is called a *diagonal* of A. In particular, $(a_{11}, a_{22}, \ldots, a_{nn})$ is called the *main diagonal* of A.

THEOREM 2.4. *Let A and B be two $n \times n$ general M-matrices. Then $A \circ B$ is a singular M-matrix if and only if one of the following conditions holds:*

(i) *$a_{ii} = 0$ for some i with $1 \leq i \leq n$.*

(ii) *$b_{ii} = 0$ for some i with $1 \leq i \leq n$.*

(iii) *There exists a permutation matrix P such that*

$$PAP^T = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where both A_{11} and B_{11} are singular and cyclic; and for every diagonal with an entry in the $(1,2)$ block, either the diagonal contains a zero for PAP^T or it contains a zero for PBP^T.

Proof. If $a_{ii} = 0$ or $b_{ii} = 0$ for some $1 \le i \le n$, then the (i,i) entry of $A \circ B$ is zero. By Theorem 3.2 of [14] and Hadamard's inequality for general M-matrices, $0 \le \det(A \circ B) \le a_{11}b_{11} \cdots a_{nn}b_{nn} = 0$. Hence $A \circ B$ is a singular M-matrix. If (iii) holds, it follows from Lemma 2.2 that $\det(A_{11} \circ B_{11}) = 0$. Hence by Fischer's inequality (1), $0 \le \det(A \circ B) = \det(PAP^T) \circ (PBP^T) \le \det(A_{11} \circ B_{11}) \det(A_{22} \circ B_{22}) = 0$. So $A \circ B$ is a singular M-matrix.

Conversely, we assume that $A \circ B$ is a singular M-matrix and $a_{ii}b_{ii} > 0$ for $i = 1, \ldots, n$. Clearly, there exists a permutation matrix P_1 such that

$$P_1(A \circ B)P_1^T = (P_1AP_1^T) \circ (P_1BP_1^T) = \begin{pmatrix} C_{11} \circ D_{11} & C_{12} \circ D_{12} & \cdots & C_{1k} \circ D_{1k} \\ 0 & C_{22} \circ D_{22} & \cdots & C_{2k} \circ D_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & C_{kk} \circ D_{kk} \end{pmatrix},$$

where $C_{ii} \circ D_{ii}$ is an $m_i \times m_i$ irreducible general M-matrix for $i = 1, \ldots, k$. Moreover, there exists some l with $1 \le l \le k$ such that $C_{ll} \circ D_{ll}$ is a singular M-matrix. By Lemma 2.2, there exists an $m_l \times m_l$ permutation matrix Q_l such that $Q_lC_{ll}Q_l^T = \widetilde{C_{ll}}$ and $Q_lD_{ll}Q_l^T = \widetilde{D_{ll}}$ are singular and cyclic. Let $P_2 = \operatorname{diag}(I_1, \ldots I_{l-1}, Q_l, I_{l+1}, \ldots, I_k)$ and let P_3 be the matrix obtained from $I = \operatorname{diag}(I_1, \ldots I_k)$ by interchanging the first block row and the lth block row of I, where I_i is the $m_i \times m_i$ identity matrix for $i = 1, 2, \ldots, k$. Let $P = P_3P_2P_1$. Then $P(A \circ B)P^T =$

$$\begin{pmatrix} F_{ll} & 0 & 0 & \cdots & 0 & 0 & F_{l,l+1} & \cdots & F_{lk} \\ F_{2l} & F_{22} & F_{23} & \cdots & F_{2,l-1} & 0 & F_{2,l+1} & \cdots & F_{2k} \\ F_{3l} & 0 & F_{33} & \cdots & F_{3,l-1} & 0 & F_{3,l+1} & \cdots & F_{3k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ F_{l-1,l} & 0 & 0 & \cdots & F_{l-1,l-1} & 0 & F_{l-1,l+1} & \cdots & F_{l-1,k} \\ F_{1l} & F_{12} & F_{13} & \cdots & F_{1,l+1} & F_{11} & F_{1,l+1} & \cdots & F_{1k} \\ 0 & 0 & 0 & \cdots & 0 & 0 & F_{l+1,l+1} & \cdots & F_{l+1,k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & F_{kk} \end{pmatrix} \equiv \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix},$$

where $F_{ij} = C_{ij} \circ D_{ij}$ for $1 \le i \le j \le n$ with $i \ne l$ and $j \ne l$, $F_{il} = \widetilde{C_{il}} \circ \widetilde{D_{il}}$ for $1 \le i \le l$, and $F_{lj} = \widetilde{C_{lj}} \circ \widetilde{D_{lj}}$ for $l \le j \le n$ and where $H_{11} = F_{ll}$. Then let $PAP^T = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ and $PBP^T = \left(\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right)$ be partitioned conformably with H. Clearly, both A_{11} and B_{11} are singular and cyclic. Moreover, by the definition of the determinant, for every diagonal with an entry in the $(1,2)$ block, either the diagonal contains a zero for PAP^T or it contains a zero for PBP^T. ∎

COROLLARY 2.5. *Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $n \times n$ general M-matrices with $a_{ii}b_{ii} > 0$ for $i = 1, \ldots, n$. If A or B is nonsingular, then $A \circ B$ is a nonsingular M-matrix.*

Proof. If $A \circ B$ is singular, then by Theorem 2.4 there exists a permutation matrix P such that $PAP^T = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ and $PBP^T = \left(\begin{smallmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{smallmatrix}\right)$, where both A_{11} and B_{11} are singular and cyclic. Hence by Fischer's inequality (1), $0 \le \det A \le \det A_{11} \det A_{22} = 0$ and $0 \le \det B \le \det B_{11} \det B_{22} = 0$, a contradiction. ∎

Remark. Fan in [6] proved that the Fan product of two nonsingular M-matrices is a nonsingular M-matrix. Lee in [10] presented an extension of Fan's result; that is,

if A is irreducible and B is nonsingular M-matrix, then $A \circ B$ is nonsingular. Clearly, Corollary 2.5 is an extension of Fan's and Lee's results.

3. Equality for Fischer's inequality. In this section, in order to obtain sufficient and necessary conditions for equality in the inequality of Fischer for nonsingular M-matrices, we need the following notation and lemma. Let $X = (x_{ij})$ be an $n \times n$ matrix. Then the *permanent* of X is defined by $perX = \sum x_{1,j_1} \cdots x_{n,j_n}$, where the summation is taken over all the permutations $(j_1 \cdots j_n)$ of the integers $1, \ldots, n$. Let $|X| = (|x_{ij}|)$ be the nonnegative matrix whose entries are given by $|x_{ij}|$. It follows from Theorem 2.6 of [13] that we have the following

LEMMA 3.1. *Let $A = (a_{ij})$ be an $n \times n$ nonsingular M-matrix. If there exists a nonzero sequence $a_{i_1,i_2} \neq 0, a_{i_2,i_3} \neq 0, \ldots, a_{i_{k-1},i_k} \neq 0$, then the (i_1, i_k) entry of A^{-1} is positive.*

We are ready to present the main result in this section.

THEOREM 3.2. *Let $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ be an $n \times n$ nonsingular M-matrix, where A_{11} is an $m \times m$ leading principal submatrix of A. Then the following statements are equivalent:*

(i) *Equality holds in Fischer's inequality; in other words,* $\det A = \det A_{11} \det A_{22}$.

(ii) *Every diagonal of A with at least one entry in A_{12} contains a zero.*

(iii) $per|A| = per|A_{11}|per|A_{22}|$.

Proof. By the definitions of the determinant and permanent, it is easy to see that (ii) and (iii) are equivalent and (ii) $\Rightarrow$ (i) holds.

If (i) holds, then A_{11} and B_{11} are nonsingular. We prove that assertion (ii) holds by induction on n. It is trivial when $n = 2$. Assume that the assertion holds for all positive integers less than n. We proceed to show that the assertion holds for any $n \times n$ nonsingular M-matrices.

Case 1. A_{11} is a 1×1 matrix. By Schur's formula, we have $\det A_{11} \det A_{22} = \det A = \det A_{22} \det(A_{11} - A_{12}A_{22}^{-1}A_{21})$. Hence, $A_{12}A_{22}^{-1}A_{21} = 0$. Suppose that there exists a diagonal $(a_{1,j_1}, a_{2,j_2}, \ldots, a_{n,j_n})$ of A such that it contains no zeros for some t with $j_1 \neq 1$ and $j_t = 1$. Then A_{22} is a nonsingular M-matrix and has a nonzero sequence $a_{j_1,i_2} \neq 0, a_{i_2,i_3} \neq 0, \ldots, a_{i_k,t} \neq 0$ with $i_2 = j_{j_1}$, $i_3 = j_{i_2}, \ldots$, and $t = j_{i_k}$. By Lemma 3.1, the (j_1, t) entry of A_{22}^{-1} is positive. So $a_{1,j_1}(A_{22}^{-1})_{j_1,t}a_{t,1} > 0$, which implies $A_{12}A_{22}^{-1}A_{21} \neq 0$, a contradiction. Hence every diagonal of A with at least one entry in A_{12} contains a zero.

Case 2. A_{11} is an $m \times m$ leading principal submatrix of A with $1 < m < n$. Let $A/(a_{11})$ be the Schur complement of (a_{11}) in A. By Lemma 1 of [5] (see also [3]),

$$A/(a_{11}) = A(1) - \frac{1}{a_{11}}(a_{21}, \ldots, a_{n1})^t(a_{12}, \ldots, a_{1n}) =: B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{9}$$

is the $(n-1) \times (n-1)$ nonsingular M-matrix, where $A(1)$ is the $(n-1) \times (n-1)$ matrix obtained from A by deleting the first row and column of A, and B is partitioned conformably with $A(1)$; i.e., B_{11} and B_{22} are $(m-1) \times (m-1)$ and $(n-m) \times (n-m)$ matrices, respectively. Moreover, $B_{22} \leq A_{22}$ implies $\det B_{22} \leq \det A_{22}$ (see [1, Corollary 4.11] or [3]). By Schur's formula and Fischer's inequality,

$$\begin{aligned} \det A &= a_{11} \det(A/(a_{11})) = a_{11} \det B \\ &\leq a_{11} \det B_{11} \det B_{22} \leq a_{11} \det B_{11} \det A_{22} \\ &= \det A_{11} \det A_{22} = \det A. \end{aligned}$$

We conclude that $\det B = \det B_{11} \det B_{22}$ and $\det A_{22} = \det B_{22}$. By the induction hypothesis, every diagonal of B with at least one entry in B_{12} contains a zero. Now we consider every diagonal $(a_{1,j_1}, \dots, a_{n,j_n})$ of A with at least one entry in A_{12}.

If $j_1 = 1$, then $a_{i,j_i} - \frac{a_{i1}a_{1,j_i}}{a_{11}} = b_{i,j_i}$ for $i = 2, \dots, n$, and $(b_{2,j_i}, \dots, b_{n,j_n})$ is a diagonal of B with at least one entry in B_{12}. By the induction hypothesis, there exists a t with $2 \le t \le n$ such that $b_{t,j_t} = a_{t,j_t} - \frac{a_{t1}a_{1,j_t}}{a_{11}} = 0$, which implies $a_{t,j_t} = 0$. Hence the diagonal $(a_{1,j_1}, \dots, a_{n,j_n})$ of A contains a zero.

If $2 \le j_1 \le n$, then there exists a k with $2 \le k \le n$ such that $j_k = 1$. Clearly, $a_{i,j_i} - \frac{a_{i1}a_{1,j_i}}{a_{11}} = b_{i,j_i}$ for $i = 2, \dots, k-1, k+1, \dots, n$, and $a_{k,j_1} - \frac{a_{k1}a_{1,j_1}}{a_{11}} = b_{k,j_1}$. Then $(b_{2,j_2}, \dots, b_{k-1,j_{k-1}}, b_{k,j_1}, b_{k+1,j_{k+1}}, \dots, b_{n,j_n})$ is a diagonal of B. We consider the following two subcases.

Subcase 1. If the diagonal $(b_{2,j_2}, \dots, b_{k-1,j_{k-1}}, b_{k,j_1}, b_{k+1,j_{k+1}}, \dots, b_{n,j_n})$ of B contains at least one entry in B_{12}, then by the induction hypothesis, either there exists a t with $2 \le t \le n$ and $t \ne k$ such that $b_{t,j_t} = a_{t,j_t} - \frac{a_{t1}a_{1,j_t}}{a_{11}} = 0$, which implies $a_{t,j_t} = 0$; or $b_{k,j_1} = 0$, which implies $a_{k,j_k} = a_{k1} = 0$ or $a_{1,j_1} = 0$. Hence the diagonal $(a_{1,j_1}, \dots, a_{n,j_n})$ of A contains a zero.

Subcase 2. If the diagonal $(b_{2,j_2}, \dots, b_{k-1,j_{k-1}}, b_{k,j_1}, b_{k+1,j_{k+1}}, \dots, b_{n,j_n})$ of B contains no entries in B_{12}, then it is easy to see that $m+1 \le j_1 \le n$ and $m+1 \le k \le n$. Hence $(b_{2,j_2}, \dots, b_{m,j_m})$ and $(b_{m+1,j_{m+1}}, \dots, b_{k,j_1}, \dots, b_{n,j_n})$ are diagonals of B_{11} and B_{22}, respectively. Let $C = \left(\begin{smallmatrix} a_{11} & \alpha^T \\ \beta & A_{22} \end{smallmatrix}\right)$, where $\alpha = (a_{1,m+1}, \dots, a_{1n})^T$ and $\beta = (a_{m+1,1}, \dots, a_{n1})^T$. Then C is the $(n-m+1) \times (n-m+1)$ nonsingular principal submatrix of A. Hence it follows from (9) that $\det C = a_{11} \det(C/(a_{11})) = a_{11} \det B_{22} = a_{11} \det A_{22}$. By the induction hypothesis, the diagonal $(a_{1,j_1}, a_{m+1,j_{m+1}}, \dots, a_{k,j_k}, \dots, a_{n,j_n})$ of C with at least one entry of α^T contains a zero. Hence the diagonal $(a_{1,j_1}, \dots, a_{m,j_m}, a_{m+1,j_{m+1}}, \dots, a_{k,j_k}, \dots, a_{n,j_n})$ of A contains a zero. □

Remark. If A is a singular M-matrix, then, in general, Theorem 3.2 does not hold. For example, let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A_{11} = \begin{pmatrix} 2 & -2 & -1 \\ -2 & 2 & -1 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_{12} = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, \quad A_{21} = (0, 0, -1),$$

and $A_{22} = (5)$. Clearly, $\det A = \det A_{11} \det A_{22} = 0$. But the diagonal $(a_{11}, a_{22}, a_{34}, a_{43})$ of A with an entry in A_{12} contains no zeros.

COROLLARY 3.3. *Let*

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \cdots & \cdots & \cdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix}$$

be a nonsingular M-matrix, where A_{ii} is square for $i = 1, \dots, k$. Then the following statements are equivalent:

(i) $\det A = \det A_{11} \cdots \det A_{kk}$.

(ii) *Every diagonal of A with at least one entry in A_{ij} for some $i \ne j$ contains a zero.*

(iii) $per|A| = per|A_{11}| \cdots per|A_{kk}|$.

Proof. It follows from induction and Theorem 3.2 that the assertion holds. □

It follows from Theorem 3.2 that necessary and sufficient conditions are obtained for equality in Hadamard's inequality.

THEOREM 3.4. *Let $A = (a_{ij})$ be an $n \times n$ nonsingular M-matrix. Then the following statements are equivalent:*

(i) *Equality holds in Hadamard's inequality, i.e.,* $\det A = a_{11} \cdots a_{nn}$.

(ii) *Every diagonal, except the main diagonal, of every principal submatrix of* A *contains a zero.*

(iii) *Every* $m \times m$ *principal submatrix of* A *contains an* $s \times (m-s)$ *zero submatrix for some* s *and* m *with* $1 \leq s < m$ *and* $2 \leq m \leq n$.

(iv) *Every* $m \times m$ *principal submatrix of* A *with* $m \geq 2$ *is reducible.*

(v) $per|A| = a_{11} \cdots a_{nn}$.

Proof. (i) $\Longrightarrow$ (ii). It follows from Corollary 3.3 that every diagonal, except the main diagonal, of A contains a zero. Moreover, for every $m \times m$ principal submatrix B of A, by Fischer's inequality and (i), it is easy to see that $\det B = b_{11} \cdots b_{mm}$, where $b_{11}, \ldots, b_{mm}$ are the main diagonal entries of B. Hence it follows from Corollary 3.3 that every diagonal, except the main diagonal, of B contains a zero. So (ii) holds.

(ii) $\Longrightarrow$ (iii) It is sufficient to prove that (iii) holds for $m = n$. Let $C = A - a_{11}E_{11}$, where E_{11} is the $n \times n$ matrix with a one in the $(1,1)$ entry and zeros elsewhere. Then every diagonal of C contains a zero. By the Frobenius–König theorem (see Theorem 2.5.5 of [3]), C contains a $t \times (n+1-t)$ zero submatrix for some t with $1 \leq t \leq n$. Hence A contains an $s \times (n-s)$ zero submatrix for some s with $1 \leq s < n$.

(iii) $\Longrightarrow$ (iv) It is sufficient to prove that (iv) holds for $m = n$. Since A contains an $s \times (n-s)$ zero submatrix, there exist $i_1 < i_2 \cdots < i_s$, $j_1 < \cdots < j_t$ and $s + t = n$ such that $B[i_1, \ldots, i_s | j_1, \ldots, j_t]$ consisting of the $i_1, \ldots, i_s$-th rows and $j_1, \ldots, j_t$-th columns of B is a zero submatrix. Notice that the main diagonal of A contains no zeros. Then $\{i_1, \ldots, i_s\} \bigcap \{j_1, \ldots, j_t\} = \emptyset$. Hence A is reducible.

(iv) $\Longrightarrow$ (v) We prove the assertion by induction on n. If $n = 2$, then A is reducible, which implies $a_{12}a_{21} = 0$. So $perA = a_{11}a_{22}$. Assume that $n \geq 2$ and that the assertion holds for all positive integers less than n. Since A is reducible, there exists a permutation matrix P such that $PAP^T = \left(\begin{smallmatrix} B & C \\ 0 & D \end{smallmatrix}\right)$, where B and D are square matrices. Hence $perA = perBperD$. By the induction hypothesis, $perB = b_{11} \cdots b_{ss}$ and $perD = d_{11} \cdots d_{tt}$, where $b_{11}, \ldots, b_{ss}, d_{11}, \ldots, d_{tt}$ are the main diagonal entries of A. So (v) holds.

(v) $\Longrightarrow$ (i) follows from the definitions of the determinant and permanent. □

COROLLARY 3.5. *Let* $A = (a_{ij})$ *be an* $n \times n$ *nonsingular* M*-matrix. If* A *is combinatorially symmetric (i.e.,* $a_{ij} = 0$ *implies* $a_{ji} = 0$*), then equality holds in Hadamard's inequality if and only if* A *is a diagonal matrix.*

Proof. It follows from Theorem 3.4. □

4. Equality for Oppenheim's inequality. We first present a preliminary result.

LEMMA 4.1. *Let* $A = (a_{ij})$ *and* $B = (b_{ij})$ *be two* $n \times n$ *general* M*-matrices with* $n > 1$. *Then* $A \circ B$ *is irreducible, and equality holds in Oppenheim's inequality; i.e.,* $\det(A \circ B) = b_{11} \cdots b_{nn} \det A$ *if and only if there exists a permutation matrix* P *such that* PAP^T *is cyclic and such that* PBP^T *is singular and cyclic.*

Proof. Sufficiency. We may assume that A is cyclic and that B is singular and cyclic. Then $b_{11} \cdots b_{nn} = (-1)^n b_{12} \cdots b_{n-1,n} b_{n,1}$. Hence by a simple calculation,

$$\begin{aligned} \det(A \circ B) &= \prod_{i=1}^{n} (a_{ii}b_{ii}) - \prod_{i=1}^{n} (a_{i,i+1}b_{i,i+1}) \\ &= \left(\prod_{i=1}^{n} b_{ii} \right) \left(\prod_{i=1}^{n} a_{ii} - (-1)^n \prod_{i=1}^{n} a_{i,i+1} \right) = \left(\prod_{i=1}^{n} b_{ii} \right) \det A. \end{aligned}$$

Moreover, $A \circ B$ is irreducible.

Necessity. If A is a singular M-matrix, then $\det(A \circ B) = b_{11} \cdots b_{nn} \det A = 0$, which implies that $A \circ B$ is an irreducible singular M-matrix. By Lemma 2.2, the assertion holds. If A is a nonsingular M-matrix, then by Lemma 3.4 of [14], $A - \delta E_{11}$ is a singular M-matrix with $\delta = \frac{\det A}{\det A(1)}$, where $A(1)$ is the $(n-1) \times (n-1)$ principal submatrix of A obtained from A by deleting the first row and column of A, and where E_{11} is the $n \times n$ matrix with a one in the $(1,1)$ entry and zeros elsewhere. Hence, by Theorem 3.2 of [14], $(A - \delta E_{11}) \circ B$ is an irreducible general M-matrix. Therefore,

$$
\begin{aligned}
0 \le \det((A - \delta E_{11}) \circ B) &= \begin{vmatrix} a_{11}b_{11} - \delta b_{11} & -a_{12}b_{12} & \cdots & -a_{1n}b_{1n} \\ -a_{21}b_{21} & a_{22}b_{22} & \cdots & -a_{2n}b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{n1}b_{n1} & -a_{n2}b_{n2} & \cdots & a_{nn}b_{nn} \end{vmatrix} \\
&= \begin{vmatrix} a_{11}b_{11} & -a_{12}b_{12} & \cdots & -a_{1n}b_{1n} \\ -a_{21}b_{21} & a_{22}b_{22} & \cdots & -a_{2n}b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ -a_{n1}b_{n1} & -a_{n2}b_{n2} & \cdots & a_{nn}b_{nn} \end{vmatrix} + \begin{vmatrix} -\delta b_{11} & -a_{12}b_{12} & \cdots & -a_{1n}b_{1n} \\ 0 & a_{22}b_{22} & \cdots & -a_{2n}b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & -a_{n2}b_{n2} & \cdots & a_{nn}b_{nn} \end{vmatrix} \\
&= \det(A \circ B) - \delta b_{11} \det(A(1) \circ B(1)) \\
&\le b_{11} \cdots b_{nn} \det A - \delta b_{11} b_{22} \cdots b_{nn} \det A(1) = 0,
\end{aligned}
$$

since $\det(A \circ B) = b_{11} \cdots b_{nn} \det A$ and $\det(A(1) \circ B(1)) \ge b_{22} \cdots b_{nn} \det(A(1))$ by Oppenheim's inequality. Hence $(A - \delta E_{11}) \circ B$ is an irreducible singular M-matrix. By Lemma 2.2, there exists a permutation matrix P such that both $P(A - \delta E_{11})P^T$ and PBP^T are singular and cyclic. Hence the assertion holds. □

THEOREM 4.2. *Let A and B be two $n \times n$ general M-matrices. Then equality holds in Oppenheim's inequality if and only if one of the following conditions holds:*

(i) $n = 1$.

(ii) $a_{ii} = 0$ *for some* i *with* $1 \le i \le n$.

(iii) $b_{ii} = 0$ *for some* i *with* $1 \le i \le n$.

(iv) *There exists a permutation matrix P such that*

$$
PAP^T = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},
$$

where both A_{11} and B_{11} are $m \times m$ singular and cyclic. Moreover, for every diagonal with an entry in the $(1,2)$ block, either the diagonal contains a zero for PAP^T or it contains a zero for PBP^T.

(v) *There exists a permutation matrix P such that*

$$
PAP^T = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \cdots & \cdots & \cdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} B_{11} & \cdots & B_{1k} \\ \cdots & \cdots & \cdots \\ B_{k1} & \cdots & B_{kk} \end{pmatrix},
$$

where either A_{ii} is an $m_i \times m_i$ cyclic matrix and B_{ii} is an $m_i \times m_i$ singular and cyclic matrix, or A_{ii} and B_{ii} are 1×1 matrices for $i = 1, \ldots, k$. Moreover, every diagonal of A with at least one entry in A_{ij} for some $i \ne j$ contains a zero.

Proof. Sufficiency. It is obvious that the result holds for $n = 1$. Hence we assume that $n > 1$. If $a_{ii} = 0$ or $b_{ii} = 0$ for some i with $1 \le i \le n$, then it is easy to see that $\det(A \circ B) = b_{11} \cdots b_{nn} \det A = 0$. If (iv) holds, then it follows from Theorem 2.4 that $A \circ B$ is singular. Hence $\det(A \circ B) = b_{11} \cdots b_{nn} \det A = 0$. If (v) holds, then it follows from Corollary 3.3 and Lemma 4.1 that $\det(A \circ B) = \det(A_{11} \circ B_{11}) \cdots \det(A_{kk} \circ B_{kk}) = b_{11} \cdots b_{nn} \det A_{11} \cdots \det A_{kk} = b_{11} \cdots b_{nn} \det A$.

Necessity. Assume that $n > 1$ and $a_{ii}b_{ii} > 0$ for $i = 1, \dots, n$. If A is a singular M-matrix, then $\det(A \circ B) = b_{11} \cdots b_{nn} \det A = 0$, which implies $A \circ B$ is a singular M-matrix. By Theorem 2.4, (iv) holds. If A is nonsingular, then by Corollary 2.5, $A \circ B$ is a nonsingular M-matrix. Hence there exists a permutation matrix Q such that

$$QAQ^T = \begin{pmatrix} C_{11} & \cdots & C_{1k} \\ \cdots & \cdots & \cdots \\ C_{k1} & \cdots & C_{kk} \end{pmatrix}, \quad QBQ^T = \begin{pmatrix} D_{11} & \cdots & D_{1k} \\ \cdots & \cdots & \cdots \\ D_{k1} & \cdots & D_{kk} \end{pmatrix},$$

where $C_{ii} \circ D_{ii}$ is an $m_i \times m_i$ irreducible nonsingular matrix for $i = 1, \dots, k$, and where $C_{ij} \circ D_{ij} = 0$ for $1 \le j < i \le k$. Moreover, $Q(A \circ B)Q^T$ is in block upper triangular form. Then it follows from Lemma 3.1 of [14] that $\det(A \circ B) = \det((PAP^T) \circ (PBP^T)) = \det(C_{11} \circ D_{11}) \cdots \det(C_{kk} \circ D_{kk})$. Since C_{ii} and D_{ii} are general M-matrices for $i = 1, \dots, k$, by Oppenheim's inequality,

$$\det(C_{ii} \circ D_{ii}) \ge d_{i_1,i_1} \cdots d_{i_{m_i},i_{m_i}} \det C_{ii} \tag{10}$$

for $i = 1, \dots, k$, where $(d_{i_1,i_1}, \dots, d_{i_{m_i},i_{m_i}})$ is the main diagonal of D_{ii}. Clearly,

$$\prod_{i=1}^{k} (d_{i_1,i_1} \cdots d_{i_{m_i},i_{m_i}}) = b_{11} \cdots b_{nn}.$$

Hence, by Fischer's inequality,

$$\left(\prod_{i=1}^{n} b_{ii} \right) \det A = \det(A \circ B) = \prod_{i=1}^{k} \det(C_{ii} \circ D_{ii})$$
$$\ge \prod_{i=1}^{n} b_{ii} \prod_{i=1}^{k} \det C_{ii} \ge \left(\prod_{i=1}^{n} b_{ii} \right) \det A.$$

Therefore equality holds in (10) for $i = 1, \dots, k$ and $\det A = \det C_{11} \cdots \det C_{kk}$. By Lemma 4.1, either both C_{ii} and D_{ii} are 1×1 matrices or there exists a permutation matrix Q_i such that $Q_i C_{ii} Q_i^T$ is cyclic and that $Q_i D_{ii} Q_i^T$ is singular and cyclic. For $i = 1, \dots, k$, let $R_i = I_1$, the 1×1 identity matrix, if both C_{ii} and D_{ii} are 1×1 matrices; otherwise let $R_i = Q_i$. Let $P = \operatorname{diag}(R_1, \dots, R_k)Q$. Then

$$PAP^T = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \cdots & \cdots & \cdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} B_{11} & \cdots & B_{1k} \\ \cdots & \cdots & \cdots \\ B_{k1} & \cdots & B_{kk} \end{pmatrix},$$

where either A_{ii} is an $m_i \times m_i$ cyclic matrix and B_{ii} is an $m_i \times m_i$ singular and cyclic matrix; or A_{ii} and B_{ii} are 1×1 matrices for $i = 1, \dots, k$. Moreover, by Corollary 3.3, every diagonal of A with at least one entry in A_{ij} $(i \ne j)$ contains a zero. So (v) holds. The proof is finished. □

5. Equality for Ando's inequality. In this section, we characterize necessary and sufficient conditions for equality in Ando's inequality. We need the following lemmas.

LEMMA 5.1. *Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $n \times n$ cyclic general M-matrices. Then equality holds in Ando's inequality* (3).

Proof. Since A and B are cyclic,

$$\begin{aligned}\det(A\circ B) &= \prod_{i=1}^{n}(a_{ii}b_{ii}) - \prod_{i=1}^{n}(a_{i,i+1}b_{i,i+1})\\ &= \left(\prod_{i=1}^{n}a_{ii}\right)\left(\prod_{i=1}^{n}b_{ii} - (-1)^n\prod_{i=1}^{n}b_{i,i+1}\right) + \left(\prod_{i=1}^{n}b_{ii}\right)\left(\prod_{i=1}^{n}a_{ii} - (-1)^n\prod_{i=1}^{n}a_{i,i+1}\right)\\ &\quad - \left(\prod_{i=1}^{n}a_{ii} - (-1)^n\prod_{i=1}^{n}a_{i,i+1}\right)\left(\prod_{i=1}^{n}b_{ii} - (-1)^n\prod_{i=1}^{n}b_{i,i+1}\right)\\ &= \left(\prod_{i=1}^{n}a_{ii}\right)\det B + \left(\prod_{i=1}^{n}b_{i,i+1}\right)\det A - \det A\det B.\end{aligned}$$

So equality holds in Ando's equality. ∎

LEMMA 5.2. *Let A and B be two $n\times n$ nonsingular M-matrices. If $A\circ B$ is irreducible, then equality holds in Ando's inequality if and only if there exists a permutation matrix P such that PAP^T and PBP^T are cyclic.*

Proof. If there exists a permutation matrix P such that PAP^T and PBP^T are cyclic, then it follows from Lemma 5.1 and [14, Corollary 3.1] that equality holds in Ando's inequality.

Conversely, we may assume that $n>1$. By Lemma 3.4 of [14], $A-\frac{\det A}{\det A(1)}E_{11}$ and $B-\frac{\det B}{\det B(1)}E_{11}$ are singular M-matrices, where $A(1)$ and $B(1)$ are the $(n-1)\times(n-1)$ principal submatrices of A and B obtained from A and B by deleting the first row and column of A and B, respectively; and where E_{11} is the $n\times n$ matrix with a one in the $(1,1)$ entry and zeros elsewhere. Hence, by Theorem 3.2 of [14], $(A-\frac{\det A}{\det A(1)}E_{11})\circ(B-\frac{\det B}{\det B(1)}E_{11})$ is a general M-matrix. Hence it follows from the proof of Lemma 4.1 that

$$\begin{aligned}0 &\le \det\left(\left(A-\frac{\det A}{\det A(1)}E_{11}\right)\circ\left(B-\frac{\det B}{\det B(1)}E_{11}\right)\right)\\ &= \det(A\circ B) - \left(b_{11}\frac{\det A}{\det A(1)} + a_{11}\frac{\det B}{\det B(1)} - \frac{\det A\det B}{\det A(1)\det B(1)}\right)\det(A(1)\circ B(1)).\end{aligned}$$

Therefore, by applying Ando's inequality to $A(1)\circ B(1)$ and performing some calculations,

$$\begin{aligned}0 &\le \det(A\circ B) - \left(b_{11}\frac{\det A}{\det A(1)} + a_{11}\frac{\det B}{\det B(1)} - \frac{\det A\det B}{\det A(1)\det B(1)}\right)\det(A(1)\circ B(1))\\ &\le \left(\prod_{i=1}^{n}b_{ii}\right)\det A + \left(\prod_{i=1}^{n}a_{ii}\right)\det B - \det A\det B\\ &\quad - \left(b_{11}\frac{\det A}{\det A(1)} + a_{11}\frac{\det B}{\det B(1)} - \frac{\det A\det B}{\det A(1)\det B(1)}\right)\\ &\quad \times\left(\left(\prod_{i=2}^{n}b_{ii}\right)\det A(1) + \left(\prod_{i=2}^{n}a_{ii}\right)\det B(1) - \det A(1)\det B(1)\right)\\ &= -\left\{\frac{\det B}{\det B(1)}(a_{11}\det A(1) - \det A)\left(\prod_{i=2}^{n}b_{ii} - \det B(1)\right)\right.\\ &\quad \left.+\frac{\det A}{\det A(1)}(b_{11}\det B(1) - \det B)\left(\prod_{i=2}^{n}a_{ii} - \det A(1)\right)\right\} \le 0,\end{aligned}$$

since both $A(1)$ and $B(1)$ are general M-matrices and satisfy Fischer's and Hadamard's inequalities. Hence $(A - \frac{\det A}{\det A(1)}E_{11}) \circ (B - \frac{\det B}{\det B(1)}E_{11})$ is an irreducible singular M-matrix. By Lemma 2.2, there exists a permutation matrix P such that both $P(A - \frac{\det A}{\det A(1)}E_{11})P^T$ and $P(B - \frac{\det B}{\det B(1)}E_{11})P^T$ are singular and cyclic, which implies that PAP^T and PBP^T are cyclic. □

THEOREM 5.3. *Let A and B be two $n \times n$ general M-matrices. Then equality holds in Ando's inequality if and only if one of the following conditions holds:*

(i) $n = 1$.

(ii) $a_{ii} = 0$ *for some* i *with* $1 \le i \le n$.

(iii) $b_{ii} = 0$ *for some* i *with* $1 \le i \le n$.

(iv) *There exists a permutation matrix P such that*

$$PAP^T = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where A_{11} and B_{11} are both singular, cyclic $m \times m$ matrices. Moreover, for every diagonal with an entry in the $(1,2)$ block, either the diagonal contains a zero for PAP^T or it contains a zero for PBP^T.

(v) *There exists a permutation matrix P such that*

$$PAP^T = \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \cdots & \cdots & \cdots \\ A_{k1} & \cdots & A_{kk} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} B_{11} & \cdots & B_{1k} \\ \cdots & \cdots & \cdots \\ B_{k1} & \cdots & B_{kk} \end{pmatrix},$$

where PAP^T and PBP^T satisfy one of the following conditions:

(v.a) *Either A_{ii} is an $m_i \times m_i$ cyclic matrix and B_{ii} is an $m_i \times m_i$ singular and cyclic matrix, or both A_{ii} and B_{ii} are 1×1 matrices for $i = 1, \ldots, k$. Moreover, every diagonal of PAP^T with at least one entry in A_{ij} for some $i \neq j$ contains a zero.*

(v.b) *Either A_{ii} is an $m_i \times m_i$ singular and cyclic matrix and B_{ii} is an $m_i \times m_i$ cyclic matrix, or both A_{ii} and B_{ii} are 1×1 matrices for $i = 1, \ldots, k$. Moreover, every diagonal of PBP^T with at least one entry in B_{ij} for some $i \neq j$ contains a zero.*

(v.c) *$k = n$, and both A_{ii} and B_{ii} are 1×1 matrices for $i = 1, \ldots, k$. Moreover, either every diagonal, except the main diagonal, of PAP^T contains a zero, or every diagonal, except the main diagonal, of PBP^T contains a zero.*

(v.d) *Both A_{11} and B_{11} are $m_1 \times m_1$ cyclic matrices, and both A_{ii} and B_{ii} are 1×1 matrices for $i = 2, \ldots, k$. Moreover, every diagonal of PAP^T with at least one entry in A_{ij} for some $i \neq j$ contains a zero, and every diagonal of PBP^T with at least one entry in B_{ij} for some $i \neq j$ contains a zero.*

Proof. Sufficiency. If one of (i), (ii), (iii), (iv), (v.a), (v.b), and (v.c) holds, it follows from Theorem 4.2 that equality holds in Ando's inequality. If (v.d) holds, it follows from the definition of the determinant that $\det A = a_{m_1+1,m_1+1} \cdots a_{n,n} \det A_{11}$, $\det B = b_{m_1+1,m_1+1} \cdots b_{n,n} \det B_{11}$, and $\det(A \circ B) = \det(A_{11} \circ B_{11}) \cdots \det(A_{kk} \circ B_{kk})$. On the other hand, by Lemma 5.1,

$$\det(A_{11} \circ B_{11}) = \left(\prod_{i=1}^{m_1} a_{ii}\right) \det B_{11} + \left(\prod_{i=1}^{m_1} b_{ii}\right) \det A_{11} - \det A_{11} \det B_{11}. \tag{11}$$

Hence it follows from (11) and $\det(A_{ii} \circ B_{ii}) = a_{m_1+i-1,m_1+i-1} b_{m_1+i-1,m_1+i-1}$ for $i = 2, \dots, k$ that

$$\begin{aligned}\det(A \circ B) &= \det(A_{11} \circ B_{11}) \cdots \det(A_{kk} \circ B_{kk}) \\ &= \left\{ \left(\prod_{i=1}^{m_1} a_{ii} \right) \det B_{11} + \left(\prod_{i=1}^{m_1} b_{ii} \right) \det A_{11} - \det A_{11} \det B_{11} \right\} \prod_{i=m_1+1}^{n} (a_{ii} b_{ii}) \\ &= \left(\prod_{i=1}^{n} a_{ii} \right) \det B + \left(\prod_{i=1}^{n} b_{ii} \right) \det A - \det A \det B.\end{aligned}$$

So equality holds in Ando's inequality.

Necessity. Without loss of generality, we may assume that $n > 1$ and $a_{ii} b_{ii} > 0$ for $i = 1, \dots, n$. We consider the following four cases.

Case 1. If A and B are singular, then $\det(A \circ B) = 0$, which implies that $A \circ B$ is singular. By Theorem 2.4, (iv) holds.

Case 2. If A is nonsingular and B is singular, then $\det(A \circ B) = (\prod_{i=1}^{n} b_{ii}) \det A$. Hence, by Theorem 4.2, (v.a) holds.

Case 3. If A is singular and B is nonsingular, then $\det(A \circ B) = (\prod_{i=1}^{n} a_{ii}) \det B$. Hence, by Theorem 4.2, (v.b) holds.

Case 4. A and B are nonsingular. If $A \circ B$ is irreducible, then, by Lemma 5.2, (v.d) holds for $k = 1$. If $A \circ B$ is reducible and every irreducible block in Frobenius form is a 1×1 submatrix, then every diagonal, except the main diagonal, of $A \circ B$ contains a zero. Hence $(\prod_{i=1}^{n} a_{ii})(\prod_{i=1}^{n} b_{ii}) = \det(A \circ B) = (\prod_{i=1}^{n} a_{ii}) \det B + (\prod_{i=1}^{n} b_{ii}) \det A - \det A \det B$. So $(\prod_{i=1}^{n} a_{ii} - \det A)(\prod_{i=1}^{n} b_{ii} - \det B) = 0$, which implies $\prod_{i=1}^{n} a_{ii} - \det A = 0$ or $\prod_{i=1}^{n} b_{ii} - \det B = 0$. Then, by Theorem 3.4, (v.c) holds. Hence we may assume that there exists a permutation matrix P such that

$$PAP^T = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \quad PBP^T = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix},$$

where $C_{11} \circ D_{11}$ is an $m_1 \times m_1$ irreducible matrix with $m_1 > 1$, and where $C_{21} \circ D_{21} = 0$. Hence $\det(A \circ B) = \det(C_{11} \circ D_{11}) \det(C_{22} \circ D_{22})$. By applying Oppenheim's inequality to $C_{11} \circ D_{11}$ and $C_{22} \circ D_{22}$, respectively, we have

$$\begin{aligned}0 &= \det(A \circ B) - \left(\prod_{i=1}^{n} b_{ii} \right) \det A - \left(\prod_{i=1}^{n} a_{ii} \right) \det B + \det A \det B \\ &\geq \left\{ \left(\prod_{i=1}^{m_1} d_{ii} \right) \det C_{11} + \left(\prod_{i=1}^{m_1} c_{ii} \right) \det D_{11} - \det C_{11} \det D_{11} \right\} \\ &\quad \times \left\{ \left(\prod_{i=m_1+1}^{n} d_{ii} \right) \det C_{22} + \left(\prod_{i=m_1+1}^{n} c_{ii} \right) \det D_{22} - \det C_{22} \det D_{22} \right\} \\ &\quad - \left(\prod_{i=1}^{n} b_{ii} \right) \det A - \left(\prod_{i=1}^{n} a_{ii} \right) \det B + \det A \det B,\end{aligned}$$

where $(c_{11}, \dots, c_{m_1,m_1})$ and $(d_{11}, \dots, d_{m_1,m_1})$ are the main diagonal of C_{11} and D_{11}, respectively; $(c_{m_1+1,m_1+1}, \dots, c_{nn})$ and $(d_{m_1+1,m_1+1}, \dots, d_{nn})$ are the main diagonals of C_{22} and D_{22}, respectively. By performing some calculations, the right-hand side of

the above inequality is equal to

$$\begin{aligned}
&(\det C_{11}\det C_{22}-\det A)(\det D_{11}\det D_{22}-\det B)\\
&+\left\{\left(\prod_{i=1}^{m_1}c_{ii}\right)\det C_{22}-\det\right\}\det D_{11}\left\{\left(\prod_{i=m_1+1}^{n}d_{ii}\right)-\det D_{22}\right\}\\
&+\det C_{11}\left\{\prod_{i=m_1+1}^{n}c_{ii}-\det C_{22}\right\}\left\{\left(\prod_{i=1}^{m_1}d_{ii}\right)\det D_{22}-\det B\right\}\\
&+\left\{\prod_{i=1}^{m_1}c_{ii}-\det C_{11}\right\}\left(\prod_{i=m_1+1}^{n}c_{ii}\right)\{\det D_{11}\det D_{22}-\det B\}\\
&+(\det C_{11}\det C_{22}-\det A)\left(\prod_{i=1}^{m_1}d_{ii}-\det D_{11}\right)\prod_{i=m_1+1}^{n}d_{ii}\\
\geq\ &0.
\end{aligned}$$

Hence, $\det(C_{11}\circ D_{11})=(\prod_{i=1}^{m_1}d_{ii})\det C_{11}+(\prod_{i=1}^{m_1}c_{ii})\det D_{11}-\det C_{11}\det D_{11}$. By Lemma 5.2, there exists a permutation Q_1 such that both $Q_1C_{11}Q_1^T$ and $Q_1D_{11}Q_1^T$ are cyclic and nonsingular matrices. Then $\det C_{11}<\prod_{i=1}^{m_1}c_{ii}$ and $\det D_{11}<\prod_{i=1}^{m_1}d_{ii}$. Moreover,

$$\left(\prod_{i=1}^{m_1}c_{ii}\right)\det C_{22}>\det C_{11}\det C_{22}\geq\det A,\quad \left(\prod_{i=1}^{m_1}d_{ii}\right)\det D_{22}>\det D_{11}\det D_{22}\geq\det B.$$

Hence $\det A=\det C_{11}\det C_{22}$, $\det B=\det D_{11}\det D_{22}$, $\det C_{22}=\prod_{i=m_1+1}^{n}c_{ii}$, and $\det D_{22}=\prod_{i=m_1+1}^{n}d_{ii}$. Let $P=\operatorname{diag}(Q_1,I)$, where I is the $(n-m_1)\times(n-m_1)$ identity matrix. By Theorems 3.2 and 3.4, (v.d) holds. □

COROLLARY 5.4. *Let A and B be two nonsingular M-matrices. If A and B are combinatorially symmetric, then equality holds in Ando's inequality if and only if there exists a permutation matrix P such that*

$$PAP^T=\begin{pmatrix}a_{11}&a_{12}&0&\cdots&0\\a_{21}&a_{22}&0&\cdots&0\\0&0&a_{33}&\cdots&0\\\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&\cdots&a_{nn}\end{pmatrix},\quad PBP^T=\begin{pmatrix}b_{11}&b_{12}&0&\cdots&0\\b_{21}&b_{22}&0&\cdots&0\\0&0&b_{33}&\cdots&0\\\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&\cdots&b_{nn}\end{pmatrix}.$$

Proof. It follows from Theorem 5.3 that the assertion holds. □

Now we may summarize the main results in this paper as follows.

THEOREM 5.5. *Let $A=(a_{ij})$ and $B=(b_{ij})$ be two $n\times n$ general M-matrices. Then*

$$\begin{aligned}
\det(AB)&\leq\left(\prod_{i=1}^{n}b_{ii}\right)\det A\leq\left(\prod_{i=1}^{n}a_{ii}\right)\det B+\left(\prod_{i=1}^{n}b_{ii}\right)\det A-\det A\det B\\
&\leq\det(A\circ B)\leq\prod_{i=1}^{n}(a_{ii}b_{ii}).
\end{aligned}\tag{12}$$

Further,

(i) *the first equality holds in* (12) *if and only if A is singular or $b_{ii}=0$ for some i with $1\leq i\leq n$, or B satisfies the conditions of Theorem* 3.4*;*

(ii) *the second and third equalities hold in* (12) *if and only if* A *and* B *satisfy the conditions of Theorem* 4.2*;*

(iii) *the third equality holds in* (12) *if and only if* A *and* B *satisfy the conditions of Theorem* 5.3*;*

(iv) *the fourth equality holds in* (12) *if and only if* $a_{ii} = 0$ *or* $b_{ii} = 0$ *for some* i *with* $1 \leq i \leq n$*, or either* A *or* B *satisfies the conditions of Theorem* 3.4.

Acknowledgments. The author would like to thank the anonymous referees very much for valuable suggestions, corrections, and comments that have resulted in a great improvement over the original manuscript.

REFERENCES

[1] T. Ando, *Inequalities for M-matrices*, Linear and Multilinear Algebra, 8 (1980), pp. 291–316.
[2] R. B. Bapat and T. E. S. Raghavan, *Nonnegative Matrices and Applications*, Cambridge University Press, Cambridge, UK, 1997.
[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979; reprinted as Classics Appl. Math. 9 by SIAM, Philadelphia, 1994.
[4] S. M. Fallat and C. R. Johnson, *Determinantal inequalities: Ancient history and recent advances*, in Algebra and Its Applications (Athens, OH, 1999), Contemp. Math. 259, AMS, Providence, RI, 2000, pp. 199–212.
[5] K. Fan, *Note on M-matrices*, Quart. J. Math., 11 (1960), pp. 43–49.
[6] K. Fan, *Inequalities for M-matrices*, Indag. Math., 26 (1964), pp. 602–610.
[7] K. Fan, *Inequalities for the sum of two matrices*, in Inequalities, O. Shisha, ed., Academic Press, New York, 1967.
[8] M. Fiedler and V. Pták, *Some inequalities related to M-matrices*, Math. Inequal. Appl., 1 (1998), pp. 171–176.
[9] C. R. Johnson and R. L. Smith, *Almost principal minors of inverse M-matrices*, Linear Algebra Appl., 337 (2001), pp. 253–265.
[10] Y. S. Lee, *On the sum of two N_0-matrices*, Linear and Multilinear Algebra, 26 (1990), pp. 215–221.
[11] Y. S. Lee, *Inequalities for M- and N_0-matrices*, Linear and Multilinear Algebra, 29 (1991), pp. 149–154.
[12] J. Z. Liu and L. Zhu, *Some improvements of Oppenheim's inequality for M-matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 305–311.
[13] J. J. McDonald, M. Neumann, H. Schneider, and M. J. Tsatsomeros, *Inverses of unipathic M-matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1025–1036.
[14] R. L. Smith, *On Fan products of Z-matrices*, Linear and Multilinear Algebra, 37 (1994), pp. 297–302.
[15] O. Taussky, *A recurring theorem on determinants*, Amer. Math. Monthly, 56 (1948), pp. 672–676.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 766–783

SEMISMOOTHNESS OF SPECTRAL FUNCTIONS*

HOUDUO QI† AND XIAOQI YANG‡

Abstract. Any spectral function can be written as a composition of a symmetric function $f : \mathbb{R}^n \mapsto \mathbb{R}$ and the eigenvalue function $\lambda(\cdot) : \mathcal{S} \mapsto \mathbb{R}^n$, often denoted by $(f \circ \lambda)$, where $\mathcal{S}$ is the subspace of $n \times n$ symmetric matrices. In this paper, we present some nonsmooth analysis for such spectral functions. Our main results are (a) $(f \circ \lambda)$ is directionally differentiable if f is semidifferentiable, (b) $(f \circ \lambda)$ is LC^1 if and only if f is LC^1, and (c) $(f \circ \lambda)$ is SC^1 if and only if f is SC^1. Result (a) is complementary to a known (negative) fact that $(f \circ \lambda)$ might not be directionally differentiable if f is directionally differentiable only. Results (b) and (c) are particularly useful for the solution of LC^1 and SC^1 minimization problems which often can be solved by fast (generalized) Newton methods. Our analysis makes use of recent results on continuously differentiable spectral functions as well as on nonsmooth symmetric–matrix-valued functions.

Key words. symmetric function, spectral function, nonsmooth analysis, semismooth function

AMS subject classifications. 49M45, 90C25, 90C33

DOI. 10.1137/S0895479802417921

1. Introduction.

There has been growing interest in the variational analysis of spectral functions. This growing trend is probably due to the following reasons. On one hand, spectral functions have important applications to some fundamental problems in applied mathematics such as semidefinite programs and engineering problems. See a survey paper by Lewis and Overton [14] for many such applications. On the other hand, efficient nonsmooth analysis tools have only been available in the past few years; see the book by Rockafellar and Wets [26]. In this paper, we study some nonsmooth properties of spectral functions which have not been reported in the literature. Our study is inspired by recent progress on spectral functions [13, 15, 16] and progress on symmetric–matrix-valued functions [2, 27, 3, 28, 11].

Let $\mathcal{S}$ be the space of $n \times n$ real symmetric matrices endowed with the inner product $\langle X, Y \rangle := \text{trace}(XY)$ for any $X, Y \in \mathcal{S}$. $\|X\|$ is the Frobenius norm of X. Let $\lambda(\cdot) : \mathcal{S} \to \mathbb{R}^n$ be the eigenvalue function such that $\lambda_i(X)$, $i = 1, \ldots, n$, yield eigenvalues of X for any $X \in \mathcal{S}$ and are patterned in nonincreasing order, i.e., $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is symmetric on an open set $\Omega \subseteq \mathbb{R}^n$ if f is invariant under coordinate permutation, i.e.,

$$f(x) = f(Px) \quad \text{for any permutation matrix } P \text{ and any } x \in \Omega.$$

For simplicity, we assume that Ω is $\mathbb{R}^n$ in this paper (all results remain valid when restricted to some open symmetric set Ω). Formally, a *spectral function* is a composition of a symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ and the eigenvalue function $\lambda(\cdot) : \mathcal{S} \to \mathbb{R}^n$;

*Received by the editors November 15, 2002; accepted for publication (in revised form) by M. L. Overton July 28, 2003; published electronically February 24, 2004.
http://www.siam.org/journals/simax/25-3/41792.html

†School of Mathematics, The University of New South Wales, Sydney NSW 2052, Australia (hdqi@maths.unsw.edu.au). The research of this author was supported by the Australian Research Council.

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maqilq@polyu.edu.hk). The research of this author was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (project PolyU 5141/01E).

that is, the spectral function $(f \circ \lambda) : \mathcal{S} \to \mathbb{R}$ is given by

$$(f \circ \lambda)(X) := f(\lambda(X)), \qquad X \in \mathcal{S}.$$

For more explanation leading to this definition, see [16]. Typical spectral functions include the kth largest eigenvalue of a symmetric matrix [14, 15] and the Schatten p-norm of a symmetric matrix ($p \geq 1$).

It is well known that the eigenvalue function $\lambda(\cdot)$ is not everywhere differentiable. So it is natural to expect that the composite function $(f \circ \lambda)$ could be not everywhere differentiable no matter how smooth f is. It was therefore surprising when Lewis claimed in [13] that $(f \circ \lambda)$ is indeed (strictly) differentiable at $X \in \mathcal{S}$ if and only if f is (strictly) differentiable at $\lambda(X)$. Moreover, it is further proved in [16] that $(f \circ \lambda)$ is twice (continuously) differentiable at $X \in \mathcal{S}$ if and only if f is twice (continuously) differentiable at $\lambda(X)$. Those two results on derivatives play an important role in this paper. It is also known that $(f \circ \lambda)$ is convex if and only if f is convex [5]. Since the eigenvalue function is Lipschitz continuous, $(f \circ \lambda)$ is locally Lipschitzian if f is. Then the generalized gradient $\partial(f \circ \lambda)$ in the sense of Clarke [4] is well defined. A beautiful formula for calculating elements in $\partial(f \circ \lambda)$ can be found in [13]. Several other subgradients of $(f \circ \lambda)$ are studied in [15]; see also [8].

The above results show that $(f \circ \lambda)$ inherits smoothness properties from f. However, this is not the case for directional differentiability. The punctured hyperbola example constructed by Lewis [13] shows that $(f \circ \lambda)$ is not necessarily directionally differentiable if f is directionally differentiable only. We will show that a sufficient condition for directional differentiability of $(f \circ \lambda)$ at $X \in \mathcal{S}$ is the semidifferentiability of f at $\lambda(X)$ (see Proposition 3.2). This result suggests that f should have differentiability properties stronger than directional differentiability in order for $(f \circ \lambda)$ to inherit the same properties from f. In fact, we will show that $(f \circ \lambda)$ is $\min(1, \rho)$-order semismooth if and only if f is ρ-order semismooth (see Proposition 3.5), generalizing a recent result of Sun and Sun [28] which proves that the eigenvalue function is strongly semismooth. As mentioned earlier, $(f \circ \lambda)$ is (twice) differentiable if and only if f is (twice) differentiable. We are also interested in the case when f is an LC^1 function (also called a $C^{1,1}$ function in the literature), i.e., f is once continuously differentiable and its derivative function $\nabla f(\cdot)$ is locally Lipschitz. Another interesting case is when f is an SC^1 function, i.e., f is not only an LC^1 function, but also its derivative function is *semismooth*. For both cases, we will show that $(f \circ \lambda)$ is an LC^1 (respectively, SC^1) function (see Propositions 4.3 and 4.5). The importance of LC^1 and SC^1 functions is that they constitute a class of minimization problems which can be solved by Newton-type methods (see [6, 20, 22]) and by penalty-type methods (see [31, 30]).

The property of semismoothness, as introduced by Mifflin [17, 18] for functionals and scalar-valued functions and further extended by Qi and Sun [23] for vector-valued functions, is of particular interest due to the key role it plays in the superlinear convergence analysis of certain generalized Newton methods [10, 21, 23]. Recent attention in research on semismoothness is on symmetric–matrix-valued functions which have important applications to semidefinite complementarity problems [29, 27, 2, 3, 28, 11]. Several important results have been established and inspired our research in this paper. For example, the absolute matrix-valued function $|X| := \sqrt{X^2}$, $X \in \mathcal{S}$, is strongly semismooth [27, 3]; the eigenvalue function $\lambda(\cdot)$ is strongly semismooth [28]. This latter result is found to be particularly useful in quadratic convergence analysis of Newton methods for inverse eigenvalue problems. Another useful result is

a lemma of Chen and Tseng [2] about the locally upper Lipschitzian property of certain orthogonal matrices yielding the spectral decomposition of a symmetric matrix.

Notation used in this paper is as follows: vectors in $\mathbb{R}^n$ are viewed as columns and capital letters such as X, Y, etc. always denote matrices in $\mathcal{S}$. For $X \in \mathcal{S}$, we denote by X_{ij} the (i,j)th entry of X. We use $\circ$ to denote the Hadamard product between two matrices, i.e.,

$$X \circ Y = [X_{ij}Y_{ij}]_{i,j=1}^n.$$

Let the operator diag : $\mathcal{S} \to \mathbb{R}^n$ be defined by $\mathrm{diag}[X] := (X_{11}, \ldots, X_{nn})^T$, while for $\mu \in \mathbb{R}^n$, $\mathrm{Diag}[\mu_1, \ldots, \mu_n]$ denotes the diagonal matrix with its ith diagonal entry μ_i. Sometimes we write $\mathrm{Diag}[\mu]$ instead of $\mathrm{Diag}[\mu_1, \ldots, \mu_n]$ for simplicity. Let $\mathcal{P}$ denote the set of all permutation matrices in $\mathbb{R}^{n\times n}$. For any given $\mu \in \mathbb{R}^n$, $\mathcal{P}_\mu$ denotes the stabilizer of μ defined by

$$\mathcal{P}_\mu := \{P \in \mathcal{P} |\ P\mu = \mu\}.$$

Throughout, $\|\cdot\|$ denotes the Frobenius norm for matrices and the 2-norm for vectors. For any linear mapping $\mathcal{L} : \mathcal{S} \to \mathcal{S}$, we define its operator norm $|||\mathcal{L}||| := \max_{\|X\|=1} \|\mathcal{L}X\|$. For any $x \in \mathbb{R}^n$, $X \in \mathcal{S}$, and any scalar $\gamma > 0$, we denote the γ-ball around x in $\mathbb{R}^n$ and the γ-ball around X in $\mathcal{S}$, respectively, by

$$\mathcal{N}(x,\gamma) := \{y \in \mathbb{R}^n |\ \|y - x\| \le \gamma\},$$
$$\mathcal{B}(X,\gamma) := \{Y \in \mathcal{S} |\ \|Y - X\| \le \gamma\}.$$

For any $\mu \in \mathbb{R}^n$ and $P \in \mathcal{P}$, we will frequently use the following fact:

$$\mathrm{Diag}[P\mu] = P\mathrm{Diag}[\mu]P^T.$$

2. Miscellaneous. In this section, we review some basic concepts on continuity and differentiability of vector-valued functions in order to avoid confusion with other concepts not treated in this paper. Those concepts also apply to the spectral function $(f \circ \lambda)$ and its gradient map $\nabla(f \circ \lambda)$ (if it exists) since the symmetric matrix space $\mathcal{S}$ can be cast as a vector space of dimension $n(n+1)/2$. All those concepts except semismoothness and their equivalent characterizations can be found in the book [26]. We also list some perturbation results on symmetric matrices for later use.

2.1. Basic concepts. Consider the mapping $F : \mathbb{R}^k \mapsto \mathbb{R}^\ell$. We say F is continuous at $x \in \mathbb{R}^k$ if $F(y) \to F(x)$ as $y \to x$, and F is continuous if F is continuous at every $x \in \mathbb{R}^k$. F is strictly continuous (also called "locally Lipschitz continuous") at $x \in \mathbb{R}^k$ [26, Chap. 9] if there exist scalars $\kappa > 0$ and $\delta > 0$ such that

$$\|F(y) - F(z)\| \le \kappa\|y - z\| \quad \forall y, z \in \mathbb{R}^k \text{ with } \|y - x\| \le \delta,\ \|z - x\| \le \delta,$$

and F is strictly continuous if F is strictly continuous at every $x \in \mathbb{R}^k$. If δ can be taken to be ∞, then F is Lipschitz continuous with Lipschitz constant κ. Define the function $\mathrm{lip}F : \mathbb{R}^k \to [0, \infty]$ by

$$\mathrm{lip}F(x) := \limsup_{\substack{y,z\to x \\ y\neq z}} \frac{\|F(y) - F(z)\|}{\|y - z\|}.$$

Then F is strictly continuous at x if and only if $\mathrm{lip}F(x)$ is finite.

We say F is directionally differentiable at $x \in \mathbb{R}^k$ if

$$F'(x;h) \ := \ \lim_{\tau \to 0^+} \frac{F(x+\tau h) - F(x)}{\tau} \quad \text{exists} \qquad \forall h \in \mathbb{R}^k,$$

and F is directionally differentiable if F is directionally differentiable at every $x \in \mathbb{R}^k$. We say F is semidifferentiable at $x \in \mathbb{R}^k$ if the limit

$$\lim_{\substack{\tau \searrow 0 \\ \hat{h} \to h}} \frac{F(x+\tau \hat{h}) - F(x)}{\tau}$$

exists for every direction $h \in \mathbb{R}^n$. It is easy to see that the limit (if it exists) equals $F'(x;h)$. F is differentiable (in the Fréchet sense) at $x \in \mathbb{R}^k$ if there exists a linear mapping $\nabla F(x) : \mathbb{R}^k \mapsto \mathbb{R}^\ell$ such that

$$F(x+h) - F(x) - \nabla F(x)h \ = \ o(\|h\|).$$

We say that F is continuously differentiable if F is differentiable at every $x \in \mathbb{R}^k$ and ∇F is continuous. If F is strictly continuous, then F is almost everywhere differentiable by Rademacher's theorem; see [4] and [26, Sec. 9J]. Then the generalized Jacobian $\partial F(x)$ of F at x (in the Clarke sense) is well defined.

DEFINITION 2.1 (semismoothness). *Suppose that $F : \mathbb{R}^k \to \mathbb{R}^\ell$ is a strictly continuous function. F is said to be* semismooth *at $x \in \mathbb{R}^k$ if F is directionally differentiable at x and for any $V \in \partial F(x+h)$,*

$$F(x+h) - F(x) - Vh = o(\|h\|). \tag{1}$$

F is said to be ρ-order semismooth $(0 < \rho < \infty)$ at x if F is semismooth at x and

$$F(x+h) - F(x) - Vh = O(\|h\|^{1+\rho}). \tag{2}$$

In particular, F is called strongly semismooth at x if F is 1-order semismooth at x.

We say F is semismooth (respectively, ρ-order semismooth) if F is semismooth (respectively, ρ-order semismooth) at every $x \in \mathbb{R}^k$. Convex functions and piecewise continuously differentiable functions are examples of semismooth functions. The composition of two (respectively, ρ-order) semismooth functions is also a (respectively, ρ-order) semismooth function. The characterization below obtained by Sun and Sun [27, Thm. 3.7] provides a convenient way for proving ρ-order semismoothness and semismoothness as well. For more applications of this result, see [3, 28].

LEMMA 2.2. *Suppose that $F : \mathbb{R}^k \to \mathbb{R}^\ell$ is strictly continuous and directionally differentiable in a neighborhood of x. Then for any $\rho \in (0, \infty)$ the following two statements are equivalent:*

(a) *for any $V \in \partial F(x+h)$,*

$$F(x+h) - F(x) - Vh = O(\|h\|^{1+\rho});$$

(b) *for any $h \in \mathbb{R}^k$ such that F is differentiable at $x+h$,*

$$F(x+h) - F(x) - \nabla F(x+h)h = O(\|h\|^{1+\rho}). \tag{3}$$

In particular, the following two statements are equivalent:

(c) *for any* $V \in \partial F(x+h)$,

$$F(x+h) - F(x) - Vh = o(\|h\|);$$

(d) *for any* $h \in \mathbb{R}^k$ *such that* F *is differentiable at* $x+h$,

$$F(x+h) - F(x) - \nabla F(x+h)h = o(\|h\|).$$

Finally we assume that $F : \mathbb{R}^k \mapsto \mathbb{R}^\ell$ is continuously differentiable. We say that F is an LC^1 function if ∇F is strictly continuous, and that F is an SC^1 function if F is an LC^1 function and ∇F is semismooth. For more discussion on LC^1 and SC^1 functions and their roles in superlinear convergence analysis of certain generalized Newton methods for some minimization problems, see [22, 20, 6]. We note that the LC^1 problem is also known as $C^{1,1}$ data in [9], where second-order analysis of the underlying function is conducted. For further development along this line, see [30, 31] and the references therein.

2.2. Perturbation results for symmetric matrices. In this subsection, we review some useful perturbation results for the spectral decomposition of real symmetric matrices. These results will be used in the next section to analyze properties of the spectral function $(f \circ \lambda)$.

Let $\mathcal{O}$ denote the group of $n \times n$ real orthogonal matrices. For each $X \in \mathcal{S}$, define the set of orthogonal matrices giving the ordered spectral decomposition of X by

$$\mathcal{O}_X := \{P \in \mathcal{O} |\ P^T X P = \mathrm{Diag}[\lambda(X)]\}.$$

Clearly $\mathcal{O}_X$ is nonempty for each $X \in \mathcal{S}$. The following lemma, proved in [2, Lem. 3], gives a key perturbation result for eigenvectors of symmetric matrices. For a different yet simple proof of this lemma, see [28].

LEMMA 2.3. *For any* $X \in \mathcal{S}$, *there exist scalars* $\eta > 0$ *and* $\epsilon > 0$ *such that*

$$\min_{P \in \mathcal{O}_X} \|P - Q\| \le \eta \|X - Y\| \quad \forall\ Y \in \mathcal{B}(X, \epsilon),\ \forall Q \in \mathcal{O}_Y. \tag{4}$$

We will also need the following perturbation results of von Neumann [19]; see also [1].

LEMMA 2.4. *For any* $X, Y \in \mathcal{S}$, *we have*

$$\|\lambda(X) - \lambda(Y)\| \le \|X - Y\| \quad \textit{and} \quad |\lambda_i(X) - \lambda_i(Y)| \le \|X - Y\|_2 \quad \forall\ i = 1, \ldots, n,$$

where $\|\cdot\|_2$ *is the* 2*-norm.*

Last, we need the following classical result [25, Thm. 1] showing that, for any $X \in \mathcal{S}$ and any $H \in \mathcal{S}$, the orthonormal eigenvectors of $X + \tau H$ may be chosen to be analytic in τ. As is remarked in [12, p. 122], the existence of such orthonormal eigenvectors depending smoothly on τ is one of the most remarkable results in the analytic perturbation theory for symmetric operators.

LEMMA 2.5. *For any* $X \in \mathcal{S}$ *and any* $H \in \mathcal{S}$, *there exist* $P(\tau) \in \mathcal{O}$, $\tau \in \mathbb{R}$, *whose entries are power series in* τ, *convergent in a neighborhood of* $\tau = 0$, *and* $P(\tau)^T (X + \tau H) P(\tau)$ *is diagonal.*

3. Directional differentiability and semismoothness of spectral functions. This section includes two main results. Proposition 3.2 says that the spectral function $(f \circ \lambda)$ is directionally differentiable if f is semidifferentiable. Without this

condition, the punctured hyperbola example [13] shows that $(f \circ \lambda)$ is not necessarily directionally differentiable. Proposition 3.5 says that $(f \circ \lambda)$ inherits semismoothness from f. The following preliminary results, which shall be used from time to time in our proofs, are due to the symmetry of f. For example, parts (a), (c), and (d) of Lemma 3.1 follow from differentiating both sides of the equality $f(\mu) = f(P\mu)$ $(P \in \mathcal{P})$ and the chain rule. Part (b) is a direct consequence from the definition of semidifferentiability and the symmetry of f.

LEMMA 3.1. *Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is symmetric. Then we have the following results:*

(a) *f is directionally differentiable at $\mu \in \mathbb{R}^n$ along $h \in \mathbb{R}^n$ if and only if f is directionally differentiable at $P\mu$ along Ph for any $P \in \mathcal{P}$.*
(b) *f is semidifferentiable at $\mu \in \mathbb{R}^n$ if and only if f is semidifferentiable at $P\mu$ for any $P \in \mathcal{P}$.*
(c) *f is differentiable at $\mu \in \mathbb{R}^n$ if and only if f is differentiable at $P\mu$ for any $P \in \mathcal{P}$. In particular, $\nabla f(P\mu) = P\nabla f(\mu)$. Moreover, if $P \in \mathcal{P}_\mu$, then $\nabla f(\mu) = P\nabla f(\mu)$. Consequently, $(\nabla f(\mu))_i = (\nabla f(\mu))_j$ if $\mu_i = \mu_j$ for some $i, j \in \{1, \ldots, n\}$.*
(d) *f is twice differentiable at $\mu \in \mathbb{R}^n$ if and only if f is twice differentiable at $P\mu$ for any $P \in \mathcal{P}$. In this case we have $\nabla^2 f(P\mu) = P\nabla^2 f(\mu)P^T$.*

The next result states that under the condition of semidifferentiability the directional differentiability of f is inherited by the spectral function $(f \circ \lambda)$. Without this condition, this result is no longer valid as the punctured hyperbola example in [13, p. 587] illustrates.

PROPOSITION 3.2. *Let $X \in \mathcal{S}$ be given. The following results hold.*

(a) *Suppose that f is semidifferentiable at $\lambda(X)$. Then $(f \circ \lambda)$ is directionally differentiable at X.*
(b) *Conversely, if $(f \circ \lambda)$ is directionally differentiable at X, then f is directionally differentiable at $\lambda(X)$.*
(c) *Suppose that f is both strictly continuous and directionally differentiable at $\lambda(X)$. Then $(f \circ \lambda)$ is directionally differentiable at X.*

Proof. (a) Let $H \in \mathcal{S}$ and define

$$X(\tau) = X + \tau H, \quad \tau \in \mathbb{R}.$$

Then by Lemma 2.5 there exists $P(\tau) \in \mathcal{O}$, $\tau \in \mathbb{R}$, whose entries are power series in τ, convergent in a neighborhood $\mathcal{I}$ of $\tau = 0$, and $P^T(\tau)X(\tau)P(\tau)$ is diagonal. Consequently the corresponding eigenvalues

$$\mu_i(\tau) = [P^T(\tau)X(\tau)P(\tau)]_{ii}, \quad i = 1, \ldots, n,$$

are also power series in τ, convergent for $\tau \in \mathcal{I}$. Denote

$$\mu(\tau) := (\mu_1(\tau), \ldots, \mu_n(\tau))^T.$$

Then we have the expansion

$$\mu(\tau) = \mu(0) + \tau\mu'(0) + o(\tau). \tag{5}$$

The fact that the elements of $\mu(\tau)$ are eigenvalues of $X(\tau)$ yields

$$\lim_{\tau \searrow 0} \frac{(f \circ \lambda)(X + \tau H) - (f \circ \lambda)(X)}{\tau}$$

$$
\begin{aligned}
&= \lim_{\tau \searrow 0} \frac{f(\mu(\tau)) - f(\mu(0))}{\tau} \\
&= \lim_{\tau \searrow 0} \frac{f(\mu(0) + \tau\mu'(0) + o(\tau)) - f(\mu(0))}{\tau} \\
&= f'(\mu(0); \mu'(0)),
\end{aligned}
$$

where the last equality uses the semidifferentiability of f at $\lambda(X)$. This proves that $(f \circ \lambda)$ is directionally differentiable at X.

(b) The proof of this part is standard and follows by restricting the spectral function to the subspace of diagonal matrices and application of Lemma 3.1(a).

(c) This part follows directly from (a) since the strict continuity and directional differentiability of f at $\lambda(X)$ imply the semidifferentiability of f at $\lambda(X)$. □

The sufficient condition of semidifferentiability in Proposition 3.2(a) cannot be replaced by directional differentiability in general. However, it can be so if f has the separable form

$$
f(x) = g(x_1) + \cdots + g(x_n), \tag{6}
$$

where $g : \mathbb{R} \to \mathbb{R}$ is directionally differentiable. The proof is simple by noticing in the preceding argument for (a) that

$$
f(\mu(\tau)) = \sum_{i=1}^{n} g(\mu_i(\tau)) = \sum_{i=1}^{n} (g(\mu_i(0) + \tau g'(\mu_i(0); \mu_i'(0)) + o(\tau))).
$$

Hence for this special case we have

$$
(f \circ \lambda)'(X; H) = \sum_{i=1}^{n} g'(\mu_i(0); \mu_i'(0)).
$$

The next result on differentiability of spectral functions will be used in our analysis of semismoothness (Proposition 3.5) and LC^1 property (Proposition 4.3) of spectral functions.

LEMMA 3.3 (see [13, Thm. 1.1 and Cor. 2.5]). *Let $X \in \mathcal{S}$. $(f \circ \lambda)$ is differentiable at X if and only if f is differentiable at $\lambda(X)$. In this case the gradient of $(f \circ \lambda)$ at X is*

$$
\nabla(f \circ \lambda)(X) = V \mathrm{Diag}[\nabla f(\mu)] V^T \tag{7}
$$

for any orthogonal matrix $V \in \mathcal{O}$ and $\mu \in \mathbb{R}^n$ satisfying $X = V \mathrm{Diag}[\mu] V^T$.

The result below shows that semismoothness implies semidifferentiability.

LEMMA 3.4. *Let $F : \mathbb{R}^k \mapsto \mathbb{R}^\ell$ and $x \in \mathbb{R}^k$. Suppose that F is semismooth at x. Then F is semidifferentiable at x.*

Proof. An equivalent characterization of semismoothness of F at x is that the limit

$$
\lim_{\substack{\hat{h} \to h \\ \tau \searrow 0}} \left\{ V\hat{h} |\ V \in \partial F(x + \tau\hat{h}) \right\} \tag{8}
$$

exists for any $h \in \mathbb{R}^k$ and equals $F'(x; h)$; see [23]. Let $h \in \mathbb{R}^k$ be given. For any $\tau \searrow 0$ and $\hat{h} \to h$, choose any element $V \in \partial F(x + \tau\hat{h})$; we then have

$$
\lim_{\substack{\hat{h} \to h \\ \tau \searrow 0}} \left(F(x + \tau\hat{h}) - F(x) \right) / \tau = \lim_{\substack{\hat{h} \to h \\ \tau \searrow 0}} \left(F(x + \tau\hat{h}) - F(x) - \tau V\hat{h} + \tau V\hat{h} \right) / \tau
$$

$$= \lim_{\substack{\hat{h}\to h\\ \tau\searrow 0}} o(\tau\|\hat{h}\|)/\tau + \lim_{\substack{\hat{h}\to h\\ \tau\searrow 0}} V\hat{h} = F'(x;h).$$

Hence F is semidifferentiable at x. □

The converse of the above result is not true, i.e., a semidifferentiable function is not necessarily semismooth. For example, let $F : \mathbb{R}^n \mapsto \mathbb{R}$ be defined by

$$F(x) := \begin{cases} \|x\|^2 \sin(\frac{1}{\|x\|}) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

This function is locally Lipschitzian, differentiable everywhere, smooth everywhere except at the origin, and semidifferentiable at 0. But it is not semismooth at 0 [24].

Now we present the second main result in this section. The sufficient part says that the spectral function $(f \circ \lambda)$ inherits semismoothness from f, which can also be obtained by using a recent result of Sun and Sun [28] that the eigenvalue function $\lambda(\cdot)$ is strongly semismooth and the fact that compositions of ρ-order semismooth functions are ρ-order semismooth [7]. However, we include a different proof here because it is direct and suggests a proof technique in analyzing SC^1 property of spectral functions in the next section.

PROPOSITION 3.5. *For any symmetric function $f : \mathbb{R}^n \mapsto \mathbb{R}$, the spectral function $(f \circ \lambda)$ is semismooth if and only if f is semismooth. If f is ρ-order semismooth $(0 < \rho < \infty)$, then $(f \circ \lambda)$ is $\min\{1, \rho\}$-order semismooth.*

Proof. Suppose f is semismooth. Then f is strictly continuous and semidifferentiable (Lemma 3.4). Hence $(f\circ\lambda)$ is strictly continuous and directionally differentiable (Lemma 3.2). Let $\mathcal{D} := \{X \in \mathcal{S} |\ (f \circ \lambda) \text{ is differentiable at } X\}$.

Fix any $X \in \mathcal{S}$. By Lemma 2.3, there exist scalars $\eta > 0$ and $\epsilon > 0$ such that (4) holds. We will show that, for any $H \in \mathcal{S}$ with $X + H \in \mathcal{D}$ and $\|H\| \leq \epsilon$, we have

$$(f \circ \lambda)(X + H) - (f \circ \lambda)(X) - \langle \nabla(f \circ \lambda)(X + H), H\rangle = o(\|H\|), \tag{9}$$

where $o(\cdot)$ and $O(\cdot)$ depend on f and X only. Then it follows from Lemma 2.2 that $(f \circ \lambda)$ is semismooth at X. Since the choice of $X \in \mathcal{S}$ was arbitrary, $(f \circ \lambda)$ is semismooth. Now choose any $Q \in \mathcal{O}_{X+H}$. Then Lemma 2.3 implies that there exists $P \in \mathcal{O}_X$ satisfying

$$\|P - Q\| \leq \eta\|H\|.$$

For simplicity, let r denote the left-hand side of (9), i.e.,

$$r := (f \circ \lambda)(X + H) - (f \circ \lambda)(X) - \langle \nabla(f \circ \lambda)(X + H), H\rangle.$$

We also let

$$\Delta_1 := f(\lambda(X + H)) - f(\lambda(X)) - \langle \nabla f(\lambda(X + H)), \lambda(X + H) - \lambda(X)\rangle$$

and

$$\Delta_2 := \langle \nabla f(\lambda(X + H)), \lambda(X + H) - \lambda(X) - \mathrm{diag}[Q^T HQ]\rangle.$$

Since $(f \circ \lambda)$ is differentiable at $X + H \in \mathcal{D}$, it follows from Lemma 3.3 that f is differentiable at $\lambda(X + H)$. Hence, Δ_1 and Δ_2 are well defined. Note that $\nabla f(\lambda(X + H))$ and $\lambda(X + H) - \lambda(X)$ are column vectors. We write their inner product in the

form of $\langle\cdot,\cdot\rangle$ rather than x^Ty for $x, y \in \mathbb{R}^n$ in order to be consistent with the inner product in $\mathcal{S}$. Using the gradient formula (7), we then have

$$\begin{aligned}&\langle\nabla(f\circ\lambda)(X+H), H\rangle = \langle Q\mathrm{Diag}[\nabla f(\lambda(X+H))]Q^T, H\rangle\\ &= \langle \mathrm{Diag}[\nabla f(\lambda(X+H))], Q^THQ\rangle = \langle \nabla f(\lambda(X+H)), \mathrm{diag}[Q^THQ]\rangle,\end{aligned}$$

yielding

$$r = \Delta_1 + \Delta_2.$$

Since f is semismooth at $\lambda(X)$ and $\lambda(X+H) \to \lambda(X)$ as $\|H\| \to 0$, it follows from Lemmas 2.2 and 2.4 that

$$\Delta_1 = o(\|\lambda(X+H) - \lambda(X)\|) = o(\|H\|).$$

It remains to show $\Delta_2 = o(\|H\|)$ in order to show $r = o(\|H\|)$. Let $\tilde{H} := Q^THQ$ and $O := P^TQ$. For simplicity, we let $\mu := \lambda(X+H)$ and $\beta := \lambda(X)$. Since

$$\mathrm{Diag}[\mu] = Q^T(X+H)Q = O^T\mathrm{Diag}[\beta]O + \tilde{H},$$

we have

$$\sum_{k=1}^n O_{ki}O_{kj}\beta_k + \tilde{H}_{ij} = \begin{cases}\mu_i & \text{if } i = j,\\ 0 & \text{else,}\quad i,j = 1,\dots,n.\end{cases} \tag{10}$$

Since $O = P^TQ = (P-Q)^TQ + I$ and $\|P-Q\| \le \eta\|H\|$, it follows that

$$O_{ij} = O(\|H\|) \quad \text{for } i \ne j. \tag{11}$$

Since $P, Q \in \mathcal{O}$, we have $O \in \mathcal{O}$ so that $O^TO = I$. This and (11) imply

$$1 = O_{ii}^2 + \sum_{k\ne i} O_{ki}^2 = O_{ii}^2 + O(\|H\|^2), \quad i = 1,\dots,n. \tag{12}$$

Then, for $i = 1,\dots,n$, the relations (10)–(12) yield

$$\begin{aligned}&\mu_i - \beta_i - (Q^THQ)_{ii} = \sum_{k=1}^n O_{ki}^2\beta_k + \tilde{H}_{ii} - \beta_i - \tilde{H}_{ii}\\ &= O_{ii}^2\beta_i + \sum_{k\ne i} O_{ki}^2\beta_k - \beta_i = \beta_i - \beta_i + O(\|H\|^2) = O(\|H\|^2).\end{aligned}$$

Hence we have

$$\lambda(X+H) - \lambda(X) - \mathrm{diag}[Q^THQ] = O(\|H\|^2),$$

which in turn implies $\Delta_2 = O(\|H\|^2)$. This proves that $(f\circ\lambda)$ is semismooth.

Suppose that f is ρ-order semismooth at $\lambda(X)$ $(0 < \rho < \infty)$. Then the preceding argument shows that

$$r = \Delta_1 + \Delta_2 = O(\|H\|^{1+\rho}) + O(\|H\|^2) = O(\|H\|^{1+\min\{1,\rho\}}).$$

This shows that $(f\circ\lambda)$ is $\min\{1,\rho\}$-order semismooth at X.

Suppose now $(f\circ\lambda)$ is semismooth. Then $(f\circ\lambda)$ is directionally differentiable and strictly continuous. By Proposition 3.2, f is directionally differentiable. It is well known that $(f\circ\lambda)$ is strictly continuous if and only if f is. Then the semismoothness of f follows from restricting the spectral function $(f\circ\lambda)$ to the subspace of diagonal matrices and using the property of semismoothness of $(f\circ\lambda)$ and Lemma 2.2. □

4. LC^1 and SC^1 spectral functions. The purpose of this section is to show that the spectral function $(f\circ\lambda)$ inherits LC^1 and SC^1 properties from f. To establish those properties, we need two more known results. One is a result of Rockafellar and Wets saying that any Lipschitz function has a uniform approximation of a sequence of continuously differentiable functions (on compact domain). The other is a result of Lewis and Sendov on twice continuously differentiable spectral functions.

LEMMA 4.1 (see [26, Thm. 9.67]). *Given $f:\mathbb{R}^n\to\mathbb{R}$, and Ω is an open subset in $\mathbb{R}^n$. If f is strictly continuous on Ω, then there exist functions $f^\nu:\mathbb{R}^n\to\mathbb{R}$, $\nu=1,2,\ldots$, continuously differentiable and converging uniformly to f on any compact set contained in Ω. Moreover, if f is an LC^1 function on Ω, then there are twice continuously differentiable functions f^ν such that $\{\nabla f^\nu\}$ converge uniformly to ∇f on any compact set C contained in Ω, and*

$$\left|\left|\left|\nabla^2 f^\nu(x)\right|\right|\right| \le \operatorname{lip}\nabla f(x) \quad \forall\nu. \tag{13}$$

If f is symmetric, then the smooth approximants $\{f^\nu\}$ can also be selected to be symmetric.

In fact, [26, Thm. 9.67] contains only the first part of Lemma 4.1. But the second part can be obtained from its proof. To see this, let $\psi^\nu:\mathbb{R}^n\to\mathbb{R}$, $\nu=1,2,\ldots$, be nonnegative, measurable, and bounded with $\int_{\mathbb{R}^n}\psi^\nu(z)dz=1$, and the sets $\mathbb{B}^\nu:=\{z\in\mathbb{R}^n|\ \psi^\nu(z)>0\}$ form a bounded sequence that converges to $\{0\}$. Let C be a compact set contained in Ω. We assume $\mathbb{B}^\nu+C\subseteq\Omega$. Define

$$f^\nu(x):=\int_{\mathbb{R}^n} f(x-z)\psi^\nu(z)dz=\int_{\mathbb{B}^\nu} f(x-z)\psi^\nu(z)dz.$$

We observe that for $x\in\Omega$

$$\nabla f^\nu(x)=\int_{\mathbb{B}^\nu}\nabla f(x-z)\psi^\nu(z)dz.$$

Then the proof argument of [26, Thm. 9.67] can be applied to the functions ∇f^ν and ∇f, establishing that $\{\nabla f^\nu\}$ converge uniformly to ∇f on any compact set C contained in Ω and (13) holds. Suppose f is symmetric. We further assume that the measurable functions $\{\psi^\nu\}$ are symmetric; it follows from the symmetry of f and ψ^ν that for any $P\in\mathcal{P}$

$$f^\nu(Px)=\int_{\mathbb{R}^n} f(Px-z)\psi^\nu(z)dz=\int_{\mathbb{R}^n} f(x-P^Tz)\psi^\nu(P^Tz)d(P^Tz)=f^\nu(x),$$

i.e., $\{f^\nu\}$ are also symmetric.

To present Lewis and Sendov's result, we suppose the symmetric function $f:\mathbb{R}^n\mapsto\mathbb{R}$ is twice differentiable at some points. Letting $\mu\in\mathbb{R}^n$ be such a point, we define a matrix map $\mathcal{A}(\cdot)$ mapping μ to an $n\times n$ matrix:

$$(\mathcal{A}(\mu))_{ij}:=\begin{cases}0 & \text{if } i=j,\\ \big(\nabla^2 f(\mu)\big)_{ii}-\big(\nabla^2 f(\mu)\big)_{ij} & \text{if } i\ne j \text{ and } \mu_i=\mu_j,\\ \dfrac{(\nabla f(\mu))_i-(\nabla f(\mu))_j}{\mu_i-\mu_j} & \text{else.}\end{cases} \tag{14}$$

The following results are [16, Thms. 3.3, 4.2].

LEMMA 4.2. *For any $X \in \mathcal{S}$, $(f \circ \lambda)$ is twice (continuously) differentiable at X if and only if f is twice (continuously) differentiable at $\lambda(X)$. Moreover, in this case the Hessian of the spectral function at X is*

$$\nabla^2(f \circ \lambda)(X)[H] = U(\mathrm{Diag}[\nabla^2 f(\lambda(X))\,\mathrm{diag}[\tilde{H}]] + \mathcal{A}(\lambda(X)) \circ \tilde{H})U^T, \ \forall H \in \mathcal{S}, \tag{15}$$

where U is any orthogonal matrix in $\mathcal{O}_X$ and $\tilde{H} = U^T H U$.

If f is twice continuously differentiable in a neighborhood of $\lambda(X)$, say, $\mathcal{N}(\lambda(X), \epsilon)$ for some $\epsilon > 0$, and

$$\|\nabla^2 f(\mu)\| \le \kappa \tag{16}$$

for any μ in this neighborhood for some $\kappa > 0$, then, according to Lemmas 2.4 and 4.2, $(f \circ \lambda)$ is twice continuously differentiable in the neighborhood $\mathcal{B}(X, \epsilon)$ of X and for any Y in this neighborhood

$$\begin{aligned}
&\||\nabla^2(f \circ \lambda)(Y)|\| = \sup_{\|H\|=1} \|\nabla^2(f \circ \lambda)(Y)(H)\| \\
&= \sup_{\|H\|=1} \|U(\mathrm{Diag}[\nabla^2 f(\lambda(Y))\,\mathrm{diag}[U^T H U]] + \mathcal{A}(\lambda(Y)) \circ (U^T H U))U^T\| \\
&= \sup_{\|H\|=1} \|\mathrm{Diag}[\nabla^2 f(\lambda(Y))\,\mathrm{diag}[U^T H U]] + \mathcal{A}(\lambda(Y)) \circ (U^T H U)\| \\
&\le \sup_{\|H\|=1} \|\mathrm{Diag}[\nabla^2 f(\lambda(Y))\,\mathrm{diag}[U^T H U]]\| + \sup_{\|H\|=1} \|\mathcal{A}(\lambda(Y)) \circ (U^T H U)\| \\
&\le \bar{\kappa} \sup_{\|H\|=1} \|U^T H U\| = \bar{\kappa}, \qquad (17)
\end{aligned}$$

for some $\bar{\kappa} > 0$ which depends only on κ. Here we use the facts $\lambda(Y) \in \mathcal{N}(\lambda(X), \epsilon)$, (16), and the twice continuous differentiability of f.

Now we present our first main result on LC^1 spectral functions.

PROPOSITION 4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable in an open set $\Omega \subseteq \mathbb{R}^n$. Let $X \in \mathcal{S}$ with $\lambda(X) \in \Omega$. The following results hold.*

(a) *$\nabla(f \circ \lambda)$ is strictly continuous at X if and only if ∇f is strictly continuous at $\lambda(X)$.*

(b) *$(f \circ \lambda)$ is an LC^1 function in $\mathcal{S}$ if and only if f is an LC^1 function in $\mathbb{R}^n$.*

Proof. For any $\epsilon > 0$ such that $\mathcal{N}(\lambda(X), \epsilon) \subset \Omega$, it is noted that f is differentiable at every point in $\mathcal{N}(\lambda(X), \epsilon)$ and $(f \circ \lambda)$ is also differentiable at every point in $\mathcal{B}(X, \epsilon)$ by Lemmas 2.4 and 3.3.

(a) Suppose that ∇f is strictly continuous at $\lambda(X)$. Then there exist scalars $\kappa > 0$ and $\delta > 0$ such that

$$\|\nabla f(y) - \nabla f(z)\| \le \kappa \|y - z\| \qquad \forall\, y, z \in \mathcal{N}(\lambda(X), \delta) \subset \Omega.$$

We note that $\lambda(Y) \in \mathcal{N}(\lambda(X), \delta)$ for any $Y \in \mathcal{B}(X, \delta)$. By letting $C := \mathcal{N}(\lambda(X), \delta)$ in Lemma 4.1, there exists a sequence of twice continuously differentiable and symmetric functions $f^\nu : \mathbb{R}^n \to \mathbb{R}$, $\nu = 1, 2, \ldots$, satisfying that ∇f^ν converges uniformly to ∇f on C and

$$\|\nabla^2 f^\nu(\xi)\| \le \kappa \qquad \forall\, \xi \in C, \ \forall \nu. \tag{18}$$

By Lemma 4.2, we know that each $(f^\nu \circ \lambda)$ is twice continuously differentiable. Letting $Y \in \mathcal{B}(X, \delta)$, it follows from Lemma 3.3 that for any $P \in \mathcal{O}_Y$ we have

$$\begin{aligned}
\|\nabla(f^\nu \circ \lambda)(Y) - \nabla(f \circ \lambda)(Y)\| &= \|P\mathrm{Diag}[\nabla f^\nu(\lambda(Y))]P^T - P\mathrm{Diag}[\nabla f(\lambda(Y))]P^T\| \\
&= \|\mathrm{Diag}[\nabla f^\nu(\lambda(Y)) - \nabla f(\lambda(Y))]\|,
\end{aligned}$$

where we use $PP^T = I$ and the properties of the Frobenius norm. Since $\{\nabla f^\nu\}_1^\infty$ converge uniformly to ∇f on C, this shows that $\{\nabla(f^\nu \circ \lambda)\}_1^\infty$ converge uniformly to $\nabla(f \circ \lambda)$ on $\mathcal{B}(X, \delta)$. Moreover, by repeating arguments for (17) to the function f^ν (noting that $\{f^\nu\}_1^\infty$ are twice continuously differentiable with the bound of (18)), we have for any $Y \in \mathcal{B}(X, \delta)$,

$$|||\nabla^2(f^\nu \circ \lambda)(Y)||| \le \bar{\kappa} \qquad \forall \nu, \tag{19}$$

for some $\bar{\kappa} > 0$, depending only on κ. Fix any $Y, Z \in \mathcal{B}(X, \delta)$ with $Y \neq Z$. Since $\{\nabla(f^\nu \circ \lambda)\}_1^\infty$ converges uniformly to $\nabla(f \circ \lambda)$ on $\mathcal{B}(X, \delta)$, for any $\epsilon > 0$ there exists an integer $\nu_1 > 0$ such that for all $\nu > \nu_1$ we have

$$\|\nabla(f^\nu \circ \lambda)(W) - \nabla(f \circ \lambda)(W)\| \le \epsilon \|Y - Z\| \qquad \forall W \in \mathcal{B}(X, \delta).$$

Then by (19) and the mean value theorem for continuously differentiable functions, we have

$$\begin{aligned}
&\|\nabla(f \circ \lambda)(Y) - \nabla(f \circ \lambda)(Z)\| \\
&= \|\nabla(f \circ \lambda)(Y) - \nabla(f^\nu \circ \lambda)(Y) + \nabla(f^\nu \circ \lambda)(Y) - \nabla(f^\nu \circ \lambda)(Z) \\
&\quad + \nabla(f^\nu \circ \lambda)(Z) - \nabla(f \circ \lambda)(Z)\| \\
&\le \|\nabla(f \circ \lambda)(Y) - \nabla(f^\nu \circ \lambda)(Y)\| + \|\nabla(f^\nu \circ \lambda)(Y) - \nabla(f^\nu \circ \lambda)(Z)\| \\
&\quad + \|\nabla(f^\nu \circ \lambda)(Z) - \nabla(f \circ \lambda)(Z)\| \\
&\le 2\epsilon \|Y - Z\| + \| \int_0^1 \nabla^2(f^\nu \circ \lambda)(Y + \tau(Y - Z))(Y - Z) d\tau \| \\
&\le (\bar{\kappa} + 2\epsilon)\|Y - Z\| \quad \forall \ \nu > \nu_1.
\end{aligned}$$

Since $Y, Z \in \mathcal{B}(X, \delta)$ and ϵ are arbitrary, and by letting $\nu \to \infty$, this yields

$$\|\nabla(f \circ \lambda)(Y) - \nabla(f \circ \lambda)(Z)\| \le \bar{\kappa}\|Y - Z\| \quad \forall\, Y, Z \in \mathcal{B}(X, \delta).$$

Thus $\nabla(f \circ \lambda)$ is strictly continuous at X.

Suppose instead that $\nabla(f \circ \lambda)$ is strictly continuous at X. Then the strict continuity of f follows from restricting $(f \circ \lambda)$ to the subspace of diagonal matrices and using formula (7).

(b) is an immediate consequence of (a) by choosing $\Omega = \mathbb{R}^n$. □

In addition to the LC^1 property, another prerequisite for being an SC^1 function is the directional differentiability of the gradient map. The following result concerns this prerequisite.

PROPOSITION 4.4. *Suppose f is differentiable on an open set $\Omega \subseteq \mathbb{R}^n$. Let $X \in \mathcal{S}$ and $\lambda(X) \in \Omega$. Then the following results hold.*

(a) *$\nabla(f \circ \lambda)$ is directionally differentiable at X provided that ∇f is semidifferentiable at $\lambda(X)$.*

(b) *∇f is directionally differentiable at $\lambda(X)$ if $\nabla(f \circ \lambda)$ is directionally differentiable at X.*

Proof. We emphasize again that for any $\epsilon > 0$ such that $\mathcal{N}(\lambda(X), \epsilon) \subset \Omega$, $(f \circ \lambda)$ is differentiable at every point in $\mathcal{B}(X, \epsilon)$. Fix such an ϵ. In the following, we will consider point $X + \tau H$ for $\tau \in \mathbb{R}$ and $H \in \mathcal{S}$. Then $X + \tau H \in \mathcal{B}(X, \epsilon)$ for all small $|\tau|$. Hence, f and $(f \circ \lambda)$ are differentiable at $\lambda(X + \tau H)$ and $X + \tau H$, respectively, for all small $|\tau|$.

(a) Let $H \in \mathcal{S}$, and define $X(\tau) = X + \tau H$, $\tau \in \mathbb{R}$. Then by Lemma 2.5 there exists $P(\tau) \in \mathcal{O}$, $\tau \in \mathbb{R}$, whose entries are power series in τ convergent in a neighborhood $\mathcal{I}$ of $\tau = 0$, and $P^T(\tau)X(\tau)P(\tau)$ is diagonal. Then the corresponding eigenvalues

$$\mu_i(\tau) := [P^T(\tau)X(\tau)P(\tau)]_{ii}, \qquad i = 1, \ldots, n,$$

are also power series in τ, convergent for $\tau \in \mathcal{I}$. Denote $\mu(\tau) := (\mu_1(\tau), \ldots, \mu_n(\tau))^T$. Then we have the expansions

$$\mu(\tau) = \mu(0) + \tau\mu'(0) + o(\tau) \quad \text{and} \quad P(\tau) = P(0) + \tau P'(0) + o(\tau).$$

We note that $\mu(0) = Q\lambda(X)$ for some $Q \in \mathcal{P}$. Hence ∇f is semidifferentiable at $\mu(0)$ by Lemma 3.1(b). In particular, we have

$$\nabla f(\mu(\tau)) = \nabla f(\mu(0) + \tau\mu'(0) + o(\tau)) = \nabla f(\mu(0)) + \tau(\nabla f)'(\mu(0); \mu'(0)) + o(\tau).$$

Then from those expansions above and the formula (7) we have

$$\begin{aligned}
&\nabla(f \circ \lambda)(X + \tau H) - \nabla(f \circ \lambda)(X) \\
&= P(\tau)\mathrm{Diag}[\nabla f(\mu(\tau))]P^T(\tau) - P(0)\mathrm{Diag}[\nabla f(\mu(0))]P^T(0) \\
&= \tau\left(P(0)\mathrm{Diag}[(\nabla f)'(\mu(0); \mu'(0))]P^T(0) + P(0)\mathrm{Diag}[\nabla f(\mu(0))](P'(0))^T\right. \\
&\quad \left. + P'(0)\mathrm{Diag}[\nabla f(\mu(0))]P(0)^T\right) + o(\tau).
\end{aligned}$$

Hence

$$\begin{aligned}
&\lim_{\tau \searrow 0} \left(\nabla(f \circ \lambda)(X + \tau H) - \nabla(f \circ \lambda)(X)\right)/\tau \\
&= P(0)\mathrm{Diag}[(\nabla f)'(\mu(0); \mu'(0))]P^T(0) + P(0)\mathrm{Diag}[\nabla f(\mu(0))](P'(0))^T \\
&\quad + P'(0)\mathrm{Diag}[\nabla f(\mu(0))]P^T(0).
\end{aligned}$$

This implies that the directional derivative $(\nabla(f \circ \lambda))'(X; H)$ is well defined.

(b) Suppose now that $\nabla(f \circ \lambda)$ is directionally differentiable at X. Then the directional differentiability of f follows again from restricting $(f \circ \lambda)$ to the subspace of diagonal matrices and using formula (7). □

Our last main result is on SC^1 property of spectral functions.

PROPOSITION 4.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set Ω in $\mathbb{R}^n$. Let $X \in \mathcal{S}$ with $\lambda(X) \in \Omega$. Then the following results hold.*

(a) *∇f is semismooth at $\lambda(X)$ if and only if $\nabla(f \circ \lambda)$ is semismooth at X. If ∇f is ρ-order semismooth at $\lambda(X)$ $(0 < \rho < \infty)$, then $\nabla(f \circ \lambda)$ is $\min\{1, \rho\}$-order semismooth at X.*

(b) *$(f \circ \lambda)$ is an SC^1 function in $\mathcal{S}$ if and only if f is an SC^1 function in $\mathbb{R}^n$.*

Proof. First we note that there exist $\eta > 0$ and $\epsilon > 0$ such that both f and $(f \circ \lambda)$ are differentiable in $\mathcal{N}(\lambda(X), \epsilon)$ and $\mathcal{B}(X, \epsilon)$, respectively, and Lemma 2.3 holds for all $Y \in \mathcal{B}(X, \epsilon)$. For simplicity, we let $F(\cdot) = \nabla f(\cdot)$.

(a) Suppose F is semismooth at $\lambda(X)$. Then F is semidifferentiable, strictly continuous, and directionally differentiable at $\lambda(X)$. By Propositions 4.4 and 4.3, $\nabla(f \circ \lambda)$ is directionally differentiable at X and locally Lipschitz continuous in a neighborhood $\mathcal{B}(X, \delta)$ for some $\delta \le \epsilon$. Let $\mathcal{D} := \{Y \in \mathcal{B}(X, \delta) |\ \nabla(f \circ \lambda) \text{ is differentiable at } Y\}$ and $\lambda := \lambda(X)$. By taking ϵ smaller if necessary, we can assume that $\epsilon < (\lambda_i - \lambda_{i+1})/2$

whenever $\lambda_i \neq \lambda_{i+1}$. We will show that, for any $H \in \mathcal{S}$ with $X+H \in \mathcal{D}$ and $\|H\| \leq \epsilon$, we have

$$\nabla(f \circ \lambda)(X+H) - \nabla(f \circ \lambda)(X) - \nabla^2(f \circ \lambda)(X+H)H = o(\|H\|). \tag{20}$$

Then, it follows from Lemma 2.2 that $\nabla(f \circ \lambda)$ is semismooth at X. Let $\mu := \lambda(X+H)$, and choose any $Q \in \mathcal{O}_{X+H}$; then there exists $P \in \mathcal{O}_X$ satisfying

$$\|P - Q\| \leq \eta \|H\|.$$

Since $X + H \in \mathcal{D}$, Lemma 4.2 implies ∇f is differentiable at μ. For simplicity, let R denote the left-hand side of (20); then we have from (7) and (15) that

$$\begin{aligned} R &:= \nabla(f \circ \lambda)(X+H) - \nabla(f \circ \lambda)(X) - \nabla^2(f \circ \lambda)(X+H)H \\ &= Q\mathrm{Diag}[\nabla f(\mu)]Q^T - P\mathrm{Diag}[\nabla f(\lambda)]P^T \\ &\quad - Q(\mathrm{Diag}[\nabla^2 f(\mu)\,\mathrm{diag}[Q^T H Q]] + \mathcal{A}(\mu) \circ (Q^T H Q))Q^T. \end{aligned}$$

Once again for simplicity, we let

$$\tilde{R} := Q^T R Q,\ \ \tilde{H} := Q^T H Q,\ \ A := \mathrm{Diag}[F(\mu)],\ \ B := \mathrm{Diag}[F(\lambda)],\ \ D := P^T Q,\ \ C := \mathcal{A}(\mu).$$

Consequently we have

$$\tilde{R} = A - D^T B D - \mathrm{Diag}[\nabla F(\mu)\,\mathrm{diag}[\tilde{H}]] - C \circ \tilde{H}. \tag{21}$$

Since $\mathrm{Diag}[\mu_1, \ldots, \mu_n] = Q^T(X+H)Q = D^T \mathrm{Diag}[\lambda_1, \ldots, \lambda_n]D + \tilde{H}$, we have

$$\sum_{k=1}^n D_{ki} D_{kj} \lambda_k + \tilde{H}_{ij} = \begin{cases} \mu_i & \text{if } i = j, \\ 0 & \text{else,} \end{cases} \qquad i, j = 1, \ldots, n. \tag{22}$$

Since $D = P^T Q = (P-Q)^T Q + I$ and $\|P - Q\| \leq \eta \|H\|$, it follows that

$$D_{ij} = O(\|H\|) \quad \forall i \neq j. \tag{23}$$

Since $P, Q \in \mathcal{O}$, we have $D \in \mathcal{O}$ so that $D^T D = I$. This implies

$$1 = D_{ii}^2 + \sum_{k \neq i} D_{ki}^2 = D_{ii}^2 + O(\|H\|^2), \quad i = 1, \ldots, n, \tag{24}$$

$$0 = D_{ii} D_{ij} + D_{ji} D_{jj} + \sum_{k \neq i,j} D_{ki} D_{kj} = D_{ii} D_{ij} + D_{ji} D_{jj} + O(\|H\|^2) \quad \forall i \neq j. \tag{25}$$

We now show that $\tilde{R} = o(\|H\|)$, which, by $\|R\| = \|\tilde{R}\|$, would prove (20). For any $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned} \tilde{R}_{ii} &\overset{(21)}{=} F_i(\mu) - \sum_{k=1}^n D_{ki}^2 F_k(\lambda) - \sum_{j=1}^n ((\nabla F(\mu))_{ij} \tilde{H}_{jj}) \\ &\overset{(22)}{=} F_i(\mu) - \sum_{k=1}^n D_{ki}^2 F_k(\lambda) - \sum_{j=1}^n \left((\nabla F(\mu))_{ij} \left(\mu_j - \sum_{k=1}^n D_{kj}^2 \lambda_k \right) \right) \\ &\overset{(23)}{=} F_i(\mu) - D_{ii}^2 F_i(\lambda) - \sum_{j=1}^n (\nabla F(\mu))_{ij} (\mu_j - D_{jj}^2 \lambda_j) + O(\|H\|^2) \end{aligned}$$

$$\begin{aligned}
&\overset{(24)}{=} F_i(\mu) - (1+O(\|H\|^2))F_i(\lambda)\\
&\quad -\sum_{j=1}^{n}((\nabla F(\mu))_{ij}(\mu_j - (1+O(\|H\|^2))\lambda_j)) + O(\|H\|^2)\\
&= F_i(\mu) - F_i(\lambda) - \sum_{j=1}^{n}(\nabla F(\mu))_{ij}(\mu_j-\lambda_j) + O(\|H\|^2)\\
&= F_i(\mu) - F_i(\lambda) - (\nabla F_i(\mu))^T(\mu-\lambda) + O(\|H\|^2),
\end{aligned}$$

where we use local boundedness of F and ∇F. Since F is semismooth at λ, each of its components is also semismooth at λ. Lemma 2.4 implies that $\|\lambda-\mu\| \le \|H\|$. Then clearly the right-hand side of the preceding relation is $o(\|H\|)$. For any $i,j \in \{1,\dots,n\}$ with $i \ne j$, we have

$$\begin{aligned}
\tilde{R}_{ij} &\overset{(21)}{=} -\sum_{k=1}^{n} D_{ki}D_{kj}F_k(\lambda) - C_{ij}\tilde{H}_{ij}\\
&\overset{(22)}{=} -\sum_{k=1}^{n} D_{ki}D_{kj}F_k(\lambda) + C_{ij}\sum_{k=1}^{n} D_{ki}D_{kj}\lambda_k\\
&\overset{(23)}{=} -(D_{ii}D_{ij}F_i(\lambda) + D_{ji}D_{jj}F_j(\lambda)) + C_{ij}(D_{ii}D_{ij}\lambda_i + D_{ji}D_{jj}\lambda_j) + O(\|H\|^2)\\
&= -((D_{ii}D_{ij}+D_{ji}D_{jj})F_i(\lambda) + D_{ji}D_{jj}(F_j(\lambda)-F_i(\lambda)))\\
&\quad +C_{ij}((D_{ii}D_{ij}+D_{ji}D_{jj})\lambda_i + D_{ji}D_{jj}(\lambda_j-\lambda_i)) + O(\|H\|^2)\\
&\overset{(25)}{=} -D_{ji}D_{jj}(F_j(\lambda)-F_i(\lambda)-C_{ij}(\lambda_j-\lambda_i)) + O(\|H\|^2).
\end{aligned}$$

Thus, if $\lambda_i = \lambda_j$, Lemma 3.1(c) implies that $F_i(\lambda) = F_j(\lambda)$, which with the preceding relation, yields

$$\tilde{R}_{ij} = O(\|H\|^2).$$

If $\lambda_i \ne \lambda_j$, then Lemma 2.4 implies $\|\mu-\lambda\| \le \|H\|$, $|\mu_i-\lambda_i| \le \|H\|$, and $|\mu_j-\lambda_j| \le \|H\|$ so that $|\mu_i-\mu_j| = |\lambda_i-\lambda_j-(\lambda_i-\mu_i)+(\lambda_j-\mu_j)| \ge |\lambda_i-\lambda_j| - 2\|H\| > 2\epsilon - 2\|H\| \ge 0$. Hence $\mu_i \ne \mu_j$, so $C_{ij} = (F_j(\mu)-F_i(\mu))/(\mu_j-\mu_i)$ and the preceding relation yield

$$\tilde{R}_{ij} = -D_{ji}D_{jj}\left(F_j(\lambda)-F_i(\lambda)-\frac{F_j(\mu)-F_i(\mu)}{\mu_j-\mu_i}(\lambda_j-\lambda_i)\right)+O(\|H\|^2) = O(\|H\|^2),$$

where the second equality uses (23) and the strict continuity of F_i and F_j at λ, so that $F_i(\mu)-F_i(\lambda) = O(\|\mu-\lambda\|) = O(\|H\|)$ and $F_j(\mu)-F_j(\lambda) = O(\|\mu-\lambda\|) = O(\|H\|)$.

Suppose F is ρ-order semismooth at $\lambda(X)$ $(0<\rho<\infty)$. Then the preceding argument shows that $\tilde{R}_{ii} = O(\max\{\|H\|^{1+\rho}, \|H\|^2\}) = O(\|H\|^{1+\min\{1,\rho\}})$ for all i while we still have $\tilde{R}_{ij} = O(\|H\|^2)$ for all $i \ne j$. This shows that $\nabla(f\circ\lambda)$ is $\min\{1,\rho\}$-order semismooth at X.

Suppose $\nabla(f\circ\lambda)$ is semismooth at X. Then $\nabla(f\circ\lambda)$ is strictly continuous and directionally differentiable at X. By Propositions 4.3 and 4.4, $F := \nabla f$ is strictly continuous and directionally differentiable at $\lambda(X)$. For any $h \in \mathbb{R}^n$ such that F is differentiable at $\lambda(X)+h$, i.e., f is twice differentiable at $\lambda(X)+h$, let $H := Q\mathrm{Diag}[h]Q^T$ for some $Q \in \mathcal{O}_X$. Then there exists $P \in \mathcal{P}$ such that $P(\lambda(X)+h) =$

$\lambda(X+H)$. Lemma 3.1(d) implies that f is twice differentiable at $\lambda(X+H)$. In turn, Lemma 4.2 yields that $\nabla(f\circ\lambda)$ is twice differentiable at $X+H$. We note that

$$Q^T(X+H)Q = \mathrm{Diag}[\lambda(X)+h] = \mathrm{Diag}[P^T\lambda(X+H)] = P^T\mathrm{Diag}[\lambda(X+H)]P,$$

which is equivalent to

$$X+H = QP^T\mathrm{Diag}[\lambda(X+H)]PQ^T = U\mathrm{Diag}[\lambda(X+H)]U^T,$$

where $U := QP^T$, and hence $U\in\mathcal{O}$ since $Q,P\in\mathcal{O}$. For simplicity, let $\mu := \lambda(X+H)$; then we have

$$\begin{aligned} U^THU &= PQ^TQ\mathrm{Diag}[h]Q^TQP^T \\ &= P\mathrm{Diag}[h]P^T = \mathrm{Diag}[Ph] \qquad (\text{using } \ P\in\mathcal{P}) \end{aligned}$$

and

$$\begin{aligned} &\mathrm{Diag}[\nabla^2 f(\mu)\,\mathrm{diag}[U^THU]] = \mathrm{Diag}[\nabla^2 f(\mu)Ph] \\ &= \mathrm{Diag}[\nabla^2 f(P(\lambda(X)+h))Ph] \qquad (\text{using } \ \mu = P(\lambda(X)+h)) \\ &= \mathrm{Diag}[P\nabla^2 f(\lambda(X)+h)P^TPh] \qquad (\text{using Lemma 3.1(d)}) \\ &= \mathrm{Diag}[P\nabla^2 f(\lambda(X)+h)h] \\ &= P\mathrm{Diag}[\nabla^2 f(\lambda(X)+h)h]P^T \qquad (\text{using } \ P\in\mathcal{P}). \end{aligned} \tag{26}$$

Since $\nabla(f\circ\lambda)$ is semismooth at X, it follows from Lemma 2.2 that

$$R := \nabla(f\circ\lambda)(X+H) - \nabla(f\circ\lambda)(X) - \nabla^2(f\circ\lambda)(X+H)H = o(\|H\|),$$

which, by (7), (15), and (26), is equivalent to

$$\begin{aligned} R &= Q\mathrm{Diag}[\nabla f(\lambda(X)+h)]Q^T - Q\mathrm{Diag}[\nabla f(\lambda(X))]Q^T \\ &\quad - U(\mathrm{Diag}[\nabla^2 f(\mu)\,\mathrm{diag}[U^THU]] + \mathcal{A}(\mu)\circ(U^THU))U^T \\ &= Q\mathrm{Diag}[\nabla f(\lambda(X)+h) - \nabla f(\lambda(X)) - \nabla^2 f(\lambda(X)+h)h]Q^T. \end{aligned}$$

The second equality uses $\mathcal{A}(\mu)\circ(U^THU) = \mathcal{A}(\mu)\circ\mathrm{Diag}[Ph] = 0$. We then have

$$\tilde{R} := Q^TRQ = \mathrm{Diag}[\nabla f(\lambda(X)+h) - \nabla f(\lambda(X)) - \nabla^2 f(\lambda(X)+h)h].$$

Since $\|\tilde{R}\| = \|R\|$ and $\|H\| = \|h\|$, the preceding relation means by noting $F = \nabla f$

$$F(\lambda(X)+h) - F(\lambda(X)) - \nabla F(\lambda(X)+h)h = o(\|h\|).$$

This proves that ∇f is semismooth at $\lambda(X)$.

(b) is an immediate consequence of (a) since the choice of X is arbitrary, Ω can be chosen as $\mathbb{R}^n$. □

Remarks. In the special case where $f:\mathbb{R}^n\mapsto\mathbb{R}$ takes the form (6) and $g(\cdot):\mathbb{R}\mapsto\mathbb{R}$ is differentiable, according to Lemma 3.3 we have

$$\nabla(f\circ\lambda)(X) = U\mathrm{Diag}[g'(\lambda_1),\ldots,g'(\lambda_n)]U^T, \tag{27}$$

where $U\in\mathcal{O}_X$ and $\lambda := \lambda(X)$. Associated with this f, we define a symmetric–matrix-valued function $f^\square:\mathcal{S}\mapsto\mathcal{S}$ by

$$f^\square(X) = U\mathrm{Diag}[g'(\mu_1),\ldots,g'(\mu_n)]U^T$$

for any $U \in \mathcal{O}$ satisfying $X = U\mathrm{Diag}[\mu_1, \ldots, \mu_n]U^T$. It is pointed out in [3] that for this special case

$$f^{\square}(X) = \nabla(f \circ \lambda)(X).$$

Among many results on continuity, differentiability, and nonsmoothness obtained in [3] is the semismoothness of $f^{\square}$. It is proved [3, Prop. 4.10] that $f^{\square}(\cdot)$ is semismooth if and only if $g'(\cdot)$ is semismooth. In other words, for this special case, the SC^1 result (Proposition 4.5) follows from [3, Prop. 4.10]. But for general cases other than (6), we do not have such direct consequences. Nevertheless, the proof here is inspired by [3, Prop. 4.10]. We would also like to point out that the treatment in [3] goes beyond this special case. In fact, given a real function of one dimension $f : \mathbb{R} \mapsto \mathbb{R}$, the symmetric–matrix-valued function defined in [3] is

$$f^{\square}(X) := U\mathrm{Diag}[f(\mu_1), \ldots, f(\mu_n)]U^T,$$

where $U \in \mathcal{O}$ satisfying $X = U\mathrm{Diag}[\mu_1, \ldots, \mu_n]U^T$. There are examples where f cannot be derivative of another real function.

5. An example. As an example, we consider the positive trace function $F : \mathcal{S} \mapsto \mathbb{R}$ by

$$F(X) := (\max\{0, \mathrm{trace}(X)\})^2 \qquad \forall X \in \mathcal{S}.$$

Obviously, $F(X) = (f \circ \lambda)(X)$ with $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined by

$$f(x) := \left(\max\left\{0, \sum x_i\right\}\right)^2 \qquad \forall x \in \mathbb{R}^n.$$

It is known that $f(\cdot)$ is continuously differentiable, and its derivative map is strongly semismooth. Hence, we can conclude that $F(\cdot)$ is continuously differentiable [13], and moreover, it is an SC^1 function (Proposition 4.5).

Acknowledgments. We thank the associate editor M. Overton for his suggestion of the paper title, and the two referees for their comments which improved the presentation of the paper significantly. In particular, [19] was brought to our attention by one referee and Proposition 3.2(c) was suggested to us by the other. We also thank Defeng Sun for his comments on this result.

REFERENCES

[1] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.
[2] X. Chen and P. Tseng, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Program., 95 (2003), pp. 431–474.
[3] X. Chen, H.D. Qi, and P. Tseng, *Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems*, SIAM J. Optim., 13 (2003), pp. 960–985.
[4] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[5] C. Davis, *All convex invariant functions of hermitian matrices*, Arch. Mathematik, 8 (1957), pp. 276–278.
[6] F. Facchinei, *Minimization of SC^1 functions and the Maratos effect*, Oper. Res. Lett., 17 (1995), pp. 131–137.
[7] A. Fischer, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Program., 76 (1997), pp. 513–532.
[8] J.-B. Hiriart-Urruty and A.S. Lewis, *The Clarke and Michel–Penot subdifferentials of the eigenvalues of a symmetric matrix*, Comput. Optim. Appl., 13 (1999), pp. 13–23.

[9] J.-B. Hiriart-Urruty, J.J. Strodiot, and V. Hien Nguyen, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
[10] H. Jiang and D. Ralph, *Global and local superlinear convergence analysis of Newton-type methods for semismooth equations with smooth least squares*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Appl. Optim. 22, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 181–209.
[11] C. Kanzow and C. Nagel, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23.
[12] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1984.
[13] A.S. Lewis, *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588.
[14] A.S. Lewis and M.L. Overton, *Eigenvalue optimization*, in Acta Numerica, Acta Numer. 5, Cambridge University Press, Cambridge, UK, 1996, pp. 149–190.
[15] A.S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Program., 84 (1999), pp. 1–24.
[16] A.S. Lewis and H.S. Sendov, *Twice differentiable spectral functions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 368–386.
[17] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[18] R. Mifflin, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., 2 (1977), pp. 191–207.
[19] J. von Neumann, *Some matrix inequalities and metrization of matric space*, Tomsk University Review, 1 (1937), pp. 286–300. Also in Collected Works Vol. IV: Continuous Geometry and Other Topics, A. H. Taub, ed., Pergamon Press, Oxford, UK, 1962, pp. 205–218.
[20] J.-S. Pang and L. Qi, *A globally convergent Newton method for convex SC^1 minimization problems*, J. Optim. Theory Appl., 85 (1995), pp. 633–648.
[21] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[22] L. Qi, *Superlinearly convergent approximate Newton methods for LC^1 optimization problems*, Math. Programming, 64 (1994), pp. 277–294.
[23] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
[24] L. Qi and P. Tseng, *An analysis of piecewise smooth functions and almost smooth functions*, Math. Oper. Res., submitted.
[25] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
[26] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[27] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[28] D. Sun and J. Sun, *Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems*, SIAM J. Numer. Anal., 40 (2003), pp. 2352–2367.
[29] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.
[30] X.Q. Yang and X.X. Huang, *A nonlinear Lagrangian approach to constrained optimization problems*, SIAM J. Optim., 11 (2001), pp. 1119–1144.
[31] X.Q. Yang and V. Jeyakumar, *Generalized second-order directional derivatives and optimization with $C^{1,1}$ functions*, Optimization, 26 (1992), pp. 165–185.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 784–803

COMPUTING ISOTYPIC PROJECTIONS WITH THE LANCZOS ITERATION*

DAVID K. MASLEN†, MICHAEL E. ORRISON‡, AND DANIEL N. ROCKMORE§

Abstract. When the isotypic subspaces of a representation are viewed as the eigenspaces of a symmetric linear transformation, isotypic projections may be achieved as eigenspace projections and computed using the Lanczos iteration. In this paper, we show how this approach gives rise to an efficient isotypic projection method for permutation representations of distance transitive graphs and the symmetric group.

Key words. isotypic projection, Lanczos iteration, separating set

AMS subject classifications. 20C40, 65F10

DOI. 10.1137/S0895479801399778

1. Introduction.

Let G be a finite group acting on a finite set X. Let $L(X)$ be the vector space of complex valued functions on X. The action of G on X gives rise to a *permutation representation* ρ of G defined on $L(X)$ by

$$(\rho(g)(f))(x) = f(g^{-1}x)$$

for all $g \in G$, $f \in L(X)$, and $x \in X$. Because $L(X)$ is a representation of G, there is a basis independent decomposition

$$L(X) = V_1 \oplus \cdots \oplus V_n$$

of $L(X)$ into G-invariant subspaces known as *isotypic subspaces*. The problem addressed in this paper is the following: Given an arbitrary $f \in L(X)$, how may we efficiently compute the projection of f onto each isotypic subspace of $L(X)$?

The problem of computing projections onto isotypic subspaces arises in *spectral analysis* which is a nonmodel-based approach to the analysis of data that may be viewed as a complex valued function f on a set X that has an underlying symmetry group G. Developed by Diaconis [5, 6], the subject extends the classical spectral analysis of time series and requires the computation of projections of f onto subsets of G-invariant subspaces of $L(X)$.

As an example, let X be the set $\{x_0, \ldots, x_{n-1}\}$ and let G be the cyclic group $\mathbb{Z}/n\mathbb{Z}$ acting on X by cyclicly permuting its elements. The elements of $L(X)$ may be viewed as *signals* on n points and the isotypic subspaces of $L(X)$ as corresponding to the different *frequencies* that make up these signals. The isotypic projections of $f \in L(X)$ may be computed with the aid of the usual discrete Fourier transform (DFT). The classical fast Fourier transform (FFT) may therefore be used to efficiently compute the projections of f onto the isotypic subspaces of $L(X)$ (see, e.g., [13]).

*Received by the editors December 14, 2001; accepted for publication (in revised form) by S. A. Vavasis June 16, 2003; published electronically February 24, 2004.

http://www.siam.org/journals/simax/25-3/39977.html

†Susquehanna Partners GP, 401 City Line Ave., Bala Cynwyd, PA 19004 (david@maslen.net).

‡Department of Mathematics, Harvey Mudd College, Claremont, CA 91711 (orrison@hmc.edu). The research of this author was supported by AFOSR.

§Department of Mathematics, Dartmouth College, Hanover, NH 03755 (daniel.rockmore@dartmouth.edu). The research of this author was supported by NSF-DMS, AFOSR, and the Santa Fe Institute.

As another example, suppose voters are asked to rank k candidates in order of preference. The set X is then the set of orderings of the k candidates and G is the symmetric group S_k whose natural action on the set of candidates induces an action on the set of orderings. If $f \in L(X)$ is such that $f(x)$ is the number of voters choosing the ordering x, then there are natural statistics associated to f. For example, the *mean response* of f is the value $(1/|X|)\sum_{x\in X} f(x)$, whereas a *first order summary* of f counts the number of voters that ranked candidate i in position j. Similarly, there are *higher order summaries* associated to f. For example, we could compute the number of voters that ranked candidates i and j in positions k and l, either respectively or so that order does not matter. These higher order summaries, however, contain redundant information. Removing this redundant information, or finding the *pure higher order effects* of f, is equivalent to computing the isotypic projections of f (see [6, 17]).

A naive approach (see, e.g., [19]) to computing the n isotypic projections of $f \in L(X)$ requires $O(n|G||X|)$ operations where we count a complex multiplication followed by a complex addition as one *operation*. Diaconis and Rockmore [7] show that a careful reorganization of this approach reduces the number of necessary operations to $O(n|X|^2)$. The advantage of their approach is that it relies only on the knowledge of the characters of G. In terms of operation counts, however, the number of operations required by a direct matrix multiplication approach is also $O(n|X|^2)$, which has prompted the search for other approaches to computing isotypic projections. For example, Driscoll, Healy, and Rockmore [8] show that if X is a distance transitive graph, then fast discrete polynomial transforms may be used to compute the n isotypic projections of $f \in L(X)$ with at most $O(|X|^2 + |X|n\log^2 n)$ operations. This bound, however, assumes the use of exact arithmetic. Stability issues arise when their algorithm is implemented using finite precision arithmetic.

In this paper, we develop an approach to computing isotypic projections that relies on a method for computing projections onto the eigenspaces of a collection of simultaneously diagonalizable linear transformations. We call the collections of transformations that we use *separating sets* because they allow us to *separate* a representation into its isotypic components. The approach may be seen as a generalization of the Gentleman–Sande, or *decimation in frequency*, FFT in that we too will be iteratively computing projections of projections (see [10]). Such collections have also been used in [3], for example, where certain separating sets are known as *complete sets of commuting operators*.

As a simple example of how a separating set is used to compute isotypic projections, suppose that $L(X)$ has three isotypic subspaces V_1, V_2, and V_3. Thus $L(X) = V_1 \oplus V_2 \oplus V_3$ and each $f \in L(X)$ may be written uniquely as $f = f_1 + f_2 + f_3$, where $f_i \in V_i$. Additionally, suppose that T and T' are diagonalizable linear transformations on $L(X)$ such that the eigenspaces of T are $V_1 \oplus V_2$ and V_3, and the eigenspaces of T' are V_1 and $V_2 \oplus V_3$. As we shall see, $\{T, T'\}$ is a separating set for $L(X)$. We may therefore compute the f_i by first projecting f onto the eigenspaces of T to compute $f_1 + f_2$ and f_3, and then projecting both $f_1 + f_2$ and f_3 onto the eigenspaces of T' to compute f_1, f_2, and f_3. Note that each computation is done with respect to a fixed basis of $L(X)$. This process of decomposing $L(X) = V_1 \oplus V_2 \oplus V_3$ is illustrated in Figure 1.

The efficiency of this approach depends on an efficient eigenspace projection method. Since the separating sets we use consist of real symmetric matrices, we look to the *Lanczos iteration* for such a method. This is an algorithm that may

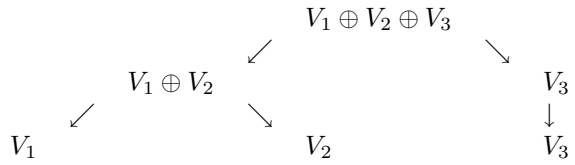


FIG. 1. *Decomposing* $L(X) = V_1 \oplus V_2 \oplus V_3$ *using* T *and* T'.

be used to efficiently compute the eigenspace projections of a real symmetric matrix when, as in all of our examples, it has relatively few eigenspaces and when it may be applied efficiently to arbitrary vectors, either directly or through a given subroutine (see, e.g., [16]).

We proceed as follows. In section 2, we describe the isotypic decomposition of a representation and introduce the idea of a separating set of diagonalizable linear transformations. In section 3, we show how an eigenspace approach to computing isotypic projections for cyclic groups leads to the Gentleman–Sande FFT. In section 4, we review how the Lanczos iteration may be used to compute the projections of a vector onto the eigenspaces of a real symmetric matrix. We then use the results of section 2 to create an isotypic projection method. This method is then shown to be efficient for certain permutation representations of distance transitive graphs in section 5 and the symmetric group in section 6.

2. Isotypic subspaces. In this section, we describe the isotypic decomposition of a representation and we introduce the idea of a separating set of diagonalizable linear transformations. We then show how these separating sets may be used to compute isotypic projections. A good reference for representations of finite groups is [19].

2.1. Complex representations. Let G be a finite group, let V be a finite dimensional vector space over $\mathbb{C}$, and let $GL(V)$ be the group of automorphisms of V. A *representation* of G is a homomorphism $\rho : G \to GL(V)$. If the homomorphism ρ is understood, then we also say that V is a representation of G. The *character* of ρ is the function $\chi : G \to \mathbb{C}$, where $\chi(g)$ is the usual trace of $\rho(g)$. Note that the character of a representation of G is constant on the conjugacy classes of G.

A subspace W of V is *invariant* if $\rho(g)(w) \in W$ for all $g \in G$, $w \in W$. A representation is said to be *simple* if it contains no nontrivial invariant subspaces. If $C_1, \ldots, C_h$ are the distinct conjugacy classes of G, then there are h distinct (up to isomorphism) simple representations $W_1, \ldots, W_h$ of G. Let d_i be the dimension of W_i, let χ_i be the character of W_i, and let $\chi_i(C_j)$ be the value of χ_i on C_j.

2.2. The isotypic decomposition. Every representation of G is a direct sum of simple representations. Thus, V is a direct sum of simple representations, say, $U_1, \ldots, U_l$. Denote by V_i the direct sum of those $U_1, \ldots, U_l$ that are isomorphic to W_i. Removing the trivial V_i (and renumbering if necessary) creates the *isotypic decomposition*

$$V = V_1 \oplus \cdots \oplus V_n,$$

where each V_i is then an *isotypic subspace* of V. Each $v \in V$ may therefore be written uniquely as

$$v = v_1 + \cdots + v_n,$$

where $v_i \in V_i$ is called the *isotypic projection* of v onto the isotypic subspace V_i. The isotypic decomposition of V is independent of the choice of U_j.

THEOREM 2.1. *The projection p_i of V onto V_i along $\oplus_{j\neq i}V_j$ is given by the formula*

$$p_i = \frac{d_i}{|G|}\sum_{g\in G}\chi_i(g)^*\rho(g).$$

Proof. See, for example, Theorem 8 in [19]. ∎

By Theorem 2.1, the isotypic projection v_i may be computed by directly applying p_i to v. There are, however, drawbacks to this approach. First, directly applying p_i to an arbitrary vector in V requires $O(\dim(V)^2)$ operations which may be prohibitive if $\dim(V)$ is large. Second, to construct p_i using the above formula requires a sum over the group G as well as an explicit knowledge of the representations of each element of G. This too may be prohibitive if G is large.

2.3. Separating sets. Suppose now that $\{T_1, \ldots, T_k\}$ is a collection of diagonalizable linear transformations on V whose eigenspaces are direct sums of the isotypic subspaces of V. For each isotypic subspace V_i, let $c_i = (\mu_{i1}, \ldots, \mu_{ik})$ be the k-tuple of eigenvalues where, for $1 \leq j \leq k$, μ_{ij} is the eigenvalue of T_j associated to V_i. If $c_i \neq c_{i'}$ whenever $V_i \neq V_{i'}$, then we say that $\{T_1, \ldots, T_k\}$ is a *separating set* for V.

The existence of a separating set $\{T_1, \ldots, T_k\}$ for V means that the computation of the isotypic projections of $v \in V$ can be achieved through a series of eigenspace projections:

Stage 1. Compute the projections of v onto each of the eigenspaces of T_1.

Stage 2. Compute the projections of each of the previously computed projections onto each of the eigenspaces of T_2.

$$\vdots$$

Stage k. Compute the projections of each of the previously computed projections onto each of the eigenspaces of T_k.

LEMMA 2.2. *The computed projections at Stage k are precisely the isotypic projections of the vector v.*

Proof. The projections at each stage are sums of the isotypic projections of v. If a projection at Stage k was the sum of two or more isotypic projections, then the corresponding isotypic subspaces must have been in the same eigenspace for each of the T_j. This, however, would contradict the fact that $\{T_1, \ldots, T_k\}$ is a separating set for V. ∎

We may easily find separating sets for V by looking to the conjugacy classes $C_1, \ldots, C_h$ of G. In particular, if $T_j = \sum_{c\in C_j}\rho(c)$ is the *class sum* of C_j (with respect to ρ) and $\mu_{ij} = |C_j|\chi_i(C_j)/d_i$, then we have the following lemma.

LEMMA 2.3. *The class sum T_j is a diagonalizable linear transformation on V whose eigenspaces are direct sums of isotypic subspaces, and μ_{ij} is the eigenvalue of T_j that is associated to the isotypic subspace V_i.*

Proof. This is a variation of Proposition 6 in [19]. ∎

The complete collection of class sums forms a separating set of V. In fact, by Theorem 2.1, every separating set for V is composed of linear combinations of class sums. We may, however, be able to find much smaller separating sets than the complete collection of class sums.

2.4. Permutation representations. Suppose now that G acts on a finite set X. Let $L(X)$ be the vector space of complex valued functions on X. The action of G on X induces a *permutation representation* $\rho : G \to GL(L(X))$ defined by

$$(\rho(g)(f))(x) = f(g^{-1}x)$$

for all $g \in G$, $f \in L(X)$, and $x \in X$. The vector space $L(X)$ has a natural basis $\{\delta_x\}_{x\in X}$, where

$$\delta_x(x') = \begin{cases} 1 & \text{if } x = x', \\ 0 & \text{otherwise.} \end{cases}$$

We will refer to $\{\delta_x\}_{x\in X}$ as the *delta basis* of $L(X)$. Note that $\dim(L(X)) = |X|$.

By choosing a basis for $L(X)$, we may identify each linear transformation on $L(X)$ with an $|X| \times |X|$ matrix. Thus, we will assume that each linear transformation on $L(X)$ is written as a matrix with respect to the delta basis of $L(X)$. In particular, if $g \in G$, then $\rho(g)$ corresponds to an $|X| \times |X|$ matrix with one 1 in each row and column, and zeros elsewhere.

3. Cyclic groups. In this section, we show how using separating sets to compute isotypic projections for cyclic groups leads to the Gentleman–Sande, or *decimation in frequency*, FFT (see [10]).

3.1. The DFT and isotypic projections. Let G be the cyclic group $\mathbb{Z}/n\mathbb{Z}$ and let X be the set $\{x_o, \ldots, x_{n-1}\}$. Let ω be a primitive nth root of unity, let g be a generator for G, and let G act on X by setting $g^j x_i = x_{i+j}$, where all subscripts are taken modulo n. The resulting permutation representation

$$\rho : \mathbb{Z}/n\mathbb{Z} \to GL(L(X))$$

has n isotypic subspaces $V_0, \ldots, V_{n-1}$, where each V_i is one-dimensional (and hence simple) with character χ_i defined by $\chi_i(g^j) = \omega^{ij}$.

Each element g^j of G forms a conjugacy class $C_j = \{g^j\}$. The eigenvalue of the class sum T_j of C_j associated to the isotypic subspace V_i is therefore $\chi_i(C_j)/d_i = \chi_i(g^j)/1 = \omega^{ij}$.

Let $f \in L(X)$ and let f_i be the isotypic projection of f onto the isotypic subspace V_i. Since ω is a primitive nth root of unity, the class sum T_1 forms a separating set for $L(X)$. The isotypic projection f_i may therefore be viewed as the projection of f onto the eigenspace of T_1 with eigenvalue ω^i. By Theorem 2.1, this may be computed as

$$f_i = \left(\frac{1}{n} \sum_{j=0}^{n-1} \omega^{-ij} \rho(g^j) \right) f.$$

Note that $f_i(x_0) = \omega^{ik} f_i(x_k)$ and that f_i is therefore determined by

$$f_i(x_0) = \left(\frac{1}{n} \sum_{j=0}^{n-1} \omega^{-ij} \rho(g^j) \right) f(x_0) = \frac{1}{n} \sum_{j=0}^{n-1} \omega^{ij} f(x_j).$$

This, however, is the ith coefficient of the usual DFT applied to f. An FFT on n points may therefore be thought of as an efficient algorithm for computing isotypic projections of vectors in $L(X)$.

3.2. The Gentleman–Sande FFT. Suppose now that $n = pq$. Since $\{T_1\}$ is a separating set for $L(X)$, so is $\{T_1, T_p\}$. We could therefore compute the isotypic projections of f by first computing the projections of f onto the eigenspaces of T_p and then projecting each of these projections onto the eigenspaces of T_1.

The eigenspaces of T_p are $W_0, \ldots, W_{q-1}$, where the eigenvalue of T_p that is associated to W_k is ω^{pk} and

$$W_k = V_k \oplus V_{k+q} \oplus \cdots \oplus V_{k+(p-1)q}.$$

The projection f'_k of f onto W_k is therefore

$$f_k + f_{k+q} + \cdots + f_{k+(p-1)q}. \tag{3.1}$$

In fact, the W_k are the isotypic subspaces of $L(X)$ with respect to the action on X of the subgroup of G that is generated by g^p. This subgroup is cyclic with order q. Thus, by Theorem 2.1,

$$f'_k = \left(\frac{1}{q} \sum_{t=0}^{q-1} \omega^{-pkt} \rho(g^{pt}) \right) f.$$

Note that $f'_k(x_s) = \omega^{pkt} f'_k(x_{s+pt})$ and that f'_k is therefore determined by the values $f'_k(x_0), \ldots, f'_k(x_{p-1})$. In this sense, since $f'_k(x_j)$ requires $O(q)$ operations to compute, f'_k requires $O(pq)$ operations to compute. The projections $f'_0, \ldots, f'_{q-1}$ may therefore be computed using $O(pq^2)$ operations.

Since $n = pq$, each $0 \le i, j \le n-1$ can be uniquely represented as $i = k + lq$ and $j = s + tp$ for some $0 \le k, t \le q-1$ and $0 \le l, s \le p-1$. Moreover, by (3.1), the isotypic projection $f_i = f_{k+lq}$ may be computed by projecting f'_k onto the eigenspace of T_1 with eigenvalue $\omega^{(k+lq)}$. Recall that f_{k+lq} is determined by $f_{k+lq}(x_0)$, which we may compute as

$$\begin{aligned}
f_{k+lq}(x_0) &= \left(\frac{1}{n} \sum_{j=0}^{n-1} \omega^{-(k+lq)j} \rho(g^j) \right) f'_k(x_0) \\
&= \frac{1}{n} \sum_{j=0}^{n-1} \omega^{(k+lq)j} f'_k(x_j) \\
&= \frac{1}{pq} \sum_{s=0}^{p-1} \sum_{t=0}^{q-1} \omega^{(k+lq)(s+tp)} f'_k(x_{s+tp}) \\
&= \frac{1}{p} \sum_{s=0}^{p-1} \omega^{(k+lq)s} \frac{1}{q} \sum_{t=0}^{q-1} \omega^{(k+lq)tp} f'_k(x_{s+tp}) \\
&= \frac{1}{p} \sum_{s=0}^{p-1} \omega^{(k+lq)s} \frac{1}{q} \sum_{t=0}^{q-1} \omega^{pkt} f'_k(x_{s+tp}) \\
&= \frac{1}{p} \sum_{s=0}^{p-1} \omega^{(k+lq)s} \frac{1}{q} \sum_{t=0}^{q-1} f'_k(x_s) \\
&= \frac{1}{p} \sum_{s=0}^{p-1} \left(\omega^{ks} f'_k(x_s) \right) (\omega^q)^{ls}.
\end{aligned}$$

This is a DFT on p points applied to the function $\omega^{ks} f'_k$. Thus, if we have computed $f'_0, \ldots, f'_{q-1}$, we may compute the isotypic projection f_i using $O(p)$ operations. Since there are pq isotypic projections and the f'_k require $O(pq^2)$ operations to compute, we may compute the isotypic projections of $f \in L(X)$ using $O(p^2q + pq^2) = O((p+q)pq)$ operations.

This particular FFT is known as the Gentleman–Sande, or *decimation in frequency*, FFT (see [10]). The approach to decomposing representations that is presented in this paper may be viewed as a generalization of decimation in frequency since we too will be iteratively computing projections of projections.

4. The Lanczos iteration. Given a separating set, isotypic projections become eigenspace projections. In this section, we show how the Lanczos iteration gives rise to an efficient isotypic projection method when the number of isotypic subspaces is relatively small and the linear transformations in a separating set are real symmetric matrices that can be applied efficiently. Good references for the Lanczos iteration are [4, 16, 22, 23].

4.1. Krylov subspaces. Let $\mathbb{C}^N$ be the usual complex vector space of N-tuples with complex coefficients. Let $M_N(\mathbb{C})$ be the set of $N \times N$ matrices with complex coefficients. We will view the elements of $\mathbb{C}^N$ as column matrices of size N. The matrices $M_N(\mathbb{C})$ may therefore be viewed as linear transformations of $\mathbb{C}^N$ with respect to the standard basis of $\mathbb{C}^N$.

Let $T \in M_N(\mathbb{C})$, let T^t denote the transpose of T, and let T^* denote the conjugate transpose of T. If $v, w \in \mathbb{C}^N$, then the usual inner product of v and w is v^*w. The norm of v is $||v|| = (v^*v)^{1/2}$. T is *symmetric* if $T = T^t$ and *hermitian* if $T = T^*$, in which case T is diagonalizable with real eigenvalues.

If $f \in \mathbb{C}^N$, then the jth *Krylov subspace* generated by T and f is the subspace $\mathcal{K}_j$ of $\mathbb{C}^N$ that is spanned by the vectors $f, Tf, \ldots, T^{j-1}f$. We write this as

$$\mathcal{K}_j = \langle f, Tf, \ldots, T^{j-1}f \rangle.$$

The T-invariant subspace $\mathcal{K} = \langle f, Tf, T^2f, \ldots \rangle$ is the *Krylov subspace* generated by T and f. Note that $\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \mathcal{K}_3 \subseteq \cdots$ and that for some m, $\mathcal{K}_m = \mathcal{K}_{m+1} = \cdots = \mathcal{K}$.

Suppose now that $T \in M_N(\mathbb{C})$ is diagonalizable with n distinct eigenvalues. Then

$$\mathbb{C}^N = V_1 \oplus \cdots \oplus V_n,$$

where the V_i are the n distinct eigenspaces of T. Each $f \in \mathbb{C}^N$ may therefore be written uniquely as $f = f_1 + \cdots + f_n$, where $f_i \in V_i$. We say that f_i is the *eigenspace projection of f onto the eigenspace V_i*. By the following lemma, we may restrict our attention to the Krylov subspace generated by T and f when computing these f_i.

LEMMA 4.1. *If $T \in M_N(\mathbb{C})$ is diagonalizable and $f \in \mathbb{C}^N$, then the nontrivial projections of f onto the eigenspaces of T form a basis for the Krylov subspace generated by T and f.*

Proof. Suppose that T has n distinct eigenvalues $\mu_1, \ldots, \mu_n$ and that $f = f_1 + \cdots + f_n$, where f_i is the projection of f onto the eigenspace corresponding to the

eigenvalue μ_i. We then have the following system of equations:

$$\begin{aligned} f &= f_1 + f_2 + \cdots + f_n, \\ Tf &= \mu_1 f_1 + \mu_2 f_2 + \cdots + \mu_n f_n, \\ T^2 f &= \mu_1^2 f_1 + \mu_2^2 f_2 + \cdots + \mu_n^2 f_n, \\ &\vdots \\ T^{n-1} f &= \mu_1^{n-1} f_1 + \mu_2^{n-1} f_2 + \cdots + \mu_n^{n-1} f_n. \end{aligned} \tag{4.1}$$

The coefficients of the f_i in (4.1) form a Vandermonde matrix

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \mu_1^2 & \mu_2^2 & \cdots & \mu_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1^{n-1} & \mu_2^{n-1} & \cdots & \mu_n^{n-1} \end{pmatrix}$$

which is invertible since the μ_i are distinct (see, e.g., [9]). We may therefore solve the system for the f_i in terms of the $T^j f$. This shows that each f_i is contained in $\mathcal{K} = \langle f, Tf, T^2 f, \dots \rangle$. On the other hand, any power of T applied to f is a linear combination of the f_i. Thus $\mathcal{K}$ is spanned by the f_i. Since the nontrivial f_i are linearly independent, the lemma follows. ∎

COROLLARY 4.2. *The dimension of $\mathcal{K} = \langle f, Tf, T^2 f, \dots \rangle$ is equal to the number of nontrivial projections of f onto the eigenspaces of T.*

COROLLARY 4.3. *Eigenvectors of the restriction of T to $\mathcal{K}$ are scalar multiples of the eigenspace projections of f.*

Proof. This follows from the fact that each eigenspace of the restriction of T to $\mathcal{K}$ is one-dimensional and is spanned by one of the nontrivial projections of f onto the eigenspaces of T. ∎

If u is an eigenvector of the restriction of T to $\mathcal{K}$, then we may scale u into an eigenspace projection of f by Corollary 4.3. If the eigenspaces of the restriction of T to $\mathcal{K}$ are orthogonal, this may be computed as

$$\frac{u^* f}{u^* u} u. \tag{4.2}$$

Moreover, these computations may be done relative to a basis of $\mathcal{K}$ allowing us to gain efficiency if the dimension of $\mathcal{K}$ is small relative to N. For example, suppose $n = \dim \mathcal{K}$. Relative to a basis of $\mathcal{K}$, the computation in (4.2) requires $3n + 1$ operations. Relative to a basis of $\mathbb{C}^N$, however, this computation requires $3N + 1$ operations.

4.2. Restricting real symmetric matrices to Krylov subspaces. Let T be an $N \times N$ real symmetric matrix. For $f \in \mathbb{C}^N$, define the jth *Lanczos matrix* L_j to be the symmetric tridiagonal matrix

$$L_j = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{j-1} \\ & & \beta_{j-1} & \alpha_j \end{pmatrix}$$

whose entries are defined recursively using the *Lanczos iteration*.

The Lanczos Iteration
(assuming exact arithmetic)

$\beta_0 = 0$, $q_0 = 0$, $q_1 = f/||f||$

```
for i = 1, 2, 3, ...
    v = T q_i
    α_i = q_i^* v
    v = v − β_{i−1} q_{i−1} − α_i q_i
    β_i = ||v||
    if β_i ≠ 0
        q_{i+1} = v/β_i
    else
        q_{i+1} = 0.
```

The Lanczos iteration is a modified version of the classical Gram–Schmidt orthogonalization process. At its heart is an efficient three-term recurrence which arises because the matrix T is real and symmetric. The usefulness of the Lanczos matrices, together with the q_i that are generated during the Lanczos iteration, is revealed in the following lemma.

LEMMA 4.4. *If the dimension of the Krylov subspace $\mathcal{K} = \langle f, Tf, T^2 f, \dots \rangle$ is m, then $\{q_1, \dots, q_m\}$ is an orthonormal basis for $\mathcal{K}$ and L_m is the restriction of T to $\mathcal{K}$ with respect to this basis.*

Although the Lanczos iteration is easily implemented, in finite precision arithmetic the q_i quickly lose their property of being orthogonal. They may even become linearly dependent (see, e.g., [22]). For this reason, some form of reorthogonalization is usually introduced. For example, the *Lanczos iteration with complete reorthogonalization*, as described in [16], reorthogonalizes v against *all* of the previous $q_1, \dots, q_i$ after computing α_i and $v = \beta_i q_{i+1}$.

The Lanczos Iteration
with Complete Reorthogonalization
(assuming finite precision arithmetic)

$\beta_0 = 0$, $q_0 = 0$, $q_1 = f/||f||$, $\epsilon =$ tolerance

```
for i = 1, 2, 3, ...
    v = T q_i
    α_i = q_i^* v
    v = v − β_{i−1} q_{i−1} − α_i q_i
    for j = 1 to i
        γ = q_{i−j+1}^* v
        v = v − γ q_{i−j+1}
    β_i = ||v||
    if β_i > ε
        q_{i+1} = v/β_i
    else
        q_{i+1} = 0.
```

Remark. The Lanczos iteration with complete reorthogonalization is much more stable than the Lanczos iteration without reorthogonalization. In fact, the numerical stability of the Lanczos iteration with reorthogonalization is comparable to that of

the Givens and Householder algorithms, which, like the Lanczos iteration, reduce a matrix to tridiagonal form (see Chapter 6, section 41 of [23]).

To get a sense of how much work it takes to compute the Lanczos iteration with complete reorthogonalization, let T^{op} be the number of operations needed to apply the matrix T to an arbitrary vector, either directly or through a given subroutine. Note that T^{op} is never more than the number of nonzero entries of T.

LEMMA 4.5. *If T is an $N \times N$ real symmetric matrix and $f \in \mathbb{C}^N$, then*

$$O(nT^{\text{op}} + n^2N)$$

operations are required to compute n iterations of the Lanczos iteration with complete reorthogonalization for T and f.

Proof. It is easy to see that the Lanczos iteration without reorthogonalization requires $O(nT^{\text{op}} + nN)$ operations. Since complete reorthogonalization requires an additional $O(n^2N)$ operations, the lemma follows. ∎

4.3. The Lanczos eigenspace projection method. We may now state the following theorem. Its proof outlines a method for computing projections onto the eigenspaces of a real symmetric matrix.

THEOREM 4.6. *If T is an $N \times N$ real symmetric matrix with n distinct eigenvalues and f is a nonzero vector in $\mathbb{C}^N$, then the projections of f onto the eigenspaces of T require $O(nT^{\text{op}} + n^2N)$ operations.*

Proof. The claim follows directly from the discussion in [16] of the Rayleigh–Ritz procedure applied to the sequence of Krylov subspaces $\mathcal{K}_1, \mathcal{K}_2, \ldots$ generated by T and f. The method is important, however, so we include the details.

Suppose that f has m nonzero projections $f_1, \ldots, f_m$ onto the eigenspaces of T. Let μ_i be the eigenvalue corresponding to the eigenspace containing f_i. Let L_m be the mth Lanczos matrix generated during the Lanczos iteration with respect to T and f. Let $\{q_1, \ldots, q_m\}$ be the corresponding orthonormal basis of the Krylov subspace $\mathcal{K}$ generated by T and f.

It is useful to express the elements of $\mathcal{K}$ with respect to the basis $\{q_1, \ldots, q_m\}$. Thus, if $v \in \mathcal{K}$, let $\tilde{v}$ denote v with respect to $\{q_1, \ldots, q_m\}$. In other words, if $v = \sum_{i=1}^m \alpha_i q_i$, then $\tilde{v} = (\alpha_1, \ldots, \alpha_m)^t$.

Since $\mathcal{K}$ is spanned by the f_i, $\mathcal{K} = \mathcal{K}_m$ and each μ_i is an eigenvalue of L_m. Let $\tilde{u}_i$ be an eigenvector of L_m with eigenvalue μ_i such that $||\tilde{u}_i|| = 1$. Since L_m is a real symmetric matrix, $\{\tilde{u}_1, \ldots, \tilde{u}_m\}$ is an orthonormal basis for $\mathcal{K}$.

Since $q_1 = ||f||^{-1} f$, $\tilde{f} = (||f||, 0, \ldots, 0)^t$. It follows that $\tilde{f}_i = (\tilde{u}_i^* \tilde{f})\tilde{u}_i$ is the eigenspace projection f_i with respect to the basis $\{q_1, \ldots, q_m\}$. Thus, if Q_m is the $N \times m$ matrix whose ith column is the vector q_i, then $f_i = Q_m \tilde{f}_i$. We may therefore compute the eigenspace projections of f as follows.

Stage 1. Generate L_m and Q_m by using the Lanczos iteration with complete reorthogonalization with T and f until a zero vector appears.

Stage 2. Compute the m eigenvalues $\mu_1, \ldots, \mu_m$ and corresponding eigenvectors $\tilde{u}_1, \ldots, \tilde{u}_m$ of L_m.

Stage 3. For $1 \le i \le m$, compute $\tilde{f}_i = (\tilde{u}_i^* \tilde{f})\tilde{u}_i$.

Stage 4. For $1 \le i \le m$, compute $f_i = Q_m \tilde{f}_i$.

Stage 1 requires $O(mT^{\text{op}} + m^2N)$ operations and Stage 2 requires $O(m^3)$ operations due to the tridiagonal form of T_m (see [23]). Stage 3 requires $O(m^2)$ operations and Stage 4 requires $O(m^2N)$ operations. Since $m \le n \le N$, the theorem follows. ∎

Remark. The coefficient implied by $O(nT^{\text{op}}+n^2N)$ in Theorem 4.6 is independent of n, T^{op}, and N. We will implicitly make use of this fact throughout the rest of the paper.

We will refer to the projection method outlined in Theorem 4.6 as the *Lanczos eigenspace projection method* or LEPM.

Remark. The LEPM is a sensible way of computing eigenspace projections only if n is much less than N and T^{op} is much less than N^2. After all, a naive algorithm that uses matrix multiplication to directly compute the f_i requires $O(nN^2)$ operations. Thus, for our method to be efficient, we must have an efficient algorithm for applying the real symmetric matrix T, and the number of distinct eigenvalues of T must be small relative to the dimension of the space upon which T acts.

4.4. The Lanczos isotypic projection method. In this section, we combine the results of sections 2.3 and 4.3 to create an isotypic projection method that relies on the use of separating sets of real symmetric matrices.

Let G be a finite group, let V be a finite dimensional representation of G, and let $\{T_1, \ldots, T_k\}$ be a separating set of real symmetric matrices for V. By Lemma 2.2, we may compute the isotypic projections of a vector $v \in V$ as follows.

Stage 1. Using the LEPM, compute the projections of v onto each of the eigenspaces of T_1.

Stage 2. Using the LEPM, compute the projections of each of the previously computed projections onto each of the eigenspaces of T_2.

$$\vdots$$

Stage k. Using the LEPM, compute the projections of each of the previously computed projections onto each of the eigenspaces of T_k.

We will refer to this approach to computing isotypic projections as the *Lanczos isotypic projection method* or LIPM.

Let $\iota(V)$ be the least number of operations needed to compute the isotypic projections of an arbitrary vector in V. We may now state our main theorem.

MAIN THEOREM 4.7. *Let G be a finite group acting on a finite set X. Let $L(X)$ be the resulting permutation representation. If $L(X) = V_1 \oplus \cdots \oplus V_n$ is the isotypic decomposition of $L(X)$ and $\{T_1, \ldots, T_k\}$ is an isotypic separating set of real symmetric matrices for $L(X)$, then*

$$\iota(L(X)) = O\left(\sum_{i=1}^{k} \left(nT_i^{\text{op}} + n^2|X|\right)\right).$$

Proof. The number of operations needed at the ith stage of the LIPM is never more than $O(nT_i^{\text{op}} + n^2|X|)$. The theorem follows immediately. □

5. Distance transitive graphs. Let X be a connected graph and denote the distance function of X by d. Let k be the *diameter* of X which is the maximum distance between any two vertices of X. A group G of automorphisms of X is said to be *distance transitive* on X if G is transitive on each of the sets $\{(x, x') \mid x, x' \in X \text{ and } d(x, x') = i\}$ for $0 \leq i \leq k$. A graph is said to be *distance transitive* if it is connected and has a distance transitive group of automorphisms. For example, the 2-element subsets of a 4-element set form a distance transitive graph where two 2-element subsets are adjacent if their intersection has size 1 (see Figure 2). A good reference for distance transitive graphs is [2].

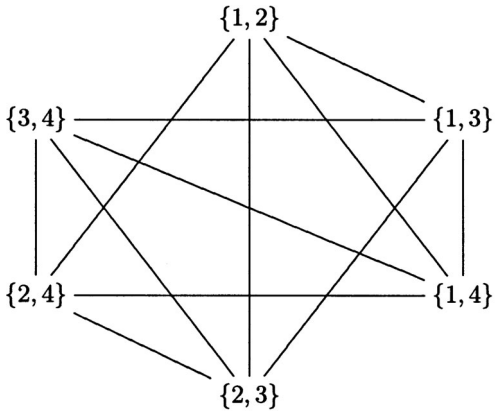


FIG. 2. *A distance transitive graph.*

Let X be a distance transitive graph, let G be a distance transitive group of automorphisms of X, and let $L(X)$ be the permutation representation of G induced by the action of G on the vertices of X. The *adjacency operator* of X is the linear transformation $A : L(X) \to L(X)$, where

$$(Af)(x) = \sum_{x' : d(x,x')=1} f(x')$$

for all $x \in X$. The operator A has $k+1$ distinct eigenvalues which are also the zeros of certain polynomials associated with the graph X (see, e.g., [2]). For example, the adjacency operator of the graph in Figure 2, relative to its delta basis (as defined in section 2.4), is

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

It has three distinct eigenvalues.

LEMMA 5.1. *The isotypic subspaces of $L(X)$ are precisely the eigenspaces of A.*

Proof. This follows from section 2 of Stanton [21]. □

THEOREM 5.2. *Let X be a distance transitive graph with diameter k, let G be a distance transitive group of automorphisms of X, and let $L(X)$ be the associated permutation representation of G. If A is the adjacency operator of X, then*

$$\iota(L(X)) = O(kA^{\mathrm{op}} + k^2|X|).$$

Proof. Relative to the delta basis of $L(X)$, the adjacency operator A is a real symmetric matrix. Thus, the result follows from Theorem 4.7 and Lemma 5.1. □

A direct matrix multiplication approach to computing isotypic projections for $L(X)$ requires $O(k|X|^2)$ operations. Although $O(kA^{\mathrm{op}} + k^2|X|)$ may yield a better upper bound, we may be able to gain even more efficiency by taking advantage of the graph structure of X. For this, the notion of a Radon transform is helpful.

5.1. Radon transforms. Let G be a finite group acting on finite sets X and Y and giving permutation representations $L(X)$ and $L(Y)$, respectively. In addition, suppose there is an incidence relation between X and Y where we write $x \sim y$ if $x \in X$ is incident to $y \in Y$. The *Radon transform* $R : L(X) \to L(Y)$ is then defined by

$$(Rf)(y) = \sum_{x:x\sim y} f(x)$$

for all $x \in X$ (see [1]). The adjoint $R^* : L(Y) \to L(X)$ of R is defined by

$$(R^*f)(x) = \sum_{y:x\sim y} f(y)$$

for all $y \in Y$.

Suppose now that X is a distance transitive graph with respect to G, and let X' be a complete subgraph of X that contains at least two vertices. Recall that a graph is said to be *complete* if every pair of distinct vertices is adjacent. Let Y be the collection of distinct images of X' under the action of G on X, and say that $x \in X$ is incident to $y \in Y$ if x is a vertex of y. Let $R : L(X) \to L(Y)$ be the associated Radon transform. For convenience, we say that Y is a *complete covering of X with Radon transform R*. Note that, with respect to the delta bases of $L(X)$ and $L(Y)$, R^*R is a matrix with integer coefficients, $R^* = R^t$, and $(R^tR)^t = R^tR^{tt} = R^tR$. Thus R^*R is a real symmetric matrix.

We will make use of the integers r and s that are defined in the following lemma.

LEMMA 5.3. *There are integers r and s such that*

$$|\{y \in Y \mid x \sim y\}| = r$$

for every vertex x of X and

$$|\{y \in Y \mid x \sim y \text{ and } x' \sim y\}| = s$$

for every edge $\{x, x'\}$ of X.

Proof. This follows from the fact that X is a distance transitive graph. ∎

LEMMA 5.4. *If $A : L(X) \to L(X)$ is the adjacency operator of X and $I : L(X) \to L(X)$ is the identity, then $A = (1/s)(R^*R - rI)$.*

Proof. This follows from the fact that, for each $x \in X$,

$$\begin{aligned}(R^*Rf)(x) &= \sum_{y:x\sim y} \sum_{x':x'\sim y} f(x') = rf(x) + s\left(\sum_{x':d(x,x')=1} f(x')\right)\\ &= ((rI + sA)f)(x). \qquad \blacksquare\end{aligned}$$

LEMMA 5.5. *If X is a distance transitive graph and Y is a complete covering of X with Radon transform R, then $\{(R^*R)\}$ is a separating set for $L(X)$ and $(R^*R)^{\mathrm{op}} \leq 2r|X|$.*

Proof. Let A be the adjacency operator of X. The product R^*R and the adjacency operator A have the same eigenspaces by Lemma 5.4; therefore $\{(R^*R)\}$ is a separating set since $\{A\}$ is a separating set by Lemma 5.1.

We may apply R^*R to a vector $f \in L(X)$ by first computing Rf and then $R^*(Rf)$. Furthermore, when regarded as a matrix with respect to the delta bases

of $L(X)$ and $L(Y)$, both R and R^* contain $r|X|$ nonzero entries. It follows that $(R^*R)^{\mathrm{op}} \leq R^{*\mathrm{op}} + R^{\mathrm{op}} \leq r|X| + r|X| = 2r|X|$. □

By Theorem 4.7 and Lemma 5.5, we have the following theorem.

THEOREM 5.6. *Let X be a distance transitive graph, and let Y be a complete covering of X. If X has diameter k and $|\{y \in Y \mid x \sim y\}| = r$ for every vertex x of X, then*

$$\iota(L(X)) = O\left(kr|X| + k^2|X|\right).$$

Remark. Since X is a distance transitive graph, there is an integer a such that, for every vertex x of X,

$$|\{x' \in X \mid d(x, x') = 1\}| = a.$$

Applying the adjacency operator of X directly therefore requires no more than $a|X|$ operations. Thus, if r is noticeably less than a, then by Theorem 5.6 we may want to use the associated Radon transform and its adjoint in the LIPM rather than the adjacency operator to compute the isotypic projections of a vector in $L(X)$. We illustrate this in the next two sections.

5.2. The Johnson graph. Let $n \geq 2$ and let $k \leq n/2$. The k-element subsets $X^{(n-k,k)}$ of $\{1, \ldots, n\}$ form a distance transitive graph with automorphism group S_n by defining two k-element subsets to be adjacent if their intersection has size $k-1$. The resulting graph is known as the *Johnson graph*. It has diameter k and is sometimes denoted by $J(n, k)$.

Each vertex of $J(n, k)$ is adjacent to $k(n-k)$ other vertices and $|X^{(n-k,k)}| = \binom{n}{k}$. The number of operations required to directly apply the adjacency operator A is therefore $k(n-k)\binom{n}{k}$. By Theorem 5.2, we therefore have that

$$\iota\left(L\left(X^{(n-k,k)}\right)\right) = O\left(k^2(n-k)\binom{n}{k}\right). \tag{5.1}$$

For each $(k-1)$-element subset $y \in X^{(n-(k-1),k-1)}$ there is a corresponding complete subgraph of $J(n, k)$ consisting of those $x \in X^{(n-k,k)}$ that contain y. The collection Y of these subgraphs forms a complete cover of $J(n, k)$ and each vertex of $J(n, k)$ is contained in k such subgraphs. Thus, by Theorem 5.6, we have the following improvement to (5.1).

THEOREM 5.7. *If $n \geq 2$, $k \leq n/2$, and $L\left(X^{(n-k,k)}\right)$ is the permutation representation of S_n associated to the Johnson graph $J(n, k)$, then*

$$\iota\left(L\left(X^{(n-k,k)}\right)\right) = O\left(k^2\binom{n}{k}\right).$$

We summarize the results of this section in Table 1. Note that the bounds involving the LIPM compare favorably to the upper bound of

$$O\left(\binom{n}{k}^2 + \binom{n}{k} k \log^2 k\right)$$

given in [8].

TABLE 1
Upper bounds on $\iota\left(L\left(X^{(n-k,k)}\right)\right)$.

LIPM with R^*R	LIPM with A	Direct matrix multiplication
$O\left(k^2\binom{n}{k}\right)$	$O\left(k^2(n-k)\binom{n}{k}\right)$	$O\left(k\binom{n}{k}^2\right)$

5.3. The Grassmann graph. Let $n \geq 2$, let $k \leq n/2$, and let V be an n-dimensional vector space over the finite field $\mathbb{F}_q$ of q elements. Let $GL(n,q)$ be the group of automorphisms of V. The k-dimensional subspaces $X_{(n-k,k)}$ of V form a distance transitive graph with respect to $GL(n,q)$ by defining two k-dimensional subspaces to be adjacent if their intersection is a $(k-1)$-dimensional subspace of V. The resulting graph is known as the *Grassmann graph*. It has diameter k and is analogous to the Johnson graph $J(n,k)$. We will denote it by $G(n,k,q)$. See [2] for details concerning the Grassmann graph.

For each nonnegative integer m, let $[m] = 1 + q + q^2 + \cdots + q^{m-1}$, let $[m]! = [m][m-1]\cdots[1]$ if $m > 0$, and let $[0]! = 1$. Note that $[m] = (q^m - 1)/(q-1)$, $[0] = 0$, and $[1] = 1$. Define

$$\binom{m}{l}_q = \begin{cases} [m]!/([l]![m-l]!) & \text{if } m \geq l \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Although not obvious, this is a polynomial in q known as a *Gaussian polynomial* (see, e.g., [20]).

Each vertex of $G(n,k,q)$ is adjacent to $q[k][n-k]$ other vertices and $|X_{(n-k,k)}| = \binom{n}{k}_q$. Direct multiplication of the adjacency operator A of $G(n,k,q)$ therefore requires $q[k][n-k]\binom{n}{k}_q$ operations. By Theorem 5.2, we have that

$$\iota\left(\mathbb{C}[X_{(n-k,k)}]\right) = O\left(kq[k][n-k]\binom{n}{k}_q\right). \tag{5.2}$$

Each $(k-1)$-dimensional subspace $y \in X_{(n-(k-1),k-1)}$, in analogy with the Johnson graph, corresponds to a complete subgraph of $G(n,k,q)$ consisting of those $x \in X_{(n-k,k)}$ that contain y. The collection Y of such subgraphs forms a complete cover of $G(n,k,q)$ and each vertex of $G(n,k,q)$ is contained in $[k]$ such subgraphs. By Theorem 5.6, we therefore have the following improvement to (5.2).

THEOREM 5.8. *Let $n \geq 2$ and $k \leq n/2$. Let $L\left(X_{(n-k,k)}\right)$ be the permutation representation of $GL(n,q)$ associated to the Grassmann graph $G(n,k,q)$. Then*

$$\iota\left(L\left(X_{(n-k,k)}\right)\right) = O\left(k[k]\binom{n}{k}_q\right).$$

We summarize the results of this section in Table 2. As with the Johnson graph, note that the bounds involving the LIPM compare favorably to the upper bound of

$$O\left(\binom{n}{k}_q^2 + \binom{n}{k}_q k\log^2 k\right)$$

given in [8].

TABLE 2
Upper bounds on $\iota\left(L\left(X_{(n-k,k)}\right)\right)$.

LIPM with R^*R	LIPM with A	Direct matrix multiplication
$O\left(k[k]\binom{n}{k}_q\right)$	$O\left(kq[k][n-k]\binom{n}{k}_q\right)$	$O\left(k\binom{n}{k}_q^2\right)$

6. The symmetric group. Spectral analysis for nonabelian groups has found its greatest success with the analysis of ranked data (see [5, 6, 17]). Ranked data arises when respondents are given a list of n items which they are asked to rank in terms of preference. We say that such a ranking is *full* if the respondents are asked to rank each element of the list. On the other hand, we say that a ranking is a *partial ranking of shape* λ if for some sequence $\lambda = (\lambda_1, \ldots, \lambda_m)$ of positive integers whose sum is n, the respondents are asked to choose their top λ_1 items, then their next top λ_2 items, and so on, with no internal ordering. Note that a full ranking is a partial ranking of shape $(1, \ldots, 1)$.

If X^λ is the set of possible partial rankings of shape λ, the *partially ranked data of shape* λ is the function $f \in L\left(X^\lambda\right)$, where, for each $x \in X^\lambda$, $f(x)$ is the number of respondents choosing the partial ranking x. For an example of partially ranked data, consider a lottery in which participants are asked to choose five numbers from the set $\{1, \ldots, 39\}$. Each lottery ticket corresponds to a partial ranking of shape $(5, 34)$, and the relevant ranked data is then the function that assigns to each such ranking the number of tickets corresponding to that ranking that were sold.

For another example of ranked data, consider the partially ranked data that arises when a film society asks its members to choose, from a list of ten movies, their three favorite movies and then their next three favorite movies. Their choices correspond to partial rankings of shape $(3, 3, 4)$, and the relevant partially ranked data is the function that assigns to each such ranking the number of members choosing that ranking.

The natural action of the symmetric group S_n on the n items in the list gives rise to an action of S_n on X^λ. Moreover, as noted in section 1, the isotypic subspaces of the resulting permutation representation $L\left(X^\lambda\right)$ correspond to certain *pure higher order effects* associated to the ranked data $f \in L\left(X^\lambda\right)$ (see [6, 17]). Computing the isotypic projections of f can therefore lead to some insight into how the respondents went about choosing their rankings.

6.1. Representation theory. Let n be a positive integer. A *composition* of n is a sequence $\lambda = (\lambda_1, \ldots, \lambda_m)$ of positive integers whose sum is n. If $\lambda_1 \geq \cdots \geq \lambda_m$, then λ is a *partition* of n. To each composition λ, there corresponds a partition $\overline{\lambda}$ obtained by arranging the parts of λ in nonincreasing order. The partitions of n form a partially ordered set under the *dominance order* where, if λ and λ' are partitions of n, then we say that λ *dominates* λ' if $\lambda_1 + \cdots + \lambda_i \geq \lambda'_1 + \cdots + \lambda'_i$ for all $i \geq 1$. If λ dominates λ', then we write $\lambda \trianglerighteq \lambda'$.

As is often the case, we identify the composition $\lambda = (\lambda_1, \ldots, \lambda_m)$ of n with the *Ferrers diagram* of shape λ, which is the left-justified array of dots with λ_i dots in the ith row (see Figure 3). If the dots of a Ferrers diagram of shape λ are replaced by the numbers $1, \ldots, n$ without repetition, then we create a *Young tableau* of shape

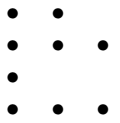


FIG. 3. *The Ferrers diagram of shape* $(2,3,1,3)$.

λ. Two Young tableaux are said to be equivalent if they differ only by a permutation of the entries within the rows of the tableaux. An equivalence class of tableaux is a *tabloid*. A tabloid is denoted by forming a representative tableau and then drawing lines between the rows (see Figure 4).

$$\begin{array}{lll} 4 & 9 & \\ 5 & 2 & 3 \\ 7 & & \\ 8 & 1 & 6 \end{array} \qquad\qquad \begin{array}{lll} 9 & 4 & \\ 2 & 5 & 3 \\ 7 & & \\ 6 & 8 & 1 \end{array}$$

$$\begin{array}{|lll|} \hline 4 & 9 & \\ \hline 2 & 3 & 5 \\ \hline 7 & & \\ \hline 1 & 6 & 8 \\ \hline \end{array}$$

FIG. 4. *Two equivalent tableaux and their tabloid.*

Let X^λ be the set of tabloids of shape λ. The set X^λ naturally corresponds to the set of rankings of shape λ since each row of a tabloid may be viewed as a ranked subset of an n-element set. Moreover, we may rearrange the subsets in each ranking so that their sizes are in nonincreasing order. We may therefore assume that λ is a partition of n.

Let λ be a partition of n. The action of S_n on $\{1,\dots,n\}$ induces an action of S_n on X^λ. For example, if $\sigma = (135)(27)$ and

$$t = \begin{array}{|lll|} \hline 5 & 2 & 3 \\ \hline 4 & 1 & 6 \\ \hline 7 & & \\ \hline \end{array},$$

then

$$\sigma t = \begin{array}{|lll|} \hline \sigma(5) & \sigma(2) & \sigma(3) \\ \hline \sigma(4) & \sigma(1) & \sigma(6) \\ \hline \sigma(7) & & \\ \hline \end{array} = \begin{array}{|lll|} \hline 1 & 7 & 5 \\ \hline 4 & 3 & 6 \\ \hline 2 & & \\ \hline \end{array}.$$

We denote the resulting permutation representation $L(X^\lambda)$ by M^λ.

For every partition μ of n, there is a simple representation W^μ of S_n. These representations form a complete (up to isomorphism) collection of simple representations of S_n. The representation M^λ is isomorphic to a direct sum of simple representations

$$M^\lambda \cong \bigoplus_{\mu \trianglerighteq \lambda} \kappa_{\mu\lambda} W^\mu,$$

where the numbers $\kappa_{\mu\lambda}$ are *Kostka numbers* and denote the multiplicity of W^μ in M^λ. (Kostka numbers also count objects known as *semistandard tableaux*. See, e.g., [18].)

Note that the subspace of M^λ that is isomorphic to $\kappa_{\mu\lambda}W^\mu$ is the isotypic subspace of M^λ that corresponds to the simple representation W^μ.

6.2. Separating sets. Let C_i be the conjugacy class of i-cycles in S_n and let T_i be the corresponding class sum with respect to M^λ. For example, if $n = 4$, $i = 3$, and $\lambda = (2, 2)$, then

$$C_3 = \{(123), (132), (124), (142), (134), (143), (234), (243)\}$$

and, under a particular order of the delta basis of $M^{(2,2)}$,

$$T_3 = \begin{pmatrix} 0 & 2 & 2 & 2 & 2 & 0 \\ 2 & 0 & 2 & 2 & 0 & 2 \\ 2 & 2 & 0 & 0 & 2 & 2 \\ 2 & 2 & 0 & 0 & 2 & 2 \\ 2 & 0 & 2 & 2 & 0 & 2 \\ 0 & 2 & 2 & 2 & 2 & 0 \end{pmatrix}.$$

THEOREM 6.1 (Katriel). *If $\lambda = (\lambda_1, \ldots, \lambda_m)$ is a partition of n, then $\{T_2, \ldots, T_m\}$ is a separating set for M^λ.*

Proof. This is Theorem 3 in Katriel [11] rewritten using the language of separating sets. □

Moreover, the number of T_i that are actually needed to form a separating set for any representation of S_n seems to be small relative to n. Katriel made this observation after calculations revealed that $\{T_2, \ldots, T_{k+1}\}$ is a separating set for any representation of the symmetric group on $\varphi(k)$ or less symbols, where $\varphi(k)$ is much larger than k. For example, $\{T_2\}$ is a separating set for S_2, S_3, S_4, and S_5 but not S_6. Thus $\varphi(1) = 5$. Similarly, calculations have shown that $\varphi(2) = 14$, $\varphi(3) = 23$, $\varphi(4) = 41$, and $\varphi(5) \geq 72$ (see [11]). We therefore have the following theorem.

THEOREM 6.2. *Let n and k be positive integers such that $n \leq \varphi(k)$. If λ is a partition of n, and ζ_λ is the number of isotypic subspaces of M^λ, then*

$$\iota\left(M^\lambda\right) = O\left(\sum_{i=2}^{k+1}\left(\zeta_\lambda(i-1)!\binom{n}{i}|X^\lambda| + \zeta_\lambda^2|X^\lambda|\right)\right).$$

Proof. The collection $\{T_2, \ldots, T_{k+1}\}$ is a separating set for M^λ since $n \leq \varphi(k)$. It is easy to show that each T_i is a real symmetric matrix with respect to the delta basis of M^λ. Thus, by Theorem 4.7,

$$\iota\left(M^\lambda\right) = O\left(\sum_{i=2}^{k+1}\left(\zeta_\lambda T_i^{\mathrm{op}} + \zeta_\lambda^2|X^\lambda|\right)\right).$$

Recall that T_i^{op} is no more than the number of nonzero entries in T_i, which is at most $|C_i||X^\lambda|$. Since $|C_i| = (i-1)!\binom{n}{i}$, the theorem follows. □

We summarize the results of this section, and include some particular examples, in Table 3.

Remarks. Note that when $n \geq 2$ and $k \leq n/2$, we were able to find a bound for $\iota\left(M^{(n-k,k)}\right)$ in section 5.2 by viewing the elements of $X^{(n-k,k)}$ as the vertices of a distance transitive graph. Moreover, the upper bound in section 5.2 is much better than the upper bound given by Theorem 6.2.

TABLE 3
Upper bounds on $\iota(M^\lambda)$.

λ	LIPM	Direct matrix multiplication
$(n-k,k)$	$O\left(k^2(n-k)\binom{n}{k}\right)$	$O\left(k\binom{n}{k}^2\right)$
$(n-k,k-1,1)$	$O\left(k^3(n-k)\binom{n}{k}\right)$	$O\left(k^3\binom{n}{k}^2\right)$
$(\lambda_1,\ldots,\lambda_m)$ $n \le \varphi(k)$	$O\left(\sum_{i=2}^{k+1}\left(\zeta_\lambda(i-1)!\binom{n}{i}\lvert X^\lambda\rvert + \zeta_\lambda^2\lvert X^\lambda\rvert\right)\right)$	$O\left(\zeta_\lambda\lvert X^\lambda\rvert^2\right)$

Additionally, an FFT and inverse for the symmetric group, both requiring $O(n^2 n!)$ operations, were constructed in [12]. Thus if $p(n)$ is the number of partitions of n, then the isotypic projections of a vector in $M^{(1,\ldots,1)}$ may be computed using $O(p(n)n^2 n!)$ operations. See [14] for an FFT for the homogeneous space $M^{(n-k,k)}$ and [15] for some generalizations of the results presented in this paper.

REFERENCES

[1] E. BOLKER, *The finite Radon transform*, in Integral Geometry (Brunswick, Maine, 1984), AMS, Providence, RI, 1987, pp. 27–50.

[2] A. BROUWER, A. COHEN, AND A. NEUMAIER, *Distance-regular Graphs*, Springer-Verlag, Berlin, 1989.

[3] J. CHEN, *Group Representation Theory for Physicists*, 2nd ed., World Scientific, Teaneck, NJ, 1989.

[4] J. CULLUM AND R. WILLOUGHBY, *Lánczos Algorithms for Large Symmetric Eigenvalue Computations, Vol.* I, *Theory*, Birkhäuser Boston, Boston, MA, 1985.

[5] P. DIACONIS, *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, CA, 1988.

[6] P. DIACONIS, *A generalization of spectral analysis with application to ranked data*, Ann. Statist., 17 (1989), pp. 949–979.

[7] P. DIACONIS AND D. ROCKMORE, *Efficient computation of isotypic projections for the symmetric group*, in Groups and Computation (New Brunswick, NJ, 1991), AMS, Providence, RI, 1993, pp. 87–104.

[8] J. DRISCOLL, D. HEALY, AND D. ROCKMORE, *Fast discrete polynomial transforms with applications to data analysis for distance transitive graphs*, SIAM J. Comput., 26 (1997), pp. 1066–1099.

[9] D. DUMMIT AND R. FOOTE, *Abstract Algebra*, 2nd ed., John Wiley, New York, 1999.

[10] W. GENTLEMAN AND G. SANDE, *Fast Fourier transforms for fun and profit*, in Proc. AFIPS, Fall Joint Computer Conference, Vol. 29, Spartan Books, NY, 1966, pp. 563–578.

[11] J. KATRIEL, *Some useful results concerning the representation theory of the symmetric group*, J. Phys. A, 24 (1991), pp. 5227–5234.

[12] D. MASLEN, *The efficient computation of Fourier transforms on the symmetric group*, Math. Comp., 67 (1998), pp. 1121–1147.

[13] D. MASLEN AND D. ROCKMORE, *Generalized FFTs—a survey of some recent results*, in Groups and Computation, II, (New Brunswick, NJ, 1995), AMS, Providence, RI, 1997, pp. 183–237.

[14] D. MASLEN AND D. ROCKMORE, *Separation of variables and the computation of Fourier transforms on finite groups.* I, J. Amer. Math. Soc., 10 (1997), pp. 169–214.

[15] M. ORRISON, *An Eigenspace Approach to Decomposing Representations of Finite Groups*, Ph.D. thesis, Dartmouth College, Hanover, NH, 2001.

[16] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall Inc., Englewood Cliffs, NJ, 1980.
[17] D. Rockmore, *Some applications of generalized FFTs*, in Groups and Computation, II (New Brunswick, NJ, 1995), AMS, Providence, RI, 1997, pp. 329–369.
[18] B. Sagan, *The Symmetric Group. Representations, Combinatorial Algorithms, and Symmetric Functions*, The Wadsworth & Brooks/Cole Mathematics Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991.
[19] J.-P. Serre, *Linear Representations of Finite Groups*, Springer-Verlag, New York, 1977.
[20] R. Stanley, *Enumerative Combinatorics. Vol.* 1, Cambridge University Press, Cambridge, UK, 1997.
[21] D. Stanton, *Orthogonal polynomials and Chevalley groups*, in Special Functions: Group Theoretical Aspects and Applications, Reidel, Dordrecht, The Netherlands, 1984, pp. 87–128.
[22] L. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[23] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 804–828

SOLUTION METHODS FOR $\mathbb{R}$-LINEAR PROBLEMS IN $\mathbb{C}^n$*

TIMO EIROLA†, MARKO HUHTANEN‡, AND JAN VON PFALER†

Abstract. We consider methods, both direct and iterative, for solving an $\mathbb{R}$-linear system $Mz + M_{\#}\overline{z} = b$ in $\mathbb{C}^n$ with a pair of matrices $M, M_{\#} \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^n$. Algorithms that avoid formulating the problem as an equivalent real linear system in $\mathbb{R}^{2n}$ are introduced. Conversely, this implies that real linear systems in $\mathbb{R}^{2n}$ can be solved with the methods proposed in this paper. Our study is motivated by Krylov subspace iterations, which when used with the real formulation can be disastrous in the standard linear case. Related matrix analysis and spectral theory are developed.

Key words. $\mathbb{R}$-linear operator in $\mathbb{C}^n$, characteristic bivariate polynomial, isometry, LU-decomposition, QR-factorization, iterative methods, consimilarity

AMS subject classifications. 15A04, 65F10

DOI. 10.1137/S0895479802415946

1. Introduction.

Consider solving, with a pair of square matrices $M, M_{\#} \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^n$, the $\mathbb{R}$-linear system

$$Mz + M_{\#}\overline{z} = b. \tag{1.1}$$

Any standard linear system is a special case of this when either $M_{\#}$ or M is zero (linear and antilinear, respectively). If both of these matrices are nonzero, we have a real linear operator in $\mathbb{C}^n$. This type of equation arises in certain engineering applications; see [23, 19, 20, 22]. See also [2], [16, Chapters 4.15 and 5.25], and the references therein.

In this paper we introduce direct and iterative methods for solving (1.1). Our study was originally motivated by iterative methods since the problem could readily be rewritten as an equivalent linear system of doubled size for its real and imaginary parts. Then any of the standard Krylov subspace methods could be executed. The usual linear case suggests, however, that this is not necessarily a good idea since the speed of convergence of iterations can be prohibitively slow; see [4, 5, 3].

To avoid the real formulation with Krylov subspace methods, one option is to generate a matrix $Q_k \in \mathbb{C}^{n\times k}$ with orthonormal columns. To this end we employ the $\mathbb{R}$-linear operator corresponding to the left-hand side of (1.1) in an Arnoldi-type iteration. Then projecting the problem to $\mathbb{C}^k$, by using the Q_k computed, gives rise to a real linear system which can be solved with dense matrix techniques. This approach can be interpreted as a Galerkin approximation. Also minimal residual methods are devised.

It is also of interest to note that any real $2n$-by-$2n$ system can be written as (1.1). Therefore all the solution methods introduced in this paper apply to real linear systems in $\mathbb{R}^{2n}$ as well. This gives rise to new direct methods as well as novel nonsymmetric iterations for real problems.

*Received by the editors October 9, 2002; accepted for publication (in revised form) by H. van der Vorst July 18, 2003; published electronically February 24, 2004.
http://www.siam.org/journals/simax/25-3/41594.html

†Institute of Mathematics, Helsinki University of Technology, Box 1100, FIN-02015, Finland (Timo.Eirola@hut.fi, Jan.von.Pfaler@hut.fi).

‡Department of Mathematics, Room 2-335, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 01239 (Marko.Huhtanen@hut.fi). Current address: Institute of Mathematics, Helsinki University of Technology, Box 1100, FIN-02015, Finland. The work of this author was supported by National Science Foundation grant DMS-0209437.

Clearly $\mathbb{R}$-linearity is a weaker assumption than $\mathbb{C}$-linearity. Therefore the problem considered involves two complex square matrices, which makes the arising matrix analysis very interesting. A large part of the paper is devoted to these questions. The spectrum of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$ is introduced. We present various canonical forms, factorizations, and respective solution formulas for problem (1.1).

The paper is organized as follows. In section 2 we develop basic matrix analysis and spectral theory for $\mathbb{R}$-linear operators in $\mathbb{C}^n$. Direct methods for solving real linear systems are derived. In section 3 we introduce iterative methods for solving the corresponding problem and give numerical examples. In section 4 some preliminary ideas are considered for computing the spectrum numerically. Properties of the spectrum are illustrated with numerical experiments.

2. Properties of $\mathbb{R}$-linear operators in $\mathbb{C}^n$. When $\mathbb{C}^n$ is regarded as a vector space over $\mathbb{R}$, an $\mathbb{R}$-linear operator in $\mathbb{C}^n$ can be represented by a $2n$-by-$2n$ matrix. However, in this paper we consider $\mathbb{C}^n$ as a vector space over $\mathbb{C}$ with its usual complex-valued inner product and associate with the system (1.1) an $\mathbb{R}$-linear mapping

$$\mathcal{M}(z) = Mz + M_{\#}\overline{z} \tag{2.1}$$

in $\mathbb{C}^n$. For the converse, when $\mathbb{C}^n$ is regarded as a vector space over $\mathbb{C}$, it is easy to verify that any real linear mapping in $\mathbb{C}^n$ can be represented in this form, after fixing a basis. We call M and $M_{\#}$ the linear and antilinear parts of $\mathcal{M}$, respectively.

Aside from the system (1.1) one can consider its real form by using the matrices M and $M_{\#}$. To this end, write $z = x + iy$ and $b = c + id$. Then equating the real and imaginary parts gives rise to the linear system

$$\begin{bmatrix} \mathrm{Re}(M + M_{\#}) & -\mathrm{Im}(M - M_{\#}) \\ \mathrm{Im}(M + M_{\#}) & \mathrm{Re}(M - M_{\#}) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c \\ d \end{bmatrix}. \tag{2.2}$$

We denote by $A \in \mathbb{R}^{2n\times 2n}$ the arising coefficient matrix. Conversely,[1] this block structuring provides the conditions on reformulating a real $2n$-by-$2n$ linear system as an $\mathbb{R}$-linear problem in $\mathbb{C}^n$.

It is readily seen that if the pairs $(M, M_{\#})$ and $(N, N_{\#})$ correspond to the matrices A and B, respectively, then the real linear map

$$\mathcal{M}(\mathcal{N}(z)) = (MN + M_{\#}\overline{N_{\#}})\, z + (MN_{\#} + M_{\#}\overline{N})\, \overline{z} \tag{2.3}$$

corresponds to the matrix AB. Hence, under sufficient assumptions on invertibility (which are generically satisfied),

$$\mathcal{M}^{-1}(z) = (M - M_{\#}\overline{M}^{-1}\overline{M_{\#}})^{-1} z + (\overline{M_{\#}} - \overline{M}{M_{\#}}^{-1}M)^{-1}\overline{z}. \tag{2.4}$$

Further, the pair $(-iI, 0)$ corresponds to $J = \left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$ and $(M^*, {M_{\#}}^T)$ to A^T.

Using these, we can express some properties of A via M and $M_{\#}$ as follows.

PROPOSITION 2.1. *Let A be the coefficient matrix of* (2.2). *Then*

1. *A is (skew-)symmetric $\iff$ $M^* = (-)\, M$ and ${M_{\#}}^T = (-)\, M_{\#}$;*

[1] Using the notation of [4, section 5.1], this corresponds to representing $A \in \mathbb{R}^{2n\times 2n}$ as the sum $A = M_{\star} + M_{\#_{\star\star}}$ in a unique way; i.e., we have an "$\mathbb{R}$-linear splitting" of A.

Since $M_{\#_{\star\star}}$ is similar to $-M_{\#_{\star\star}}$, its eigenvalues are symmetrically located with respect to the origin. Therefore, if $M_{\#_{\star\star}}$ dominates in this splitting, (2.2) can be very difficult to solve fast with iterative methods.

2. *A is orthogonal* $\iff M^*M + {M_\#}^T \overline{M_\#} = I$ *and* $M^*M_\# + {M_\#}^T \overline{M} = 0$;
3. *A is Hamiltonian* $\iff M^* = -M$ *and* ${M_\#}^T = M_\#$;
4. *A is symplectic* $\iff M^*M - {M_\#}^T \overline{M_\#} = I$ *and* $M^*M_\# - {M_\#}^T \overline{M} = 0$.

(Hamiltonian means that $A^T = JAJ$, and symplectic means that $A^T JA = J$.)

With the norm $\|\mathcal{M}\| = \max_{\|z\|=1} \|\mathcal{M}(z)\|$ the set of $\mathbb{R}$-linear operators in $\mathbb{C}^n$ is a Banach algebra over $\mathbb{R}$. However, $(M, M_\#) \mapsto (M^*, {M_\#}^T)$ is not an involution since $(\overline{\alpha}M^*, \alpha {M_\#}^T) \neq \overline{\alpha}(M^*, {M_\#}^T)$ for $\alpha \in \mathbb{C}\backslash\mathbb{R}$. In particular, we are not dealing with a $\mathbf{C}^\star$-algebra.

2.1. The spectrum of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$. For solvability of (1.1) it is natural to define the spectrum as follows.

DEFINITION 2.2. *$\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{M} : \mathbb{C}^n \mapsto \mathbb{C}^n$ if the range of $\lambda I - \mathcal{M}$ is not $\mathbb{C}^n$. The set of eigenvalues of $\mathcal{M}$ is denoted by $\sigma(\mathcal{M})$.*

If $\lambda = \alpha + i\beta \in \mathbb{C}$, with $\alpha, \beta \in \mathbb{R}$, is an eigenvalue of $\mathcal{M}$, then there exists a vector $b \in \mathbb{C}^n$ such that the equation $(\lambda I - M)z - M_\#\overline{z} = b$ does not have a solution. Then its equivalent real formulation does not have a solution either. If A is the coefficient matrix in (2.2), this implies that

$$A(\alpha, \beta) = \alpha I - \beta J - A \tag{2.5}$$

is not invertible, i.e., $\det A(\alpha, \beta) = 0$. We call $\det A(\alpha, \beta)$ the characteristic bivariate polynomial of $\mathcal{M}$. Consequently, we have an algebraic criterion for finding the eigenvalues of $\mathcal{M}$. The following gives a geometric interpretation.

PROPOSITION 2.3. *If $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{M}$, then there exists a nonzero vector $z \in \mathbb{C}^n$ such that $\mathcal{M}(z) = \lambda z$.*

It is now clear that $\lambda \notin \sigma(\mathcal{M})$ if and only if $\lambda I - \mathcal{M}$ is invertible.

Although an eigenvalue λ gives rise to an $\mathbb{R}$-linear invariant subspace for $\mathcal{M}$, we are actually dealing with a mildly nonlinear eigenproblem. More precisely, there need not be an invariant subspace associated with an eigenvector z of $\mathcal{M}$ when $\mathbb{C}^n$ is regarded as a vector space over $\mathbb{C}$. Indeed, with $\rho, \sigma \in \mathbb{R}$ we have

$$\mathcal{M}((\rho + i\sigma)z) = (\rho + i\sigma)\lambda z - i\, 2\sigma M_\#\overline{z}, \tag{2.6}$$

which belongs to $\operatorname{span}\{z, M_\#\overline{z}\}$ or, equivalently, to $\operatorname{span}\{z, Mz\}$.

PROPOSITION 2.4. *A subspace $V \subset \mathbb{C}^n$ is invariant for $\mathcal{M}$ if and only if it is simultaneously invariant for $z \mapsto Mz$ and $z \mapsto M_\#\overline{z}$.*

Proof. It is clear that the latter implies the former. For the converse, assume that $\mathcal{M}(V) \subset V$. Then with $z \in V$ and $\beta \in \mathbb{C}$ we have $V \ni \beta\mathcal{M}(z) - \mathcal{M}(\beta z) = (\overline{\beta} - \beta)M_\#\overline{z}$, so that $M_\#\overline{V} \subset V$. Therefore, also $MV \subset V$. ∎

In case V is an invariant subspace for $\mathcal{M}$, the spectrum of $\mathcal{M} : V \mapsto V$ is a subset of $\sigma(\mathcal{M})$, a property of fundamental importance in sparse matrix computations.

With an invertible $\mathbb{R}$-linear operator $\mathcal{T}$ in $\mathbb{C}^n$, consider a similarity transformation $\mathcal{T}^{-1} \circ \mathcal{M} \circ \mathcal{T}$ of $\mathcal{M}$. The spectrum of $\mathcal{M}$ remains invariant in this operation if the real form $B \in \mathbb{R}^{2n\times 2n}$ of $\mathcal{T}$ commutes with J. This is equivalent to having $\mathcal{T}(z) = Tz$ for an invertible $T \in \mathbb{C}^{n\times n}$. In this case we say that $\mathcal{T}^{-1} \circ \mathcal{M} \circ \mathcal{T}$ is a $\mathbb{C}$-linear similarity transformation of $\mathcal{M}$. The simplest such $\mathcal{T}$ is $\mathcal{T}(z) = \lambda z$ with $\lambda \in \mathbb{C}\backslash\{0\}$. Then $\mathcal{T}^{-1} \circ \mathcal{M} \circ \mathcal{T}(z) = Mz + \frac{\overline{\lambda}}{\lambda} M_\#\overline{z}$.

A general $\mathbb{R}$-linear similarity transformation in $\mathbb{C}^n$ need not preserve the spectrum except that the eigenvalues on the real axis remain invariant.

To quantify (2.6) more generally, consider the kernel of $\lambda I - \mathcal{M}$, i.e., the set $\{z \in \mathbb{C}^n : \lambda z - \mathcal{M}(z) = 0\}$. Denote by r its dimension as a subspace of $\mathbb{C}^n$ over $\mathbb{R}$

and let m be the dimension of the largest $\mathbb{C}$-linear subspace it contains. The resulting "multiplicity" index pair $(r/2, m)$ gives useful information regarding the eigenvalues of $\mathcal{M}$. Clearly, if the antilinear part of $\mathcal{M}$ vanishes, then $r/2 = m$ for every eigenvalue.

Example 1. Let $\mathcal{M}$ be upper triangular; that is, $M = \left[\begin{smallmatrix} 2 & 1 \\ 0 & 4 \end{smallmatrix}\right]$ and $M_\# = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ are upper triangular matrices. Then for any $\mu, \nu \in \mathbb{C}$ of modulus one we have the eigenvectors and eigenvalues

$$z = \left[\begin{smallmatrix} \mu \\ 0 \end{smallmatrix}\right],\ \lambda = 2 + \overline{\mu}^2 \qquad \text{and} \qquad z = \left[\begin{smallmatrix} \frac{2\nu+\nu^3+\overline{\nu}}{4+2\nu^2+2\overline{\nu}^2} \\ \nu \end{smallmatrix}\right],\ \lambda = 4 + \overline{\nu}^2\ .$$

These are the only eigenvalues and eigenvectors (up to real multiples) of $\mathcal{M}$. Thus the spectrum of $\mathcal{M}$ consists of two circles having one common point $\lambda = 3$. For $\lambda \in \sigma(\mathcal{M}) \backslash \{3\}$ we always have the index pair $(1/2, 0)$. At the intersection point we get $(1, 0)$.

The spectrum has the following algebro-geometric structure (see also [14]).

THEOREM 2.5. *The spectrum of $\mathcal{M} : \mathbb{C}^n \mapsto \mathbb{C}^n$ is a bounded algebraic curve of degree $2n$ at most. The mapping $\lambda \mapsto (\lambda I - \mathcal{M})^{-1}$ is real analytic for $\lambda \notin \sigma(\mathcal{M})$.*

Proof. Since $\lambda I - \mathcal{M}$ is invertible if and only if (2.5) is, the spectrum of $\mathcal{M}$ consists of those points $(\alpha, \beta) \in \mathbb{R}^2$ for which $\det A(\alpha, \beta)$ is zero. This is clearly a bivariate polynomial in the real variables α and β of degree $2n$. That the arising algebraic curve must be bounded follows from Proposition 2.3 and the fact that $\mathcal{M}$ is a bounded operator.

For the second claim, for a fixed λ the mapping $\mathcal{R}(\lambda) = (\lambda I - \mathcal{M})^{-1}$ is also an $\mathbb{R}$-linear operator in $\mathbb{C}^n$. Therefore $(\lambda I - \mathcal{M})^{-1}(z) = R(\lambda)z + R_\#(\lambda)\overline{z}$ for matrices R and $R_\#$ depending on λ. The inverse of (2.5) is real analytic at those points where the determinant is nonzero. Thus, R and $R_\#$ are real analytic as well. ∎

The boundedness assertion of this theorem imposes restrictions on those algebraic curves that can appear as the spectrum of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$.

If both M and $M_\#$ are upper (lower) triangular matrices, then we say that $\mathcal{M}$ is upper (lower) triangular. As we already saw in Example 1, the spectrum can contain circles. This can be stated more generally in the following theorem.

THEOREM 2.6. *Assume that $S \in \mathbb{C}^{n\times n}$ is invertible and $\mathcal{R}(z) = Rz + R_\# \overline{z} = S^{-1}\mathcal{M}(Sz)$ is upper (lower) triangular. Then $\sigma(\mathcal{M})$ is the union of the circles*

$$\left\{\lambda \in \mathbb{C} : |r_{j,j} - \lambda| = |r^\#_{j,j}|\right\}, \quad j = 1, \ldots, n.$$

Proof. Assume $\mathcal{R}$ is upper triangular. Clearly the spectra of $\mathcal{R}$ and $\mathcal{M}$ are the same.

If λ is not in the union of the circles, then the equations of type

$$(r_{k,k} - \lambda)w_k + r^\#_{k,k}\,\overline{w}_k = v_k$$

are uniquely solvable for w_k. Then $\lambda z - \mathcal{R}(z) = 0$ implies $z = 0$. Hence λ is not an eigenvalue of $\mathcal{R}$.

If λ is in the union, take the first j such that $|r_{j,j} - \lambda| = |r^\#_{j,j}|$. Set $w_j = \left(\frac{r^\#_{j,j}}{r_{j,j}-\lambda}\right)^{\frac{1}{2}}$ ($w = 1$ if $r^\#_{j,j} = 0$) and $w_k = 0$ for $k = j+1, \ldots, n$. Then $(r_{j,j} - \lambda)w_j - r^\#_{j,j}\,\overline{w}_j = 0$ and the equations for $w_k\,,\ k = j-1, \ldots, 1$, are uniquely solvable recursively to give an eigenvector of $\mathcal{R}$. ∎

In case $M_\# = 0$ we may use a Schur decomposition of M and the circles reduce to points.

Although we do not have a spectral mapping theorem for a real linear operator $\mathcal{M}$, it is clear, for a fixed $\mu \in \mathbb{C}$, what the spectra of $\mu I + \mathcal{M}$ and $\mu I \circ \mathcal{M}$ are in

terms of $\sigma(\mathcal{M})$. However, in Theorem 2.6 we can assume the diagonal entries of $R_\#$ to be nonnegative real (after performing a $\mathbb{C}$-linear diagonal unitary similarity transformation, if necessary). Consequently, we have a spectral mapping theorem in case $\mathcal{M}$ is triangularizable; i.e., by knowing only $\sigma(\mathcal{M})$ we can readily determine $\sigma\left(p(\mathcal{M})\right)$ for any polynomial p. To this end use (2.3) repeatedly.

In the situation of Theorem 2.6 there exists an increasing chain of nested invariant subspaces of $\mathcal{M}$ of dimension k for $k = 1, 2, \ldots, n$. The spectrum of $\mathcal{M}$ restricted to these subspaces consists of k circles corresponding to the first k pairs of the diagonal entries of R and $R_\#$. In this manner there arises a hierarchy among these circles since, unlike with the Schur decomposition (which exists if $M_\# = 0$), we cannot reorder the diagonal entries of R and $R_\#$ pairwise in general. To see this, it suffices to consider a 2-by-2 case. The circle corresponding to the $(1,1)$-entries always gives rise to an invariant subspace of dimension 1. The other circle need not have an invariant subspace associated with it. Consider, for example, $\mathcal{M}$ with $M = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ and $M_\# = \left[\begin{smallmatrix} 3 & 1 \\ 0 & 1 \end{smallmatrix}\right]$. The invariant subspace of M corresponding to the eigenvalue $\lambda = 0$ of M is not invariant for $z \mapsto M_\# \overline{z}$. Thus, by Proposition 2.4, the order of the diagonal entries of M and $M_\#$ cannot be swapped.

Remark. Under the assumptions of Theorem 2.6, the characteristic bivariate polynomial of $\mathcal{M}$ factors as the product of second degree bivariate polynomials. So one might consider using Krylov subspace techniques to locate just a few of these circles in case n is large. However, the prescribed hierarchy can make this a very challenging problem.

If R and $R_\#$ are diagonal matrices, it is natural to say that $\mathcal{M}$ is diagonalizable in a $\mathbb{C}$-linear similarity transformation. Equivalently, $\mathcal{M}$ has n linearly independent eigenvectors which each give rise to an invariant subspace of $\mathcal{M}$. If the matrix S can be chosen unitary, we say that $\mathcal{M}$ is unitarily diagonalizable. Then M is normal while the condition on $M_\#$ means that the matrix is unitarily condiagonalizable, i.e., complex symmetric [9, Chapter 4.6]. See [9, Chapter 4.5] for a careful study and examples of the case in which M is additionally Hermitian. To this corresponds a symmetric coefficient matrix A in (2.2).

Remark. If $\mathcal{M}$ is unitarily diagonalizable in a $\mathbb{C}$-linear similarity transformation, then its real form A lies in the unitary orbit of binormal matrices. For this, see [12].

The spectrum is not the union of circles in general.

Example 2. One readily verifies that with $M = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$ and $M_\# = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 2 \end{smallmatrix}\right]$ the eigenvalues of $\mathcal{M}$ are given by those $\lambda \in \mathbb{C}$ that satisfy the equation $\lambda^2 - 2e^{-i2\theta}\lambda - e^{-i2\theta} = 0$ for some $\theta \in [0, 2\pi)$.

To find a simpler form $S^*\mathcal{M}(Sz) = Rz + R_\# \overline{z}$ for a general $\mathcal{M}$ with a unitary matrix $S \in \mathbb{C}^{n\times n}$, we can always have an upper triangular $S^*MS = R$ (employ the Schur decomposition of M). Or alternatively, by performing Householder transformations in an obvious way, we can have a Hessenberg matrix $S^* M_\# \overline{S} = R_\#$.

To generalize the concept of unitary $\mathbb{C}$-linear similarity transformation, we call an $\mathbb{R}$-linear operator in $\mathbb{C}^n$ an *isometry* if it preserves the spectral norm. This gives us a group since, clearly, an isometry corresponds to an orthogonal matrix in $\mathbb{R}^{2n\times 2n}$. Hence, if $\mathcal{U}(z) = Uz + U_\# \, \overline{z}$ is an isometry, we have

$$\mathcal{U}^{-1}(z) = U^* z + {U_\#}^T \, \overline{z} \, .$$

Example 3. If $Q \in \mathbb{C}^{n\times k}$ satisfies $\mathrm{Re}(Q^*Q) = I$, then $\mathcal{U}(z) = (I - QQ^*)z - QQ^T\overline{z}$ is an isometry (use item 2 of Proposition 2.1). In addition, it satisfies $\mathcal{U}^2 = I$. Note that the columns of Q need not be linearly independent over $\mathbb{C}$, as $Q =$

$\frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1-i & 1+i \\ 0 & 0 \end{smallmatrix}\right]$ illustrates. Moreover, if $Q \in \mathbb{C}^{n\times 1}$ is a unit vector, then $\mathcal{U}$ corresponds to a Householder transformation in $\mathbb{R}^{2n\times 2n}$.

With an isometry we preserve the lengths but not the angles; i.e., for $z, w \in \mathbb{C}^n$ the inner product (z, w) need not equal $(\mathcal{U}(z),\mathcal{U}(w))$ unless $\mathcal{U}$ is a $\mathbb{C}$-linear unitary operator. In connection with the QR-decomposition we need isometries which map an arbitrary pair of vectors to be parallel; see section 2.2.

PROPOSITION 2.7. *Let $\mathcal{U}$ be an isometry. Then $\sigma(\mathcal{U})$ is either empty, a finite set on the unit circle, or the unit circle.*

Proof. If λ is an eigenvalue of $\mathcal{U}$, then it must have modulus one. If $\sigma(\mathcal{U})$ is not finite, then the respective algebraic curve must be closed. Thereby it is the unit circle. To see that the spectrum can be empty, consider $\mathcal{U}(z) = \left[\begin{smallmatrix} 0 & 1 \\ i & 0 \end{smallmatrix}\right]\overline{z}$. ∎

If $M^* = -M$ and ${M_\#}^T = -M_\#$, then $(\mathcal{M} + I) \circ (\mathcal{M} - I)^{-1}$ gives us an isometry, i.e., an analogy of the Cayley transform.

The following can be verified by a direct computation.

PROPOSITION 2.8. *Let $\mathcal{U}$ be an isometry. If $\mathcal{M}(z) = Mz + M_\#\overline{z}$ with $M^* = M$ and ${M_\#}^T = M_\#$ (or $M^* = -M$ and ${M_\#}^T = -M_\#$), then $\mathcal{U}^{-1} \circ \mathcal{M} \circ \mathcal{U} = Nz + N_\#\overline{z}$ with $N^* = (-)\,N$ and ${N_\#}^T = (-)\,N_\#$.*

In a translation of an antilinear operator we have $M = \kappa I$ with $\kappa \in \mathbb{C}$. This case is of particular importance in view of applications [23, 19, 20, 22] (with $\kappa = 0$ it arises in particle physics). It also appears after preconditioning the system (1.1) with the inverse of M from the left, under the assumption that M is readily invertible. We denote the corresponding operator by $\mathcal{M}_\kappa$, that is, $\mathcal{M}_\kappa(z) = \kappa z + M_\#\overline{z}$. This yields us another instance where we encounter circles.

PROPOSITION 2.9. *For $\mathcal{M}_\kappa$ the spectrum is the union of circles centered at κ.*

Proof. Repeat the arguments of [9, p. 246] with the translation κ. ∎

This case is not covered by Theorem 2.6 since $\mathcal{M}_\kappa$ may not have an upper triangular form under $\mathbb{C}$-linear similarity transformation ([9, Theorem 4.6.3] determines when this is possible). Moreover, the situation is fundamentally different now since there is an invariant subspace associated with each circle.

At least one circle appears in the following case.

PROPOSITION 2.10. *Assume that $\mathcal{M}(z) = Mz + \kappa\overline{z}$ with $\kappa \in \mathbb{C}$. If the intersection of the null spaces of M and $\overline{M}$ is nontrivial, then $\sigma(\mathcal{M})$ contains the set $\{\lambda \in \mathbb{C} : |\lambda| = |\kappa|\}$.*

Proof. If $Mv = M\overline{v} = 0$, then

$$\lambda(\alpha v + \alpha\overline{v}) - \mathcal{M}(\alpha v + \alpha\overline{v}) = (\lambda\alpha - \kappa\overline{\alpha})(v + \overline{v}).$$

If $v + \overline{v} \neq 0$, we get eigenvalues $\lambda = \kappa\overline{\alpha}/\alpha$, i.e., all complex numbers with modulus $|\kappa|$. If $\overline{v} = -v$, use the vector $v - \overline{v}$. ∎

Note that if $\operatorname{rank}(M) < \frac{n}{2}$, then the assumptions of the proposition are automatically satisfied. This is the case also if M is real and singular.

If $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{M}_\kappa$, then $M_\#\overline{z} = (\lambda - \kappa)z$ holds for a nonzero $z \in \mathbb{C}^n$. Therefore $\overline{M_\#}z = (\overline{\lambda} - \overline{\kappa})\overline{z}$, so that

$$M_\#\overline{M_\#}z = (\overline{\lambda} - \overline{\kappa})M_\#\overline{z} = |\lambda - \kappa|^2 z.$$

Consequently, a necessary condition for λ to be an eigenvalue of $\mathcal{M}_\kappa$ is that $M_\#\overline{M_\#}$ has $|\lambda - \kappa|^2$ as its eigenvalue. Since $M_\#\overline{M_\#}$ may have no real nonnegative eigenvalues, we infer that the spectrum of $\mathcal{M}_\kappa$ can be empty. See also [9, Chapter 4].

The spectrum of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$ is related to the eigenvalues of its real form (2.2) as follows.

PROPOSITION 2.11. *Let $A \in \mathbb{R}^{2n\times 2n}$ be the real form of $\mathcal{M}(z) = Mz + M_{\#}\overline{z}$. Then $\lambda = \alpha + i\beta \in \sigma(\mathcal{M}) \setminus \{0\}$ if and only if $\alpha^2 + \beta^2 \in \sigma\big(\alpha A + \beta JA\big)$.*

Proof. Assume $z = x + iy \in \mathbb{C}^n$ is an eigenvector corresponding to λ. Then rewriting the equality $\mathcal{M}(z) = \lambda z$ by using (2.5) we have

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \otimes I \begin{bmatrix} x \\ y \end{bmatrix},$$

which is equivalent to

$$\left(\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \otimes I \right) A \begin{bmatrix} x \\ y \end{bmatrix} = (\alpha^2 + \beta^2) \begin{bmatrix} x \\ y \end{bmatrix},$$

so that the claim follows. ∎

Fix $\rho \in \mathbb{R}$, assume $\mu \in \mathbb{R}$, and consider

$$A + \rho JA. \tag{2.7}$$

Then using Proposition 2.11 we obtain $\beta = \rho\alpha$ and $\alpha\mu = \alpha^2(1 + \rho^2)$, so that $\lambda = \frac{\mu}{1+\rho^2}(1 + i\rho)$ is an eigenvalue of $\mathcal{M}$. In other words, any real eigenvalue of (2.7), with $\rho \in \mathbb{R}$, gives rise to an eigenvalue of $\mathcal{M}$.

Remark. Consider $\mathcal{M}_\kappa$ with $M_\#$ to be complex symmetric. Then (2.7) is symmetric for $\mathcal{M}_0$, independently of $\rho \in \mathbb{R}$. Consequently, $\sigma(\mathcal{M}_\kappa)$ is nonempty.

Although we do not have an adjoint operator for a real linear operator in $\mathbb{C}^n$, the following operation is of interest.

PROPOSITION 2.12. *For $\mathcal{M}(z) = Mz + M_{\#}\overline{z}$ let $\widetilde{\mathcal{M}}(z) = M^*z + {M_{\#}}^T\overline{z}$. Then $\sigma(\widetilde{\mathcal{M}}) = \overline{\sigma(\mathcal{M})}$.*

Proof. The real form of $\widetilde{\mathcal{M}}$ is A^T, where A is the real form of $\mathcal{M}$. Since $(\beta J + A)^T = -\beta J + A^T$, an eigenvalue $\alpha + i\beta$ of $\widetilde{\mathcal{M}}$ gives rise to an eigenvalue $\alpha - i\beta$ of $\mathcal{M}$ and vice versa. ∎

In particular, if $M^* = M$ and ${M_{\#}}^T = M_{\#}$, then the spectrum is symmetric relative to the real axis. If $M^* = -M$ and ${M_{\#}}^T = -M_{\#}$, then $\sigma(\mathcal{M})$ is on the imaginary axis although it need not be symmetrically located with respect to the origin.

Naturally all the eigenvalues of a real linear operator $\mathcal{M}$ in $\mathbb{C}^n$ lie inside the disk $\{\lambda \in \mathbb{C} : |\lambda| \le ||\mathcal{M}||\}$. Also the field of values is defined in an obvious way. Geršgorin disks have an analogy with

$$\rho_l(M, M_\#) = |m^{\#}_{l,l}| + \sum_{j=1, j\neq l}^{n} (|m_{l,j}| + |m^{\#}_{l,j}|).$$

For this a direct adaptation, e.g., of the proof of [9, Theorem 6.1.1], can be used to show that the eigenvalues of $\mathcal{M}$ are located in the union of disks

$$\bigcup_{l=1}^{n} \{z \in \mathbb{C} : |z - m_{l,l}| \le \rho_l(M, M_\#)\}. \tag{2.8}$$

An analogy of the Bauer–Fike theorem holds as well, as follows.

PROPOSITION 2.13. *Assume that $S \in \mathbb{C}^{n\times n}$ is invertible such that $S^{-1} \circ \mathcal{M} \circ S$ is diagonal. If $\mathcal{E}$ is $\mathbb{R}$-linear in $\mathbb{C}^n$ and λ is an eigenvalue of $\mathcal{M} + \mathcal{E}$, then*

$$\operatorname{dist}(\sigma(\mathcal{M}), \lambda) \le \left|\left|S^{-1}\right|\right| \, ||S|| \, \|\mathcal{E}\|.$$

Proof. For a diagonal ℝ-linear operator the norm of the resolvent is the reciprocal of the distance of λ to the spectrum. To see this it suffices to consider the scalar case. With fixed λ_1, $\lambda_2 \in \mathbb{C}$ the inverse of $z \mapsto (\lambda - \lambda_1)z - \lambda_2\overline{z}$ is

$$z \mapsto \frac{1}{|\lambda - \lambda_1|^2 - |\lambda_2|^2}\big((\overline{\lambda} - \overline{\lambda_1})z + \lambda_2\overline{z}\big). \tag{2.9}$$

Choosing z on the unit circle such that $(\overline{\lambda} - \overline{\lambda_1})z$ and $\lambda_2\overline{z}$ are parallel, the norm of (2.9) is $|\frac{1}{|\lambda-\lambda_1|-|\lambda_2|}|$, i.e., the reciprocal of the distance of λ to the spectrum. Hence we can mimic the proof of [9, Theorem 6.3.2] together with (2.14). ∎

This is of use, for example, if M is diagonalizable and $\|M_\#\| \ll \|M\|$.

2.2. Factorizations for an ℝ-linear operator in $\mathbb{C}^n$. Consider solving a real linear system $\mathcal{M}(z) = b$ for $b \in \mathbb{C}^n$. If $\mathcal{M}$ is upper (lower) triangular, then we can use the formula (2.9) on a sequence of 1-by-1 systems together with back (forward) substitution to find the solution.[2]

For solving a general invertible real linear system in $\mathbb{C}^n$ we need to factorize $\mathcal{M}$.

LU–decomposition. For given $M, M_\# \in \mathbb{C}^{n\times n}$ consider finding a lower triangular $\mathcal{L}(z) = Lz + L_\#\overline{z}$ and an upper triangular $\mathcal{U}(z) = Uz + U_\#\overline{z}$ such that

$$\mathcal{M}(z) = Mz + M_\#\overline{z} = \mathcal{L}(\mathcal{U}(z)) = (LU + L_\#\overline{U_\#})z + (LU_\# + L_\#\overline{U})\overline{z}$$

holds for every $z \in \mathbb{C}^n$, i.e., $\mathcal{M} = \mathcal{L} \circ \mathcal{U}$. We assume that all the diagonal entries of L are equal to 1, and that $L_\#$ is strictly lower triangular.

For this LU-decomposition we need appropriate elementary ℝ-linear operators in $\mathbb{C}^n$. The following is easy to check, where, for the sake of clarity, both row and column vectors are boldfaced.

LEMMA 2.14. *If*

$$\mathcal{L}(z) = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{l} & I \end{bmatrix} z + \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{l}^\# & 0 \end{bmatrix} \overline{z}\,,$$

then

$$\mathcal{L}^{-1}(z) = \begin{bmatrix} 1 & \mathbf{0} \\ -\mathbf{l} & I \end{bmatrix} z + \begin{bmatrix} 0 & \mathbf{0} \\ -\mathbf{l}^\# & 0 \end{bmatrix} \overline{z}\,.$$

Assume $\mathcal{L}_1$ is of this type and partition M and $M_\#$ accordingly as

$$M = \begin{bmatrix} m_{1,1} & \mathbf{m}_{1,2}^T \\ \mathbf{m}_{2,1} & M_{2,2} \end{bmatrix} \text{ and } M_\# = \begin{bmatrix} m_{1,1}^\# & \mathbf{m}_{1,2}^{\#\,T} \\ \mathbf{m}_{2,1}^\# & M_{2,2}^\# \end{bmatrix}.$$

We need to determine the vectors $\mathbf{l}$ and $\mathbf{l}^\#$. The linear part of $\mathcal{L}_1 \circ \mathcal{M}$ is

$$\begin{bmatrix} m_{1,1} & \mathbf{m}_{1,2}^T \\ m_{1,1}\,\mathbf{l} + \mathbf{m}_{2,1} + \overline{m_{1,1}^\#}\,\mathbf{l}^\# & \mathbf{l}\,\mathbf{m}_{1,2}^T + \mathbf{l}^\#\,\overline{\mathbf{m}_{1,2}^\#}^T + M_{2,2} \end{bmatrix},$$

while its antilinear part is

$$\begin{bmatrix} m_{1,1}^\# & \mathbf{m}_{1,2}^{\#\,T} \\ m_{1,1}^\#\,\mathbf{l} + \mathbf{m}_{2,1}^\# + \overline{m_{1,1}}\,\mathbf{l}^\# & \mathbf{l}\,\mathbf{m}^{\#\,T}_{1,2} + \mathbf{l}^\#\,\overline{\mathbf{m}_{1,2}}^T + M_{2,2}^\# \end{bmatrix}.$$

[2]Hence the Gauss–Seidel and the Jacobi methods, as well as any other basic iterations, can be devised by splitting a given ℝ-linear operator in $\mathbb{C}^n$ in an obvious way.

In order to have zeros in the first columns of these below the diagonal we take

$$\begin{bmatrix} \mathbf{l} & \mathbf{l}^{\#} \end{bmatrix} = -\begin{bmatrix} \mathbf{m}_{2,1} & \mathbf{m}^{\#}{}_{2,1} \end{bmatrix} \begin{bmatrix} m_{1,1} & m_{1,1}^{\#} \\ \overline{m_{1,1}^{\#}} & \overline{m_{1,1}} \end{bmatrix}^{-1}. \tag{2.10}$$

Thus, we need to assume that $|m_{1,1}| \neq |m_{1,1}^{\#}|$.

This is then repeated with the blocks

$$\mathbf{l}\,\mathbf{m}_{1,2}^{T} + \mathbf{l}^{\#}\,\overline{\mathbf{m}_{1,2}^{\#}}^{T} + M_{2,2} \equiv \widetilde{M}$$

and

$$\mathbf{l}\,\mathbf{m}^{\#}{}_{1,2}^{T} + \mathbf{l}^{\#}\,\overline{\mathbf{m}_{1,2}}^{T} + M_{2,2}^{\#} \equiv \widetilde{M_{\#}}$$

of size $(n-1)$-by-$(n-1)$. If no breakdown occurs, after $n-1$ steps we have an upper triangular $\mathcal{L}_{n-1} \circ \cdots \circ \mathcal{L}_1 \circ \mathcal{M}$. Or equivalently, by using Lemma 2.14 repeatedly, we have an LU-decomposition of $\mathcal{M}$ (since products of lower triangular $\mathbb{R}$-linear operators in $\mathbb{C}^n$ remain lower triangular). The product $\mathcal{L}_1^{-1} \circ \mathcal{L}_2^{-1} \circ \cdots \circ \mathcal{L}_{n-1}^{-1}$ does not involve any computations since the lower triangular parts of its linear and antilinear part are obtained by collecting the vectors from each of its factors.

If $M_{\#} = 0$, then this gives us the standard LU-factorization of M.

Remark. The 2-by-2 matrix in (2.10) is now the "pivot." There are $n-1$ pivot matrices in all. In particular, pivoting is straightforwardly incorporated with the scheme by performing pre-/postoperations with $\mathcal{P}(z) = Pz$, where P is a permutation matrix. This is needed if the inversion in (2.10) is ill-conditioned.

Define the jth principal minor of $\mathcal{M}$ by extracting the upper left j-by-j blocks of M and $M_{\#}$ and compute the value of the corresponding characteristic bivariate polynomial at the origin. It is easy to see that if all the principal minors of $\mathcal{M}$ are nonzero, this LU-factorization exists.

Assuming no breakdown occurs, a Matlab [18] code is as follows:

```
function [L,La,U,Ua]=rl_lu(M,Ma)

% This computes lower triangular L (with unit diagonal), strictly
% lower triangular La, and upper triangular U and Ua such that
%
%    M = L*U+La*conj(Ua)   and   Ma = L*Ua+La*conj(U)
%

n=size(M,1); L=eye(n); La=zeros(n); U=M; Ua=Ma;
for k=2:n ,
  a=U(k-1,k-1);  b=Ua(k-1,k-1);
  w=[U(k:n,k-1),Ua(k:n,k-1)]/[a,b;b',a'];
  L(k:n,k-1)=w(:,1);  La(k:n,k-1)=w(:,2);
  z=zeros(n-k+1,1);  U(k:n,k-1)=z;  Ua(k:n,k-1)=z;
  U(k:n,k:n)=U(k:n,k:n)-w*[U(k-1,k:n);conj(Ua(k-1,k:n))];
  Ua(k:n,k:n)=Ua(k:n,k:n)-w*[Ua(k-1,k:n);conj(U(k-1,k:n))];
end
```

This requires $\sim \frac{4}{3}n^3$ complex flops to compute an LU-factorization of $\mathcal{M} : \mathbb{C}^n \mapsto \mathbb{C}^n$. The actual execution time depends on how well complex arithmetic is implemented on a computer. In practice a pivoting strategy is also needed.

For more symmetry, let $u_{j,j}$, $u^{\#}_{j,j}$ for $j = 1, \ldots, n$ be the diagonal entries of U and $U_{\#}$, respectively. Define a diagonal operator $\mathcal{D}(z) = Dz + D_{\#}\overline{z}$ according to $D = \operatorname{diag}(\overline{u_{j,j}})$ and $D_{\#} = \operatorname{diag}(-u^{\#}_{j,j})$. If $|u_{j,j}| \neq |u^{\#}_{j,j}|$ for $j = 1, \ldots, n$, then $\mathcal{D}$ is invertible and $\mathcal{M} = \mathcal{L} \circ \mathcal{D} \circ \widetilde{\mathcal{U}}$ with an upper triangular $\widetilde{\mathcal{U}}(z) = \mathcal{D}^{-1}(\mathcal{U}(z)) = \widetilde{U}z + \widetilde{U_{\#}}\overline{z}$ such that all the diagonal entries of $\widetilde{U}$ are equal to 1 while $\widetilde{U_{\#}}$ is strictly upper triangular. This gives us a "Cholesky factorization" if $M^* = M$ and ${M_{\#}}^T = M_{\#}$; see item 1 of Proposition 2.1. That is, then $L^* = \widetilde{U}$ and ${L_{\#}}^T = \widetilde{U_{\#}}$. This adds to the fact that this type of real linear operators have many special properties.

For further structure, when M and $M_{\#}$ are banded, the factors $\mathcal{L}$ and $\mathcal{U}$ inherit the (maximum) band structure.

A given $2n$-by-$2n$ real matrix can fail to have an LU-factorization (without pivoting) but has an LU-factorization as an $\mathbb{R}$-linear operator in $\mathbb{C}^n$.

Example 4. To the matrix $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right] \in \mathbb{R}^{2\times 2}$ corresponds $M = 1/2$ and $M_{\#} = -1/2 + i$. For this operator $L = 1$, $L_{\#} = 0$, $U = 1/2$, and $U_{\#} = -1/2 + i$.

The converse holds as well and, in fact, we obtain a curious class of $\mathbb{R}$-linear operators (and the corresponding matrices A). If all the entries of M and $M_{\#}$ have equal modulus (say 1, as in the Schur matrix), then pivoting, i.e., pre- and postoperations by permutations, does not cure a breakdown. Hence an appropriate strategy to avoid a breakdown needs to be devised.

Remark. For large problems the above algorithm can be used as a starting point for devising "ILU-preconditioners" for $\mathcal{M}$. In particular, if $A \in \mathbb{R}^{2n\times 2n}$ is regarded as an $\mathbb{R}$-linear operator in $\mathbb{C}^n$, then this gives rise to new ILU-preconditioning techniques for solving linear systems in $\mathbb{R}^{2n}$.

QR-decomposition. Here we consider slightly more general real linear operators $\mathcal{M} : \mathbb{C}^p \to \mathbb{C}^n$ defined via (2.1) by two matrices $M, M_{\#} \in \mathbb{C}^{n\times p}$. Our aim is to transform $\mathcal{M}$ to upper triangular form by operating with isometries from the left. Clearly, the standard Householder transformations in $\mathbb{C}^n$ could be applied to make either the linear or the antilinear part of $\mathcal{M}$ upper triangular, but we want them in this form simultaneously.

THEOREM 2.15. *For a given $\mathbb{R}$-linear operator $\mathcal{M} : \mathbb{C}^p \to \mathbb{C}^n$ there exists an isometry $\mathcal{Q}$ (in $\mathbb{C}^n$) such that $\mathcal{R} = \mathcal{Q}^{-1} \circ \mathcal{M}$ is upper triangular*

This is proved by the construction that follows. For this purpose we need special elementary isometries.

For given $x, y \in \mathbb{C}^n$ we want a real linear isometry that maps x and y in the direction $e = [1\,0\,\cdots\,0]^T$. If x and y are linearly dependent, a standard Householder transformation in $\mathbb{C}^n$ will do. So, let us assume that x and y are linearly independent over $\mathbb{R}$. We look for an isometry in the form $\mathcal{H}(z) = z - UU^*z - UU^T\overline{z}$, where $U \in \mathbb{C}^{n\times 2}$ is such that $\operatorname{Re}(U^*U) = I$; see Example 3. We call this a real linear Householder transformation. Writing $V = [x\ y] \in \mathbb{C}^{n\times 2}$ gives us the equation

$$V - 2\,U\,\operatorname{Re}(U^*V) = e\,a^* \tag{2.11}$$

for some $a \in \mathbb{C}^2$. Hence U is of the form $U = (V - e\,a^*)R$, where $R \in \mathbb{R}^{2\times 2}$. By multiplying (2.11) with U^* we obtain $\operatorname{Re}(U^*(V + e\,a^*)) = 0$. Therefore (2.11) holds if $\operatorname{Re}((V - e\,a^*)^*(V + e\,a^*)) = 0$. By setting $w = V^*e$ this becomes

$$\operatorname{Re}(V^*V + w\,a^* - a\,w^* - a\,a^*) = 0.$$

Vector $c = \operatorname{Re}(V^*V)^{\frac{1}{2}}\ \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$ satisfies $\operatorname{Re}(V^*V - c\,c^*) = 0$. We try $a = \eta\,c$, where $|\eta| = 1$,

so that $\mathrm{Re}(V^*V - a\,a^*) = 0$. Equation $\mathrm{Re}(w\,a^* - a\,w^*) = 0$ amounts to

$$\mathrm{Re}(w_1\overline{\eta c_2} - \eta c_1\overline{w_2}) = 0\,. \tag{2.12}$$

This is equivalent to $q\eta + \overline{q\eta} = 0$, where $q = c_2\overline{w_1} - c_1\overline{w_2}$. Thus (2.12) is satisfied for $\eta = i\,\overline{q}/|q|$. Finally, we get U by orthonormalizing the columns of $V - e\,a^*$ with respect to the inner product $\langle u, v\rangle = \mathrm{Re}(u^*v)$.

Since $\mathcal{H}^2 = I$ we also have $\mathcal{H}(e\,a^*) = V$.

The following Matlab code finds U:

```
function U=rl_H(x,y)

  V=[x,y]; n=length(x);
  c=real(V'*V)^(1/2)*[1;i];
  q=V(1,:)*[-c(2);c(1)];
  V(1,:)=V(1,:)-i*sign(q)*c';
  [Q,R]=qr([real(V);imag(V)],0);
  U=Q(1:n,:)+i*Q(n+1:2*n,:);
```

For the QR-decomposition we first want a real linear Householder transformation such that the first columns of

$$\widehat{M} = (I - UU^*)M - UU^T\,\overline{M_\#} \qquad \text{and} \qquad \widehat{M_\#} = (I - UU^*)M_\# - UU^T\,\overline{M}$$

are multiples of e. Let m and $m^\#$ be the first columns of M and $M_\#$. Then

$$p = (I - UU^*)m - UU^T\,\overline{m^\#} \qquad \text{and} \qquad q = (I - UU^*)m^\# - UU^T\,\overline{m}$$

are both multiples of e if

$$\begin{aligned} p + q &= (I - UU^*)(m + m^\#) - UU^T\,\overline{(m + m^\#)},\\ i(p - q) &= (I - UU^*)\,i(m - m^\#) - UU^T\,\overline{i(m - m^\#)} \end{aligned}$$

are such. Thus we take the real linear Householder transformation that maps $m + m^\#$ and $i(m - m^\#)$ to multiples of e. Then the first columns of $\widehat{M}$ and $\widehat{M_\#}$ have zeros below the first entries.

After this we continue similarly with the lower right $(n-1)$-by-$(p-1)$ blocks of $\widehat{M}$ and $\widehat{M_\#}$. Below is the Matlab code for this decomposition:

```
function [Q,Qa,R,Ra]=rl_qr(M,Ma)

  % This constructs a real linear isometry z -> Q*z+Qa*conj(z)
  % and upper triangular R and Ra such that
  %     M = Q*R+Qa*conj(Ra)    and    Ma = Q*Ra+Qa*conj(R)

[n,p]=size(M); R=M; Ra=Ma; Q=eye(n); Qa=zeros(n);
for k=1:min(p,n-1) , kn=k:n; kp=k:p;
  x=R(kn,k); y=Ra(kn,k);
  U=rl_H(x+y,i*(x-y));
  W=U'*R(kn,kp)+conj(U'*Ra(kn,kp));
  R(kn,kp)=R(kn,kp)-U*W; Ra(kn,kp)=Ra(kn,kp)-U*conj(W);
  W=Q(:,kn)*U+Qa(:,kn)*conj(U);
  Q(:,kn)=Q(:,kn)-W*U'; Qa(:,kn)=Qa(:,kn)-W*transpose(U);
end
```

This implementation requires $\sim \frac{40}{3}\, n^2 p$ complex flops. With back substitution this algorithm can be used to solve overdetermined $\mathbb{R}$-linear systems.

Remark. The prescribed real linear Householder transformations can also be used in computing an isometry $\mathcal{U}$ such that $\mathcal{U}^{-1} \circ \mathcal{M} \circ \mathcal{U}$ has its linear and antilinear parts in Hessenberg form.

Schur decomposition. Here we consider bringing a given real linear operator to upper triangular form under a real linear isometric similarity transformation. The construction of this part proves the following.

THEOREM 2.16. *For a given $\mathbb{R}$-linear operator $\mathcal{M}$ there exists an isometry $\mathcal{U}$ such that $\mathcal{T} = \mathcal{U}^{-1} \circ \mathcal{M} \circ \mathcal{U}$ is upper triangular.*

We need the following auxiliary result.

LEMMA 2.17. *There exist vectors $x, y \in \mathbb{C}^n$, linearly independent over $\mathbb{R}$, and $B \in \mathbb{R}^{2\times 2}$ such that $\mathcal{M}([x\ y]) = [x\ y]\, B$.*

Proof. Let $A \in \mathbb{R}^{2n\times 2n}$ correspond to $\mathcal{M}$. Take $u, v \in \mathbb{R}^{2n}$ either as

- two linearly independent real eigenvectors of A, or
- a real eigenvector u and vector v such that $Av - \lambda v = u$, or
- the real and imaginary parts of an eigenvector corresponding to a nonreal eigenvalue.

Set $[x\ y] = [I\ iI]\, [u\ v]$. ∎

Now, let x, y be as in the previous lemma and take a real linear Householder transformation such that $\mathcal{H}([x\ y]) = e\, a^*$. Then also $\mathcal{H}(e\, a^*) = [x\ y]$. Consider

$$\widehat{\mathcal{M}}(z) = \mathcal{H}(\mathcal{M}(\mathcal{H}(z))) = \widehat{M} z + \widehat{M_\#}\, \overline{z}$$

and let $\widehat{m}$ and $\widehat{m}^\#$ be the first columns of $\widehat{M}$ and $\widehat{M_\#}$. We have

$$\begin{bmatrix} \widehat{m} & \widehat{m}^\# \end{bmatrix} \begin{bmatrix} a^* \\ a^T \end{bmatrix} = \widehat{\mathcal{M}}(e\, a^*) = \mathcal{H}(\mathcal{M}([x\ y])) = \mathcal{H}([x\ y]\, B) = e\, a^*\, B\,,$$

since B is real. Here $\det \left[\begin{smallmatrix} a^* \\ a^T \end{smallmatrix}\right] = \overline{a_1}\, a_2 - a_1\, \overline{a_2} \neq 0$ unless $\overline{a_1}\, a_2 \in \mathbb{R}$. But the latter would imply that x and y are linearly dependent over $\mathbb{R}$—a contradiction. Hence both $\widehat{m}$ and $\widehat{m}^\#$ are multiples of e.

Continue similarly with the lower right $(n-1)$-by-$(n-1)$ blocks of $\widehat{M}$ and $\widehat{M_\#}$ to obtain $\mathcal{U}$ as a composition of real linear Householder transformations.

Due to Proposition 2.8, the Schur decomposition $\mathcal{U}^{-1} \circ \mathcal{M} \circ \mathcal{U}$ of $\mathcal{M}$ is diagonal in case $M^* = M$ and ${M_\#}^T = M_\#$.

Remark. Items 1 and 2 of Proposition 2.1 hence give us very special real linear operators. In view of this, to the polar decomposition of $A \in \mathbb{R}^{2n\times 2n}$ corresponds $\mathcal{M} = \mathcal{U} \circ \mathcal{S}$, where $\mathcal{U}$ is an isometry and $\mathcal{S}(z) = Sz + S_\# \overline{z}$ with $S^* = S$ and ${S_\#}^T = S_\#$.

With a small rank $M_\#$ the operator $\mathcal{M}$ can be regarded as "almost" $\mathbb{C}$-linear. So can its inverse in the following sense.

PROPOSITION 2.18. *Let $\mathcal{M}(z) = Mz + M_\# \overline{z}$ be invertible with $\mathcal{M}^{-1}(z) = Rz + R_\# \overline{z}$. If M is invertible as well, then* $\operatorname{rank}(R - M^{-1}) \le \operatorname{rank}(M_\#)$ *and* $\operatorname{rank}(R_\#) = \operatorname{rank}(M_\#)$.

Proof. We have $\mathcal{M}^{-1}(\mathcal{M}z) = (RM + R_\#\overline{M_\#})z + (RM_\# + R_\#\overline{M})\overline{z}$. For this to be the identity we obtain the conditions

$$R_\# = -RM_\#\overline{M}^{-1} \text{ and } R = (M - M_\#\overline{M}^{-1}\overline{M_\#})^{-1}. \tag{2.13}$$

With R we can use the Sherman–Morrison formula to obtain the claims. ∎

In this case one option is to use standard $\mathbb{C}$-linear algorithms with (2.13) and the Sherman–Morrison formula to find the inverse of $\mathcal{M}$.

Not all the matrix factorizations have a particularly interesting analogue for $\mathbb{R}$-linear operators in $\mathbb{C}^n$. For instance, assume the real form $A \in \mathbb{R}^{2n\times 2n}$ of $\mathcal{M}$ is nonderogatory so that $A = T^{-1}\widetilde{C}T$ with a companion matrix $\widetilde{C}$. Since all the factors are real, we have $\mathcal{M} = \mathcal{T}^{-1}\circ\mathcal{C}\circ\mathcal{T}$ with $\mathcal{C}(z) = Cz + C_{\#}\overline{z}$ such that C is a companion matrix while $C_{\#}$ is a rank-1 matrix with one nonzero column. Since the spectrum of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$ is not preserved under a general $\mathbb{R}$-linear similarity transformation, this factorization may not be very useful (aside from giving a very structured factor $\mathcal{C}$). In general T cannot be found such that the corresponding $\mathcal{T}$ would be $\mathbb{C}$-linear.

2.3. Miscellaneous remarks. We also have a Neumann-type series expansion for the inverse. Consider first the operator $\mathcal{M}_\kappa$.

THEOREM 2.19. *Assume $M_{\#} \in \mathbb{C}^{n\times n}$, and $\lambda \in \mathbb{C}$ is such that $||M_{\#}|| < |\lambda|$ holds. Then $(\lambda I - \mathcal{M}_0)^{-1}(z) = R(\lambda)z + R_{\#}(\lambda)\overline{z}$ with*

$$R(\lambda) = \sum_{j=0}^{\infty} \frac{(M_{\#}\overline{M_{\#}})^j}{\lambda|\lambda|^{2j}} \text{ and } R_{\#}(\lambda) = \sum_{j=0}^{\infty} \frac{M_{\#}(\overline{M_{\#}}M_{\#})^j}{|\lambda|^{2(j+1)}}.$$

Proof. By making an ansatz

$$z = \frac{b}{\lambda} + \frac{M_{\#}\overline{b}}{|\lambda|^2} + \frac{M_{\#}\overline{M_{\#}}b}{\lambda|\lambda|^2} + \frac{M_{\#}\overline{M_{\#}}M_{\#}\overline{b}}{|\lambda|^4} + \frac{M_{\#}\overline{M_{\#}}M_{\#}\overline{M_{\#}}b}{\lambda|\lambda|^4} + \cdots,$$

it is straightforward to verify that z converges and solves the equation $\lambda z - M_{\#}\overline{z} = b$ for any $b \in \mathbb{C}^n$. Separating the linear and antilinear terms (that is, the matrices multiplying b and $\overline{b}$, respectively) from this sequence gives $R(\lambda)$ and $R_{\#}(\lambda)$. ∎

For a general $\mathbb{R}$-linear operator $\mathcal{M}$ we have

$$(rI - \mathcal{M})^{-1} = \sum_{j=0}^{\infty} \frac{\mathcal{M}^j}{r^{j+1}}, \tag{2.14}$$

whenever $r \in \mathbb{R}$ and $||\mathcal{M}|| < |r|$. Assume $\lambda \in \mathbb{C}$. Since solving $\lambda z - Mz - M_{\#}\overline{z} = b$ is equivalent to solving $z - \frac{M}{\lambda}z - \frac{M_{\#}}{\lambda}\overline{z} = \frac{b}{\lambda}$, we can employ (2.14) with this latter problem. A substitution to (2.14) gives a series expansion for the linear and antilinear parts of $\mathcal{R}(\lambda) = (\lambda I - \mathcal{M})^{-1}$ as

$$R(\lambda) = \frac{I}{\lambda} + \frac{M}{\lambda^2} + \frac{M^2}{\lambda^3} + \frac{M_{\#}\overline{M_{\#}}}{\lambda|\lambda|^2}$$

$$+\frac{M^3}{\lambda^4} + \frac{MM_{\#}\overline{M_{\#}}}{\lambda^2|\lambda|^2} + \frac{M_{\#}\overline{M}^2}{\overline{\lambda}^2|\lambda|^2} + \frac{M_{\#}\overline{MM_{\#}}}{|\lambda|^4} + \frac{M_{\#}\overline{M_{\#}}M}{\lambda^2|\lambda|^2} + \cdots$$

and

$$R_{\#}(\lambda) = \frac{M_{\#}}{|\lambda|^2} + \frac{MM_{\#}}{\lambda|\lambda|^2} + \frac{M_{\#}\overline{M}}{\overline{\lambda}|\lambda|^2} + \frac{M^2M_{\#}}{\lambda^2|\lambda|^2} + \frac{MM_{\#}\overline{M}}{|\lambda|^4} + \frac{M_{\#}\overline{M_{\#}}M_{\#}}{|\lambda|^4} + \cdots.$$

Remark. Since the set of $\mathbb{R}$-linear operators in $\mathbb{C}^n$ is a normed algebra over $\mathbb{R}$, $\lim_{j\to\infty} ||\mathcal{M}^j||^{1/j}$ exists and gives the spectral radius of the real form of $\mathcal{M}$. However,

its connection with the spectrum of $\mathcal{M}$ is not obvious (except when $M_\# = 0$) since $\sigma(\mathcal{M})$ can even be empty.

Regardless of the size of the spectrum, the minimal polynomial of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$ is well defined.

THEOREM 2.20. *Let $\mathcal{M}$ be an $\mathbb{R}$-linear operator in $\mathbb{C}^n$. Then there exists a monic polynomial p of degree at most $2n$ such that $p(\mathcal{M}) = 0$.*

Proof. Take p to be the minimal polynomial of $A \in \mathbb{R}^{2n\times 2n}$ corresponding to the real formulation (2.2) of $\mathcal{M}$. Since p has only real coefficients, $p(\mathcal{M})$ is also zero. ∎

For iterative methods, the following is of interest.

COROLLARY 2.21. *If $\mathcal{M}$ is invertible, then $\mathcal{M}(q(\mathcal{M})) = I$ for a polynomial q of degree $2n-1$ at most.*

Proof. Take $p(\lambda)/p(0) = \lambda q(\lambda) - 1$, which clearly has real coefficients. Therefore the equivalent real operator in $\mathbb{R}^{2n}$ gives the identity. ∎

Example 5. In the context of forming polynomials in an $\mathbb{R}$-linear mapping in $\mathbb{C}^n$ many interesting classes of operators arise. In [20] there is an operator considered, the so-called Friedrichs operator [6], whose square is a $\mathbb{C}$-linear mapping in $\mathbb{C}^n$. Generalizing this, it is an interesting problem to find, for a given $\mathcal{M}$, a monic polynomial of the lowest possible degree such that $p(\mathcal{M})$ is $\mathbb{C}$-linear (or $\mathbb{C}$-antilinear).

Rank-1 matrices are fundamental for matrix computations. In fact, let $M = m_1 m_2^*$ and $M_\# = n_1 n_2^*$ both be of rank 1. Then there are three possibilities for the multiplicity indexes (see Example 1) of the eigenvalue 0 of $\mathcal{M}$; i.e., we can have four different types of real linear low rank operators, listed in Table 1.

TABLE 1
Options for $\mathbb{R}$-linear operators in $\mathbb{C}^n$ with rank-1 matrices M and $M_\#$.

	$\dim(\mathrm{span}\{m_1, n_1\}) = 2$	$\dim(\mathrm{span}\{m_1, n_1\}) = 1$
$\dim(\mathrm{span}\{m_2, \overline{n}_2\}) = 2$	$(\frac{2n-2}{2}, n-2)$	$(\frac{2n-1}{2}, n-2)$
$\dim(\mathrm{span}\{m_2, \overline{n}_2\}) = 1$	$(\frac{2n-1}{2}, n-1)$	$(\frac{2n-1}{2}, n-1)$

Example 6. Let $\sigma_1 \left[\begin{smallmatrix} u_a \\ u_b \end{smallmatrix}\right] \left[\begin{smallmatrix} v_a^* & v_b^* \end{smallmatrix}\right] \in \mathbb{R}^{2n\times 2n}$ the best rank-1 approximation to $A \in \mathbb{R}^{2n\times 2n}$ from its SVD; i.e., σ_1 is the largest singular value of A and $u_j, v_j \in \mathbb{R}^n$ for $j = a, b$. For the corresponding real linear operator in $\mathbb{C}^n$ the respective approximation is $\frac{\sigma_1}{2}(u_a + iu_b)\big((v_a^* - iv_b^*)z + (v_a^* + iv_b^*)\overline{z}\big) = \sigma_1\, u\, \mathrm{Re}(v^* z)$ with $u = u_a + iu_b$ and $v = v_a + iv_b$. In the classification of Table 1 this is in the lower right corner.

Repeating the construction of the preceding example with each rank-1 term in the SVD of A, we obtain an expansion

$$\mathcal{M}(z) = \textstyle\sum_{j=1}^{2n} \sigma_j\, u_j\, \mathrm{Re}(v_j^* z)$$

for $\mathcal{M}$. Although this is a potentially useful representation of $\mathcal{M}$, at this point we are not sure whether it should be called the SVD of $\mathcal{M}$.

Let $V_1, V_2 \subset \mathbb{C}^n$ be two subspaces of dimension k (over $\mathbb{C}$ as usual) and let $I_j : V_j \mapsto \mathbb{C}^k$ be an isometric isomorphism for $j = 1, 2$. Define $\widetilde{\mathcal{P}}$ via

$$\begin{array}{ccc} V_1 & \xrightarrow{\widetilde{\mathcal{P}}} & V_2 \\ \downarrow I_1 & & \uparrow I_2^{-1} \\ \mathbb{C}^k & \xrightarrow{\mathcal{U}} & \mathbb{C}^k \end{array}, \tag{2.15}$$

where $\mathcal{U}$ is an $\mathbb{R}$-linear isometry in $\mathbb{C}^k$. Then $\mathcal{P} = \widetilde{\mathcal{P}} \oplus 0$ gives an $\mathbb{R}$-linear *partial isometry* in $\mathbb{C}^n$; i.e., $\|\mathcal{P}(z)\| = \|z\|$ for $z \in V_1$ while $\mathcal{P}(z) = 0$ for $z \in V_1^\perp$.

3. Iterative methods for solving $\mathbb{R}$-linear problems in $\mathbb{C}^n$. Assume $Q_k \in \mathbb{C}^{n\times k}$ with orthonormal columns has been generated (typically $k \ll n$). Then a low-dimensional approximation to the problem (1.1) is given by

$$\mathcal{M}^{(k)}(w_k) = Q_k^* M Q_k w_k + Q_k^* M_\# \overline{Q}_k \overline{w}_k = Q_k^* b \tag{3.1}$$

so that $z_k = Q_k w_k$ yields the corresponding Galerkin approximation for the solution. The arising $\mathbb{R}$-linear mapping in $\mathbb{C}^k$ can also be used in approximating the spectrum of $\mathcal{M}$ via a Ritz-type construction to have "Ritz curves." In particular, the subspace spanned by the columns of Q_k is invariant for $\mathcal{M}$ if and only if

$$(I - Q_k Q_k^*) M Q_k = 0 \ \text{ and } \ (I - Q_k Q_k^*) M_\# \overline{Q}_k = 0.$$

If this holds, then we have $\sigma(\mathcal{M}^{(k)}) \subset \sigma(\mathcal{M})$. Otherwise good approximations (in some sense) can be expected when the matrices on the left-hand side are small in norm.

3.1. The case of $\mathcal{M}_\kappa$. To compute Q_k with an iterative method, consider first the simplest case involving the operator $\mathcal{M}_\kappa$ for $\kappa \in \mathbb{C}$. For this we can use a minimal residual approach once we generate Q_k with an Arnoldi-type iteration [1] and replace Q_k^* in (3.1) with Q_{k+1}^*. Then what remains is to solve the arising low order problem with the least squares methods.

To this end we first apply $\mathcal{M}_\kappa$ to a starting vector $b \in \mathbb{C}^n$. This yields us

$$\kappa b + M_\# \overline{b}.$$

Orthogonalizing this against b yields $\alpha_1^1 b + \alpha_1^2 M_\# \overline{b}$ with $\alpha_1^1, \alpha_1^2 \in \mathbb{C}$. Applying $\mathcal{M}_\kappa$ to this vector gives

$$\kappa\alpha_1^1 b + (\kappa\alpha_1^2 + \overline{\alpha_1^1}) M_\# \overline{b} + \overline{\alpha_1^2} M_\# \overline{M_\#} b.$$

Orthogonalizing this against b and $\alpha_1^1 b + \alpha_1^2 M_\# \overline{b}$ yields a vector which is a linear combination of the vectors b, $M_\#\overline{b}$, and $M_\#\overline{M_\#}b$. An application of $\mathcal{M}_\kappa$ to this vector and then performing an orthogonalization yields a linear combination of the vectors b, $M_\#\overline{b}$, $M_\#\overline{M_\#}b$, and $M_\#\overline{M_\#}M_\#\overline{b}$. Continuing this inductively proves the following.

THEOREM 3.1. *Let $\kappa \in \mathbb{C}$, $M_\# \in \mathbb{C}^{n\times n}$, and $b \in \mathbb{C}^n$. Then the Arnoldi method with $\mathcal{M}_\kappa$ gives an orthonormal basis $\{q_1, q_2, \dots\}$ of the Krylov subspace*

$$\operatorname{span}\left\{b, M_\#\overline{b}, M_\#\overline{M_\#}b, M_\#\overline{M_\#}M_\#\overline{b}, \dots\right\}. \tag{3.2}$$

Remark. Solving $\kappa z + M_\# \overline{z} = b$ with a direct method is naturally equivalent to solving $\overline{M_\#} z + \overline{\kappa}\overline{z} = \overline{b}$. However, an execution of the Arnoldi method with the complex conjugate $\overline{\mathcal{M}_\kappa}$ of $\mathcal{M}_\kappa$ using the starting vector $\overline{b}$ does not seem to generate a subspace with a simple spanning set like that of (3.2) unless simplifying assumptions are made.

By inspecting its spanning set, we can view (3.2) as a block Krylov subspace generated with the matrix $M_\#\overline{M_\#}$ by using the starting vectors $\{b, M_\#\overline{b}\}$. In particular, a matrix $M_\# \in \mathbb{C}^{n\times n}$ is congruence normal if $M_\#\overline{M_\#}$ is normal; see [7] and the references therein. In this case the ideas of [10, 14] can be used for generating this subspace with a recurrence whose length grows very slowly.

If $\deg(M_\#\overline{M_\#})$, the degree of the minimal polynomial of $M_\#\overline{M_\#}$, is moderate, then we have a nontrivial invariant subspace of $\mathcal{M}_\kappa$ with (3.2).

COROLLARY 3.2. *The dimension of* (3.2) *is* $\min\{\operatorname{rank}(M_\#)+1, 2\deg(M_\#\overline{M_\#})\}$ *at most.*

Proof. The claim follows by rewriting (3.2) as the sum of two subspaces,

$$\mathcal{K}(M_\#\overline{M_\#}; b) + M_\#\overline{\mathcal{K}(M_\#\overline{M_\#}; b)}, \tag{3.3}$$

where $\mathcal{K}(M_\#\overline{M_\#}; b) = \operatorname{span}\{b, M_\#\overline{M_\#}b, (M_\#\overline{M_\#})^2 b, \dots\}$. ∎

In view of iterative methods, this illustrates how the bound of Theorem 2.20 can be pessimistic.

Remark. Any invariant subspace of $\mathcal{M}_\kappa$ is necessarily invariant for $M_\#\overline{M_\#}$. For the converse, $\mathcal{K}(M_\#\overline{M_\#}; b)$ is an invariant subspace of $M_\#\overline{M_\#}$ for any vector $b \in \mathbb{C}^n$. Hence (3.3) is the smallest invariant subspace of $\mathcal{M}_\kappa$ containing $\mathcal{K}(M_\#\overline{M_\#}; b)$. For instance, if b is an eigenvector of $M_\#\overline{M_\#}$, then the dimension of (3.3) is either 1 or 2. Both cases are possible.

We denote by W_k the subspace spanned by the first k vectors in (3.2). Clearly, $\mathcal{M}_\kappa(W_k) \subset W_{k+1}$. This implies that the resulting canonical form (3.1) consists of a diagonal and a Hessenberg matrix for the linear and antilinear parts of $\mathcal{M}_\kappa$, respectively. Writing $Q_k = [q_1\ q_2\ \dots\ q_k]$ we get the following.

THEOREM 3.3. *The Arnoldi method with* $\mathcal{M}_\kappa$ *gives a Hessenberg matrix* $Q_k^* M_\# \overline{Q}_k$ *for* $k = 1, 2, \dots$.

Proof. If W_j denotes the subspace spanned by the first j vectors in (3.2), then $M_\#$ maps $\overline{W}_j$ into W_{j+1} for every $j > 0$. ∎

If no breakdown occurs, with $k = n$ we have performed a unitary consimilarity transformation of $M_\#$; see [8] for more on the concept of consimilarity.

With iterative methods one is always interested in the length of the recurrence to have less expensive steps.

THEOREM 3.4. *If* ${M_\#}^T = cM_\#$ *for* $c = \pm 1$, *then the Arnoldi method with* $\mathcal{M}_\kappa$ *is realizable with a* 3*-term recurrence.*

Proof. Let $q_0, \dots, q_{j-1}$ denote the orthonormal basis of W_j generated with the Arnoldi method. Then

$$(\mathcal{M}_\kappa(q_{j-1}), q_l) = (\kappa q_{j-1}, q_l) + (\overline{q}_{j-1}, {M_\#}^* q_l),$$

where the first inner product is zero for $j - l > 1$. Hence we have $(\overline{q}_{j-1}, {M_\#}^* q_l) = \overline{({q_{j-1}}, {M_\#}^T \overline{q}_l)} = c\overline{(q_{j-1}, M_\#\overline{q}_l)} = 0$ for $j - l > 2$. ∎

Under these assumptions the matrices $Q^*\kappa I Q$ and $Q^* M_\# \overline{Q}$ are diagonal and tridiagonal, respectively. With $c = -1$ the diagonal entries of $Q^* M_\# \overline{Q}$ equal zero; i.e., we then get a real skew-symmetric matrix. Hence the eigenvalues of the matrices $M_\#$ and $Q^* M_\# \overline{Q}$ can differ dramatically even though the eigenvalues of the mappings $z \mapsto M_\# \overline{z}$ and $z \mapsto Q^* M_\# \overline{Q}\overline{z}$ are the same.

Remark. Since Q is unitary, the singular values of the matrix $M_\#$ equal those of $Q^* M_\# \overline{Q}$. Therefore, under the assumptions of Theorem 3.4, the singular values of $M_\#$ can be approximated with an iterative method relying on a 3-term recurrence.

The following is of use for preconditioning the problem (1.1) with the inverse of M from the left.

PROPOSITION 3.5. *Let* $\mathcal{M}(z) = Mz + M_\#\overline{z}$ *be diagonalizable in a* $\mathbb{C}$*-linear unitary similarity transformation. If* M *is invertible, then* $M^{-1}M_\#$ *is complex symmetric.*

Returning to our original problem, consider iteratively solving the real linear system $\mathcal{M}_\kappa(z) = b$ for $b \in \mathbb{C}^n$. With the iteration prescribed we have $M_\#\,\overline{Q_k}\,\overline{w}$ in the

range of Q_{k+1} for all $w \in \mathbb{C}^k$. Therefore

$$\begin{aligned} \left\|\kappa Q_k w_k + M_\# \overline{Q}_k \overline{w}_k - b\right\| &= \left\|\kappa Q_{k+1}^* Q_k w_k + Q_{k+1}^* M_\# \overline{Q}_k \overline{w}_k - Q_{k+1}^* b\right\| \\ &= \left\|\kappa \widetilde{I}_k w_k + \widetilde{H}_k \overline{w}_k - \|b\| \, e_1\right\|, \end{aligned} \tag{3.4}$$

where $\widetilde{I}_k \in \mathbb{C}^{(k+1)\times k}$ is the identity matrix augmented with the row of zeros and $\widetilde{H}_k$ is a $(k+1)$-by-k Hessenberg matrix. Hence, solving the system $\kappa z + M_\# \overline{z} = b$ approximately with the corresponding minimal residual approach amounts to finding the minimum of the last expression in (3.4), e.g., by employing our real linear QR-decomposition.

A Matlab implementation of this method is as follows:

```
function x=rl_GMRES(kappa,Ma,b,tol)

nb=norm(b); Q=b/nb; H=[]; Ha=[]; eb=nb; err=1; j=0;
while err>tol ,j=j+1;
  r=Ma*conj(Q(:,j));
  for l=1:j,
    h=Q(:,l)'*r; r=r-Q(:,l)*h;
    Ha(l,j)=h; end
  nr=norm(r); Q=[Q,r/nr];
  jj=j:j+1; H(jj,j)=[kappa;0]; Ha(j+1,j)=nr; eb(j+1,1)=0;
  for l=1:j-1, U=UM{l};  ll=l:l+1;
    W=U'*H(ll,j)+conj(U'*Ha(ll,j));
    H(ll,j)=H(ll,j)-U*W; Ha(ll,j)=Ha(ll,j)-U*conj(W); end
  x=H(jj,j); y=Ha(jj,j);
  U=rl_H(x+y,i*(x-y)); W=U'*H(jj,j)+conj(U'*Ha(jj,j));
  H(jj,j)=H(jj,j)-U*W; Ha(jj,j)=Ha(jj,j)-U*conj(W);
  eb(jj)=eb(jj)-2*U*real(U'*eb(jj)); UM{j}=U;
  err=abs(eb(j+1))/nb;
end
w=rl_ut_solve(H(1:j,:),Ha(1:j,:),eb(1:j));
z=Q(:,1:j)*w;
```

Note that (similarly to the standard implementation of GMRES) the Hessenberg matrix is transformed isometrically to an upper triangular form while it is being built.

The work and storage needed with this method (as a function of the number of steps) are comparable to those of GMRES [21]. Further, we have the following proposition.

PROPOSITION 3.6. *The method above is at least as fast as the standard GMRES method applied to the real form* $A \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = \left[\begin{smallmatrix} \mathrm{Re}(b) \\ \mathrm{Im}(b) \end{smallmatrix}\right]$ *of the problem.*

Proof. Minimizing (3.4) amounts to finding the minimum of $\|\mathcal{M}_\kappa(z) - b\|$ for $z \in W_k$. GMRES applied to the real form minimizes the same but only in the set of real linear combinations of $b, \mathcal{M}_\kappa(b), \ldots, \mathcal{M}_\kappa^{k-1}(b)$, which is a subset of W_k. ∎

The number of steps needed for the exact solution is bounded according to Corollary 3.2.

The case of $\kappa = 0$ and ${M_\#}^T = M_\#$ has been studied in [3].

With $\kappa = 0$ we have a "conjugate GMRES" algorithm for solving $M_\# \overline{z} = b$ (which is equivalent to solving $\overline{M_\#} z = \overline{b}$ when direct methods are used). It coincides with GMRES if and only if $M_\#$ and b are real.

Example 7. One can readily construct problems in which solving $M_\# \overline{z} = b$ is much faster than solving $\overline{M_\#} z = \overline{b}$, and vice versa. To give an extreme example, if $M_\# \overline{M_\#} = I$, then `rl_GMRES` converges in one step. However, iteratively solving $\overline{M_\#} z = \overline{b}$ can be disastrous since we have $M_\# = S\overline{S}^{-1}$ for an invertible $S \in \mathbb{C}^{n\times n}$ by [9, Lemma 4.6.9]. In particular, the degree of $\overline{M_\#}$ can be n with a very unfavorable eigenvalue distribution for speedy convergence of GMRES.

The critical structure in the preceding example was the fact that any matrix $M_\#$ is congruent similar to a real matrix [8]; i.e., $M_\# = SR\overline{S}^{-1}$ for an invertible $S \in \mathbb{C}^{n\times n}$ and $R \in \mathbb{R}^{n\times n}$. In Example 7 we had $R = I$. Hence we can infer, more generally, that if the degree of the factor R is moderate, then `rl_GMRES` converges fast.

Remark. As a curiosity, because of (2.4) we could also consider separately the linear and antilinear parts of $\mathcal{M}_\kappa^{-1}$. This amounts to solving, after multiplying the second system with $M_\#$ from the left, two standard $\mathbb{C}$-linear systems in $\mathbb{C}^n$ involving translations of $M_\#\overline{M_\#}$. In particular, if $M_\#$ is congruence normal, then the 3-term recurrence of [11] can be employed to this end.

3.2. The general case. To compute Q_k to have an approximation (3.1) for a general $\mathbb{R}$-linear operator $\mathcal{M}$ using an Arnoldi-type iteration is straightforward by orthogonalizing $\mathcal{M}(q_j)$ against the vectors $q_1, \ldots, q_j$ computed so far for $j = 1, \ldots, k-1$.

As opposed to the case $\mathcal{M}_\kappa$, this iteration is less satisfactory since Hessenberg matrices do not arise. Even with $\mathcal{M}(z) = Mz + \kappa\overline{z}$ for $\kappa \in \mathbb{C}$, no particular structure seems to appear. Moreover, we do not have the critical property $\mathcal{M}(W_k) \subset W_{k+1}$ for a minimal residual approach. Here W_k denotes the span of the vectors generated after $k-1$ steps. Still, for the number of steps we have the following analogy of [15, Proposition 2.6].

PROPOSITION 3.7. *The Arnoldi method with $\mathcal{M}(z) = Mz + M_\#\overline{z}$ generates at most* $\deg(M)(\operatorname{rank}(M_\#) + 1)$ *linearly independent vectors.*

Proof. Denote by $\mathcal{K}_j(M; q_0) = \operatorname{span}\{q_0, Mq_0, \ldots, M^{j-1}q_0\}$, where $q_0 = b/\|b\|$. If $v_1, \ldots, v_k$ is a basis of the range of $M_\#$, then

$$q_2 = \alpha_1^2(Mq_0 + M_\#\overline{q}_0) - \alpha_2^2 q_0 \in \mathcal{K}_2(M; q_0) + \operatorname{span}\{v_l\}_{l=1,\ldots,k}$$

with $\alpha_1^2, \alpha_2^2 \in \mathbb{C}$. Similarly,

$$q_3 \in \mathcal{K}_3(M; q_0) + \operatorname{span}\{M^j v_l\}_{\substack{j=0,1\\ l=1,\ldots,k}}$$

so that the induction step becomes clear. Since

$$\mathcal{K}_n(M; q_0) + \operatorname{span}\{M^j v_l\}_{\substack{j=0,\ldots,n\\ l=1,\ldots,k}}$$

has dimension at most $\deg(M)(\operatorname{rank}(M_\#) + 1)$, the assertion is proved. ∎

This also implies that if $\deg(M)(\operatorname{rank}(M_\#) + 1) < n$, then $\mathcal{M}$ always has an invariant subspace.

Instead of a minimal residual approximation to the solution of the system $\mathcal{M}(z) = b$, we compute a Galerkin approximation by using (3.1). (In the numerical linear algebra community also the abbreviation FOM is used in case iterative methods are executed for generating a Galerkin approximation.)

For a minimum residual method we should augment Q_k with (typically) k extra orthonormal vectors such that the resulting span would include the range of $\mathcal{M}_{|_{W_k}}$. This seems to become rather uneconomical.

3.3. Cost, restarting, and related remarks. For an iterative method to be preferred over a direct method, typically the crucial bottleneck is the cost of matrix-vector products. Here all the standard ideas, such as using the FFT techniques, apply in an obvious way.

As with GMRES, restarting may be needed to save storage. In connection with this, there is now the additional possibility of solving the conjugated problem $\overline{\mathcal{M}(z)} = \overline{M_{\#}}z + \overline{M}\overline{z} = \overline{b}$ instead of the original system $\mathcal{M}(z) = b$. Either of these two options can be chosen before each new restart.

The Krylov subspace methods suggested above were based on an iterative generation of orthogonal projectors. These are very particular types of partial isometries (2.15). Hence the possibility of iteratively computing more general real linear partial isometries and using them in solving linear systems approximately needs to be studied further.

We have considered only methods that consume storage linearly. Devising a quasi-minimal residual type of iteration [5] is another alternative to save memory.

3.4. Numerical experiments. Next we consider iteratively solving a system $\mathcal{M}(z) = b$. In each experiment either `rl_GMRES` or `rl_Gal` applied to $\mathcal{M}$ is compared with GMRES applied to the equivalent real formulation of the problem. Here `rl_Gal` refers to the method of section 3.2. To save storage, we also compare their restarted versions `rl_GMRES`(k), `rl_Gal`(k), and GMRES(k) restarted after every k steps. The residual at the jth step (defined similarly for the real formulation) is denoted by $r_j = \mathcal{M}(z_j) - b$,

The computations were performed with Matlab, whose syntax we use.

Example 8. This family of $\mathbb{R}$-linear systems arose in connection with the inverse problem of reconstructing an unknown electric conductivity in the unit disc from boundary measurements; see [23, 19]. To this end one needs to solve repeatedly the system $\mathcal{M}_\kappa(z) = z + M_{\#}\overline{z} = \mathbf{1}$ resulting from a discretization of a weakly singular Fredholm integral equation of the second kind depending on various parameters. More precisely, $\mathcal{M}_\kappa$ depends on the measured current on the unit circle as well as on the point in the unit disc for which the reconstruction is being computed. The right-hand side is the constant vector with ones. Due to the size $n = 2^{16}$ of the system, the matrices are not represented explicitly.

The problem was iteratively solved by using the simulated boundary data on the unit circle used in [23, Problem 4] with the initial guess $z_0 = \mathbf{1}$. We executed `rl_GMRES` and GMRES as well as their restarted versions with $k = 30, 60$. Since $M_{\#}$ is the product of a Toeplitz matrix and a diagonal matrix, matrix-vector products could be computed fast by using the FFT.

After fixing values for the parameters, the relative residuals $\|r_k\| \,/\, \|r_0\|$ were compared in the $\log_{10}$ scale for all the six iterations; see Figure 1.

The experiments were repeated by varying the parameters. Each time `rl_GMRES` outperformed GMRES (see also Proposition 3.6) such that quantitatively we had approximately 30% shorter execution times in a typical case as illustrated in Figure 1. There we also see that for the restarted iterations the difference can be even more drastic: `rl_GMRES`(60) converged, whereas GMRES(60) stagnated. For $k = 30$ both methods stagnated.

Example 9. Here we illustrate the Galerkin approximation of section 3.2 for iteratively solving the system $\mathcal{M}(z) = b$ with restarts. Using Matlab syntax, we denote by $R_{n,m} = \texttt{randn}\,(n, m) \in \mathbb{R}^{n\times m}$ a normally distributed random matrix, which has been regenerated each time it is encountered. So no two matrices $R_{n,m}$ are

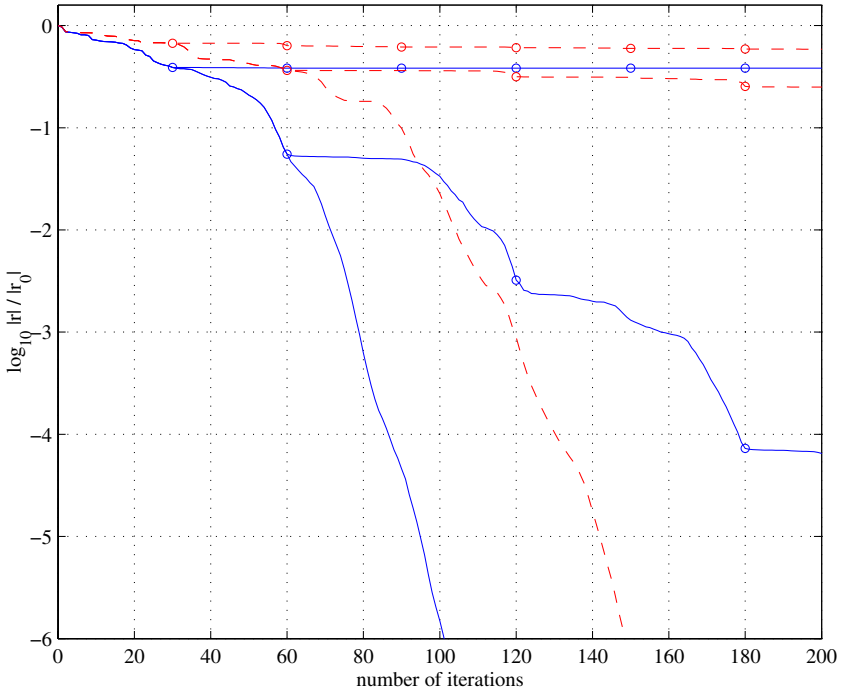


FIG. 1. *The convergence of the relative residuals in the* $\log_{10}$ *scale in Example* 8. `rl_GMRES` *and* `rl_GMRES`(k) *are depicted with solid lines and GMRES and GMRES*(k) *with dashed lines,* $k = 30, 60$. *The restart points are marked with* "o".

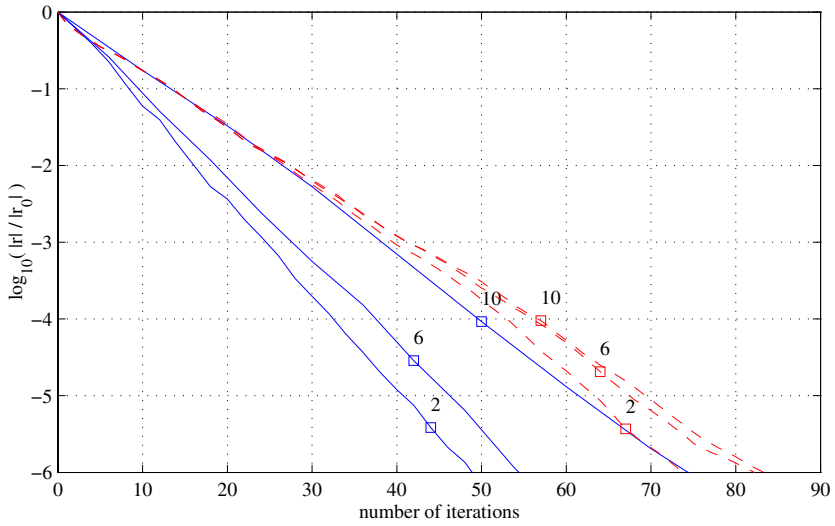


FIG. 2. *The relative residuals in the* $\log_{10}$ *scale in Example* 9. `rl_Gal`(k) *is depicted with solid lines and GMRES*(k) *with dashed lines. The labels above the squares refer to the restarting frequency* k.

the same here. In this somewhat artificial problem $M = (20+10i)I + R_{n,n} + i\,R_{n,n}$, $M_\# = (I + R_{n,n} + i\,R_{n,n})/10$, and $b = R_{n,1} + i\,R_{n,1}$ with $n = 150$. By using the initial guess $z_0 = 0$, we executed `rl_Gal`(k) and GMRES(k) with $k = 2,\ 6,\ 10$.

Short restarting frequency seems to be optimal in this problem for both methods. See Figure 2 for the behavior of the relative residuals in the $\log_{10}$ scale.

4. Computing the spectrum of an $\mathbb{R}$-linear operator in $\mathbb{C}^n$. In this final section we consider some ideas for locating the spectrum of a real linear operator $\mathcal{M}$.

In the one-parameter family of real matrices $A(0,\beta) = -\beta J - A$ in (2.5), every real eigenvalue $-\alpha$ of $A(0,\beta)$ corresponds to an eigenvalue $\alpha + i\beta$ of $\mathcal{M}$. Hence a brute force method to find the spectrum of $\mathcal{M}$ is to compute the eigenvalues of $A(0,\beta)$ for those β that are of interest. Since the spectrum is bounded by the norm, we need to consider only the interval $\{\beta \in \mathbb{R} : |\beta| \le \|\mathcal{M}\|\}$, which can be further reduced (to be possibly nonconvex) by using the Geršgorin disks (2.8). There are also many alternatives to benefit from Geršgorin disks by performing $\mathbb{C}$-linear similarity transformations in a clever way.

Remark. For a fixed $\mu \in \mathbb{C}$, the spectra of $\mathcal{M}$ and $\mu I \circ \mathcal{M}$ are related in an obvious way. However, for computations a multiplication by a scalar makes a difference. For instance, if $\sigma(\mathcal{M})$ is locally tangential to the real axis, then the prescribed approach is numerically less stable. The choice $\mu = i$ rotates the spectrum by $\pi/2$ and removes the problem in that particular neighborhood.

To get a rough picture of $\sigma(\mathcal{M})$, one option is to use a coarse grid for β over an interval of interest. The grid can then be refined in those areas where the spectrum appears to be changing rapidly while β varies. However, with a coarse tracking of the spectrum we face the risk of missing entire isolated subsets of $\sigma(\mathcal{M})$. For example, if $M_\# = 0$, then the spectrum consists of isolated points which are missed almost certainly. Also, in a nearly $\mathbb{C}$-linear case with $\|M_\#\| \ll \|M\|$ the isolated subsets of the spectrum can be very small and could thus be overlooked.

To locate tiny subsets better, we employ the information also in the nonreal eigenvalues and the corresponding eigenvectors of $A(0,\beta)$. To this end, set $\phi(w) = \frac{i\,w^*Jw}{w^*w}$ for $w \in \mathbb{C}^{2n}$ with $w \neq 0$. This satisfies $\phi(\mu u) = \phi(u) \in \mathbb{R}$ for any nonzero $u \in \mathbb{C}^{2n}$ and $\mu \in \mathbb{C}$. Also $|\phi(w)| \le \|J\| = 1$.

With ϕ define the set-valued function

$$\Phi(\beta) = \{\phi(w)\ \mathrm{Im}\,\lambda\ :\ A(0,\beta)w = \lambda w\,,\ w \neq 0\} \subset \mathbb{R}.$$

Obviously, if $A(0,\beta)$ has a real eigenvalue, then $0 \in \Phi(\beta)$.

LEMMA 4.1. *If $|\beta| > 2\|A\|$, then $\beta\,\Phi(\beta) \subset \mathbb{R}_+$. If for all β the nonreal eigenvalues of $A(0,\beta)$ are simple, then Φ is continuous.*

Proof. For the first claim, assume $|\beta| > 2\|A\|$ and that $A(0,\beta)w = \lambda w\,,\ w \neq 0$. Since J is normal, $\min|\lambda \pm i\beta| \le \|A\|$ by the Bauer–Fike theorem. Hence $|\,\mathrm{Im}\,\lambda| > \|A\|$ with $\mathrm{Im}\,\lambda$ having the same sign as β. If $\mathrm{Im}\,\lambda \ge 0$, then, since $\beta\,\phi(w) \in \mathbb{R}$,

$$\beta\,\phi(w) = \frac{i\,w^*\beta Jw}{w^*w} = \frac{i\,w^*(-\lambda + A)w}{w^*w} = \frac{w^*(\mathrm{Im}\,\lambda + \frac{1}{2}(iA + (iA)^*))w}{w^*w} > 0.$$

In the case $\mathrm{Im}\,\lambda < 0$ we get $\beta\,\phi(w)z < 0$. Hence, $\mathrm{Im}\,\lambda\,\beta\,\phi(w) > 0$ in both cases.

Eigenvalues depend continuously on β and eigenvectors corresponding to simple eigenvalues can be chosen continuous. Hence the assumptions imply that the numbers $\phi(w)\ \mathrm{Im}\,\lambda$ depend continuously on β for $\mathrm{Im}\,\lambda \neq 0$. Further, these tend to zero when λ approaches $\mathbb{R}$ since $|\phi(w)| \le 1$. ∎

Since $\beta\,\Phi(\beta) \subset \mathbb{R}_+$ far from the origin, we see that all the elements of $\Phi(\beta)$ cross the origin as β runs over $\mathbb{R}$. In fact, the vanishing elements (often) seem to take opposite signs as we step over an isolated subset of spectrum. If we order Φ along decaying $\operatorname{Re}\lambda$, we may use a simple bisection method to refine the grid on every change (of sign or length) of the vector Φ. If, at a change of sign in Φ, $\operatorname{Im}\lambda$ stays away from zero, we skip the interval. If $\operatorname{Im}\lambda$ crosses the origin, we look for a subset of the spectrum. This way we are able to locate isolated points and horizontal parts of the spectrum more accurately while, simultaneously, decreasing execution times. Of course, there is no guarantee that this tool will manage to pick every isolated subset of the spectrum.

Using the lemma we can save computational work by refining the β-grid only on intervals of interest, but this still is a rather tedious way to visualize the spectrum. The same technique, with a very coarse grid, can also be used to only locate a point on each isolated subset of the spectrum. These points, in turn, can be extended to find the corresponding piece of the curve $\{\alpha + i\beta : \det A(\alpha,\beta) = 0\}$ using standard continuation techniques (see [17]).

Once sufficiently many points of $\sigma(\mathcal{M})$ have been computed accurately, one can also use the information to find the characteristic bivariate polynomial of $\mathcal{M}$ approximately. To this end, for instance, the algorithms proposed in [13, section 4.1] can be employed.

Next we consider numerical examples. All our matrices are artificially constructed and small since we aim at illustrating only certain aspects of the spectrum. The matrices $R_{n,n}$ are defined as in Example 9.

Example 10. The spectrum of an $\mathbb{R}$-linear operator can be profuse and very arresting. We illustrate this with $\mathcal{M} : \mathbb{C}^{10} \mapsto \mathbb{C}^{10}$ having $M = R_{10,10} + i\,R_{10,10}$ and $M_\# = R_{10,10} + i\,R_{10,10}$. See Figure 3(a).

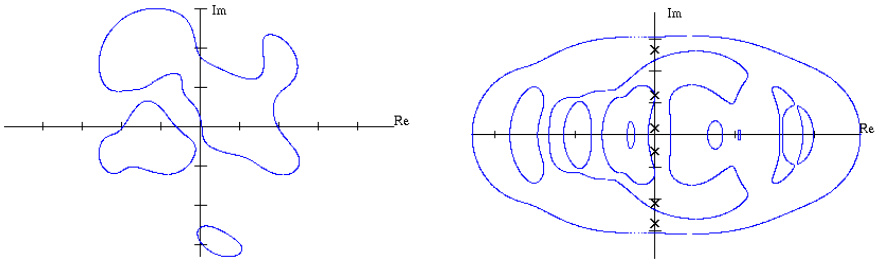
Example 11. To illustrate Proposition 2.12, we take $\mathcal{M}_1$ and $\mathcal{M}_2$ with the real forms $A_1 = R + R^T$ and $A_2 = R - R^T$ with $R = R_{20,20}$. The spectrum of $\mathcal{M}_1$ is symmetric relative to the real axis. The spectrum of $\mathcal{M}_2$ consists of at most $2n = 20$ isolated points. See Figure 3(b).

Example 12. To see how the spectrum varies, let $M = M_1 + \frac{1}{20}M_2$ such that $M_1 \in \mathbb{C}^{10\times 10}$ is a diagonal matrix having the eigenvalues $z_j = 6e^{i\theta_j}$, with $\theta_j = \frac{2\pi}{10}j$ for $j = 0,\dots,9$, and $M_2 = \frac{1}{2}\,(R_{10,10} + i\,R_{10,10})$. The antilinear part is $M_\# = \frac{1}{4}\,(R_{10,10} + i\,R_{10,10})$. In Figure 4 we have plotted $\sigma(M)$ and $\sigma(\mathcal{M})$ together with the Geršgorin disks. The Bauer–Fike bound of Proposition 2.13 is also plotted by regarding $z \mapsto M_\#\overline{z}$ as the perturbation $\mathcal{E}$ of the $\mathbb{C}$-linear operator $\widehat{\mathcal{M}}(z) = Mz$. Rounding to four digits, we had $\left\|S^{-1}\right\|\,\|S\| = 1.055$ and $\|\mathcal{E}\| = 3.431$.

Example 13. We illustrate the fact that the spectrum is not preserved, in general, in an $\mathbb{R}$-linear similarity transformation. We take $\mathcal{M} : \mathbb{C}^2 \mapsto \mathbb{C}^2$ where $M = \frac{1}{10}\left[\begin{smallmatrix} 0 & 11 \\ 3i & 10i \end{smallmatrix}\right]$ and $M_\# = \left[\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right]$. The spectrum of $\mathcal{M}$ is a curve encircling the origin. Let

$$E_0 = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right],\ E_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right],\ E_2 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right],\ \text{and } E_3 = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right].$$

Then we consider $\mathcal{M}_{s,i,t} = \mathcal{T}_{i,t,s}^{-1} \circ \mathcal{M} \circ \mathcal{T}_{i,t,s}$, where the real forms of $\mathcal{T}_{i,t,+}$ and $\mathcal{T}_{i,t,-}$ are $\exp\left(\left[\begin{smallmatrix} tE_i & 0 \\ 0 & -tE_i \end{smallmatrix}\right]\right)$ and $\exp\left(\left[\begin{smallmatrix} 0 & tE_i \\ tE_i & 0 \end{smallmatrix}\right]\right)$, respectively. Unless $t = 0$, $\mathcal{T}_{i,t,+}$ and $\mathcal{T}_{i,t,-}$ are not $\mathbb{C}$-linear. In Figure 5 we have plotted $\sigma(\mathcal{M}_{i,t,s})$. The spectrum is shown for four pairs (s,i), $s = 0, 2$ and $i = \pm$, each on a separate plot. On each plot, the horizontal copies of the complex plane correspond to values of $t = -0.7,\dots,0.7$. Note that the two real eigenvalues remain invariant here as in any similarity transformation.



(a) (b)

FIG. 3. *The spectra of the* $\mathbb{R}$*-linear operators of Examples* 10–11. (a) *The spectrum of* $\mathcal{M}$ *of Example* 10. (b) *The spectra of "symmetric"* $\mathcal{M}_1$ *(solid lines) and "antisymmetric"* $\mathcal{M}_2$ *(crosses) of Example* 11.

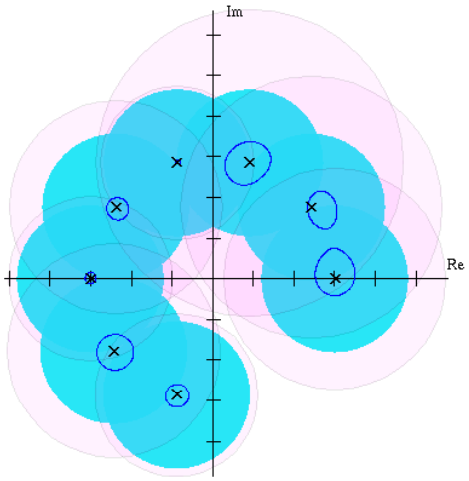


FIG. 4. *The spectrum of an* $\mathbb{R}$*-linear operator* $\mathcal{M}$*, where* M *is almost diagonal. The Bauer–Fike bound of Proposition* 2.13 *has been shaded darker by regarding* $z \mapsto M_{\#}\overline{z}$ *as the perturbation* $\mathcal{E}$*. The Geršgorin disks of* $\mathcal{M}$ *are shaded lighter. See Example* 12.

5. Conclusions. Matrix analysis for $\mathbb{R}$-linear operators in $\mathbb{C}^n$ has been studied. Although we are dealing with a weaker assumption than $\mathbb{C}$-linearity, a large part of the familiar theory could be recovered. In particular, most of the matrix factorizations aimed at solving linear systems can be regarded as special cases of our more general results.

Basics of the spectral theory for $\mathbb{R}$-linear operators in $\mathbb{C}^n$ were developed together with some preliminary computational ideas for finding the spectrum.

Since the initial motivation for our study was Krylov subspace methods, we have introduced new iterative schemes that avoid using an equivalent real formulation.

Acknowledgment. We are grateful to Dr. Samuli Siltanen for useful discussions and providing the Matlab codes used in our experiments in Example 8.

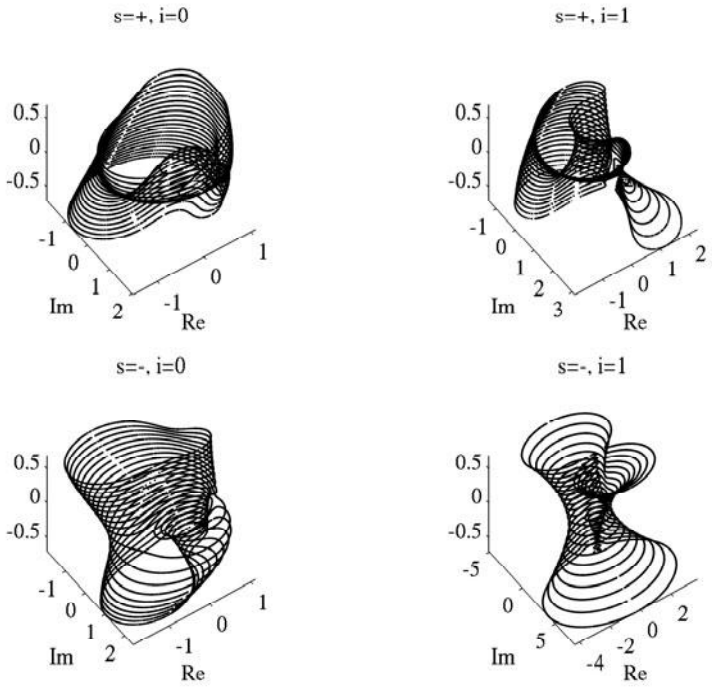


FIG. 5. *The spectrum of an $\mathbb{R}$-linear operator under $\mathbb{R}$-linear, but not $\mathbb{C}$-linear, similarity transformations. The horizontal copies of the complex plane correspond to the values of the parameter t. The darkest curves correspond to value $t = 0$. See Example* 13.

REFERENCES

[1] W.E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem,* Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] J.H. BEVIS, F.J. HALL, AND R.E. HARTWIG, *The matrix equation $A\overline{X} - XB = C$ and its special cases,* SIAM J. Matrix Anal. Appl., 9 (1988), pp. 348–359.

[3] A. BUNSE-GERSTNER AND R. STÖVER, *On a conjugate gradient-type method for solving complex symmetric linear systems,* Linear Algebra Appl., 287 (1999), pp. 105–123.

[4] R.W. FREUND, *Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices,* SIAM J. Sci. Statist. Comput., 13 (1992), pp. 425–448.

[5] R. FREUND, G.H. GOLUB, AND N. NACHTIGAL, *Iterative solution of linear systems,* Acta Numerica, 1 (1992), pp. 57–100.

[6] K. FRIEDRICHS, *On certain inequalities for analytic functions and for functions of two variables,* Trans. Amer. Math. Soc., 1 (1937), pp. 321–364.

[7] Y.P. HONG AND R.A. HORN, *A characterization of unitary congruence*, Linear and Multilinear Algebra, 25 (1989), pp. 105–119.

[8] Y.P. HONG AND R.A. HORN, *A canonical form for matrices under consimilarity*, Linear Algebra Appl., 102 (1988), pp. 143–168.

[9] R.A. HORN AND C.R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1987.

[10] M. HUHTANEN *A stratification of the set of normal matrices,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 349–367.

[11] M. HUHTANEN, *A Hermitian Lanczos method for normal matrices*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1092–1108.

[12] M. HUHTANEN, *Aspects of Nonnormality for Iterative Methods,* Report A453, Helsinki University of Technology, Espoo, Finland, 2002.

[13] M. Huhtanen and R.M. Larsen, *On generating discrete orthogonal bivariate polynomials,* BIT, 42 (2002), pp. 393–407.
[14] M. Huhtanen and R.M. Larsen, *Exclusion and inclusion regions for the eigenvalues of a normal matrix*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1070–1091.
[15] M. Huhtanen and O. Nevanlinna, *Minimal decompositions and iterative methods,* Numer. Math., 86 (2000), pp. 257–281.
[16] N. Karapetiants and S. Samko, *Equations with Involutive Operators*, Birkhäuser Boston, Boston, MA, 2001.
[17] H.B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Tata Inst. Fund. Res. Lectures on Math. and Phys. 79, Springer-Verlag, Berlin, 1987.
[18] Mathworks, *Matlab*, www.mathworks.com/products/matlab.
[19] J.L. Mueller and S. Siltanen, *Direct reconstructions of conductivities from boundary measurements,* SIAM J. Sci. Comput., 24 (2003), pp. 1232–1266.
[20] M. Putinar and H.S. Shapiro, *The Friedrichs operator of a planar domain*, in Complex Analysis, Operators, and Related Topics, Oper. Theory Adv. Appl. 113, Birkhäuser, Basel, 2000, pp. 303–330.
[21] Y. Saad and M.H. Schultz, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems,* SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[22] H.S. Shapiro, *The Schwarz Function and Its Generalization to Higher Dimensions*, Univ. Arkansas Lecture Notes Math. Sci. 9., John Wiley and Sons, New York, 1992.
[23] S. Siltanen, J. Mueller, and D. Isaacson, *An implementation of the reconstruction algorithm of A. Nachman for the 2D inverse conductivity problem,* Inverse Problems, 16 (2000), pp. 681–699.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 829–841

KRONECKER PRODUCT APPROXIMATIONS FOR IMAGE RESTORATION WITH REFLEXIVE BOUNDARY CONDITIONS*

JAMES G. NAGY†, MICHAEL K. NG‡, AND LISA PERRONE†

Abstract. Many image processing applications require computing approximate solutions of very large, ill-conditioned linear systems. Physical assumptions of the imaging system usually dictate that the matrices in these linear systems have exploitable structure. The specific structure depends on (usually simplifying) assumptions of the physical model and other considerations such as boundary conditions. When reflexive (Neumann) boundary conditions are used, the coefficient matrix is a combination of Toeplitz and Hankel matrices. Kronecker products also occur, but this structure is not obvious from measured data. In this paper we discuss a scheme for computing a (possibly approximate) Kronecker product decomposition of structured matrices in image processing, which extends previous work by Kamm and Nagy [*SIAM J. Matrix Anal. Appl.*, 22 (2000), pp. 155–172] to a wider class of image restoration problems.

Key words. image restoration, Kronecker product, singular value decomposition

AMS subject classifications. 65F20, 65F30

DOI. 10.1137/S0895479802419580

1. Introduction.

Image restoration is the process of reconstructing an image of an unknown scene from an observed image, where

(1.1) `observed image = distortion(original scene) + noise.`

The "distortion" can arise from many sources; atmospheric turbulence, out of focus lens, and motion blurs are but a few examples. Typically the distortion is described mathematically as a *point spread function* (PSF). Specifically, a PSF is a function that specifies how points in the image are distorted. PSFs are often classified as either *spatially invariant* or *spatially variant*. *Spatially invariant* means that the distortion is independent of position, while *spatially variant* means that the distortion does depend on position. Spatially invariant PSFs occur most frequently in applications [8], so this is what we consider in this paper.

A PSF can be further classified as *separable* or *nonseparable*. Separable means that the distortion in the horizontal and vertical directions is independent. That is, a two-dimensional distortion is a composition of two one-dimensional distortions. The topic of separability is often ignored when discussing image restoration problems, but, as we will see, by exploiting this structure, more choices are available in terms of image restoration algorithms.

We begin with a mathematical model of the spatially invariant image restoration problem. The image of an object can be modeled as

$$\mathbf{g} = K\mathbf{f} + \mathbf{n}, \tag{1.2}$$

*Received by the editors December 4, 2002; accepted for publication (in revised form) by M. Hanke July 3, 2003; published electronically February 24, 2004. This work was supported by the National Science Foundation under grant DMS 00-75239.

http://www.siam.org/journals/simax/25-3/41958.html

†Department of Mathematics and Computer Science, Emory University, Atlanta, GA (nagy@mathcs.emory.edu, perrone@mathcs.emory.edu).

‡Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong (mng@maths.hku.hk). The research of this author was supported in part by RGC grants 7132/00P and 7130/02P and HKU CRCG grants 10203501, 10203907, and 10204437.

where $\mathbf{g}$ is an n^2-vector representing the distorted image of size $n \times n$, $\mathbf{f}$ is a vector representing the true image, and $\mathbf{n}$ is a vector representing additive noise. K is an $n^2 \times n^2$ blurring matrix constructed from the PSF, but it has structure that can be exploited in computations. Because the blurring model is a convolution, $\mathbf{g}$ is not completely determined by $\mathbf{f}$ in the same domain where $\mathbf{g}$ is defined. Thus in solving $\mathbf{f}$ from $\mathbf{g}$, we need some assumptions on the values of $\mathbf{f}$ outside the domain of $\mathbf{g}$. These assumptions are called the boundary conditions. The structure of the blurring matrix K depends on the boundary conditions [11].

- *Periodic boundary conditions.* The image outside the domain of consideration is a repeat, in all directions, of the image inside [4]. In this case, K will be a block-circulant-circulant-block (BCCB) matrix. We can use two-dimensional fast Fourier transforms (FFTs) to diagonalize the matrix [4], but this boundary condition may be unrealistic in many situations.
- *Zero boundary conditions.* The values of $\mathbf{f}$ outside the domain of consideration are zero [1]. In this case, K will be a block-Toeplitz-Toeplitz-block (BTTB) matrix. FFTs can be used to implement fast matrix vector multiplications for K.
- *Reflexive boundary conditions.* The scene immediately outside the boundary is a reflection of the original scene inside. In this case, the matrix K is block-Toeplitz-plus-Hankel with Toeplitz-plus-Hankel-blocks (BTHTHB) [11]. In the following discussion, we express the matrix K as the sum of a block-Toeplitz-Toeplitz-block (BTTB) matrix, a block-Toeplitz-Hankel-block (BTHB) matrix, a block-Hankel-Toeplitz-block (BHTB) matrix, and a block-Hankel-Hankel-block (BHHB) matrix. Although the matrix K has a complicated structure, it can be diagonalized by the two-dimensional fast cosine transform (FCT) when the PSF is symmetric [11].

In [5, 10, 11], it has been shown that using reflexive boundary conditions in image restoration or reconstruction can be better than using periodic or zero boundary conditions.

Aside from the issue of boundary conditions, it is well known that blurring matrices are in general very ill-conditioned and image restoration algorithms will be extremely sensitive to noise [4]. The ill-conditioning of the blurring matrices stems from the wide range of magnitudes of their eigenvalues [3]. Therefore excess amplification of the noise at small eigenvalues can occur. In [11], classical Tikhonov regularization is employed to attain the stability of image restoration algorithms. A fast image restoration algorithm with the reflexive boundary conditions is developed and proposed. Since the size of the matrix K is very large, iterative methods with cosine transform based preconditioners are used to speed up the convergence of the algorithm.

We note that if the blur is separable, then the matrix K can be further decomposed into a Kronecker product of smaller matrices. In this case we are not restricted (by size constraints) to using only iterative methods. In particular, we can use singular value decomposition (SVD) based methods [6] to perform the regularization in the image restoration process. The problem is determining when a PSF is separable. We may not have an explicit mathematical formula for the PSF, and thus must recognize separability from the image data. This has been done in the case of zero boundary conditions [9]. One aim of this paper is to consider how to do this for reflexive boundary conditions.

The outline of the paper is as follows. In section 2, definitions and notation are set up. In section 3, the Kronecker product approximation of the blurring matrix K

is studied, and we provide an algorithm for constructing this approximation from the given PSF. In section 4, simulation results are presented to demonstrate the effectiveness of using this Kronecker approximation. Finally, some concluding remarks are given in section 5.

2. Definitions and notation. In order to prove the main result of this paper, we need the following definitions and notation.

2.1. Toeplitz and Hankel matrices. Banded Toeplitz and Hankel matrices arise frequently in image restoration applications. Here we demonstrate how to use a column vector to represent these matrices:

- $\mathbf{toep}(\mathbf{a}, k)$ is an $n \times n$ banded Toeplitz matrix whose kth column is $\mathbf{a} \in \Re^n$. For example,

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{toep}(\mathbf{a}, 3) = \begin{bmatrix} a_3 & a_2 & a_1 & 0 \\ a_4 & a_3 & a_2 & a_1 \\ 0 & a_4 & a_3 & a_2 \\ 0 & 0 & a_4 & a_3 \end{bmatrix}.$$

- $\mathbf{hank}(\mathbf{a}, k)$ is an $n \times n$ Hankel matrix with its first row and its last column defined by $[a_{k+1}, \ldots, a_n, 0, \ldots, 0]$ and $[0, \ldots, 0, a_1, \ldots, a_{k-1}]^T$, respectively, where $\mathbf{a} \in \Re^n$. For example,

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{hank}(\mathbf{a}, 3) = \begin{bmatrix} a_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & a_1 \\ 0 & 0 & a_1 & a_2 \end{bmatrix}.$$

- We use the notation $\mathbf{Toep}(A, k)$ and $\mathbf{Hank}(A, k)$ for similar definitions with block matrices. For example,

$$A = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{bmatrix} \quad \text{implies} \quad \mathbf{Toep}(A, 3) = \begin{bmatrix} A_3 & A_2 & A_1 & 0 \\ A_4 & A_3 & A_2 & A_1 \\ 0 & A_4 & A_3 & A_2 \\ 0 & 0 & A_4 & A_3 \end{bmatrix}$$

and

$$\mathbf{Hank}(A, 3) = \begin{bmatrix} A_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_1 \\ 0 & 0 & A_1 & A_2 \end{bmatrix}.$$

- With the above notation, we can describe the blurring matrices that arise in image restoration. Let P be an $n \times n$ array containing the image of a point spread function. Suppose the center of the PSF (location of the point source) is at p_{ij}. Let $\mathbf{p}_k^T$ be the kth row of P, and define

$$T_k = \mathbf{toep}(\mathbf{p}_k, j) \quad \text{and} \quad H_k = \mathbf{hank}(\mathbf{p}_k, j),$$

$$T = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_n \end{bmatrix}.$$

Then, we can formulate the blurring matrix under
- zero boundary conditions as

$$K = \mathbf{Toep}(T, i),$$

- or under reflexive boundary conditions as

$$K = K_{tt} + K_{th} + K_{ht} + K_{hh}, \tag{2.1}$$

where $K_{tt} = \mathbf{Toep}(T, i)$, $K_{th} = \mathbf{Toep}(H, i)$, $K_{ht} = \mathbf{Hank}(T, i)$, and $K_{hh} = \mathbf{Hank}(H, i)$.

2.2. The shift matrix. We also need to use the shift matrix:

$$Z = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & \ddots & \ddots & \\ \vdots & & \ddots & 1 \\ 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

The name *shift matrix* comes from the fact that if we multiply a vector by Z, the entries are shifted up, and if we multiply by Z^T, the entries are shifted down. The following properties of the shift matrix will be used in section 3.

1. $\mathbf{toep}(\mathbf{a}, k) = \begin{bmatrix} Z^{k-1}\mathbf{a} & \cdots & Z^0\mathbf{a} & \cdots (Z^{n-k})^T\mathbf{a} \end{bmatrix}$.
2. If $\mathbf{e}_l$ is the lth column of the identity matrix, then

$$Z^k\mathbf{e}_l = \begin{cases} 0, & l = 1, 2, \ldots, k, \\ \mathbf{e}_{l-k}, & l = k+1, \ldots, n, \end{cases}$$

and

$$(Z^k)^T\mathbf{e}_l = \begin{cases} \mathbf{e}_{l+k}, & l = 1, 2, \ldots, n-k, \\ 0, & l = n-k+1, \ldots, n. \end{cases}$$

3. From Property 2, it is easy to show that

$$(Z^k)^T(Z^k) = \operatorname{diag}(\begin{bmatrix} 0 & \cdots & 0 & \underset{\substack{\uparrow \\ k+1 \text{ entry}}}{1} & \cdots & 1 \end{bmatrix})$$

and

$$(Z^k)(Z^k)^T = \operatorname{diag}(\begin{bmatrix} 1 & \cdots & \underset{\substack{\uparrow \\ n-k \text{ entry}}}{1} & 0 & \cdots & 0 \end{bmatrix}).$$

4. From Property 3, it follows that

$$(Z^k)^T(Z^k) + (Z^{n-k})(Z^{n-k})^T = (Z^k)(Z^k)^T + (Z^{n-k})^T(Z^{n-k}) = I,$$

and thus,

$$\sum_{k=1}^{n-1} \left((Z^k)^T(Z^k) + (Z^k)(Z^k)^T \right) = (n-1)I. \tag{2.2}$$

5. For $a, b < n$,

$$(Z^a)^T Z^b + Z^{n-a}(Z^{n-b})^T = \begin{cases} Z^{b-a} & \text{if b} > \text{a}, \\ (Z^{a-b})^T & \text{if a} > \text{b}. \end{cases} \tag{2.3}$$

2.3. Kronecker product matrices. Here we state some properties and definitions related to Kronecker products. A Kronecker product is defined to be

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nn}B \end{bmatrix}.$$

Two transformations that we need follow.

- The **vec** operator transforms two-dimensional arrays into one-dimensional vectors by stacking columns. For example,

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{vec}(X) = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ x_{12} \\ x_{22} \\ x_{32} \\ x_{13} \\ x_{23} \\ x_{33} \end{bmatrix}.$$

- The **tilde** transformation reorders the entries of a block matrix as follows. If K is a block matrix,

$$K = \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & K_{22} & \cdots & K_{2n} \\ \vdots & \vdots & & \vdots \\ K_{n1} & K_{n2} & \cdots & K_{nn} \end{bmatrix},$$

then

$$\tilde{K} = \mathbf{tilde}(K) = \begin{bmatrix} \mathbf{vec}(K_{11})^T \\ \vdots \\ \mathbf{vec}(K_{n1})^T \\ \vdots \\ \mathbf{vec}(K_{1n})^T \\ \vdots \\ \mathbf{vec}(K_{nn})^T \end{bmatrix}.$$

Note that in image restoration, with reflexive boundary conditions, from (2.1), we have

$$\tilde{K} = \tilde{K}_{tt} + \tilde{K}_{th} + \tilde{K}_{ht} + \tilde{K}_{hh}. \tag{2.4}$$

Van Loan and Pitsianis [12] show, for a general block matrix, that

$$\left\| K - \sum_{k=1}^{s} (A_k \otimes B_k) \right\|_F = \left\| \tilde{K} - \sum_{k=1}^{s} \tilde{\mathbf{a}}_k \tilde{\mathbf{b}}_k^T \right\|_F,$$

where $\tilde{\mathbf{a}}_k = \mathbf{vec}(A_k)$ and $\tilde{\mathbf{b}}_k = \mathbf{vec}(B_k)$. The best Kronecker product approximation is obtained by finding the SVD of $\tilde{K}$. In particular, if

$$\tilde{K} = \sum_{k=1}^{r} \tilde{\sigma}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^T,$$

then the above Frobenius norm is minimized by taking

$$\tilde{\mathbf{a}}_k = \sqrt{\tilde{\sigma}_k}\tilde{\mathbf{u}}_k \quad \text{and} \quad \tilde{\mathbf{b}}_k = \sqrt{\tilde{\sigma}_k}\tilde{\mathbf{v}}_k.$$

The problem with this approach is that we need to compute the principal singular values and vectors of an $n^2 \times n^2$ matrix, $\tilde{K}$. The purpose of this paper is to show, for image restoration problems, how this computational effort can be reduced substantially by computing principal singular values and vectors of arrays of size at most $n \times n$.

3. Kronecker product approximation. Let K be the $n^2 \times n^2$ blurring matrix for a spatially invariant image restoration problem using reflexive boundary conditions (see section 1), and suppose P is the $n \times n$ PSF image array. In this section we show how a Kronecker product approximation of this $n^2 \times n^2$ matrix can be accomplished by computing the principal singular values and vectors of an $n \times n$ array related to P. This has been done for zero boundary conditions [9]. The case for reflexive boundary conditions is a bit more difficult to derive. To simplify notation, we consider only one term in the sum of Kronecker products; we describe later how to extend this to an arbitrary number of terms.

Our aim is, given the PSF, P, with center p_{ij}, to find vectors $\mathbf{a}$ and $\mathbf{b}$ of length n such that the matrices

$$A_t = \mathbf{toep}(\mathbf{a}, i)\,, \quad A_h = \mathbf{hank}(\mathbf{a}, i)\,,$$
$$B_t = \mathbf{toep}(\mathbf{b}, j)\,, \quad B_h = \mathbf{hank}(\mathbf{b}, j)$$

minimize $||K - (A_t + A_h) \otimes (B_t + B_h)||_F$ over all such Kronecker products. We first state the main result and the corresponding algorithm that comes from it. The proof will come later.

THEOREM 3.1. *Let P be an $n \times n$ PSF, with center p_{ij}. Let R be the Cholesky factor of the $n \times n$ symmetric Toeplitz matrix with its first row $[n, 1, 0, 1, 0, 1, \ldots]$. Then*

$$||K - (A_t + A_h) \otimes (B_t + B_h)||_F = ||RPR^T - (R\mathbf{a})(R\mathbf{b})^T||_F.$$

We prove this theorem later. First we note that the Frobenius norm in the left-hand side involves matrices with dimension $n^2 \times n^2$, and the Frobenius norm in the right-hand side involves matrices with dimension $n \times n$. Based on this theorem, the algorithm for constructing the Kronecker product approximation of K is as follows.

ALGORITHM. `Construct the approximation` $K \approx A \otimes B$.

- `Compute` R
- `Construct` $P_r = RPR^T$
- `Compute the SVD:` $P_r = \sum \sigma_k \mathbf{u}_k \mathbf{v}_k^T$
- `Construct the vectors:` $\mathbf{a} = \sqrt{\sigma_1} R^{-1}\mathbf{u}_1$ `and` $\mathbf{b} = \sqrt{\sigma_1} R^{-1}\mathbf{v}_1$
- `Construct the matrices:` $A_t = \mathbf{toep}(\mathbf{a}, i)$, $A_h = \mathbf{hank}(\mathbf{a}, i)$, $B_t = \mathbf{toep}(\mathbf{b}, j)$, `and` $B_h = \mathbf{hank}(\mathbf{b}, j)$

In our experience, for real PSFs, the singular values of P_r decay very quickly to zero (see the numerical example in section 4). In fact, it is often the case that $\sigma_1 >> \sigma_2 \approx \cdots \approx \sigma_n \approx 0$. Thus, $K \approx A \otimes B$ is generally a very good approximation. However, if a rank s approximation is desired, where $1 < s \leq \text{rank}(P_r)$, the last two steps of the algorithm can be easily modified, as follows, to produce the approximation $K \approx \sum_{k=1}^{s} A_k \otimes B_k$.

- `For` $k = 1, 2, \ldots, s$, `construct the vectors:`
  $\mathbf{a}_k = \sqrt{\sigma_k} R^{-1} \mathbf{u}_k$ `and` $\mathbf{b}_k = \sqrt{\sigma_k} R^{-1} \mathbf{v}_k$
- `For` $k = 1, 2, \ldots, s$, `construct the matrices:`
  $A_{tk} = \mathbf{toep}(\mathbf{a}_k, i)$, $A_{hk} = \mathbf{hank}(\mathbf{a}_k, i)$, $B_{tk} = \mathbf{toep}(\mathbf{b}_k, j)$, `and` $B_{hk} = \mathbf{hank}(\mathbf{b}_k, j)$

In this case, the statement of Theorem 3.1 becomes

$$\left\| K - \sum_{k=1}^{s} (A_{tk} + A_{hk}) \otimes (B_{tk} + B_{hk}) \right\|_F = \left\| RPR^T - \sum_{k=1}^{s} (R\mathbf{a}_k)(R\mathbf{b}_k)^T \right\|_F .$$

We now proceed to prove Theorem 3.1. From Van Loan and Pitsianis [12], we have

$$\begin{aligned} ||K - (A_t + A_h) \otimes (B_t + B_h)||_F &= ||\tilde{K} - \mathbf{vec}(A_t + A_h)\mathbf{vec}(B_t + B_h)^T||_F \\ &= ||\tilde{K} - (\tilde{\mathbf{a}}_t + \tilde{\mathbf{a}}_h)(\tilde{\mathbf{b}}_t + \tilde{\mathbf{b}}_h)^T||_F, \end{aligned}$$

where $\tilde{\mathbf{a}}_t = \mathbf{vec}(A_t)$, $\tilde{\mathbf{a}}_h = \mathbf{vec}(A_h)$, $\tilde{\mathbf{b}}_t = \mathbf{vec}(B_t)$, and $\tilde{\mathbf{b}}_h = \mathbf{vec}(B_h)$.

LEMMA 3.2. *Let P be an $n \times n$ PSF with center p_{ij}, let Z be the $n \times n$ shift matrix, and define $\tilde{K}_{tt}$, $\tilde{K}_{th}$, $\tilde{K}_{ht}$, and $\tilde{K}_{hh}$ as in* (2.4). *Then*

(i) $\tilde{K}_{tt} = D_{t,i} \tilde{P} D_{t,j}^T$,
(ii) $\tilde{K}_{th} = D_{t,i} \tilde{P} D_{h,j}^T$,
(iii) $\tilde{K}_{ht} = D_{h,i} \tilde{P} D_{t,j}^T$,
(iv) $\tilde{K}_{hh} = D_{h,i} \tilde{P} D_{h,j}^T$,

where

$$\tilde{P} = \begin{bmatrix} P & P & \cdots & P \\ P & P & \cdots & P \\ \vdots & \vdots & & \vdots \\ P & P & \cdots & P \end{bmatrix} \in \Re^{n^2 \times n^2}, \tag{3.1}$$

$D_{t,k}$ and $D_{h,k}$ are block-diagonal matrices given by

$$D_{t,k} = \text{ diag } [Z^{k-1}, Z^{k-2}, \ldots, Z^1, Z^0, Z^T, \ldots, (Z^{n-k})^T] \in \Re^{n^2 \times n^2} \tag{3.2}$$

and

$$D_{h,k} = \text{ diag } [Z^k, Z^{k+1}, \ldots, Z^{n-1}, Z^n, (Z^{n-1})^T, \ldots, (Z^{n-k+1})^T] \in \Re^{n^2 \times n^2}. \tag{3.3}$$

Proof. We only prove (i); similar techniques can be used to establish the other relations. First observe that

$$K_{tt} = \begin{bmatrix} \mathbf{toep}(\mathbf{p}_i, j) & \cdots & \mathbf{toep}(\mathbf{p}_1, j) & & \\ \vdots & \ddots & \vdots & \ddots & \\ \mathbf{toep}(\mathbf{p}_n, j) & \cdots & \mathbf{toep}(\mathbf{p}_i, j) & \cdots & \mathbf{toep}(\mathbf{p}_1, j) \\ & \ddots & \vdots & \ddots & \vdots \\ & & \mathbf{toep}(\mathbf{p}_n, j) & \cdots & \mathbf{toep}(\mathbf{p}_i, j) \end{bmatrix},$$

where $\mathbf{p}_k^T$ is the kth row of P. Denote the kth block column of K_{tt} as $[K_{tt}]_k$; that is,

$$[K_{tt}]_k = \left[\begin{array}{c} \mathbf{toep}(\mathbf{p}_{i-k+1}, j) \\ \vdots \\ \mathbf{toep}(\mathbf{p}_n, j) \\ 0 \\ \vdots \\ 0 \end{array}\right] \left.\begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}\right\} i-k \qquad \text{if} \quad 1 \le k \le i$$

and

$$[K_{tt}]_k = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \mathbf{toep}(\mathbf{p}_1, j) \\ \vdots \\ \mathbf{toep}(\mathbf{p}_{n-k+i}, j) \end{array}\right] \left.\begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}\right\} k-i \qquad \text{if} \quad i+1 \le k \le n.$$

Then, for $1 \le k \le i$, we have

$$\begin{aligned} [\widetilde{K_{tt}}]_k &= \left[\begin{array}{c} \mathbf{vec}(\mathbf{toep}(\mathbf{p}_{i-k+1}, j))^T \\ \vdots \\ \mathbf{vec}(\mathbf{toep}(\mathbf{p}_n, j))^T \\ 0^T \\ \vdots \\ 0^T \end{array}\right] \\ &= \left[\begin{array}{ccc} \mathbf{p}_{i-k+1}^T & \cdots & \mathbf{p}_{i-k+1}^T \\ \vdots & & \vdots \\ \mathbf{p}_n^T & \cdots & \mathbf{p}_n^T \\ 0^T & \cdots & 0^T \\ \vdots & & \vdots \\ 0^T & \cdots & 0^T \end{array}\right] \left[\begin{array}{ccccccc} (Z^{j-1})^T & & & & & & \\ & \ddots & & & & & \\ & & Z^0 & & & & \\ & & & Z & & & \\ & & & & \ddots & & \\ & & & & & Z^{n-j} & \end{array}\right] \\ &= Z^{i-k} \left[\begin{array}{ccc} P & \cdots & P \end{array}\right] D_{t,j}^T. \end{aligned}$$

Similarly, for $i+1 \le k \le n$, we have

$$[\widetilde{K_{tt}}]_k = (Z^{k-i})^T \left[\begin{array}{ccc} P & \cdots & P \end{array}\right] D_{t,j}^T \,,$$

and, therefore,

$$\tilde{K}_{tt} = \left[\begin{array}{c} [\widetilde{K_{tt}}]_1 \\ \vdots \\ [\widetilde{K_{tt}}]_i \\ \vdots \\ [\widetilde{K_{tt}}]_n \end{array}\right] = \left[\begin{array}{c} Z^{i-1} \left[\begin{array}{ccc} P & \cdots & P \end{array}\right] D_{t,j}^T \\ \vdots \\ Z^0 \left[\begin{array}{ccc} P & \cdots & P \end{array}\right] D_{t,j}^T \\ \vdots \\ (Z^{n-i})^T \left[\begin{array}{ccc} P & \cdots & P \end{array}\right] D_{t,j}^T \end{array}\right] = D_{t,i} \tilde{P} D_{t,j}^T \,. \qquad \square$$

According to Lemma 3.2, we have

$$\tilde{K} = (D_{t,i} + D_{h,i})\tilde{P}(D_{t,j} + D_{h,j})^T .$$

The next lemma states the forms of the vectors $\tilde{\mathbf{a}}_t$, $\tilde{\mathbf{a}}_h$, $\tilde{\mathbf{b}}_t$, and $\tilde{\mathbf{b}}_h$.

LEMMA 3.3. *Let $\tilde{\mathbf{a}}_t$, $\tilde{\mathbf{a}}_h$, $\tilde{\mathbf{b}}_t$, and $\tilde{\mathbf{b}}_h$ be defined as above. Then*

$$\tilde{\mathbf{a}}_t + \tilde{\mathbf{a}}_h = (D_{t,i} + D_{h,i}) \begin{bmatrix} \mathbf{a} \\ \mathbf{a} \\ \vdots \\ \mathbf{a} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{b}}_t + \tilde{\mathbf{b}}_h = (D_{t,j} + D_{h,j}) \begin{bmatrix} \mathbf{b} \\ \mathbf{b} \\ \vdots \\ \mathbf{b} \end{bmatrix} .$$

The proof of Lemma 3.3 is similar to that of Lemma 3.2. We just note that $\tilde{\mathbf{a}}_t = \mathbf{vec}(A_t)$, $\tilde{\mathbf{a}}_h = \mathbf{vec}(A_h)$, $\tilde{\mathbf{b}}_t = \mathbf{vec}(B_t)$, and $\tilde{\mathbf{b}}_h = \mathbf{vec}(B_h)$ where $A_t = \mathbf{toep}(\mathbf{a}, i)$, $A_h = \mathbf{hank}(\mathbf{a}, i)$, $B_t = \mathbf{toep}(\mathbf{b}, j)$, and $B_h = \mathbf{hank}(\mathbf{b}, j)$.

We now have all of the tools needed to prove our main theorem.

Proof of Theorem 3.1. Using Lemmas 3.2 and 3.3, we obtain

$$\begin{aligned}
&\tilde{K} - (\tilde{\mathbf{a}}_t + \tilde{\mathbf{a}}_h)(\tilde{\mathbf{b}}_t + \tilde{\mathbf{b}}_h)^T \\
&= (D_{t,i} + D_{h,i}) \left(\tilde{P} - \begin{bmatrix} \mathbf{a} \\ \mathbf{a} \\ \vdots \\ \mathbf{a} \end{bmatrix} \begin{bmatrix} \mathbf{b}^T & \mathbf{b}^T & \cdots & \mathbf{b}^T \end{bmatrix} \right) (D_{t,j} + D_{h,j})^T \\
&= (D_{t,i} + D_{h,i}) \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} (P - \mathbf{a}\mathbf{b}^T) \begin{bmatrix} I & I & \cdots & I \end{bmatrix} (D_{t,j} + D_{h,j})^T .
\end{aligned}$$

Let

$$\hat{Q} = \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix}$$

and note that if we find the QR factorizations

$$(D_{t,i} + D_{h,i})\hat{Q} = Q_i R_i , \quad (D_{t,j} + D_{h,j})\hat{Q} = Q_j R_j , \tag{3.4}$$

then

$$||\tilde{K} - (\tilde{\mathbf{a}}_t + \tilde{\mathbf{a}}_h)(\tilde{\mathbf{b}}_t + \tilde{\mathbf{b}}_h)^T||_F = ||R_i(P - \mathbf{a}\mathbf{b}^T)R_j^T||_F .$$

The next task is to determine the matrices R_i and R_j. By (3.4), we have

$$R_i^T R_i = \begin{bmatrix} I & I & \cdots & I \end{bmatrix} (D_{t,i} + D_{h,i})^T (D_{t,i} + D_{h,i}) \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} ,$$

and similarly for $R_j^T R_j$. Let us consider $R_i^T R_i$ for the case that $i \leq \frac{n}{2}$. To simplify the presentation, we define a matrix W as

$$W = (D_{t,i} + D_{h,i}) \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix}.$$

Using (3.2) and (3.3) with $k = i$, we have

$$W = \begin{bmatrix} Z^{i-1} & + & Z^i \\ \vdots & & \vdots \\ Z^1 & + & Z^{2i-2} \\ Z^0 & + & Z^{2i-1} \\ (Z^1)^T & + & Z^{2i} \\ \vdots & & \vdots \\ (Z^{n-2i})^T & + & Z^{n-1} \\ (Z^{n-2i+1})^T & + & Z^n \\ (Z^{n-2i+2})^T & + & (Z^{n-1})^T \\ \vdots & & \vdots \\ (Z^{n-i})^T & + & (Z^{n-i+1})^T \end{bmatrix}.$$

After multiplication and rearrangement of terms,

$$\begin{aligned} R_i^T R_i &= W^T W \\ &= Z^0 Z^0 + \sum_{k=1}^{n-1} \big((Z^k)^T Z^k + Z^k (Z^k)^T \big) + \sum_{k=1}^{i-1} \big(Z^{2k-1} + (Z^{2k-1})^T \big) \\ &\quad + \sum_{k=i}^{n-i} \big(Z^{2k-1} + (Z^{2k-1})^T \big), \end{aligned}$$

where the second summation utilizes (2.3) and the remaining terms arise directly from multiplication. Using (2.2) and simplifying, we get

$$R_i^T R_i = nI + \sum_{k=1}^{n-i} \big(Z^{2k-1} + (Z^{2k-1})^T \big).$$

Since in this case $i \leq \frac{n}{2}$, we have $2n - 2i - 1 \geq n - 1$, so the largest exponent on Z in the sum will be *at least* large enough to "fill" every other off-diagonal of $R_i^T R_i$ with 1's. Therefore, $R_i^T R_i$ is the $n \times n$ symmetric Toeplitz matrix with first row given by

$$\begin{bmatrix} n & 1 & 0 & 1 & 0 & 1 & \cdots \end{bmatrix}.$$

Using j in place of i in the argument above, $R_j^T R_j$ yields the same matrix when $j \leq \frac{n}{2}$. In the cases of $i > \frac{n}{2}$ and $j > \frac{n}{2}$, similar proofs generate identical results. Thus, for all possible values of i and j, $R_i^T R_i = R_j^T R_j =$ the $n \times n$ symmetric Toeplitz matrix described above. By setting $R = R_i = R_j$, we complete the proof of our main theorem.

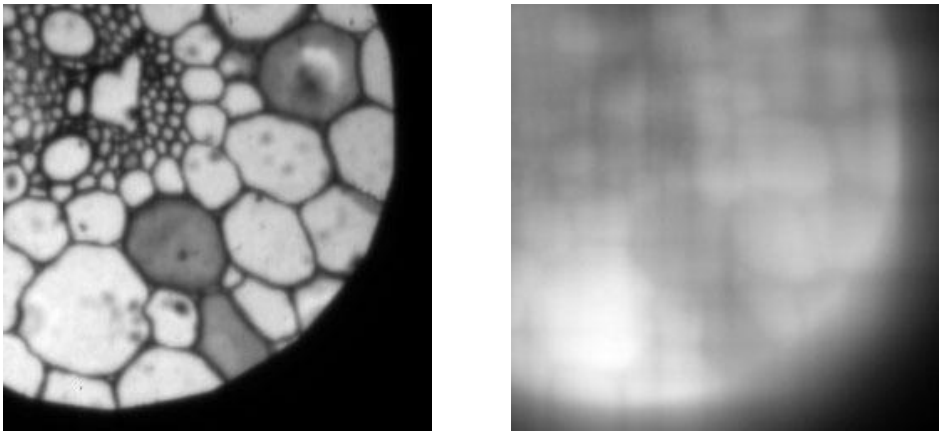


FIG. 4.1. *The true image, and the blurred, noisy image to be restored.*

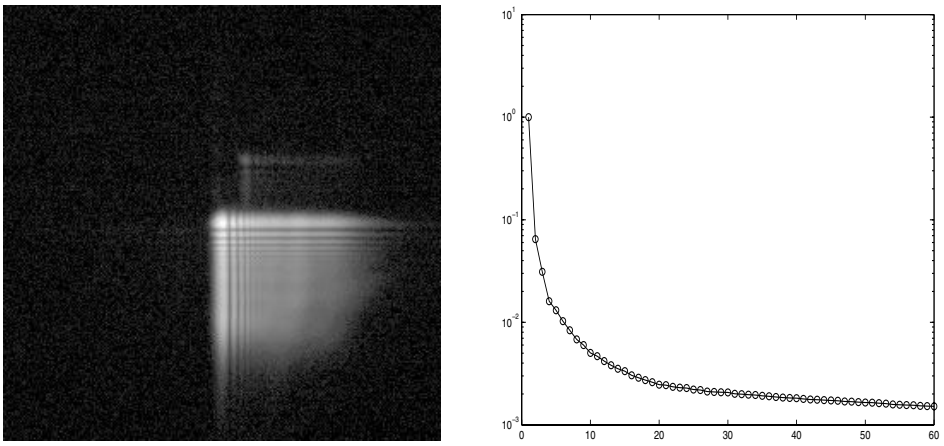


FIG. 4.2. *Image of the nonsymmetric PSF, and a plot of its first* 60 *singular values.*

4. Numerical examples. Now that we have a Kronecker sum approximation for the blurring matrix under reflexive boundary conditions, we illustrate how it can be used for an image restoration example. Recall that the image formation model is given in (1.2). The test data we use is shown in Figures 4.1 and 4.2. The 256×256 blurred and noisy image, $\mathbf{g}$, shown on the right side of Figure 4.1, and its corresponding true image, $\mathbf{f}$, on the left, have been excised from larger 512×512 images. Blurring was performed on the larger image so that the natural boundary elements would contribute to the blur, and 0.1% Gaussian white noise was added to the pixel values. All numerical tests reported here were performed on the smaller image using fabricated (reflexive and zero [9]) boundary conditions. All computations were done in Matlab 6.1.

The PSF, shown in Figure 4.2, is an example of blurring that occurs in wavefront coding, where a cubic phase filter is used to improve depth of field resolution in light efficient wide aperture optical systems [2]. A plot of the first 60 singular values of the PSF is also shown in Figure 4.2. Note that the largest singular value dominates the spectrum by an order of magnitude. In fact, for all singular values smaller

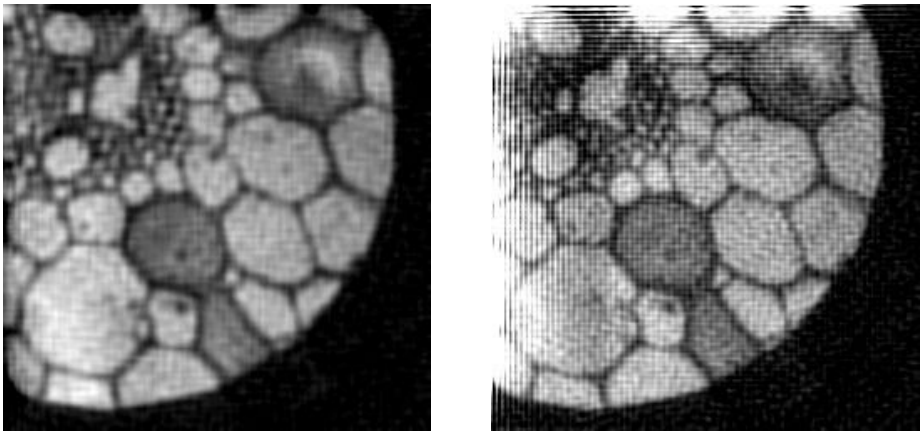


FIG. 4.3. *TSVD restoration with reflexive boundary conditions and TSVD restoration with zero boundary conditions.*

than σ_5, it dominates the spectrum by two orders of magnitude. In our experience, this is typical in image restoration problems whether the PSF is symmetric or nonsymmetric. For this reason, a Kronecker sum approximation with $s \leq 5$ can generally provide excellent restorations. We remark that since the PSF is nonsymmetric (and cannot be approximated well by a symmetric PSF), using a cosine transform based preconditioner with Tikhonov regularization [10] is not effective.

As in [9], the Kronecker product decomposition is used to construct an approximate SVD of K, which can then be used in image restoration algorithms. That is, suppose K is approximated by $T = \sum_{k=1}^{s} A_k \otimes B_k$, where A_k and B_k are $n \times n$ Toeplitz plus Hankel matrices computed according to the algorithm in section 3. An approximate SVD for K can be computed as

$$\begin{aligned} K &\approx U\Sigma V^T, \\ U &= U_A \otimes U_B, \\ V &= V_A \otimes V_B, \\ \Sigma &= \operatorname{diag}(U^T T V) \\ &= \operatorname{diag}(U^T(A_1 \otimes B_1 + A_2 \otimes B_2 + \cdots + A_s \otimes B_s)V), \end{aligned}$$

where $A_1 = U_A \Sigma_A V_A^T$ and $B_1 = U_B \Sigma_B V_B^T$. Since s is usually small ($s \leq 5$), the cost of the above scheme is only $O(n^3)$ (as opposed to $O(n^6)$, which is the cost of computing an SVD of K directly). It is therefore computationally viable to consider, for example, using this approximation with the truncated singular value decomposition (TSVD). The TSVD solution is given by

$$\mathbf{f}_{TSVD} = V\Sigma^{+}U^T\mathbf{g}\,, \quad \Sigma^{\dagger} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_t}, 0, \ldots, 0\right),$$

where t is called a truncation index, or regularization parameter. The truncation index is problem dependent; several approaches may be used to choose an appropriate value [3, 7]. For our experiments, we use generalized cross validation (GCV):

$$t = \arg\min_k G(k) = \arg\min_k \frac{||K\mathbf{f}_k - \mathbf{g}||_2^2}{(N-k)^2}\,,$$

where N is the number of pixels in the image, and

$$\mathbf{f}_k = V \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k}, 0, \ldots, 0\right) U^T \mathbf{g}\,.$$

We computed TSVD restorations using the SVD approximation based on sums of s Kronecker products, for several values of s, but the results were visually indistinguishable, so only those for $s = 1$ are reported here. Figure 4.3 shows TSVD restorations using reflexive (relative error 0.3358) and zero (relative error 0.6862) boundary conditions. In each case we used GCV to choose the truncation index, t; in particular, we obtained $t = 4852$ for reflexive and $t = 7813$ for zero boundary conditions. As expected, the reflexive boundary condition has addressed the problem of ringing effects at the image boundary.

5. Concluding remarks. In this paper, we have studied SVD-based regularization methods for solving image restoration problems with reflexive boundary conditions. We have shown that a Kronecker product decomposition of block-Toeplitz-plus-Hankel with Toeplitz-plus-Hankel-block matrices from image restoration problems can be determined by computing the singular value decomposition of weighted point spread functions. Numerical results suggest that the reflexive boundary condition provides an effective model for image restoration problems in terms of the minimization of the ringing effects near the boundary. We also find that the SVD-based regularization method using the Kronecker product decomposition is efficient in terms of the computational cost of solving image restoration problems.

REFERENCES

[1] H. Andrews and B. Hunt, *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs, NJ, 1977.
[2] E. R. Dowski and W. T. Cathey, *Extended depth of field through wavefront coding*, Appl. Optics, 34 (1995), pp. 1859–1866.
[3] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
[4] R. Gonzalez and R. Woods, *Digital Image Processing*, Addison-Wesley, Reading, MA, 1992.
[5] D. Grishin and T. Strohmer, *Fast multi-dimensional scattered data approximation with Neumann boundary conditions*, Linear Algebra Appl., to appear.
[6] P. C. Hansen, *Regularization tools: A Matlab package for the analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
[7] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, PA, 1997.
[8] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[9] J. Kamm and J. G. Nagy, *Optimal Kronecker product approximation of block Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 155–172.
[10] M. Ng, R. Chan, T. Chan, and A. Yip, *Cosine transform preconditioners for high resolution image reconstruction*, Linear Algebra Appl., 316 (2000), pp. 89–104.
[11] M. K. Ng, R. H. Chan, and W. Tang, *A fast algorithm for deblurring models with Neumann boundary conditions*, SIAM J. Sci. Comput., 21 (1999), pp. 851–866.
[12] C. F. Van Loan and N. P. Pitsianis, *Approximation with Kronecker products*, in Linear Algebra for Large Scale and Real Time Applications, M. S. Moonen and G. H. Golub, eds., Kluwer Academic Publications, Dordrecht, The Netherlands, 1993, pp. 293–314.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 842–857

NEW FAST ALGORITHMS FOR TOEPLITZ-PLUS-HANKEL MATRICES*

GEORG HEINIG† AND KARLA ROST‡

Abstract. New fast algorithms for the solutions of linear systems with a Toeplitz-plus-Hankel coefficient matrix M, of both Levinson- and Schur-type, are presented that require fewer operations than previous ones. The Schur-type algorithm produces a ZW-factorization of M, and the Levinson-type algorithm produces a WZ-factorization of M^{-1}. The new algorithms are in spirit close to the split Levinson and Schur algorithms of Delsarte and Genin.

Key words. Toeplitz-plus-Hankel matrix, fast algorithm, ZW-factorization, WZ-factorization

AMS subject classifications. 65F05, 15A06, 15A23, 15A09

DOI. 10.1137/S0895479802410074

1. Introduction. In this paper we present and compare new fast algorithms for the solution of linear systems $M_n\mathbf{f} = \mathbf{b}$ with a nonsingular Toeplitz-plus-Hankel (T+H) coefficient matrix $M_n = [\, a_{j-k} + h_{j+k-1} \,]_{j,k=1}^{n}$ with entries from a field $\mathbb{F}$. It is well known that linear systems with such a coefficient matrix can be solved with $O(n^2)$ computational complexity compared with $O(n^3)$ for a general system. Fast algorithms have been designed and studied in [30], [31], [16], [14], [10], [35], [36], [34], [9], [13] and other papers. Note that none of these algorithms works for all nonsingular T+H matrices, except under some conditions such as strong nonsingularity. An algorithm that works without additional conditions was recently presented in [12].

The main focus of the present paper is not to remove additional restrictions but to reduce the complexity of the algorithms. Delsarte and Genin showed in [5] and [6] that in the classical Levinson and Schur algorithms for the solution of real symmetric Toeplitz systems the number of multiplications can be reduced by about one-half via splitting the solutions into their symmetric and skewsymmetric parts. Yagle observed in [35] that the split Levinson algorithm for symmetric Toeplitz matrices has an analogue for general T+H matrices, even though there is no decomposition into symmetric and skewsymmetric parts.

In the series of papers [27], [28], and [29] Melman showed that the number of operations in the split Levinson algorithm for symmetric Toeplitz matrices can be reduced further by considering double steps. Note that Melman's double step Levinson-type algorithm is closely related to an algorithm in [11] that is assigned for more general centrosymmetric T+H matrices. The latter paper also proposes Schur-type algorithms. We conclude from [11] and from Melman's papers that it might lead to more efficient algorithms if recursions for the central submatrices rather than recursions for the leading principal submatrices are considered. We have shown in a series of papers that this is the case for skewsymmetric Toeplitz matrices [23] and [24], symmetric Toeplitz matrices [25], and centrosymmetric or centro-skewsymmetric T+H matrices

*Received by the editors June 24, 2002; accepted for publication (in revised form) by D. Calvetti July 22, 2003; published electronically March 30, 2004. This work was supported by research grant SM05/02 of Kuwait University.

http://www.siam.org/journals/simax/25-3/41007.html

†Department of Mathematics and Computer Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait (georg@mcs.sci.kuniv.edu.kw).

‡Department of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany (krost@mathematik.tu-chemnitz.de).

[26]. Typically these algorithms are based on three-term recursions, such as standard algorithms for Hankel matrices, whereas standard algorithms for Toeplitz matrices are based on two-term recursions.

The main aim of the present paper is to show that the general T+H structure is, like the symmetric Toeplitz and centrosymmetric T+H structure, a perfect background to consider recursions for the central instead of leading principal submatrices, and to demonstrate how this leads to more efficient algorithms. We will assume that all central submatrices are nonsingular. A matrix with this property will be called *centro-nonsingular*. In particular, positive definite matrices are centro-nonsingular. If a T+H matrix has singular central submatrices, one can apply the (slower) algorithms from [12].

In the same way the classical Schur algorithm for the leading principal submatrices is related to an LU-factorization of the matrix and the classical Levinson algorithm to a UL-factorization of the inverse, the corresponding algorithms for the central submatrices are related to a ZW-factorization of the matrix and a WZ-factorization of the inverse, respectively. For symmetric Toeplitz matrices this was mentioned by Demeure in [7], and for centrosymmetric and centro-skewsymmetric T+H matrices this was shown in our recent paper [26] (for skewsymmetric Toeplitz matrices see [23]). Note that a nonsingular matrix admits a ZW-factorization if and only if it is centro-nonsingular.

WZ-factorizations, which are also called "quadrant interlocking" or "bow tie" factorizations, were originally introduced and studied by Evans and his coworkers in connection with the parallel solution of tridiagonal systems (see [33], [8] and references therein).

Let us outline of the rest of the paper. In section 2 we discuss inversion formulas for T+H matrices, which are representations of the inverse of a T+H matrix M_n with the help of $O(n)$ parameters. We offer a new version in which the inverse is represented by the first and last columns and rows of the inverse of M_n and an $(n+2)\times(n+2)$ extension of it. This version is particularly appropriate for the application of our algorithms.

In section 3 we develop a three-term recursion formula for the first and last columns of the inverses of the central submatrices of the T+H matrix M_n. This recursion leads to a Levinson-type algorithm for finding the parameters involved in the inversion formula. In section 4 we describe a recursion for the residuals of these vectors leading to a Schur-type algorithm. The Schur-type algorithm can replace the inner product calculations in the Levinson-type algorithm, which seems to be more appropriate in parallel processing.

In section 5 we give a sketch of a "superfast" algorithm with complexity $O(n\log^2 n)$ (in the case where $\mathbb{F}$ are real or complex numbers) based on a combination of the Levinson-type and Schur-type algorithms and a divide-and-conquer approach. We refrain from presenting details because at the moment this algorithm still seems impractical.

Section 6 is dedicated to factorizations. In particular, we show that the Schur-type algorithm leads to a ZW-factorization of the matrix M_n, whereas the Levinson-type algorithm leads to a WZ-factorization of its inverse.

In section 7 different methods for solving a linear system with a T+H coefficient matrix are discussed and compared from the viewpoint of complexity in sequential processing. This concerns Levinson-type, Schur-type, and hybrid algorithms based on the inversion formula, factorizations, and direct recursions. In contrast to the

solution of pure Toeplitz systems, for general T+H systems the algorithms based on direct recursion are more efficient than the algorithms based on the inversion formula. The pure Schur-type algorithm based on the ZW-factorization of the matrix might have some advantages from the viewpoint of stability. Special attention is paid to the symmetric case in which a further reduction of the complexity is possible in some cases.

Notation. Throughout the paper, let $\mathbf{e}_k$ stand for the kth vector in the standard basis of $\mathbb{F}^n$. We set $\mathbf{e}_- = \mathbf{e}_1$ and $\mathbf{e}_+ = \mathbf{e}_n$. J_n denotes the $n \times n$ matrix of the counteridentity

$$J_n = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix},$$

and $\widehat{\mathbf{x}} = J_n\mathbf{x}$ for $\mathbf{x} \in \mathbb{F}^n$.

For a given vector $\mathbf{a} = (a_j)_{j=1-n}^{n-1} \in \mathbb{F}^{2n-1}$ we denote by $T_n(\mathbf{a})$ the $n \times n$ Toeplitz matrix

$$T_n(\mathbf{a}) = [\, a_{i-j} \,]_{i,j=1}^{n}.$$

Since for an $n \times n$ Hankel matrix H_n the matrix H_nJ_n is Toeplitz, any $n \times n$ T+H matrix can be represented in the form

$$M_n = T_n(\mathbf{a}) + T_n(\mathbf{b})J_n, \tag{1.1}$$

where $\mathbf{a} = (\, a_j \,)_{j=1-n}^{n-1}, \mathbf{b} = (\, b_j \,)_{j=1-n}^{n-1}$. Note that this representation is not unique, since the spaces of Toeplitz and Hankel matrices have a nontrivial (two-dimensional) intersection. Throughout the paper we assume that M_n is a nonsingular T+H matrix given by (1.1).

2. Inversion formula. It was shown in [17] that inverses of T+H matrices have a similarity in structure to inverses of Toeplitz and Hankel matrices. They are so-called *T+H-Bezoutians*. Practically, this means that the n^2 entries of the matrix are given by $O(n)$ parameters, which reduces the storage amount for the inverse matrix significantly. Furthermore and more importantly, matrix representations of T+H-Bezoutians, i.e., inversion formulas for T+H matrices, allow an efficient solution of linear systems of equations. They are also useful in iteration methods.

The parameters in the inversion formula can be described in different ways (see [18]): 1. with the help of a basis of the nullspace of a related T+H matrix, 2. with the help of the solution of certain "fundamental equations" emerging from the displacement structure of the matrix, and 3. from columns and rows of the inverse matrix. Note that the latter is possible only under some conditions.

Here we are going to present another version which is particularly convenient in connection with the algorithms we will discuss in the forthcoming sections. We represent the inverse of the T+H matrix M_n with the help of the first and last columns and rows of M_n^{-1} and of the inverse of an extension of M_n.

To be more precise, let $a_{\pm n}$, $a_{\pm(n+1)}$, and $b_{\pm n}$, $b_{\pm(n+1)}$ be arbitrary but chosen in such a way that the $(n+2) \times (n+2)$ T+H matrix

$$M_{n+2} = T_{n+2}(\widetilde{\mathbf{a}}) + T_{n+2}(\widetilde{\mathbf{b}})J_{n+2}\,, \quad \widetilde{\mathbf{a}} = (a_j)_{j=-n-1}^{n+1},\ \widetilde{\mathbf{b}} = (b_j)_{j=-n-1}^{n+1}$$

is nonsingular. It is easily checked that for almost all choices this is the case. Note that M_{n+2} has M_n as the central $n \times n$ submatrix.

Let $\mathbf{x}_-$ be the first and $\mathbf{x}_+$ the last column of M_n^{-1}, $(\mathbf{x}'_\pm)^T$ the last and the first rows of M_n^{-1}, and $\widetilde{\mathbf{x}}_\pm$ and $\widetilde{\mathbf{x}}'_\pm$ the corresponding quantities for M_{n+2}^{-1}. We show how the inverse of M_n is given by these eight vectors.

In order to simplify the inversion formula it is convenient to normalize the vectors $\widetilde{\mathbf{x}}_\pm$ and $\widetilde{\mathbf{x}}'_\pm$. Let Q be the 2×2 matrix given by

$$Q = \begin{bmatrix} \mathbf{e}_-^T \\ \mathbf{e}_+^T \end{bmatrix} [\, \widetilde{\mathbf{x}}_- \ \widetilde{\mathbf{x}}_+ \,],$$

the entries of which are just the first and last components of $\widetilde{\mathbf{x}}_\pm$. Using a Schur complement argument, one can show that Q is nonsingular, since M_n is nonsingular. Now we define $\widetilde{\mathbf{u}}_\pm$ and $\widetilde{\mathbf{u}}'_\pm$ by

$$[\, \widetilde{\mathbf{u}}_- \ \widetilde{\mathbf{u}}_+ \,] = [\, \widetilde{\mathbf{x}}_- \ \widetilde{\mathbf{x}}_+ \,]\, Q^{-1}, \quad [\, \widetilde{\mathbf{u}}'_- \ \widetilde{\mathbf{u}}'_+ \,] = [\, \widetilde{\mathbf{x}}'_- \ \widetilde{\mathbf{x}}'_+ \,]\, Q^{-T}.$$

The vectors $\widetilde{\mathbf{u}}_\pm$ have the form

$$\widetilde{\mathbf{u}}_- = \begin{bmatrix} 1 \\ \mathbf{z}_- \\ 0 \end{bmatrix}, \quad \widetilde{\mathbf{u}}_+ = \begin{bmatrix} 0 \\ \mathbf{z}_+ \\ 1 \end{bmatrix}$$

for some $\mathbf{z}_\pm \in \mathbb{F}^{\,n}$.

It is convenient to present the inversion formula in terms of the generating functions, which are defined as follows. If $A = [\, a_{ij} \,]_{i,j=1}^n$ is a matrix, then the *generating function of* A is the bivariate polynomial

$$A(t,s) = \sum_{i,j=1}^{n} a_{ij} t^{i-1} s^{j-1}.$$

In the same spirit we define $\mathbf{x}(t) = \sum_{j=1}^m x_j t^{j-1}$ for a vector $\mathbf{x} = (x_j)_{j=1}^m \in \mathbb{F}^{\,m}$.

The following is obtained from Theorem 3.1 in [17] (see also [22]) after some elementary rearrangements.

THEOREM 2.1. *The inverse of M_n is given by*

$$M_n^{-1}(t,s) = \frac{t\mathbf{x}_+(t)\widetilde{\mathbf{u}}'_+(s) - \widetilde{\mathbf{u}}_+(t)s\mathbf{x}'_+(s) + t\mathbf{x}_-(t)\widetilde{\mathbf{u}}'_-(s) - \widetilde{\mathbf{u}}_-(t)s\mathbf{x}'_-(s)}{(t-s)(1-ts)}. \tag{2.1}$$

Formula (2.1) can be written as a recursion for the entries of the matrix M_n^{-1} (see [17]). In this way M_n^{-1} can be constructed in $O(n^2)$ operations. More important are explicit matrix representations. The first representations of this kind were presented in [18]. More efficient formulas were given in [19] for the case where $\mathbb{F}$ is the field of complex numbers and in [20] and [21] for the case where $\mathbb{F}$ is the field of reals (see also [1], [2], [3], and references therein for related results). These formulas include only diagonal matrices and matrices of the discrete Fourier or related trigonometric transformations. Using these representations, matrix-vector multiplication can be carried out in $O(n \log n)$ operations. We suppose that the formulas in [19], [20], and [21] can also be written in terms of generalized circulants, so that the restriction to the real or complex numbers is not important.

In the inversion formula (2.1) eight vectors are involved, whereas the matrix depends on only $4n - 4$ parameters. That means that there is a lot of redundancy. It was noticed in [14] (see also [15, II, Prop. 2.4]) that the matrix M_n^{-1} actually can

be constructed with the help of only four vectors. Let us mention the corresponding result adapted to our notation. We introduce the vectors

$$\mathbf{g}_- = -(\, a_{j-n-1} + b_j \,)_{j=1}^n, \quad \mathbf{g}_+ = -(\, a_j + b_{j-n-1} \,)_{j=1}^n,$$

and the $n \times 4$ matrices

$$X = [\, \mathbf{x}_- \ \mathbf{z}_- \ \mathbf{z}_+ \ \mathbf{x}_+ \,], \quad G = [\, \mathbf{e}_1 \ \mathbf{g}_- \ \mathbf{g}_+ \ \mathbf{e}_n \,] \,.$$

Let S_n denote the $n \times n$ tridiagonal matrix with zeros on the main diagonal and ones on the adjoining diagonals.

THEOREM 2.2. *The columns* $\mathbf{c}_k = M_n^{-1}\mathbf{e}_k$ *of the inverse matrix can recursively be computed via*

$$\mathbf{c}_0 = \mathbf{0}\,, \quad \mathbf{c}_1 = \mathbf{x}_-\,, \quad \mathbf{c}_{k+1} = (S_n - XG^T)\mathbf{c}_k - \mathbf{c}_{k-1} \quad (k = 1, 2, \ldots, n). \tag{2.2}$$

A complexity analysis shows, however, that the application of Theorem 2.2 instead of Theorem 2.1 does not lead to more efficient algorithms, so this theorem is more of theoretical interest.

3. Levinson-type algorithm. Besides the $n \times n$ T+H matrix, matrix M_n and its nonsingular $(n+2) \times (n+2)$ extension $M_{n+2} = [\, c_{ij} \,]_{i,j=0}^{n+1}$, we consider the central submatrices $M_{n-2l} = [\, c_{ij} \,]_{i,j=l+1}^{n-l}$ for $l = 1, \ldots, m$, where $m = \left[\frac{n-1}{2}\right]$ and $[\,\cdot\,]$ denotes the integer part. Throughout the rest of the paper we assume that all submatrices M_{n-2l} are nonsingular. Recall that a matrix with this property will be called *centro-nonsingular*.

It is a crucial fact for the design of our algorithms that the central submatrices M_k of $M_n = T_n(\mathbf{a}) + T_n(\mathbf{b})J_n$ are given by

$$M_k = T_k(\mathbf{a}) + T_k(\mathbf{b})J_k,$$

where $T_k(\mathbf{a}) = [a_{i-j}]_{i,j=1}^k, k = 2, 4, \ldots, n/2$ if n is even, and $k = 1, 3, \ldots, (n+1)/2$ if n is odd.

We now describe a three-term recursion $(k-2, k) \to k+2$ for the first column $\mathbf{x}_k^-$ and last column $\mathbf{x}_k^+$ of M_k^{-1}. Arriving at $k = n$ we will obtain the vectors $\mathbf{x}_\pm = \mathbf{x}_n^\pm$ and $\widetilde{\mathbf{x}}_\pm = \mathbf{x}_{n+2}^\pm$ which are involved in the formula for the inverse matrix in Theorem 2.1.

In what follows we use the notation

$$\mathbf{c}(i:j) = [\, a_i + b_j \ \ldots \ a_j + b_i \,] \,.$$

We start with a simple observation that can easily be checked.

LEMMA 3.1. *If* $M_k\,\mathbf{f} = (g_j)_{j=1}^k$ *and* $\mathbf{f}_1(t) = (t^2+1)\mathbf{f}(t)$, *then*

$$M_{k+2}\mathbf{f}_1 = (g_j + g_{j-2})_{j=1}^{k+2},$$

where

$$g_0 = \mathbf{c}(-1:-k)\,\mathbf{f}, \ g_{-1} = \mathbf{c}(-2:-k-1)\,\mathbf{f}, \ g_{k+1} = \mathbf{c}(k:1)\,\mathbf{f}, \ g_{k+2} = \mathbf{c}(k+1:2)\,\mathbf{f}\,.$$

Applying this observation to the vectors $\mathbf{f} = \mathbf{x}_k^\pm$ for $k \geq 3$ and introducing the vectors $\check{\mathbf{x}}_k^\pm$ by

$$\check{\mathbf{x}}_k^\pm(t) = (t^2+1)\mathbf{x}_k^\pm(t) - t^2\mathbf{x}_{k-2}^\pm(t), \tag{3.1}$$

we obtain

$$M_{k+2}\check{\mathbf{x}}_k^{\pm} = \begin{bmatrix} r_{-3,k}^{\pm} - r_{-3,k-2}^{\pm} + \tau_{\mp} \\ r_{-2,k}^{\pm} - r_{-2,k-2}^{\pm} \\ \mathbf{0} \\ r_{2,k}^{\pm} - r_{2,k-2}^{\pm} \\ r_{3,k}^{\pm} - r_{3,k-2}^{\pm} + \tau_{\pm} \end{bmatrix},$$

where $\tau_+ = 1, \tau_- = 0$, and

$$\begin{aligned} r_{-2,k}^{\pm} &= \mathbf{c}(-1:-k)\,\mathbf{x}_k^{\pm}, & r_{-3,k}^{\pm} &= \mathbf{c}(-2:-k-1)\,\mathbf{x}_k^{\pm}, \\ r_{2,k}^{\pm} &= \mathbf{c}(k:1)\,\mathbf{x}_k^{\pm}, & r_{3,k}^{\pm} &= \mathbf{c}(k+1:2)\,\mathbf{x}_k^{\pm}. \end{aligned}$$

We denote $\alpha_{j,k}^{\pm} = r_{j,k}^{\pm} - r_{j,k-2}^{\pm}$ $(j = \pm 2, \pm 3)$ and define vectors

$$\mathbf{y}_k^{\pm} = \check{\mathbf{x}}_k^{\pm} - \alpha_{2,k}^{\pm} \begin{bmatrix} 0 \\ \mathbf{x}_k^{+} \\ 0 \end{bmatrix} - \alpha_{-2,k}^{\pm} \begin{bmatrix} 0 \\ \mathbf{x}_k^{-} \\ 0 \end{bmatrix}.$$

Then

$$M_{k+2} \begin{bmatrix} \mathbf{y}_k^{-} & \mathbf{y}_k^{+} \end{bmatrix} = \begin{bmatrix} \gamma_k^{--} & \gamma_k^{-+} \\ \mathbf{0} & \mathbf{0} \\ \gamma_k^{+-} & \gamma_k^{++} \end{bmatrix}$$

with

$$\begin{aligned} \gamma_k^{+\pm} &= \alpha_{3,k}^{\pm} + \tau_{\pm} - \alpha_{2,k}^{\pm} r_{2,k}^{+} - \alpha_{-2,k}^{\pm} r_{2,k}^{-}, \\ \gamma_k^{-\pm} &= \alpha_{-3,k}^{\pm} + \tau_{\mp} - \alpha_{2,k}^{\pm} r_{-2,k}^{+} - \alpha_{-2,k}^{\pm} r_{-2,k}^{-}. \end{aligned} \tag{3.2}$$

Due to the centro-nonsingularity of M_n, the matrix on the right-hand side has rank 2.

We introduce the matrices

$$A_k = \begin{bmatrix} \alpha_{-2,k}^{-} & \alpha_{-2,k}^{+} \\ \alpha_{2,k}^{-} & \alpha_{2,k}^{+} \end{bmatrix} \quad \text{and} \quad \Gamma_k = \begin{bmatrix} \gamma_k^{--} & \gamma_k^{-+} \\ \gamma_k^{+-} & \gamma_k^{++} \end{bmatrix}.$$

Then Γ_k is nonsingular, and we obtain the recursion

$$[\,\mathbf{x}_{k+2}^{-} \;\; \mathbf{x}_{k+2}^{+}\,] = [\,\mathbf{y}_k^{-} \;\; \mathbf{y}_k^{+}\,]\,\Gamma_k^{-1} = [\,\check{\mathbf{x}}_k^{-} \;\; \check{\mathbf{x}}_k^{+}\,]\,\Gamma_k^{-1} - \begin{bmatrix} 0 & 0 \\ \mathbf{x}_k^{-} & \mathbf{x}_k^{+} \\ 0 & 0 \end{bmatrix} A_k \Gamma_k^{-1}. \tag{3.3}$$

This recursion can be written in polynomial language. To give it a more compact form we introduce

$$\mathbf{x}_k(t) = [\,\mathbf{x}_k^{-}(t) \;\; \mathbf{x}_k^{+}(t)\,].$$

THEOREM 3.2. *The vector polynomials* $\mathbf{x}_k(t)$ *satisfy the recursion*

$$\mathbf{x}_{k+2}(t) = (\mathbf{x}_k(t)((t^2+1)I_2 - tA_k) - t^2\mathbf{x}_{k-2}(t))\Gamma_k^{-1}.$$

This theorem leads to a Levinson-type algorithm for computing the vectors $\mathbf{x}_n^{\pm}$ and $\mathbf{x}_{n+2}^{\pm}$ that are needed for applying the inversion formula. The recursion starts with $k = 3$ or $k = 4$, depending on whether n is even or odd. The initialization requires the solution of systems of order ≤ 4. For even n one can also start the recursion with $k = 2$ by artificially setting $r_{\pm 2,0}^{\mp} = -1$, $r_{\pm 2,0}^{\pm} = 0$, and $r_{\pm 3,0}^{\pm} = r_{\pm 3,0}^{\mp} = 0$. For the recursion from $k-2$ and k to $k+2$ one has to calculate first the eight numbers $r_j^{\pm}$ for $j = \pm 2, \pm 3$ which provide the entries of the matrices A_k and Γ_k. Then Theorem 3.2 can be applied.

Replacing $a_{\pm j}$ and $b_{\pm j}$ by $a_{\mp j}$ and $b_{\mp j}$, respectively, for $j = 1, 2, \dots, n$, we obtain a recursion for the first and last rows of the matrices M_k^{-1}. Recall that the first and last rows of M_n^{-1} and M_{n+2}^{-1} are involved in the inversion formula (2.1).

We introduce the 2×2 matrices

$$Q_k = \begin{bmatrix} \mathbf{e}_-^T \\ \mathbf{e}_+^T \end{bmatrix} [\, \mathbf{x}_k^- \ \mathbf{x}_k^+ \,]$$

consisting of the first and last components of $\mathbf{x}_k^{\pm}$.

COROLLARY 3.3. *The matrices Q_k satisfy the recursion*

$$Q_{k+2} = Q_k \, \Gamma_k^{-1}. \tag{3.4}$$

Instead of the vectors $\mathbf{x}_k^{\pm}$ we may consider their normalizations $\mathbf{u}_k^{\pm}$ defined by

$$[\, \mathbf{u}_k^- \ \ \mathbf{u}_k^+ \,] = [\, \mathbf{x}_k^- \ \ \mathbf{x}_k^+ \,] \, Q_k^{-1}.$$

The first component of $\mathbf{u}_k^-$ equals 1 and the last component zero, and the first component of $\mathbf{u}_k^+$ equals zero and the last component equals 1. The corresponding vector polynomials $\mathbf{u}_k(t) = [\, \mathbf{u}_k^-(t) \ \ \mathbf{u}_k^+(t) \,]$ satisfy the three-term recursion

$$\mathbf{u}_{k+2}(t) = \mathbf{u}_k(t)((t^2+1)I_2 - tB_k) - t^2 \mathbf{u}_{k-2}(t) C_k, \tag{3.5}$$

where

$$B_k = Q_k A_k Q_k^{-1}, \quad C_k = Q_{k-2} Q_k^{-1}.$$

The entries of the 2×2 matrices B_k and C_k can be computed via products of rows of M_k and the vectors $\mathbf{u}_k$, in the same way the matrices A_k and Γ_k are formed.

In the classical Levinson-type algorithms for Toeplitz matrices the replacement of the columns of the inverse matrix by the normalized vectors leads to a reduction in the complexity of the algorithm. A complexity analysis shows, however, that this is not the case in the present situation.

Complexity. We express the complexity for the computation of $\mathbf{x}_n^{\pm}$ and $\mathbf{x}_{n+2}^{\pm}$ in terms of additions (A) and of multiplications (M) and neglect lower order terms. The recursion from k to $k+2$ requires the computation of eight inner products of length k, $4k$ additions to get $\check{\mathbf{x}}_k^{\pm}$, and six vector additions and eight multiplications of a vector by a constant to compute $\mathbf{x}_{k+2}^{\pm}$. Taking into account that we carry out double steps this results in

$$\frac{9}{2}\, n^2 \ (A) \quad \text{plus} \quad 4\, n^2 \ (M).$$

This is less than for the fastest algorithm in [14], where $5\, n^2$ (A) plus $\frac{11}{2}\, n^2$ (M) are needed.

4. Schur-type algorithm. One of several motivations to consider Schur-type algorithms for T+H systems is that the Levinson-type recursion (3.3) includes inner product calculations to find the residuals $r_{j,k}^{\pm}$ $(j = \pm 2, \pm 3)$, which might not be convenient in parallel processing. But these parameters can also be computed by a Schur-type algorithm which can be carried out in parallel in an obvious manner. The Schur-type algorithm is also of independent interest since it provides a matrix factorization. This will be explained in section 5.

We start with some general definitions and observations. For $n-k$ even and $\mathbf{u} \in \mathbb{F}^k$, we denote

$$\rho_j(\mathbf{u}) = \begin{cases} \mathbf{c}(j+k-2 : j-1)\,\mathbf{u} & : \quad j = 1, \ldots, n+3-k\,, \\ \mathbf{c}(j+1 : j-k+2)\,\mathbf{u} & : \quad j = -1, \ldots, -(n+3-k)\,, \end{cases}$$

and $\rho^{\pm}(\mathbf{u}) = (\rho_{\pm j}(\mathbf{u}))_{j=1}^{n+3-k}$. We have, in particular,

$$M_{n+2} \begin{bmatrix} \mathbf{0} \\ \mathbf{u} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} (\rho_j(\mathbf{u}))_{j=-l-2}^{-1} \\ * \\ (\rho_j(\mathbf{u}))_{j=1}^{l+2} \end{bmatrix},$$

where $l = \frac{n-k}{2}$, the asterisk denotes a vector of length $k-2$, and $\mathbf{0}$ is a zero vector of length $l+1$. Therefore, we call $\rho^{\pm}(\mathbf{u})$ the *residual vectors* for $\mathbf{u}$.

We consider the vectors

$$\mathbf{u}^{(0)} = \begin{bmatrix} 0 \\ \mathbf{u} \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{u}^{(1)} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{u} \end{bmatrix} + \begin{bmatrix} \mathbf{u} \\ 0 \\ 0 \end{bmatrix}$$

in $\mathbb{F}^{k+2}$. In polynomial language, $\mathbf{u}^{(0)}(t) = t\mathbf{u}(t)$ and $\mathbf{u}^{(1)}(t) = (t^2+1)\mathbf{u}(t)$. It is easily checked that

$$\rho_j(\mathbf{u}^{(0)}) = \rho_{j+1}(\mathbf{u}) \quad \text{and} \quad \rho_j(\mathbf{u}^{(1)}) = \rho_{j+2}(\mathbf{u}) + \rho_j(\mathbf{u})$$

for $j = 1, \ldots, n+1-k$, and

$$\rho_j(\mathbf{u}^{(0)}) = \rho_{j-1}(\mathbf{u}) \quad \text{and} \quad \rho_j(\mathbf{u}^{(1)}) = \rho_{j-2}(\mathbf{u}) + \rho_j(\mathbf{u})$$

for $j = -1, \ldots, -(n+1-k)$. Translating this into polynomial language we obtain

$$\rho^{\pm}(\mathbf{u}^{(1)})(t) = P(1+t^{-2})\rho^{\pm}(\mathbf{u})(t) \quad \text{and} \quad \rho^{\pm}(\mathbf{u}^{(0)})(t) = Pt^{-1}\rho^{\pm}(\mathbf{u})(t)\,, \tag{4.1}$$

where P denotes the projection cutting off all negative powers of t.

We specify this observation for $\mathbf{x}_k^{\pm}$. Note that, clearly, $n-k$ is even, and that $\rho_j(\mathbf{x}_k^{\pm}) = r_{jk}^{\pm}$ for $j = \pm 2, \pm 3$, and $\tau_{\pm} = \rho_{\pm 1}(\mathbf{x}_k^{+})$, $\tau_{\mp} = \rho_{\pm 1}(\mathbf{x}_k^{-})$, as $r_{jk}^{\pm}$ and $\tau_{\pm}$ were defined in section 3.

We introduce vectors $\mathbf{r}_k^{+\pm} = \rho^{\pm}(\mathbf{x}_k^{+})$ and $\mathbf{r}_k^{-\pm} = \rho^{\pm}(\mathbf{x}_k^{-}) \in \mathbb{F}^{n+3-k}$ and the 2×2 matrix polynomials

$$\mathbf{r}_k(t) = \begin{bmatrix} \mathbf{r}_k^{--}(t) & \mathbf{r}_k^{-+}(t) \\ \mathbf{r}_k^{+-}(t) & \mathbf{r}_k^{++}(t) \end{bmatrix}.$$

From Theorem 3.2 and (4.1) we obtain the following theorem.

THEOREM 4.1. *The matrix polynomials* $\mathbf{r}_k(t)$ *satisfy the recursion*

$$\mathbf{r}_{k+2}(t) = (\mathbf{r}_k(t)((t^{-2}+1)I_2 - t^{-1}A_k)\Gamma_k^{-1} - t^{-2}\mathbf{r}_{k-2}(t))\Gamma_k^{-1}. \tag{4.2}$$

Actually, the projection P should be on the right-hand side of this recursion. However, according to construction of the matrices A_k and Γ_k, no negative powers of t appear, so $\mathbf{r}_{k+2}(t)$ given by (4.2) is really a polynomial and there is no need to write P.

The initialization for this recursion follows from the initialization of the recursion for the Levinson-type algorithm.

The algorithm emerging from Theorem 4.1 can be used to compute the factors in the recursion of Theorem 3.2 instead of using inner product calculations. But it also can be used without any reference to the vectors $\mathbf{x}_k^{\pm}$ to obtain a factorization of M_n. This will be discussed in section 6.

Let us point out that in each step of the algorithm the lengths of the four residual vectors are decreased by 1.

Complexity. For the computation of the residuals $\mathbf{r}_k$,

$$5\,n^2\ (A) \quad \text{plus} \quad 4\,n^2\ (M)$$

are needed. Using this recursion to compute the solutions $\mathbf{x}_k^{\pm}$ without inner product calculations, the overall complexity is

$$\frac{15}{2}\,n^2\ (A) \quad \text{plus} \quad 6\,n^2\ (M)\,.$$

5. Superfast algorithm. In this section we show how a combination of Theorems 3.2 and 4.1 leads to an $O(\mu(n)\log n)$ complexity algorithm for the solution of an $n \times n$ T+H system. Here $\mu(n)$ denotes the complexity of the multiplication of two polynomials with coefficients in $\mathbb{F}$ of degree n. In the case where $\mathbb{F}$ is the field of real or complex numbers we have $\mu(n) = O(n\log n)$ if FFT is employed, so the overall complexity will be $O(n\log^2 n)$. We refrain from presenting details because, first, the approach is quite standard (see [32] and references therein) and, second, it is not clear to us at the moment to what extent the algorithm is practical.

We introduce the quadratic matrix polynomial

$$\Phi_{k,k+2}(t) = \begin{bmatrix} 0 & -t^2\Gamma_k^{-1} \\ I_2 & ((t^2+1)I_2 - tA_k)\Gamma_k^{-1} \end{bmatrix}$$

and define

$$X_k(t) = [\,\mathbf{x}_{k-2}(t)\ \ \mathbf{x}_k(t)\,], \quad R_k(t) = [\,\mathbf{r}_{k-2}(t)\ \ \mathbf{r}_k(t)\,].$$

Then

$$X_{k+2}(t) = X_k(t)\Phi_{k,k+2}(t) \quad \text{and} \quad R_{k+2}(t) = R_k(t)\Phi_{k,k+2}(t^{-1}).$$

Note that $\Phi_{k,k+2}(t)(t)$ can be obtained from $R_k(t)$.

Now, for $m > k$ and even $m-k$, there is a matrix polynomial $\Phi_{k,m}(t)$ of degree $m-k$ for which

$$X_m(t) = X_k(t)\Phi_{k,m}(t) \quad \text{and} \quad R_m(t) = R_k(t)\Phi_{k,m}(t^{-1}). \tag{5.1}$$

For the application of the inversion formula we need $X_{n+2}(t)$. We can obtain this easily if we have $\Phi_{1,n+2}(t)$ in case n is odd or $\Phi_{2,n+2}(t)$ in case n is even.

We compute this using a divide-and-conquer strategy. To that aim we consider an algorithm "ALG" with input $(k, m, R_k(t))$ and output $\Phi_{k,m}(t)$. The algorithm works as follows.

If $m - k = 2$, then ALG simply applies the formulas from above to get $\Phi_{k,k+2}(t)$.

If $m - k > 2$, then we choose an l approximately in the middle of k and m, with even $l - k$, and apply the algorithm ALG for $(k, l, R_k(t))$ to get $\Phi_{k,l}(t)$. Then we compute $R_l(t)$ by (5.1) using fast polynomial multiplication. Then we apply ALG to $(l, m, R_l(t))$. The output is $\Phi_{l,m}(t)$. Finally we compute $\Phi_{k,m}(t) = \Phi_{k,l}(t)\Phi_{l,m}(t)$, again using fast polynomial multiplication.

Thus the algorithm ALG reduces the problem to two problems of approximately half the size plus a bounded number of polynomial multiplications. Hence the complexity to compute $\Phi_{1,n+2}(t)$ or $\Phi_{2,n+2}(t)$ is $O(\mu(n) \log n)$.

6. ZW- and WZ-factorization. We are going to use the algorithms above to produce some matrix factorizations. In particular we will see that the Levinson-type algorithm is related to a WZ-factorization of M_n^{-1}, whereas the Schur-type algorithm is related to a ZW-factorization of M_n.

Let us recall some concepts concerning ZW- and WZ-factorization. A matrix $A = [\, a_{ij} \,]_{i,j=1}^{n}$ is called a *W-matrix* (or a bow tie matrix) if $a_{ij} = 0$ for all (i, j) for which $i > j$ and $i + j > n$ or $i < j$ and $i + j \le n$. It will be called a *unit W-matrix* if in addition $a_{ii} = 1$ for $i = 1, \ldots, n$ and $a_{i,n+1-i} = 0$ for $i \ne (n + 1)/2$. The transpose of a W-matrix is called a *Z-matrix* (or hourglass matrix). A matrix which is both a Z- and a W-matrix is, by definition, an *X-matrix*. These names arise from the shapes of the possible positions for nonzero entries, which are as follows:

$$W = \begin{bmatrix} \bullet & & & & & \bullet \\ \bullet & \circ & & & \circ & \bullet \\ \bullet & \circ & \circ & \circ & \circ & \bullet \\ \bullet & \circ & \bullet & \bullet & \circ & \bullet \\ \bullet & \bullet & & & \bullet & \bullet \\ \bullet & & & & & \bullet \end{bmatrix}, \quad Z = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ & \circ & \circ & \circ & \bullet & \\ & & \circ & \bullet & & \\ & & \bullet & \circ & & \\ & \bullet & \circ & \circ & \circ & \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix},$$

$$X = \begin{bmatrix} \bullet & & & & & \bullet \\ & \bullet & & & \bullet & \\ & & \bullet & \bullet & & \\ & & \bullet & \bullet & & \\ & \bullet & & & \bullet & \\ \bullet & & & & & \bullet \end{bmatrix}.$$

A unit Z- or W-matrix is obviously nonsingular and a linear system with such a coefficient matrix can be solved in the same way as triangular systems by back substitution.

Any centro-nonsingular matrix A admits a unique factorization $A = ZXW$ in which W is a unit W-matrix, Z is a unit Z-matrix, and X is an X-matrix, and any matrix with such factorization is centro-nonsingular. Such a factorization will be called (unit) ZW-factorization of A. Analogously, a (unit) WZ-factorization is defined. A matrix admits a WZ-factorization if and only if the inverse is centro-nonsingular.

For simplicity of notation we assume that n is even, $n = 2m$. The case of odd n is similar.

We arrange the vectors $\mathbf{x}_k^{\pm}$ introduced in section 3 in a W-matrix V as follows:

$$V = \begin{bmatrix} & 0 & \cdots & \mathbf{0} & \mathbf{0} & \cdots & 0 & \\ \mathbf{x}_n^- & \mathbf{x}_{n-2}^- & \cdots & \mathbf{x}_2^- & \mathbf{x}_2^+ & \cdots & \mathbf{x}_{n-2}^+ & \mathbf{x}_n^+ \\ & 0 & \cdots & \mathbf{0} & \mathbf{0} & \cdots & 0 & \end{bmatrix}.$$

From the residuals of the vectors $\mathbf{x}_k^{\pm}$ we form vectors of length $l+1$, $l = \frac{n-k}{2}$,

$$\mathbf{z}_k^{+\pm} = \begin{bmatrix} \tau_{\pm} \\ r_{2,k}^{\pm} \\ r_{3,k}^{\pm} \\ \vdots \\ r_{l+1,k}^{\pm} \end{bmatrix}, \quad \mathbf{z}_k^{-\pm} = \begin{bmatrix} r_{-l-1,k}^{\pm} \\ r_{-l+1,k}^{\pm} \\ \vdots \\ r_{-2,k}^{\pm} \\ \tau_{\mp} \end{bmatrix}.$$

Now we observe that $M_n V = Z$ with

$$Z = \begin{bmatrix} 1 & \mathbf{z}_{n-2}^{--} & \cdots & \mathbf{z}_4^{--} & \mathbf{z}_2^{--} & \mathbf{z}_2^{-+} & \mathbf{z}_4^{-+} & \cdots & \mathbf{z}_{n-2}^{-+} & 0 \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & & & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{z}_{n-2}^{+-} & \cdots & \mathbf{z}_4^{+-} & \mathbf{z}_2^{+-} & \mathbf{z}_2^{++} & \mathbf{z}_4^{++} & \cdots & \mathbf{z}_{n-2}^{++} & 1 \end{bmatrix}. \tag{6.1}$$

We see that Z is a unit Z-matrix. Since $M_n = ZV^{-1}$ and V^{-1} is a W-matrix, Z is the left factor in the ZW-factorization of M_n. Since Z contains only residuals, it can be computed with the help of the Schur-type algorithm emerging from Theorem 4.1. In order to obtain the right factor one has to run the Schur-type algorithm for the transpose of the matrix M_n. This gives a unit Z-matrix which will be denoted by Z'. Now the right factor in the ZW-factorization of M_n is equal to Z'^T.

It remains to describe the computation of the middle factor X. For this we introduce the X-matrix

$$\Lambda = \operatorname{diag}(x_{1n}^-, \ldots, x_{12}^-, x_{22}^+, \ldots, x_{nn}^+) + J_n \operatorname{diag}(x_{nn}^-, \ldots, x_{22}^-, x_{12}^+, \ldots, x_{1n}^+),$$

where $x_{jk}^{\pm}$ denotes the jth component of $\mathbf{x}_k^{\pm}$. Then $W = V\Lambda^{-1}$ is a unit W-matrix and $M_n = Z\Lambda^{-1}W^{-1}$. Thus the middle factor X in the ZW-factorization of M_n is equal to Λ^{-1}.

Introducing Q_k as in section 3,

$$Q_k = \begin{bmatrix} x_{1k}^- & x_{1k}^+ \\ x_{kk}^- & x_{kk}^+ \end{bmatrix}, \quad P_k = Q_k^{-1} = \begin{bmatrix} \xi_k^{--} & \xi_k^{-+} \\ \xi_k^{+-} & \xi_k^{++} \end{bmatrix},$$

we have

$$X = \operatorname{diag}(\xi_n^{--}, \ldots, \xi_2^{--}, \xi_2^{++}, \ldots, \xi_n^{++}) + J_n \operatorname{diag}(\xi_n^{+-}, \ldots, \xi_2^{+-}, \xi_2^{-+}, \ldots, \xi_n^{-+}). \tag{6.2}$$

According to Corollary 3.3, $Q_{k+2} = Q_k \Gamma_k^{-1}$. Hence the following is true.

PROPOSITION 6.1. *The middle factor X of the ZW-factorization of M_n is given by* (6.2)*, where the entries can be obtained by the recursion*

$$P_{k+2} = \Gamma_k P_k, \quad P_2 = M_2.$$

We now discuss the WZ-factorization of M_n^{-1}. The left factor is given by $W = V\Lambda$. To describe the right factor we denote by V' the W-matrix built from the vectors $\mathbf{x}'^{\pm}_k$ analogously to V. The right factor is now ${W'}^T = \Lambda {V'}^T$. The middle factor is equal to Λ and can be computed with the help of Corollary 3.3.

If one is interested in the (unit) WZ-factorization of M_n^{-1}, it is more appropriate to consider the recursion of the monic normalizations $\mathbf{u}_k^{\pm}$ of the vectors $\mathbf{x}_k^{\pm}$, which were introduced at the end of section 3. Then the additional multiplications $W = V\Lambda$ and ${W'}^T = \Lambda {V'}^T$ can be avoided.

Furthermore, without these additional multiplications we can establish the nonunit WZ-factorization

$$M_n^{-1} = V X {V'}^T.$$

7. Solution of linear systems. In this section we discuss different possibilities for solving a linear system

$$M_n \mathbf{f} = \mathbf{b}$$

with a centro-nonsingular T+H matrix M_n. We compare these possibilities by counting the additions (A) and multiplications (M) in sequential processing.

All possibilities use either the Levinson-type algorithm emerging from Theorem 3.2, which will be called *Alg A*, or the Schur-type algorithm emerging from Theorem 4.1, which will be called *Alg B*, or both. Thus we have three types of algorithms: Levinson-type, mixed Levinson–Schur-type, and pure Schur-type algorithms. Levinson-type algorithms are labeled with (a), Schur-type algorithms with (b), and mixed algorithms with (c).

Let us note that all mixed and Schur-type algorithms have parallel complexity $O(n)$, since they do not include inner product calculations. Levinson-type algorithms have parallel complexity $O(n \log n)$. Furthermore, pure Schur-type algorithms can be expected to be more stable than the others for ill-conditioned systems.

7.1. Solution via inversion formula. A first possibility for solving a system with coefficient matrix M_n is to compute the data in the inversion formula (2.1) and to solve the system by fast matrix-vector multiplication. In the case where $\mathbb{F}$ is the field of real or complex numbers, the matrix-vector multiplication can be carried out efficiently with complexity $O(n \log n)$ if FFT or fast algorithms for trigonometric transforms are employed (see [19], [20], and [21]).

(a) To compute the data in formula (2.1) *Alg A* is applied to the matrix M_{n+2} and M_{n+2}^T. This requires

$$9\,n^2 \text{ (A)} \quad \text{plus} \quad 8\,n^2 \text{ (M)}.$$

In the case of a symmetric matrix this reduces to $\frac{9}{2}\,n^2$ (A) plus $4\,n^2$ (M). The alternative inversion formula (2.2) requires the application of *Alg A* only once, even if the matrix is not symmetric. However, to establish the inverse matrix with the help of this formula is much more costly than to run *Alg A* a second time. Therefore, we refrain from a further discussion.

(c) We apply *Alg A*, but instead of inner product calculations, *Alg B* is used. In doing so, it is not necessary to also run the algorithm for the transpose of the matrix, since the vectors $\mathbf{u}'_{\pm}$ and $\mathbf{x}'_{\pm}$ can be obtained from the residuals. In

fact, the application of *Alg A* and *Alg B* produces a factorization $M_{n+2}V_{n+2} = Z_{n+2}$, where Z_{n+2} is a unit Z-matrix and V_{n+2} is a W-matrix. Hence

$$(\widetilde{\mathbf{x}}'_-)^T = \mathbf{e}_1^T M_{n+2}^{-1} = [\,\widetilde{x}^-_{1,n+2}\ 0\ \ldots\ 0\ \widetilde{x}^+_{1,n+2}\,]Z_{n+2}^{-1}$$

and

$$(\widetilde{\mathbf{x}}'_+)^T = \mathbf{e}_{n+2}^T M_{n+2}^{-1} = [\,\widetilde{x}^-_{n+2,n+2}\ 0\ \ldots\ 0\ \widetilde{x}^+_{n+2,n+2}\,]Z_{n+2}^{-1}.$$

Similarly,

$$(\mathbf{x}'_-)^T = \mathbf{e}_1^T M_n^{-1} = [\,x^-_{1,n}\ 0\ \ldots\ 0\ x^+_{1,n}\,]Z_n^{-1}$$

and

$$(\mathbf{x}'_+)^T = \mathbf{e}_n^T M_n^{-1} = [\,x^-_{n,n}\ 0\ \ldots\ 0\ x^+_{n,n}\,]Z_n^{-1},$$

where Z_n is the central $n \times n$ submatrix of Z_{n+2}. Now, in order to obtain $\widetilde{\mathbf{x}}'_\pm$ and $\mathbf{x}'{}_\pm$, four Z-systems have to be solved, which requires $2\,n^2$ (A) plus $2\,n^2$ (M). This results in the overall complexity

$$\frac{19}{2}\,n^2\ \text{(A)}\quad \text{plus}\quad 8\,n^2\ \text{(M)}\,.$$

The symmetric case does not lead to a further reduction.

7.2. Solution via factorization.

(a) The application of *Alg A* to M_n and M_n^T produces, together with the recursion for the diagonal factor, a factorization $M_n^{-1} = VXV'^T$ with W-matrices V and V' and an X-matrix X. The solution of the system can be obtained by multiplying $\mathbf{b}$ first by a W-matrix, then by an X-matrix and, finally, by a Z-matrix. This requires n^2 (A) plus n^2 (M). The overall complexity is

$$10\,n^2\ \text{(A)}\quad \text{plus}\quad 9\,n^2\ \text{(M)}\,.$$

For a symmetric matrix M_n this reduces to $\frac{11}{2}\,n^2$ (A) plus $5\,n^2$ (M).

(b) The application of *Alg B* to M_n and M_n^T produces, together with the recursion of the X-factor, a factorization $M_n = ZXZ'^T$ with unit Z-matrices Z and Z'. The system can now be solved by back substitution, which requires n^2 (A) plus n^2 (M). Altogether we need

$$11\,n^2\ \text{(A)}\quad \text{plus}\quad 9\,n^2\ \text{(M)}\,.$$

For a symmetric matrix M_n this reduces to $6\,n^2$ (A) plus $5\,n^2$ (M).

(c) The application of *Alg A* and *Alg B* to M_n produces a factorization $M_nV = Z$, where V is a W-matrix and Z is a unit Z-matrix. Hence $\mathbf{f} = M_n^{-1}\mathbf{b} = VZ^{-1}\mathbf{b}$. That means the solution is obtained by first solving a unit Z-system and then multiplying the result by a W-matrix. This requires n^2 (A) plus n^2 (M). The overall complexity is

$$\frac{17}{2}\,n^2\ \text{(A)}\quad \text{plus}\quad 7\,n^2\ \text{(M)}\,.$$

There is no further reduction in the symmetric case.

7.3. Direct recursion. The application of the inversion formula or a factorization is particularly useful if several systems for the same coefficient matrix have to be solved. If only one system has to be solved, then it is reasonable to do it via direct recursion, which will be discussed now.

Suppose that $\mathbf{b} = (b_j)_{j=1}^{n}$ and $\mathbf{b}_k = (b_j)_{j=l+1}^{n-l}$, $l = \frac{n-k}{2}$. We solve the systems

$$M_k\,\mathbf{f}_k = \mathbf{b}_k$$

recursively in double steps $k \longrightarrow k+2$. We have

$$\mathbf{f}_{k+2} = \begin{bmatrix} 0 \\ \mathbf{f}_k \\ 0 \end{bmatrix} - \sigma_k^- \mathbf{x}_{k+2}^- - \sigma_k^+ \mathbf{x}_{k+2}^+ , \tag{7.1}$$

where

$$\sigma_k^- = \mathbf{c}(-1:-k)\,\mathbf{f}_k - b_l\,, \quad \sigma_k^+ = \mathbf{c}(k:1)\,\mathbf{f}_k - b_{n-l+1}\,.$$

(a) We apply *Alg A* to M_n and compute $\mathbf{f}_k$ by the recursion (7.1). The resulting algorithm will require

$$\frac{11}{2}\,n^2 \text{ (A)} \quad \text{plus} \quad 5\,n^2 \text{ (M)}\,.$$

(c) We apply *Alg A*, but instead of the inner product calculations, we use *Alg B*. The numbers $\sigma_k^{\pm}$ will be precomputed during the recursion. For this we introduce

$$\sigma_{j,k} = \begin{cases} \mathbf{c}(j+k-1:j)\,\mathbf{f}_k & : \quad j = 1,\dots,l-1\,, \\ \mathbf{c}(j:j-k+1)\,\mathbf{f}_k & : \quad j = -1,\dots,1-l\,. \end{cases}$$

Then $\sigma_k^- = \sigma_{-1,k} - b_{l-1}$, $\sigma_k^+ = \sigma_{1,k} - b_{n-l-2}$, and

$$\sigma_{j,k+2} = \begin{cases} \sigma_{j+1,k} - \sigma_k^- r_{j+1,k+2}^- - \sigma_k^+ r_{j+1,k+2}^+ & : \quad j > 0\,, \\ \sigma_{j-1,k} - \sigma_k^- r_{j-1,k+2}^- - \sigma_k^+ r_{j-1,k+2}^+ & : \quad j < 0\,, \end{cases}$$

where $r_{j\pm1,k+2}^{\pm}$ are computed with the help of Theorem 4.1. The overall complexity will be

$$\frac{17}{2}\,n^2 \text{ (A)} \quad \text{plus} \quad 7\,n^2 \text{ (M)}\,.$$

In both (a) and in (c) we do not have a reduction in the case where the matrix is symmetric.

We collect the coefficients of n^2 of the complexity estimates in a table. The lower order terms are $O(n \log n)$ for 7.1 and $O(n)$ for 7.2 and 7.3.

Method	General		Symmetric	
	(A)	(M)	(A)	(M)
7.1 (a)	9	8	4.5	4
7.1 (c)	9.5	8	9.5	8
7.2 (a)	10	9	5.5	5
7.2 (b)	11	9	6	5
7.2 (c)	8.5	7	8	7
7.3 (a)	5.5	5	5.5	5
7.3 (c)	8.5	7	8.5	7

Let us reiterate that a lower complexity does not mean that this method is always preferable to the others; other important issues also have to be taken into account. For example, in parallel processing methods (b) and (c) are preferable to methods (a). Furthermore, practical experience and theoretical results (see [4] and references therein) suggest that, as a rule, Schur-type algorithms are more stable than Levinson-type algorithms. This means that method 7.2 (b) is preferable to the others from this point of view.

REFERENCES

[1] D. A. Bini and V. Y. Pan, *Polynomial and Matrix Computations*, Birkhäuser Verlag, Basel, Boston, Berlin, 1994.

[2] E. Bozzo, *Algebras of higher dimension for displacement decompositions and computations with Toeplitz-plus-Hankel matrices*, Linear Algebra Appl., 230 (1995), pp. 127–150.

[3] E. Bozzo and C. Di Fiore, *On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decompositions*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 312–326.

[4] R. P. Brent, *Stability of fast algorithms for structured linear systems*, in Fast Reliable Algorithms for Matrices with Structure, T. Kailath and A. H. Sayed, eds., SIAM, Philadelphia, 1999, pp. 103–116.

[5] P. Delsarte and Y. Genin, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 470–478.

[6] P. Delsarte and Y. Genin, *On the splitting of classical algorithms in linear prediction theory*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 645–653.

[7] C. J. Demeure, *Bowtie factors of Toeplitz matrices by means of split algorithms*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 1601–1603.

[8] D. J. Evans and M. Hatzopoulos, *A parallel linear system solver*, Internat. J. Comput. Math., 7 (1979), pp. 227–238.

[9] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.

[10] I. Gohberg and I. Koltracht, *Efficient algorithm for Toeplitz plus Hankel matrices*, Integral Equations Operator Theory, 12 (1989), pp. 136–142.

[11] G. Heinig, *Chebyshev-Hankel matrices and the splitting approach for centrosymmetric Toeplitz-plus-Hankel matrices*, Linear Algebra Appl., 327 (2001), pp. 181–196.

[12] G. Heinig, *Inversion of Toeplitz-plus-Hankel matrices with any rank profile*, in Fast Algorithms for Structured Matrices, Contemp. Math. 323, V. Olshevsky, ed., AMS, Providence, RI, 2004, pp. 75–90.
[13] G. Heinig and A. Bojanczyk, *Transformation techniques for Toeplitz and Toeplitz-plus-Hankel matrices.* II. *Algorithms*, Linear Algebra Appl., 278 (1998), pp. 11–36.
[14] G. Heinig, P. Jankowski, and K. Rost, *Fast inversion algorithms of Toeplitz-plus-Hankel matrices*, Numer. Math., 52 (1988), pp. 665–682.
[15] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhäuser Verlag, Basel, Boston, Stuttgart and Akademie-Verlag, Berlin, 1984.
[16] G. Heinig and K. Rost, *Fast inversion of Toeplitz-plus-Hankel matrices*, Wiss. Z. TH Karl-Marx-Stadt, 27 (1985), pp. 66–71.
[17] G. Heinig and K. Rost, *On the inverses of Toeplitz-plus-Hankel matrices*, Linear Algebra Appl., 106 (1988), pp. 39–52.
[18] G. Heinig and K. Rost, *Matrix representations of Toeplitz-plus-Hankel matrix inverses*, Linear Algebra Appl., 113 (1989), pp. 65-78.
[19] G. Heinig and K. Rost, *DFT representations of Toeplitz-plus-Hankel Bezoutians with application to fast matrix-vector multiplication*, Linear Algebra Appl., 284 (1998), pp. 157–175.
[20] G. Heinig and K. Rost, *Hartley transform representations of inverses of real Toeplitz-plus-Hankel matrices*, Numer. Funct. Anal. Optim., 21 (2000), pp. 175–189.
[21] G. Heinig and K. Rost, *Efficient inversion formulas for Toeplitz-plus-Hankel matrices using trigonometric transformations*, in Structured Matrices in Mathematics, Computer Science, and Engineering, Vol. 2, Contemp. Math. 281, V. Olshevsky, ed., AMS, Providence, RI, 2001, pp. 247–264.
[22] G. Heinig and K. Rost, *Centrosymmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices and Bezoutians*, Linear Algebra Appl., 366 (2003), pp. 257–281.
[23] G. Heinig and K. Rost, *Fast algorithms for skewsymmetric Toeplitz matrices*, in Toeplitz matrices and singular integral equations, Oper. Theory Adv. Appl. 135, A. Böttcher, I. Gohberg, and P. Junghanns, eds., Birkhäuser, Basel, 2002, pp. 193–208.
[24] G. Heinig and K. Rost, *Split algorithms for skewsymmetric Toeplitz matrices with arbitrary rank profile*, Theoret. Comput. Sci., to appear.
[25] G. Heinig and K. Rost, *Split algorithms for symmetric Toeplitz matrices with arbitrary rank profile*, Numer. Linear Algebra Appl., to appear.
[26] G. Heinig and K. Rost, *Fast algorithms for centro-symmetric and centro-skewsymmetric Toeplitz-plus-Hankel matrices*, Numer. Algorithms, 33 (2003), pp. 305–317.
[27] A. Melman, *A symmetric algorithm for Toeplitz systems*, Linear Algebra Appl., 301 (1999), pp. 145–152.
[28] A. Melman, *The even-odd split Levinson Algorithm for Toeplitz systems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 256–270.
[29] A. Melman, *A two-step even-odd split Levinson algorithm for Toeplitz systems*, Linear Algebra Appl., 338 (2001), pp. 219–237.
[30] G. A. Merchant and T. W. Parks, *Efficient solution of a Toeplitz-plus-Hankel coefficient matrix system of equations*, IEEE Trans. Acoust. Speech Signal Process., 30 (1982), pp. 40–44.
[31] A. B. Nersesjan and A. A. Papoyan, *Construction of the matrix inverse to the sum of Toeplitz and Hankel matrices*, Izv. AN Arm. SSR, Matematika, 8 (1983), pp. 150–160 (in Russian).
[32] V. Y. Pan, *Structured Matrices and Polynomials*, Birkhäuser Verlag, Boston and Springer-Verlag, New York, 2001.
[33] S. Chandra Sekhara Rao, *Existence and uniqueness of WZ factorization*, Parallel Comput., 23 (1997), pp. 1129–1139.
[34] A. H. Sayed, H. Lev-Ari, and T. Kailath, *Fast triangular factorization of the sum of quasi-Toeplitz and quasi-Hankel matrices*, Linear Algebra Appl., 191 (1993), pp. 77–106.
[35] A. E. Yagle, *New analogs of split algorithms for arbitrary Toeplitz-plus-Hankel matrices*, IEEE Trans. Signal Process., 39 (1991), pp. 2457–2463.
[36] C. J. Zarowski, *A Schur algorithm and linearly connected processor array for Toeplitz-plus-Hankel matrices*, IEEE Trans. Inform. Theory, 40 (1992), pp. 2065–2078.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 858–899

ORTHOGONAL EIGENVECTORS AND RELATIVE GAPS*

INDERJIT S. DHILLON† AND BERESFORD N. PARLETT‡

Abstract. This paper presents and analyzes a new algorithm for computing eigenvectors of symmetric tridiagonal matrices factored as LDL^t, with D diagonal and L unit bidiagonal. If an eigenpair is well behaved in a certain sense with respect to the factorization, the algorithm is shown to compute an approximate eigenvector which is accurate to working precision. As a consequence, all the eigenvectors computed by the algorithm come out numerically orthogonal to each other without making use of any reorthogonalization process. The key is first running a qd-type algorithm on the factored matrix LDL^t and then applying a fine-tuned version of inverse iteration especially adapted to this situation. Since the computational cost is $O(n)$ per eigenvector for an $n \times n$ matrix, the proposed algorithm is the central step of a more ambitious algorithm which, at best (i.e., when all eigenvectors are well-conditioned), would compute all eigenvectors of an $n \times n$ symmetric tridiagonal at $O(n^2)$ cost, a great improvement over existing algorithms.

Key words. eigenvectors, orthogonality, symmetric tridiagonal matrix, bidiagonal factorization, high relative accuracy

AMS subject classifications. 15A18, 15A23

DOI. 10.1137/S0895479800370111

1. Setting the scene.

This section is addressed to a broader audience than is the rest of the paper.

A real symmetric matrix has a full set of orthogonal eigenvectors, and users of software expect computed eigenvectors to be orthogonal to working accuracy. Excellent programs are available to diagonalize real symmetric matrices, so we might be tempted to say that the problem of computing orthogonal eigenvectors is solved. The best approach has three phases: (1) reducing the given dense symmetric matrix A to tridiagonal form T, (2) computing the eigenvalues and eigenvectors of T, and (3) mapping T's eigenvectors into those of A. For an $n \times n$ matrix the first and third phases require $O(n^3)$ arithmetic operations each. There are several choices for the second phase. The QR algorithm is simple but rather slow. The time-consuming part is the accumulation of $O(n^2)$ plane rotations, each of which requires $O(n)$ operations. Yet we must remember that it is this accumulation that guarantees numerically orthogonal eigenvectors, however close some of the eigenvalues may be, and that is a beautiful feature of the QR algorithm [15, 26]. An attractive feature of QR is that it requires only $O(n^2)$ operations to compute the eigenvalues alone. In principle, once the eigenvalues are known, one can invoke inverse iteration to independently compute each eigenvector at a cost of $O(n)$ per eigenvector. For distributed memory computers this feature would permit computation of the eigenvectors in parallel. The blemish in this approach is that the computed eigenvectors may not be numerically orthogonal when some eigenvalues are close, say agreeing to more than three decimals. So inverse

*Received by the editors March 31, 2000; accepted for publication (in revised form) July 8, 2003; published electronically March 30, 2004.

http://www.siam.org/journals/simax/25-3/37011.html

†Department of Computer Sciences, University of Texas, Austin, TX 78712-1188 (inderjit @cs.utexas.edu). Part of this author's research was supported while the author was at the University of California, Berkeley, by DARPA contract DAAL03-91-C-0047 and NSF grants ASC-9313958 and CDA-9401156. At the University of Texas, this research was supported by a University startup grant and NSF CAREER Award ACI-0093404.

‡Mathematics Department and Computer Science Division, EECS Department, University of California, Berkeley, CA 94720 (parlett@math.berkeley.edu).

iteration is augmented with a Gram–Schmidt process to enforce orthogonality, thus making it an $O(n^3)$ procedure in the worst case. A very careful implementation of this approach is available in the LAPACK library [25]. Since the mid 1990s, stable versions of the divide and conquer method for Phase 2 have been available [21]. Divide and conquer is a fast method much of the time but can reduce to an $O(n^3)$ process for rather uniform eigenvalue distributions.

With n near 1000 there are cases where the $O(n^3)$ reduction of a dense matrix to tridiagonal form T takes *much less* time (10–20%) than computing T's eigenvectors by inverse iteration. For example, the tridiagonal reduction of a certain 1000×1000 dense matrix takes about 10 seconds, while LAPACK's bisection and inverse iteration software takes 93 seconds to compute all the eigenvalues and eigenvectors of the tridiagonal. The timings for a 2000×2000 matrix clearly show the $O(n^3)$ behavior: 101 seconds for tridiagonalization and 839 seconds for solving the tridiagonal eigenproblem; detailed timing results are given in section 8.1. This behavior was an unpleasant surprise for the guardians of LAPACK and was brought to our attention in 1995 by a group of computational quantum chemists who were interested in a parallel solution. For one of their examples of order 966 there was a cluster of 931 eigenvalues deemed to be close to each other, and the Gram–Schmidt process in inverse iteration was consuming all the time (see section 8.1). It was time to re-evaluate Phase 2.

As values of n near 10^3 become common and values exceeding 10^4 do occur, it is hard to stop people dreaming of an algorithm that is guaranteed to compute a numerically orthogonal set of eigenvectors of T in $O(n^2)$ operations in the worst case. The presence of parallel distributed memory computer systems has vitalized the search for such an algorithm. Ideally the n eigenvalues would be equally distributed among all processors, along with a copy of the tridiagonal, and the eigenvectors would be computed independently at the same time and would turn out to be orthogonal to working accuracy.

There are formidable obstacles that impede the realization of this dream, and these will be reviewed in the next section.

This paper presents a central step towards the goal. The method proposed, in section 4, is Algorithm Getvec for computing a single eigenvector in $O(n)$ time. The main theorem, Theorem 15 in section 7, shows that in special, but important, situations (see below) our new algorithm produces an eigenvector that is guaranteed to be within $O(n\varepsilon)$ of the true eigenvector whenever the eigenvalue has a *relative* separation from its neighbors that exceeds a threshold *tol*, say 10^{-3}. It has been known for years that inverse iteration can produce fully accurate eigenvectors whenever the eigenvalue has an *absolute* separation that is above the threshold. So our contribution is to change absolute to relative in the separation condition. Our examples show that the resulting speedups can be dramatic (from 839 seconds to 4.6 seconds). Section 8 contains detailed experimental results. To establish our result, roundoff errors included, we were obliged to jettison the traditional representation of a tridiagonal matrix by its diagonal and next-to-diagonal entries. Instead, we use a bidiagonal factorization LDL^t of a carefully chosen translate of the original tridiagonal T. The crucial properties that must be satisfied in order for Algorithm Getvec to compute an accurate approximation to eigenvector $\boldsymbol{v}$ (corresponding to eigenvalue λ of LDL^t) are that (i) both λ and $\boldsymbol{v}$ must be determined to high relative accuracy by the parameters in L and D and (ii) the *relative gap* between λ and its nearest neighbor μ in the spectrum should exceed *tol*; $|\lambda - \mu| > tol \cdot |\lambda|$. The phrase "determined to high relative accuracy" is explained in section 2. We say that an LDL^t factorization that satisfies

property (i) is a *relatively robust representation* (RRR) for the eigenpair $(\lambda, \boldsymbol{v})$. A positive (or negative) definite LDL^t factorization is known to be an RRR for all its eigenpairs [7]; in section 6.1 we give conditions for an indefinite LDL^t to be an RRR.

The proof of the main theorem, Theorem 15, rests on the existence of relative perturbation results for the bidiagonal factors and on a special interpretation of the roundoff errors in differential qd algorithms that yields what is called mixed stability: carefully selected small relative perturbations of both the input and the output of our subroutines reveal the existence of an exact relationship of the form $\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I = \bar{N}\bar{D}\bar{N}^t$, where $\bar{N}$ is a "twisted" factor defined in section 4 and $\hat{\lambda}$ is an approximation to λ. The translation by $\hat{\lambda}$ preserves eigenvectors while shifting the eigenvalue of interest very close to 0. The middle part of this paper presents the relevant error analysis. Although essential for our results, this analysis will be indigestible for most readers, but it tells us that changes of only 3 or 4 units in the last digit of each entry of the input L, D and the output twisted factors suffice to give the exact relation.

The algorithm presented in this paper, Algorithm Getvec, allows us to compute a very accurate eigenvector for each eigenvalue that has a large relative separation from its neighbors ($> tol$). How do we compute numerically orthogonal approximations to eigenvectors when relative gaps are smaller? The full method, Algorithm MRRR or MR^3 (Algorithm of Multiple Relatively Robust Representations), crucially depends on Algorithm Getvec but is beyond the scope of the paper and is described in detail in [10]. Here we briefly sketch the outline of Algorithm MR^3 to show how it is based on the results of this paper: Compute the extreme eigenvalues of T and start with a base τ at one end of the spectrum. Compute the positive (or negative) definite factorization $LDL^t = \pm(T - \tau I)$ and find all its eigenvalues to high relative accuracy. Next invoke Algorithm Getvec on LDL^t to compute eigenvectors for all the eigenvalues λ of LDL^t that have large relative gaps. For each cluster of relatively close eigenvalues, pick a new base τ_c at, or close to, one end of the cluster. Perform a careful factorization $L_cD_cL_c^t = LDL^t - \tau_c I$ to get a new RRR. Shifting by τ_c increases the relative separations of eigenvalues in the cluster. Refine, to high relative accuracy, the shifted eigenvalues (of $L_cD_cL_c^t$) that now have relative gaps exceeding tol, and invoke Algorithm Getvec on the new factorization $L_cD_cL_c^t$ to compute their eigenvectors. Repeat the process with suitable bases τ until all eigenvectors have been computed. For more details on Algorithm MR^3 the reader is referred to [10], which also addresses the tricky aspect of showing that the eigenvectors computed using the various LDL^t factorizations are numerically orthogonal.

We now give a brief outline of the paper. Section 2 elaborates on the difficulties in achieving our goals, and section 3 demonstrates the need to use an LDL^t factorization to replace T. We present the proposed algorithm in section 4, while the "mixed" roundoff error analyses and associated commutative diagrams are given in section 5. To prove the correctness of the proposed algorithm, the various parts of a commutative diagram are analyzed in detail in section 6, and the main theorem is proved in section 7. Numerical results and timing comparisons are given in section 8. Finally, section 9 discusses an extension to the computation of singular vectors.

Householder notation (capital letters for matrices, Greek lowercase for scalars, and lowercase bold roman for vectors) is generally followed. The norm $\|\cdot\|$ will refer to the 2-norm. Sections 3, 4, 5, and 6.2 are derived from [9].

2. Difficulties. The quality of an approximate eigenvector $\boldsymbol{z}$ is measured by its residual. The following basic result, which goes back to Temple in the 1930s, if not earlier, will be needed later. See [28, Chapters 10 and 11] for details and a proof.

THEOREM 1. *Let $A = A^t$ be a real matrix that has a simple eigenvalue λ with normalized eigenvector $\boldsymbol{v}$. For any unit vector $\boldsymbol{z}$ and a scalar μ, closer to λ than to any other eigenvalue,*

$$|\sin\angle(\boldsymbol{v},\boldsymbol{z})| \le \|A\boldsymbol{z} - \boldsymbol{z}\mu\|/\mathrm{gap}(\mu), \tag{1}$$

where $\mathrm{gap}(\mu) = \min\{|\nu - \mu| : \nu \ne \lambda,\ \nu \in \mathit{spectrum}(A)\}$. *In addition, the error in the eigenvalue is bounded by the residual norm, i.e.,*

$$|\mu - \lambda| \ \le\ \|A\boldsymbol{z} - \boldsymbol{z}\mu\|.$$

The sad fact is that a small residual norm does not guarantee an accurate eigenvector when $\mathrm{gap}(\mu)$ is also small. On the other hand, accurate approximations $\boldsymbol{y}$ to $\boldsymbol{u}$ and $\boldsymbol{z}$ to $\boldsymbol{v}$ (where $\boldsymbol{u}$ is an eigenvector orthogonal to $\boldsymbol{v}$), in the strong sense that

$$|\sin\angle(\boldsymbol{u},\boldsymbol{y})| < n\varepsilon \ \text{ and } \ |\sin\angle(\boldsymbol{v},\boldsymbol{z})| < n\varepsilon, \tag{2}$$

where ε is the machine precision, do ensure numerical orthogonality of the *computed* eigenvectors since

$$|\cos\angle(\boldsymbol{y},\boldsymbol{z})| \le |\sin\angle(\boldsymbol{u},\boldsymbol{y})| + |\sin\angle(\boldsymbol{v},\boldsymbol{z})| < 2n\varepsilon.$$

Thus accuracy yields orthogonality. This observation is not as vacuous as it appears. In the QR algorithm the computed eigenvectors are acceptable because they are orthogonal (numerically) and their residuals are small *but* they are not always accurate in the sense of (2). Part of the explanation for this anomaly is that A may not determine some of its eigenpairs to high accuracy. Thus the eigenvector $\boldsymbol{v}$ used above may be highly sensitive as soon as there is uncertainty in the entries of A and so the concept of accuracy goes out of focus. That is why, in the sense of (2), accuracy is not the only way to compute numerically orthogonal eigenvectors; the QR algorithm is a good example.

Let us return to the residual norm. *In general*, the best we can hope for is to produce residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{z}) = A\boldsymbol{z} - \boldsymbol{z}\mu$ satisfying

$$\|\boldsymbol{r}\| \le n\varepsilon(\lambda_{max} - \lambda_{min}). \tag{3}$$

By (1) and (3), if $\mathrm{gap}(\mu) \ge \mathit{tol}\cdot(\lambda_{max} - \lambda_{min})$, where tol is the gap threshold (say 10^{-3}), then

$$|\sin\angle(\boldsymbol{v},\boldsymbol{z})| \le n\varepsilon/\mathit{tol}$$

and accuracy is assured (throughout the paper we assume that $n^3\varepsilon \le 1$). On the other hand, in the many cases when $\mathrm{gap}(\mu) \ll \mathit{tol}$, the residual norm must be much smaller than the right-hand side of (3) in order to deliver the accuracy of (2).

In general we see no possibility for reducing the residuals without using higher precision arithmetic in parts of the computation. Instead we turn to special matrices and special situations, in particular, to a symmetric tridiagonal matrix T. Our goal is to compute residuals satisfying

$$\|\boldsymbol{r}\| = \|T\boldsymbol{z} - \boldsymbol{z}\hat{\lambda}\| \le Kn\varepsilon|\hat{\lambda}| \tag{4}$$

for some modest constant K independent of $\boldsymbol{z}$ and $\hat{\lambda}$, so that, by Theorem 1,

$$|\sin\angle(\boldsymbol{v},\boldsymbol{z})| \le \frac{Kn\varepsilon|\hat{\lambda}|}{\mathrm{gap}(\hat{\lambda})} = \frac{Kn\varepsilon}{\mathrm{relgap}(\hat{\lambda})},$$

where relgap$(\hat{\lambda}) := \text{gap}(\hat{\lambda})/|\hat{\lambda}|$. Note that if $\hat{\lambda} = O(\varepsilon(\lambda_{max} - \lambda_{min}))$, then (4) requires $\|\boldsymbol{r}\| = O(\varepsilon^2(\lambda_{max} - \lambda_{min}))$. How is that possible since even the rounded version of the "true" eigenvector may not achieve (4)?

To achieve (4) we need three separate properties:

I. The eigenpair $(\lambda, \boldsymbol{v})$ must be determined to high relative accuracy by the matrix parameters. See the next paragraph for definitions.

II. The computed $\hat{\lambda}$ must approximate λ to high relative accuracy, i.e., $|\lambda - \hat{\lambda}| = O(n\varepsilon|\hat{\lambda}|)$.

III. The vector $\boldsymbol{z}$ must then be computed to satisfy (4).

To achieve property I we discard the traditional representation of T in favor of a suitable LDL^t factorization of T or some translate $T - \tau I$. Write l_i for $L(i+1, i)$ and d_i for $D(i, i)$. We say that $(\lambda, \boldsymbol{v})$ is *determined to high relative accuracy* by L and D if small relative changes, $l_i \to l_i(1 + \eta_i)$, $d_i \to d_i(1 + \delta_i)$, $|\eta_i| < \xi$, $|\delta_i| < \xi$, cause changes $\delta\lambda$ and $\delta\boldsymbol{v}$ that satisfy

$$\frac{|\delta\lambda|}{|\lambda|} \leq K_1 n\xi, \quad \lambda \neq 0, \tag{5}$$

$$|\sin\angle(\boldsymbol{v}, \boldsymbol{v} + \delta\boldsymbol{v})| \leq \frac{K_2 n\xi}{\text{relgap}(\lambda)} \tag{6}$$

for modest constants K_1 and K_2, say, smaller than 100. We call such an LDL^t factorization an RRR for $(\lambda, \boldsymbol{v})$. The backward stable QR algorithm on T cannot guarantee such accuracy.

Section 3 shows the necessity for the change of representation to LDL^t. Property II is then easily achieved by a bisection algorithm that uses differential qd transforms (see section 4.1) or, in the positive definite case, by the dqds algorithm; see [13]. Given properties I and II, we can think of satisfying property III by using inverse iteration. While traditional inverse iteration often works well in practice, we employ an elegant alternative that uses a rank-revealing twisted factorization of $T - \hat{\lambda}I$ to obtain a starting vector that is guaranteed to be good.

A subtle point in our analysis is that (4) is achieved, not for T or LDL^t but for a small relative perturbation of LDL^t.

Much of this paper, from section 4 onwards, is devoted to the algorithm and a proof to show that property III can be achieved in the presence of roundoff error.

3. Standard tridiagonal form is inadequate. In this section we show that the standard representation of tridiagonals is inadequate for our purpose of computing highly accurate eigenvectors. Recent work has shown that some tridiagonal classes do determine all their eigenvalues to high relative accuracy [8]. However, for many tridiagonals small relative changes in the diagonal and off-diagonal entries can cause huge relative changes in the small eigenvalues.

We now give a carefully constructed example which exhibits this relative instability even when $n = 3$.

Example 1. Consider the tridiagonal

$$T_1 = \begin{bmatrix} 1 - \sqrt{\varepsilon} & \varepsilon^{1/4}\sqrt{1 - 7\varepsilon/4} & 0 \\ \varepsilon^{1/4}\sqrt{1 - 7\varepsilon/4} & \sqrt{\varepsilon} + 7\varepsilon/4 & \varepsilon/4 \\ 0 & \varepsilon/4 & 3\varepsilon/4 \end{bmatrix}$$

and a small relative perturbation to the off-diagonals of T_1,

$$T_1 + \delta T_1 = \begin{bmatrix} 1-\sqrt{\varepsilon} & \varepsilon^{1/4}(1+\varepsilon)\sqrt{1-7\varepsilon/4} & 0 \\ \varepsilon^{1/4}(1+\varepsilon)\sqrt{1-7\varepsilon/4} & \sqrt{\varepsilon}+7\varepsilon/4 & \varepsilon(1+\varepsilon)/4 \\ 0 & \varepsilon(1+\varepsilon)/4 & 3\varepsilon/4 \end{bmatrix},$$

where ε is a small quantity of the order of the machine precision. The two smallest eigenvalues of T_1 and $T_1 + \delta T_1$ are[1]

$$\lambda_1 = \varepsilon/2 + \varepsilon^{3/2}/8 + O(\varepsilon^2), \quad \lambda_1 + \delta\lambda_1 = \varepsilon/2 - 7\varepsilon^{3/2}/8 + O(\varepsilon^2),$$

$$\lambda_2 = \varepsilon - \varepsilon^{3/2}/8 + O(\varepsilon^2), \quad \lambda_2 + \delta\lambda_2 = \varepsilon - 9\varepsilon^{3/2}/8 + O(\varepsilon^2),$$

while

$$\lambda_3 = 1 + \varepsilon + O(\varepsilon^2), \qquad \lambda_3 + \delta\lambda_3 = 1 + \varepsilon + 2\varepsilon^{3/2} + O(\varepsilon^2).$$

Thus

$$\left|\frac{\delta\lambda_i}{\lambda_i}\right| = (3-i)\sqrt{\varepsilon} + O(\varepsilon), \quad i = 1, 2,$$

and the relative change in these eigenvalues is much larger than the initial relative perturbations in the entries of T_1. Similarly the corresponding eigenvectors of T_1 and $T_1 + \delta T_1$ are

$$\boldsymbol{v}_1 = \begin{bmatrix} \frac{\varepsilon^{1/4}}{\sqrt{2}}(1+\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ -\frac{1}{\sqrt{2}}(1-\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon) \\ \frac{1}{\sqrt{2}}(1-\frac{3\varepsilon}{4}) + O(\varepsilon^{3/2}) \end{bmatrix}, \quad \boldsymbol{v}_1 + \delta\boldsymbol{v}_1 = \begin{bmatrix} \frac{\varepsilon^{1/4}}{\sqrt{2}}(1+\frac{5\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ -\frac{1}{\sqrt{2}}(1+\frac{3\sqrt{\varepsilon}}{2}) + O(\varepsilon) \\ \frac{1}{\sqrt{2}}(1-2\sqrt{\varepsilon}) + O(\varepsilon) \end{bmatrix}$$

and

$$\boldsymbol{v}_2 = \begin{bmatrix} -\frac{\varepsilon^{1/4}}{\sqrt{2}}(1+\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ \frac{1}{\sqrt{2}}(1-\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon) \\ \frac{1}{\sqrt{2}}(1+\frac{3\varepsilon}{4}) + O(\varepsilon^{3/2}) \end{bmatrix}, \quad \boldsymbol{v}_2 + \delta\boldsymbol{v}_2 = \begin{bmatrix} -\frac{\varepsilon^{1/4}}{\sqrt{2}}(1-\frac{3\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ \frac{1}{\sqrt{2}}(1-\frac{5\sqrt{\varepsilon}}{2}) + O(\varepsilon) \\ \frac{1}{\sqrt{2}}(1+2\sqrt{\varepsilon}) + O(\varepsilon) \end{bmatrix},$$

whereby

$$\left|\frac{\delta v_i(j)}{v_i(j)}\right| = O(\sqrt{\varepsilon}) \text{ for } i = 1, 2 \text{ and } j = 1, 2, 3.$$

Since a small relative change of ε in the off-diagonal entries of T_1 results in a much larger relative change in its eigenvalues and eigenvectors, we say that T_1 does not determine its eigenvalues and eigenvectors to high relative accuracy. Consequently, in the face of roundoff errors, it is unlikely that we can compute numerically orthogonal eigenvectors without explicit orthogonalization. To corroborate this, we gave the best possible approximations to λ_1 and λ_2 as input to the EISPACK and LAPACK implementations of inverse iteration but turned off all orthogonalization within these

[1] We carefully constructed this matrix to have the desired behavior, which may be verified by using a symbol manipulator such as Maple [5] or Mathematica [37].

procedures. As expected, we found the computed vectors to have dot products as large as $O(\sqrt{\varepsilon})$. □

In contrast, when T is positive definite, the representations LDL^t and $\tilde{L}\tilde{L}^t$, where $\tilde{L} = LD^{1/2}$, each determine *all* the eigenpairs to high relative accuracy. See [8, Theorem 5.13] for more details. Thus these factored forms are preferable to the standard form for eigenvalue/eigenvector calculations.

When D is not positive definite the situation is more complicated. Extensive testing shows that even in the face of element growth, LDL^t determines its small eigenpairs to high relative accuracy; see Example 2 in section 6.1. Of course we may also use the representation UDU^t derived from Gaussian elimination in reverse order or even a twisted factorization (see section 4). The important point is that the positive definite case is not the only one in which some eigenpairs are determined to high relative accuracy by a factored form.

Let $LDL^t\boldsymbol{v} = \boldsymbol{v}\lambda$, $\lambda \neq 0$. A relative condition number defined in [9] is

$$\boldsymbol{\kappa}_{rel}(\lambda) := \boldsymbol{v}^t L|D|L^t\boldsymbol{v}/|\lambda|.$$

In section 6.1 the true relative condition number covering all the relative perturbations in L and D is shown to be $1 + \boldsymbol{\kappa}_{rel}(\lambda)$. Note that when D is positive definite, then $\boldsymbol{\kappa}_{rel}(\lambda) = 1$, but we do not need such strong stability for our results. A value of $\boldsymbol{\kappa}_{rel}(\lambda)$ such as 10 or 100 is adequate.

The focus of this paper is on how to exploit high relative accuracy when it occurs, not to give conditions for its occurrence. See section 6.1 for some discussion on the latter.

4. Algorithm Getvec. In this section we present our procedure, Algorithm Getvec, for computing an eigenvector.

If $\hat{\lambda}$ is an accurate approximation to an eigenvalue λ of T, then $T - \hat{\lambda}I$ is almost singular. In order to compute the eigenvector, i.e., to solve $(T - \hat{\lambda}I)\boldsymbol{z} \approx 0$, we seek a factorization that reveals this singularity. As we show below, in the tridiagonal case we can easily construct such a "twisted" factorization from the forward and backward triangular factors [31].

Suppose that

$$LDL^t - \hat{\lambda}I = L_+D_+L_+^t = U_-D_-U_-^t,$$

where L_+ is unit lower bidiagonal and U_- is unit upper bidiagonal. The $L_+D_+L_+^t$ and $U_-D_-U_-^t$ factorizations may be obtained by Gaussian elimination in the forward and backward directions, respectively. Note that by the discussion in section 3, we have replaced T by LDL^t. It may happen that neither D_+ nor D_- reveals the singularity of $LDL^t - \hat{\lambda}I$. A twisted factorization, written as

$$LDL^t - \hat{\lambda}I = N_kD_kN_k^t,$$

may be constructed by factoring the matrix from top down and from bottom up meeting at row k. The twisted factor N_k takes rows $1:k$ of L_+ and rows $k:n$ of U_-. Thus all rows of N_k have two nonzeros, except row k, which has three nonzero entries

$$(L_+(k-1) \quad 1 \quad U_-(k)),$$

while D_k is diagonal,

$$D_k = \mathrm{diag}(D_+(1), \ldots, D_+(k-1), \gamma_k, D_-(k+1), \ldots, D_-(n)),$$

where we denote $L_+(i+1,i)$ by $L_+(i)$, $U_-(i,i+1)$ by $U_-(i)$, and the ith diagonal entries of D_+ and D_- by $D_+(i)$ and $D_-(i)$, respectively. We will continue to use this notation for the rest of the paper.

Clearly, there are n twisted factorizations, one for each $k = 1, \ldots, n$. One such twisted factor, with $n = 6$ and $k = 3$, is shown in Figure 1.

$$\begin{bmatrix} \mathrm{x} & & & & & \\ \mathrm{x} & \mathrm{x} & & & & \\ & \mathrm{x} & \mathrm{x} & \mathrm{x} & & \\ & & & \mathrm{x} & \mathrm{x} & \\ & & & & \mathrm{x} & \mathrm{x} \\ & & & & & \mathrm{x} \end{bmatrix}$$

FIG. 1. *Twisted triangular factor N_k with $n = 6$, $k = 3$.*

The only new entry is γ_k, the kth diagonal element of D_k, and it is of great importance. As we show in section 6.2, the nearness to singularity of $LDL^t - \hat{\lambda}I$ is revealed by a small value of $|\gamma_k|$ for an appropriate choice of k. There are several formulae for γ_k, for example,

$$\gamma_k = \begin{cases} D_+(k) + D_-(k) - (d_{k-1}l_{k-1}^2 + d_k - \hat{\lambda}), \\ D_+(k) - (d_k l_k)^2 / D_-(k+1), \end{cases} \tag{7}$$

where $d_k = D(k,k)$, $l_{k-1} = L(k,k-1)$, and so $(LDL^t)_{kk} = d_{k-1}l_{k-1}^2 + d_k$ and $(LDL^t)_{k,k+1} = d_k l_k$.

Naive ways of computing twisted factorizations will not satisfy our demands of high relative accuracy. The so-called "differential qd transforms" allow accurate computation of twisted factorizations, including more robust expressions for γ_k. We will, however, wait until section 4.1 to give details of the qd transforms. Without further ado, we present Algorithm Getvec, which computes an approximate eigenvector by first forming the appropriate twisted factorization. In the following we assume that LDL^t is an irreducible tridiagonal, i.e., $d_i l_i \neq 0$ for $1 \leq i \leq n-1$.

ALGORITHM Getvec(LDL^t, $\hat{\lambda}$).

I. Factor $LDL^t - \hat{\lambda}I = L_+D_+L_+^t$ by the *dstqds* transform (Algorithm 4.2 in section 4.1).

II. Factor $LDL^t - \hat{\lambda}I = U_-D_-U_-^t$ by the *dqds* transform (Algorithm 4.4 in section 4.1).

III. Compute γ_k for $k = 1, \ldots, n$ (by the top formula of (18)). Pick an r such that $|\gamma_r| = \min_k |\gamma_k|$. Then

$$N_r D_r N_r^t = LDL^t - \hat{\lambda}I$$

is the desired twisted factorization (see Algorithm 4.5 in section 4.1).

IV. Form the approximate eigenvector $\boldsymbol{z}$ by solving $N_r^t \boldsymbol{z} = \boldsymbol{e}_r$, where $\boldsymbol{e}_r$ is the rth column of the identity matrix, which is equivalent to solving

$$(LDL^t - \hat{\lambda}I)\boldsymbol{z} = N_r D_r N_r^t \boldsymbol{z} = \boldsymbol{e}_r \gamma_r \quad (\text{since } D_r \boldsymbol{e}_r = \boldsymbol{e}_r \gamma_r \text{ and } N_r \boldsymbol{e}_r = \boldsymbol{e}_r)$$

via

$$z(r) = 1,$$

$$\text{for } i = r-1, \ldots, 1, \quad z(i) = \begin{cases} -L_+(i)z(i+1), & D_+(i) \neq 0, \\ -(d_{i+1}l_{i+1}/d_i l_i)z(i+2), & \text{otherwise,} \end{cases}$$

$$\text{for } j = r, \ldots, n-1, \quad z(j+1) = \begin{cases} -U_-(j)z(j), & D_-(j+1) \neq 0, \\ -(d_{j-1}l_{j-1}/d_j l_j)z(j-1), & \text{otherwise.} \end{cases}$$

V. If wanted, compute $znrm = \|\boldsymbol{z}\|$ and set $\tilde{\boldsymbol{z}} = \boldsymbol{z}/znrm$.

Remark 1. Steps I–III above employ differential qd transforms that are essential in order to exploit the RRR properties of the bidiagonal representation LDL^t. The choice of r in step III ensures that $|\gamma_r| \leq 2n|\hat{\lambda} - \lambda|$ and the residual norm $\|(LDL^t - \hat{\lambda}I)\boldsymbol{z}\|/\|\boldsymbol{z}\| \leq \sqrt{n}|\hat{\lambda} - \lambda|$ under suitable conditions; see Theorems 10 and 11 in section 6.2.

Remark 2. No pivoting is done in steps I and II since the computation assumes IEEE arithmetic [1]. If some $D_+(i)$ (or $D_-(i)$) equals zero, then infinity is produced at the next step, and the computation of $\boldsymbol{z}$ in step IV handles this special case. See Remark 3 below.

Remark 3. Let us explain the special handling in step IV above of the case of a zero entry in D_+ or D_-. In exact arithmetic, when $\hat{\lambda}$ is an eigenvalue, zero entries in D_+ and D_- can occur if and only if the corresponding eigenvector has a zero entry. In particular, when $\hat{\lambda}$ is an eigenvalue, then $D_+(i) = 0$ and $D_-(i+2) = 0$ if and only if $z(i+1) = 0$. See [31] for more details. Thus, when $D_+(i) = 0$, $i < r$, we use the $(i+1)$st equation of the tridiagonal system $(LDL^t - \hat{\lambda}I)\boldsymbol{z} = \boldsymbol{e}_r\gamma_r$ to connect $z(i)$ with $z(i+2)$. The case when $D_-(j+1) = 0$, $j > r$, is handled similarly.

Remark 4. The index r is desired to be such that the rth component of the eigenvector $\boldsymbol{v}$ is largest in magnitude [31]. It is possible to avoid up to half of the $2n$ divisions in steps I and II by observing that $v(i)$ cannot be the largest in magnitude if the eigenvalue is not contained in the ith Gerschgorin disk. This observation enables us to identify the smallest and largest indices that are candidates for the twist index r. The savings are often real when n is large since eigenvectors of large matrices often have negligible entries at either end. See [9, section 3.4.1] for details.

Remark 5. In addition to computing an eigenvector approximation, the above algorithm can also be used to improve the accuracy of $\hat{\lambda}$. By Lemma 12 in section 6.2, $\gamma_r/\|\boldsymbol{z}\|^2$ is the Rayleigh quotient correction to $\hat{\lambda}$ and so it can double the number of correct digits when $\hat{\lambda}$ is not quite acceptable, for example, when $|\hat{\lambda} - \lambda| = O(\sqrt{\varepsilon}|\lambda|)$ where λ is the eigenvalue closest to $\hat{\lambda}$. Indeed, the refinement of eigenvalues in Algorithm MR^3 of [10] is done by switching from bisection to this Rayleigh quotient correction for increased efficiency.

Remark 6. The vector $\boldsymbol{z}$ often has small numerical support (defined below) when n is large. This situation can be detected when consecutive entries in $\boldsymbol{z}$ are small enough in magnitude. Then the remaining entries in $\boldsymbol{z}$ may be set to zero. Suppose all elements $z(j)$, $j < i-1 < r-1$, are set to zero; then equations $i-2$ and $i-1$ of $(LDL^t - \hat{\lambda}I)\boldsymbol{z} = \boldsymbol{e}_r\gamma_r$ are no longer satisfied and result in a residual $\beta_{i-2}(z(i-1)\boldsymbol{e}_{i-2} - z(i-2)\boldsymbol{e}_{i-1})$, where $\beta_{i-2} = D_+(i-2)L_+(i-2)$. For the vector $\boldsymbol{z}$ to be an accurate eigenvector (see Theorem 1), it suffices to ensure that $|z(i-1)|$ and $|z(i-2)|$ are small enough that

$$|D_+(i-2)L_+(i-2)|\,(|z(i-1)| + |z(i-2)|) \;<\; \varepsilon \cdot \mathrm{gap}(\hat{\lambda}),$$

where $z(i-2) = -L_+(i-2)z(i-1)$. Similarly when $i > r$ we set the elements $z(j)$, $j > i$, to 0 if $|z(i)|$ and $|z(i+1)|$ are small enough that

$$|D_-(i)U_-(i-1)|\,(|z(i)| + |z(i+1)|) \;<\; \varepsilon \cdot \mathrm{gap}(\hat{\lambda}),$$

where $z(i+1) = -U_-(i)z(i)$. Thus all our computed vectors have a first and last nonzero component and we call the index set {first:last} the *numerical support* of $\boldsymbol{z}$ and so

$$|\mathrm{supp}(\boldsymbol{z})| = \mathrm{last} - \mathrm{first} + 1.$$

Note that in exact arithmetic the first and last entries of an eigenvector of an unreduced tridiagonal matrix are nonzero but in practice they are often extremely small, and so the above situation is not uncommon.

There is more to be said about the support. Before $\boldsymbol{z}$ is computed all the $\{\gamma_i\}$ are computed in order to find the smallest among them. By Lemma 11 in [31], as $\hat{\lambda} \to \lambda$,

$$\frac{\gamma_r}{\gamma_i} \;\to\; \frac{v(i)^2}{v(r)^2}, \tag{8}$$

where $\boldsymbol{v}$ is λ's eigenvector. This suggests that if $|\gamma_i| > |\gamma_r|/\varepsilon^2$, then $z(i)$ may be neglected and it might be argued that this gives us a better way to approximate $\mathrm{supp}(\boldsymbol{z})$ at the time r is chosen. Unfortunately, machine precision is often not sufficient to put $\hat{\lambda}$ close enough to λ for (8) to hold for indices where $|v(i)| \ll \sqrt{\varepsilon}$, and so this strategy does not work in practice.

Remark 7. It is not essential that $|\gamma_r|$ be minimal. In principle one keeps a list of indices i such that $|\gamma_{min}| < |\gamma_i| < 2|\gamma_{min}|$, and can choose r to be any of these indices.

Remark 8. Suppose $\hat{\lambda}$ approximates λ. As will be shown in section 7, in the presence of roundoff errors, the best we can hope for is that the computed vector $\hat{\boldsymbol{z}}$ satisfies

$$|\sin\angle(\hat{\boldsymbol{z}}, \boldsymbol{v})| \;=\; O\left(\frac{|\lambda - \hat{\lambda}|}{\mathrm{gap}(\hat{\lambda})}\right) \;=\; O\left(\frac{n\varepsilon|\hat{\lambda}|}{\mathrm{gap}(\hat{\lambda})}\right) \;=\; O\left(\frac{n\varepsilon}{\mathrm{relgap}(\hat{\lambda})}\right).$$

Such a $\hat{\boldsymbol{z}}$ will be an accurate eigenvector when $\mathrm{relgap}(\hat{\lambda}) \geq tol$. A natural question to ask is: can such an accurate approximation be computed when the relative gap is smaller, say, $\mathrm{relgap}(\hat{\lambda}) = \sqrt{\varepsilon}$? A tempting solution is to extend Algorithm Getvec to do a step of inverse iteration: $(LDL^t - \hat{\lambda}I)\boldsymbol{y} = \hat{\boldsymbol{z}} \;\Rightarrow (LDL^t - \hat{\lambda}I)^2\boldsymbol{y} \approx \gamma_r \boldsymbol{e}_r$. The tempting argument is that by doing so,

$$|\sin\angle(\boldsymbol{y}, \boldsymbol{v})| \;=\; O\left(\frac{|\lambda - \hat{\lambda}|^2}{\mathrm{gap}(\hat{\lambda})^2}\right) \;=\; O\left(\frac{n^2\varepsilon^2}{\mathrm{relgap}(\hat{\lambda})^2}\right),$$

since the eigenvalues of $(LDL^t - \hat{\lambda}I)^2$ are just $(\lambda_i - \hat{\lambda})^2$. When $\mathrm{relgap}(\hat{\lambda}) = \sqrt{\varepsilon}$ and n is modest, this strategy appears to yield an accurate eigenvector $\boldsymbol{y}$.

Unfortunately this simple solution does not work. In our experience, even for small n the extra step of inverse iteration increases the accuracy by a factor of .1 or .01 but not by a factor of $\sqrt{\varepsilon}$ as the above reasoning indicates. The failure is due to the presence of roundoff errors and limitations due to relative perturbation theory.

The case of relgap$(\hat{\lambda}) \ll tol$ requires radically different strategies. One strategy is to take a new shift to improve the relative gaps and then invoke Algorithm Getvec. Small relative gaps are not the concern of this paper, but the interested reader may see [9, 10] for details. Very tight clusters of eigenvalues that are well-separated from the rest of the spectrum may also be handled by the overlapping submatrix ideas of [29].

4.1. Differential qd transforms. This section completes the description of Algorithm Getvec by presenting the differential qd transforms that are needed to compute the $L_+D_+L_+^t$, $U_-D_-U_-^t$ and $N_rD_rN_r^t$ decompositions in steps I–III of the algorithm. Algorithm 4.1 given below is a straightforward implementation of the transformation

$$LDL^t - \mu I \;=\; L_+D_+L_+^t. \tag{9}$$

We call this the "**st**ationary **q**uotient-**d**ifference with **s**hift"(stqds) transform for historical reasons. The term was first coined by Rutishauser for similar transformations that formed the basis of his qd algorithm first developed in 1954 [34, 35, 36]. Although (9) is not identical to the stationary transformation given by Rutishauser, the differences are not significant enough to warrant inventing new terminology. The term "stationary" is used for (9) since it represents an identity transformation when $\mu = 0$. Rutishauser used the term "progressive" instead for the formation of $U_-D_-U_-^t$ from $LDL^t - \mu I$ or of $L_+D_+L_+^t$ from $UDU^t - \mu I$.

ALGORITHM 4.1. **(stqds)**-stationary qd transform.

$D_+(1) \;:=\; d_1 - \mu$
for $i = 1, n-1$
$\quad L_+(i) \;:=\; (d_il_i)/D_+(i)$ (10)
$\quad D_+(i+1) \;:=\; d_il_i^2 + d_{i+1} - L_+(i)d_il_i - \mu$ (11)
end for

This algorithm loses accuracy when there is element growth. Next we show how to eliminate some of the additions and subtractions from Algorithm 4.1. We introduce the intermediate variable $s_i := D_+(i) - d_i$, $1 \le i \le n$. A two-term recurrence between s_i and s_{i+1}, $1 \le i \le n-1$, may be obtained as follows:

$$\begin{aligned} s_{i+1} &= D_+(i+1) - d_{i+1} \\ &= d_il_i^2 - L_+(i)d_il_i - \mu \quad \text{by (11)} \\ &= L_+(i)l_i(D_+(i) - d_i) - \mu \quad \text{by (10)} \\ &= L_+(i)l_is_i - \mu. \end{aligned} \tag{12}$$

Using this intermediate variable, we get the so-called *differential form* of the stationary qd transform (dstqds). This term was again coined by Rutishauser in the appendix of [36].

ALGORITHM 4.2. **(dstqds)**-*differential form* of the stationary qd transform.

$$
\begin{aligned}
& s_1 := -\mu \\
& \text{for } i = 1, n-1 \\
(13) \qquad & \quad D_+(i) := s_i + d_i \\
& \quad L_+(i) := (d_i l_i)/D_+(i) \\
& \quad s_{i+1} := L_+(i) l_i s_i - \mu \\
& \text{end for} \\
& D_+(n) := s_n + d_n
\end{aligned}
$$

In section 5 we will show that the differential transforms, in the face of roundoff errors, have attractive properties which play a crucial role in proving the main result of the paper, Theorem 15.

We also need to compute the transformation

$$LDL^t - \mu I = U_- D_- U_-^t,$$

which we call the "progressive **q**uotient-**d**ifference with **s**hift"(qds) transform. The following algorithm gives an obvious way to implement this transformation.

ALGORITHM 4.3. **(qds)**-progressive qd transform.

$$
\begin{aligned}
& D_-(n) := d_{n-1} l_{n-1}^2 + d_n - \mu \\
& \text{for } i = n-1, 1, -1 \\
(14) \qquad & \quad U_-(i) := (d_i l_i)/D_-(i+1) \\
(15) \qquad & \quad D_-(i) := d_{i-1} l_{i-1}^2 + d_i - (d_i l_i) U_-(i) - \mu \\
& \text{end for}
\end{aligned}
$$

Here we have adopted the convention that $d_0 = l_0 = 0$, which justifies (15) for $i = 1$. As in the stationary transformation, we introduce the intermediate variable $p_i := D_-(i) - d_{i-1} l_{i-1}^2$, $1 \le i \le n$. A two-term recurrence between p_i and p_{i+1}, $1 \le i \le n-1$, may be obtained as follows:

$$
\begin{aligned}
(16) \qquad p_i &= D_-(i) - d_{i-1} l_{i-1}^2 \\
&= d_i - U_-(i) d_i l_i - \mu \quad \text{by (15)} \\
&= \frac{d_i}{D_-(i+1)} (D_-(i+1) - d_i l_i^2) - \mu \quad \text{by (14)} \\
(17) \qquad &= \frac{d_i}{D_-(i+1)} \cdot p_{i+1} - \mu.
\end{aligned}
$$

Using this intermediate variable, we get the *differential form* of the progressive qd transform.

ALGORITHM 4.4. **(dqds)**-*differential form* of the progressive qd transform.

$$
\begin{aligned}
&p_n \;:=\; d_n - \mu \\
&\text{for } i = n-1, 1, -1 \\
&\quad D_-(i+1) \;:=\; d_i l_i^2 + p_{i+1} \\
&\quad t \;:=\; d_i / D_-(i+1) \\
&\quad U_-(i) \;:=\; l_i t \\
&\quad p_i \;:=\; p_{i+1} t - \mu \\
&\text{end for} \\
&D_-(1) \;:=\; p_1
\end{aligned}
$$

Note that we have denoted the intermediate variables by the symbols s_i and p_i to stand for *stationary* and *progressive*, respectively.

We also need to find all the γ_k's in order to choose the appropriate twisted factorization for computing the eigenvector. By (7),

$$
\begin{aligned}
\gamma_k &= D_+(k) - \frac{(d_k l_k)^2}{D_-(k+1)} \\
&= s_k + d_k - \frac{(d_k l_k)^2}{D_-(k+1)} \quad \text{by (13)} \\
&= s_k + \frac{d_k}{D_-(k+1)} \left(D_-(k+1) - d_k l_k^2 \right).
\end{aligned}
$$

Substituting from (16), (17), and (12) in the above equation, we can express γ_k by any of the following formulae:

$$
\gamma_k \;=\; \begin{cases} s_k + \frac{d_k}{D_-(k+1)} \cdot p_{k+1}, \\ s_k + p_k + \mu, \\ p_k + L_+(k-1) l_{k-1} s_{k-1}. \end{cases} \tag{18}
$$

In section 5, we will see that the top and bottom formulae in (18) are "better" in the presence of roundoff. When μ is close to an eigenvalue of LDL^t, the near-singularity of $LDL^t - \mu I$ can be revealed by choosing $r = \text{argmin}_k |\gamma_k|$.The twisted factorization at position r is given by

$$
LDL^t - \mu I \;=\; N_r D_r N_r^t,
$$

where $D_r = \text{diag}(D_+(1), \ldots, D_+(r-1), \gamma_r, D_-(r+1), \ldots, D_-(n))$, and N_r is the corresponding twisted factor that takes rows $1:r$ of L_+ and rows $r:n$ of U_- (see the beginning of section 4). It may be formed by the following "**d**ifferential **tw**isted **q**uotient-**d**ifference with **s**hift"(dtwqds) transform which is just the appropriate blend of Algorithms 4.2 and 4.4.

ALGORITHM 4.5. **(dtwqds)**-differential twisted qd transform.

$s_1 := -\mu$
for $i = 1, r-1$
  $D_+(i) := s_i + d_i$
  $L_+(i) := (d_i l_i)/D_+(i)$
  $s_{i+1} := L_+(i) l_i s_i - \mu$
end for
$p_n := d_n - \mu$
for $i = n-1, r, -1$
  $D_-(i+1) := d_i l_i^2 + p_{i+1}$
  $t := d_i/D_-(i+1)$
  $U_-(i) := l_i t$
  $p_i := p_{i+1} t - \mu$
end for
if $r < n$
  $$\gamma_r := s_r + \frac{d_r}{D_-(r+1)} \cdot p_{r+1}$$
else
  $\gamma_r := s_n + d_n$
end if

Note: In cases where we have already computed the stationary and progressive transformations, i.e., we have computed L_+, D_+, U_-, and D_-, the only additional work needed for dtwqds is one multiplication and one addition to compute γ_r.

We emphasize that *the particular qd transforms presented in this section are new.* Similar qd recurrences have been studied by Rutishauser [34, 35, 36]; Henrici [22], [23, Chapter 7]; Fernando and Parlett [13]; and Yao Yang [38].

4.2. Relation to previous work. Algorithm Getvec presented earlier is close in spirit to the one presented by Godunov and his co-workers in the USSR in 1985; see [18] and [17]. They formulated the idea of taking the top entries in the vector from one sequence and the bottom entries from another one and then choosing the right index at which to join the two pieces. Independently, Fernando discovered a similar idea in terms of running the well-known two-term recurrence for D_+, both forward from $D_+(1)$ and backward from $D_+(n) = 0$, and then joining the two sequences where they are closest. In [31], Parlett and Dhillon formulated and proved the double factorization theorem that gave a formula for computing γ_k, and showed the relationship of γ_k to the diagonal of the inverse. Further, [31] showed that at least one twisted factorization must reveal the size of the smallest eigenvalue thus yielding an accurate eigenvector (see Theorem 11 in section 6.2).

However, neither Godunov nor Fernando reaps the full reward for choosing the best place to join two pieces.

The reasons are quite different in the two cases. Godunov et al. carefully select approximate eigenvalues on opposite sides of the true eigenvalue for the two sequences that provide the eigenvector entries. However, they need *directed* rounding in order to establish their bounds in finite precision arithmetic. Directed rounding is available

in most modern computer hardware since it is part of the IEEE floating point standard [1]; however, the only programming language that makes it available in 2002 is C99. It is not yet implemented in Fortran 2000. Fernando does not consider the effects of roundoff error but, as with Godunov et. al., computes the two factorizations from a translate of the original matrix T that may not define its eigenvalues to high relative accuracy. The 3×3 example in section 3 illustrates the problem: the algorithm given by Fernando in section 5 of [14], even with highly accurate eigenvalue approximations, can yield eigenvectors with error exceeding $\sqrt{\varepsilon}$.

Thus we use the LDL^t representation instead of the diagonal and off-diagonal elements of T. Even use of a good representation is not enough to ensure that the residual norm $\|(LDL^t - \hat{\lambda}I)\boldsymbol{z}\| = O(\varepsilon|\lambda - \hat{\lambda}|)$ for the computed $\boldsymbol{z}$. For example, if Rutishauser's stationary qd algorithm (stqds) were used to compute L_+ and D_+ satisfying $LDL^t - \hat{\lambda}I = L_+D_+L_+^t$ we could not prove our main result, Theorem 15 in section 7. That result requires a second innovation, beyond the use of LDL^t, namely use of the *differential* qd algorithms introduced in section 4.1 to compute the entries of the twisted factors. The crucial relative mixed error analyses as will be shown by the commutative diagrams in section 5 are not valid for Rutishauser's implementation. Hence the LDL^t representation and differential qd transforms are both crucial to our goal of computing orthogonal eigenvectors when relative gaps are large.

5. Roundoff error analysis. In this section, we exhibit desirable properties of the differential qd transforms of section 4.1 in the face of roundoff errors. The error analysis that follows is somewhat daunting and a trustful reader may wish to skip the proofs. However, the very special "interpretation" of the roundoff errors is the rock on which our main result, Theorem 15, is built.

First, we introduce our model of arithmetic. We assume that the floating point result of a basic arithmetic operation $\circ$ satisfies

$$fl(x \circ y) = (x \circ y)(1 + \eta) = (x \circ y)/(1 + \delta),$$

where η and δ depend on x, y, $\circ$, and the arithmetic unit but satisfy

$$|\eta| < \varepsilon, \quad |\delta| < \varepsilon$$

for a given ε that depends only on the arithmetic unit. We shall choose freely the form (η or δ) that suits the analysis. As usual, we will ignore $O(\varepsilon^2)$ terms in our analyses. We also adopt the convention of denoting the computed value of x by $\hat{x}$.

Ideally, we would like to show that the differential qd transforms introduced in section 4.1 produce an output that is exact for data that is very close to the input matrix. Since we desire relative accuracy, we would like this backward error to be relative. However, our algorithms do not admit such a pure backward analysis (see [38] for a backward analysis where the backward errors are absolute but not relative). Nevertheless, we will give a hybrid interpretation involving both backward and forward relative errors.

The best way to understand our first result is by studying Figure 2. Following Rutishauser, we merge elements of L and D into a single array,

$$Z := \{d_1, l_1, d_2, l_2, \ldots, d_{n-1}, l_{n-1}, d_n\}.$$

Likewise, the array $\vec{Z}$ is made up of elements $\vec{d_i}$ and $\vec{l_i}$; $\hat{Z}_+$ contains elements $\hat{D}_+(i)$, $\hat{L}_+(i)$, and so on. The acronym *ulp* in Figure 2 stands for *u*nits in the *l*ast *p*lace held. It is the natural way to refer to *relative* differences between numbers. When a result is correctly rounded the error is not more than half an *ulp*.

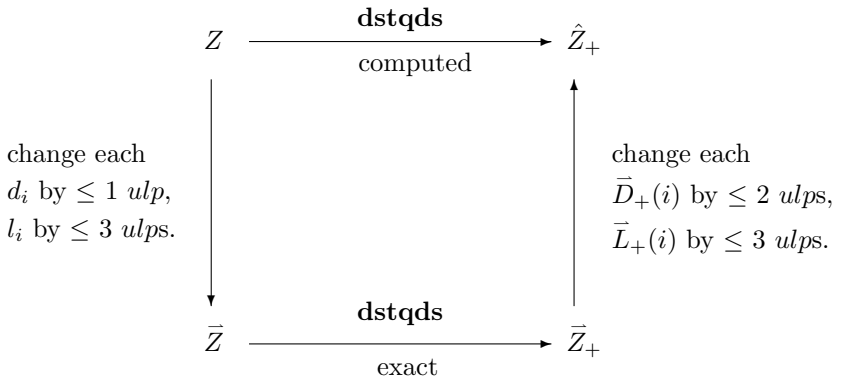


FIG. 2. *Effects of roundoff*—**dstqds** *transform.*

Notational Guide. In all results of this section, numbers in the computer are represented either by letters without any overbar, such as Z, or by "hatted" symbols, such as $\hat{Z}_+$. For example, in Figure 2, Z represents the input data, while $\hat{Z}_+$ represents the output data obtained by executing the dstqds algorithm in finite precision. Intermediate arrays, such as $\vec{Z}$ and $\vec{Z}_+$, are introduced for our analysis but are typically unrepresentable in a computer's limited precision. Note that we have chosen the symbol $\rightharpoonup$ in Figure 2 to indicate a process that takes rows and columns of a tridiagonal in increasing order, i.e., from "left to right." Later, in Figure 3 we use $\leftharpoonup$ to indicate a "right to left" process.

Figure 2 states that the computed outputs of the dstqds transform (Algorithm 4.2), $\hat{D}_+(i)$ and $\hat{L}_+(i)$, are small relative perturbations of the quantities $\vec{D}_+(i)$ and $\vec{L}_+(i)$ which in turn are the results of an EXACT dstqds transform applied to the perturbed matrix represented by $\vec{Z}$. The elements of $\vec{Z}$ are obtained by small relative changes in the inputs L and D. Analogous results hold for the dqds and dtwqds transforms (Algorithms 4.4 and 4.5). As we mentioned above, this is not a pure backward error analysis. We have put small perturbations not only on the input but also on the output in order to obtain an exact dstqds transform. This property is called mixed stability in [4, 6] and numerical stability in [24] but note that our perturbations are relative, not absolute.

THEOREM 2. *Let the dstqds transform be computed as in Algorithm* 4.2. *In the absence of overflow and underflow, the diagram in Figure* 2 *commutes and* $\vec{d}_i$ $(\vec{l}_i)$ *differs from* d_i (l_i) *by at most* 1 (3) *ulps, while* $\hat{D}_+(i)$ $(\hat{L}_+(i))$ *differs from* $\vec{D}_+(i)$ $(\vec{L}_+(i))$ *by at most* 2 (3) *ulps.*

Proof. We write the exact equations satisfied by the computed quantities:

$$\hat{D}_+(i) = (\hat{s}_i + d_i)/(1+\varepsilon_+),$$

$$\hat{L}_+(i) = d_i\, l_i(1+\varepsilon_*)(1+\varepsilon_/)/\hat{D}_+(i) \;=\; \frac{d_i\, l_i(1+\varepsilon_*)(1+\varepsilon_/)(1+\varepsilon_+)}{\hat{s}_i + d_i},$$

$$\text{and} \quad \hat{s}_{i+1} = \frac{\hat{L}_+(i)\, l_i \hat{s}_i(1+\varepsilon_\circ)(1+\varepsilon_{**}) - \mu}{1+\varepsilon_{i+1}}.$$

In the above, all ε's depend on i but we have chosen to single out the one that accounts

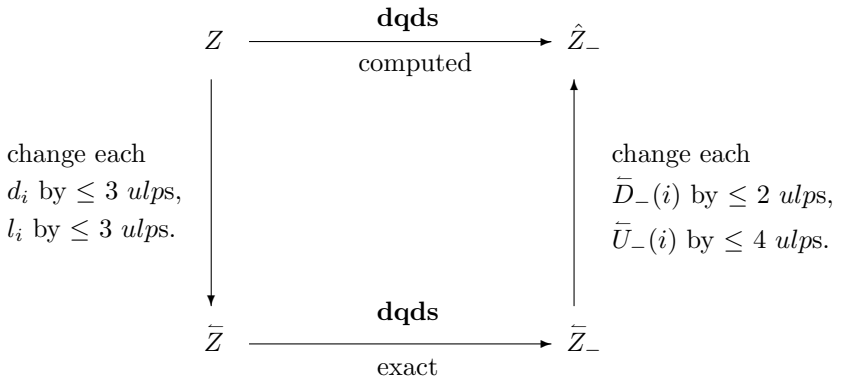


FIG. 3. *Effects of roundoff*—**dqds** *transform.*

for the subtraction as it is the only one where the dependence on i must be made explicit. In more detail the last relation is

$$(1+\varepsilon_{i+1})\hat{s}_{i+1} = \frac{d_i\, l_i^2\, \hat{s}_i}{\hat{s}_i + d_i}\,(1+\varepsilon_*)(1+\varepsilon_{/})(1+\varepsilon_+)(1+\varepsilon_\circ)(1+\varepsilon_{**}) - \mu.$$

The trick is to define $\bar{d}_i$ and $\bar{l}_i$ so that the exact dstqds relation

$$\bar{s}_{i+1} = \frac{\bar{d}_i\, \bar{l}_i^2\, \bar{s}_i}{\bar{s}_i + \bar{d}_i} - \mu \tag{19}$$

is satisfied. This may be achieved by setting

$$\begin{aligned} \bar{d}_i &= d_i(1+\varepsilon_i), \\ \bar{s}_i &= \hat{s}_i(1+\varepsilon_i), \\ \bar{l}_i &= l_i \sqrt{\frac{(1+\varepsilon_*)(1+\varepsilon_{/})(1+\varepsilon_+)(1+\varepsilon_\circ)(1+\varepsilon_{**})}{1+\varepsilon_i}}. \end{aligned} \tag{20}$$

In order to satisfy the exact mathematical relations of dstqds,

$$\bar{D}_+(i) = \bar{s}_i + \bar{d}_i, \tag{21}$$

$$\bar{L}_+(i) = \frac{\bar{d}_i\, \bar{l}_i}{\bar{s}_i + \bar{d}_i}, \tag{22}$$

we set

$$\begin{aligned} \bar{D}_+(i) &= \hat{D}_+(i)(1+\varepsilon_+)(1+\varepsilon_i), \\ \bar{L}_+(i) &= \hat{L}_+(i)\sqrt{\frac{(1+\varepsilon_\circ)(1+\varepsilon_{**})}{(1+\varepsilon_*)(1+\varepsilon_{/})(1+\varepsilon_+)(1+\varepsilon_i)}}, \end{aligned} \tag{23}$$

and the result holds. □

A similar result holds for the dqds transform.

THEOREM 3. *Let the dqds transform be computed as in Algorithm* 4.4. *In the absence of overflow and underflow, the diagram in Figure* 3 *commutes and* $\bar{d}_i$ $(\bar{l}_i)$

differs from d_i (l_i) *by at most* 3 (3) *ulps, while* $\hat{D}_-(i)$ ($\hat{U}_-(i)$) *differs from* $\bar{\bar{D}}_-(i)$ ($\bar{\bar{U}}_-(i)$) *by at most* 2 (4) *ulps.*

Proof. The proof is similar to that of Theorem 2. The computed quantities satisfy

$$\begin{aligned}
\hat{D}_-(i+1) &= (d_i\, l_i^2(1+\varepsilon_*)(1+\varepsilon_{**}) + \hat{p}_{i+1})/(1+\varepsilon_+), \\
\hat{t} &= d_i(1+\varepsilon_/)/\hat{D}_-(i+1), \\
\hat{U}_-(i) &= l_i\hat{t}(1+\varepsilon_\circ) \;=\; \frac{d_i\, l_i(1+\varepsilon_/)(1+\varepsilon_\circ)(1+\varepsilon_+)}{d_i\, l_i^2(1+\varepsilon_*)(1+\varepsilon_{**}) + \hat{p}_{i+1}}, \\
\hat{p}_i &= \frac{(d_i/\hat{D}_-(i+1))\hat{p}_{i+1}(1+\varepsilon_/)(1+\varepsilon_{\circ\circ}) - \mu}{1+\varepsilon_i}, \\
\Rightarrow \quad (1+\varepsilon_i)\hat{p}_i &= \frac{d_i\,\hat{p}_{i+1}}{d_i\, l_i^2(1+\varepsilon_*)(1+\varepsilon_{**}) + \hat{p}_{i+1}}(1+\varepsilon_/)(1+\varepsilon_{\circ\circ})(1+\varepsilon_+) - \mu.
\end{aligned} \tag{24}$$

Note that the above ε's are different from the ones in the proof of the earlier Theorem 2. As in Theorem 2, the trick is to satisfy the exact relation,

$$\bar{\bar{p}}_i = \frac{\bar{\bar{d}}_i\,\bar{\bar{p}}_{i+1}}{\bar{\bar{d}}_i\,\bar{\bar{l}}_i^2 + \bar{\bar{p}}_{i+1}} \;-\; \mu, \tag{25}$$

which is achieved by setting

$$\bar{\bar{d}}_i = d_i(1+\varepsilon_/)(1+\varepsilon_{\circ\circ})(1+\varepsilon_+), \tag{26}$$

$$\bar{\bar{p}}_i = \hat{p}_i(1+\varepsilon_i),$$

$$\text{and} \quad \bar{\bar{l}}_i = l_i\sqrt{\frac{(1+\varepsilon_*)(1+\varepsilon_{**})(1+\varepsilon_{i+1})}{(1+\varepsilon_/)(1+\varepsilon_{\circ\circ})(1+\varepsilon_+)}}, \tag{27}$$

$$\text{so that} \quad \bar{\bar{d}}_i\,\bar{\bar{l}}_i^2 = d_i\, l_i^2(1+\varepsilon_*)(1+\varepsilon_{**})(1+\varepsilon_{i+1}).$$

The other dqds relations,

$$\bar{\bar{D}}_-(i+1) = \bar{\bar{d}}_i\,\bar{\bar{l}}_i^2 \;+\; \bar{\bar{p}}_{i+1}, \tag{28}$$

$$\bar{\bar{U}}_-(i) = \frac{\bar{\bar{d}}_i\,\bar{\bar{l}}_i}{\bar{\bar{d}}_i\,\bar{\bar{l}}_i^2 + \bar{\bar{p}}_{i+1}}, \tag{29}$$

may be satisfied by setting

$$\begin{aligned}
\bar{\bar{D}}_-(i+1) &= \hat{D}_-(i+1)(1+\varepsilon_+)(1+\varepsilon_{i+1}), \\
\bar{\bar{U}}_-(i) &= \frac{\hat{U}_-(i)}{1+\varepsilon_\circ}\sqrt{\frac{(1+\varepsilon_*)(1+\varepsilon_{**})(1+\varepsilon_{\circ\circ})}{(1+\varepsilon_/)(1+\varepsilon_+)(1+\varepsilon_{i+1})}}.
\end{aligned} \tag{30}$$

∎

By combining parts of the analyses for the dstqds and dqds transforms, we can also exhibit a similar result for the twisted factorization computed by Algorithm 4.5. In Figure 4, the various Z arrays represent corresponding twisted factors that may be obtained by "concatenating" the stationary and progressive factors. In particular, for any twist position k,

$$\begin{aligned}
\hat{Z}_k &:= \{\hat{D}_+(1), \hat{L}_+(1), \ldots, \hat{L}_+(k-1), \hat{\gamma}_k, \hat{U}_-(k), \hat{D}_-(k+1), \ldots, \hat{U}_-(n-1), \hat{D}_-(n)\}, \\
\bar{Z}_k &:= \{\vec{D}_+(1), \vec{L}_+(1), \ldots, \vec{L}_+(k-1), \bar{\gamma}_k, \bar{U}_-(k), \bar{\bar{D}}_-(k+1), \ldots, \bar{\bar{U}}_-(n-1), \bar{\bar{D}}_-(n)\},
\end{aligned}$$

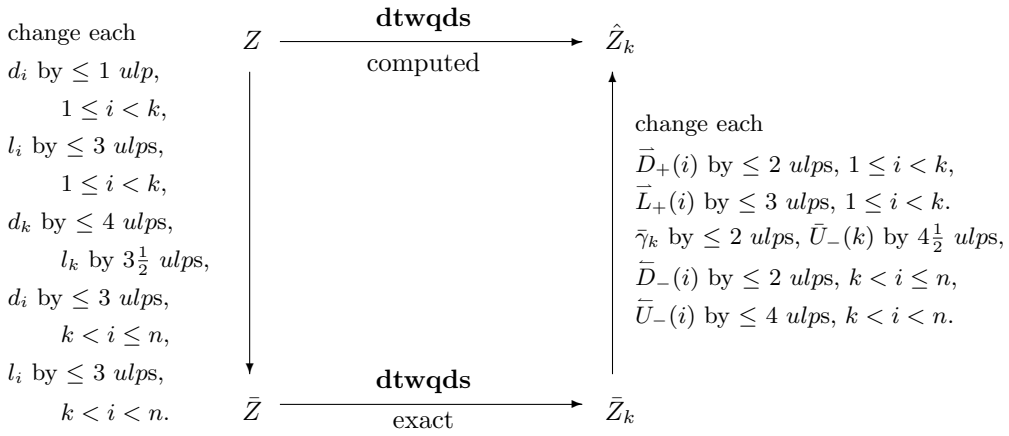


FIG. 4. *Effects of roundoff*—**dtwqds** *transform.*

while

$$\bar{Z} := \{\vec{d}_1, \vec{l}_1, \ldots, \vec{l}_{k-1}, \bar{d}_k, \bar{l}_k, \overleftarrow{d}_{k+1}, \ldots, \overleftarrow{l}_{n-1}, \overleftarrow{d}_n\}.$$

$\hat{Z}_k$ and $\bar{Z}_k$ represent the twisted factorizations

$$\hat{N}_k \hat{D}_k \hat{N}_k^t \qquad \text{and} \qquad \bar{N}_k \bar{D}_k \bar{N}_k^t,$$

respectively.

THEOREM 4. *Let the dtwqds transform be computed as in Algorithm* 4.5. *In the absence of overflow and underflow, the diagram in Figure* 4 *commutes.*

Proof. The crucial observation is that for the exact stationary transform (i.e., (19), (21), and (22)) to be satisfied for $1 \leq i \leq k-1$, roundoff errors need to be put only on $d_1, d_2, \ldots, d_{k-1}$ and $l_1, l_2, \ldots, l_{k-1}$. Similarly for the progressive transform (i.e., (25), (28) and (29)) to hold for $k+1 \leq i < n$, roundoff errors need to be put only on the bottom part of the matrix, i.e., on $d_{k+1}, \ldots, d_n$ and $l_{k+1}, \ldots, l_{n-1}$.

Next we turn to the entries associated with the twist position k. By the top formula in (18),

$$\hat{\gamma}_k = \left(\hat{s}_k + \frac{d_k}{\hat{D}_-(k+1)} \hat{p}_{k+1}(1+\varepsilon_{/}^-)(1+\varepsilon_{\circ\circ}^-) \right) \Big/ (1+\varepsilon_k).$$

Note that in the above, we have put the superscript "$-$" on some ε's to indicate that they are identical to the corresponding ε's in the proof of Theorem 3. By (20) and (24),

$$(1+\varepsilon_k)\hat{\gamma}_k = \frac{\vec{s}_k}{1+\varepsilon_k^+} + \frac{\hat{p}_{k+1} \cdot d_k (1+\varepsilon_{/}^-)(1+\varepsilon_{\circ\circ}^-)(1+\varepsilon_{+}^-)}{d_k\, l_k^2 (1+\varepsilon_{*}^-)(1+\varepsilon_{**}^-) + \hat{p}_{k+1}}$$

$$\Rightarrow \quad (1+\varepsilon_k)(1+\varepsilon_k^+)\hat{\gamma}_k = \vec{s}_k + \frac{\hat{p}_{k+1}(1+\varepsilon_{k+1}^-) \cdot d_k (1+\varepsilon_{/}^-)(1+\varepsilon_{\circ\circ}^-)(1+\varepsilon_{+}^-)(1+\varepsilon_k^+)}{d_k\, l_k^2 (1+\varepsilon_{*}^-)(1+\varepsilon_{**}^-)(1+\varepsilon_{k+1}^-) + \hat{p}_{k+1}(1+\varepsilon_{k+1}^-)},$$

where the superscript "+" indicates that the corresponding ε's are identical to those in the proof of Theorem 2. Note that we are free to attribute roundoff errors to d_k and l_k in order to preserve exact mathematical relations at the twist position k. In particular, by setting

$$\begin{aligned}
\bar{\gamma}_k &= \hat{\gamma}_k(1+\varepsilon_k)(1+\varepsilon_k^+),\\
\bar{d}_k &= d_k(1+\varepsilon_{/}^-)(1+\varepsilon_{\circ\circ}^-)(1+\varepsilon_{+}^-)(1+\varepsilon_k^+),\\
\bar{l}_k &= l_k\sqrt{\frac{(1+\varepsilon_{*}^-)(1+\varepsilon_{**}^-)(1+\varepsilon_{k+1}^-)}{(1+\varepsilon_{/}^-)(1+\varepsilon_{\circ\circ}^-)(1+\varepsilon_{+}^-)(1+\varepsilon_k^+)}}
\end{aligned}$$

and recalling that $\bar{\bar{p}}_{k+1} = \hat{p}_{k+1}(1+\varepsilon_{k+1}^-)$ (see (26)), the following exact relation holds:

$$\bar{\gamma}_k = \bar{\bar{s}}_k + \frac{\bar{d}_k\,\bar{\bar{p}}_{k+1}}{\bar{\bar{d}}_k\,\bar{l}_k^2 + \bar{\bar{p}}_{k+1}}.$$

In addition, the exact relation

$$\bar{U}_-(k) = \frac{\bar{d}_k\,\bar{l}_k}{\bar{\bar{d}}_k\,\bar{l}_k^2 + \bar{\bar{p}}_{k+1}}$$

holds if we set

$$\bar{U}_-(k) = \frac{\hat{U}_-(k)}{1+\varepsilon_{\circ}^-}\sqrt{\frac{(1+\varepsilon_{*}^-)(1+\varepsilon_{**}^-)(1+\varepsilon_{\circ\circ}^-)(1+\varepsilon_k^+)}{(1+\varepsilon_{/}^-)(1+\varepsilon_{k+1}^-)(1+\varepsilon_{+}^-)}}, \tag{31}$$

where $\varepsilon_{\circ}^-$ is identical to the $\varepsilon_{\circ}$ of (30). Note that since $\bar{d}_k\,\bar{l}_k^2 = \bar{\bar{d}}_k\,\bar{\bar{l}}_k^2$, the $(k+1)$st diagonal element in $\bar{Z}_k$ remains $\bar{\bar{D}}_-(k+1)$ as

$$\bar{d}_k\,\bar{l}_k^2 + \bar{\bar{p}}_{k+1} \;=\; \bar{\bar{d}}_k\,\bar{\bar{l}}_k^2 + \bar{\bar{p}}_{k+1} \;=\; \bar{\bar{D}}_-(k+1) \qquad \text{from (28).} \qquad \square$$

Note: A similar result may be obtained if γ_k is computed by the last formula in (18).

6. Analysis of the commutative diagram. The roundoff error analysis of the previous section shows that the commutative diagram of Figure 4 holds, with $k = r$, for Algorithm Getvec's computation, which forms the twisted factorization $N_r D_r N_r^t$ and then computes an approximate eigenvector. Figure 5 lays out the essentials given in Figure 4 and shows that the computed vector $\hat{\boldsymbol{z}}$ can be connected to the eigenvector $\boldsymbol{v}$ in three steps: (i) the right side relates $\hat{\boldsymbol{z}}$ to a vector $\bar{\boldsymbol{z}}$, (ii) the bottom arrow connects $\bar{\boldsymbol{z}}$ to an eigenvector $\bar{\boldsymbol{v}}$, and (iii) the left side relates $\bar{\boldsymbol{v}}$ to the desired eigenvector $\boldsymbol{v}$. In the rest of this section, we analyze each of these relationships in detail, before bringing it all together in section 7.

6.1. The left side—relative perturbation theory. The left side of Figure 5 examines the closeness of the eigenvector $\boldsymbol{v}$ to $\bar{\boldsymbol{v}}$ when small relative changes are made to the nontrivial entries of L and D. When LDL^t is positive (or negative) definite, it is well known that it determines its eigenvalues and eigenvectors to high relative accuracy [7], i.e., LDL^t is an RRR; see (5) and (6). However, in many cases, an indefinite LDL^t factorization also determines its eigenpairs to high relative accuracy. This section discusses conditions under which this can happen.

In the following analysis LDL^t should be thought of as the most familiar of the n twisted factorizations and the results below extend, with small modifications, to any twisted factorization.

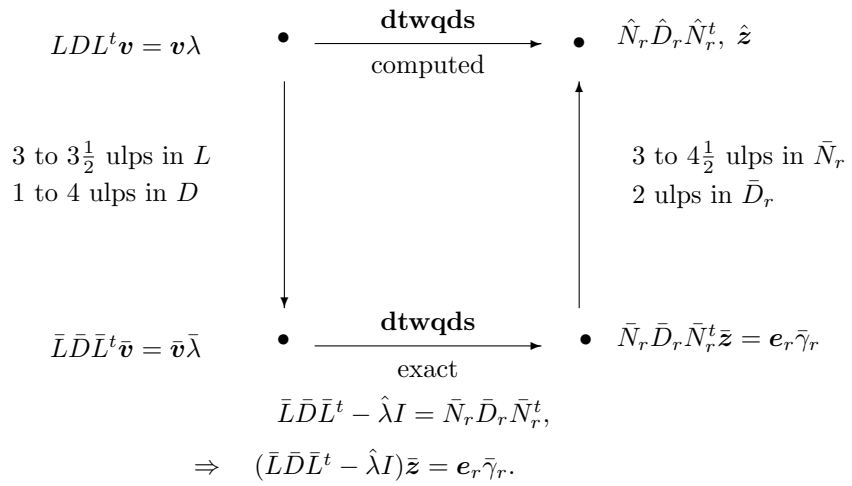


FIG. 5. *Relationships connecting* $\boldsymbol{v}$ *to* $\hat{\boldsymbol{z}}$.

6.1.1. Multiplicative form. For the sake of completeness, we present the following well-known lemma; see [8, Lemma 5.7] and its proof.

LEMMA 5. *Let L be a unit bidiagonal matrix with no zero off-diagonal entries. Independent relative perturbations in the off-diagonals may be represented by the two-sided scaling*

$$E^{-1}LE,$$

where $E = \mathrm{diag}(e_1, \ldots, e_n)$ is a diagonal scaling matrix unique to within a constant multiple, and independent of L.

Proof. Let $L_{ij}\alpha_{ij}$ represent the perturbation of L_{ij}. The equations to be solved are

$$\frac{L_{i+1,i}e_i}{e_{i+1}} = L_{i+1,i}\alpha_{i+1,i}, \qquad 1 \le i < n.$$

Letting $e_n = 1$ we get $e_{n-1} = \alpha_{n,n-1}$. Decreasing the index i further, we get

$$e_i = e_{i+1} \cdot \alpha_{i+1,i} = \prod_{j=i}^{n-1} \alpha_{j+1,j}, \qquad i = n-1, n-2, \ldots, 1. \qquad \square$$

Independent relative perturbations to nonzero entries of D are directly represented by a diagonal scaling matrix that we choose to write as F^2. Thus independent relative perturbations to the nontrivial entries of L and D lead to the perturbed matrix

$$\bar{T} = \bar{L}\bar{D}\bar{L}^t = (E^{-1}LE)(FDF)(EL^tE^{-1}). \tag{32}$$

LEMMA 6. *Let Algorithm dtwqds be executed in finite precision arithmetic. The matrices E and F that account for changes in the input L and D in order for the commutative diagram of Figure 4 to hold (see Theorem 10) satisfy*

$$(1-\varepsilon)^{6n-1} < \|(EF)^2\| < (1+\varepsilon)^{6n-1},$$

where ε is the machine precision.

Proof. Let $\bar{l}_i = l_i(1+\eta_i)$ and $\bar{d}_i = d_i(1+\delta_i)$. With twist index k, the bounds on η_i and δ_i satisfy

$$\begin{aligned}
i<k &: (1-\varepsilon)^3 < 1+\eta_i < (1+\varepsilon)^3, && 1-\varepsilon < 1+\delta_i < 1+\varepsilon,\\
i=k &: (1-\varepsilon)^{7/2} < 1+\eta_i < (1+\varepsilon)^{7/2}, && (1-\varepsilon)^4 < 1+\delta_i < (1+\varepsilon)^4,\\
i>k &: (1-\varepsilon)^3 < 1+\eta_i < (1+\varepsilon)^3, && (1-\varepsilon)^3 < 1+\delta_i < (1+\varepsilon)^3.
\end{aligned}$$

From Lemma 5 the bound on $\|E\| = \max_i |e_i|$ is maximized for e_1, the product of all the independent perturbations, which is upper bounded by $(1+\varepsilon)^{3(n-1)+1/2}$. In contrast $d_i \longrightarrow d_i f_i^2$, $i = 1, \ldots, n$, and $f_1^2 < 1+\varepsilon$ unless the twist is at $k=1$ when $f_1^2 < (1+\varepsilon)^4$. Thus

$$\begin{aligned}
\|(EF)^2\| &= \max_i (e_i f_i)^2 < (1+\varepsilon)^{6(n-1)+1}(1+\varepsilon)^4 = (1+\varepsilon)^{6n-1},\\
\|(EF)^2\| &= \max_i (e_i f_i)^2 > (1-\varepsilon)^{6(n-1)+1}(1-\varepsilon)^4 = (1-\varepsilon)^{6n-1}. \qquad \square
\end{aligned}$$

Let $(\lambda, \boldsymbol{v})$ be an eigenpair of LDL^t, $\lambda \neq 0$, $\|\boldsymbol{v}\| = 1$. We may write $\bar{T}$ in (32) in standard multiplicative form, i.e., with "outer" perturbations only, as

$$\begin{aligned}
\bar{T} &= G^t T G = G^t L D L^t G,\\
\text{where} \quad G &:= L^{-t} F E L^t E^{-1}
\end{aligned} \tag{33}$$

is an upper triangular matrix *sometimes* close to I. There is an eigenpair $(\bar{\lambda}, \bar{\boldsymbol{v}})$ of $\bar{T}$ associated with $(\lambda, \boldsymbol{v})$ and we want to investigate the closeness of $\bar{\lambda}$ to λ and $\bar{\boldsymbol{v}}$ to $\boldsymbol{v}$.

All the published results in relative perturbation theory known to us (see [11, 27, 2]) consider the form (33) above and do not require bidiagonal form for L. The perturbation bounds depend on $\|G^t - G^{-1}\|$ and/or $\|G^t G - I\|$; when these quantities are small, LDL^t can be shown to be an RRR for all the eigenvalues [12, 11, 27]. Yet G depends on L and will be far from orthogonal when L is ill-conditioned for inversion. However, in our experiments, we have often encountered situations where LDL^t is indefinite, L is ill-conditioned and the small eigenvalues in the interior of the spectrum are relatively robust while some of the larger ones are very sensitive.

So the desired bounds must not be uniform over the eigenvalues. In the work we have examined (see [2]), the bounds either are uniform or do not treat eigenvectors or are not computable. The value of the representation (32), along with Lemma 6 above, is that E and F are independent of L in the bidiagonal case. The price we pay for (32) is the presence of the "inner" scalings EF that bring us to new territory. There is a way to turn this inner scaling into a standard congruence and it was introduced in the earliest papers on computing singular values of a matrix C. Thus $C \longrightarrow XCY^t$ corresponds to

$$\begin{pmatrix} O & C \\ C^t & O \end{pmatrix} \longrightarrow \begin{pmatrix} X & O \\ O & Y \end{pmatrix} \begin{pmatrix} O & C \\ C^t & O \end{pmatrix} \begin{pmatrix} X^t & O \\ O & Y^t \end{pmatrix},$$

and the eigenvalues of the double matrix are the singular values of C and their negations, while the eigenvectors contain the right and left singular vectors. All the extensive perturbation theory for symmetric matrices has been brought to bear on the double matrix [11].

Our case $T = LDL^t$ is more difficult. In order to follow the approach indicated above, let $\Omega := \text{sign}(D) = \text{diag}(\pm 1)$ and observe that (32) corresponds to

(34)
$$\begin{pmatrix} O & L|D|^{1/2} \\ \Omega|D|^{1/2}L^t & O \end{pmatrix} \longrightarrow \begin{pmatrix} E^{-1} & O \\ O & EF \end{pmatrix} \begin{pmatrix} O & L|D|^{1/2} \\ \Omega|D|^{1/2}L^t & O \end{pmatrix} \begin{pmatrix} E^{-1} & O \\ O & EF \end{pmatrix},$$

where we have used the commutativity $EF\Omega = \Omega EF$. When $\Omega \neq I$, our double matrix is not normal and its eigenvalues are the square roots of those of LDL^t together with their negations. So the spectrum of the double matrix lies on both the real and the imaginary axes.

The first order perturbation analyses in [32, 30] foreshadow the upcoming results in section 6.1.2 and give realistic (relative) condition numbers that discriminate among the eigenpairs. Nevertheless those first order expressions do not yield bounds; the higher order terms are not controlled. One of us has developed bounds [33] on the change in both λ and $\boldsymbol{v}$, under mild conditions, and these bounds are close in form to the first order perturbation results in [32, 30], and are close to Demmel and Kahan's results in [7] when T is positive definite. In the next section we adapt these bounds to our situation. Of particular interest is the bound on the change in an eigenvector $\boldsymbol{v}$.

6.1.2. Perturbation bounds. We present here the quantities that govern the sensitivity of λ and $\boldsymbol{v}$ to the special perturbations $L \longrightarrow E^{-1}LE, D \longrightarrow FDF$ as given in section 6.1.1. In [9] Dhillon used first order perturbation theory to introduce a relative condition number corresponding to (only) the inner perturbations EF in (32),

$$\boldsymbol{\kappa}_{rel}(\lambda) := \frac{\boldsymbol{v}^t L|D|L^t\boldsymbol{v}}{|\lambda|} = \frac{\boldsymbol{v}^t L|D|L^t\boldsymbol{v}}{|\boldsymbol{v}^t LDL^t\boldsymbol{v}|},$$

and it plays the dominant role in λ's sensitivity to inner and outer perturbations.

This section gives bounds that account for both inner and outer perturbations and are derived in [33] using the double matrix form of (34); however, we quote results from [33] without giving the derivations, as they are much too long to be included here. In (34), when $\Omega = I$ the spectral decomposition of the symmetric double matrix is intimately related to the SVD of $L|D|^{1/2}$. For general Ω, we need to introduce the hyperbolic SVD (HSVD) of a matrix.

In the definitions that follow, K denotes a general real square matrix, and unlike $L|D|^{1/2}$, K need not be bidiagonal. Given a signature matrix $\Omega = \text{diag}(\omega_1, \ldots, \omega_n)$, $\omega_i = \pm 1$, and the spectral decomposition $K\Omega K^t = V\Lambda V^t, V^t = V^{-1}$, the hyperbolic SVD (HSVD) of K, introduced in [3], is defined as

$$K = V\Sigma P^t \quad \text{with} \quad P^t\Omega P = \bar{\Omega},$$

where $\bar{\Omega}$ is another signature matrix congruent to Ω. Note that $\Lambda = \Sigma^2\bar{\Omega}$. Without loss of generality, we can order the eigenvalues in Λ so that $\bar{\Omega} = \Omega$; hence the HSVD can be written as

$$K = V\Sigma P^t \quad \text{with} \quad P^t\Omega P = \Omega. \tag{35}$$

When $\Omega = I$, the standard SVD is recovered. ΩP holds the right Ω-singular vectors of K since $K\Omega P = V\Sigma\Omega$ while V holds the left Ω-singular vectors since $K^tV = P\Sigma$. Thus $(\sigma, \boldsymbol{v}, \Omega\boldsymbol{p})$ can be called an Ω-singular triple since

$$K\Omega\boldsymbol{p} = \boldsymbol{v}\sigma\omega, \quad K^t\boldsymbol{v} = \boldsymbol{p}\sigma. \tag{36}$$

Note that there are alternate interpretations: (i) P is the eigenvector matrix of the definite pair (K^tK, Ω) as $K^tK - \mu\Omega = P(\Sigma^2 - \mu\Omega)P^t, \Sigma^2 = |\Lambda|$, and (ii) ΩP is the eigenvector matrix of $\Omega K^tK = (\Omega P)\Lambda(\Omega P)^{-1}$ since, by (35), $(\Omega P)^{-1} = \Omega P^t$.

It is not hard to see that with $K = L|D|^{1/2}$, then, since $K^t\boldsymbol{v} = \boldsymbol{p}\sigma$,

$$\boldsymbol{\kappa}_{rel}(\lambda) = \frac{\boldsymbol{v}^tKK^t\boldsymbol{v}}{\sigma^2} = \boldsymbol{p}^t\boldsymbol{p} \geq 1.$$

We now present the main theorem of [33] which quantifies the change in the HSVD when K is perturbed to D_lKD_r. Before we do so, we must introduce some terminology.

Write $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n], P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n]$. One ingredient in the bounds on the change in $\boldsymbol{v}_j$ is the set $\{\|\boldsymbol{p}_i\|^2\}$. The other ingredient is the relative separation between eigenvalues. Actually it is the separations between the $\sigma_i = \sqrt{|\lambda_i|}$ that emerge naturally in the theory in [33],

$$\delta_{ji} := \frac{|\lambda_j - \lambda_i|}{\sigma_j + \sigma_i} = \begin{cases} |\sigma_j - \sigma_i| & \text{if } \omega_i = \omega_j, \\ \frac{\sigma_j^2 + \sigma_i^2}{\sigma_j + \sigma_i} & \text{if } \omega_i \neq \omega_j, \end{cases} \tag{37}$$

but the factors that govern the sensitivity of $\boldsymbol{v}_j$ involve both relative separations $(\delta_{ji}/\sigma_j)^2$ and $(\delta_{ji}/\sigma_i)^2$ for $i \neq j$. Note that

$$\left(\frac{\delta_{jk}}{\sigma_j}\right)^2 = \left(\frac{|\lambda_j - \lambda_k|}{\sigma_j(\sigma_j + \sigma_k)}\right)^2 = \frac{|\lambda_j - \lambda_k|}{|\lambda_j|} \cdot \frac{|\lambda_j - \lambda_k|}{(\sigma_j + \sigma_k)^2} \geq \frac{|\lambda_j - \lambda_k|}{|\lambda_j|} \cdot \frac{|\lambda_j - \lambda_k|}{2(|\lambda_j| + |\lambda_k|)}.$$

The upcoming bounds concern a particular eigenpair $(\lambda_j, \boldsymbol{v}_j)$, $\|\boldsymbol{v}_j\| = 1$, and as noted above, $\|\boldsymbol{p}_j\|$ plays a leading role. The other quantities of interest are

$$\begin{aligned} \mathrm{rgap}_j &:= \min_{i \neq j} \frac{\delta_{ji}}{\sigma_j}, \\ \|\boldsymbol{m}_j\|^2 &:= \sum_{i \neq j} \left(\frac{\|\boldsymbol{p}_i\|}{\delta_{ji}/\sigma_j}\right)^2, \\ \|\boldsymbol{m}_{\langle j \rangle}\|^2 &:= \sum_{i \neq j} \left(\frac{\|\boldsymbol{p}_i\|}{\delta_{ji}/\sigma_i}\right)^2, \\ \|P_{\langle j \rangle}\|_F^2 &:= \sum_{i \neq j} \|\boldsymbol{p}_i\|^2, \end{aligned}$$

where $\langle j \rangle$ denote the index set complementary to j in $\{1, \ldots, n\}$. The actual vectors $\boldsymbol{m}_j$ and $\boldsymbol{m}_{\langle j \rangle}$ play no role in the bounds; only their norms are needed. Note that $1/\mathrm{rgap}_j \leq \|\boldsymbol{m}_j\| \leq \|P_{\langle j \rangle}\|_F/\mathrm{rgap}_j$. Of the four expressions given above, the second appears in the bounds and the other three in the restriction on the perturbation level of Theorem 7.

We cite the needed part of the relevant theorem from [33] and then adapt it to our situation.

THEOREM 7. *Let Ω be a signature matrix, $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n)$, $\omega_i = \pm 1$. Consider an invertible matrix K with HSVD as in* (35). *Define $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) := \Sigma^2\Omega$ and assume that $\lambda_i \neq \lambda_j$, $i \neq j$. Let K be perturbed to D_lKD_r, with D_l and D_r diagonal, and let $\bar{\varepsilon}_d := \max\{\|D_l^{\pm 2} - I\|, \|D_r^{\pm 2} - I\|\}$. If $\bar{\varepsilon}_d$ is small enough that*

$$8\bar{\varepsilon}_d\|\boldsymbol{p}_j\|^2 \leq \mathrm{rgap}_j, \qquad 4\bar{\varepsilon}_d\|\boldsymbol{m}_{\langle j \rangle}\|\,\|P_{\langle j \rangle}\|_F \leq 1, \tag{38}$$

then there is an Ω-singular triple $(\bar{\sigma}_j, \bar{\boldsymbol{v}}_j, \Omega\bar{\boldsymbol{v}}_j)$ of D_lKD_r such that

$$|\sin\angle(\bar{\boldsymbol{v}}_j, \boldsymbol{v}_j)| \le \bar{\varepsilon}_d \left(2\|\boldsymbol{p}_j\|\|\boldsymbol{m}_j\| + \frac{1}{2}\right)/(1-\bar{\varepsilon}_d),$$

$$\text{and} \quad \frac{|\bar{\sigma}_j^2 - \sigma_j^2|}{\sigma_j^2} \le \frac{\bar{\varepsilon}_d(\|\boldsymbol{p}_j\|^2+1)+\beta_2}{1-\bar{\varepsilon}_d(\|\boldsymbol{p}_j\|^2+1)-\beta_2},$$

where $0 < \beta_2 \le 2(\bar{\varepsilon}_d\|\boldsymbol{p}_j\|)^2\{(\|P_{\langle j\rangle}\|_F + 2\|\boldsymbol{m}_j\|)^2 + 4\|\boldsymbol{m}_j\|\,\|\boldsymbol{m}_{\langle j\rangle}\|\}$.

Theorem 7 reveals how we should define the condition numbers:

$$\text{relcond}(\boldsymbol{v}_j) := 2\|\boldsymbol{p}_j\|\,\|\boldsymbol{m}_j\| + 1/2, \tag{39}$$

$$\text{relcond}(\lambda_j) := \|\boldsymbol{p}_j\|^2 + 1. \tag{40}$$

Note that relcond(λ_j) exceeds Dhillon's $\boldsymbol{\kappa}_{rel}(\lambda_j)$ by 1 and it accounts for both the inner and outer perturbations.

The following lemma shows that in the indefinite well-conditioned case and in the definite case, these relative condition numbers are small. For any invertible M define $\text{cond}_F(M) := \|M\|_F\|M^{-1}\|_F$ and $\text{cond}_2(M) := \|M\|_2\|M^{-1}\|_2$.

LEMMA 8. *With the notation developed above (setting $K = L|D|^{1/2}$, $\Omega = \text{sign}(D)$),*

$$\text{relcond}(\boldsymbol{v}_j) \;\le\; \frac{\text{cond}_F(L|D|^{1/2})}{\text{rgap}_j} + 1/2, \;\text{ and }\; \text{relcond}(\lambda_j) \;\le\; \text{cond}_2(L|D|^{1/2}).$$

When LDL^t is definite, then

$$\text{relcond}(\boldsymbol{v}_j) \;\le\; \frac{2\sqrt{n-1}}{\text{rgap}_j} + 1/2 \;\text{ and }\; \text{relcond}(\lambda_j) \;=\; 2.$$

Proof. In [33, Lemma 2.1] it is shown that

$$\|P\|_F^2 = \sum \|\boldsymbol{p}_i\|^2 \le \text{cond}_F(L|D|^{1/2}).$$

Observe, from the definition of $\|\boldsymbol{m}_j\|$, that $H_j^2 := \|P_{\langle j\rangle}\|_F^2/\|\boldsymbol{m}_j\|^2$ is a weighted harmonic mean of the relative gaps $(\delta_{ji}/\sigma_j)^2$ and, as for any mean, $\text{rgap}_j^2 \le H_j^2$. From the definition of relcond$(\boldsymbol{v}_j)$ in (39),

$$\begin{aligned}\text{relcond}(\boldsymbol{v}_j) &= 2\|\boldsymbol{p}_j\|\frac{\|P_{\langle j\rangle}\|_F}{H_j} + 1/2, \\ &\le \frac{\|P\|_F^2}{\text{rgap}_j} + 1/2, \\ &\le \frac{\text{cond}_F(L|D|^{1/2})}{\text{rgap}_j} + 1/2.\end{aligned} \tag{41}$$

The bound on relcond(λ_j) is an easy consequence of (36). When D is definite, then $\|\boldsymbol{p}_i\| = 1$ for all i, and $\|P_{\langle j\rangle}\|_F^2 = n-1$. Thus, by (41), the result holds. ∎

Next we apply Theorem 7 to our situation.

COROLLARY 9. *With the notation presented above let invertible tridiagonal $T = LDL^t$ have eigenpairs $(\lambda_i, \boldsymbol{v}_i)$, $\|\boldsymbol{v}_i\| = 1$, with L a proper bidiagonal matrix (all*

$l_i \neq 0$) with 1's on the diagonal, and D a diagonal matrix. Consider perturbations $L \longrightarrow E^{-1}LE$, $D \longrightarrow FDF$ with $\|(EF)^2 - I\| = \bar{\varepsilon}_d$. If $\bar{\varepsilon}_d$ is small enough that (38) *holds, then there is an eigenpair $(\bar{\lambda}_j, \bar{\boldsymbol{v}}_j)$ of $\bar{T} = E^{-1}LEFDFEL^tE^{-1}$ such that*

$$|\sin \angle(\bar{\boldsymbol{v}}_j, \boldsymbol{v}_j)| \leq \mathrm{relcond}(\boldsymbol{v}_j)\bar{\varepsilon}_d/(1-\bar{\varepsilon}_d),$$
$$\textit{and} \quad \frac{|\bar{\lambda}_j - \lambda_j|}{|\lambda_j|} \leq \frac{\bar{\varepsilon}_d \mathrm{relcond}(\lambda_j) + \beta_2}{1 - \bar{\varepsilon}_d \mathrm{relcond}(\lambda_j) - \beta_2},$$

where $\mathrm{relcond}(\boldsymbol{v}_j)$ *and* $\mathrm{relcond}(\lambda_j)$ *are defined in* (39) *and* (40)*, respectively, and β_2 is as in Theorem* 7.

Proof. L is proper so that no subdiagonal entry vanishes, T is invertible, and thus its eigenvalues are simple and do not vanish. Therefore we may apply Theorem 7 with $\bar{\varepsilon}_d := \max\{\|E^{-2} - I\|, \|(EF)^2 - I\|\} = \|(EF)^2 - I\|$. The left Ω-singular vectors V of $K = L|D|^{1/2}$ are the eigenvectors of $K\Omega K^t = T$. ∎

The bounds in Theorem 7 make no explicit reference to n, the order of the matrix. This factor appears when we relate $\bar{\varepsilon}_d$ to the machine precision ε. In our application, Lemma 6 above gives $1 + \bar{\varepsilon}_d < (1+\varepsilon)^{6n-1}$. Corollary 9 and this bound on $\bar{\varepsilon}_d$ will be applied in Theorem 15.

Lemma 8 covers the "easy" cases. However, in the case when LDL^t is indefinite and ill-conditioned the small eigenpairs are often relatively robust, while the large ones are not. Corollary 9 covers such cases. The following example illustrates one such situation.

Example 2. To bring life to the quantities of this section we exhibit a 4×4 symmetric matrix $T = T(\eta)$ that depends on a parameter η (think of η as 10^{-8}):

$$T = T(\eta) := \begin{bmatrix} \eta & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & -2\eta & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 3\eta & \eta \\ 0 & 0 & \eta & 2\eta \end{bmatrix},$$
$$T = LDL^t = V(\Sigma^2\Omega)V^t,$$
$$\Omega = \mathrm{diag}(1,-1,1,1), \quad L|D|^{1/2} = V\Sigma P^t.$$

T is indefinite but permits triangular factorization $T = LDL^t$ with large element growth, like $1/\eta$, in the multipliers. This ill-conditioned $L|D|^{1/2}$ has HSVD: $L|D|^{1/2} = V\Sigma P^t$; recall from (40) that the $\|\boldsymbol{p}_i\|$ govern the relative conditioning of the eigenvalues. We present only the leading terms in the quantities shown below. The eigenvalues $\omega\sigma^2$ of T are not presented in monotonic order because of the constraint that $P^t\Omega P = \Omega$. For full details see [33].

$$\Lambda = \Sigma^2\Omega = \mathrm{diag}\left(\frac{4-\sqrt{2}}{2}\eta, -1, \frac{4+\sqrt{2}}{2}\eta, +1\right),$$
$$\Sigma = \mathrm{diag}(\sqrt{\eta}\,\mu_-, 1, \sqrt{\eta}\,\mu_+, 1) \quad \mu_-^2 := \frac{4-\sqrt{2}}{2}, \quad \mu_+^2 := \frac{4+\sqrt{2}}{2},$$
$$\left\{\|\boldsymbol{p}_i\|^2\right\} = \left\{\ 2 - \frac{\sqrt{2}}{4}, \quad \frac{1}{2\eta}, \quad 2 + \frac{\sqrt{2}}{4}, \quad \frac{1}{2\eta}\ \right\}.$$

The two small σ's are close, $\sigma_1 \approx 1.137\eta$ and $\sigma_3 \approx 1.645\eta$, while the other σ's are almost 1 but have differing ω values. The large singular values are extremely

sensitive, "condition number" $1/\eta$, but the two small σ's are relatively robust and the associated $\boldsymbol{v}$'s also turn out to be relatively robust. We demonstrate the latter for $\boldsymbol{v}_1$.

For $\sigma_1 = \sqrt{\eta}\,\mu_- \approx 1.137\sqrt{\eta}$,

$$\text{rgap}_1 = \frac{\mu_+ - \mu_-}{\mu_-} \approx .447,$$

$$\delta_{12} = \frac{1+\eta\mu_-^2}{1+\sqrt{\eta}\mu_-}, \quad \delta_{13} = \sqrt{\eta}(\mu_+ - \mu_-), \quad \delta_{14} = 1 - \sqrt{\eta}\mu_-.$$

Recall from (37) that δ_{12} is a quotient, not just a difference, because $\omega_1 \neq \omega_2$.

The point of this example is the neutralizing of large $\|\boldsymbol{p}\|$ values in the first and third terms by equally large relative separations in $\|\boldsymbol{m}_1\|$ below:

$$\begin{aligned}
\|\boldsymbol{m}_1\| &= \left[\sum_{i\neq 1} \|\boldsymbol{p}_i\|^2 \left(\sigma_1/\delta_{1i}\right)^2\right]^{1/2} \\
&= \left[\frac{1}{2\eta}\left(\frac{\sqrt{\eta}\mu_-}{1-\sqrt{\eta}\mu_-}\right)^2 + \left(2+\frac{\sqrt{2}}{4}\right)\left(\frac{\sqrt{\eta}\mu_-}{\sqrt{\eta}(\mu_+-\mu_-)}\right)^2 + \frac{1}{2\eta}\left(\frac{\sqrt{\eta}\mu_-}{1-\sqrt{\eta}\mu_-}\right)^2\right]^{1/2} \\
&\approx \left[\frac{1}{2}\mu_-^2 + \left(2+\frac{\sqrt{2}}{4}\right)\left(\frac{\mu_-}{\mu_+-\mu_-}\right)^2 + \frac{1}{2}\mu_-^2\right]^{1/2} \\
&\approx (\mu_-)\left[1+(8+\sqrt{2})\right]^{1/2} \approx 3.44.
\end{aligned}$$

Combined with the modest value of $\|\boldsymbol{p}_1\|$, relcond$(\boldsymbol{v}_1)$ defined in (39) is less than 10. Similarly $\boldsymbol{v}_3$ is seen to be relatively robust, but $\boldsymbol{v}_2$ and $\boldsymbol{v}_4$ are not. ∎

6.2. The bottom arrow—rank-revealing twisted factorization. The bottom arrow in Figure 5 represents an exact relation $\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I = \bar{N}_r\bar{D}_r\bar{N}_r^t$. Consider the vector $\bar{\boldsymbol{z}}^{(r)}$ (denoted as $\bar{\boldsymbol{z}}$ in Figure 5) such that $\bar{\boldsymbol{z}}^{(r)}(r) = 1$ and $(\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I)\bar{\boldsymbol{z}}^{(r)} = \boldsymbol{e}_r\bar{\gamma}_r$. This section presents the desired bound on the residual norm $\|(\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I)\bar{\boldsymbol{z}}^{(r)}\|/\|\bar{\boldsymbol{z}}^{(r)}\|$ when r is chosen appropriately. For ease of notation, we drop the overbars for the rest of section 6.2; thus the quantities L, D, γ_k, and $\boldsymbol{z}$ below can be thought of as $\bar{L}$, $\bar{D}$, $\bar{\gamma_k}$, and $\bar{\boldsymbol{z}}$, respectively, in Figure 5. We first establish that in cases of interest, when $\hat{\lambda}$ approximates λ, then one of the γ_k, $1 \leq k \leq n$, reveals that $LDL^t - \hat{\lambda}I$ is nearly singular.

Let $\hat{\lambda} \neq \lambda$. Since $\boldsymbol{e}_k^t\boldsymbol{z}^{(k)} = 1 = \gamma_k\boldsymbol{e}_k^t(LDL^t - \hat{\lambda}I)^{-1}\boldsymbol{e}_k$,

$$\gamma_k^{-1} = \boldsymbol{e}_k^t(LDL^t - \hat{\lambda}I)^{-1}\boldsymbol{e}_k. \tag{42}$$

We present next the relation of γ_k to the spectral factorization of $LDL^t - \hat{\lambda}I$ using an eigenvector expansion. These results do not require the tridiagonal form.

Let $LDL^t = V\Lambda V^t$. Replace LDL^t with $V\Lambda V^t$ in (42) to find, for each k,

$$\frac{1}{\gamma_k} = \frac{|v_j(k)|^2}{\lambda_j - \hat{\lambda}} + \sum_{i\neq j}\frac{|v_i(k)|^2}{\lambda_i - \hat{\lambda}}, \tag{43}$$

where $\lambda = \lambda_j$ is an eigenvalue closest to $\hat{\lambda}$ and its normalized eigenvector is $\boldsymbol{v}_j$. The following theorem shows that when λ_j is isolated the twist index k for which the eigenvector component $|v_j(k)|$ is large leads to a small $|\gamma_k|$.

THEOREM 10. *Let γ_k be as in (43), where $\hat{\lambda}$ approximates λ_j, and let λ_j be isolated enough, i.e.,*

$$\frac{|\lambda_j - \hat{\lambda}|}{\mathrm{gap}(\hat{\lambda})} \;\le\; \frac{1}{M} \cdot \frac{1}{n-1}, \tag{44}$$

where $M > 1$ and $\mathrm{gap}(\hat{\lambda}) := \min_{i \neq j} |\lambda_i - \hat{\lambda}|$. Then, for k such that $v_j(k) \ge 1/\sqrt{n}$,

$$|\gamma_k| \;\le\; \frac{|\lambda_j - \hat{\lambda}|}{|v_j(k)|^2} \cdot \frac{M}{M-1} \;\le\; n|\lambda_j - \hat{\lambda}| \cdot \frac{M}{M-1}. \tag{45}$$

Proof. A proof is given in [9, Section 3.2], which we repeat here for the sake of completeness. By (43),

$$\frac{1}{\gamma_k} = \frac{|v_j(k)|^2}{\lambda_j - \hat{\lambda}} \left[1 + \sum_{i \neq j} \left| \frac{v_i(k)}{v_j(k)} \right|^2 \left(\frac{\lambda_j - \hat{\lambda}}{\lambda_i - \hat{\lambda}} \right) \right]. \tag{46}$$

Since

$$\sum_{i \neq j} \left| \frac{v_i(k)}{v_j(k)} \right|^2 = \frac{1 - |v_j(k)|^2}{|v_j(k)|^2},$$

we can rewrite (46) as

$$\frac{1}{\gamma_k} = \frac{|v_j(k)|^2}{\lambda_j - \hat{\lambda}} \left[1 + \left(|v_j(k)|^{-2} - 1 \right) \mathcal{A}_1 \right], \tag{47}$$

where

$$\mathcal{A}_1 = \sum_{i \neq j} w_i \left(\frac{\lambda_j - \hat{\lambda}}{\lambda_i - \hat{\lambda}} \right), \quad 1 = \sum_{i \neq j} w_i, \;\; w_i \ge 0,$$

and so

$$|\mathcal{A}_1| \le |\lambda_j - \hat{\lambda}| / \mathrm{gap}(\hat{\lambda}), \quad \text{with } \mathrm{gap}(\hat{\lambda}) = \min_{i \neq j} |\lambda_i - \hat{\lambda}|. \tag{48}$$

If (44) holds, then by (47) and (48),

$$|\gamma_k| \le \frac{|\lambda_j - \hat{\lambda}|}{|v_j(k)|^2} \left| 1 - \left(|v_j(k)|^{-2} - 1 \right) \left(\frac{1}{M \cdot (n-1)} \right) \right|^{-1}.$$

For k such that $|v_j(k)| \ge 1/\sqrt{n}$,

$$|\gamma_k| \le \frac{|\lambda_j - \hat{\lambda}|}{|v_j(k)|^2} \left[1 - \frac{1}{M} \right]^{-1},$$

and so the result holds. ∎

In general, the case $\gamma_k = \infty$ for all k can occur, but we are free to choose $\hat{\lambda}$ to avoid such situations; see also [9, section 3.3]. In cases of interest, $|\lambda_j - \hat{\lambda}|/\mathrm{gap}(\hat{\lambda}) = O(\varepsilon)$, implying that $M \gg 1$ and $M/(M-1) \approx 1$, whence (45) shows that when

$|v_j(k)|$ is above average, $|\gamma_k|$ reveals the near singularity. This justifies step III in Algorithm Getvec.

We now show that under suitable conditions the vector $\boldsymbol{z}^{(k)}$ enjoys a small residual norm and serves as an excellent approximation to the eigenvector $\boldsymbol{v}_j$ [14, 31].

THEOREM 11. *Let* $\boldsymbol{z}^{(k)}$ *satisfy*

$$(LDL^t - \hat{\lambda}I)\boldsymbol{z}^{(k)} = \boldsymbol{e}_k\gamma_k$$

with $z^{(k)}(k) = 1$, *and let* γ_k *be as in* (43) *where* $\hat{\lambda}$ *approximates* λ_j, $\hat{\lambda} \neq \lambda_j$. *Then, if* $v_j(k) \neq 0$, *the residual norm*

$$\frac{\|(LDL^t - \hat{\lambda}I)\boldsymbol{z}^{(k)}\|}{\|\boldsymbol{z}^{(k)}\|} \;=\; \frac{|\gamma_k|}{\|\boldsymbol{z}^{(k)}\|} \;\leq\; \frac{|\lambda_j - \hat{\lambda}|}{|v_j(k)|},$$

and thus for at least one k,

$$\frac{|\gamma_k|}{\|\boldsymbol{z}^{(k)}\|} \;\leq\; \sqrt{n}|\lambda_j - \hat{\lambda}|.$$

Proof. A proof is given in [31, section 5] and [9, section 3.2], but we repeat it here for the sake of completeness. Recall that $LDL^t = V\Lambda V^t$. Then

$$\begin{aligned}
\boldsymbol{z}^{(k)} &= (LDL^t - \hat{\lambda}I)^{-1}\boldsymbol{e}_k\gamma_k, \\
\Rightarrow \|\boldsymbol{z}^{(k)}\|^2 &= |\gamma_k|^2\boldsymbol{e}_k^t V(\Lambda - \hat{\lambda}I)^{-2}V^t\boldsymbol{e}_k, \\
&= |\gamma_k|^2 \sum_{i=1}^{n} \frac{|v_i(k)|^2}{|\hat{\lambda} - \lambda_i|^2}, \\
\Rightarrow \frac{|\gamma_k|}{\|\boldsymbol{z}^{(k)}\|} &\leq \frac{|\lambda_j - \hat{\lambda}|}{|v_j(k)|} \quad \text{for all } k \text{ with } v_j(k) \neq 0.
\end{aligned}$$

Noting that $|v_j(k)| \geq 1/\sqrt{n}$ for at least one k completes the proof. ∎

However, $(\hat{\lambda}, \boldsymbol{z}^{(k)})$ is not the best approximate eigenpair because $\hat{\lambda}$ is not the Rayleigh quotient of $\boldsymbol{z}^{(k)}$. By using the Rayleigh quotient we obtain a useful decrease in residual norm.

LEMMA 12. *Let* $LDL^t = T$ *and* $(T - \hat{\lambda}I)\boldsymbol{z}^{(k)} = \boldsymbol{e}_k\gamma_k$, $z^{(k)}(k) = 1$. *Then the Rayleigh quotient* ρ *with respect to* $T - \hat{\lambda}I$ *is*

$$\rho(\boldsymbol{z}^{(k)}) = \gamma_k/\|\boldsymbol{z}^{(k)}\|^2,$$

$$\text{and} \quad \|(T - (\hat{\lambda} + \rho)I)\boldsymbol{z}^{(k)}\|/\|\boldsymbol{z}^{(k)}\| = \frac{\gamma_k}{\|\boldsymbol{z}^{(k)}\|^2}\left(\|\boldsymbol{z}^{(k)}\|^2 - 1\right)^{1/2}.$$

Proof. Write $\boldsymbol{z}$ for $\boldsymbol{z}^{(k)}$ and γ for γ_k, and note that

$$\boldsymbol{z}^t(T - \hat{\lambda}I)\boldsymbol{z} = \boldsymbol{z}^t\boldsymbol{e}_k\gamma = \gamma, \quad \text{since } z(k) = 1,$$

and

$$\begin{aligned}
(T - (\hat{\lambda} + \rho)I)\boldsymbol{z} &= \boldsymbol{e}_k\gamma - \boldsymbol{z}\rho, \\
\|(T - (\hat{\lambda} + \rho)I)\boldsymbol{z}\|^2 &= \gamma^2 + \|\boldsymbol{z}\|^2\rho^2 - 2\gamma\rho, \\
&= \frac{\gamma^2}{\|\boldsymbol{z}\|^2}\left(\|\boldsymbol{z}\|^2 - 1\right).
\end{aligned}$$

∎

The above lemma justifies the use of Algorithm Getvec in increasing $\hat{\lambda}$'s accuracy; see Remark 5 in section 4.

6.3. The right side—computing the eigenvector by multiplications. This section looks at the right side of Figure 5 and shows that the vector $\hat{\boldsymbol{z}}$ computed by Algorithm Getvec is very close to a vector $\bar{\boldsymbol{z}}$ that obeys the exact relationship (49), where $\bar{N}_r$ and $\bar{D}_r$ are perturbed factors determined by step IV of Algorithm Getvec.

THEOREM 13. *Let $\hat{N}_r$ and $\hat{D}_r$, $\bar{N}_r$ and $\bar{D}_r$ be the twisted factors represented by $\hat{Z}_r$ and $\bar{Z}_r$, respectively, in Figure* 4 *(see also Theorem* 4 *and Figure* 5*). Let $\hat{\boldsymbol{z}}$ be the vector computed in step* IV *of Algorithm* Getvec*, and let $\bar{\boldsymbol{z}}$ be the exact solution of*

$$\bar{N}_r \bar{D}_r \bar{N}_r^t \bar{\boldsymbol{z}} = \bar{\gamma}_r \boldsymbol{e}_r, \tag{49}$$

where $\bar{z}(r) = 1$. Then, barring underflow, $\hat{\boldsymbol{z}}$ is a small relative perturbation of $\bar{\boldsymbol{z}}$. Specifically,

$$\begin{aligned} &\hat{z}(r) = \bar{z}(r) \;=\; 1, \\ &\hat{z}(i) = \bar{z}(i)\cdot(1+\eta_i), \;\; i \neq r, \quad (1-\varepsilon)^{5|i-r|}\varepsilon \leq 1+\eta_i \leq (1+\varepsilon)^{5|i-r|\varepsilon}, \end{aligned} \tag{50}$$

where ε is the machine precision.

Proof. The above bound accounts for the roundoff errors in the recurrence in step IV of Algorithm Getvec. For now, assume that no component of D_+ or D_- is zero (so that only the top formulae for $\hat{z}(i)$ and $\hat{z}(j+1)$ in step IV are used). The matrix $\bar{N}_r$, built out of $\vec{L}_+$ and $\overleftarrow{U}_-$, was defined in Theorem 4 so that the equality $\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I = \bar{N}_r \bar{D}_r \bar{N}_r^t$ holds. Thus $\bar{N}_r$ is a given matrix, not to be modified, in the context of this theorem. Because of the roundoff error in multiplication the top entries of $\hat{\boldsymbol{z}}$ computed in step IV of Algorithm Getvec satisfy

$$\hat{z}(i) = -\hat{L}_+(i)\hat{z}(i+1)(1+\varepsilon_i), \quad i < r,$$

and the bottom entries satisfy

$$\hat{z}(i) = -\hat{U}_-(i-1)\hat{z}(i-1)(1+\varepsilon_i), \quad i > r, \tag{51}$$

where $|\varepsilon_i| < \varepsilon$. In contrast, the ideal vector $\bar{\boldsymbol{z}}$ satisfies

$$\begin{aligned} \bar{z}(i) &= -\vec{L}_+(i)\bar{z}(i+1), && i < r, \\ \text{and} \quad \bar{z}(i) &= -\overleftarrow{U}_-(i-1)\bar{z}(i-1), && i > r. \end{aligned} \tag{52}$$

Since $\hat{z}(r) = \bar{z}(r) = 1$, we may define $\eta_r = 0$ and trivially write $\hat{z}(r) = \bar{z}(r)(1+\eta_r)$ with $|\eta_r| \leq 4(r-r)\varepsilon$. Now proceed by induction as i decreases in order to prove (50). Examine (23) to find that

$$\hat{L}_+(i) = \vec{L}_+(i)(1+\delta_i), \quad (1-\varepsilon)^3 < 1+\delta_i < (1+\varepsilon)^3 \quad \text{for all } i < r.$$

Thus

$$\begin{aligned} \hat{z}(i-1) &= -\vec{L}_+(i-1)(1+\delta_{i-1})\hat{z}(i)(1+\varepsilon_{i-1}), \\ &= -\vec{L}_+(i-1)(1+\delta_{i-1})\bar{z}(i)(1+\eta_i)(1+\varepsilon_{i-1}), \\ &\quad \text{where } (1-\varepsilon)^{4(r-i)} \leq 1+\eta_i \leq (1+\varepsilon)^{4(r-i)} \text{ by induction}, \\ &= \bar{z}(i-1)(1+\delta_{i-1})(1+\eta_i)(1+\varepsilon_{i-1}), \;\; \text{by (52)} \\ &= \bar{z}(i-1)(1+\eta_{i-1}), \quad \text{thus defining } 1+\eta_{i-1} := (1+\eta_i)(1+\delta_{i-1})(1+\varepsilon_{i-1}), \end{aligned}$$

and $(1-\varepsilon)^{4(r-i)+4} \leq 1+\eta_{i-1} \leq (1+\varepsilon)^{4(r-i)+4}$, as claimed.

For the lower half of $\hat{\boldsymbol{z}}$, $i \geq r$, the argument is similar with $\hat{U}_-$ and $\overleftarrow{U}_-$ involved instead of $\hat{L}_+$ and $\vec{L}_+$. Note that $\hat{U}_-$ is related to $\overleftarrow{U}_-$ by (31) and (30), which, respectively, involve $1\frac{1}{2}$ and 1 more ulps than (23).

To begin, define $\eta_r = 0$ so that $|\eta_r| \leq 5(r-r)\varepsilon$. For $i = r+1$, (50) holds since (31) gives 4.5 ulps for $\overleftarrow{U}_-(r)$ in (51), while $\varepsilon_{r+1} = 0$ (because $\hat{z}(r) = 1$). For $i > r+1$, (30) gives 4 ulps and ε_i in (51) gives one more ulp for an increase of at most 5 ulps each time i increases. Thus (50) holds for all values of i.

We now consider the case when an eigenvector entry vanishes, i.e., $D_+(i) = 0$ (see Remark 3 in section 4). In this case the alternate formulae in step IV of Algorithm Getvec are used to compute the next eigenvector entry, i.e., if $i < r$, then

$$z(i) = -(d_{i+1}l_{i+1}/d_i l_i)z(i+2), \tag{53}$$

where d_i and l_i are elements of the input matrices L and D. Examining the relations between d_i and $\vec{d}_i$, and between l_i and $\vec{l}_i$ in the proof of Theorem 2, we can see that the product

$$d_i l_i \;=\; \vec{d}_i\vec{l}_i(1+\xi_i) \;=\; \vec{D}_+(i)\vec{L}_+(i)(1+\xi_i), \quad (1-\varepsilon)^3 < 1+\xi_i < (1+\varepsilon)^3, \quad i < r.$$

Thus the term $(d_{i+1}l_{i+1}/d_i l_i)$ in (53) contributes 6 ulps, and combining these with the 4 arithmetic operations in (53), we can write

$$\hat{z}(i) = -(\vec{d}_{i+1}\vec{l}_{i+1}/\vec{d}_i\vec{l}_i)\hat{z}(i+2)\cdot(1+\delta_i),$$

where $(1-\varepsilon)^{10} < 1+\delta_i < (1+\varepsilon)^{10}$. (A closer analysis reveals that $(1-\varepsilon)^8 < 1+\delta_i < (1+\varepsilon)^8$.) Thus (50) holds in this case also. The case when $D_-(i+1) = 0$, $i > r$, is similar. ∎

Corollary 14 (to Theorem 13). *Under the hypotheses of Theorem* 13,

$$|\sin\angle(\bar{\boldsymbol{z}}, \hat{\boldsymbol{z}})| \leq \frac{(1+\varepsilon)^{5(n-1)} - 1}{(1-\varepsilon)^{5(n-1)}}.$$

Proof. First we establish a general result on elementwise perturbation of vectors which shows that the term $(n-1)$ above could be replaced by a weighted standard deviation of the relative changes to $\hat{\boldsymbol{z}}$'s entries.

Let $\mathbf{0} \neq \boldsymbol{u} \in \mathbb{R}^n$ and let $\bar{\boldsymbol{u}}$ be given by $\bar{u}(i) = (1+\eta_i)u(i)$. For expressions concerning the angle $\angle(\boldsymbol{u}, \bar{\boldsymbol{u}})$ there is no loss in assuming that $\|\boldsymbol{u}\|^2 = \boldsymbol{u}^t\boldsymbol{u} = 1$.

Now,

$$\begin{aligned}
\cos^2\angle(\boldsymbol{u}, \bar{\boldsymbol{u}}) &= \frac{(\bar{\boldsymbol{u}}^t\boldsymbol{u})^2}{\bar{\boldsymbol{u}}^t\bar{\boldsymbol{u}}} \\
&= \frac{1 + 2\sum \eta_i u(i)^2 + (\sum \eta_i u(i)^2)^2}{1 + 2\sum \eta_i u(i)^2 + \sum \eta_i^2 u(i)^2}, \\
\sin^2\angle(\boldsymbol{u}, \bar{\boldsymbol{u}}) &= \frac{\sum \eta_i^2 u(i)^2 - (\sum \eta_i u(i)^2)^2}{1 + 2\sum \eta_i u(i)^2 + \sum \eta_i^2 u(i)^2}.
\end{aligned}$$

The numerator is a weighted variance of the η_i which we denote by (std. dev.$(\eta_i; \boldsymbol{u})$)^2. The denominator exceeds $(1+\text{avg})^2$, where avg $=$ avg$(\eta_i; \boldsymbol{u}) = \sum \eta_i u(i)^2$ because, by Cauchy–Schwarz, avg$^2 = (\sum \eta_i u(i)\cdot u(i))^2 \leq \sum \eta_i^2 u(i)^2$. On taking square roots,

$$|\sin\angle(\boldsymbol{u}, \bar{\boldsymbol{u}})| \leq \frac{\text{std. dev.}(\eta_i; \boldsymbol{u})}{1 + \text{avg}}. \tag{54}$$

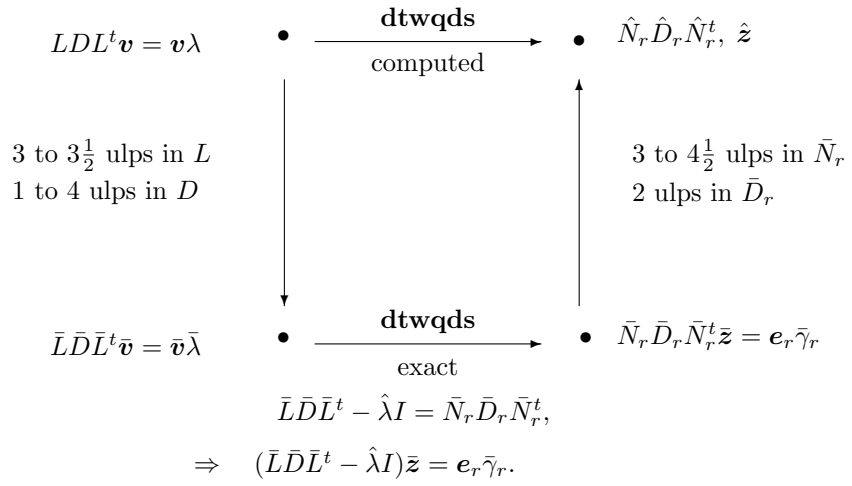


FIG. 6. *Relationships connecting* $\boldsymbol{v}$ *to* $\hat{\boldsymbol{z}}$.

A crude but simple bound on a standard deviation of the η_i is $\max_i |\eta_i|$. Finally substitute $\bar{\boldsymbol{z}}$ for $\boldsymbol{u}$ and $\hat{\boldsymbol{z}}$ for $\bar{\boldsymbol{u}}$ and use (50) to verify that

$$1 + \max_i |\eta_i| \ \le\ \max\left((1+\varepsilon)^{5(n-r)}, (1+\varepsilon)^{5(r-1)}\right),$$

and, by Theorem 13,

$$(1-\varepsilon)^{5(n-1)} \ \le\ 1 + \text{avg.}$$

Since r might be 1 or n the corollary is established. ∎

Note: The standard deviation in (54) *is weighted by the squares of the eigenvector entries. So, in practice, for localized eigenvectors we can replace* $n-1$ *in the above bound with the size of the numerical support of* $\hat{\boldsymbol{z}}$.

7. Bounds on accuracy (proof of correctness). The following is the main theorem of the paper. Figure 6 is identical to Figure 5 and we repeat it here so that it can be readily consulted.

THEOREM 15. *Let* $(\lambda, \boldsymbol{v}) = (\lambda_j, \boldsymbol{v}_j)$ *be an eigenpair of the real symmetric unreduced* $n \times n$ *tridiagonal matrix* LDL^t *with* $\|\boldsymbol{v}\| = 1$. *Let* $\hat{\lambda}$ *be an accurate approximation closer to* λ *than to any other eigenvalue of* LDL^t *and let* $\hat{\boldsymbol{z}}$ *be the vector computed in step* IV *of Algorithm* Getvec *in section* 4 *using* $\hat{\lambda}$, $\hat{N}_r$, $\hat{D}_r$, *and twist index* r. *Let* $\bar{L}$ *and* $\bar{D}$ *be the perturbations of* L *and* D *determined by the error analysis of section* 5, *and let* $(\bar{\lambda}, \bar{\boldsymbol{v}})$ *be the eigenpair of* $\bar{L}\bar{D}\bar{L}^t$ *with* $\bar{\lambda}$ *the closest eigenvalue to* $\hat{\lambda}$, *and* $\|\bar{\boldsymbol{v}}\| = 1$. *Let* ε *denote the machine precision, and for convenience, let* $\varepsilon_* := (1+\varepsilon)^{6n-1} - 1$. *If* ε *is small enough that* (38) *holds with* ε_* *instead of* $\bar{\varepsilon}_d$, *then*

$$|\sin\angle(\hat{\boldsymbol{z}}, \boldsymbol{v})| \le \frac{(1+\varepsilon)^{5(n-1)} - 1}{(1-\varepsilon)^{5(n-1)}} + \frac{|\bar{\lambda} - \hat{\lambda}|}{|\bar{v}(r)|\text{gap}(\hat{\lambda})} + \frac{\varepsilon_* \text{relcond}(\boldsymbol{v})}{1-\varepsilon_*}, \tag{55}$$

where relcond$(\boldsymbol{v})$ *is as in* (39), *and*

$$\text{gap}(\hat{\lambda}) := \min\{|\hat{\lambda} - \bar{\mu}|,\ \bar{\lambda} \ne \bar{\mu} \in \textit{spectrum of } \bar{L}\bar{D}\bar{L}^t\}.$$

Proof. There are three terms in the upper bound on $\sin\angle(\hat{\boldsymbol{z}}, \boldsymbol{v})$ because we connect $\hat{\boldsymbol{z}}$ to $\boldsymbol{v}$ via two "ideal" vectors $\bar{\boldsymbol{z}}$, $\bar{\boldsymbol{v}}$ and each transition contributes a term: $\hat{\boldsymbol{z}} \to \bar{\boldsymbol{z}}$, $\bar{\boldsymbol{z}} \to \bar{\boldsymbol{v}}$, $\bar{\boldsymbol{v}} \to \boldsymbol{v}$; see Figure 6. Recall from Theorem 4 that the matrices $\bar{L}$, $\bar{D}$, $\bar{N}_r$, $\bar{D}_r$ depend on $\hat{\lambda}$ and were defined so that the equality

$$\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I = \bar{N}_r\bar{D}_r\bar{N}_r^t \tag{56}$$

holds. That was the culmination of the error analysis in section 5. Recall that $\bar{D}_r(r) = \bar{\gamma}_r$. Then $\bar{\boldsymbol{z}}$ is defined as the exact solution of

$$\bar{N}_r\bar{D}_r\bar{N}_r^t\bar{\boldsymbol{z}} = \boldsymbol{e}_r\bar{\gamma}_r, \tag{57}$$

with $\bar{z}(r) = 1$. First consider $\hat{\boldsymbol{z}}$ and $\bar{\boldsymbol{z}}$. Theorem 13 shows that each $\bar{z}(i)$ is of the form $\hat{z}(i)(1+\eta_i)$ and Corollary 14 proves that

$$|\sin\angle(\hat{\boldsymbol{z}}, \bar{\boldsymbol{z}})| < \frac{(1+\varepsilon)^{5(n-1)} - 1}{(1-\varepsilon)^{5(n-1)}}. \tag{58}$$

Next consider $\bar{\boldsymbol{z}}$ and $\bar{\boldsymbol{v}}$. Combine (56) and (57) and then invoke Theorem 11 in section 6.2 to find that

$$\frac{\|(\bar{L}\bar{D}\bar{L}^t - \hat{\lambda}I)\bar{\boldsymbol{z}}\|}{\|\bar{\boldsymbol{z}}\|} = \frac{|\bar{\gamma}_r|}{\|\bar{\boldsymbol{z}}\|} \le \frac{|\bar{\lambda} - \hat{\lambda}|}{|\bar{v}(r)|}.$$

So by Theorem 1 (the gap theorem),

$$|\sin\angle(\bar{\boldsymbol{v}}, \bar{\boldsymbol{z}})| \le \frac{|\bar{\lambda} - \hat{\lambda}|}{|\bar{v}(r)|\mathrm{gap}(\hat{\lambda})}. \tag{59}$$

Finally consider $\bar{\boldsymbol{v}}$ and $\boldsymbol{v}$. The left side of Figure 6 indicates that $\bar{\boldsymbol{v}}$ and $\boldsymbol{v}$ are related through the matrix perturbations given in section 6.1 (see Lemma 5):

$$LDL^t \longrightarrow \bar{L}\bar{D}\bar{L}^t = E^{-1}LEFDFEL^tE^{-1}.$$

Theorem 4 bounded the entries in the specific matrices E and F and, by Lemma 6 in section 6.1.1,

$$\bar{\varepsilon}_d := \|(EF)^2 - I\| < \varepsilon_* := (1+\varepsilon)^{6n-1} - 1.$$

Thus Corollary 9 yields

$$|\sin\angle(\boldsymbol{v}, \bar{\boldsymbol{v}})| \le \frac{\mathrm{relcond}(\boldsymbol{v})\varepsilon_*}{1-\varepsilon_*}. \tag{60}$$

Add (58), (59), and (60) to obtain the theorem's bound on $|\sin\angle(\hat{\boldsymbol{z}}, \boldsymbol{v})|$. ∎

Next we discuss the implications of Theorem 15 for computing numerically orthogonal eigenvectors from Algorithm Getvec. The first term is essentially $5n\varepsilon$ and the last is $\mathrm{relcond}(\boldsymbol{v})6n\varepsilon$, and we are concerned only with cases when $\mathrm{relcond}(\boldsymbol{v})$ is $O(1)$. The middle term is the delicate one. If we bound each term separately, we would have $|\hat{\lambda} - \bar{\lambda}| \le Kn\varepsilon|\lambda|$, $1/|v(r)| \le \sqrt{n}$, and $\mathrm{relgap}(\hat{\lambda}) \ge tol$, giving a bound that exceeds $O(n\varepsilon)$. However for symmetric tridiagonal matrices the three terms are not independent. Moreover Getvec is often invoked for small isolated eigenvalues that have very large $\mathrm{relgap}(\lambda)$. For example, let us consider the extreme example introduced by

Demmel and Kahan in [7] that shows the bound of $Kn\varepsilon$ on the change in eigenvalue is attainable. The matrix is LL^t, L bidiagonal, with $L_{ii} = 1$ and $L_{i+1,i} = \beta \gg 1$. Think of $\beta = 10$. Small relative changes of $1+\varepsilon$ to the off-diagonals and $1-\varepsilon$ to the diagonal entries change $\lambda_{min} = \lambda_1 \approx \beta^{2(1-n)}$ by $2(2n-1)\varepsilon|\lambda| < 4n\varepsilon|\lambda|$. Only the smallest eigenvalue suffers this degree of sensitivity but $\mathrm{relgap}(\lambda_1) \approx \beta^2 n - 1$ while $|v(r)| \approx 1$. In this case the middle term in (55) is negligible compared to the other two terms. In fact, the corresponding eigenvector $\boldsymbol{v}$ decays very rapidly and very low accuracy in the eigenvalue (correct exponent) is sufficient to produce a very good eigenvector. All the other eigenvalues of this example are clustered in $[\beta^2 - 2/\beta, \beta^2 + 1]$ and so their eigenvectors should be calculated from a factorization of $LL^t - (\beta^2+1)I$, not from LL^t (see Algorithm MR^3 in [10]).

In general this middle term warrants further study, but we must recall from Theorem 11 that part of the middle term is a bound on the quantity $|\gamma_r|/\|\boldsymbol{z}\|$, and in practice, we have good approximations (up to order of magnitude) on it as well as on $\mathrm{gap}(\lambda)$. So any situation in which the middle term is too large is detectable. The algorithm monitors this term before accepting an eigenvector.

The reader may have noticed that the bound (55) contains quantities from both the factorizations LDL^t and $\bar{L}\bar{D}\bar{L}^t$; for example $\mathrm{gap}(\hat{\lambda})$ in the middle term is with respect to the eigenvalues of $\bar{L}\bar{D}\bar{L}^t$. However, the entries in L and $\bar{L}$ differ by at most 3 ulps and those of D and $\bar{D}$ by at most 4 ulps. Our application is only to well-conditioned eigenpairs, and so such λ and $\bar{\lambda}$ will differ only by a few ulps and the computed $\hat{\lambda}$ must be a good approximation to each of them. We feel that it is satisfactory to present our results in this form.

The following corollary summarizes a typical situation in which Algorithm Getvec is invoked.

COROLLARY 16. *In addition to the assumptions of Theorem* 15 *suppose that* (i) r *is such that* $\bar{v}(r) \geq 1/\sqrt{n}$, (ii) $\hat{\lambda}$ *is computed to satisfy* $|\hat{\lambda} - \bar{\lambda}|/|\hat{\lambda}| \leq K\varepsilon$, (iii) $\mathrm{relgap}(\hat{\lambda})$ *exceeds* 2^{-8}, *and* (iv) $\mathrm{relcond}(\boldsymbol{v}) \leq M/\mathrm{relgap}(\hat{\lambda})$. *Then*

$$|\sin\angle(\hat{\boldsymbol{z}}, \boldsymbol{v})| \leq 5n\varepsilon + 2^8 K\sqrt{n}\varepsilon + 2^8 M\varepsilon + O(\varepsilon^2). \qquad \square$$

8. Numerical examples. We first compare and contrast the behavior of Algorithm Getvec on two 3×3 tridiagonals. These aptly illustrate various aspects of the theory.

Example 3. First consider the matrix

$$T_0 = \begin{bmatrix} 1 & \sqrt{\varepsilon} & 0 \\ \sqrt{\varepsilon} & 7\varepsilon/4 & \varepsilon/4 \\ 0 & \varepsilon/4 & 3\varepsilon/4 \end{bmatrix},$$

where ε is the machine precision ($\varepsilon \approx 2.2 \times 10^{-16}$ in IEEE double precision). The eigenvalues of T_0 are

$$\lambda_1 = \varepsilon/2 + O(\varepsilon^2), \quad \lambda_2 = \varepsilon + O(\varepsilon^2), \quad \lambda_3 = 1 + \varepsilon + O(\varepsilon^2),$$

while the corresponding normalized eigenvectors are

$$\boldsymbol{v}_1 = \begin{bmatrix} -\sqrt{\varepsilon/2} + O(\varepsilon^{3/2}) \\ \frac{1}{\sqrt{2}}(1 + \frac{\varepsilon}{4}) + O(\varepsilon^2) \\ -\frac{1}{\sqrt{2}}(1 - \frac{3\varepsilon}{4}) + O(\varepsilon^2) \end{bmatrix}, \qquad \boldsymbol{v}_2 = \begin{bmatrix} -\sqrt{\varepsilon/2} + O(\varepsilon^{3/2}) \\ \frac{1}{\sqrt{2}}(1 - \frac{5\varepsilon}{4}) + O(\varepsilon^2) \\ \frac{1}{\sqrt{2}}(1 + \frac{3\varepsilon}{4}) + O(\varepsilon^2) \end{bmatrix},$$

$$\boldsymbol{v}_3 = \begin{bmatrix} 1 - \frac{\varepsilon}{2} + O(\varepsilon^2) \\ \sqrt{\varepsilon} + O(\varepsilon^{3/2}) \\ \frac{\varepsilon^{3/2}}{4} + O(\varepsilon^{5/2}) \end{bmatrix}.$$

The exact triangular factorization is given by $T_0 = L_0^{exact} D_0^{exact} (L_0^{exact})^t$, where

$$L_0^{exact} = \begin{bmatrix} 1 & 0 & 0 \\ \sqrt{\varepsilon} & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix} \quad \text{and} \quad D_0^{exact} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3\varepsilon/4 & 0 \\ 0 & 0 & 2\varepsilon/3 \end{bmatrix}.$$

When applying Algorithm Getvec to the above matrix, we observe the following.

1. The factorization computed in IEEE double precision arithmetic, $L_0 D_0 L_0^t$, turns out to be exact, i.e., $L_0 = L_0^{exact}$ and $D_0 = D_0^{exact}$.
2. The computed eigenvalues $\hat{\lambda}_i$ satisfy
$$|\hat{\lambda}_i - \lambda_i| \le 2\varepsilon |\hat{\lambda}_i|, \qquad 1 \le i \le 3.$$
3. For each $\hat{\lambda}_i$, $\gamma_k^{(i)}$ can be computed by applying steps I–III of Algorithm Getvec. The computed values are
$$\gamma^{(1)} = \begin{bmatrix} 1.11 \cdot 10^{-16} \\ 2.46 \cdot 10^{-32} \\ 2.46 \cdot 10^{-32} \end{bmatrix}, \quad \gamma^{(2)} = \begin{bmatrix} 2.22 \cdot 10^{-16} \\ 4.93 \cdot 10^{-32} \\ 4.93 \cdot 10^{-32} \end{bmatrix}, \quad \gamma^{(3)} = \begin{bmatrix} 4.44 \cdot 10^{-16} \\ -2.00 \\ -1.00 \end{bmatrix}.$$
Algorithm Getvec chooses $r = 2$ for $\hat{\lambda}_1$, $r = 2$ for $\hat{\lambda}_2$, and $r = 1$ for $\hat{\lambda}_3$. Note that for the first two eigenvalues $|\gamma_r| = O(\varepsilon^2) = O(\varepsilon|\lambda_i|) \ll \varepsilon\|T_0\|$.
4. The eigenvectors $\hat{\boldsymbol{v}}_i$ computed by Algorithm Getvec are such that
$$\max |\hat{\boldsymbol{v}}_i^t \hat{\boldsymbol{v}}_j| = 1.66 \cdot 10^{-16} < \varepsilon, \qquad 1 \le i \le 3, \;\; 1 \le j < i,$$
$$\max \frac{|\hat{\boldsymbol{v}}_i(k) - \boldsymbol{v}_i(k)|}{|\boldsymbol{v}_i(k)|} = 8.88 \cdot 10^{-16} < 4\varepsilon, \qquad 1 \le i \le 3, \;\; 1 \le k \le 3.$$
Amazingly each eigenvector entry is computed to high relative accuracy, even the tiny $v_3(3)$ entry.
5. Instead of Algorithm Getvec, we can use one step of inverse iteration,
$$(L_0 D_0 L_0^t - \hat{\lambda}_i I)\, \boldsymbol{x}_i \;=\; \text{random vector},$$
to compute the eigenvectors. These computed vectors also turn out to be accurate and numerically orthogonal (however, the tiny $v_3(3)$ entry is not computed to high relative accuracy). Note that the analysis of section 7 does not extend to random right-hand sides.
6. Both $|\gamma_2^{(3)}|$ and $|\gamma_3^{(3)}|$ are large while the corresponding eigenvector entries are $O(\sqrt{\varepsilon})$ and $O(\varepsilon^{3/2})$, respectively. Thus the numerical support of an eigenvector cannot solely be determined by the magnitudes of γ_i, and illustrates our comments at the end of Remark 6 in section 4. ∎

Example 4. The above matrix T_0 is a "benign" example. The following, also discussed in section 3, is a harder case:

$$T_1 = \begin{bmatrix} 1-\sqrt{\varepsilon} & \varepsilon^{1/4}\sqrt{1-7\varepsilon/4} & 0 \\ \varepsilon^{1/4}\sqrt{1-7\varepsilon/4} & \sqrt{\varepsilon}+7\varepsilon/4 & \varepsilon/4 \\ 0 & \varepsilon/4 & 3\varepsilon/4 \end{bmatrix}.$$

The eigenvalues of T_1 are

$$\lambda_1 = \frac{\varepsilon}{2} + \frac{\varepsilon^{3/2}}{8} + O(\varepsilon^2), \quad \lambda_2 = \varepsilon - \frac{\varepsilon^{3/2}}{8} + O(\varepsilon^2), \quad \lambda_3 = 1 + \varepsilon + O(\varepsilon^2),$$

while the corresponding normalized eigenvectors are

$$\boldsymbol{v}_1 = \begin{bmatrix} \frac{\varepsilon^{1/4}}{\sqrt{2}}(1+\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ -\frac{1}{\sqrt{2}}(1-\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon) \\ \frac{1}{\sqrt{2}}(1-\frac{3\varepsilon}{4}) + O(\varepsilon^{3/2}) \end{bmatrix}, \qquad \boldsymbol{v}_2 = \begin{bmatrix} \frac{\varepsilon^{1/4}}{\sqrt{2}}(1+\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ -\frac{1}{\sqrt{2}}(1-\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon) \\ -\frac{1}{\sqrt{2}}(1+\frac{3\varepsilon}{4}) + O(\varepsilon^{3/2}) \end{bmatrix},$$

$$\boldsymbol{v}_3 = \begin{bmatrix} 1-\frac{\sqrt{\varepsilon}}{2} + O(\varepsilon) \\ \varepsilon^{1/4}(1+\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon^{5/4}) \\ \frac{\varepsilon^{5/4}}{4}(1+\frac{\sqrt{\varepsilon}}{2}) + O(\varepsilon^{9/4}) \end{bmatrix}.$$

In exact arithmetic, $T_1 = L_1^{exact} D_1^{exact} (L_1^{exact})^t$, where

$$L_1^{exact} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{\varepsilon^{1/4}\sqrt{1-7\varepsilon/4}}{1-\sqrt{\varepsilon}} & 1 & 0 \\ 0 & \frac{1-\sqrt{\varepsilon}}{3} & 1 \end{bmatrix} \quad \text{and}$$

$$D_1^{exact} = \begin{bmatrix} 1-\sqrt{\varepsilon} & 0 & 0 \\ 0 & \frac{3\varepsilon}{4(1-\sqrt{\varepsilon})} & 0 \\ 0 & 0 & \frac{\varepsilon(8+\sqrt{\varepsilon})}{12} \end{bmatrix}.$$

On this example, Algorithm Getvec behaves quite differently than on T_0 from Example 3:

1. The computed factorization $L_1 D_1 L_1^t$ does not have high relative accuracy. The relative errors in $L_1(2), D_1(2)$, and $D_1(3)$ are as large as $4.97 \cdot 10^{-9}$.
2. Consequently, some of the computed eigenvalues $\hat{\lambda}_i$ do not have high relative accuracy with respect to the eigenvalues of T_1. In particular,

   $$|\hat{\lambda}_i - \lambda_i| \approx 10^{-9} |\hat{\lambda}_i| \quad \text{for } i = 1, 2.$$

   Unlike λ_1 and λ_2, the third eigenvalue λ_3 is computed to high relative accuracy, i.e., $|\hat{\lambda}_3 - \lambda_3| = O(\varepsilon)$. However, the important point is that all the $\hat{\lambda}_i$ have high relative accuracy with respect to the eigenvalues of $L_1 D_1 L_1^t$.
3. The values of $\gamma_k^{(i)}$ computed by steps I–III of Algorithm Getvec are

   $$\gamma^{(1)} = \begin{bmatrix} -4.13 \cdot 10^{-24} \\ -7.40 \cdot 10^{-32} \\ -9.86 \cdot 10^{-32} \end{bmatrix}, \qquad \gamma^{(2)} = \begin{bmatrix} -6.62 \cdot 10^{-24} \\ -9.86 \cdot 10^{-32} \\ -9.86 \cdot 10^{-32} \end{bmatrix},$$

   $$\gamma^{(3)} = \begin{bmatrix} 2.22 \cdot 10^{-16} \\ 1.49 \cdot 10^{-8} \\ -1.00 \end{bmatrix}.$$

Algorithm Getvec chooses $r = 2$ for $\hat{\lambda}_1$, $r = 2$ for $\hat{\lambda}_2$, and $r = 1$ for $\hat{\lambda}_3$. Note that for the first two eigenvalues $|\gamma_r| = O(\varepsilon^2) \ll \varepsilon\|T\|$.

4. The eigenvectors $\hat{\boldsymbol{v}}_i$ computed in step IV of Algorithm Getvec are numerically orthogonal, i.e.,

$$\max |\hat{\boldsymbol{v}}_i^t \hat{\boldsymbol{v}}_j| \;=\; 5.55 \cdot 10^{-17} < \varepsilon, \qquad 1 \le i \le 3,\;\; 1 \le j < i.$$

But as in the case of the computed eigenvalues, the relative errors in the computed eigenvectors (with respect to the eigenvectors of T_1) are much larger than $O(\varepsilon)$, i.e.,

$$\max \frac{|\hat{\boldsymbol{v}}_i(k) - \boldsymbol{v}_i(k)|}{|\boldsymbol{v}_i(k)|} \;=\; 3.72 \cdot 10^{-9}, \qquad 1 \le i \le 2,\;\; 1 \le k \le 3.$$

All components of the third eigenvector $\boldsymbol{v}_3$ are computed to high relative accuracy.

5. The inverse iteration step

$$\begin{aligned} L_1 D_1 L_1^t - \hat{\lambda}_i I &= L_+ D_+ L_+^t, \\ L_+ D_+ L_+^t \, \boldsymbol{x}_i &= \text{random vector} \end{aligned} \tag{61}$$

also leads to computed eigenvectors that are numerically orthogonal when the dstqds transform is used to compute (61). From our experience, the use of a twisted factorization in Algorithm Getvec does not appear to be essential in practice; inverse iteration using dstqds also works well. However, twisted factorizations are more elegant to use, have better numerical behavior, and allow us to prove the accuracy of our algorithm.

6. When the diagonal and off-diagonal elements of T_1 are directly used to compute eigenvalues and eigenvectors (either by using inverse iteration or twisted factorizations as in Algorithm Getvec), the dot products between the computed eigenvectors are as large as 10^{-8}. See Example 1 in section 3 for an explanation of this failure. Thus the use of $L_1 D_1 L_1^t$ is essential for achieving numerical orthogonality in this case. ∎

The above example beautifully illustrates our techniques. We do not promise high relative accuracy for eigenvalues and eigenvectors of the given tridiagonal matrix. In fact, it is unrealistic to hope for such accuracy as explained in section 3. However, we get a "good" factorization of the tridiagonal and then proceed to compute its eigenvalues and eigenvectors to high accuracy, which automatically leads to orthogonality.

Example 5. Our next example is

$$T_2 = \begin{bmatrix} .520000005885958 & .519230209355285 & & \\ .519230209355285 & .589792290767499 & .36719192898916 & \\ & .36719192898916 & 1.89020772569828 & 2.7632618547882 \cdot 10^{-8} \\ & & 2.7632618547882 \cdot 10^{-8} & 1.00000002235174 \end{bmatrix}$$

with eigenvalues

$$\lambda_1 \approx \varepsilon, \quad \lambda_2 \approx 1 + \sqrt{\varepsilon}, \quad \lambda_3 \approx 1 + 2\sqrt{\varepsilon}, \quad \lambda_4 \approx 2.0.$$

Note that the interior eigenvalues have $\text{relgap}(\lambda_i) = O(\sqrt{\varepsilon})$. When we apply Algorithm Getvec to the LDL^t factorization of T_2, the corresponding computed eigenvectors have

$$|\hat{\boldsymbol{v}}_2^t \hat{\boldsymbol{v}}_3| \;=\; 1.12 \cdot 10^{-8} \;=\; O(\sqrt{\varepsilon}).$$

TABLE 1
Timing comparisons for computing all eigenvalues and eigenvectors.

Matrix type	Matrix size	Time taken (in seconds)				
		Lapack DSTEBZ +DSTEIN	DSTEBZ + Eispack TINVIT	Lapack DSTEDC	Lapack DSTEQR	Lapack DLASQ1 +Algorithm Getvec
Arithmetic progression (ε apart)	125	0.20 (.09+.11)	0.14 (.09+.05)	0.01	0.13	**0.05** (.01+.04)
	250	1.12 (.32+.80)	0.67 (.32+.35)	0.03	0.98	**0.08** (.02+.06)
	500	7.81 (1.25+6.06)	4.17 (1.25+2.92)	0.20	7.46	**0.39** (.11+.28)
	1000	93.97 (4.87+89.10)	39.44 (4.87+34.57)	1.22	74.33	**1.19** (.36+.83)
	2000	839.5 (20.9+818.6)	343.9 (20.9+323.0)	6.03	913.7	**4.50** (1.5+3.0)
Uniform distribution (ε to 1)	125	0.11 (.08+.03)	0.10 (.08+.02)	0.05	0.13	**0.05** (.01+.05)
	250	0.45 (.33+.12)	0.38 (.33+.05)	0.24	0.99	**0.11** (.04+.07)
	500	1.74 (1.24+.50)	1.47 (1.24+.23)	1.51	7.51	**0.31** (.11+.20)
	1000	93.37 (4.87+88.50)	5.68 (4.87+.81)	10.47	74.07	**1.16** (.40+.76)
	2000	839.5 (20.9+818.6)	344.8 (20.9+323.8)	159.3	648.2	**4.57** (1.6+3.0)
Uniform distribution (ε to 1 with random signs)	125	0.11 (.08+.03)	0.09 (.08+.01)	0.05	0.13	**0.05** (.01+.04)
	250	0.45 (.33+.12)	0.39 (.33+.06)	0.25	0.96	**0.13** (.04+.09)
	500	1.78 (1.27+.51)	1.50 (1.27+.23)	1.51	7.40	**0.37** (.09+.28)
	1000	7.21 (4.96+2.25)	5.89 (4.96+.93)	10.25	68.85	**1.51** (.37+1.14)
	2000	31.40 (21.3+10.1)	25.20 (21.3+4.10)	160.8	955.6	**5.10** (1.5+3.6)
(1,2,1) Matrix	125	0.12 (.09+.03)	0.10 (.09+.01)	0.05	0.13	**0.04** (.01+.03)
	250	0.44 (.32+.12)	0.37 (.32+.05)	0.16	0.92	**0.08** (.03+.05)
	500	1.85 (1.24+.61)	1.49 (1.24+.25)	1.02	7.01	**0.39** (.09+.30)
	1000	12.38 (4.88+7.50)	7.06 (4.88+2.18)	7.26	71.75	**1.22** (.32+.90)
	2000	840.0 (21.0+819.0)	128.8 (21.0+107.8)	105.8	678.7	**4.71** (1.6+3.1)
Biphenyl	966	77.61 (4.6+73.0)	33.02 (4.6+28.4)	7.66	73.96	**1.33** (.3+1.03)

As discussed in Remark 8 in section 4, inverse iteration appears to be a natural remedy to cure the problem. However, even after ten inverse iteration steps

$$|\hat{\boldsymbol{v}}_2^t \hat{\boldsymbol{v}}_3| \;=\; 3.45 \cdot 10^{-9} \;=\; O(\sqrt{\varepsilon}).$$

Thus the simple approach of using multiple inverse iteration steps does not lead to numerical orthogonality, as explained in Remark 8. For an approach that can achieve orthogonality in this situation, the reader is referred to [10]; also see Chapter 5 in [9]. □

8.1. Timing comparisons. Algorithm Getvec can lead to substantial speedups over earlier LAPACK software[2] to compute eigenvectors when the relative gaps between eigenvalues exceed *tol* ($= 10^{-3}$) but the absolute gaps are smaller. We illustrate this speedup on various examples in Table 1. Matrices of the first type have eigenvalues in an arithmetic progression,

$$\lambda_i = i \cdot \varepsilon, \quad i = 1, 2, \ldots, n-1, \;\text{ and }\; \lambda_n = 1.$$

The second type has eigenvalues that come from a uniform random distribution in the interval $[\varepsilon, 1]$, while the third type has a similar eigenvalue distribution as the second

[2]Since we first wrote this paper, our software has been incorporated in the latest release of LAPACK, where Algorithm Getvec appears as subroutine DLAR1V.

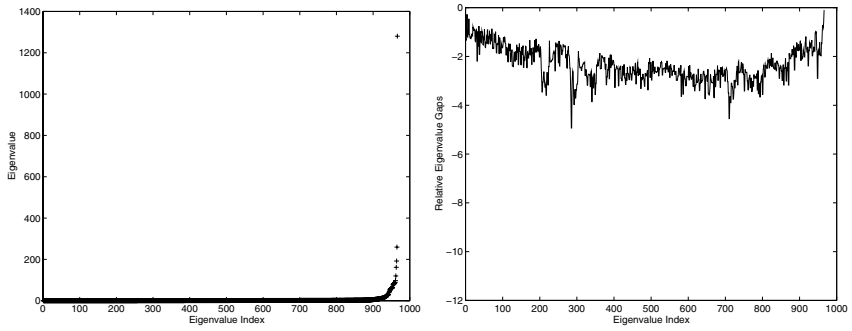


FIG. 7. *Left, eigenvalue distribution; right, relative gaps for the* Biphenyl *matrix.*

type except random $\pm$ signs are placed on the eigenvalues. The fourth type are the Toeplitz tridiagonal matrices with 2's on the diagonals and 1's as the off-diagonal elements, with eigenvalues $\lambda_i = 4\sin^2[\pi i/(2(n+1))]$. The final example comes from a real application in computational quantum chemistry—more specifically it arises in the modeling of the biphenyl molecule using Møller–Plesset theory [9]. Most of the eigenvalues of this positive definite 966×966 Biphenyl matrix are small compared to its norm. See Figure 7 for a plot of the eigenvalues and their relative gaps.

In Table 1 we compare the speed of Algorithm Getvec to various existing algorithms. In our implementation, we factor $T = LDL^t$ and then use the dqds software in LAPACK (subroutine DLASQ1) to compute all eigenvalues of LDL^t to high relative accuracy before invoking Algorithm Getvec to compute all eigenvectors. DSTEBZ is the bisection routine in LAPACK, while DSTEIN and TINVIT are inverse iteration routines from LAPACK and EISPACK, respectively, that perform Gram–Schmidt orthogonalization when eigenvalues have small absolute gaps, in particular, when $|\lambda_{i+1} - \lambda_i| \leq 10^{-3}\|T\|$ (actually TINVIT uses $\max_i |T_{i,i}| + |T_{i,i+1}|$ instead of $\|T\|$ while DSTEIN uses the 1-norm of T). DSTEQR uses the QR iteration to compute orthogonal eigenvectors [19], while DSTEDC is the divide and conquer code in LAPACK [21]. The QR algorithm and divide and conquer method compute eigenvalues and eigenvectors simultaneously, while our strategy and the inverse iteration routines first find the eigenvalues and then the eigenvectors—for these cases Table 1 gives the breakup of the time needed to compute eigenvalues as compared to eigenvectors.

The QR code DSTEQR always takes $O(n^3)$ time irrespective of the eigenvalue distribution. Due to the orthogonalization criterion, DSTEIN and TINVIT take $O(n^3)$ time on matrices of type 1 and the Biphenyl matrix and on large matrices of type 2 and 4. Table 1 shows that on these examples, Algorithm Getvec can be about two orders of magnitude faster than DSTEIN, TINVIT, and DSTEQR. Even on matrices where DSTEIN and TINVIT show $O(n^2)$ behavior, such as matrices of type 3 and matrices of type 2 and 4 with $n \leq 500$, Algorithm Getvec is generally faster. Also see that Algorithm Getvec is several times faster than DSTEDC on four of the five matrix types, and is comparable in speed on the first example, where DSTEDC is very fast due to deflation of clustered eigenvalues. The reader should observe the $O(n^2)$ behavior of Algorithm Getvec, whereas the other subroutines, in general, show an $O(n^3)$ behavior (all timings were measured using Fortran BLAS on a 333-MHz UltraSPARC processor with 1 GByte main memory). All algorithms delivered adequate numerical orthogonality on the test cases.

9. Singular vectors. A natural application of the procedures analyzed in this paper is to compute the SVD of a bidiagonal matrix L^t: $L^t = U\Sigma V^t$, $U^t = U^{-1}$, $V^t = V^{-1}$. Since $LL^t = V\Sigma^2 V^t$, the Cholesky factor of the symmetric positive definite matrix LL^t is the initial input and so the output of our method is V whose columns are the right singular vectors of L^t.

What must be done to compute U? The tempting formula

$$\begin{aligned} \boldsymbol{u} &= L^t\boldsymbol{v}/\sigma, \quad \sigma \neq 0, \\ \text{solve } \ L\boldsymbol{u} &= \mathbf{0}, \qquad\quad \sigma = 0, \end{aligned}$$

is well known to be treacherous. Orthogonal $\boldsymbol{v}$'s do not give rise to orthogonal $\boldsymbol{u}$'s because of the cancellation in forming $L^t\boldsymbol{v}$.

A better way is to invoke Algorithm Getvec again, as shown below. Note that a natural operation on bidiagonal and diagonal arrays is to "flip" them: $L \longrightarrow\sim L$. In practice the order of the entries is reversed. Formally

$$\sim L = \tilde{I}L^t\tilde{I},$$

where $\tilde{I}$ is the reversal matrix, $\tilde{I} = (\boldsymbol{e}_n, \ldots, \boldsymbol{e}_1)$ when $I = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$. For diagonal matrices, flipping is just reversal. If cost were of no consequence, then U could be computed by flipping the given L^t, calling our algorithm, and reversing the output. The justification is that

$$\begin{aligned} (\sim L)(\sim L^t) &= (\tilde{I}L^t\tilde{I})(\tilde{I}L^t\tilde{I})^t \\ &= \tilde{I}L^tL\tilde{I} = \tilde{I}U\Sigma^2U^t\tilde{I}. \end{aligned}$$

The reversal mechanism needs to be applied locally. When an eigenvalue (σ^2) has been computed Algorithm Getvec invokes Algorithms 4.2 and 4.4 to obtain a double factorization and, after selecting an index, the desired singularity-revealing twisted factorization. From this comes the singular vector $\boldsymbol{v}$. In order to compute $\boldsymbol{u}$ it is only necessary to reverse L, apply Algorithms 4.2 and 4.4 again, select a possibly different index, and form the corresponding twisted factorization. Then Algorithm Getvec, in section 4, will yield $\{\tilde{I}\boldsymbol{u}\}$. In other words very little extra code is needed in order to compute $\boldsymbol{u}$ as well as $\boldsymbol{v}$.

However, even the use of Getvec outlined in the above paragraph is *not* adequate. It produces matrices U and V that are orthogonal to working precision, but the extra coupling relations $\|L^t\boldsymbol{v} - \boldsymbol{u}\sigma\| = O(\varepsilon\|L\|)$ and $\|L\boldsymbol{u} - \boldsymbol{v}\sigma\| = O(\varepsilon\|L\|)$ may fail when singular values are clustered.

In recent work [20], Großer and Lang have presented coupling relations that connect factorizations of $LL^t - \mu^2 I$ and $L^tL - \mu^2 I$. By forcing these relations to hold for the computed factorizations they found a way to use Algorithm Getvec and satisfy all the desired properties to working accuracy:

$$L^t\boldsymbol{v} - \boldsymbol{u}\sigma \approx \mathbf{0}, \quad L\boldsymbol{u} - \boldsymbol{v}\sigma \approx \mathbf{0}, \quad U^tU - I \approx 0, \quad V^tV - I \approx 0.$$

This algorithm is to become part of the LAPACK library.

Acknowledgment. We would like to thank an anonymous referee for an extraordinarily detailed reading of our original manuscript and for several constructive suggestions that, at the cost of considerable delay, greatly improved the presentation of this paper.

REFERENCES

[1] ANSI/IEEE, *IEEE Standard for Binary Floating Point Arithmetic*, Std 754-1985 ed., New York, 1985.

[2] J. Barlow, B. Parlett, and K. Veselić, eds., Linear Algebra Appl., 309, no. 1–3 (2000), pp. 1–361.

[3] A. W. Bojanczyk, R. Onn, and A. O. Steinhardt, *Existence of the hyperbolic singular value decomposition*, Linear Algebra Appl., 185 (1993), pp. 21–30.

[4] J. R. Bunch, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[5] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt, *Maple* V *Library Reference Manual*, Springer-Verlag, Berlin, 1991.

[6] L. S. DeJong, *Towards a formal definition of numerical stability*, Numer. Math., 28 (1977), pp. 211–220.

[7] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[8] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[9] I. S. Dhillon, *A New* $O(n^2)$ *Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Ph.D. thesis, Computer Science Division, University of California, Berkeley, California, 1997. Available as UC Berkeley Technical Report UCB//CSD-97-971.

[10] I. S. Dhillon and B. N. Parlett, *Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices*, Linear Algebra Appl., to appear.

[11] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[12] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation bounds for eigenspaces and singular vector subspaces*, in Proceedings of the Fifth SIAM Conference on Applied Linear Algebra, J. G. Lewis, ed., SIAM, 1994, pp. 62–66.

[13] K. Fernando and B. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

[14] K. V. Fernando, *On computing an eigenvector of a tridiagonal matrix. Part* I: *Basic results*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1013–1034.

[15] J. G. F. Francis, *The QR transformation, part* I, Comput. J., 4 (1961–62), pp. 265–271.

[16] J. G. F. Francis, *The QR transformation, part* II, Comput. J., 4 (1961–62), pp. 332–345.

[17] S. K. Godunov, A. G. Antonov, O. P. Kiriljuk, and V. I. Kostin, *Guaranteed Accuracy in Numerical Linear Algebra*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

[18] S. K. Godunov, V. I. Kostin, and A. D. Mitchenko, *Computation of an eigenvector of symmetric tridiagonal matrices*, Siberian Math. J., 26 (1985), pp. 71–85.

[19] A. Greenbaum and J. Dongarra, *Experiments with QL/QR Methods for the Symmetric Tridiagonal Eigenproblem*, Computer Science Dept. Technical Report CS-89-92, University of Tennessee, Knoxville, 1989. (LAPACK Working Note 17, available electronically on netlib.)

[20] B. Grosser and B. Lang, *An* $O(n^2)$ *algorithm for the bidiagonal SVD*, Linear Algebra Appl., 358 (2002), pp. 45–70.

[21] M. Gu and S. C. Eisenstat, *A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 172–191.

[22] P. Henrici, *The quotient-difference algorithm*, Nat. Bur. Standards Appl. Math. Series, 19 (1958), pp. 23–46.

[23] P. Henrici, *Applied and Computational Complex Analysis*, Vol. I, John Wiley & Sons, New York, 1974.

[24] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[25] E. R. Jessup and I. Ipsen, *Improving the accuracy of inverse iteration*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 550–572.

[26] V. N. Kublanovskaya, *On some algorithms for the solution of the complete eigenvalue problem*, Zh. Vychisl. Mat., 1 (1961), pp. 555–570.

[27] R.-C. Li, *Relative perturbation theory:* II. *Eigenspace and singular subspace variations*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 471–492.

[28] B. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, 1997.

[29] B. N. Parlett, *Invariant subspaces for tightly clustered eigenvalues of tridiagonals*, BIT, 36 (1996), pp. 542–562.

[30] B. N. Parlett, *Spectral sensitivity of products of bidiagonals*, Linear Algebra Appl., 275/276 (1998), pp. 417–431.
[31] B. N. Parlett and I. S. Dhillon, *Fernando's solution to Wilkinson's problem: An application of double factorization*, Linear Algebra Appl., 267 (1997), pp. 247–279.
[32] B. N. Parlett and I. S. Dhillon, *Relatively robust representations of symmetric tridiagonals*, Linear Algebra Appl., 309 (2000), pp. 121–151.
[33] B. N. Parlett, *A bidiagonal matrix determines its hyperbolic SVD to (varied) relative accuracy*, SIAM J. Matrix Anal. Appl., submitted.
[34] H. Rutishauser, *Der Quotienten-Differenzen-Algorithmus*. Z. Angew. Math. Phys., 5 (1954), pp. 223–251.
[35] H. Rutishauser, *Vorlesungen über numerische Mathematik*, Birkhäuser, Basel, 1976.
[36] H. Rutishauser, *Lectures on Numerical Mathematics*, Birkhäuser, Boston, 1990.
[37] S. Wolfram, *Mathematica: A System for Doing Mathematics by Computer*, 2nd ed., Addison-Wesley, Reading, MA, 1991.
[38] Y. Yang, *Error Analysis of the dqds Algorithm*, Ph.D. thesis, University of California, Berkeley, California, 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 3, pp. 900–900

ERRATUM: SUCCESSIVELY ORDERED ELEMENTARY BIDIAGONAL FACTORIZATION*

CHARLES R. JOHNSON†, D. D. OLESKY‡, AND P. VAN DEN DRIESSCHE§

DOI. 10.1137/S0895479803435269

An error in the statement of Theorem 15 in [1] was recently discovered. That statement is correct if the matrix A is a unit lower triangular matrix, as was assumed in an earlier version, but is not true in general. A correct statement of the theorem is as follows.

THEOREM 15. *Let A be an n-by-n nonsingular matrix with the generic factorization*

$$A = \left(\prod_{k=1}^{n-1} \prod_{j=n}^{k+1} L_j(s_{jk}) \right) D \left(\prod_{k=n-1}^{1} \prod_{j=k+1}^{n} U_j(t_{kj}) \right).$$

Then for $1 \le k \le n-1$,

$$s_{k+1,1} = \frac{a_{k+1,1}}{a_{k1}},$$

and for $2 \le q \le n-1$ and $1 \le k \le n-q$,

$$s_{k+q,q} = \frac{1}{P_{kq}} \, \frac{\det(A[k+1,\ldots,k+q|1,\ldots,q])}{\det(A[k+1,\ldots,k+q-1|1,\ldots,q-1])},$$

where

$$P_{1q} = \frac{\det A[1,\ldots,q|1,\ldots,q]}{\det A[1,\ldots,q-1|1,\ldots,q-1]}$$

and for $k \ge 2$, $P_{kq} = s_{k+q-1,q} \cdots s_{q+1,q} \, P_{1q}$. Similar formulae hold for $t_{q,k+q}$ with the rows and columns of the minors of A interchanged.

Note that if A is a unit lower triangular matrix, then $P_{1q} = 1$.

Another correction in [1] is as follows:
$\binom{n}{2}$ in line 6 of the paragraph after Example 1 on page 1080 should be $\binom{n+1}{2}$.

Acknowledgment. The authors wish to thank Nelli Marselina Siktimu, Jakarta, Indonesia for inquiring about the statement of the above theorem.

REFERENCE

[1] C. R. JOHNSON, D. D. OLESKY, AND P. VAN DEN DRIESSCHE, *Successively ordered elementary bidiagonal factorization*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1079–1088.

*Received by the editors September 18, 2003; accepted for publication by H. van der Vorst October 10, 2003; published electronically March 30, 2004.
http://www.siam.org/journals/simax/25-3/43526.html

†Department of Mathematics, College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (crjohnso@math.wm.edu).

‡Department of Computer Science, University of Victoria, Victoria, BC V8W 3P6, Canada (dolesky@cs.uvic.ca).

§Department of Mathematics and Statistics, University of Victoria, Victoria, BC V8W 3P4, Canada (pvdd@math.uvic.ca).

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 901–920

LOW-RANK APPROXIMATIONS WITH SPARSE FACTORS II: PENALIZED METHODS WITH DISCRETE NEWTON-LIKE ITERATIONS*

ZHENYUE ZHANG†, HONGYUAN ZHA‡, AND HORST SIMON§

Abstract. In [*SIAM J. Matrix Anal. Appl.*, 23 (2002), pp. 706–727], we developed numerical algorithms for computing *sparse* low-rank approximations of matrices, and we also provided a detailed error analysis of the proposed algorithms together with some numerical experiments. The low-rank approximations are constructed in a certain factored form with the degree of sparsity of the factors controlled by some user-specified parameters. In this paper, we cast the sparse low-rank approximation problem in the framework of penalized optimization problems. We discuss various approximation schemes for the penalized optimization problem which are more amenable to numerical computations. We also include some analysis to show the relations between the original optimization problem and the reduced one. We then develop a globally convergent *discrete* Newton-like iterative method for solving the approximate penalized optimization problems. We also compare the reconstruction errors of the sparse low-rank approximations computed by our new methods with those obtained using the methods in the earlier paper and several other existing methods for computing sparse low-rank approximations. Numerical examples show that the penalized methods are more robust and produce approximations with factors which have fewer columns and are sparser.

Key words. low-rank matrix approximation, singular value decomposition, sparse factorization, perturbation analysis

AMS subject classifications. 15A18, 15A23, 65F15, 65F50

DOI. 10.1137/S0895479801394477

1. Introduction.

Low-rank approximations of matrices have many applications in information retrieval, data mining, and solving ill-posed problems, to name a few [5, 7]. The theory of singular value decomposition (SVD) provides the best rank-k approximation, $\text{best}_k(A)$, of a given $m \times n$ matrix A in terms of its singular values and singular vectors:

$$\text{best}_k(A) \equiv U_k \Sigma_k V_k^T = [u_1, \ldots, u_k] \, \text{diag}(\sigma_1, \ldots, \sigma_k) [v_1, \ldots, v_k]^T,$$

where $\sigma_i, i = 1, \ldots, k$, are the k largest singular values of A, and u_i and v_i are the corresponding left and right singular vectors [1]. Notice that, even when A is sparse, there is in general no guarantee that $\text{best}_k(A)$ will be sparse, not even the factors U_k and V_k. To remedy this drawback of the low-rank approximations computed by

*Received by the editors August 28, 2001; accepted for publication (in revised form) by D. P. O'Leary October 24, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/simax/25-4/39447.html

†Department of Mathematics, Zhejiang University, Yu-Quan Campus, Hangzhou, 310027, People's Republic of China (zyzhang@zju.edu.cn). The work of this author was supported in part by NSFC (project 60372033), the Special Funds for Major State Basic Research Projects of China (project G19990328), and the Foundation for University Key Teacher by the Ministry of Education, China. This work also was supported in part by NSF grant CCR-9619452 and by the Director, Office of Science, Office of Laboratory Policy and Infrastructure Management, of the U.S. Department of Energy under contract DE-AC03-76SF00098.

‡Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 (zha@cse.psu.edu). The work of this author was supported in part by NSF grant CCR-9901986.

§National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, One Cyclotron Road, M/S: 50B, Berkeley, CA 94720 (HDSimon@lbl.gov). The research of this author was supported by the Director, Office of Science, Office of Laboratory Policy and Infrastructure Management, of the U.S. Department of Energy under contract DE-AC03-76SF00098.

SVD, it was proposed to construct low-rank approximations B_k of A in a factored form $B_k = X_k D_k Y_k^T$ without imposing the orthogonality constraints on columns of the left and right factors X_k and Y_k [2, 4, 6]. In [7] we further developed this idea and proposed to compute $B_k = X_k D_k Y_k^T$, with sparse factors X_k and Y_k, that solves the following optimization problem:

$$\min\{\|A - X_k D_k Y_k^T\|_F \mid D_k \text{ diagonal}, X_k \in \mathcal{R}^{m\times k} \text{ and } Y_k \in \mathcal{R}^{n\times k} \text{ sparse}\}. \tag{1.1}$$

In [7], several algorithms are developed for controlling the degree of sparsity of the factors X_k and Y_k, and a detailed error analysis of our proposed algorithms is given that compares the computed sparse low-rank approximations with those obtained from SVD and some of the previous methods developed in [2, 6]. The basic idea is to use a sequence of rank-one deflation steps to construct the approximation

$$B_k = X_k D_k Y_k^T = [x_1, \ldots, x_k]\,\mathrm{diag}(d_1, \ldots, d_k)[y_1, \ldots, y_k]^T = \sum_{i=1}^{k} x_i d_i y_i^T.$$

At each deflation step, approximations to the largest left and right singular vectors $\hat{u}_i$ and $\hat{v}_i$ of the deflated matrix $A_{i-1} = A - B_{i-1}$ are used to construct a sparse rank-one approximation $x_i d_i y_i^T$ for matrix A_{i-1}, here $A_0 = A$. Specifically, the sparse vectors x_i and y_i are obtained by discarding small components of $\hat{u}_i$ and $\hat{v}_i$. We proved that if the norm of the vector consisting of the discarded components is no greater than $\sqrt{2}\epsilon$ at each step, then the computed sparse low-rank approximation B_k has *reconstruction error*, defined as $\|A - B_k\|_F$, no greater than that of the best rank-k approximation $\mathrm{best}_k(A)$ by a factor $(1 + b_k\epsilon)^{1/2}$, i.e.,

$$\|A - B_k\|_F \le (1 + b_k\epsilon)^{1/2}\|A - \mathrm{best}_k(A)\|_F, \tag{1.2}$$

where $b_1 = \sigma_1^2/\left(\sigma_2^2 + \cdots + \sigma_n^2\right)$ (assuming $n \le m$) and

$$b_k = \frac{\sum_{i=1}^{k}\sigma_i\sigma_{i+1}}{\sum_{i=k+1}^{n}\sigma_i^2} + O(\epsilon), \quad k \ge 2.$$

The tolerance parameter ϵ specified by the user can balance the trade-off between the degree of sparsity and good reconstruction error of the low-rank approximations. We suggested in [7] that the size of the tolerance ϵ_i used at each deflation step be a variable determined by

$$\epsilon_i = \frac{\|A_{i-1}\|_F}{\|A\|_F}\epsilon, \quad i = 1, \ldots, k. \tag{1.3}$$

Numerical results in [7] show that the variable-tolerance scheme (1.3) works better than an alternative constant-tolerance scheme whereby the tolerance is fixed at each step. In general, if we preset the desirable reconstruction error, reducing ϵ in (1.3) will yield an approximation with smaller rank k, but the degree of sparsity of X_k and Y_k tend to increase, while increasing ϵ will have the opposite effect. Moreover, we also observed that the ranks and the degree of sparsity of the low-rank approximations computed by the methods in [7] sometimes can be quite sensitive to the choice of ϵ (ϵ_j); i.e., a slight change of ϵ, having little effect on the reconstruction error, can have a much greater effect on the rank of the low-rank approximation and the degree of sparsity of its factors. This less robust behavior is rather undesirable.

Our goal is to develop more robust methods for sparse low-rank approximations. Our basic idea for the improved algorithms is to use penalty terms to penalize low-rank approximations with factors X_k and Y_k having large numbers of nonzeros. In a rather general framework, we can consider the optimization problem

$$\min\{\lambda_x \mathtt{nnz}(X_k) + \lambda_y \mathtt{nnz}(Y_k) + \lambda_e \|A - X_k D_k Y_k^T\|_F^2 \mid D_k \text{ diagonal}\}, \tag{1.4}$$

where λ_x, λ_y, and λ_e are user-determined penalty parameters; here $\mathtt{nnz}(\cdot)$ denotes the number of nonzeros of a matrix. In essence, we want a low-rank approximation $B_k = X_k D_k Y_k^T$ to have a small reconstruction error $\|A - X_k D_k Y_k^T\|_F$, and at the same time we also penalize those B_k for which the X_k and Y_k factors have large numbers of nonzeros. In particular, we can use a variation of the deflation technique to reduce the problem (1.4) to the problem of finding a sequence of sparse rank-one approximations, and build the low-rank approximation one rank at a time [2, 4, 6, 7]. Therefore, in this paper we focus on the rank-one case of problem (1.4):

$$\min_{x,d,y}\{\lambda_x \mathtt{nnz}(x) + \lambda_y \mathtt{nnz}(y) + \lambda_e \|A - x d y^T\|_F^2\}. \tag{1.5}$$

This is an optimization problem with an objective function over $(m+n+1)$-dimensional space, assuming the matrix A has m rows and n columns. We can certainly use some general techniques to solve the optimization problem (1.5). Because of the large dimensionality and the unsmooth property of the objective function, general techniques for solving optimization problems will be not efficient and cost much for the optimization problem (1.5). However, the problem itself possesses many useful structures that deserve exploitation. The approach proposed in this paper is to approximately reduce the $(m + n + 1)$-dimensional optimization problem (1.5) to a much simpler one-dimensional (1-D) discrete problem by several reduction steps, each of which involves a certain tight approximation. Indeed, some structures of the 1-D discrete optimization problem can be further exploited by a continuous-discrete relaxation technique (smoothly interpolating and discretely retransforming), so that it can be solved easily by using a discrete Newton-like iteration (DNI). We will discuss the continuous-discrete relaxation technique and derive the discrete Newton-like iteration in this paper.

The rest of the paper is organized as follows. In section 2, we first examine the rank-one case of the penalized optimization problem (1.5) and then, following the idea of constructing sparse rank-one approximations proposed in [7], we discuss the 1-D discrete optimization problem that will be solved eventually. To show the relations between the 1-D problem actually solved and the original problem (1.5), we describe in section 3 the reduction process in several steps, giving the motivation behind each step and showing the degree of approximation involved with respect to (1.5) by theoretical analysis together with numerical experiments. The continuous-discrete relaxation technique mentioned above is discussed in section 4. In section 5, we propose a *discrete* globally convergent Newton-like method for solving the resulting discrete optimization problem. Several computational issues arise in the proposed method: We focus on how to compute the Newton-like directions at each iterative step and how to select the next iterate to guarantee the existence of a so-called bracketing interval. In section 6, we present some numerical examples and make comparison with several existing methods for computing sparse low-rank approximations.

2. Penalized optimization problems for sparse rank-one approximations. In this section, we consider the rank-one version (1.5) of the optimization

problem (1.4). To make the penalty parameters clearer, we rewrite (1.5) in the format

$$\min_{x,y,d}\left\{\lambda\left(\mu\frac{\mathtt{nnz}(x)}{m}+(1-\mu)\frac{\mathtt{nnz}(y)}{n}\right)+(1-\lambda)\frac{\|A-xdy^T\|_F^2}{\|A\|_2^2}\right\}, \tag{2.1}$$

where the parameters λ and μ are two user-specified penalty terms chosen in the interval (0, 1) and have the following interpretations: λ balances the degree of sparsity and the reconstruction error of the rank-one approximation, while μ balances the degrees of sparsity of the left and right factors of the rank-one approximation xdy^T. In general, we choose $\mu = 1/2$ to keep the sparsity structure of the approximation *symmetric* if no other reasons dictate doing otherwise.

It is easy to verify that, for fixed vectors x and y,

$$\min_d \|A-xdy^T\|_F^2 = \|A\|_F^2 - d^2(x,y),$$

where

$$d(x,y)=\frac{x^TAy}{\|x\|\,\|y\|}. \tag{2.2}$$

Therefore, ignoring a constant factor, (2.1) can be written in the form

$$\min_{x,y}\left\{\lambda\left(\mu\frac{\mathtt{nnz}(x)}{m}+(1-\mu)\frac{\mathtt{nnz}(y)}{n}\right)-(1-\lambda)\frac{d^2(x,y)}{\|A\|_2^2}\right\}. \tag{2.3}$$

For simplicity, let us introduce

$$\alpha=\frac{\lambda\mu}{(1-\lambda)m},\quad \beta=\frac{\lambda(1-\mu)}{(1-\lambda)n} \tag{2.4}$$

and rewrite the optimization problem (2.3) in the following equivalent form:

$$G(\alpha,\beta)=\min_{x,\,y}\left\{\alpha*\mathtt{nnz}(x)+\beta*\mathtt{nnz}(y)-\frac{d^2(x,y)}{\|A\|_2^2}\right\}. \tag{2.5}$$

The objective function of the optimization problem above is a combination of an integer-valued function and a continuous function over an $(m+n)$-dimensional space of continuous variables. Following the idea of constructing sparse rank-one approximations proposed in [7], we consider an approximate solution with x and y constructed from the components of the left and right singular vectors u and v corresponding to the largest singular value of A. To make the presentation self-contained as well as for later reference, we briefly review the basic ideas discussed in [7] for sparse rank-one approximations.

Assume that the largest singular vectors u and v of A are such that $\|u\|_2 = \|v\|_2 = 1$. Let $\tilde{w}$ be the sorted vectors of the vector $w = [u^T, v^T]^T$ in nonincreasing order,

$$|\tilde{w}(1)| \ge |\tilde{w}(2)| \ge \cdots \ge |\tilde{w}(m+n)|.$$

For a given $\xi > 0$, let $k = k(\xi) \le m+n$ be the smallest integer satisfying

$$\tilde{w}^2(1)+\tilde{w}^2(2)+\cdots+\tilde{w}^2(k)\ge 2(1-\xi),$$

and also let $i = i(\xi)$ and $j = j(\xi) = k(\xi) - i(\xi)$ be, respectively, the numbers of the u-components and the v-components in the subset $\{\tilde{w}(1), \ldots, \tilde{w}(k)\}$. Then denote by x_u the sparse vector consisting of the i u-components in the same component positions and zeros elsewhere. Similarly, denote by y_v the sparse vector consisting of the j v-components in the same component positions and zeros elsewhere. Clearly, if P_u and P_v are the permutations determined by sorting w to $\tilde{w}$ such that $\tilde{u} = P_u u$ and $\tilde{v} = P_v v$ are in nondecreasing orders in absolute value, then

$$x_u = P_u^T \begin{bmatrix} \tilde{u}(1:i) \\ 0 \end{bmatrix}, \quad y_v = P_v^T \begin{bmatrix} \tilde{v}(1:j) \\ 0 \end{bmatrix}. \tag{2.6}$$

In [7], we prove that if $\xi \le 1/3$, then

$$\|A - x_u d(x_u, y_v) y_v^T\|_F^2 \le (1 + b_1 \tau) \|A - u\sigma v^T\|_F^2, \tag{2.7}$$

where $b_1 = \sigma_1^2 / \left(\sigma_2^2 + \cdots + \sigma_n^2\right)$ and $\tau = 4\xi\left(1 - \xi^2/(1-\xi)^2\right) < 4\xi$. Note that $\|A - u\sigma v^T\|_F$ is the smallest reconstruction error for any rank-one (sparse or dense) approximation of A. In [7], the degree of sparsity of the rank-one approximation $xd(x,y)y^T$, or the integer k controlled by the parameter ξ, is closely related to the reconstruction error since

$$\|x_u - u\|_2^2 + \|y_v - v\|_2^2 = \tilde{w}^2(k+1) + \cdots + \tilde{w}^2(m+n) \le 2\xi.$$

In general, if ξ is relatively small, we cannot expect to generate rank-one approximations with left and right vectors having many zero entries. On the other hand, if ξ is relatively large, the vectors x_u and y_v can be chosen to be very sparse, but $x_u d(x_u, y_v) y_v^T$ may not be a good rank-one approximation, compared with the best rank-one approximation $u\sigma v^T$.

It turns out that the parameter ξ can be approximately determined by the optimal solution to the following penalized optimization problem:

$$F(\alpha, \beta, p) = \min_{\xi} \left\{ \alpha * i(\xi) + \beta * j(\xi) - \left(\frac{|h(\xi)|}{\|A\|_2} \right)^p \right\}, \tag{2.8}$$

where $h(\xi) = d(x_u, y_v)$ for the sparsified vectors x_u and y_v that depend on the parameter ξ with $i(\xi)$ and $j(\xi)$ defined as above.

Comparing with the $(n+m)$-dimensional problem (2.5), the 1-D optimization problem (2.8) is much simpler. It can be derived from the original optimal problem (2.5) using several reduction steps. In the next section, we will describe the reduction process. To solve the optimization problem (2.8), we will present a discrete Newton-like method in section 5.

3. Relations between $G(\alpha, \beta)$ and $F(\alpha, \beta, p)$. We write (2.5) in a general form

$$\min_{x,y} \left\{ \alpha \mathtt{nnz}(x) + \beta \mathtt{nnz}(y) - \left(\frac{|d(x,y)|}{\|A\|_2} \right)^p \right\} \tag{3.1}$$

for $p > 0$. Clearly, the space $\mathcal{R}^m$ of m-dimensional vectors can be divided into m sets $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_m$ such that each vector in $\mathcal{U}_i$ has i nonzeros. Similar partition also holds for the space $\mathcal{R}^n$ of n-dimensional vectors. In the case when such a set $\mathcal{U}_i$ of m-dimensional vectors and a set $\mathcal{V}_j$ of n-dimensional vectors are fixed, the objective function of the optimization problem above for $x \in \mathcal{U}_i$ and $y \in \mathcal{V}_j$ is minimized by

maximizing $|d(x,y)|$ over the two sets $\mathcal{U}_i$ and $\mathcal{V}_j$. This property explicitly exhibits a *combinatorial* nature of the above optimization problem. To illustrate it in a clearer fashion, for $x \in \mathcal{U}_i$ and $y \in \mathcal{V}_j$, denote by I and J the index sets defined by the corresponding nonzero components of vectors x and y, respectively. Then the size of I is $|I| = i$, and the nonzero section of x can be represented by $x(I)$; similarly with $y(J)$ for y and $|J| = j$. We also use the notation $A(I,J)$ to denote the submatrix of A, the elements of which are at the intersection of the rows and columns in I and J. Therefore

$$|d(x,y)| = \left| \frac{x^T(I)A(I,J)y(J)}{\|x(I)\|_2 \|y(J)\|_2} \right| \le \|A(I,J)\|_2,$$

and the minimum of the objective function in (3.1) over $\mathcal{U}_i \times \mathcal{V}_j$ is given by $\alpha * i + \beta * j - (s_{i,j}/\|A\|_2)^p$ with

$$s_{ij} = \max_{|I|=i,\,|J|=j} \|A(I,J)\|_2. \tag{3.2}$$

It follows that the optimization problem (3.1) can be written equivalently in the following form:

$$G_S(\alpha,\beta,p) = \min_{i,j} \left\{ \alpha * i + \beta * j - \left(\frac{s_{ij}}{\|A\|_2} \right)^p \right\}, \tag{3.3}$$

where $S = (s_{ij}) \in \mathcal{R}^{m\times n}$ and consequently $G(\alpha,\beta) = G_S(\alpha,\beta,2)$.

Compared with (2.5), problem (3.3) is only an optimization problem over a finite 2-D index set. However, the corresponding objective function of (3.3) is more difficult to evaluate when the size of matrix A is large. In the next subsections, we will derive a sequence of nonnegative matrices $H = (h_{i,j})$ that are easier to evaluate than S and additionally satisfy $S \ge H \ge 0$, i.e.,

$$s_{i,j} \ge h_{i,j} \ge 0 \quad \text{for all } i \text{ and } j.$$

This will lead to a sequence of upper bounds for $G_S(\alpha,\beta,p)$ defined by the optimization problem

$$G_H(\alpha,\beta,p) = \min_{i\,j} \left\{ \alpha * i + \beta * j - \left(\frac{h_{ij}}{\|A\|_2} \right)^p \right\}. \tag{3.4}$$

This sequence of upper-bounding will lead to the problem (2.8), which we will eventually solve approximately.

3.1. Approximation using truncated Rayleigh quotients. As denoted in the previous section, let $\{u,\sigma,v\}$ be the largest singular triplet of matrix A. A lower bound of S is the matrix $R = (r_{ij})$ of Rayleigh quotients

$$r_{ij} = \max_{|I|=i,\,|J|=j} \left| \frac{u(I)^T A(I,J)v(J)}{\|u(I)\|\,\|v(J)\|} \right|,$$

which gives rise to the following optimization problem:

$$G_R(\alpha,\beta,p) = \min_{i\,j} \left\{ \alpha * i + \beta * j - \left(\frac{r_{ij}}{\|A\|_2} \right)^p \right\}.$$

This problem can be further reduced by choosing the index sets of I and J in a specific way determined by the *truncated* Rayleigh quotients as follows. Let $\tilde{u}$ and $\tilde{v}$ be the sorted versions of u and v, respectively, such that

$$|\tilde{u}(1)| \geq \cdots \geq |\tilde{u}(m)| \quad \text{and} \quad |\tilde{v}(1)| \geq \cdots \geq |\tilde{v}(n)|,$$

and let π_u and π_v be the permutation index vectors satisfying $\tilde{u} = u(\pi_u)$ and $\tilde{v} = v(\pi_v)$. With I and J defined by

$$I = \pi_u(1:i), \quad J = \pi_v(1:j), \tag{3.5}$$

we obtain a lower bound of r_{ij}, which we denote as

$$t_{ij} = \left| \frac{u(I)^T A(I,J) v(J)}{\|u(I)\| \, \|v(J)\|} \right|.$$

Setting $T = (t_{t_ij})$ yields the following optimization problem:

$$G_T(\alpha, \beta, p) = \min_{ij} \left\{ \alpha * i + \beta * j - \left(\frac{t_{ij}}{\|A\|_2} \right)^p \right\}. \tag{3.6}$$

Obviously, $s_{ij} \geq r_{ij} \geq t_{ij}$ and

$$G_S(\alpha, \beta, p) \leq G_R(\alpha, \beta, p) \leq G_T(\alpha, \beta, p).$$

Now back to $F(\alpha, \beta, p)$ defined in (2.8). It is not difficult to verify that the vectors x_u and y_v defined in (2.6) satisfy

$$x_u(\pi_u) = \begin{bmatrix} \tilde{u}(1:i) \\ 0 \end{bmatrix} = \begin{bmatrix} u(I) \\ 0 \end{bmatrix}, \quad y_v(\pi_v) = \begin{bmatrix} \tilde{v}(1:j) \\ 0 \end{bmatrix} = \begin{bmatrix} v(J) \\ 0 \end{bmatrix},$$

where I and J are those defined in (3.5). By setting $i = i(\xi)$ and $j = j(\xi)$, we have $|d(x_u, y_v)| = t_{ij}$. Therefore, $F(\alpha, \beta, p)$ is the minimal value of the objective function defined in (3.6) over the subset $\{ (i(\xi),\, j(\xi)) \,|\, \xi > 0 \}$.

In the same way, $F(\alpha, \beta, p)$ can be reviewed as a reduced form of the original problem $G_S(\alpha, \beta, p)$ by several reduction steps and thus it can satisfy

$$G_S(\alpha, \beta, p) \leq G_R(\alpha, \beta, p) \leq G_T(\alpha, \beta, p) \leq F(\alpha, \beta, p).$$

The difference between the optimal solutions corresponding to $F(\alpha, \beta, p)$ and $G_S(\alpha, \beta, p)$ will naturally be nonzero. In the next subsection, we will illustrate the quality of $G_R(\alpha, \beta, p)$ and $G_T(\alpha, \beta, p)$ as approximations to $G_S(\alpha, \beta, p)$, and give some numerical results to compare the sparsities of the low-rank approximations obtained based on $G_T(\alpha, \beta, p)$ and $F(\alpha, \beta, p)$. In section 4, we will concentrate on computing $F(\alpha, \beta, p)$ to find approximations to $G_S(\alpha, \beta, p)$.

3.2. How good are the approximations? In the above subsection, we showed that G_R and G_T can be used to provide upper bounds of G_S. The following theorem shows that they can also be used to give lower bounds of G_S.

THEOREM 3.1. *Let c_{RS}, c_{TR}, and c_{TS} be*

$$c_{RS} = \max_{ij} \frac{s_{ij}^p - r_{ij}^p}{s_{ij}^p}, \quad c_{TR} = \max_{ij} \frac{r_{ij}^p - t_{ij}^p}{r_{ij}^p}, \quad c_{TS} = \max_{ij} \frac{s_{ij}^p - t_{ij}^p}{s_{ij}^p}.$$

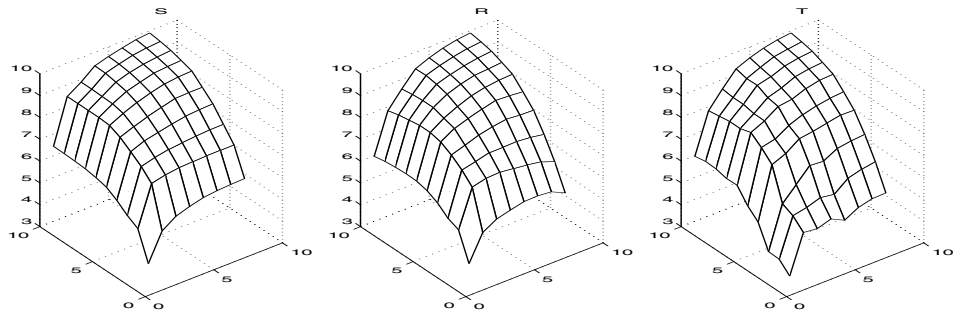


FIG. 1. *Plots of the matrices S, R, and T.*

Then we have $0 \le G_R - G_S \le c_{RS}$, $0 \le G_T - G_R \le c_{TR}$, $0 \le G_T - G_S \le c_{TS}$.

Proof. Denote by (i_S, j_S) the optimal index-pair for the problem (3.3). Then

$$\begin{aligned}
0 &\le G_R(\alpha,\beta,p) - G_S(\alpha,\beta,p) \\
&\le \left(\alpha * i_S + \beta * i_S - \left(\frac{r_{i_S,j_S}}{\|A\|_2}\right)^p\right) - \left(\alpha * i_S + \beta * i_S - \left(\frac{s_{i_S,j_S}}{\|A\|_2}\right)^p\right) \\
&= \left(\frac{s_{i_S,j_S}}{\|A\|_2}\right)^p \left(1 - \left(\frac{r_{i_S,j_S}}{s_{i_S,j_S}}\right)^p\right) \\
&\le \left(\frac{s_{i_S,j_S}}{\|A\|_2}\right)^p c_{RS}.
\end{aligned}$$

The result follows from the inequalities $s_{i,j} \le \|A\|_2$ for all i and j. The proofs for other inequalities are similar. □

The theorem above shows that $G_T(\alpha,\beta,p)$ will be a good approximation to $G_S(\alpha,\beta,p)$ if c_{TS} is small. We are, however, unable to provide a useful bound for c_{TS}. In general T seems to differ from S only slightly. To illustrate this, let us consider a small numerical example. (We have also tested other small matrices and found similar behavior.[1])

Example 1. Let $m = 10$, $n = 8$, and $l = \min(m,n)$. We construct A as (using the notation of MATLAB)

```
[U, r] = qr(rand(m,l));
[V, r] = qr(rand(n,l));
A = U*diag(10*rand(l,1))*V';
```

We take $\mu = 0.5$ and $\lambda = 0.3$ and then construct the three matrices S, R, and T. Figure 1 plots the matrices S, R, and T. In general, s_{ij}, r_{ij}, and t_{ij} are close to each other if i and j are not small; i.e., a large discrepancy in s_{ij}, r_{ij}, and t_{ij} may occur only when the indexes i and j are small. Below we list the average values and the maximums of the relative errors between s_{ij}, r_{ij}, and t_{ij}. Note that the maximums occur generally with small indexes i and j.

Relative error	Average	Max
$(s_{ij} - r_{ij})/s_{ij}$	2.5809e-02	1.0719e-01
$(r_{ij} - t_{ij})/r_{ij}$	4.3807e-02	2.8485e-01

[1] Only small matrices are used in our examples, since computing S, R, and T involves exhaustive search.

On the other hand, for the tested matrices, the index-pair (i_H, j_H) of the problem (3.4) and corresponding values h_{i_H,j_H} change only slightly for different choices of $H = S$, R, T and $p = 1$, 2, respectively. Below we list the computed results.

	$p=1$			$p=2$		
H	i_H	j_H	h_{i_H,j_H}	i_H	j_H	h_{i_H,j_H}
S	6	3	8.72648	7	6	9.50874
R	5	3	8.35768	8	6	9.52802
T	6	4	8.79437	8	6	9.51918

It seems that in general we can expect the following:

$$i_S + j_S \lesssim i_R + j_R \lesssim i_T + j_T \quad \text{and} \quad s_{i_S,j_S} \lesssim r_{i_R,j_R} \lesssim t_{i_T,j_T}.$$

It is difficult to prove the above assertion in a rigorous fashion. We now, however, provide a proof for a weaker form.

THEOREM 3.2. *Using the notation as above, we have*

$$\big(\alpha * i_T + \beta * j_T\big) - \big(\alpha * i_S + \beta * j_S\big) \le \frac{t_{i_T,j_T} - t_{i_s,j_s}}{s_{i_S,j_S}}.$$

Proof. Notice that

$$\begin{aligned}
&\big(\alpha * i_T + \beta * j_T\big) - \big(\alpha * i_S + \beta * j_S\big) \\
&\quad = G_T(\alpha,\beta,p) - G_S(\alpha,\beta,p) - \big(t^p_{i_T,j_T} - s^p_{i_S,j_S}\big)/\|A\|_2^p \\
&\quad \le \big(t^p_{i_T,j_T} - t^p_{i_S,j_S}\big)/\|A\|_2^p \\
&\quad \le \big(t^p_{i_T,j_T} - t^p_{i_S,j_S}\big)/s^p_{i_S,j_S},
\end{aligned}$$

completing the proof. ∎

Similar results can also be proved for r_{i_R,j_R} and t_{i_T,j_T}.

Finally, we show some numerical results to illustrate the difference in sparsity of the low-rank approximations obtained based on the optimal solutions of $G_T(\alpha,\beta,p)$ and $F(\alpha,\beta,p)$. The tested matrices A that we considered are constructed as those in Example 1 with $m = 200$ and $n = 160$. We test 100 matrices and compare the total number of nonzeros in the optimal solutions x and y for $G_T(\alpha,\beta,p)$ and $F(\alpha,\beta,p)$ with $\alpha = 1/(2m)$ and $\beta = 1/(2n)$ corresponding to $\lambda = 0.2$ and $\mu = 0.5$. The difference is quite small. Figure 2 plots the sorted relative difference of the total number of nonzeros for the 100 tested matrices.

4. Continuous relaxation for $F(\alpha,\beta,p)$. Now we discuss how to solve the discrete optimization problem (2.8). Clearly the objective function $f(\xi)$ of problem $F(\alpha,\beta,p)$ has at most $m+n$ different values at $\xi = \xi_k$ defined by

$$\xi_k = 1 - \frac{1}{2}\big(\tilde{w}(1)^2 + \ldots + \tilde{w}(k)^2\big), \quad k = 1, 2, \ldots, (m+n), \tag{4.1}$$

for the sorted vector $\tilde{w}$ of $w = (u^T, v^T)^T$ with computed singular vectors u and v.[2] Therefore, the discrete problem $F(\alpha,\beta,p)$ can be solved by a direct method, shown as the following two steps: (1) evaluate the $m+n$ function values $f(\xi_k)$,

[2]It is not required to evaluate $f(\xi_k)$ if $\tilde{w}(1:k)$ has only u-components or v-components.

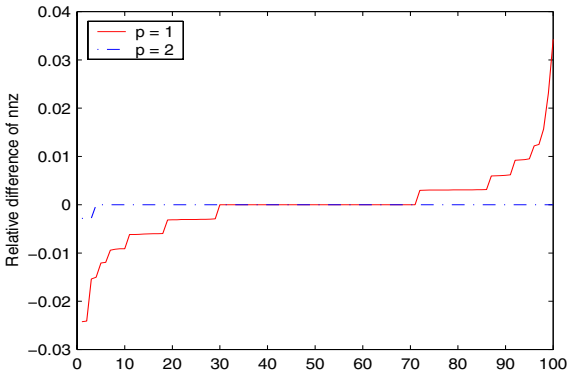


FIG. 2. *Plots of the difference in sparsity of the optimal solutions based on G_T and F for* 100 *test matrices.*

$k = 1, 2, \ldots, (m + n)$, and (2) find the minimum and the optimal ξ_{k^*} or k^* for constructing the sparse vectors $x_u(\xi_{k^*})$ and $y_v(\xi_{k^*})$. Clearly, to evaluate the $m + n$ function values $f(\xi_k)$, $k = 1, 2, \ldots, (m+n)$, $m+n$ vector inner-products are required if some updating techniques are used. Roughly speaking, it needs the same cost as that for $x^T Ay$,[3] together with the cost of checking the u-components or v-components for the components of w.

To solve $F(\alpha, \beta, p)$ efficiently, we will reduce $F(\alpha, \beta, p)$ to a discrete form which does not need the matrix-vector product and can be solved easily using a Newton-like iterative method discussed in the next section. Our strategy is to first implicitly construct a continuous approximation of the discrete optimization problem by smoothly interpolating the discrete objective function in (2.8) and finding the conditions that characterize the solution to the continuous problem. The conditions are then transformed back to the discrete problem that we will solve.

To this end, let $\phi(\xi)$, $\psi(\xi)$, and $\omega(\xi)$ be approximations of the piecewise constant functions $i(\xi)$, $j(\xi)$, and $|h(\xi)|/\|A\|_2$ with $h(\xi) = d(x_u, y_v)$ defined as before, respectively,

$$\phi(\xi) \approx i(\xi), \quad \psi(\xi) \approx j(\xi), \quad \omega(\xi) \approx \frac{|h(\xi)|}{\|A\|_2}.$$

Then we seek to solve the corresponding approximate continuous optimization problem,

$$C(\alpha, \beta) = \min_{\xi} \Big\{ \alpha\phi(\xi) + \beta\psi(\xi) - \omega^p(\xi) \Big\}. \tag{4.2}$$

Obviously, the optimal ξ^* satisfies

$$\alpha\phi'(\xi) + \beta\psi'(\xi) - p\omega^{p-1}(\xi)\omega'(\xi) = 0, \tag{4.3}$$

provided that ϕ, ψ, and ω are differentiable.

[3] In the $(i + 1)$th rank-one iteration of our sparse low-rank approximation method, the matrix corresponding to $F(\alpha, \beta, p)$ is the reconstruction error matrix $A_i = A - X_i D_i Y_i^T$.

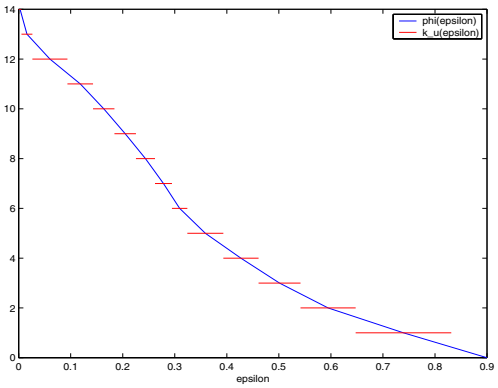


FIG. 3. *Plot for the piece-constant function $i(\xi)$ and its interpolation $\phi(\xi)$.*

Now we show how to construct *smooth* functions that interpolate $i(\xi)$ and $j(\xi)$. For simplicity, we set $p = 1$. We choose functions $\phi(\xi)$ and $\psi(\xi)$ defined by the integral equations

$$\int_0^{\phi(\xi)} f_u(t)dt = 1 - \xi, \quad \int_0^{\psi(\xi)} f_v(t)dt = 1 - \xi,$$

where $f_u(t)$ and $f_v(t)$ are continuous functions interpolating $\{\tilde{u}^2(i)\}$ and $\{\tilde{v}^2(i)\}$, respectively; see Figure 3 for an illustration.

By the proof of Theorem 3.1 in [7], we have

$$1 \geq \frac{|h(\xi)|}{\|A\|_2} \geq 1 - \frac{2\xi}{1-\xi}.$$

Therefore we can choose $\omega(\xi)$ to be the mean of the lower bound and the upper bound,[4]

$$\omega(\xi) = 1 - \frac{\xi}{1-\xi}. \tag{4.4}$$

The functions $\phi(\xi)$ and $\psi(\xi)$ are differentiable because f_u and f_v are continuous, and

$$\phi'(\xi) = -\frac{1}{f_u(\phi(\xi))} \approx -\frac{1}{\tilde{u}^2(i(\xi))}, \quad \psi'(\xi) = -\frac{1}{f_v(\psi(\xi))} \approx -\frac{1}{\tilde{v}^2(j(\xi))},$$

where in the above we have used $f_u(\phi(\xi)) \approx \tilde{u}^2(i(\xi))$ and $f_u(\psi(\xi)) \approx \tilde{v}^2(j(\xi))$. On the other hand, $\omega'(\xi) = -(1-\xi)^{-2}$. Hence the optimal ξ approximately satisfies

$$\frac{\alpha}{\tilde{u}^2(i(\xi))} + \frac{\beta}{\tilde{v}^2(j(\xi))} - \frac{1}{(1-\xi)^2} = 0,$$

or equivalently,

$$1 - \xi = \left(\alpha\tilde{u}^{-2}(i(\xi)) + \beta\tilde{v}^{-2}(j(\xi))\right)^{-1/2}.$$

[4]One can in general choose $\omega(\xi) = 1 - c_0\xi/(1-\xi)$ with a constant $c_0 \in (0, 2)$.

Note that if the optimal ξ satisfies $\xi_k \le \xi < \xi_{k+1}$, where ξ_k is defined in (4.1), then $i(\xi) = i(\xi_k)$, $j(\xi_k) = j(\xi_k)$, and

$$\frac{1}{2}\left(\tilde{w}^2(1) + \cdots + \tilde{w}^2(k)\right) = 1 - \xi_k \approx 1 - \xi.$$

For simplicity, we define $i_k = i(\xi_k)$ and $j_k = j(\xi_k)$. We conclude that the integer k, such that the interval $[\xi_k, \xi_{k+1})$ contains the optimal ξ, approximately satisfies

$$\frac{1}{2}\left(\tilde{w}^2(1) + \cdots + \tilde{w}^2(k)\right) = \left(\alpha\tilde{u}^{-2}(i_k) + \beta\tilde{v}^{-2}(j_k)\right)^{-1/2}.$$

Now we can transform the continuous form (4.3) back to the following discrete optimization problem:

$$D(\alpha, \beta) = \min_k \left|\Phi(k) - \Psi(k)\right|, \tag{4.5}$$

where

$$\Phi(k) = \frac{1}{2}\left(\tilde{w}^2(1) + \cdots + \tilde{w}^2(k)\right), \quad \Psi(k) = \left(\frac{\alpha}{\tilde{u}^2(i_k)} + \frac{\beta}{\tilde{v}^2(j_k)}\right)^{-1/2}.$$

It is not difficult to verify that the integers $i_k \ge 1$ and $j_k \ge 1$ determined by k satisfy the following equations:

$$\begin{cases} i_k + j_k &= k, \\ \min\left\{\tilde{u}^2(i_k), \tilde{v}^2(j_k)\right\} &= \tilde{w}^2(k). \end{cases} \tag{4.6}$$

It should be pointed out that the indexes i_k and j_k may not be uniquely determined by k if $|\tilde{w}_k| = |\tilde{w}_{k+1}|$. We will have a detailed discussion about this in the next section. Ignoring, for the moment, the possibility of those indexes being multivalued, we can easily see that the discrete function $\Psi(k)$ is *decreasing*, while $\Phi(k)$ is *increasing*. Figure 4 plots the graphs of $\Psi(k)$ and $\Phi(k)$ for a matrix of order 2331×1398. Obviously, the monotonicity of $\Phi(k)$ and $\Psi(k)$ is very helpful for constructing a *globally* convergent method to solve the minimization problem (4.5). Such an iterative method will be considered in the next section.

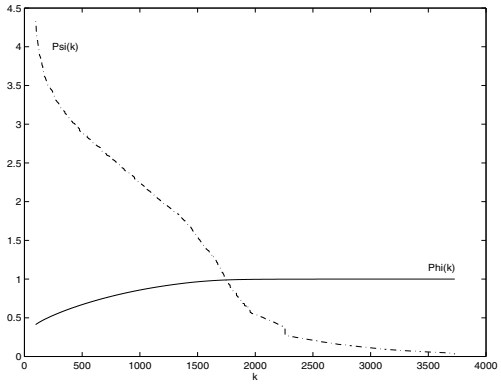


FIG. 4. *Plot for the discrete curves* $\Phi(k)$ *(solid line) and* $\Psi(k)$ *(dashdot line).*

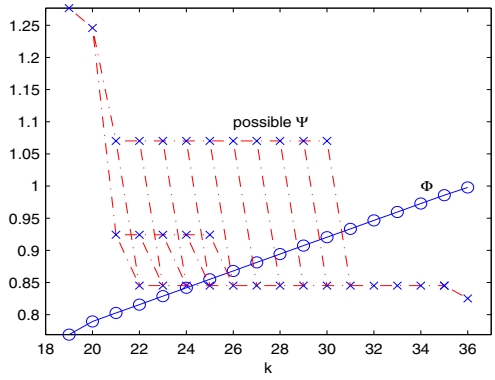


FIG. 5. *The curves $\Phi(k)$ (solid circle line) and the possible curve $\Psi(k)$ (dashed x-mark lines) in a bracketing interval.*

5. Discrete Newton-like iteration. In this section, we discuss a Newton-like iteration for solving the optimization problem (4.5).

Assume we have an approximation k_ℓ to the optimal k^*. Define $\Psi_\ell = \Psi(k_\ell)$ and $\Phi_\ell = \Phi(k_\ell)$. We choose suitably two secants $\delta\Psi_\ell$ and $\delta\Phi_\ell$ for $\Psi(k)$ and $\Phi(k)$, respectively. Replacing $\Psi(k)$ and $\Phi(k)$ in (4.5) by the secant lines $\Psi_\ell + (k - k_\ell)\delta\Psi_\ell$ and $\Phi_\ell + (k - k_\ell)\delta\Phi_\ell$, and solving the problem

$$\min_k |(\Psi_\ell + (k - k_\ell)\delta\Psi_\ell) - (\Phi_\ell + (k - k_\ell)\delta\Phi_\ell)|, \tag{5.1}$$

yields the following DNI:

$$k_{l+1} = k_\ell + \left\lfloor \frac{\Psi_\ell - \Phi_\ell}{\delta\Phi_\ell - \delta\Psi_\ell} \right\rfloor, \tag{5.2}$$

where $\lfloor a \rfloor$ is the floor function giving the largest integer no greater than a. The monotonicity of $\Phi(k)$ and $\Psi(k)$ also ensures the uniqueness of the solution to (4.5).

As mentioned at the end of the last section, we need to investigate whether the objective function for (4.5) is well defined. Note that $\Phi(k)$ is always well defined.

If $|\tilde{w}(k)| > |\tilde{w}(k+1)|$, i_k and j_k are uniquely determined by k; i_k is the number of the u-components of subvector $\tilde{w}(1:k)$, and j_k is the number of the v-components of $\tilde{w}(1:k)$. In this case, $\Psi(k)$ is also well defined. If $|\tilde{w}(k)| = |\tilde{w}(k+1)|$, i_k and j_k may be not unique, and then $\Psi(k)$ may have different values depending on the choices of i_k and j_k; see Figure 5 for a detailed illustration. There are several important computational details that we need to discuss for the DNI.

w-constant intervals. We call (a, b) a w-constant interval if $b - a > 1$ and

$$|\tilde{w}(a)| > |\tilde{w}(a+1)| = \cdots = |\tilde{w}(b)| > |\tilde{w}(b+1)|.$$

Obviously, if all components in subvector $\tilde{w}(a+1:b)$ are u-components or all components are v-components, $\Psi(k)$ is well defined for $k \in [a+1, b-1]$. Otherwise, if $\tilde{w}(a+1:b)$ has both u-components and v-components, then for any integer $k \in [a+1, b-1]$, i_k can be any integer in the interval $[i_a, \min(i_b, k - j_a)]$ and $j_k = k - i_k \in [\max(j_a, k - i_b), k - i_a]$. This will lead to ill-defining of $\Psi(k)$ because

$\Psi(k)$ may have multiple values depending on the choice of i_k and j_k if $\alpha \neq \beta$. However, we can ignore the multidefinition of $\Psi(k)$ if $k^* \notin [a, b]$ because in this case the choice of pair (i_k, j_k) does not affect the convergence of the DNI.

Bracketing intervals. We should pay more attention to w-constant interval $(a, b]$ if the optimal $k^* \in [a+1, b-1]$. Such a w-constant interval is called a bracketing interval. For a bracketing interval $(a, b]$, the optimal k^* should be $k^* = i^* + j^*$ with i^* and j^* solving

$$\min_{i_a \le i \le i_b,\, j_a \le j \le j_b} \big|\Phi(i+j) - \Psi(i+j)\big|. \tag{5.3}$$

It is easy to see that $\Phi(k)$ is uniquely defined and is a linear function in the interval (a, b), while $\Psi(k)$ will have multiple values depending on i_k and j_k. Figure 5 plots the graph of $\Phi(k)$ and many possible graphs for $\Psi(k)$ in a bracketing interval.

The existence of bracketing intervals makes problem (4.5) more complicated because it is necessary to solve (5.3) with possible multivalued $\Psi(k)$ for each $k \in [a+1, b-1]$. Note that there are no easy methods for solving (5.3) except for exhaustive enumeration. Fortunately, bracketing intervals seldom occur, and the length of any bracketing interval that does occur is generally very small. For all the numerical experiments we did, the largest length of the bracketing interval seen was $b - a = 3$.

Choosing the secants $\delta\Phi_\ell$ and $\delta\Psi_\ell$. There are many ways to choose the secants $\delta\Phi_\ell$ and $\delta\Psi_\ell$, which are to be used to construct the next iterate. Let k_0 be the smallest integer such that $\min(i_{k_0}, j_{i_0}) = 1$. For the initial $\delta\Phi_0$ and $\delta\Psi_0$, we set

$$\delta\Phi_0 = \frac{\Phi(k_0 + d) - \Phi(k_0)}{d} \quad \text{and} \quad \delta\Psi_0 = \frac{\Psi(k_0 + d) - \Psi(k_0)}{d},$$

with a step size $d \le \frac{1}{2}(m + n - k_0)$; for example, we have used $d = \min(100, \lfloor \frac{1}{2}(m + n - k_0) \rfloor)$. In general, one can choose $\delta\Phi_\ell = (\Phi_\ell - \Phi_{l-1})/(k_\ell - k_{l-1})$ and $\delta\Psi_\ell = (\Psi_\ell - \Psi_{l-1})/(k_\ell - k_{l-1})$. However, a better way is to compute $\delta\Phi_\ell$ and $\delta\Psi_\ell$ by

$$\delta\Phi_\ell = \frac{\Phi_{\max} - \Phi_{\min}}{k_{\max} - k_{\min}} \quad \text{and} \quad \delta\Psi_\ell = \frac{\Psi_{\max} - \Psi_{\min}}{k_{\max} - k_{\min}}$$

if we know an interval $[k_{\min}, k_{\max}]$ that contains the optimal k^*, where

$$\Phi_{\min} = \Phi(k_{\min}), \quad \Phi_{\max} = \Phi(k_{\max}),$$
$$\Psi_{\min} = \Psi(k_{\min}), \quad \Psi_{\max} = \Psi(k_{\max}).$$

Initially, we can set $k_{\min} = k_0$ and $k_{\max} = m + n$. The interval $[k_{\min}, k_{\max}]$ will be updated at each iteration step as follows. If $\Psi_\ell < \Phi_\ell$, we then decrease $k_{\max}$ and reset $k_{\max} = k_\ell$; otherwise if $\Psi_\ell \le \Phi_\ell$, we increase $k_{\min}$ and reset $k_{\min} = k_\ell$.

Avoiding infinite loop. To avoid infinite loop of the Newton-like iteration in the case when $\lfloor \frac{\Psi_\ell - \Phi_\ell}{\delta\Phi_\ell - \delta\Psi_\ell} \rfloor = 0$, we need to slightly modify k_{l+1} at the lth iteration so that

$$\begin{array}{rcccll} k_\ell + 1 & \le & k_{l+1} & \le & k_{\max} - 1 & \text{if} \quad k_\ell < k^*, \quad \text{or} \\ k_{\min} + 1 & \le & k_{l+1} & \le & k_\ell - 1 & \text{if} \quad k_\ell > k^*. \end{array} \tag{5.4}$$

It is not difficult to check whether the inequality $k_\ell \le k^*$ holds for given k_ℓ by comparing Φ_ℓ and Ψ_ℓ. Now we are ready to present our algorithm for solving the discrete optimization problem (4.5).

The convergence property of Algorithm DNI is relatively easy to establish. Notice that the length of the interval $[k_{\min}, k_{\max}]$ will be reduced by at least one at each iteration. Therefore the above algorithm is guaranteed to converge in at most $m + n - k_0$ iterations. For all the numerical experiments we did, Algorithm DNI converged within about 10 iterations on average.

Algorithm DNI (discrete Newton-like iteration). Given two vectors u and v, this algorithm solves the minimization problem (4.5).

1. [Initialization]
- **1.1** Sort $\tilde{w} = Pw$, $\tilde{w}(k) \leftarrow \tilde{w}(k)^2$, $\tilde{u}(i) \leftarrow \tilde{u}(i)^2$, $\tilde{v}(j) \leftarrow \tilde{v}(j)^2$.
- **1.2** Determine the smallest k_0 such that $i_{k_0} \geq 1$, $j_{k_0} \geq 1$, and set an initial trust interval $[k_{\min}, k_{\max}]$.

2. For $\ell = 0, 1, 2, \ldots,$ until convergence,
- **2.1** Compute $\Phi_\ell = \Phi(k_\ell)$ and $\Psi_\ell = \Psi(k_\ell)$.
- **2.2** Modify the trust interval $[k_{\min}, k_{\max}]$ and check convergence. If $k_{\min} = k_{\max}$, or $|\Psi_\ell - \Phi_\ell| < \tau$, stop.
- **2.3** Determine the secants $\delta\Phi_0$ and $\delta\Psi_0$.
- **2.4** One Newton iteration, $k_{\ell+1} = k_\ell + \lfloor \frac{\Psi_\ell - \Phi_\ell}{\delta\Phi_\ell - \delta\Psi_\ell} \rfloor$. Slightly modify k_{l+1} as in (5.4) if necessary.
- **2.5** Determine the w-constant interval $\big(a, b\big]$ which covers k, and compute $\Psi_a = \Psi(a)$, $\Psi_b = \Psi(b)$, $\Phi_a = \Phi(a)$, $\Phi_b = \Phi(b)$.
- **2.6** If $\big[a, b\big]$ does not cover the optimal k^*, compute $k_{\ell+1}$ by

$$k_{\ell+1} = \begin{cases} a & \text{if } \Psi_a \leq \Phi_a, \\ b & \text{if } \Psi_b > \Phi_b; \end{cases}$$

otherwise turn to step 3.

3. Compute $k^* \in \big[a, b\big]$.

The iterative method DNI is very efficient. In Table 1, we list the computation costs of Algorithm DNI and the direct method for solving (2.8) with $\lambda = 0.1$ when they were used, respectively, together with the sparse low-rank approximation algorithm (SLRA) for computing sparse low-rank approximations of the matrix A. Each sparse low-rank approximation consists of 83 rank-one sparse approximations to $A_0 = A$ and the reconstruction error matrices $A_i = A - X_i D_i Y_i^T$ of the previously computed sparse approximations $X_i D_i Y_i^T$, $i = 1, 2, \ldots, 82$.

TABLE 1
Computation cost of Algorithm DNI and the direct method for matrix `cisi`.

	Min flops	Max flops	Mean flops	Total flops	Total CPU(s)
DNI	9.1500e+3	4.1130e+4	2.6966e+4	2.2382e+6	1.7600e+0
Direct method	2.4500e+5	2.4546e+5	2.4525e+5	2.0110e+7	9.3340e+2

Combining Algorithm DNI with SLRA, we have the following overall penalized algorithm for computing spare low-rank approximations (see next page).

6. Numerical experiments. In this section, we will present several numerical experiments to illustrate the penalized SLRA with DNI (SLRA-DNI). We will compare SLRA-DNI with the SLRA using a strategy of mixed-sorting and variable-tolerance (SLRA-MV) proposed in [7]. The test matrices we will use are the same as in [7]. (The reader is referred to [7] for detailed descriptions of those test matrices.) In all of the tests we always use four Lanczos bidiagonalization iterations for computing the approximate largest singular vectors of the deflated matrices $A_i = A - X_i D_i Y_i^T$ at each iteration. The penalty factors α and β are chosen as in (2.4) with $\mu = 1/2$ and

$$\lambda = \texttt{0.05:0.05:0.5}. \tag{6.1}$$

In general, larger values of λ produce sparser factors X_k and Y_k.

Algorithm Penalized SLRA with DNI. Given a matrix A, penalty parameters α, β, and tolerance τ, this algorithm produces a positive diagonal matrix D_k and sparse matrices X_k and Y_k such that $B_k \equiv X_k D_k Y_k^T$ is a low-rank approximation of A with relative reconstruction error τ.

1. Set $A_0 = A$ and $\eta_0 = \|A\|_F^2$.

2. For $i = 1, 2, \ldots,$ until convergence,

2.1 Compute approximately the largest singular triplet (σ, u, v) of matrix A_{i-1}.

2.2 Determine integer k^* and corresponding $i^* = i_{k^*}$ and $j^* = j_{k^*}$, using Algorithm DNI.

2.3 Construct the sparse vectors x_i and y_i as defined in (2.6) with i^* and j^*, and determine the optimal d_i for the minimization problem $\min_d \|A_{i-1} - x_i d y_i^T\|$.

2.4 Set $A_i = A_{i-1} - x_i d_i y_i^T$ and $\eta_i = \eta_{i-1} - d_i^2$.

2.5 Check convergence: if $\eta_i < (\tau\|A\|_F)^2$, then set $k = i$ and turn to step 3.

3. Set $D_k = \mathrm{diag}(d_1, \ldots, d_k)$, $X_k = [x_1, \ldots, x_k]$, and $Y_k = [y_1, \ldots, y_k]$.

Behavior of DNI iterations. Compared with SLRA-MV, the penalized SLRA-DNI requires additional computations for the DNI part. The main cost of DNI is the evaluations of $\Psi(k)$ and $\Phi(k)$ for different k. Clearly they can be easily updated with about $2(m+n)$ flops, and the total cost of DNI is about $2k_{DNI}(n+m)$, where k_{DNI} is the iteration number of DNI. In all the numerical experiments we did, the iteration number k_{DNI} required is relatively small. In Table 2 we list the average number of iterations for each matrix and λ, and generally $k_{DNI} \approx 10$. Compared with the cost for computing approximately the singular vectors with four Lanczos bidiagonalization iterations, the cost for DNI can be negligible. Therefore the SLRA-DNI requires a cost similar to that for SLRA with the variable-mixed sorting technique if the approximate ranks (the number of columns of the block X_k) of the computed low-rank approximations are approximately equal.

TABLE 2
Average of the discrete secant iterations of DNI.

λ	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ash958	9.5	8.6	9.2	8.3	8.1	7.9	7.7	7.0	7.2	6.6
illc1033	11.2	10.3	9.2	9.6	9.0	8.7	8.4	8.1	7.7	70
cisi	12.8	11.3	10.4	10.0	9.9	9.4	9.4	8.9	88	8.4
cacm	11.3	10.6	10.4	10.2	9.9	9.8	9.8	97	9.1	9.3
med	14.6	12.8	12.0	11.1	11.0	10.2	97	9.5	9.1	8.9
npl	11.2	11.5	11.9	12.0	11.7	114	11.2	11.0	11.0	10.7
orsirr2	10.3	9.9	9.6	8.7	84	8.4	8.1	7.9	7.6	7.5
e20r1000	12.8	12.1	11.3	111	10.9	10.6	10.4	10.0	9.9	9.6

Although it is possible that the optimal integer k^* is in a w-constant interval which leads to complication in the algorithm, the bracketing intervals seldom occur, and the lengths $b - a$ of the bracketing intervals that do occur are rather small in general. Among the eight tested matrices, no bracketing intervals occurred for `illc1033`, `npl`, `orsirr2`, and `e20r1000` for the choice of λ `= 0.05:0.05:0.50`, while for the other four matrices, `ash958`, `cisi`, `cacm` and `med`, the bracketing intervals are rather small. See the next table, where the integer p in the form $p(q)$ is the maximal length of the q bracketing intervals that occurred.

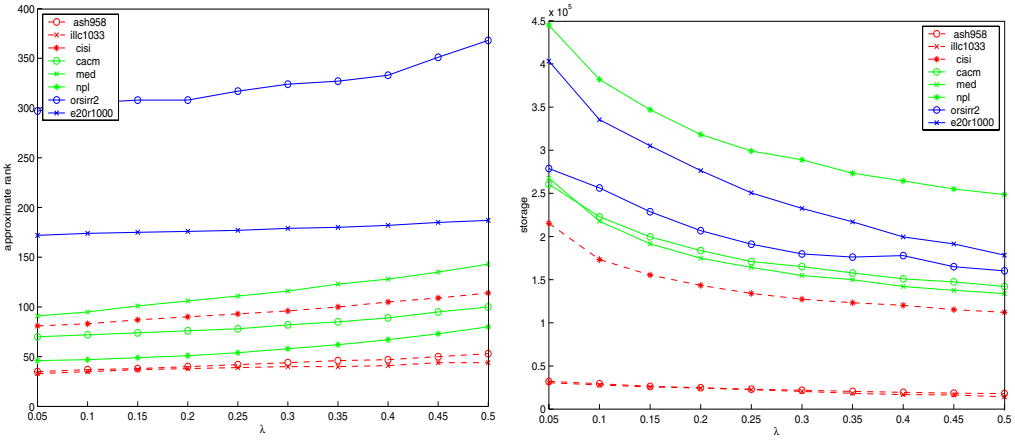


FIG. 6. *Plots for ranks (left) and numbers of nonzeros of X_k and Y_k (right) versus λ for the penalized SLRA-DNI.*

λ	0.05	0.10	015	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ash958			2(1)		2(2)				2(1)	
cisi			3(1)			2(1)				
cacm	2(1)	2(3)	2(1)	2(2)	2(1)	2(1)		2(2)	3(1)	2(1)
med										2(1)

Sensitivity of SLRA-DNI to parameter λ. We now show how the parameter λ affects the approximation rank and the storage $\mathtt{nnz}(X_k) + \mathtt{nnz}(Y_k) + k$ required for the factors X_k, Y_k, and D_k computed by SLRA-DNI. For the eight test matrices we considered, the approximation ranks and storages remain relatively stable with respect to the choice $\lambda = 0.05 : 0.05 : 0.5$. See Figure 6 for the results computed by SLRA-DNI. SLRA-DNI seems to be more stable than SLRA-MV. To show this, we choose the parameter ϵ for SLRA-MV to be

$$\epsilon = 0.05 : 0.05 : 0.5,$$

as we did in [7]. In Figure 7 we plot the storage versus the approximate rank for the two SLRA algorithms. The results show that SLRA-DNI always gives sparser factors than SLRA-MV when the computed approximations have approximately equal ranks. Figure 8 plots the storages versus reconstruction errors for the four term-document matrices in our test set for the methods TSVD, SLRA-MV, SLRA-DNI, and SDD. The storage requirement of SLRA-DNI grows relatively slowly compared with that of TSVD. Notice that the storage requirement for SDD is accounted as $(\mathtt{nnz}(X_k) + \mathtt{nnz}(Y_k))/32 + \mathtt{nnz}(D_k)$, because SDD gives factors X_k and Y_k with elements chosen from set $\{-1, 0, 1\}$, and the storage for each nonzero in X_k and Y_k is only 2 bits (or 1/32 the storage of a double).

Comparison with other methods. Now we make a comparison of the computational costs for several of the existing methods including ours. Obviously, in the $(j+1)$th rank-one approximation step, the major cost is the matrix-vector product $A_j x = Ax - X_j(D_j(Y_j^T x))$. Seven matrix-vector products are required for four Lanczos bidiagonalization iterations for approximately computing the largest left and right singular vectors of A_j. Let r be the average number of nonzeros for each column

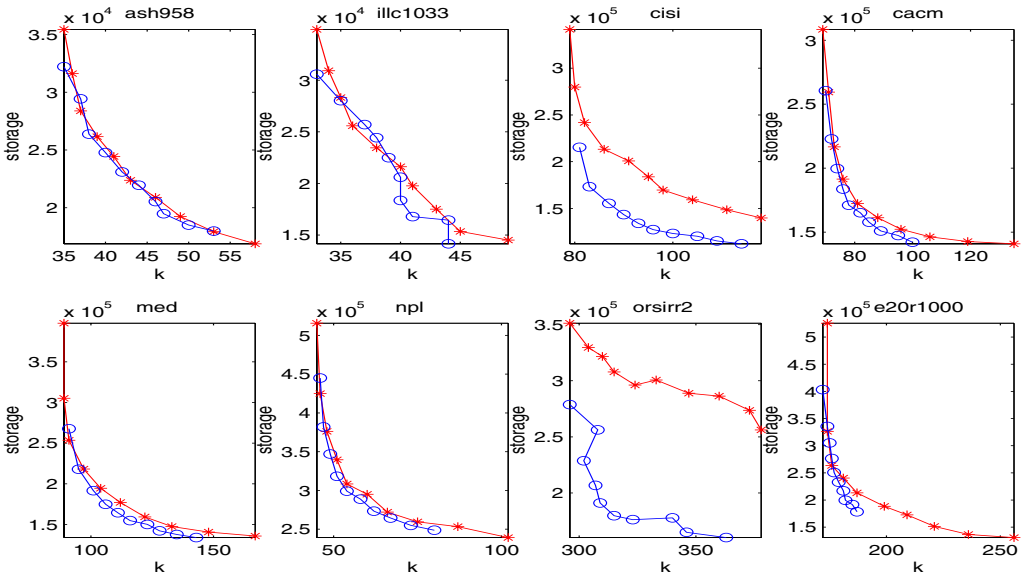


FIG. 7. *Plots storage versus approximate rank for the penalized SLRA-DNI (−o−) and SLRA-MV (− ∗ −).*

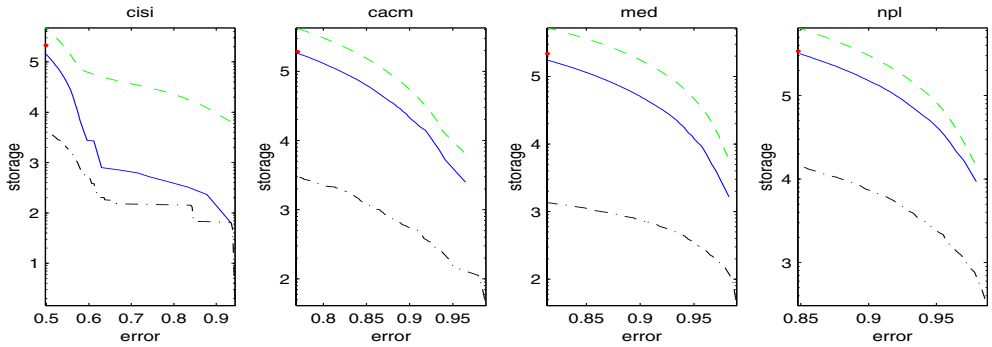


FIG. 8. *Plots storage versus reconstruction error for penalized SLRA-DNI (solid line), SLRA-MV (points), and truncated SVD (dashed line), and SDD (dashdot line).*

in the factor X_k or Y_k. Then each rank-one approximation requires seven matrix-vectors of the original matrix A, $7(4r+1)j$ flops (multiplications and divisions count one flop each), and the normalization cost for four m-dimensional vectors and four n-dimensional vectors. Thus the total cost for the sparse approximation $B_k = X_k D_k Y_k^T$ is about $7k$ matrix-vector multiplications of the original matrix A and another $7(2r+1/2)k(k+1)+36k(m+n)$ flops. Including the cost of DNI, that is about $2k_{DNI}(n+m)$ with $k_{DNI} \approx 10$; the total flops required are about

$$(14r + 3.5)k(k + 1) + 36k(m + n) + 14k\mathtt{nnz}(A).$$

In Table 3, we list computation results for several of the existing methods with different parameters. As we did in [7], for each sparse test matrix, we first compute a sparse approximation using 300 columns/rows of the test matrix that are determined by SPQR [6]. The reconstruction error then is used as the tolerance for other methods.

TABLE 3
Comparisons on cost for different methods.

Matrix	Method		k	nnz	CPUtime(s)	Flops
cisi	TSVD		72	471600	668.61	8.6541e+9
	SLRA-MV	$\epsilon = 0.1$	80	279695	33.89	2.4947e+8
		$\epsilon = 0.3$	95	183809	26.25	2.2194e+8
	SLRA-DNI	$\lambda = 01$	83	173233	18.34	1.9706e+8
		$\lambda = 0.3$	96	127298	17.25	1.9711e+8
	SDD	$choice = 1$	318	88503	254.75	5.1115e+8
		$choice = 2$	300	123804	664.76	7.2086e+8
	SPQR		300	129480	49.55	3.2661e+8
cacm	TSVD		63	422982	414.30	5.3256e+9
	SLRA-MV	$\epsilon = 01$	71	259321	30.81	2.2781e+8
		$\epsilon = 0.3$	88	161162	24.66	2.1046e+8
	SLRA-DNI	$\lambda = 0.1$	72	222786	20.32	2.1361e+8
		$\lambda = 0.3$	82	165043	17.80	2.0415e+8
	SDD	$choice = 1$	217	98932	198.34	4.1030e+8
		$choice = 2$	205	89624	630.60	3.6730e+8
	SPQR		300	134004	51.14	2.9662e+8
med	TSVD		80	522960	726.83	9.3770e+9
	SLRA-MV	$\epsilon = 0.1$	89	305015	42.08	2.9025e+8
		$\epsilon = 0.3$	112	176964	29.33	2.4306e+8
	SLRA-DNI	$\lambda = 0.1$	95	217701	23.68	2.4540e+8
		$\lambda = 0.3$	116	154644	22.14	2.4288e+8
	SDD	$choice = 1$	335	31357	178.23	2.5482e+8
		$choice = 2$	338	32368	654.60	2.5075e+8
	SPQR		300	120319	43.06	2.9364e+8
npl	TSVD		41	645791	529.71	58064e+9
	SLRA-MV	$\epsilon = 0.1$	46	424917	56.03	3.3373e+8
		$\epsilon = 0.3$	60	295024	46.86	3.6177e+8
	SLRA-DNI	$\lambda = 0.1$	47	381847	31.04	3.2627e+8
		$\lambda = 0.3$	58	288838	29.49	3.5174e+8
	SDD	$choice = 1$	—	—	—	—
		$choice = 2$	120	460073	1047.93	8.8916e+8
	SPQR		300	227649	116.38	5.8994e+8

For SLRA-MV, we take the starting error $\epsilon = 0.1$ and $\epsilon = 0.3$, respectively. For SLRA-DNI, we set $\lambda = 0.1$ and $\lambda = 0.3$. There are several initialization strategies used for the inner loops of SDD. We choose the "threshold" strategy ($choice = 1$) and the "cycling" strategy ($choice = 2$). In the table, column nnz is the number of nonzeros in factors X_k and Y_k for methods except SPQR, for which nnz is the number of nonzeros in X_k, Y_k, and H_k of the approximation $B_k = X_k H_k Y_k^T$. For the four tested term-document matrices, penalized SLRA is relatively faster than other methods and has in general less storage requirements than other methods except SDD. As we mentioned before, SDD gives $B_k = X_k D_k Y_k^T$ with factors X_k and Y_k with elements chosen from set $\{-1, 0, 1\}$ and certainly has the smallest storage requirement. However, SDD depends on the initialization strategy used for the inner iterations. For different initialization strategies, SDD tends to give different sparse approximations. At each outer iteration step of SDD, the "threshold" strategy, for example, is to determine a column of the previously constructed matrix $A_k = A - X_k D_k Y_k^T$ such that the column has its squared norm larger than the average of those of all the columns. This process will tend to be expensive if the columns of A_k have almost the same norms. It may also be possible for an infinite loop to occur if certain methods for selecting a column candidate, such as that described in SDDPACK, are used [3].[5] Such a phenomenon

[5]The MATLAB implementation is significantly slower than the C implementation.

seem to have occurred for matrix `npl`. In general, to achieve the same reconstruction error as that of TSVD or SLRA, SDD needs to compute an approximation with a much larger rank k.

7. Concluding remarks. Computing low-rank approximations of matrices is a very important matrix computation problem that has a variety of applications. The large sizes and sparsity properties of the matrices arising from some of the applications entail that we find low-rank approximations that themselves also possess some sparsity properties. We continue our research on this problem following the general framework proposed in [7]: We formulate the sparse low-rank approximation problem as an $(m + n + 1)$-dimensional penalized optimization problem and successfully reduce the penalized optimization problem to a simpler 1-D form that can be solved numerically by a discrete Newton-like iteration (DNI) method. Numerical experiments show that our penalized method is more robust and produces approximations with lower ranks (fewer columns) and/or sparser factors.

Acknowledgment. The authors want to thank the anonymous referees for their comments and suggestions that greatly improved the presentation of the paper.

REFERENCES

[1] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[2] T. G. Kolda and D. P. O'Leary, *A semidiscrete matrix decomposition for latent semantic indexing in information retrieval*, ACM Trans. Inform. Systems, 16 (1998), pp. 322–346.

[3] T. G. Kolda and D. P. O'Leary, *Algorithm* 805: *Computation and uses of the semidiscrete matrix decomposition*, ACM Trans. Math. Software, 26 (2000), pp. 416–437.

[4] D. P. O'Leary and S. Peleg, *Digital image compression by outer product expansion*, IEEE Trans. Commun., 31 (1983), pp. 441–444.

[5] H. D. Simon and H. Zha, *Low-rank matrix approximation using the Lanczos bidiagonalization process with applications*, SIAM J. Sci. Comput., 21 (2000), pp. 2257–2274.

[6] G. W. Stewart, *Four algorithms for the efficient computation of truncated pivoted QR approximation to a sparse matrix*, Numer. Math., 83 (1999), pp. 313–323.

[7] Z. Zhang, H. Zha, and H. Simon, *Low-rank approximations with sparse factors* I: *Basic algorithms and error analysis*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 706–727.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 921–929

GENERALIZED EXPONENTIAL VANDERMONDE DETERMINANT AND HERMITE MULTIPOINT DISCRETE BOUNDARY VALUE PROBLEM*

RAGHIB ABU-SARIS† AND WAJDI AHMAD‡

Abstract. We introduce a determinant that encompasses the classical Vandermonde determinant, the generalized Vandermonde determinant, and the recently introduced exponential Vandermonde determinant when the exponents are nonnegative integers. An explicit factorization of such a determinant will be established. This factorization enables us to develop a computationally tractable necessary and sufficient condition for the existence of a unique solution of a Hermite (ℓ-point) discrete boundary value problem.

Key words. discrete boundary value problems, existence and uniqueness theorems, generalized exponential Vandermonde determinant

AMS subject classifications. 39A10, 65F40

DOI. 10.1137/S0895479803421264

1. Introduction.

First, consider a linear difference equation with constant coefficients of order $k \geq 2$:

$$\sum_{j=0}^{k} a_j \, y(t+j) = 0, \qquad t = 0, 1, 2, \ldots, \qquad a_0 a_k \neq 0, \tag{1.1}$$

subject to separable auxiliary conditions (AC) of the form

$$y(n_i) = y_{n_i}, \qquad i = 1, 2, \ldots, k, \tag{1.2}$$

where n_i's are nonnegative integers such that $0 = n_1 < \cdots < n_k$.

If the integers $n_i,\ i = 1, \ldots, k$, are consecutive, then (1.1), together with conditions (1.2), is called a *(discrete) initial value problem (DIVP)*. However, if there are two n_i's that are not consecutive, then (1.1) together with conditions (1.2) is called a *(discrete) boundary value problem (DBVP)*. Boundary conditions as described in (1.2) are equivalent to the *Hermite (ℓ-point) conditions* (see [3, p. 12]):

$$\Delta^j y(t_i) = \alpha_{ij}, \qquad 1 \leq i \leq \ell, \quad 0 \leq j \leq k_i - 1,$$

where $k_i \geq 1$,

$$0 = n_1 = t_1 < t_1 + k_1 < t_2 < \cdots < t_{\ell-1} < t_{\ell-1} + k_{\ell-1} < t_\ell \leq t_\ell + k_\ell - 1 = n_k,$$

and $\sum_{j=1}^{\ell} k_j = k$. Of particular interest are the *Hermite (2-point) conditions*. In this case, conditions (1.2) split into *initial* and *final conditions:*

$$\text{initial condition(s): } y(i) = y_i, \qquad i = 0, \ldots, k_1 - 1, \tag{1.3}$$

$$\text{final condition(s): } y(j) = y_j, \qquad j = N, \ldots, N + k_2 - 1, \tag{1.4}$$

*Received by the editors January 10, 2003; accepted for publication (in revised form) by D. Bini October 23, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/simax/25-4/42126.html

†Department of Basic Sciences, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates (rabusaris@sharjah.ac.ae).

‡Department of Electrical and Electronics Engineering, University of Sharjah, P.O. Box 27272, Sharjah, United Arab Emirates (wajdi@sharjah.ac.ae).

where N is a positive integer such that $N > k_1$. A necessary and sufficient condition for the existence of a unique solution of a Hermite 2-point DBVP has been established in [1]. It is worth mentioning that boundary conditions (1.2) include, as a special case, the *Niccoletti conditions* [3, p. 12]:

$$y(n_i) = y_{n_i}, \qquad i = 1, \ldots, k,$$

where

$$0 = n_1 < n_1 + 1 < n_2 < \cdots < n_{k-1} < n_{k-1} + 1 < n_k,$$

i.e., $n_{i+1} - n_i \geq 2$ for $i = 1, \ldots, k-1$.

Unlike DIVP, the theory and the construction of the solutions of DBVP are more difficult [3, p. 629]. In fact, there is a need for a simple criterion to determine whether a DBVP is well posed [4, p. 43]. In the literature, there are many papers concerned with DBVP. For a comprehensive survey on linear and nonlinear DBVP, we refer the reader to the books by Agarwal [3, pp. 629–634 and 684–691] and by Kelley and Peterson [8, pp. 279–289 and 327–340] and the references cited therein. Our interest in DBVP was initiated by an open problem due to Trigiante in [9]. Sufficient conditions on the distribution of the characteristic roots that ensure the existence of a unique solution of DBVP (1.1)–(1.2) were established in [2]. Certainly, we are not the first to tackle this problem. However, the existence and uniqueness criterion presented in this paper has two favorable properties: *being necessary and sufficient* and *easy to apply*.

To start, let $p(\lambda)$ be the characteristic polynomial associated with (1.1), i.e.,

$$p(\lambda) = \sum_{j=0}^{k} a_j \lambda^j. \tag{1.5}$$

Observe that if the characteristic polynomial (1.5) has $r \geq 1$ distinct characteristic roots, then the general solution of (1.1) is given by

$$y(t) = \sum_{j=1}^{r} q_j(t) z_j^t, \qquad t \geq 0,$$

where $q_j(t) = \sum_{i=0}^{m_j - 1} c_{ji} t^i$ is a polynomial in t of degree $m_j - 1$, and $m_j \geq 1$ is the multiplicity of characteristic root z_j (cf. [6, pp. 63–65] and [8, pp. 54–63] for proofs and details) such that $\sum_{j=1}^{r} m_j = k$. Applying AC (1.2), we obtain

$$\begin{aligned} \sum_{j=1}^{r} q_j(0) &= y_0, \\ \vdots & \quad \vdots \\ \sum_{j=1}^{r} q_j(n_k) z_j^{n_k} &= y_{n_k}, \end{aligned}$$

which is a system of linear equations in which the coefficient matrix M is given by

$$M = \begin{cases} \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ z_1^{n_k} & \cdots & n_k^{k-1} z_1^{n_k} \end{pmatrix} & \text{if} \quad r = 1, \\ \begin{pmatrix} M_1 & \cdots & M_r \end{pmatrix} & \text{if} \quad r > 1, \end{cases} \tag{1.6}$$

where M_j is the $k \times m_j$ matrix given by

$$M_j = \begin{cases} \begin{pmatrix} 1 \\ \vdots \\ z_j^{n_k} \end{pmatrix} & \text{if} \quad m_j = 1, \\ \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ z_j^{n_k} & \cdots & n_k^{m_j-1} z_j^{n_k} \end{pmatrix} & \text{if} \quad m_j > 1 \end{cases} \tag{1.7}$$

for $j = 1, \ldots, r$. Therefore, the existence of a unique solution of DBVP (1.1)–(1.2) is equivalent to the nonsingularity of the block matrix M.

Now, let $\vec{z} = (z_1, \ldots, z_r)$, $\vec{m} = (m_1, \ldots, m_r)$, and $\vec{n} = (n_1, \ldots, n_k)$, and define

$$G_k(\vec{z}, \vec{m}, \vec{n}) = \det(M) = |M|. \tag{1.8}$$

Observe the following:

- If $n_j = j-1$ for $j = 2, \ldots, k$, then G_k reduces to the *generalized Vandermonde determinant* introduced by Flowe and Harris in [7]; it is worth noticing that

$$\det(M) = \left(\prod_{i=1}^{r} z_i^{m_i(m_i-1)/2} \right) \left(\prod_{i=1}^{r} \prod_{j_i=0}^{m_i-1} j_i! \right) \left(\prod_{j=2}^{r} \prod_{i=1}^{j-1} (z_j - z_i)^{m_j m_i} \right),$$

  which is nonvanishing whenever z_j are distinct and nonzero. Moreover, it should be noted that the last factor includes the classical Vandermonde determinant $\prod_{j>i}(z_j - z_i)$.
- If $r = k$ and thus $m_i = 1$ for $i = 1, 2, \ldots, k$, then G_k reduces to an *exponential Vandermonde determinant* with nonnegative integer exponents (cf. [10, 11]).

Therefore, it is natural to call the matrix M defined in (1.6) *a generalized exponential Vandermonde matrix* and the determinant defined in (1.8) *a generalized exponential Vandermonde determinant (GEVD).*

This paper is organized as follows. In section 2, we state and illustrate the applicability of our main results. Preliminary results needed to establish the explicit factorization of GEVD will be recalled in section 3. We establish the proof of the explicit factorization result in section 4. We conclude in section 5 with worthy remarks about the results obtained.

2. The main results. Our main results are the following two theorems.

THEOREM 2.1. *Let* $s_i = n_i - n_{i-1} - 1$, $i = 2, 3, \ldots,$ $n_1 = 0$, *and* $s = \sum_{i=2}^{k} s_i$. *Let* A_i *be the matrix of size* $s_i \times s$ *defined by*

$$A_i = \begin{pmatrix} a_{n_{i-1}+1} & a_{n_{i-1}} & \cdots & a_{n_{i-1}-s+2} \\ \vdots & \vdots & & \vdots \\ a_{n_i-1} & a_{n_i-2} & \cdots & a_{n_i-s} \end{pmatrix},$$

where a_j, $j = 0, \ldots, k$, *are the coefficients of the monic polynomial* $p(\lambda) = \prod_{i=1}^{r} (\lambda - z_i)^{m_i} = \sum_{j=0}^{k} a_j \lambda^j$, $a_k = 1$, *and* $a_j = 0$ *if* $j < 0$ *or* $j > k$. *Then*

$$G_k(\vec{z}, \vec{m}, \vec{n}) = \left[\prod_{i=2}^{k} (-1)^{s_i(k-i+1)} \right] \left(\prod_{i=1}^{r} z_i^{m_i(m_i-1)/2} \right) |A| \; V_k(\vec{z}, \vec{m}),$$

where

$$V_k(\vec{z},\vec{m}) = |\overbrace{\vec{u}(z_1)\cdots\vec{u}^{(m_1-1)}(z_1)}^{m_1\ columns}\ \cdots\ \overbrace{\vec{u}(z_r)\cdots\vec{u}^{(m_r-1)}(z_r)}^{m_r\ columns}| = \left(\prod_{i=1}^{r}\prod_{j_i=0}^{m_i-1} j_i!\right)\ \prod_{j=2}^{r}\prod_{i=1}^{j-1}(z_j-z_i)^{m_i m_j},$$

$\vec{u}(z) = (1, z, \ldots, z^{k-1})^T \in \mathbb{C}^k$, *and* A *is the* $s \times s$ *block matrix*

$$A = \begin{pmatrix} A_2 \\ \vdots \\ A_k \end{pmatrix}.$$

Since the matrix M is nonsingular whenever $G_k(\vec{z},\vec{m},\vec{n}) \neq 0$, which is in turn equivalent to $|A| \neq 0$, we have the following important result.

THEOREM 2.2. *Let the matrix* A *be as defined in Theorem* 2.1. *DBVP* (1.1)–(1.2) *has a unique solution if and only if* A *is nonsingular.*

The following result is related to the Niccoletti DBVP and is an immediate corollary of Theorem 2.2.

COROLLARY 2.1. *If the associated characteristic polynomial of* (1.1) *is even and* $n_i - n_{i-1} = 2,\ i = 2, \ldots, k$ *(all gaps are of unit size), then there exists no unique solution of DBVP* (1.1)–(1.2).

For the sake of definiteness, we emphasize the following points.

Remark 2.1.

- If $s_i = 0$ for some i, the corresponding block will be missing from the matrix A, as will be seen in Example 2.2 below.
- Also, the matrix A introduced in Theorem 2.1 has a nice structure that makes it easy to construct. Namely, the first column is made up of the coefficients of the missing powers in G_k, and the indices in each row are sequential and in descending order.

And for the sake of clarity, we present the following examples.

Example 2.1. If $k = 4$, $n_2 = 2$, $n_3 = 4$, and $n_4 = 6$, then $s_2 = s_3 = s_4 = 1$, and $s = 3$. Therefore,

$$A_2 = \begin{pmatrix} a_1 & a_0 & a_{-1} \end{pmatrix} = \begin{pmatrix} a_1 & a_0 & 0 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} a_3 & a_2 & a_1 \end{pmatrix},$$

and

$$A_4 = \begin{pmatrix} a_5 & a_4 & a_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & a_3 \end{pmatrix}.$$

Hence

$$A = \begin{pmatrix} A_2 \\ A_3 \\ A_4 \end{pmatrix} = \begin{pmatrix} a_1 & a_0 & 0 \\ -- & -- & -- \\ a_3 & a_2 & a_1 \\ -- & -- & -- \\ 0 & 1 & a_3 \end{pmatrix}.$$

Example 2.2. If $k = 6,\ n_2 = 1,\ n_3 = 4,\ n_4 = 5,\ n_5 = 9$, and $n_6 = 10$, then $s_2 = 0,\ s_3 = 2,\ s_4 = 0,\ s_5 = 3,\ s_6 = 0$, and $s = 5$. Therefore,

$$A_3 = \begin{pmatrix} a_2 & a_1 & a_0 & a_{-1} & a_{-2} \\ a_3 & a_2 & a_1 & a_0 & a_{-1} \end{pmatrix} = \begin{pmatrix} a_2 & a_1 & a_0 & 0 & 0 \\ a_3 & a_2 & a_1 & a_0 & 0 \end{pmatrix},$$

$$A_5 = \begin{pmatrix} a_6 & a_5 & a_4 & a_3 & a_2 \\ a_7 & a_6 & a_5 & a_4 & a_3 \\ a_8 & a_7 & a_6 & a_5 & a_4 \end{pmatrix} = \begin{pmatrix} 1 & a_5 & a_4 & a_3 & a_2 \\ 0 & 1 & a_5 & a_4 & a_3 \\ 0 & 0 & 1 & a_5 & a_4 \end{pmatrix},$$

and $A_2,\ A_4,\ A_6$ will not appear in A. Hence

$$A = \begin{pmatrix} A_3 \\ A_5 \end{pmatrix} = \begin{pmatrix} a_2 & a_1 & a_0 & 0 & 0 \\ a_3 & a_2 & a_1 & a_0 & 0 \\ -- & -- & -- & -- & -- \\ 1 & a_5 & a_4 & a_3 & a_2 \\ 0 & 1 & a_5 & a_4 & a_3 \\ 0 & 0 & 1 & a_5 & a_4 \end{pmatrix}.$$

3. Preliminary results. To prove Theorem 2.1, several results are needed. The first one is the following lemma.

LEMMA 3.1. *If* $\vec{u}(z) = (z^{n_1}, z^{n_2}, \ldots, z^{n_k})^T$ *and* $\vec{v}_j(z) = (n_1^j z^{n_1}, n_2^j z^{n_2}, \ldots, n_k^j z^{n_k})^T$, *then*

$$\vec{v}_j(z) = \sum_{\ell=0}^{j} \frac{1}{\ell!}\ (\Delta^\ell n^j)|_{n=0}\ z^\ell\ \vec{u}^{(\ell)}(z),$$

where $\vec{u}^{(\ell)}$ *denotes the* ℓ*th derivative of* $\vec{u}$, *the superscript* T *denotes the transpose, and* Δ *is the forward difference operator, i.e.,* $\Delta f(n) = f(n+1) - f(n)$.

Proof. First, by Newton's forward interpolating polynomial [5, p. 128],

$$\begin{aligned} n^j &= \sum_{\ell=0}^{j} \begin{pmatrix} n \\ \ell \end{pmatrix}\ (\Delta^\ell n^j)|_{n=0} \\ &= \sum_{\ell=0}^{j} \frac{(\Delta^\ell n^j)|_{n=0}}{\ell!}\ \prod_{i=0}^{\ell-1}(n-i). \end{aligned}$$

Therefore,

$$n^j z^n = \sum_{\ell=0}^{j} \frac{(\Delta^\ell n^j)|_{n=0}}{\ell!}\ z^\ell\ \frac{d^\ell z^n}{dz^\ell}.$$

Now, since $(\Delta n_i^j)|_{n_i=0}$ is independent of i,

$$\begin{aligned} \vec{v}_j(z) &= \left(\sum_{\ell=0}^{j} \frac{(\Delta^\ell n_1^j)|_{n_1=0}}{\ell!}\ z^\ell\ \frac{d^\ell z^{n_1}}{dz^\ell}, \cdots, \sum_{\ell=0}^{j} \frac{(\Delta^\ell n_k^j)|_{n_k=0}}{\ell!}\ z^\ell\ \frac{d^\ell z^{n_k}}{dz^\ell} \right)^T \\ &= \sum_{\ell=0}^{j} \frac{(\Delta^\ell n^j)|_{n=0}}{\ell!}\ z^\ell\ \left(\frac{d^\ell z^{n_1}}{dz^\ell}, \cdots, \frac{d^\ell z^{n_k}}{dz^\ell} \right)^T \\ &= \sum_{\ell=0}^{j} \frac{(\Delta^\ell n^j)|_{n=0}}{\ell!}\ z^\ell\ \vec{u}^{(\ell)}(z). \qquad \square \end{aligned}$$

We also need the following two lemmas from [1]. The first is a possible generalization of the *Leibnitz differentiation formula*, and the second is related to Hermite 2-point conditions.

LEMMA 3.2. *If c, d are positive integers, and $\Phi(z) = |C_1(z) \ \cdots \ C_c(z)|$ such that $C_j(z) : \mathbb{C} \longrightarrow \mathbb{C}^c$ are sufficiently differentiable functions, then*

$$\Phi^{(d)}(z) = \sum_{\|\alpha\|=d} \frac{d!}{\alpha_1! \cdots \alpha_c!} \ \left| C_1^{(\alpha_1)}(z) \ \cdots \ C_c^{(\alpha_c)}(z) \right|,$$

where $\|\alpha\| = \alpha_1 + \cdots + \alpha_c$ and $\alpha_i \in \{0, 1, \ldots, d\}$.

LEMMA 3.3. *Let k_1, k_2, and N be positive integers such that $k_1 + k_2 = k$ and $k_1 < N$. If*

$$\vec{u}(z) = (1, \ldots, z^{k_1-1}, z^N, \ldots, z^{N+k_2-1})^T \in \mathbb{C}^k$$

and

$$E_k(\vec{z}, \vec{m}) = | \underbrace{\vec{u}(z_1) \cdots \vec{u}^{(m_1-1)}(z_1)}_{m_1 \ \text{columns}} \ \cdots \ \underbrace{\vec{u}(z_r) \cdots \vec{u}^{(m_r-1)}(z_r)}_{m_r \ \text{columns}} |,$$

then

$$E_k(\vec{z}, \vec{m}) = (-1)^{(N-k_1)k_2} \ |A_{k_1+1}| \ V_k(\vec{z}, \vec{m}),$$

where A_{k_1+1} is a matrix of size $(N - k_1) \times (N - k_1)$ as defined in Theorem 2.1.

4. Proof of Theorem 2.1. First, by Lemma 3.1 and the usual properties of determinants, we have

$$G_k(\vec{z}, \vec{m}, \vec{n}) = \left(\prod_{i=1}^{r} z_i^{m_i(m_i-1)/2} \right) \ F_k(\vec{z}, \vec{m}, \vec{n}),$$

where

$$F_k(\vec{z}, \vec{m}, \vec{n}) = |\vec{u}(z_1) \cdots \vec{u}^{(m_1-1)}(z_1) \ \cdots \ \vec{u}(z_r) \cdots \vec{u}^{(m_r-1)}(z_r)|,$$

and $\vec{u}(z) = (1, z^{n_2}, \ldots, z^{n_k})^T$. Therefore, we will conclude once we show that

$$F_k(\vec{z}, \vec{m}, \vec{n}) = \left[\prod_{i=2}^{k} (-1)^{s_i(k-i+1)} \right] \ |A| \ V_k(\vec{z}, \vec{m}).$$

Now, observe that Lemma 3.3 deals with the case when the powers are consecutive except at one place where there is a gap of arbitrary size between the powers; i.e., the result is true when the number of gaps is 1. However, Theorem 2.1 allows the possibility of more than one gap between the powers. To this end, suppose that the result holds when the number of gaps is $h - 1 \geq 1$. To prove that it is true when the number of gaps is h, let the powers be given by $\vec{n} = (0, \ldots, k_1 - 1, N_1, \ldots, N_1 + k_2 - 1, \ldots, N_h, \ldots, N_h + k_{h+1} - 1)$, where $k_1, \ldots, k_{h+1}$, $N_1, \ldots, N_h$ are positive integers such that $k_1 + \cdots + k_{h+1} = k$, $k_1 < N_1$, and $N_j + k_{j+1} < N_{j+1}$, $j = 1, \ldots, h - 1$. Further, define $g_1 = N_1 - k_1$, $g_j = N_j - N_{j-1} - k_j$, $j = 2, \ldots, h$, $s = g_1 + \cdots + g_h$, and $K_j = \sum_{i=1}^{j} k_i$, $j = 1, \ldots, h$. We claim that

$$F_k(\vec{z}, \vec{m}, \vec{n}) = \left[\prod_{j=1}^{h} (-1)^{g_j(k-K_j)} \right] \ |A| \ V_k(\vec{z}, \vec{m}),$$

where

$$A = \begin{pmatrix} A_{K_1+1} \\ - - - \\ \vdots \\ - - - \\ A_{K_h+1} \end{pmatrix}.$$

We will justify our claim by mathematical induction on the size of the first gap. To start, suppose that $g_1 = 1$ (i.e., $N_1 = k_1 + 1$). By the induction hypothesis, we have

$$\begin{aligned} F_{k+1}\left((\vec{z}, z_0), (\vec{m}, 1), \tilde{n}\right) &= \left[\prod_{j=2}^{h} (-1)^{g_j(k+1-(K_j+1))}\right] \; |B| \; V_{k+1}\left((\vec{z}, z_0), (\vec{m}, 1)\right) \\ &= \left[\prod_{j=2}^{h} (-1)^{g_j(k-K_j)}\right] \; |B| \; V_{k+1}\left((\vec{z}, z_0), (\vec{m}, 1)\right), \end{aligned}$$

where $\tilde{n} = (0, \ldots, k_1 - 1, k_1, N_1, \ldots, N_1 + k_2 - 1, \ldots, N_h, \ldots, N_h + k_{h+1} - 1)$, z_0 is an additional zero of the polynomial $q(\lambda) = (\lambda - z_0)p(\lambda) = \sum_{i=0}^{k+1} b_i \; \lambda^i$, and B is the block matrix of size $(s-1) \times (s-1)$ given by

$$B = \begin{pmatrix} B_{K_2+2} \\ - - - \\ \vdots \\ - - - \\ B_{K_h+2} \end{pmatrix} = \begin{pmatrix} b_{\tilde{n}_{K_2+1}+1} & b_{\tilde{n}_{K_2+1}} & \cdots & b_{\tilde{n}_{K_2+1}-s+3} \\ \vdots & \vdots & & \vdots \\ b_{\tilde{n}_{K_2+2}-1} & b_{\tilde{n}_{K_2+2}-2} & \cdots & b_{\tilde{n}_{K_2+2}-s+1} \\ - - - & - - - & -- & - - - - - \\ \vdots & \vdots & \cdots & \vdots \\ - - - & - - - & -- & - - - - - \\ b_{\tilde{n}_{K_h+1}+1} & b_{\tilde{n}_{K_h+1}} & \cdots & b_{\tilde{n}_{K_h+1}-s+3} \\ \vdots & \vdots & & \vdots \\ b_{\tilde{n}_{K_h+2}-1} & b_{\tilde{n}_{K_h+2}-2} & \cdots & b_{\tilde{n}_{K_h+2}-s+1} \end{pmatrix}.$$

Differentiating k_1 times with respect to z_0 and substituting $z_0 = 0$, we obtain

$$\left.\frac{\partial^{k_1} F_{k+1}((\vec{z}, z_0), (\vec{m}, 1), \tilde{n})}{\partial z_0^{k_1}}\right|_{z_0=0} = (-1)^{k+k_1} \; k_1! \; F_k(\vec{z}, \vec{m}, \vec{n}). \tag{4.1}$$

However, by the Leibnitz rule of differentiation, we have

$$\begin{aligned} \left.\frac{\partial^{k_1} F_{k+1}((\vec{z}, z_0), (\vec{m}, 1), \tilde{n})}{\partial z_0^{k_1}}\right|_{z_0=0} &= \left[\prod_{j=2}^{h} (-1)^{g_j(k-K_j)}\right] \\ \times \sum_{\ell=0}^{k_1} \begin{pmatrix} k_1 \\ \ell \end{pmatrix} \; \left.\frac{\partial^{\ell} |B|}{\partial z_0^{\ell}}\right|_{z_0=0} & \left.\frac{\partial^{k_1-\ell} \; V_{k+1}((\vec{z}, z_0), (\vec{m}, 1))}{\partial z_0^{k_1-\ell}}\right|_{z_0=0}. \end{aligned} \tag{4.2}$$

But $\partial^j \; V_{k+1}(\vec{z}, z_0, \vec{m}, 1)/\partial z_0^j$ evaluated at $z_0 = 0$, and after expanding through the $(k+1)^{st}$ column, has exactly one gap between the powers of size 1. Therefore, by

Lemma 3.3,

$$\left.\frac{\partial^j\ V_{k+1}((\vec{z},z_0),(\vec{m},1))}{\partial z_0^j}\right|_{z_0=0} = (-1)^{j+k}\ j!\ E_k(\vec{z},\vec{m}) = j!\ a_j\ V_k(\vec{z},\vec{m}), \tag{4.3}$$

and by Lemma 3.2,

$$\left.\frac{\partial^j|B|}{\partial z_0^j}\right|_{z_0=0} = \sum_{\|\alpha\|=j} \frac{j!}{\alpha_1!\cdots\alpha_{s-1}!}\ \left|C_1^{(\alpha_1)}(0)\cdots C_{s-1}^{(\alpha_{s-1})}(0)\right|,$$

where $C_i(z_0) = (b_{\tilde{n}_{K_2+1}-i+2},\ldots,b_{\tilde{n}_{K_2+2}-i},\ldots,b_{\tilde{n}_{K_h+1}-i+2},\ldots,b_{\tilde{n}_{K_h+2}-i})^T$. Since $b_i = a_{i-1} - a_i z_0$ for $i = 0,1,2,\ldots,k,\ a_{-1} = 0$, i.e., linear in z_0, we have $\alpha_i \in \{0,1\}$ and so $\alpha_i! = 1$. Furthermore, at $z_0 = 0$, we have $b_i = a_{i-1}$. Therefore, applying the usual properties of determinants and using the fact that $b_i' = -a_i$, we get

$$\begin{aligned}\left.\frac{\partial^j\ |B|}{\partial z_0^j}\right|_{z_0=0} &= j!\ \left|C_1'(0)\ \cdots\ C_j'(0)\ C_{j+1}(0)\ \cdots\ C_{s-1}(0)\right| \\ &= (-1)^j\ j!\ M_{1,j+1}, \end{aligned}\tag{4.4}$$

where M_{ij} is the ij-minor of A. Substituting (4.3) and (4.4) into (4.2), equating (4.1) and (4.2), and simplifying, we obtain

$$\begin{aligned} F_k(\vec{z},\vec{m},\vec{n}) &= (-1)^{-(k+k_1)}\ \left[\prod_{j=2}^{h}(-1)^{g_j(k-K_j)}\right]\ \sum_{\ell=0}^{k_1}(-1)^{\ell}\ a_{k_1-\ell}\ M_{1,\ell+1}\ V_k(\vec{z},\vec{m}) \\ &= \left[\prod_{j=1}^{h}(-1)^{g_j(k-K_j)}\right]\ |A|\ V_k(\vec{z},\vec{m}). \end{aligned}$$

Hence the result is true for $g_1 = 1$.

As a final step, suppose that the statement is true for gap size g_1. To prove that it holds for gap size $g_1 + 1$, the same steps above can be applied exactly the same way. However, two modifications are needed. To be specific, the lower index in the product of powers of -1 factor will be 1 rather than 2, and the matrix B will have an extra block of size $g_1 \times (s-1)$. This completes the proof. ∎

5. Conclusions and final remarks. We conclude our paper with the following remarks.

1. The results obtained are self-standing, as they provide a solution to the fundamental question of existence and uniqueness of a solution of DBVP (1.1)–(1.2) which is a critical issue for a problem to be well posed. Nonetheless, from an application standpoint, DIVP and DBVP arise in various fields of interest such as analysis of electric circuits and power electronic converters, discrete modelling of economic and biological phenomena, and interpolation with exponential functions, to name a few. The reader is referred to the books by Agarwal [3, pp. 6–11 and 13–26] and Kelley and Peterson [8, Ch. 1] for detailed examples on such applications. In addition, the criterion developed in Theorem 2.2 can be easily programmed using a symbolic computational package such as MAPLE.

2. Theorem 2.2 extends and generalizes Theorem 2.1 which was established by the authors in [1]. Furthermore, it applies to a wider range of boundary conditions, namely,

$$T \begin{pmatrix} y(0) \\ y(n_2) \\ \vdots \\ y(n_k) \end{pmatrix} = \begin{pmatrix} y_0 \\ y_{n_2} \\ \vdots \\ y_{n_k} \end{pmatrix},$$

where T is a square matrix of full rank. In the literature, this type of boundary conditions are called *implicit separated conditions* [3, p. 12].

REFERENCES

[1] R. Abu-Saris and W. Ahmad, *A necessary and sufficient condition for the existence of a unique solution of a discrete boundary value problem*, Int. J. Math. Math. Sci., 39 (2003), pp. 2455–2463.

[2] R. Abu-Saris and H. Yousef, *Existence theorems for boundary value problems in difference equations*, J. Differ. Equations Appl., 7 (2001), pp. 255–263.

[3] R. Agarwal, *Difference Equations and Inequalities: Theory, Methods, and Applications*, 2nd ed., Marcel Dekker, New York, 2000.

[4] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory*, Springer-Verlag, New York, 1999.

[5] J. Faires and R. Burden, *Numerical Methods*, 2nd ed., Brooks/Cole, New York, 1998.

[6] S. Elaydi, *An Introduction to Difference Equations*, 2nd ed., Springer-Verlag, New York, 1999.

[7] R. Flowe and G. Harris, *A note on generalized Vandermonde determinants*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1146–1151.

[8] W. Kelley and A. Peterson, *Difference Equations: An Introduction with Applications*, 2nd ed., Harcourt/Academic Press, New York, 2001.

[9] G. Ladas, *Open problems and conjectures*, J. Differ. Equations Appl., 4 (1998), pp. 395–396.

[10] J. Robbin and D. Salamon, *The exponential Vandermonde matrix*, Linear Algebra Appl., 317 (2000), pp. 225–226.

[11] S. Yang, H. Wu, and Q. Zhang, *Generalization of Vandermonde determinants*, Linear Algebra Appl., 336 (2001), pp. 201–204.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 930–946

THE SPECTRA OF PRECONDITIONED TOEPLITZ MATRIX SEQUENCES CAN HAVE GAPS*

THOMAS HUCKLE† AND STEFANO SERRA-CAPIZZANO‡

Abstract. Different than for the case of Toeplitz matrix sequences $\{T_n(f)\}$, $f \in L^1$, we can prove that the closure of the union of all the spectra of preconditioned matrix sequences of the form $\{T_n^{-1}(g)T_n(f)\}$, $f, g \in L^1$, $g \geq 0$, can have gaps if the essential range of f/g is not connected. The result has important consequences on the practical use of band Toeplitz preconditioners widely used in the literature both for (multilevel) ill-conditioned positive definite and (multilevel) indefinite Toeplitz linear systems.

Key words. Toeplitz matrix, generating function, preconditioning, spectral distribution and localization results

AMS subject classifications. 65F10, 15A18

DOI. 10.1137/S0895479802419269

1. Introduction. Let f be a real valued function of k variables, integrable on the k-cube $I_k := (0, 2\pi)^k$. Throughout, the symbol $\fint_{I_k}$ stands for $(2\pi)^{-k} \int_{I_k}$, and the symbol L^1 stands for $L^1(I^k, (2\pi)^{-k}dx)$. The Fourier coefficients of f, given by

$$\widehat{f}_j := \fint_{I_k} f(x)e^{-\mathbf{i}\,\langle j,x\rangle}\,dx, \quad \mathbf{i}^2 = -1, \quad j \in \mathbf{Z}^k, \quad \langle j, x\rangle = \sum_{t=1}^{k} j_t x_t, \tag{1}$$

are the entries of the k-level Toeplitz matrices generated by f. More precisely, if $n = (n_1, \ldots, n_k)$ is a k-index with positive entries, then $T_n(f)$ denotes the matrix of order $\widehat{n} := \prod_{i=1}^{k} n_i$ given by

$$T_n(f) = \sum_{|j_1|<n_1} \cdots \sum_{|j_k|<n_k} \left[J_{n_1}^{(j_1)} \otimes \cdots \otimes J_{n_k}^{(j_k)}\right] \widehat{f}_{(j_1,\ldots,j_k)}. \tag{2}$$

In this case, we say that the sequence $\{T_n(f)\}$ is *generated* by f. In the above equation, $\otimes$ denotes the tensor product, while $J_m^{(l)}$ denotes the matrix of order m whose (i, j) entry equals 1 if $j - i = l$ and equals zero otherwise: the reader is referred to [36] for more details on multilevel Toeplitz matrices. Furthermore, in many cases the generating function is known. If it is unknown the reader is referred to [29] for the one-dimensional case and [16] for the two-dimensional case.

The spectral properties of the sequence $\{T_n(f)\}$ and of related preconditioned sequences are completely understood and characterized in terms of the underlying generating functions (see [13, 3, 37, 33, 31]). For instance, it is immediate to deduce that $T_n(f)$ is Hermitian for any n since f is real valued and therefore $\overline{\widehat{f}_{-(j_1,\ldots,j_k)}} = \widehat{f}_{(j_1,\ldots,j_k)}$.

*Received by the editors December 4, 2002; accepted for publication (in revised form) by L. Reichel September 18, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/simax/25-4/41926.html

†Institut für Informatik, TU Munchen, Arcisstrasse 21, 80290 Munchen, Germany (huckle@informatik.tu-muenchen.de).

‡Dipartimento di Chimica, Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100 Como, Italy (stefano.serrac@uninsubria.it, serra@mail.dm.unipi.it).

Further results are contained in the following theorem.

THEOREM 1.1. *Let f be a k variate Lebesgue integrable function defined over I_k. Then the following facts hold:*

1. [13] *If f is not identically constant, then every eigenvalue of $T_n(f)$ lies in (m, M), where*

$$m = \text{essinf } (f)$$

and

$$M = \text{esssup } (f),$$

where essinf *and* esssup *denote* inf *and* sup *(respectively) up to zero Lebesgue measure sets; if f is identically constant, then $m = M$ and $T_n(f) = mI$ with I being the identity matrix of size $\widehat{n}$.*

2. [13] *If we denote by $\lambda_{\min}(n)$ and by $\lambda_{\max}(n)$ the minimal and the maximal eigenvalues of $T_n(f)$, then*

$$\lim_{n\to\infty} \lambda_{\min}(n) = m, \quad \lim_{n\to\infty} \lambda_{\max}(n) = M.$$

In addition, if $n_i \sim n_j$ for any i and j, then [22, 2] $\lambda_{\min}(n) - m \sim \widehat{n}^{-\alpha/k}$ and $M - \lambda_{\max}(n) \sim \widehat{n}^{-\beta/k}$, where α is the maximum among the orders of the zeros of $f(x) - m$ and β is the maximum among the orders of the zeros of $M - f(x)$.

More asymptotics are known and concern the global spectral behavior of the sequence $\{T_n(f)\}$ when f is just Lebesgue integrable: we start with a necessary definition.

DEFINITION 1.2. *Let $\{A_n\}$ be a sequence of matrices of increasing dimensions d_n ($d_n < d_{n+1}\ \forall n$) and let θ be a real valued measurable function defined over a set K of finite Lebesgue measure $m\{K\}$. We write that $\{A_n\}$ is distributed as the measurable function θ in the sense of the eigenvalues, i.e., $\{A_n\} \sim_\lambda \theta$ if for every F continuous, real valued, and with bounded support we have*

$$\lim_{n\to\infty} \frac{1}{d_n} \sum_{j=1}^{d_n} F\Big(\lambda_j\big(A_n\big)\Big) = \frac{1}{m\{K\}} \int_K F\big(\theta(s)\big)\, ds, \tag{3}$$

where $\lambda_j\big(A_n\big)$, $j = 1, \ldots, d_n$, are the eigenvalues of A_n.

THEOREM 1.3 (see [37]). *Let f be a k variate Lebesgue integrable function defined over I_k. Then*

$$\{T_n(f)\} \sim_\lambda f.$$

This kind of result goes back to Szegö [13] for the case where the symbol is L^∞ (essentially bounded). The complete generalization to the most general L^1 setting is due to Tyrtyshnikov and Zamarashkin [37]. We also mention the paper [33] by Tilli for the technique used which is truly elegant and essentially based on the notion of matrix valued linear positive operators (see also [31, 27]). Further extensions concerning the class of test functions can be found in [28] where it is proved that the largest possible class of test functions is made by continuous functions defined on the whole real axis and satisfying a growth condition of the form $|F(z)| \le B + A|z|$, for some given constants A and B. Finally, very exotic generalizations concerning nonfunctional symbols (measures, distributions, integrals in principal value, etc.) are recently considered by Tyrtyshnikov and Zamarashkin, Tilli, and Trench (see [38, 35, 34]).

The interesting fact is that all these properties stand also for sequences of preconditioned matrices of the form $\{T_n^{-1}(g)T_n(f)\}$ with nonnegative and not identically zero g. More precisely, the following counterpart of Theorem 1.1 holds.

THEOREM 1.4. *Let f and g be two k variate Lebesgue integrable functions defined over I_k and assume that g is nonnegative with positive essential supremum. Set $h = g^{-1}f$; then we have the following:*

1. [22] *$T_n(g)$ is Hermitian positive definite and the eigenvalues of $T_n^{-1}(g)T_n(f)$ are contained in (r, R) if $r < R$ and*

$$r = \text{essinf } (h)$$

and

$$R = \text{esssup } (h);$$

if $r = R$, then h is identically constant and $T_n^{-1}(g)T_n(f) = rI$, with I denoting the identity matrix of size $\widehat{n}$.

2. [9, 21] *If we denote by $\lambda_{\min}(n)$ and by $\lambda_{\max}(n)$ the minimal and the maximal eigenvalues of $T_n^{-1}(g)T_n(f)$, then*

$$\lim_{N\to\infty} \lambda_{\min}(n) = r, \quad \lim_{N\to\infty} \lambda_{\max}(n) = R.$$

Furthermore, we have a counterpart to Theorem 1.3 as well.

THEOREM 1.5 (see [21]). *Let f and g be two k variate Lebesgue integrable functions defined over I_k and assume that g is nonnegative with positive essential supremum. Set $h = g^{-1}f$; then we have*

$$\{T_n^{-1}(g)T_n(f)\} \sim_\lambda h.$$

We just mention that these global spectral results (Theorems 1.3 and 1.5) are of interest in asymptotic (numerical) linear algebra and, for instance, they play a central role to prove precise asymptotic bounds on the convergence rate of (preconditioned) conjugate gradient-like algorithms (see the recent work by Beckermann and Kuijlaars [1]).

In conclusion, it seems that every property of Toeplitz sequences is enjoyed by preconditioned Toeplitz sequences as well. The following section, section 2, is devoted to showing that this is "essentially" false since an important property holding for Toeplitz sequence is violated by concrete examples of preconditioned Toeplitz sequences. We recall that the closure of the union of all the spectra of the matrix sequence $\{T_n(f)\}$ coincides with the convex hull of the essential range of f, i.e., with the interval $[m, M]$ (see [40]), where m and M are the constants indicated in Theorem 1.1. On the other hand, we prove that the closure of the union of all the spectra of the preconditioned matrix sequence $\{T_n^{-1}(g)T_n(f)\}$, with given $f, g \in L^1$, $g \geq 0$ and not identically zero, can have gaps if the essential range of f/g is not connected.

Section 3 then discusses practical (positive) consequences on the preconditioning technique known as band Toeplitz preconditioning, in both positive definite (ill-conditioned) and indefinite (ill-posed) settings.

2. Statement and proof of the result. We introduce some useful notation. For a given measurable function h the set $\mathcal{ER}(h)$ denotes the essential range and is defined as

$$\{y \in \mathbf{R} : \ \forall \epsilon > 0, \ m\{x : \ h(x) \in (y - \epsilon, y + \epsilon)\} > 0\},$$

where $m\{\cdot\}$ denotes the Lebesgue measure on $\mathbf{R}^k$ with integer positive k. It is evident that $\mathcal{ER}(h)$ is a closed set. By $Coh(X)$ we indicate the convex hull of a set $X \subset \mathbf{R}$. Furthermore, for a given sequence $\{A_n\}$ of matrices of increasing size d_n, we define $U(\{A_n\})$ as the closure of the union over n of all the spectra of A_n. By definition $U(\{A_n\})$ is a closed set. Finally a set X is a cluster for $\{A_n\}$ if for every ϵ extension of X all the eigenvalues of A_n belong to that extended set except at most $o(d_n)$ outliers.

From the results of the preceding section (Theorems 1.1–1.5) we deduce that

$$\mathcal{ER}(f) \subset U(\{T_n(f)\}) \subset Coh[\mathcal{ER}(f)]$$

for every $f \in L^1$ and that

$$\mathcal{ER}(h) \subset U(\{T_n^{-1}(g)T_n(f)\}) \subset Coh[\mathcal{ER}(h)]$$

for every $f, g \in L^1$, g nonnegative, not identically zero and with $h = g^{-1}f$. A result by Widom tells us that $U(\{T_n(f)\}) = Coh[\mathcal{ER}(f)]$ for every $f \in L^1$, while some counterexamples, with simple f and g, show that $U(\{T_n^{-1}(g)T_n(f)\})$ can be not connected and indeed can be a strict subset of $Coh[\mathcal{ER}(h)]$.

THEOREM 2.1. *Let f be a k variate Lebesgue integrable function defined over I_k and assume that the essential range $\mathcal{ER}(f)$ is not connected. Then the following facts hold:*

1. *$\mathcal{ER}(f)$ is a cluster for the eigenvalues of $\{T_n(f)\}$,*
2. *$U(\{T_n(f)\}) = Coh[\mathcal{ER}(f)]$.*

Proof. The first statement is a direct consequence of Theorem 1.3, while the second is a known result (see [40]). □

THEOREM 2.2. *Let f and g be two k variate Lebesgue integrable functions defined over I_k and assume that g is nonnegative with positive essential supremum. Set $h = g^{-1}f$; then the following facts hold:*

1. *$\mathcal{ER}(h)$ is a cluster for the eigenvalues of $\{T_n^{-1}(g)T_n(f)\}$,*
2. *f and g exist such that both the essential range $\mathcal{ER}(h)$ and $U(\{T_n^{-1}(g)T_n(f)\})$ are not connected, i.e., $U(\{T_n^{-1}(g)T_n(f)\}) \neq Coh[\mathcal{ER}(h)]$.*

Proof. The first statement follows from Theorem 1.5. For the second we construct some examples.

First construction. Let $f = \sin(x)$, $g = \sin^2(x)$, $T_1 := T_n(f)$, and $T_2 := T_n(g)$. We will prove that for even n the eigenvalues of the preconditioned matrix $T_2^{-1}T_1$ are contained in the intervals $(-\infty, -1)$ and $(1, \infty)$. For odd n there is an additional eigenvalue at $\lambda = 0$. Therefore $U(\{T_n^{-1}(g)T_n(f)\}) \subset (-\infty, -1] \bigcup \{0\} \bigcup [1, \infty)$ while $Coh[\mathcal{ER}(h)]$, $h = g^{-1}f = \frac{1}{\sin(x)}$, coincides with the whole real line since

$$\lim_{x \to 0^+} h(x) = \infty$$

and

$$\lim_{x \to 0^-} h(x) = -\infty.$$

Therefore a remarkable thing is that, up to at most the outlier $\lambda = 0$, the set $U(\{T_n^{-1}(g)T_n(f)\})$ is the same as $\mathcal{ER}(h)$.

In the following we will use Matlab-like notation. More specifically, Toeplitz(r) denotes the Hermitian Toeplitz matrix whose first row is r, and tridiag$[a_1, a_0, a_{-1}]$,

pentadiag$[a_2, a_1, a_0, a_{-1}, a_{-2}]$, etc., denote a Toeplitz matrix which is tridiagonal, pentadiagonal, etc. and whose (j,k) entry is given by a_{j-k}. We get

$$T_1 = T_n(\sin(x)) = \frac{1}{2}\text{Toeplitz}(0, \mathbf{i}, 0, \ldots, 0)$$

and

$$T_2 = T_n(\sin^2(x)) = \frac{1}{2}T_n(1-\cos(2x)) = \frac{1}{2}\text{Toeplitz}(1, 0, -0.5, 0, \ldots, 0)\ .$$

Taking into account the relation

$$T_1^2 = T_n^2(\sin(x)) = \frac{1}{4}\text{tridiag}(-\mathbf{i}, 0, \mathbf{i})^2 = \frac{1}{4}\begin{pmatrix} 1 & 0 & -1 & & & \\ 0 & 2 & 0 & \ddots & & \\ -1 & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & 2 & 0 & -1 \\ & & -1 & 0 & 2 & 0 \\ & & & -1 & 0 & 1 \end{pmatrix},$$

it holds that

$$T_1^2 - T_2 = -\frac{1}{4}(e_1 e_1^T + e_n e_n^T) =: -\frac{1}{4}E$$

with $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ being the jth canonical unit vector. Hence, for even n

$$f_1 := T_1^{-1}e_1 = -2\mathbf{i}\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad f_n := T_1^{-1}e_n = 2\mathbf{i}\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix},$$

with $f_n^H f_n = f_1^H f_1 = 2n$, $f_1^H e_1 = f_n^H e_n = 0$, and $f_1^H e_n = 2\mathbf{i} = -f_n^H e_1$.

First argument (Gershgorin). For even n, the eigenvalues of $T_2^{-1}T_1$ are the inverse of those of $T_1^{-1}T_2$, and the expression of $T_1^{-1}T_2$ can be computed explicitly. Indeed

$$\begin{aligned} T_1^{-1}T_2 &= T_1^{-1}\left(T_1^2 + \frac{1}{4}E\right) \\ &= T_1 + \frac{1}{4}T_1^{-1}E \\ &= T_1 + \frac{1}{4}(f_1 e_1^H + f_n e_n^H), \end{aligned}$$

and therefore all the Gershgorin disks are centered in zero with radius 1.5 except for the first and the last whose radii equal 1. Consequently, since the resulting matrix is irreducible, the application of the first and third Gershgorin theorems (see, e.g., [12, 39]) tells one that the eigenvalues of $T_1^{-1}T_2$ belong to $(-1.5, 1.5)$. Finally all the eigenvalues of $T_2^{-1}T_1$ lie in $(-\infty, -2/3)$ or in $(2/3, \infty)$. For n odd we have that T_1 is

singular and therefore we have only an extra eigenvalue equal to 0 (for the others it is enough to use a perturbation argument).

It is evident that this argument is sufficient to prove the existence of gaps in the set $U(\{T_n^{-1}(g)T_n(f)\})$. However, we can obtain a tight estimate by using further tools.

Second argument. Let n be even. We denote by λ the eigenvalues of $T_2^{-1}T_1$ and by F the matrix $T_1^{-1}ET_1^{-1} = f_1 f_1^H + f_n f_n^H$. Now it holds that

$$T_2 = T_1^2 + \frac{1}{4}E = T_1\Big(I + \frac{1}{4}F\Big)T_1.$$

Therefore, λ is also an eigenvalue of

$$T_1^{-1}x = \lambda\Big(I + \frac{1}{4}F\Big)x,$$

or $1/\lambda$ is an eigenvalue of

$$T_1 x = \frac{1}{\lambda}\Big(I + \frac{1}{4}F\Big)^{-1}x = \frac{1}{\lambda}\Big(I + \tau F\Big)x$$

with $\tau = -1/(2n+4)$.

Now we consider this last eigenvalue problem for variable $\tilde{\tau}$, i.e.,

$$T_1 x = \frac{1}{\lambda}\Big(I + \tilde{\tau}F\Big)x.$$

For $\tilde{\tau} = 0$ the eigenvalues are given by those of T_1^{-1}, and therefore for all eigenvalues it holds that $|\lambda| > 1$. By contradiction, we suppose that for $\tilde{\tau} = \tau = -1/(2n+4)$ there exists an eigenvalue $\lambda(\tau)$ such that $|\lambda(\tau)| < 1$. Therefore, we have to find values $\tilde{\tau} \in (-1/(2n+4), 0)$ for which $|\lambda(\tilde{\tau})| = 1$. Consequently we are especially interested in the values $\tilde{\tau}$ that allow an eigenvalue 1 or an eigenvalue -1. We consider the case of $\lambda = 1$: therefore,

$$T_1 x = \big(I + \tilde{\tau}F\big)x.$$

Then $\tilde{\tau}$ is an eigenvalue of the generalized eigenvalue problem

$$(I - T_1)x = -\tilde{\tau}Fx.$$

Now, this problem has only two solutions $\tilde{\tau}_\pm$. Furthermore, the matrix $A := I - T_1$ is a diagonal similarity transformation of tridiag$(-0.5, 1, -0.5)$ with diagonal matrix $D = \mathrm{diag}(\mathbf{i}^k)_{k=0,\dots,n-1}$. Therefore, $A = I - T_1 = \tilde{S}^H \Lambda \tilde{S}$, where $\tilde{S}$ denotes the so-called modified sine-transform and Λ is a diagonal matrix with positive eigenvalues. The solution of $Ag_1 = f_1$ is given by

$$g_1 = \frac{1}{n+1}\,(\,2n \quad 4n\mathbf{i} \quad 4-2n \quad 8\mathbf{i} \quad 2n-8 \quad 4(n-2)\mathbf{i} \quad 12-2n \quad 16\mathbf{i} \quad \cdots\,)^H.$$

In view of these properties the two eigenvalues $\tilde{\tau}_\pm$ are given by

$$-\frac{1}{\tilde{\tau}_\pm} = |f_1^H A^{-1} f_1| \pm |f_1^H A^{-1} f_n| = |f_1^H g_1| \pm |f_n^H g_1| = \frac{2n(n+2)}{n+1} \pm \frac{8\lfloor \frac{n+2}{4} \rfloor}{n+1}.$$

First case $4|n$: Then we get $\tilde{\tau}_- < \tilde{\tau}_+ < \tau$, and therefore $\lambda = 1$ can be no eigenvalue.

Second case $4|(n+2)$: Then we get $\tilde{\tau}_- < \tilde{\tau}_+ = \tau$, and therefore 1 is no eigenvalue for all $\tilde{\tau}$ with $\tau < \tilde{\tau} < 0$. Hence, for τ the positive eigenvalues satisfy $\lambda \geq 1$.

The same analysis can be used for $\lambda = -1$. Hence, for τ it holds that $|\lambda| \geq 1$ and the case $|\lambda| = 1$ happens only for $4|(n+2)$.

For odd n a similar analysis can be developed based on the regular matrix $T_1 + \rho E$ for $\rho = (\sqrt{5}-2)/2$.

Second construction. Define $T_\alpha := T_1 + \alpha T_2$. Then for $\alpha \neq \pm 1$ the eigenvalues of $T_\alpha^{-1} T_\beta$ satisfy the following conditions: for $\alpha > 1$ and $\beta < \alpha$ the eigenvalues are contained in

$$\left(-\infty, \frac{1-\beta}{1-\alpha}\right] \bigcup \left\{\frac{\beta}{\alpha}\right\} \bigcup \left[\frac{1+\beta}{1+\alpha}, 1\right)$$

with $\frac{1-\beta}{1-\alpha} < \frac{\beta}{\alpha} < \frac{1+\beta}{1+\alpha} < 1$. For $\alpha > 1$ and $\beta > \alpha$ the eigenvalues are contained in

$$(-\infty, -1) \bigcup \left(1, \frac{1-\beta}{1-\alpha}\right]$$

with $1 < \frac{1+\beta}{1+\alpha} \leq \frac{\beta}{\alpha} < \frac{1-\beta}{1-\alpha}$.

For $-1 < \alpha < 1$ and $\beta < \alpha$ the eigenvalues are contained in

$$\left[\frac{1+\beta}{1+\alpha}, 1\right) \bigcup \left(1, \frac{1-\beta}{1-\alpha}\right] \bigcup \left\{\frac{\beta}{\alpha}\right\}$$

with $\frac{\beta}{\alpha} > \frac{1-\beta}{1-\alpha}$ for negative α, respectively, $\frac{\beta}{\alpha} < \frac{1+\beta}{1+\alpha}$ for positive α. For $-1 < \alpha < 1$ and $\beta > \alpha$ the eigenvalues are contained in

$$\left[\frac{1-\beta}{1-\alpha}, 1\right) \bigcup \left(1, \frac{1+\beta}{1+\alpha}\right] \bigcup \left\{\frac{\beta}{\alpha}\right\}$$

with $\frac{\beta}{\alpha} < \frac{1-\beta}{1-\alpha}$ for negative α, respectively, $\frac{\beta}{\alpha} > \frac{1+\beta}{1+\alpha}$ for positive α. For $\alpha < -1$ and $\beta < \alpha$ the spectrum is contained in

$$(-\infty, 1) \bigcup \left(1, \frac{1-\beta}{1-\alpha}\right] \bigcup \left\{\frac{\beta}{\alpha}\right\} \bigcup \left[\frac{1+\beta}{1+\alpha}, \infty\right)$$

with $\frac{1-\beta}{1-\alpha} < \frac{\beta}{\alpha} < \frac{1+\beta}{1+\alpha}$. For $\alpha < -1$ and $\beta > \alpha$ the spectrum is contained in

$$\left(-\infty, \frac{1+\beta}{1+\alpha}\right] \bigcup \left\{\frac{\beta}{\alpha}\right\} \bigcup \left[\frac{1-\beta}{1-\alpha}, 1\right) \bigcup (1, \infty)$$

with $\frac{1+\beta}{1+\alpha} < \frac{\beta}{\alpha} < \frac{1-\beta}{1-\alpha}$. Thereby, the eigenvalue β/α can occur only for odd n.

Indeed the argument is simple: if λ is an eigenvalue of $T_\alpha^{-1} T_\beta$, then

$$\mu = \frac{\alpha\lambda - \beta}{1-\lambda}$$

is an eigenvalue of $T_2^{-1} T_1$. Therefore, $\mu \in (-\infty, -1) \bigcup \{0\} \bigcup (1, \infty)$, and $\mu = 0$ occurs only for odd n.

Finally we remark that some of the above tools and statements can be extended to the multilevel case. For instance, the machinery considered in the *first construction*

can be easily generalized to a multivariate context, thus proving the occurrence of the gap phenomenon in the multilevel case as well. □

Remark (some relation between the example in Gershgorin's argument and orthogonal polynomials). The beginning is the same as for Gershgorin's argument. For even n, the eigenvalues of $T_2^{-1}T_1$ are the inverse of those of $T_1^{-1}T_2$, where

$$T_1^{-1}T_2 = T_1 + \frac{1}{4}T_1^{-1}E.$$

For every given paramenter $\epsilon > 0$, let us define $E(\epsilon)$ as the positive definite diagonal matrix whose diagonal entries are $(E(\epsilon))_{1,1} = (E(\epsilon))_{n,n} = 1$ and $(E(\epsilon))_{j,j} = \epsilon$, $j = 2, \ldots, n-1$. Consequently we define

$$K(\epsilon) = T_1 + \frac{1}{4}T_1^{-1}E(\epsilon)$$

so that

$$\lim_{\epsilon \to 0} K(\epsilon) = T_1^{-1}T_2, \qquad \lim_{\epsilon \to 0} E(\epsilon) = E.$$

By considering the similarity relation $\sim_{\mathrm{S}}$ (that, of course, maintains the eigenvalues unchanged), we have

$$\begin{aligned} K(\epsilon) &= E^{-1/2}(\epsilon)\left[E^{1/2}(\epsilon)K(\epsilon)E^{-1/2}(\epsilon)\right]E^{1/2}(\epsilon) \\ &\sim_{\mathrm{S}} E^{1/2}(\epsilon)K(\epsilon)E^{-1/2}(\epsilon) \\ &= E^{1/2}(\epsilon)T_1E^{-1/2}(\epsilon) + \frac{1}{4}E^{1/2}(\epsilon)T_1^{-1}E^{1/2}(\epsilon) \\ &= E^{1/2}(\epsilon)T_1E^{-1/2}(\epsilon) + \frac{1}{4}E^{1/2}T_1^{-1}E^{1/2} + R(\epsilon), \end{aligned}$$

where $R(\epsilon)$ is defined implicitly. For our purposes it is enough to observe that each entry of $R(\epsilon)$ is $O(\sqrt{\epsilon})$ with constant depending only on n. Moreover, the matrix $\Theta(\epsilon) = E^{1/2}(\epsilon)T_1E^{-1/2}(\epsilon) + \frac{1}{4}E^{1/2}T_1^{-1}E^{1/2}$ has the following explicit expression:

$$\Theta(\epsilon) = \frac{1}{2}\begin{pmatrix} 0 & \mathbf{i}\epsilon^{-1/2} & 0 & \cdots & 0 & \mathbf{i} \\ -\mathbf{i}\epsilon^{1/2} & \ddots & \mathbf{i} & \ddots & & 0 \\ 0 & -\mathbf{i} & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \mathbf{i} & 0 \\ 0 & & \ddots & -\mathbf{i} & \ddots & \mathbf{i}\epsilon^{1/2} \\ -\mathbf{i} & 0 & \cdots & 0 & -\mathbf{i}\epsilon^{-1/2} & 0 \end{pmatrix}.$$

The crucial point is the computation of the characteristic polynomial of $\Theta(\epsilon)$ that, as we will see, is not dependent on ϵ and is essentially the same as the nth Chebyshev polynomial of the first kind. Indeed we have

$$\begin{aligned} \det(\theta I - \Theta(\epsilon)) &= \theta\ \det(\theta I - T_{n-1}(\sin(x))) - \frac{1}{4}\det(\theta I - T_{n-2}(\sin(x))) \\ &\quad + (-1)^{n-1}\left(-\frac{\mathbf{i}}{2}(-1)^{n-2}\frac{\mathbf{i}}{2}\right)\det(\theta I - T_{n-2}(\sin(x))) \\ &= \theta\ \det(\theta I - T_{n-1}(\sin(x))) - \frac{1}{2}\det(\theta I - T_{n-2}(\sin(x))). \end{aligned}$$

Since $T_k(\sin(x))$ is similar to $T_k(\cos(x))$, it is evident that $\det(\theta I - \Theta(\epsilon))$ formally coincides with $\det(\theta I - X))$, where

$$X = \frac{1}{2}\begin{pmatrix} 0 & \sqrt{2} & & & & \\ \sqrt{2} & \ddots & 1 & & & \\ & 1 & & & \ddots & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & 1 \\ & & & & 1 & 0 \end{pmatrix}.$$

Since X is the Jacobi matrix of the Chebyshev polynomial of the second kind, it follows that the eigenvalues of $\Theta(\epsilon)$ are given by

$$\cos\left(\frac{\pi j}{n+1}\right), \quad j = 1, \ldots, n. \tag{4}$$

In conclusion, a further qualitative argument tells us that the spectrum of $T_n^{-1}(g)T_n(f)$ is contained in the interval $(-\infty, -1]\bigcup[1, \infty)$ with the additional eigenvalue 0 only in the case of odd n.

Remark (nonpolynomial examples). After the examples considered in the above proof, one may ask if the presence of gaps can be obtained working only with simple polynomials, i.e., with simple banded Toeplitz matrices. The answer is no and, in fact, starting from our basic example in Theorem 2.2, we easily construct a whole family of examples where one of the matrices is full: let f be a nonnegative function such that the essential supremum of $h = f(x)/\sin^2(x)$ is strictly positive. Then all the eigenvalues of $\{T_n^{-1}(f)T_n(\sin(x))\}$ belong to

$$\left(-\infty, -\|h\|_\infty^{-1}\right]\bigcup\{0\}\bigcup\left[\|h\|_\infty^{-1}, \infty\right).$$

To prove the previous statement we consider the inverse matrices $\{T_n^{-1}(\sin(x))T_n(f)\}$ and prove that all their eigenvalues are in the bounded interval $[-\|h\|_\infty, \|h\|_\infty]$. In order to have this it is enough to consider the generalized Rayleigh quotient: more precisely, every eigenvalue λ of $T_n^{-1}(\sin(x))T_n(f)$ belongs to the set

$$\begin{aligned}
\lambda \in & \left[\min_{y\neq 0}\frac{y^H T_n^{1/2}(f)T_n^{-1}(\sin(x))T_n^{1/2}(f)y}{y^H y}, \max_{y\neq 0}\frac{y^H T_n^{1/2}(f)T_n^{-1}(\sin(x))T_n^{1/2}(f)y}{y^H y}\right] \\
= & \left[\min_{y\neq 0}\frac{y^H T_n^{-1}(\sin(x))y}{y^H T_n^{-1}(f)y}, \max_{y\neq 0}\frac{y^H T_n^{-1}(\sin(x))y}{y^H T_n^{-1}(f)y}\right] \\
\subset & \left[\min_{y\neq 0}\frac{y^H T_n^{-1}(\sin(x))y}{y^H T_n^{-1}(\sin^2(x))y}\cdot\max_{y\neq 0}\frac{y^H T_n^{-1}(\sin^2(x))y}{y^H T_n^{-1}(f)y},\right. \\
& \left.\max_{y\neq 0}\frac{y^H T_n^{-1}(\sin(x))y}{y^H T_n^{-1}(\sin^2(x))y}\cdot\max_{y\neq 0}\frac{y^H T_n^{-1}(\sin^2(x))y}{y^H T_n^{-1}(f)y}\right] \\
\subset & \left[-1\cdot\|h\|_\infty, 1\cdot\|h\|_\infty\right] = \left[-\|h\|_\infty, \|h\|_\infty\right].
\end{aligned}$$

An interesting practical consequence of the above results concerns the indefinite preconditioning. Indeed we can prove that there exist cases where the eigenvalues of $T_n^{-1}(g)T_n(f)$ belong to the range of f/g with no outliers even when both f and g are

indefinite. Following the analysis in [15], we know that this is not a trivial result, so we report it as an independent corollary.

COROLLARY 2.3. *For even n and $\gamma \geq 0$ the spectrum of $T_n^{-1}(\sin(x))T_n(\sin(x) + \gamma \sin^2(x))$ is contained in the interval $[1-\gamma, 1+\gamma]$.*

3. A discussion on the practical impact of the result. First we briefly mention that Toeplitz matrices are of great interest in many fields of pure and applied mathematics (see, e.g., [6, 40] and references therein). In some of these applications, large Toeplitz linear systems have to be solved in real time, and consequently, it is crucial to have fast solvers.

In the positive one level case a lot of optimal iterative solvers are known [6] (in the ill-conditioned case as well) based on band [4, 9, 20, 24] and matrix algebra [8, 7, 14, 23, 17] preconditioners. Some specialized multigrid strategies [10, 5, 25] are available and can be very efficient: their implementation is more tricky but their optimal convergence holds in the multilevel setting as well. We stress that the latter remark is not trivial since the matrix algebra preconditioning loses optimality when the number of levels k exceeds 1 (refer to [32, 26]). Furthermore, difficult problems are encountered in the indefinite setting and in the case of multilevel structures. For these multilevel/indefinite problems, one of the more promising strategies is based on a band Toeplitz preconditioning (see [18, 19, 30, 15]): here the main problem is that we know where most of the eigenvalues are contained (see [15]) but until now we did not have good information on the position of the outliers. Therefore the result of the preceding section gives a positive message since it informs us that it is possible to get a stronger control on the position of the possible outliers. In particular this is very important in the nondefinite case, where often the range of the function f/g is not connected and has a positive part and a negative part both well separated from zero. The possibility of restricting the outliers from approaching zero as n becomes large is very important for the convergence feature of methods such as the MINRES or the GMRES.

3.1. Positive definite preconditioning for nondefinite problems. In this subsection, we revisit the numerical test performed in [19] in light of the new results on the "gaps" for preconditioned Toeplitz sequences. Moreover we also check the numerical behavior of the MINRES in order to give convincing evidence of the use of our theoretical results. In the following, H_n will denote the preconditioned matrix and $\Sigma(X)$ will denote the complete spectrum of the square matrix X.

Example 1. Let

$$f(x) \equiv x = \sum_{k=1}^{\infty} \frac{\mathbf{i}(-1)^k}{k}(e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I_1,$$

and choose $T_n(g)$ generated by

$$g(x) \equiv |x| = \frac{\pi}{2} - \sum_{k=1}^{\infty} \frac{1-(-1)^k}{\pi k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx}), \quad x \in I_1.$$

According to Theorem 2.2 we expect that the eigenvalues of $H_n = T_n^{-1}(g)T_n(f)$ form two clusters around -1 and 1 since $f/g = \text{sign}(x)$:
For $n = 16$ we have

$$\Sigma(H_n) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9997, \pm 0.9946, \pm 0.9287, \pm 0.4773\}. \tag{5}$$

TABLE 1
P[MINRES].

Size $= n$	16	64	64	128	256	512
Ex1	7	9	9	11	11	13
Ex2a	7	9	9	11	11	13
Ex2b	15	25	33	41	43	47
Ex3	8	10	11	13	17	19

For $n = 64$ we have

$$\Sigma(H_n) = \{\pm 1.000 \text{ (27 times)}, \pm 0.9995, \pm 0.9963, \pm 0.9737, \pm 0.8470, \pm 0.3830\}. \tag{6}$$

For solving the corresponding linear system, we apply the MINRES with stop criterion of the relative residual and with tolerance $\epsilon = 10^{-7}$. The resulting number of iterations is reported in first row of Table 1. The slight increase of the number of the outliers and of the spectral condition number demonstrated in (5)–(6) is perfectly reflected in the logarithmic-like growth of the number of iterations.

Example 2. Let

$$f(x) \equiv \text{sign}(x)x^2 = \sum_{k=1}^{\infty} \frac{\mathbf{i}}{\pi k}\left((-1)^k\pi^2 + \frac{2}{k^2}\big(1+(-1)^{(k+1)}\big)\right) (e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I_1;$$

we propose two different functions g_1 and g_2:

$$g_1(x) \equiv x^2 = \frac{\pi^2}{3} + 2\sum_{k=1}^{\infty} \frac{(-1)^k}{k^2}(e^{\mathbf{i}kx} + e^{-\mathbf{i}kx}), \quad x \in I_1,$$

$$g_2(x) = 2 - 2\cos(x), \quad x \in I_1.$$

According to the results of the previous section we expect that $\Sigma(H_n) = \Sigma(T_n^{-1}(g_1)T_n(f))$ forms two clusters around -1 and 1 since $f/g_1 = \text{sign}(x)$. For $n = 16$ we have

$$\Sigma(H_n) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9998, \pm 0.9961, \pm 0.9412, \pm 0.500\}. \tag{7}$$

For $n = 64$ we have

$$\Sigma(H_n) = \{\pm 1.000 \text{ (27 times)}, \pm 0.9997, \pm 0.9972, \pm 0.9787, \pm 0.8640, \pm 0.4002\}. \tag{8}$$

In the case of $H_n = T_n^{-1}(g_2)T_n(f)$ we expect (Theorem 2.2, part 1) that most of the eigenvalues belong to $\mathcal{ER}(f/g_2) = [-\pi^2/4, -1] \bigcup [1, \pi^2/4]$. For $n = 16$ we obtain

14 eigenvalues in $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$,
2 eigenvalues $= \pm 0.7078$ in $(-1, 1)$.

For $n = 64$ we have

60 eigenvalues in $[-\pi^2/4, -1] \bigcup [1, \pi^2/4]$,
2 eigenvalues $= \pm 0.9938$ in a small neighborhood of -1 and 1, respectively, and
2 eigenvalues $= \pm 0.5698$ in $(-1, 1)$.

Also in this case, we apply the MINRES with the same stopping criterion as before. The resulting number of iterations is reported in rows 2 and 3 of Table 1. The moderate increase of the number of the outliers and of the spectral condition number shown imply a logarithmic-like growth of the number of iterations. Of course, in the case of $g = g_2$ the absolute number of iterations is bigger due to fact that the tightest clustering set is not a finite number of nonzero points but is represented by the two nontrivial intervals $[-\pi^2/4, -1]$ and $[1, \pi^2/4]$.

Example 3. Let

$$f(x) \equiv e^x - 1 \quad = \quad \sum_{k=-\infty}^{\infty} \frac{(-1)^k(e^{\pi} - e^{-\pi})}{2\pi(1+k^2)} (1 + \mathbf{i}k)e^{\mathbf{i}kx} \; - 1, \quad x \in I_1,$$

and

$$g(x) \equiv |e^x - 1| \quad = \quad \sum_{k=-\infty}^{\infty} t_k + \frac{1 + \mathbf{i}k}{2\pi(1+k^2)} ((e^{\pi} - e^{-\pi})(-1)^k - 2)e^{\mathbf{i}kx}, \quad x \in I_1,$$

where t_k is $\frac{2\mathbf{i}}{\pi k}$ if k is odd and 0 elsewhere.

According to Theorem 2.2 we expect two clusters around -1 and 1 for the spectrum of $H_n = T_n^{-1}(g)T_n(f)$:
For $n = 16$ we have

$$\Sigma(H_n) = \{\pm 1.000 \text{ (4 times)}, 0.9950, -0.9987, 0.9741, -0.9740, 0.6755, -0.6842, 0.3395, -0.3384\}. \tag{9}$$

For $n = 64$ we have

$$\Sigma(H_n) = \{1.000 \text{ (27 times)}, -1 \text{ (26 times)}, 0.9999, -0.9998, 0.9993, -0.9978, 0.9950, -0.9824, 0.9710, -0.8764, 0.4212, -0.4076\}. \tag{10}$$

The numerical results reported in the last row of Table 1 are of the same type as in Example 1, and therefore we do not add further comments.

In the cases (5)–(6), (7)–(8), (9)–(10), f/g is the function $\operatorname{sign}(x)$ (because $g(x) = |f(x)|$ and $f(0) = 0$), and it is interesting to compare these spectra with the spectrum of $T_n(\operatorname{sign}(x))$. For $n = 16$ we have

$$\Sigma(T_n(\operatorname{sign}(x))) = \{\pm 1.000 \text{ (4 times)}, \pm 0.9995, \pm 0.9913, \pm 0.9013, \pm 0.4294\}.$$

For $n = 64$ we have

$$\Sigma(T_n(\operatorname{sign}(x))) = \{\pm 1.000 \text{ (26 times)}, \pm 0.9999, \pm 0.9993, \pm 0.9945, \pm 0.9636, \pm 0.8122, \pm 0.3487\}.$$

The similarities are very deep, but strangely enough, the behavior of the preconditioned sequences $\{H_n\}$ is better than that of $\{T_n(\operatorname{sign}(x))\}$ in the sense that the outliers (with respect to the cluster $\{\pm 1\}$) are more separated from zero in the case of the preconditioned sequences $\{H_n\}$. This apparently curious situation finds its explanation in the "gap phenomenon" just studied in the preceding section.

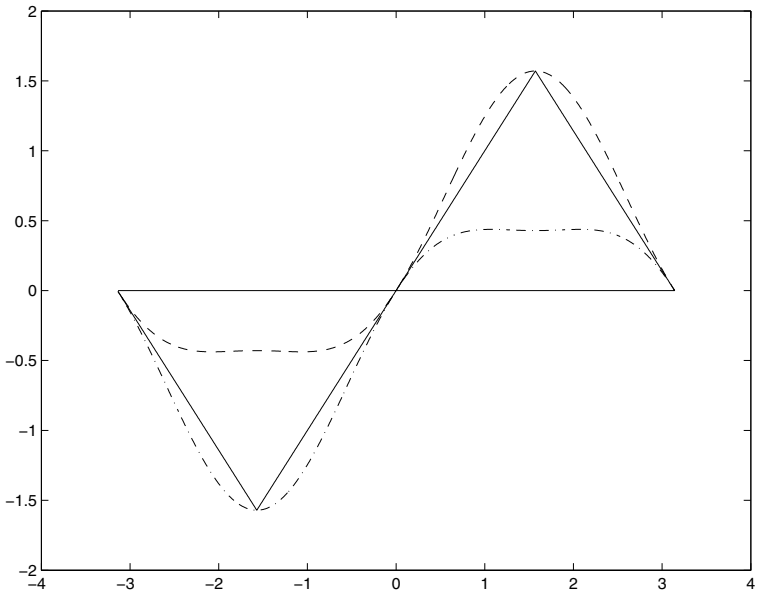


FIG. 1. *$f(x)$ from Example 4 with upper/lower bound $g(x) = \sin(x) + \gamma \sin^2(x)$, $\gamma = \pm(\pi/2 - 1)$.*

TABLE 2
P[CG/GMRES/QMR] with $\gamma = \pi/2 - 1$ for $f(x)$ from Example 4.

Size $= n$	$\lambda_{\min}$	$\lambda_{\max}$	κ_{sp}	#(it)	#(outliers)	cond$(T_n(f))$
16	0.7623	3.3237	4.3599	12/12/12	0	15.0665
32	0.7622	3.4902	4.5793	16/16/16	0	31.0667
64	0.7622	3.5750	4.6904	18/18/19	0	63.0667
128	0.7622	3.6175	4.7460	19/19/20	0	127.067
256	0.7622	3.6387	4.7738	19/20/20	0	255.067
512	0.7622	3.6492	4.7877	19/20/20	0	511.067

3.2. Nondefinite preconditioning for nondefinite problems.

Example 4. We consider the odd function $f(x)$ with $f(x) = x$ for $x \in [-\pi/2, \pi/2]$ and $f(x) = \pi/2 - x$ for $x \in [\pi/2, 3\pi/2]$, given by

$$f(x) \equiv x = -\frac{2}{\pi} \sum_{k=0}^{\infty} \frac{\mathbf{i}(-1)^k}{(2k-1)^2} (e^{\mathbf{i}kx} - e^{-\mathbf{i}kx}), \quad x \in I_1 \ .$$

Therefore f has zeros of order 1 at 0 and π. As preconditioners we choose $g(x) := \sin(x) + \gamma \sin^2(x)$. For $g_1(x) = \sin(x) + (\pi/2 - 1)\sin^2(x)$ and $g_2(x) = \sin(x) - (\pi/2 - 1)\sin^2(x)$, it holds that $g_2(x) \le f(x) \le g_1(x)$ and g_1 and g_2 also have zeros of order 1 at 0 and π; see Figure 1. Because of the second result in Theorem 2.2 the eigenvalues of $T_n^{-1}(g_1)T_n(g_2)$ are for even n contained in the interval

$$[4/\pi - 1, \pi/(4 - \pi)] \approx [0.2732, 3.6598].$$

For odd n there occurs an additional eigenvalue at -1. We expect that the eigenvalues of $T_n^{-1}(g_1)T_n(f)$ are contained in the interval $\mathcal{ER}(f/g_1) = [0.7622, \frac{\pi}{4-\pi}]$. Besides PCG and GMRES, in Table 2 we also consider the QMR-method because it can take advantage of the symmetry of the matrices [11].

TABLE 3
P[CG/GMRES/QMR] with $g(x) = \sin(x)$ for GMRES, respectively, $g(x) = \sin(x)+0.01\sin^2(x)$ for PCG and QMR and $f(x)$ from Example 4.

Size $= n$	$\lambda_{\min}$	$\lambda_{\max}$	κ_{sp}	#(it)	#(outliers)
16	1.0178	1.4557	1.4303	9/7/8	0
32	1.0053	1.5071	1.4992	9/8/9	0
64	1.0015	1.5371	1.5348	9/9/9	0
128	1.0004	1.5534	1.5528	9/9/9	0
256	1.0001	1.5619	1.5618	9/9/9	0
512	1.0000	1.5663	1.5663	9/9/9	0

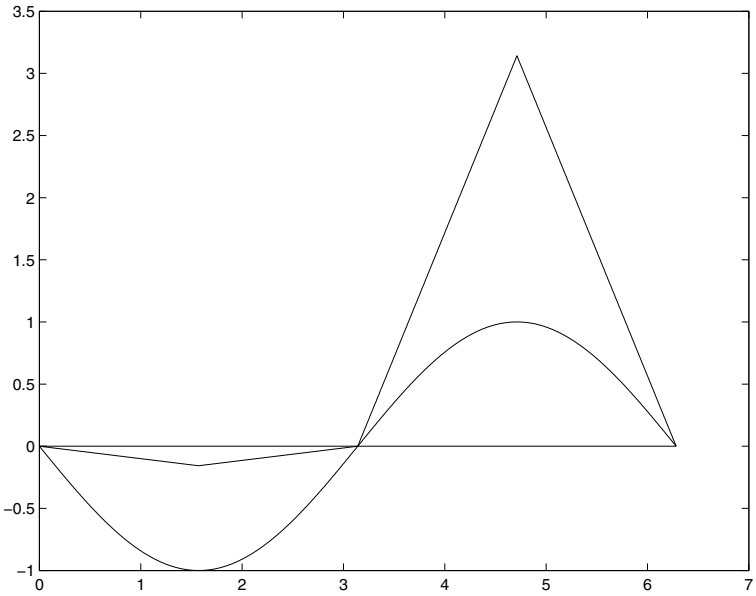


FIG. 2. *$f(x)$ from Example 5 with upper bound $g(x) = \sin(x)$.*

Next we choose $g(x) = \sin(x)$. Then the eigenvalues of $T_n^{-1}(g)T_n(f)$ are expected to be in the interval $\mathcal{ER}(f/g) = [1, \pi/2]$. In this form there always occurred a breakdown in PCG, respectively, QMR. Hence, for PCG and QMR we replaced $g(x)$ by $\tilde{g}(x) := \sin(x) + 0.01\sin^2(x)$; see Table 3.

Example 5. With $\rho = 0.1$ and $\delta = 2$ we define $f(x) := -\rho x$ in $[0, \pi/2]$, $f(x) := \rho(x - \pi)$ in $[\pi/2, \pi]$, $f(x) := \delta(x - \pi)$ in $[\pi, 3\pi/2]$, and $f(x) := \delta(2\pi - x)$ in $[3\pi/2, 2\pi]$ with $g(x) = -\sin(x)$. Then $g(x)$ is a lower bound of $f(x)$: $g(x) \le f(x)$, but we can get no trigonometric polynomial that is an upper bound; see Figure 2. For PCG and QMR a breakdown again occurred, and therefore in Table 4 we again replace $g(x)$ by $\tilde{g}(x) := \sin(x) + 0.01\sin^2(x)$.

To improve the convergence we also consider preconditioners of the form $g(x) = -\gamma\sin(x) - \beta\sin^2(x) - \alpha\sin^3(x)$. Especially for $\gamma = 1.05$, $\beta = 0.01$, and $\alpha = 0.5$ the number of iterations is reduced; see Table 5.

Finally, we consider two more examples.

Example 6. We take $f(x) = (x^2 + 5)\sin(x)$, $g(x) = \sin(x)$, and we observe that

$$\inf f/g = 5, \qquad \sup f/g = 14.8696.$$

TABLE 4
P[CG/GMRES/QMR] Example 5 *with* $g(x) = \sin(x)$ *for GMRES, respectively,* $g(x) = \sin(x) + 0.01\sin^2(x)$ *for PCG and QMR.*

Size $= n$	#(outliers < 0)	#(outliers > 0)	#(singular value outliers)	#(it)	cond$(T_n(f))$
16	1	1	4	18/15/17	201.56
32	1	3	4	27/25/26	1289.4
64	2	3	4	34/30/32	1446.4
128	2	3	4	35/31/35	957.3
256	2	4	4	37/34/38	2350.1
512	2	5	4	42/35/41	7728.8
1024				46/39/47	
2048				45/38/44	

TABLE 5
P[CG/GMRES/QMR] Example 5 *with* $g(x) = -1.05\sin(x) - 0.01\sin^2(x) - 0.5\sin^3(x)$.

Size $= n$	#(it)
16	16/14/16
32	22/19/21
64	23/21/23
128	24/21/23
256	26/23/26
512	26/24/30
1024	32/27/32
2048	32/26/32

TABLE 6
P[CG/GMRES] Example 6.

Size $= n$	$\lambda_{\min}$	$\lambda_{\max}$	κ_{sp}	#(it)	#(outliers)
16	5.1252	13.0190	2.5402	8/8	0
32	5.0346	13.8726	2.7554	12/12	0
64	5.0091	14.3511	2.8650	15/12	0
128	5.0023	14.6050	2.9196	13/13	0
256	5.0006	14.7359	2.9468	13/13	0
512	5.0002	14.8024	2.9604	14/13	0

In this case it is worth mentioning that the range of f/g contains, perfectly, all the spectra (with no outliers): the related experimental behavior is reported in Table 6, with regard to both the preconditioned GMRES and the CG method. It is evident that both methods are optimal.

Example 7. Setting $f(x) = (\beta x^2(2 - 2\cos(x) + \alpha)\sin(x) + \gamma(2 - 2\cos(x))$ with $g(x) = \sin(x)$, $\alpha = 1$, $\beta = 0.1$, and $\gamma = 0.2$, we have

$$\inf f/g = 0.63, \qquad \sup f/g = 2.29.$$

Here we observe a unique outlier with respect to the interval $[0.63, 2.29]$ described by the range of f/g. However this unique outlier seems to converge to a positive constant, and therefore it is unable to spoil the numerical behavior of the considered iterative techniques. The associated experiments are reported in Table 7 with regard to both the preconditioned GMRES and the CG method.

TABLE 7
P[CG/GMRES] Example 7.

Size $= n$	$\lambda_{\min}$	$\lambda_{\max}$	κ_{sp}	#(it)	#(outliers)
16	0.5146	2.2336	4.3397	13/13	1
32	0.5146	2.2489	4.3696	15/15	1
64	0.5146	2.2526	4.3767	16/15	1
128	0.5146	2.2540	4.3794	16/15	1
256	0.5146	2.2543	4.3801	16/15	1
512	0.5146	2.2544	4.3803	16/15	1

REFERENCES

[1] B. BECKERMANN AND A. KUIJLAARS, *Superlinear convergence of conjugate gradients*, SIAM J. Numer. Anal., 39 (2001), pp. 300–329.

[2] A. BÖTTCHER AND S. GRUDSKY, *On the condition numbers of large semi-definite Toeplitz matrices*, Linear Algebra Appl., 279 (1998), pp. 285–301.

[3] A. BÖTTCHER AND B. SILBERMANN, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, 1999.

[4] R. H. CHAN, *Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions*, IMA J. Numer. Anal., 11 (1991), pp. 333–345.

[5] R. H. CHAN, Q. CHANG, AND H. SUN, *Multigrid method for ill-conditioned symmetric Toeplitz systems*, SIAM J. Sci. Comput., 19 (1998), pp. 516–529.

[6] R. H. CHAN AND M. NG, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[7] R. H. CHAN AND G. STRANG, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[8] T. F. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[9] F. DI BENEDETTO, G. FIORENTINO, AND S. SERRA-CAPIZZANO, *C.G. Preconditioning for Toeplitz Matrices*, Comput. Math. Appl., 25 (1993), pp. 35–45.

[10] G. FIORENTINO AND S. SERRA-CAPIZZANO, *Multigrid methods for Toeplitz matrices*, Calcolo, 28 (1991), pp. 283–305.

[11] R. FREUND AND N. NACHTIGAL, *A new Krylov-subspace method for symmetric indefinite linear systems*, in Proceedings of the 14th IMACS World Congress on Computational and Applied Mathematics, W. F. Ames, ed., IMACS, 1994, pp. 1253–1256.

[12] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, 1983.

[13] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, 2nd ed., Chelsea, New York, 1984.

[14] T. HUCKLE, *Some aspects of circulant preconditioners*, SIAM J. Sci. Comput., 14 (1993), pp. 531–541.

[15] T. HUCKLE, S. SERRA-CAPIZZANO, AND C. TABLINO-POSSIO, *Preconditioning strategies for Hermitian nondefinite Toeplitz linear systems*, SIAM J. Sci. Comput., 25 (2004), pp. 1633–1654.

[16] D. NOUTSOS, S. SERRA CAPIZZANO, AND P. VASSALOS, *A preconditioning proposal for ill-conditioned Hermitian block Toeplitz systems*, Numer. Linear Algebra Appl., to appear.

[17] D. POTTS AND G. STEIDL, *Preconditioners for ill-conditioned Toeplitz systems constructed from positive kernels*, SIAM J. Sci. Comput., 22 (2000), pp. 1741–1761.

[18] S. SERRA-CAPIZZANO, *Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems*, BIT, 34 (1994), pp. 579–594.

[19] S. SERRA-CAPIZZANO, *Preconditioning strategies for Hermitian Toeplitz systems with nondefinite generating functions*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1007–1019.

[20] S. SERRA-CAPIZZANO, *Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems*, Math. Comp., 66 (1997), pp. 651–665.

[21] S. SERRA-CAPIZZANO, *An ergodic theorem for classes of preconditioned matrices*, Linear Algebra Appl., 282 (1998), pp. 161–183.

[22] S. SERRA-CAPIZZANO, *On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1998), pp. 109–129.

[23] S. Serra-Capizzano, *A Korovkin-type theory for finite Toeplitz operators via matrix algebras*, Numer. Math, 82 (1999), pp. 117–142.
[24] S. Serra-Capizzano, *How to choose the best iterative strategy for symmetric Toeplitz systems*, SIAM J. Numer. Anal., 36 (1999), pp. 1078–1103.
[25] S. Serra-Capizzano, *Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix sequences*, Numer. Math., 92 (2002), pp. 433–465.
[26] S. Serra-Capizzano, *Matrix algebra preconditioners for multilevel Toeplitz matrices are not superlinear*, Linear Algebra Appl., 343/344 (2002), pp. 303–319.
[27] S. Serra-Capizzano, *More inequalities and asymptotics on matrix valued linear positive operators: The noncommutative case*, in Toeplitz Matrices and Singular Integral Equations, A. Böttcher, I. Gohberg, and P. Junghanns, eds., Oper. Theory Adv. Appl. 135, Birkhäuser, Basel, 2002, pp. 286–308.
[28] S. Serra-Capizzano, *Test functions, growth conditions and Toeplitz matrices*, Rend. Circ. Mat. Palermo Ser. (2), 68 (2002), pp. 791–795.
[29] S. Serra-Capizzano, *How to choose the best iterative strategy for symmetric Toeplitz systems*, SIAM J. Numer. Anal., 36 (1999), pp. 1078–1103.
[30] S. Serra-Capizzano and P. Tilli, *Extreme singular values and eigenvalues of non Hermitian block Toeplitz matrices*, J. Comput. Appl. Math., 108 (1999), pp. 113–130.
[31] S. Serra-Capizzano and P. Tilli, *On unitarily invariant norms of matrix-valued linear positive operators*, J. Inequalities Appl., 7 (2002), pp. 309–330.
[32] S. Serra-Capizzano and E. Tyrtyshnikov, *Any circulant-like preconditioner for multilevel Toeplitz matrices is not superlinear*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 431–439.
[33] P. Tilli, *A note on the spectral distribution of Toeplitz matrices*, Linear and Multilinear Algebra, 45 (1998), pp. 147–159.
[34] P. Tilli, *Eigenvalues of Toeplitz Matrices*, talk presented at the 1st AMS-UMI Conference, Pisa, 2002.
[35] W. Trench, *Properties of some generalization of Kac-Murdock-Szegö matrices*, in Structured Matrices in Mathematics, Computer Science, and Engineering II, Contemp. Math. 281, V. Olshevskey, ed., AMS, Providence, RI, 2001, pp. 193–212.
[36] E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[37] E. Tyrtyshnikov and N. Zamarashkin, *Spectra of multilevel Toeplitz matrices: Advanced theory via simple matrix relationships*, Linear Algebra Appl., 270 (1998), pp. 15–27.
[38] E. Tyrtyshnikov and N. Zamarashkin, *Toeplitz eigenvalues of radon measures*, Linear Algebra Appl., 343 (2002), pp. 345–354.
[39] R. S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[40] H. Widom, *Toeplitz matrices*, in Studies in Real and Complex Analysis, I. Hirshman, Jr., ed., Stud. Math. 3, Math. Ass. Amer., Washington, DC, 1965, pp. 179–201.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 947–963

COMPUTATION OF THE SMALLEST EVEN AND ODD EIGENVALUES OF A SYMMETRIC POSITIVE-DEFINITE TOEPLITZ MATRIX*

A. MELMAN†

Abstract. We propose an algorithm to compute the smallest even and odd eigenvalues of a real symmetric positive-definite Toeplitz matrix, which is based on the factorization of the characteristic polynomial into an even and an odd polynomial. Newton's method is used to compute the smallest even and odd eigenvalues as the smallest roots of the even and odd characteristic polynomials, respectively.

Key words. Toeplitz matrix, even, odd, eigenvalue, characteristic polynomial, Newton's method

AMS subject classifications. 65F15, 15A18

DOI. 10.1137/S0895479803423354

1. Introduction.

The computation of the smallest eigenvalue of a positive-definite Toeplitz matrix continues to be of interest, mainly due to its application in signal processing. In this respect, we mention the paper by Pisarenko [21] and the many subsequent papers devoted to this subject, e.g., [7], [12], [14], [16], [17], [19], and [22].

Most of the methods in these papers compute the smallest eigenvalue by solving the so-called *secular* or *spectral* equation (see [10]). The present paper is motivated by one method that does not, namely, the one proposed in [16], in which the smallest eigenvalue is found by computing the smallest root of the characteristic polynomial.

Symmetric centrosymmetric matrices, of which Toeplitz matrices are a special case, have two kinds of eigenvalues: even and odd, and the methods we just mentioned can be divided into two groups: those that take this special spectral structure into account and those that do not. The papers in the former category include [12], [17], [19], and [22] and can, in fact, all be considered as "even-odd" versions of the method in [7], which belongs to the latter category. The method in [16] also belongs to the latter category and, since taking the spectral structure into account generally leads to better numerical methods, we have developed here the "even-odd" equivalent of this method. Throughout this paper, we will refer to the method from [16] as the "MB method," after its authors Nicola Mastronardi and Daniel Boley.

The MB method computes the smallest root of the characteristic equation of a symmetric positive-definite Toeplitz matrix, to which we will refer throughout as an "SPD" Toeplitz matrix, with Newton's method, which in this case converges monotonically from any point to the left of the smallest root. In our method, we will do the same, but for two different polynomials, namely, the even and the odd characteristic polynomials. It can be shown quite easily that the new method must be faster than the MB method, and this is borne out by the numerical experiments.

The paper is organized as follows. In section 2 we review the basic properties of Toeplitz matrices and introduce our notation. In sections 3 and 4 we derive the

*Received by the editors February 24, 2003; accepted for publication (in revised form) by L. Reichel September 17, 2003; published electronically April 21, 2004.
http://www.siam.org/journals/simax/25-4/42335.html
†Department of Mathematics and Computer Science, St. Mary's College, Moraga, CA 94575 (melman@stmarys-ca.edu).

results, which are used in section 5 for the construction of the algorithm. Numerical results are presented in section 6.

2. Preliminaries. The notation we introduce in this section will be used throughout the paper.

A matrix $T \in \mathbb{R}^{(n,n)}$ is said to be *Toeplitz* if its elements T_{ij} satisfy $T_{ij} = \rho_{j-i}$ for some vector $(\rho_0, \ldots, \rho_{n-1})^T \in \mathbb{R}^n$. Many early results about such matrices can be found in, e.g., [3], [6], and [8].

Toeplitz matrices are *persymmetric*; i.e., they are symmetric about their southwest-northeast diagonal. For such a matrix T, this is the same as requiring that $JTJ = T^T$, where J, the exchange matrix, is a matrix with ones on its southwest-northeast diagonal and zeros everywhere else. It is easy to see that the inverse of a persymmetric matrix is also persymmetric. We will concentrate on symmetric Toeplitz matrices, i.e., matrices which satisfy $JTJ = T$ and are therefore also centrosymmetric.

An *even vector* v is defined as a vector satisfying $Jv = v$ and an *odd vector* w as one that satisfies $Jw = -w$. If these vectors are eigenvectors, then their associated eigenvalues are called *even* and *odd*, respectively. It was shown in [6] that, given a real symmetric centrosymmetric matrix T of order n, there exists an orthonormal basis for $\mathbb{R}^n$, composed of $n - \lfloor n/2 \rfloor$ even and $\lfloor n/2 \rfloor$ odd eigenvectors of T, where $\lfloor \alpha \rfloor$ denotes the integral part of α.

Finally, we note that for any $\lambda \in \mathbb{R}$, the matrix $(T - \lambda I)$ is symmetric and centrosymmetric whenever T is. We will use these results in the special case of an SPD Toeplitz matrix.

For simplicity's sake, our notation will not specifically indicate the dimensions of the identity matrix I and the exchange matrix J. It will also not differentiate between vectors and scalars. Usually, the context suffices to make matters clear and in the few instances where it might not, we will specifically indicate the relevant dimensions.

Both as an illustration of our notation and because it will be useful later on, we point out that any symmetric Toeplitz matrix of dimension $n \times n$ can be written as

$$\begin{pmatrix} A & B \\ JBJ & A \end{pmatrix} \quad \text{or as} \quad \begin{pmatrix} A & s & B \\ s^T & \rho_0 & s^T J \\ JBJ & Js & A \end{pmatrix},$$

depending on whether n is even or odd, respectively. For even n, the blocks in the matrix T have $\frac{n}{2}$ rows and columns. For odd n, the blocks have $\frac{n-1}{2}$ rows and columns. The column vector s has dimension $\frac{n-1}{2}$.

The following result is a special case of Lemma 3 in [6] (our notation is slightly different).

LEMMA 2.1. *For a symmetric Toeplitz matrix T, defined by $(\rho_0, \ldots, \rho_{n-1})^T$ when n is even, the following holds:*

$$KTK^T = \begin{pmatrix} A - BJ & 0 \\ 0 & A + BJ \end{pmatrix}, \text{ with } K = \frac{1}{\sqrt{2}} \begin{pmatrix} I & -J \\ I & J \end{pmatrix}.$$

When n is odd, then

$$KTK^T = \begin{pmatrix} A - BJ & 0 & 0 \\ 0^T & \rho_0 & \sqrt{2}s^T \\ 0 & \sqrt{2}s & A + BJ \end{pmatrix}, \text{ with } K = \frac{1}{\sqrt{2}} \begin{pmatrix} I & 0 & -J \\ 0^T & \sqrt{2} & 0^T \\ I & 0 & J \end{pmatrix}.$$

The matrix K satisfies $KK^T = I = K^T K$. The matrix T can therefore be split into two parts. The eigenvalues associated with $A - BJ$ are odd and those associated

with the part containing $A + BJ$ are even. This means that the characteristic polynomial of T can be factored into two polynomials, one corresponding to the even and the other to the odd eigenvalues, i.e.,

$$\det(T - \lambda I) = \det(A - BJ - \lambda I)\det(A + BJ - \lambda I) \quad \text{(even dimension)}$$

and

$$\det(T - \lambda I) = \det(A - BJ - \lambda I)\det\begin{pmatrix} \rho_0 - \lambda & \sqrt{2}s^T \\ \sqrt{2}s & A + BJ - \lambda I \end{pmatrix} \quad \text{(odd dimension)}.$$

In both the even and the odd case, we can write this concisely as $p(\lambda) = p^e(\lambda)p^o(\lambda)$. Throughout this paper, the superscripts "e" and "o" refer to even and odd, respectively.

From now on we will denote by T_k a real SPD Toeplitz matrix of dimension $k \times k$ and its characteristic polynomial will be written as $p_k(\lambda) = p_k^e(\lambda)p_k^o(\lambda)$. We note that the index k refers to the matrix T_k and not necessarily to the degree of the polynomial to which it is attached.

The Cauchy interlacing theorem states that the eigenvalues of T_n interlace those of its principal submatrix T_{n-1}, and also that the even and odd eigenvalues of T_n interlace the even and odd eigenvalues, respectively, of its principal submatrix T_{n-2} (see, e.g., [8]). Just as in [16] it was assumed that the smallest eigenvalue of T_n is not an eigenvalue of T_{n-1}, we will assume throughout this paper that the smallest even or odd eigenvalue of T_n is not an eigenvalue of T_{n-2}.

As an illustration, Figure 1 shows the even and odd characteristic polynomials, as well as the characteristic polynomial itself for the symmetric Toeplitz matrix, generated by $(1, 2, 0.2, 0.002)$.

Both in the method from [16] and in ours, an important role is played by the so-called *Yule–Walker equations*. For an $n \times n$ SPD Toeplitz matrix T_n, defined by $(\rho_0, \rho_1, \ldots, \rho_{n-1})$, this system of linear equations is given by $T_n y^{(n)} = -t_n$, where $t_n = (\rho_1, \ldots, \rho_n)^T$. Durbin's algorithm solves this system by recursively computing the solutions to lower dimensional systems. We now describe a basic step of Durbin's algorithm, while referring to [11, pp. 194–196] for full details.

Assuming that the solution to $T_{k-1}y^{(k-1)} = -t_{k-1}$ is available, the algorithm computes the solution to $T_k y^{(k)} = -t_k$ as follows.

Compute $\bar{y}^{(k-1)}$, by which we denote the first $k-1$ components of $y^{(k)}$, and α_{k-1}, the last component of $y^{(k)}$, from

$$\begin{pmatrix} T_{k-1} & Jt_{k-1} \\ (Jt_{k-1})^T & \rho_0 \end{pmatrix}\begin{pmatrix} \bar{y}^{(k-1)} \\ \alpha_{k-1} \end{pmatrix} = -\begin{pmatrix} t_{k-1} \\ \rho_k \end{pmatrix},$$

which leads to

$$\bar{y}^{(k-1)} = T_{k-1}^{-1}(-t_{k-1} - \alpha_{k-1}Jt_{k-1}) = y^{(k-1)} + \alpha_{k-1}Jy^{(k-1)} \tag{1}$$

and

$$\alpha_{k-1} = -\rho_{k+1} - t_{k-1}^T J\bar{y}^{(k-1)} = -\frac{\rho_{k+1} + t_{k-1}^T Jy^{(k-1)}}{\rho_0 + t_{k-1}^T y^{(k-1)}}\,. \tag{2}$$

In addition, we define, as in [11], $\beta_k = \rho_0 + t_k^T y^{(k)}$. The following recursion then holds (see [11, p. 195]):

$$\beta_k = (1 - \alpha_{k-1}^2)\beta_{k-1}. \tag{3}$$

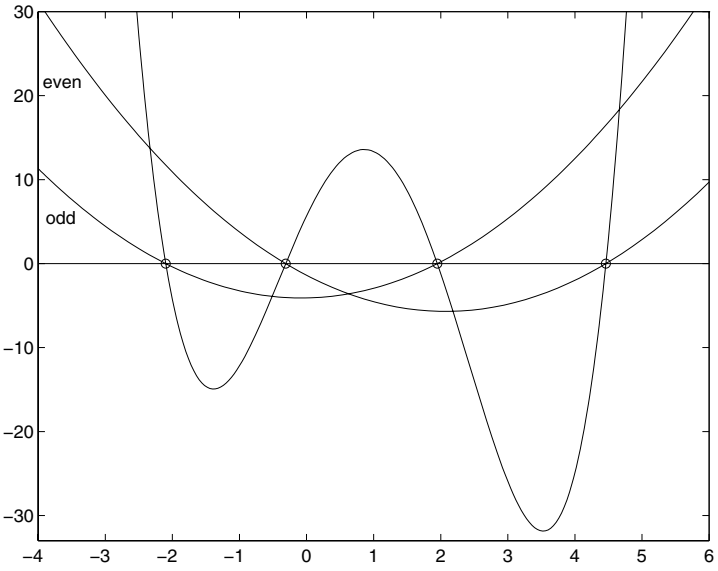


FIG. 1. *Even and odd characteristic polynomials.*

The first step of the method consists of solving a trivial 1×1 system, whereas in the final step, $y^{(n)}$ is computed from $y^{(n-1)}$, β_{n-1}, and α_{n-1}. The quantities α_k are called reflection coefficients, or Schur–Szegö parameters. Durbin's algorithm requires $2n^2 + \mathcal{O}(n)$ flops, which we define as in [11].

A more efficient method than Durbin's algorithm is what we will call the *split Durbin algorithm* from [9], where it is called the "split Levinson algorithm." We prefer this terminology because "Durbin" usually refers to the Yule–Walker equations, which have a special right-hand side, whereas "Levinson" usually refers to a system with an arbitrary right-hand side. In this we also follow [11].

To explain this algorithm, we define an *even* solution $u^{(k)}$ of the Yule–Walker equations $T_k y^{(k)} = -t_k$ as the solution of $T_k u^{(k)} = -(t_k + Jt_k)$, or $u^{(k)} = y^{(k)} + Jy^{(k)}$, and an *odd* solution as the solution of $T_k v^{(k)} = -(t_k - Jt_k)$, or $v^{(k)} = y^{(k)} - Jy^{(k)}$. The algorithm is based on the remarkable observation that the solution $y^{(k)}$ can be written either as a combination of the two successive even solutions $u^{(k)}$ and $u^{(k-1)}$ or as a combination of the two successive odd solutions $v^{(k)}$ and $v^{(k-1)}$. It is therefore sufficient to compute either the even or the odd solutions. For full details, we refer to [9], or [20] where it is summarized in the same notation as here. Let us just state the recursions for the even and odd solutions. For the even solutions, we have

$$u_1^{(k)} = u_k^{(k)} = u_1^{(k-1)} - \frac{\rho_0 + t_{k-1}^T u^{(k-1)} + \rho_k}{\rho_0 + t_{k-2}^T u^{(k-2)} + \rho_{k-1}} + 1,$$

$$u_j^{(k)} = u_j^{(k-1)} + u_{j-1}^{(k-1)} - \frac{\rho_0 + t_{k-1}^T u^{(k-1)} + \rho_k}{\rho_0 + t_{k-2}^T u^{(k-2)} + \rho_{k-1}} u_{j-1}^{(k-2)} \quad (2 \le j \le k-1).$$

For the odd solutions, one obtains

$$v_1^{(k)} = v_k^{(k)} = v_1^{(k-1)} - \frac{\rho_0 + t_{k-1}^T v^{(k-1)} - \rho_k}{\rho_0 + t_{k-2}^T v^{(k-2)} - \rho_{k-1}} + 1,$$

$$v_j^{(k)} = v_j^{(k-1)} + v_{j-1}^{(k-1)} - \frac{\rho_0 + t_{k-1}^T v^{(k-1)} - \rho_k}{\rho_0 + t_{k-2}^T v^{(k-2)} - \rho_{k-1}} v_{j-1}^{(k-2)} \quad (2 \le j \le k-1).$$

The split Durbin algorithm requires $\frac{3}{2}n^2 + \mathcal{O}(n)$ flops.

Finally, we mention that there also exist so-called superfast methods to solve the Yule–Walker equations (see, e.g., [1], [2]). However, for matrices with dimensions up to several hundred, they are less efficient than the algorithms mentioned here, which are usually referred to as "fast methods."

3. Recursions for the even and odd characteristic polynomials. In this section we will derive separate recursions for the even and odd characteristic polynomials. We start by mentioning the following recursion for the characteristic polynomials of an SPD Toeplitz matrix, which is stated in [16, Proposition 2.1]:

$$p_k(\lambda) = p_{k-1}(\lambda)\left(\rho_0 - \lambda - t^T \left(T_{k-1} - \lambda I\right)^{-1} t\right),$$

where $t = (\rho_1, \ldots, \rho_{k-1})^T$. Instead of this recursion, we derive separate recursions for the even and odd characteristic polynomials. But before we do, we will first introduce some definitions, notation, and a lemma.

We write the matrix K of dimensions $k \times k$ from Lemma 2.1 for even k as follows:

$$K = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0^T & -1 \\ a & \tilde{K} & a \\ 0 & b^T & 0 \end{pmatrix},$$

with $q = (1, 0, \ldots, 0)^T$, $a = \binom{0}{q}$, $b = \binom{Jq}{q}$, and $\tilde{K} = \left(\begin{smallmatrix} I & -J \\ S & SJ \end{smallmatrix}\right)$, where $S = \left(\begin{smallmatrix} 0^T & 0 \\ I & 0 \end{smallmatrix}\right)$. The vector q has dimension $(\frac{k}{2} - 1)$, and the vectors a and b have dimension $(k-2)$. In general, when S is an $n \times n$ matrix, then, in this matrix, I is the $(n-1) \times (n-1)$ identity matrix, 0^T is the $(n-1)$-dimensional null row-vector, the 0 immediately to the right of I is the $(n-1)$-dimensional null column vector, and the 0 in the upper right-hand corner is the scalar 0. We note that for a vector $w = (w_1, \ldots, w_n)^T$, $Sw = (0, w_1, \ldots, w_{n-1})^T$. When W is a matrix of dimensions $n \times n$, SWS^T results in a matrix with zero first row and column and with the lower right principal $(n-1)\times(n-1)$ submatrix equal to the upper left principal $(n-1) \times (n-1)$ submatrix of W. In this particular case, where K has (even) dimensions $k \times k$, S has dimensions $(\frac{k}{2}-1)\times(\frac{k}{2}-1)$, as do the other three submatrices of $\tilde{K}$: I, J, and SJ.

For $w \in \mathbb{R}^n$, we define its last $n-1$ components as the vector $w_* \in \mathbb{R}^{n-1}$, i.e., $w_* = (w_2, \ldots, w_n)^T$, and for even n, we also denote by $w_<$ and $w_>$ its first and last $n/2$ components, respectively.

We have the following lemma.

LEMMA 3.1. *For $w \in \mathbb{R}^{k-2}$ and k even,*

$$w = Jw \Rightarrow \tilde{K}w = \begin{pmatrix} 0 \\ 2Sw_< \end{pmatrix},$$

$$w = -Jw \Rightarrow \tilde{K}w = \begin{pmatrix} 2w_< \\ 0 \end{pmatrix}.$$

Proof.

$$\tilde{K}w = \begin{pmatrix} I & -J \\ S & SJ \end{pmatrix} \begin{pmatrix} w_< \\ w_> \end{pmatrix} = \begin{pmatrix} w_< - Jw_> \\ S(w_< + Jw_>) \end{pmatrix}.$$

The proof follows from the fact that $w = Jw$ implies $w_< = Jw_>$, whereas $w = -Jw$ implies $w_< = -Jw_>$. ∎

We now derive the aforementioned even and odd recursions in the following proposition.

PROPOSITION 3.2. *The even and odd characteristic polynomials of an SPD Toeplitz matrix T_k satisfy the following recursion relations for λ not an eigenvalue of its principal submatrix T_{k-2}:*

$$p_k^e(\lambda) = p_{k-2}^e(\lambda)\left(\rho_0 + \rho_{k-1} - \lambda - \frac{1}{2}(\tilde{t} + J\tilde{t})^T (T_{k-2} - \lambda I)^{-1} (\tilde{t} + J\tilde{t})\right), \tag{4}$$

$$p_k^o(\lambda) = p_{k-2}^o(\lambda)\left(\rho_0 - \rho_{k-1} - \lambda - \frac{1}{2}(\tilde{t} - J\tilde{t})^T (T_{k-2} - \lambda I)^{-1} (\tilde{t} - J\tilde{t})\right), \tag{5}$$

where $\tilde{t} = (\rho_1, \rho_2, \ldots, \rho_{k-2})^T$.

Proof. We will prove the proposition for matrices of even dimension. For odd-dimensional matrices, the proof is entirely analogous. Setting $T = T_k - \lambda I$ and $G = T_{k-2} - \lambda I$, we have

$$T = \begin{pmatrix} \rho_0 - \lambda & \tilde{t}^T & \rho_{k-1} \\ \tilde{t} & G & J\tilde{t} \\ \rho_{k-1} & \tilde{t}^T J & \rho_0 - \lambda \end{pmatrix} = \begin{pmatrix} 1 & \tilde{t}^T & 0 \\ 0 & G & 0 \\ 0 & \tilde{t}^T J & 1 \end{pmatrix} \begin{pmatrix} \alpha & 0^T & \beta \\ g & I & Jg \\ \beta & 0^T & \alpha \end{pmatrix}, \tag{6}$$

where $g = G^{-1}\tilde{t}$, $\alpha = \rho_0 - \lambda - \tilde{t}^T G^{-1}\tilde{t}$, and $\beta = \rho_{k-1} - \tilde{t}^T G^{-1} J\tilde{t}$.

We write the SPD Toeplitz matrices T and G as follows:

$$T = \begin{pmatrix} A & B \\ JBJ & A \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} \tilde{A} & \tilde{B} \\ J\tilde{B}J & \tilde{A} \end{pmatrix}.$$

The blocks in T and G are of size $\frac{k}{2} \times \frac{k}{2}$ and $(\frac{k}{2} - 1) \times (\frac{k}{2} - 1)$, respectively. From Lemma 2.1, we then have

$$\begin{aligned} \begin{pmatrix} A - BJ & 0 \\ 0 & A + BJ \end{pmatrix} &= KTK^T \\ &= \left(K \begin{pmatrix} 1 & \tilde{t}^T & 0 \\ 0 & G & 0 \\ 0 & \tilde{t}^T J & 1 \end{pmatrix} K^T\right)\left(K \begin{pmatrix} \alpha & 0^T & \beta \\ g & I & Jg \\ \beta & 0^T & \alpha \end{pmatrix} K^T\right). \end{aligned} \tag{7}$$

Recalling that

$$K = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0^T & -1 \\ a & \tilde{K} & a \\ 0 & b^T & 0 \end{pmatrix},$$

the first factor in the right-hand side of (7) gives

$$\begin{aligned} &K \begin{pmatrix} 1 & \tilde{t}^T & 0 \\ 0 & G & 0 \\ 0 & \tilde{t}^T J & 1 \end{pmatrix} K^T \\ &= \frac{1}{2} \begin{pmatrix} 2 & (\tilde{t} - J\tilde{t})^T \tilde{K}^T & (\tilde{t} - J\tilde{t})^T b \\ 0 & 2aa^T + a(\tilde{t} + J\tilde{t})^T \tilde{K}^T + \tilde{K} G \tilde{K}^T & a(\tilde{t} + J\tilde{t})^T b + \tilde{K} G b \\ 0 & b^T G \tilde{K}^T & b^T G b \end{pmatrix}. \end{aligned} \tag{8}$$

The second factor in the right-hand side of (7) gives

$$K \begin{pmatrix} \alpha & 0^T & \beta \\ g & I & Jg \\ \beta & 0^T & \alpha \end{pmatrix} K^T$$

$$= \frac{1}{2} \begin{pmatrix} 2(\alpha-\beta) & 0^T & 0 \\ \tilde{K}(g-Jg) & 2(\alpha+\beta)aa^T + \tilde{K}(g+Jg)a^T + \tilde{K}\tilde{K}^T & \tilde{K}b \\ b^T(g-Jg) & b^T((g+Jg)a^T + \tilde{K}^T) & b^Tb \end{pmatrix}. \tag{9}$$

We now show that the matrices in (8) and (9) are block-diagonal. To this end, we make the following nine observations:

(i) $(\tilde{t}-J\tilde{t})^T b = (g-Jg)^T b = 0$ because these are scalar products of an even and an odd vector.

(ii) As a direct consequence of Lemma 3.1, we have that $\tilde{K}(\tilde{t}+J\tilde{t}) = \begin{pmatrix} 0 \\ 2S(\tilde{t}+J\tilde{t})_< \end{pmatrix}$, $\tilde{K}(\tilde{t}-J\tilde{t}) = \begin{pmatrix} 2(\tilde{t}-J\tilde{t})_< \\ 0 \end{pmatrix}$, and $\tilde{K}(g-Jg) = \begin{pmatrix} 2(g-Jg)_< \\ 0 \end{pmatrix}$. In addition, $\tilde{K}b = \begin{pmatrix} 0 \\ 2SJq \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

(iii) $\tilde{K}G\tilde{K}^T = \begin{pmatrix} 2(\tilde{A}-\tilde{B}J) & 0 \\ 0 & 2S(\tilde{A}+\tilde{B}J)S^T \end{pmatrix}$.

(iv) $Gb = \begin{pmatrix} \tilde{A} & \tilde{B} \\ J\tilde{B}J & \tilde{A} \end{pmatrix} \begin{pmatrix} Jq \\ q \end{pmatrix} = \begin{pmatrix} (\tilde{A}+\tilde{B}J)Jq \\ (\tilde{A}+J\tilde{B})q \end{pmatrix}$. We therefore have $b^TGb = (q^TJ \quad q^T) \begin{pmatrix} (\tilde{A}+\tilde{B}J)Jq \\ (\tilde{A}+J\tilde{B})q \end{pmatrix} = 2(\rho_0 - \lambda + \rho_1)$.

(v) $\tilde{K}Gb = \begin{pmatrix} 0 \\ 2S(\tilde{A}+\tilde{B}J)Jq \end{pmatrix}$.

(vi) $\tilde{K}\tilde{K}^T = \begin{pmatrix} 2I & 0 \\ 0 & 2SS^T \end{pmatrix}$.

(vii) $aa^T = \begin{pmatrix} 0 & 0 \\ 0 & qq^T \end{pmatrix}$.

(viii) $\tilde{K}(g+Jg)a^T = \begin{pmatrix} 0 \\ 2S(g+Jg)_< \end{pmatrix} (0 \quad q^T) = \begin{pmatrix} 0 & 0 \\ 0 & 2(S(g+Jg)_<)q^T \end{pmatrix}$.

(ix) $a(\tilde{t}+J\tilde{t})^T b = 2((\tilde{t}+J\tilde{t})_<^T Jq)a = \begin{pmatrix} 0 \\ \rho_{\frac{k}{2}-1} + \rho_{\frac{k}{2}} \\ 0 \end{pmatrix}$, where the upper zero vector is of dimension $\frac{k}{2}-1$, whereas the lower zero vector has dimension $\frac{k}{2}-2$. We note that $\rho_{\frac{k}{2}-1} + \rho_{\frac{k}{2}}$ is the $(\frac{k}{2}-1)$th component of $(\tilde{t}+J\tilde{t})_<$.

Using all of this, we obtain for (8)

$$K \begin{pmatrix} 1 & \tilde{t}^T & 0 \\ 0 & G & 0 \\ 0 & \tilde{t}^T J & 1 \end{pmatrix} K^T = \begin{pmatrix} 1 & (\tilde{t}-J\tilde{t})_<^T & 0 & 0^T \\ 0 & (\tilde{A}-\tilde{B}J) & 0 & 0 \\ 0 & 0^T & 1 & (\tilde{t}+J\tilde{t})_<^T \\ 0 & 0 & 0 & (\tilde{A}+\tilde{B}J) \end{pmatrix}, \tag{10}$$

where the first and third column in the right-hand side are k-dimensional vectors. More specifically, the first row in the right-hand side of (10) follows from observations (i) and (ii) and the last row is explained by (iv) and (v). The last column follows

from (iv), (v), and (ix). The $(2,2)$ block is an immediate consequence of (iii) and the lower right four blocks are explained by (iii) and (iv), by the first equation in (ii), and by observation (vii).

For (9), we obtain

$$K \begin{pmatrix} \alpha & 0^T & \beta \\ g & I & Jg \\ \beta & 0^T & \alpha \end{pmatrix} K^T = \begin{pmatrix} \alpha - \beta & 0^T & 0^T & 0 \\ (g - Jg)_< & I & 0 & 0 \\ 0 & 0 & H & 0 \\ 0 & 0^T & h^T & 1 \end{pmatrix}, \tag{11}$$

where the first and last columns in the right-hand side are k-dimensional vectors and where

$$H = \begin{pmatrix} \alpha + \beta & 0 \\ (S(g + Jg)_<)_* & I \end{pmatrix} \qquad \text{and} \qquad h = \begin{pmatrix} g_{\frac{k}{2}-1} + g_{\frac{k}{2}} \\ 0 \end{pmatrix}.$$

We note that the identity matrix I in H is a $(\frac{k}{2} - 2) \times (\frac{k}{2} - 2)$ matrix and that the zero in h is a zero vector of dimension $\frac{k}{2} - 2$. The first column in the right-hand side of (11) follows from observation (i) and the third equation in observation (ii). The last column follows from the last equation in (ii) and because $b^T b = 2$. The $(2,2)$ block is a consequence of (vi), (vii), and (ix). The blocks in H follow from (vi), (vii), and (viii), and the components of the vector h follow from the last equation in (ii) and from the definitions of a and b.

From (7), (10), and (11), we conclude that

$$\begin{pmatrix} A - BJ & 0 \\ 0 & A + BJ \end{pmatrix}$$
$$= \begin{pmatrix} 1 & (\tilde{t} - J\tilde{t})_<^T & 0 & 0^T \\ 0 & \tilde{A} - \tilde{B}J & 0 & 0 \\ 0 & 0^T & 1 & (\tilde{t} + J\tilde{t})_<^T \\ 0 & 0 & 0 & \tilde{A} + \tilde{B}J \end{pmatrix} \begin{pmatrix} \alpha - \beta & 0^T & 0^T & 0 \\ (g - Jg)_< & I & 0 & 0 \\ 0 & 0 & H & 0 \\ 0 & 0^T & h^T & 1 \end{pmatrix}.$$

Therefore, for the even part, we have

$$A + BJ = \begin{pmatrix} 1 & (\tilde{t} + J\tilde{t})_<^T \\ 0 & \tilde{A} + \tilde{B}J \end{pmatrix} \begin{pmatrix} H & 0 \\ h^T & 1 \end{pmatrix},$$

which means that the even characteristic polynomial satisfies

$$p_k^e(\lambda) = \det(A + BJ) = \det(\tilde{A} + \tilde{B}J)\det(H) = p_{k-2}^e(\lambda)(\alpha + \beta).$$

Recalling the definitions of α, β, and G, we obtain

$$p_k^e(\lambda) = p_{k-2}^e(\lambda) \left(\rho_0 + \rho_{k-1} - \lambda - \frac{1}{2}(\tilde{t} + J\tilde{t})^T \left(T_{k-2} - \lambda I\right)^{-1} (\tilde{t} + J\tilde{t}) \right),$$

where we have used the fact that

$$\tilde{t}^T \left(T_{k-2} - \lambda I\right)^{-1} (\tilde{t} + J\tilde{t}) = \frac{1}{2}(\tilde{t} + J\tilde{t})^T \left(T_{k-2} - \lambda I\right)^{-1} (\tilde{t} + J\tilde{t}).$$

Analogously, we find for the odd characteristic polynomial

$$p_k^o(\lambda) = p_{k-2}^o(\lambda) \left(\rho_0 - \rho_{k-1} - \lambda - \frac{1}{2}(\tilde{t} - J\tilde{t})^T \left(T_{k-2} - \lambda I\right)^{-1} (\tilde{t} - J\tilde{t}) \right). \qquad \square$$

The following proposition shows that the even and odd characteristic polynomials have the same basic properties as the characteristic polynomial itself, as summarized in [16, Proposition 2.1].

PROPOSITION 3.3. *The even and odd characteristic polynomials* $p_k^e(\lambda)$ *and* $p_k^o(\lambda)$ *of any principal submatrix* T_k *of the SPD Toeplitz matrix* T_n *satisfy the following:*

$$\forall \lambda < \lambda_{min}^e(T_n) : p_k^e(\lambda) > 0, \quad (p_k^e(\lambda))' < 0, \quad (p_k^e(\lambda))'' > 0, \tag{12}$$

$$\forall \lambda < \lambda_{min}^o(T_n) : p_k^o(\lambda) > 0, \quad (p_k^o(\lambda))' < 0, \quad (p_k^o(\lambda))'' > 0. \tag{13}$$

Proof. We have that $p_k^e(\lambda) = \Pi_{j=1}^{\lceil \frac{k}{2} \rceil}(\lambda_j^e(T_k) - \lambda)$ and $p_k^o(\lambda) = \Pi_{j=1}^{\lfloor \frac{k}{2} \rfloor}(\lambda_j^o(T_k) - \lambda)$. Since the interlacing properties of the even and odd eigenvalues imply that, for all $1 \le k \le n$,

$$\lambda_{min}^e(T_n) \le \lambda_{min}^e(T_k) \quad \text{and} \quad \lambda_{min}^o(T_n) \le \lambda_{min}^o(T_k),$$

the proof follows immediately. ∎

4. Bounds on the even and odd eigenvalues. As was mentioned in the introduction, several methods for computing the smallest eigenvalue of an $n \times n$ SPD Toeplitz matrix T_n are based on a so-called secular equation. Apparently, the first such method was proposed in [7]. Assuming that $\lambda_{min}(T_n) \neq \lambda_{min}(T_{n-1})$, the secular equation there takes the form

$$h(\lambda) \stackrel{\triangle}{=} -\rho_0 + \lambda + t^T(T_{n-1} - \lambda I)^{-1}t = 0,$$

where $t = (\rho_1, \rho_2, \ldots, \rho_{n-1})^T$, as before. The smallest root of $h(\lambda)$ is the smallest eigenvalue of T_n and its singularities are the eigenvalues of T_{n-1}. Later, in [17], [19], and [22], this equation was replaced by two similar equations: one for the even and one for the odd eigenvalues. Assuming that $\lambda_{min}^e(T_n) \neq \lambda_{min}^e(T_{n-2})$ and $\lambda_{min}^o(T_n) \neq \lambda_{min}^o(T_{n-2})$, they are given, respectively, by

$$h^e(\lambda) \stackrel{\triangle}{=} -\rho_0 - \rho_{k-1} + \lambda + \frac{1}{2}(\tilde{t} + J\tilde{t})^T (T_{n-2} - \lambda I)^{-1} (\tilde{t} + J\tilde{t}) = 0,$$

$$h^o(\lambda) \stackrel{\triangle}{=} -\rho_0 + \rho_{k-1} + \lambda + \frac{1}{2}(\tilde{t} - J\tilde{t})^T (T_{n-2} - \lambda I)^{-1} (\tilde{t} - J\tilde{t}) = 0,$$

where, as before, $\tilde{t} = (\rho_1, \rho_2, \ldots, \rho_{n-2})^T$. The smallest roots of $h^e(\lambda)$ and $h^o(\lambda)$ are the smallest even and odd eigenvalues of T_n, respectively, and their singularities are the even and odd eigenvalues of T_{n-2}, respectively.

Those equations can be used to derive upper bounds on the smallest even and odd eigenvalues, which we will later use to predict the parity of the smallest eigenvalue. Since these bounds are not new, we will only give the basic results and refer to the proper references for the details. The functions $h(\lambda)$, $h^e(\lambda)$, and $h^o(\lambda)$ are all of the same basic form and can all be represented, for appropriate values of its parameters, by the function $f(\lambda)$, given by

$$f(\lambda) \stackrel{\triangle}{=} -\rho + \lambda + v^T(T - \lambda I)^{-1}v.$$

We denote the smallest singularity of $f(\lambda)$ by α_0. If the values of f and f' are known at a point $x_0 < \alpha_0$, then, following [18], we can compute an upper bound on the smallest root of f by approximating $f(\lambda)$ by a function $\psi(\lambda)$ of the form

$$\psi(\lambda) = -\rho + \lambda + \frac{a}{b - \lambda}\,.$$

The parameters a and b are determined by the approximation conditions $\psi(x_0) = f(x_0)$ and $\psi'(x_0) = f'(x_0)$. It was shown in [18] (point (1) on p. 368 and Theorem 4.1 on p. 371), that $\psi(\lambda) \leq f(\lambda)$ on the interval $(-\infty, \alpha_0)$ and that $b > \alpha_0$. In addition, the function $f(\lambda)$ increases monotonically from $-\infty$ to $+\infty$ on this interval and $\psi(\lambda)$ does the same on $(-\infty, b)$. The root of $\psi(\lambda)$ on $(-\infty, b)$ must therefore be an upper bound on the root of $f(\lambda)$ in $(-\infty, \alpha_0)$. We note that $x_0 = 0$ in [18], but this is not essential. The computation of the upper bound can therefore be achieved by solving a simple quadratic equation.

If the values of f and f' are known at two distinct points $x_0, x_1 < \alpha_0$, then we can construct a more accurate approximation of f. In this case, we write $f(\lambda)$ as in [14]:

$$f(\lambda) = f(x_0) + f'(x_0)(\lambda - x_0) + (\lambda - x_0)^2 g(\lambda),$$

where $g(\lambda)$ is a rational function of the same general form and with the same singularities as $f(\lambda)$. We now define the following approximation $\phi(\lambda)$ of $f(\lambda)$:

$$\phi(\lambda) = f(x_0) + f'(x_0)(\lambda - x_0) + (\lambda - x_0)^2 \left(\frac{a}{b - \lambda}\right),$$

where a and b are such that $\frac{a}{b-\lambda}$ is a first order approximation to $g(\lambda)$ at x_1, i.e.,

$$g(x_1) = \frac{a}{b - x_1} \quad \text{and} \quad g'(x_1) = \frac{a}{(b - x_1)^2}.$$

We note that $\phi(\lambda)$ is a first order approximation to $f(\lambda)$ at both x_0 and x_1. In [19, Theorem 4.1 on p. 657] it was shown that $\phi(\lambda)$ exhibits the same properties as $\psi(\lambda)$. The root of $\phi(\lambda)$ on $(-\infty, b)$ must therefore be an upper bound on the root of $f(\lambda)$ in $(-\infty, \alpha_0)$. It can be computed, once again, by solving a simple quadratic equation.

If we apply these procedures to $h^e(\lambda)$ and $h^o(\lambda)$ for given points x_0 and x_1, then the approximations thus obtained allow us to compute upper bounds on the smallest even and odd eigenvalues by computing their smallest roots.

5. The algorithm. The smallest eigenvalue $\lambda_{min}(T_n)$ of T_n is, of course, given by the smallest root of $p_n(\lambda)$. In [16], this root is computed with Newton's method, starting from a point to the left of $\lambda_{min}(T_n)$. Property 2.1 in that paper ensures that the method will converge monotonically. The Newton step $\frac{p_n(\lambda)}{p_n'(\lambda)}$ is computed, using the recursion in [16, p. 1924], and the various quantities that need to be computed are obtained from Durbin's algorithm. Because Durbin's algorithm is used, each Newton step requires $3n^2 + \mathcal{O}(n)$ operations, namely, $2n^2 + \mathcal{O}(n)$ operations for Durbin's algorithm, and $n^2 + \mathcal{O}(n)$ because of the extra computation of the scalar product $\|y^{(k)}\|^2$ for each Yule–Walker subsystem.

The algorithm we now propose normally has two phases. In the first phase, Newton's method is simultaneously applied to the computation of the smallest roots of both $p_n^e(\lambda)$ and $p_n^o(\lambda)$. During this phase, each step generates two iterates, one for the even and one for the odd characteristic polynomial. The algorithm proceeds with the smallest of those two iterates while at the same time generating upper bounds on the even and odd eigenvalues. The first phase ends if and when these bounds have become sufficiently accurate to determine the parity of the smallest eigenvalue of T_n. At this point, the second phase begins: starting from the last iterate, Newton's method is applied to the computation of the smallest root of either $p_n^e(\lambda)$ or $p_n^o(\lambda)$, depending

on whether in the first phase the smallest eigenvalue was found to be even or odd, respectively. If the bounds in the first phase fail to determine the correct parity, then the algorithm will simply run its course without ever making the transition to the second phase.

As in the MB method, we first derive a recursion to compute the Newton step. We will have two such recursions: one for the even and one for the odd characteristic polynomial. They are given in the following proposition.

PROPOSITION 5.1. *With* $\tilde{t} = (\rho_1, \rho_2, \ldots, \rho_{k-2})^T$, $\lambda < \lambda_{min}(T_n)$, *and* $u^{(k)}$ *and* $v^{(k)}$ *denoting the even and odd solutions of the* k*-dimensional Yule–Walker subsystem, respectively, one has the following recursions for the Newton steps for the even and odd characteristic polynomials:*

$$\frac{p_k^e(\lambda)}{(p_k^e(\lambda))'} = \frac{\rho_0 + \rho_{k-1} - \lambda + \tilde{t}^T u^{(k-2)}}{\left(\frac{p_{k-2}^e(\lambda)}{(p_{k-2}^e(\lambda))'}\right)^{-1}\left(\rho_0 + \rho_{k-1} - \lambda + \tilde{t}^T u^{(k-2)}\right) - \left(1 + \frac{1}{2}\|u^{(k-2)}\|^2\right)}, \tag{14}$$

$$\frac{p_k^o(\lambda)}{(p_k^o(\lambda))'} = \frac{\rho_0 - \rho_{k-1} - \lambda + \tilde{t}^T v^{(k-2)}}{\left(\frac{p_{k-2}^o(\lambda)}{(p_{k-2}^o(\lambda))'}\right)^{-1}\left(\rho_0 - \rho_{k-1} - \lambda + \tilde{t}^T v^{(k-2)}\right) - \left(1 + \frac{1}{2}\|v^{(k-2)}\|^2\right)}. \tag{15}$$

Proof. From Proposition 3.1 we have

$$(p_k^e(\lambda))' = (p_{k-2}^e(\lambda))'\left(\rho_0 + \rho_{k-1} - \lambda - \frac{1}{2}(\tilde{t} + J\tilde{t})^T (T_{k-2} - \lambda I)^{-1}(\tilde{t} + J\tilde{t})\right) - p_{k-2}^e(\lambda)\left(1 + \frac{1}{2}(\tilde{t} + J\tilde{t})^T (T_{k-2} - \lambda I)^{-2}(\tilde{t} + J\tilde{t})\right).$$

In our notation, this can be written as

$$(p_k^e(\lambda))' = (p_{k-2}^e(\lambda))'\left(\rho_0 + \rho_{k-1} - \lambda + \frac{1}{2}(\tilde{t} + J\tilde{t})^T u^{(k-2)}\right) - p_{k-2}^e(\lambda)\left(1 + \frac{1}{2}\|u^{(k-2)}\|^2\right).$$

Because $\frac{1}{2}(\tilde{t} + J\tilde{t})^T u^{(k-2)} = \tilde{t}^T u^{(k-2)}$, we have

$$\frac{p_k^e(\lambda)}{(p_k^e(\lambda))'} = \frac{p_{k-2}^e(\lambda)\left(\rho_0 + \rho_{k-1} - \lambda + \tilde{t}^T u^{(k-2)}\right)}{(p_{k-2}^e(\lambda))'\left(\rho_0 + \rho_{k-1} - \lambda + \tilde{t}^T u^{(k-2)}\right) - p_{k-2}^e(\lambda)\left(1 + \frac{1}{2}\|u^{(k-2)}\|^2\right)},$$

which leads to

$$\frac{p_k^e(\lambda)}{(p_k^e(\lambda))'} = \frac{\rho_0 + \rho_{k-1} - \lambda + \tilde{t}^T u^{(k-2)}}{\frac{(p_{k-2}^e(\lambda))'}{p_{k-2}^e(\lambda)}\left(\rho_0 + \rho_{k-1} - \lambda + \tilde{t}^T u^{(k-2)}\right) - \left(1 + \frac{1}{2}\|u^{(k-2)}\|^2\right)},$$

which is the same as (14). The recursion for the odd characteristic polynomial is obtained analogously. ∎

In the first phase, once again as in the MB algorithm, we use Durbin's algorithm to compute the various required quantities. However, instead of computing one extra scalar product per step as in the MB method, our method computes two scalar products of half the length for *every other* Yule–Walker subsystem, namely, $\|u^{(k)}\|^2$ and $\|v^{(k)}\|^2$. Taking into account that $u^{(k)}$ and $v^{(k)}$ need to be obtained from $y^{(k)}$, this means that each step of our algorithm costs $\frac{11}{4}n^2 + \mathcal{O}(n)$ operations, namely, $2n^2 + \mathcal{O}(n)$ operations for Durbin's algorithm, and $\frac{3}{4}n^2 + \mathcal{O}(n)$ because of the

extra scalar products. This is a savings of $\frac{1}{4}n^2 + \mathcal{O}(n)$ operations per step when compared to the MB method. The computation of the upper bounds on the eigenvalues during this phase does not entail any significant additional computational cost because all the required ingredients have already been computed.

Once the parity of the smallest eigenvalue has been determined, the second phase proceeds with the appropriate version (even or odd) of the split Durbin algorithm. During this phase, the number of computations per step is only $\frac{7}{4}n^2+\mathcal{O}(n)$ operations, namely, $\frac{3}{2}n^2+\mathcal{O}(n)$ operations for the split Durbin algorithm and $\frac{1}{4}n^2+\mathcal{O}(n)$ for the extra scalar product. This represents significant additional savings.

Finally, we note that our method can be improved in exactly the same way that the MB method was improved in [15], namely, by modifying Newton's method. We also mention here once again that superfast methods could be used instead of the fast algorithms. This does not affect the general structure of our algorithm.

We took $\lambda = 0$ as the starting point for our algorithm and for the stopping criterion (analogously to the MB method) we used the ratios

$$\beta^e = \frac{p_n^e(\bar{\lambda})}{p_{n-2}^e(\bar{\lambda})} = (\lambda_{min}^e(T_n) - \bar{\lambda}) \prod_{j=1}^{\lceil \frac{n}{2} \rceil - 1} \frac{\lambda_j^e(T_n) - \bar{\lambda}}{\lambda_j^e(T_{n-2}) - \bar{\lambda}},$$

$$\beta^o = \frac{p_n^o(\bar{\lambda})}{p_{n-2}^o(\bar{\lambda})} = (\lambda_{min}^o(T_n) - \bar{\lambda}) \prod_{j=1}^{\lfloor \frac{n}{2} \rfloor - 1} \frac{\lambda_j^o(T_n) - \bar{\lambda}}{\lambda_j^o(T_{n-2}) - \bar{\lambda}},$$

where $\bar{\lambda}$ is an approximation to the smallest eigenvalue of T_n. Since in our algorithm $\lambda_{min}^e(T_n) - \bar{\lambda} > 0$ and $\lambda_{min}^o(T_n) - \bar{\lambda} > 0$, and because of the interlacing properties of the even and odd eigenvalues, all the factors in the right-hand sides are larger than one. This means that $\beta^e > \lambda_{min}^e(T_n) - \bar{\lambda}$ and $\beta^o > \lambda_{min}^o(T_n) - \bar{\lambda}$. As a stopping criterion, we therefore used $\beta^e < \epsilon$ and $\beta^o < \epsilon$ for the even and odd smallest eigenvalues, respectively, where ϵ is a given tolerance. This then guarantees $\lambda_{min}^e(T_n) - \bar{\lambda} < \epsilon$ and $\lambda_{min}^o(T_n) - \bar{\lambda} < \epsilon$, respectively.

When one looks at Figure 1, one can intuitively understand why the split into an even and odd characteristic polynomial should increase the Newton step and therefore decrease the number of iterations. Since either the even or odd characteristic polynomial has to go through fewer points than the characteristic polynomial itself, they should oscillate less "wildly" at the endpoints and therefore climb less steeply, which in turns increases the Newton step. This is confirmed by the following proposition.

PROPOSITION 5.2. *The Newton step* $|\frac{p_n(\lambda)}{p_n'(\lambda)}|$ *for the characteristic polynomial of* T_n*, when* $\lambda < \lambda_{min}(T_n)$*, is always less than the corresponding Newton steps for either the even or the odd characteristic polynomial.*

Proof.

$$\frac{p_n'(\lambda)}{p_n(\lambda)} = \frac{(p_n^e(\lambda))' p_n^o(\lambda) + p_n^e(\lambda)(p_n^o(\lambda))'}{p_n^e(\lambda) p_n^o(\lambda)} = \frac{(p_n^e(\lambda))'}{p_n^e(\lambda)} + \frac{(p_n^o(\lambda))'}{p_n^o(\lambda)} \cdot$$

However, for $\lambda < \lambda_{min}(T_n)$, we know from [16, Propositions 3.2 and 2.1] that

$$\frac{p_n'(\lambda)}{p_n(\lambda)}, \quad \frac{(p_n^e(\lambda))'}{p_n^e(\lambda)}, \quad \text{and} \quad \frac{(p_n^o(\lambda))'}{p_n^o(\lambda)}$$

are all negative. One therefore obtains that

$$\left|\frac{p_n'(\lambda)}{p_n(\lambda)}\right| > \left|\frac{(p_n^e(\lambda))'}{p_n^e(\lambda)}\right| \quad \text{and} \quad \left|\frac{p_n'(\lambda)}{p_n(\lambda)}\right| > \left|\frac{(p_n^o(\lambda))'}{p_n^o(\lambda)}\right| \cdot$$

Taking the reciprocal of each side in these inequalities completes the proof. ∎

We now briefly discuss some of the factors that affect the numerical stability of our algorithm. The Newton step is determined by the recursion relations in Proposition 5.1 and its accuracy is therefore influenced by the accuracy with which $u^{(k-2)}$ and $v^{(k-2)}$ can be calculated. This, in turn, depends on the method that is used to compute these quantities. We have used two different methods to do that: the Durbin and the split Durbin algorithms. Both types of algorithms are weakly stable for SPD matrices (see [4], [5], and [13]), but, as was observed from numerical experiments in [20], Levinson's algorithm is sometimes significantly more accurate than its split counterparts and the same seems therefore plausible for Durbin's algorithm. In turn, one can also expect Durbin's algorithm to be less accurate than the much more expensive Cholesky factorization.

On the other hand, as was shown by the error analysis in section 5 of [18], the smaller the gap between the smallest even or odd eigenvalues of T_n and T_{n-2}, the more difficult the accurate computation of $u^{(n-2)}$ or $v^{(n-2)}$ will be. Likewise, the smaller the gap between the smallest eigenvalues of T_n and T_{n-1}, the more difficult it will be to accurately compute $y^{(n-1)}$. Because of the interlacing properties of the eigenvalues, the gap between the smallest even or odd eigenvalues of T_n and T_{n-2} is at least as large as the gap between the smallest eigenvalues of T_n and T_{n-1}, which gives our method an advantage over the MB method. Moreover, because they proceed in steps of two, the number of recursions in our method is half that of the MB method.

However, it would require more study, which would be beyond the scope of the present paper, to understand how these various factors affect the relative accuracy of the different methods.

To conclude this section, we summarize our algorithm below. It computes the smallest eigenvalue of an SPD Toeplitz matrix, defined by the vector $(\rho_0, \ldots, \rho_{n-1})^T$, which we assume is available to all functions in the algorithm. The notation we use is fairly self-explanatory: e.g., $h_1^e = h^e(x_1)$, and $dh_1^e = (h^e)'(x_1)$, and analogously for similar quantities. Even and odd Newton steps are denoted by N^e and N^o, respectively. The number of iterations in the first and second phase is given by k_1 and k_2, respectively, and ϵ is a tolerance, which is assumed to be given. We define the following functions.

DURBIN: a function with one argument, representing a point for which this function uses Durbin's algorithm and the recurrence relations in Proposition 5.1 to return the even and odd Newton steps, the function value and the derivative of the even and odd secular function, and also β^e and β^o, which were defined just before Proposition 5.2.

EVENSPLITDURBIN: a function with one argument, representing a point for which this function uses the even split Durbin algorithm and the even recurrence relation in Proposition 5.1 to return the even Newton step and β^e.

ODDSPLITDURBIN: a function with one argument. It is the odd equivalent of the EVENSPLITDURBIN function.

INITIALEVENBOUND: a function with three arguments, namely, a point and the function value and derivative of the even secular function at that point. It returns an even upper bound. This is the upper bound, based on one point, described in section 4.

INITIALODDBOUND: a function with three arguments. It is the odd equivalent of the INITIALEVENBOUND function.

EVENBOUND: a function with four arguments, namely, two points and the function

value and derivative of the even secular function at those two points. It returns an even upper bound. This is the upper bound, based on two points, described in section 4.

ODDBOUND: a function with four arguments. It is the odd equivalent of the EVENBOUND function.

Algorithm 5.1.

$x_0 = 0$
$k_1 = 1$; $k_2 = 0$
$(N_0^e, N_0^o, h_0^e, dh_0^e, h_0^o, dh_0^o, \beta^e, \beta^o) = \text{DURBIN}(x_0)$
$B^e = \text{INITIALEVENBOUND}(x_0, h_0^e, dh_0^e)$
$B^o = \text{INITIALODDBOUND}(x_0, h_0^o, dh_0^o)$
$x_1^e = x_0 - N_0^e$
$x_1^o = x_0 - N_0^o$
$\beta = \min(\beta^e, \beta^o)$
WHILE $(x_1^e \le B^o$ & $x_1^o \le B^e$ & $\beta > \epsilon)$
  $x_1 = \min(x_1^e, x_1^o)$
  $k_1 = k_1 + 1$
  $(N_1^e, N_1^o, h_1^e, dh_1^e, h_1^o, dh_1^o, \beta^e, \beta^o) = \text{DURBIN}(x_1)$
  $B^e = \text{EVENBOUND}(x_0, h_0^e, dh_0^e, x_1, h_1^e, dh_1^e)$
  $B^o = \text{ODDBOUND}(x_0, h_0^o, dh_0^o, x_1, h_1^o, dh_1^o)$
  $x_1^e = x_1 - N_1^e$
  $x_1^o = x_1 - N_1^o$
  $\beta = \min(\beta^e, \beta^o)$
  $x_0 = x_1$; $h_0^e = h_1^e$; $dh_0^e = dh_1^e$; $h_0^o = h_1^o$; $dh_0^o = dh_1^o$
END
IF $(x_1^e > B^o)$
  WHILE $(\beta^o > \epsilon)$
    $x^o = x_1^o$
    $k_2 = k_2 + 1$
    $(N^o, \beta^o) = \text{ODDSPLITDURBIN}(x^o)$
    $x^o = x^o - N^o$
  END
ELSE
  WHILE $(\beta^e > \epsilon)$
    $x^e = x_1^e$
    $k_2 = k_2 + 1$
    $(N^e, \beta^e) = \text{EVENSPLITDURBIN}(x^e)$
    $x^e = x^e - N^e$
  END
END

6. Numerical results. In this section we will compare our method to the one in [16] for 500 random matrices of the form $T = \mu \sum_{k=1}^{n} \xi_k T_{2\pi\theta_k}$, where n is the dimension of T, and θ_k, ξ_k are uniformly distributed random numbers in $(0, 1)$. The parameter μ is chosen such that $T_{kk} = 1$ for $k = 1, \ldots, n$, and $(T_\theta)_{ij} = \cos(\theta(i - j))$. These matrices are positive semidefinite and they were also used in [16], before which they were used in [7]. Even though they could theoretically be singular, we did not encounter such cases, nor did we encounter cases where the smallest eigenvalue of T_n was an eigenvalue of T_{n-1} or where the smallest even or odd eigenvalue of T_n was an eigenvalue of T_{n-2}. We ran our method exactly as in Algorithm 5.1, except that we did not use the initial bound, based on one point. In almost all cases, not surprisingly, this bound was not sufficient to determine the parity of the smallest eigenvalue. We used the recursion formulas exactly as they are stated in Proposition 5.1. In [16], it was mentioned that the particular form in which they are cast avoids cancellation problems, but in [15] it was claimed that this was not really necessary. We did not check this, since whatever conclusion we would reach would be valid for these particular matrices only, and different situations could arise when a different

TABLE 1
Comparison of algorithms using 500 *random matrices for each dimension.*

Dimension	MB		EO		EO (no pred)		EO (both)	
	Steps	Flops	Steps	Flops	Steps	Flops	Steps	Flops
16	9.12	7.55×10^3	$6.69\ (2.44 + 4.25)$	5.21×10^3	6.67	6.22×10^3	$14.95\ (7.66 + 7.28)$	9.88×10^3
32	10.41	3.33×10^4	$7.08\ (2.67 + 4.42)$	1.91×10^4	7.21	2.39×10^4	$15.55\ (7.76 + 7.79)$	3.55×10^4
64	11.49	1.44×10^5	$7.66\ (2.98 + 4.69)$	0.75×10^5	7.82	0.96×10^5	$16.33\ (8.13 + 8.20)$	1.34×10^5
128	11.70	5.81×10^5	$7.61\ (2.87 + 4.73)$	2.82×10^5	9.49	4.48×10^5	$16.11\ (7.95 + 8.16)$	4.96×10^5
256	12.83	2.54×10^6	$7.56\ (2.96 + 4.60)$	1.10×10^6	8.32	1.54×10^6	$15.63\ (7.88 + 7.75)$	1.86×10^6
512	14.77	1.16×10^7	$8.56\ (3.29 + 5.27)$	0.49×10^7	8.73	0.64×10^7	$16.38\ (7.99 + 8.38)$	0.77×10^7
1024	14.06	4.43×10^7	$7.51\ (3.33 + 4.17)$	1.74×10^7	9.39	2.72×10^7	$15.54\ (7.92 + 7.62)$	2.88×10^7

type of matrix is involved. We stress that our numerical results serve only to illustrate the potential of our method. Far more extensive numerical results would be needed to seriously compare all existing methods and their variations.

We summarize our results in Table 1, where we have reported the average number of Newton steps and the average total flop count for 500 randomly generated matrices of dimensions 16, 32, 64, 128, 512, and 1024. The abbreviation "EO" stands for "even-odd" and refers to our method. In the first column for this method, we have added in parentheses the average number of iterations in the first and second phases, i.e., $8\,(3+5)$ means a total of eight iterations, three in the first phase and five in the second phase. For comparison, we have also run phase one of our algorithm from beginning to end without using bounds to predict the parity of the smallest eigenvalue. These results appear under the heading "EO (no pred)." In the columns headed by "EO (both)," we have reported the combined average number of iterations and the total average number of flops needed to compute *both* the smallest even and the smallest odd eigenvalues. We have added the average number of even and odd iterations in parentheses, i.e., $10\,(6+4)$ means a total of ten iterations, six for the even and four for the odd eigenvalue. All experiments were run in double precision using Matlab and the flops were counted by Matlab's internal flop counter.

The tolerance was chosen to be $\epsilon = 10^{-14}$, the same as the tolerance in the numerical results of [16]. In general, it may make more sense to consider a relative rather than an absolute tolerance, but since we are using it only for comparison purposes, we decided to follow [16] in this matter.

As one can see from the results, there is a significant advantage to our method compared to the MB algorithm, both in the number of iterations and in the flop count, which becomes more pronounced as the size of the matrices increases: for matrices of dimension 1024, our method needs 2.5 times fewer flops than the MB algorithm. Because our algorithm needs fewer iterations, it produces a lower flop count than the MB method, regardless of whether fast or superfast methods are used for solving the Yule–Walker equations. Moreover, already for moderate matrix dimensions (> 64), our method computes two eigenvalues with less flops than it takes the MB method to compute just one. Here too, the difference becomes more significant as the size of the matrices increases. Finally we note that in our method, the number of iterations seems to be less sensitive to the size of the matrix than in the MB method. It is also clear that using bounds to predict the parity of the smallest eigenvalue lowers both the number of iterations and the flop count. For these particular matrices, we found that the parity can be predicted after slightly less than half the total number of iterations.

As a final remark we note that in [16] the MB method was already compared to the method in [7] and was found to be more efficient, at least for this class of matrices. The same therefore holds true for our method.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *The generalized Schur algorithm for the superfast solution of Toeplitz systems*, in Rational Approximations and Applications in Mathematics and Physics, Lecture Notes in Math. 1237, J. Gilewicz, M. Pindor, and W. Siemaszko, eds., Springer, 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] A. L. Andrew, *Eigenvectors of certain matrices*, Linear Algebra Appl., 7 (1973), pp. 151–162.

[4] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.
[5] J. R. Bunch, *Matrix properties of the Levinson and Schur algorithms*, J. Numer. Linear Algebra Appl., 1 (1992), pp. 183–198.
[6] A. Cantoni and P. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.
[7] G. Cybenko and C. Van Loan, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 123–131.
[8] P. Delsarte and Y. Genin, *Spectral properties of finite Toeplitz matrices*, in Mathematical Theory of Networks and Systems, Proc. MTNS-83 International Symposium, Beer-Sheva, Israel, 1983, Springer, London, 1984, pp. 194–213.
[9] P. Delsarte and Y. V. Genin, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., ASSP-34, 1986, pp. 470–478.
[10] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.
[11] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, London, 1996.
[12] D. Huang, *Symmetric solutions and eigenvalue problems of Toeplitz systems*, IEEE Trans. Signal Process., 40 (1992), pp. 3069–3074.
[13] H. Krishna and Y. Wang, *The split Levinson algorithm is weakly stable*, SIAM J. Numer. Anal., 30 (1993), pp. 1498–1508.
[14] W. Mackens and H. Voss, *The minimum eigenvalue of a symmetric positive-definite Toeplitz matrix and rational Hermitian interpolation*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 521–534.
[15] W. Mackens and H. Voss, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix by Newton-type methods*, SIAM J. Sci. Comput., 21 (2000), pp. 1650–1656.
[16] N. Mastronardi and D. Boley, *Computing the smallest eigenpair of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Comput., 20 (1999), pp. 1921–1927.
[17] A. Melman, *Spectral functions for real symmetric Toeplitz matrices*, J. Comput. Appl. Math., 98 (1998), pp. 233–243.
[18] A. Melman, *Bounds on the extreme eigenvalues of real symmetric Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 362–378.
[19] A. Melman, *Extreme eigenvalues of real symmetric Toeplitz matrices*, Math. Comp., 70 (2000), pp. 649–669.
[20] A. Melman, *A two-step even-odd split Levinson algorithm for Toeplitz systems*, Linear Algebra Appl., 338 (2001), pp. 219–237.
[21] V. F. Pisarenko, *The retrieval of harmonics from a covariance function*, Geophys. J. Royal Astron. Soc., 33 (1973), pp. 347–366.
[22] H. Voss, *Symmetric schemes for computing the minimum eigenvalue of a symmetric Toeplitz matrix*, Linear Algebra Appl., 287 (1999), pp. 359–371.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 964–985

JACOBI-LIKE ALGORITHMS FOR THE INDEFINITE GENERALIZED HERMITIAN EIGENVALUE PROBLEM*

CHRISTIAN MEHL†

Abstract. We discuss structure-preserving Jacobi-like algorithms for the solution of the indefinite generalized Hermitian eigenvalue problem. We discuss a method based on the solution of Hermitian 4×4 subproblems which generalizes the Jacobi-like method of Bunse-Gerstner and Faßbender for Hamiltonian matrices. Furthermore, we discuss structure-preserving Jacobi-like methods based on the solution of non-Hermitian 2×2 subproblems. For these methods a local convergence proof is given. Numerical test results for the comparison of the proposed methods are presented.

Key words. Jacobi-like method, Hermitian pencil, eigenvalues

AMS subject classification. 65F15

DOI. 10.1137/S089547980240947X

1. Introduction. The generalized Hermitian eigenvalue problem arises in many applications. One important example is the linear quadratic optimal control problem (see [19, 20, 25] and the references therein). This is the problem of minimizing the cost functional

$$\frac{1}{2}\int_{t_0}^{\infty} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^* \mathcal{M} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt, \qquad \mathcal{M} = \begin{bmatrix} Q & S \\ S^* & R \end{bmatrix}, \tag{1.1}$$

subject to the dynamics

$$E\dot{x}(t) = Ax(t) + Bu(t), \quad t_0 < t, \qquad x(t_0) = x_0, \tag{1.2}$$

where $A, E, Q \in \mathbb{C}^{n\times n}$, $B, S \in \mathbb{C}^{n\times m}$, $R \in \mathbb{C}^{m\times m}$, Q, R Hermitian, $x_0, x(t), u(t) \in \mathbb{C}^n$, and $t_0, t \in \mathbb{R}$. (In many applications, there are additional restrictions such as $\mathcal{M}$ in (1.1) being positive semidefinite.) It is known that solutions of (1.1)–(1.2) can be obtained via the solution of a boundary value problem (see [24, 25] and the references therein). For the solution of this boundary value problem one has to compute deflating subspaces of the matrix pencil

$$\lambda\mathcal{A} - \mathcal{B} = \lambda \begin{bmatrix} 0 & -E^* & 0 \\ E & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} Q & A^* & S \\ A & 0 & B \\ S^* & B^* & R \end{bmatrix}.$$

Setting $\mathcal{G}_0 = i\mathcal{A}$ and $\mathcal{H}_0 = \mathcal{B}$, we find that $\lambda\mathcal{G}_0 - \mathcal{H}_0$ is a Hermitian pencil; i.e., both $\mathcal{G}_0$ and $\mathcal{H}_0$ are Hermitian. Clearly, both pencils $\lambda\mathcal{A} - \mathcal{B}$ and $\lambda\mathcal{G}_0 - \mathcal{H}_0$ have the same right deflating subspaces, and the eigenvalues of $\lambda\mathcal{G}_0 - \mathcal{H}_0$ coincide with the eigenvalues of $\lambda\mathcal{A} - \mathcal{B}$ multiplied by i.

Another important application is the Hermitian quadratic eigenvalue problem, i.e., the problem of finding $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^n\backslash\{0\}$ such that $(\lambda^2 M + \lambda C + K)x = 0$, where $M, C, K \in \mathbb{C}^{n\times n}$ are Hermitian. This problem arises, for example, in the

*Received by the editors June 11, 2002; accepted for publication (in revised form) by I. Dhillon June 22, 2003; published electronically April 21, 2004. This research was supported by the DFG research center "Mathematics for key technologies" (FZT 86) in Berlin.
http://www.siam.org/journals/simax/25-4/40947.html
†TU Berlin, Institut für Mathematik, D-10623 Berlin, Germany (mehl@math.tu-berlin.de).

analysis of geometrical nonlinear buckling structures with finite element methods (see [1, 12]) or in the theory of damped oscillatory systems (see [9, 18]). With the substitution $\mu = \frac{1}{\lambda}$ for $\lambda \neq 0$, the problem can be linearized such that it reduces to the generalized Hermitian eigenvalue problem

$$\mu \begin{bmatrix} C & M \\ M & 0 \end{bmatrix} \begin{bmatrix} \lambda x \\ x \end{bmatrix} = \begin{bmatrix} -K & 0 \\ 0 & M \end{bmatrix} \begin{bmatrix} \lambda x \\ x \end{bmatrix}. \tag{1.3}$$

There are numerous algorithms for the solution of the generalized Hermitian eigenvalue problem $\lambda \mathcal{G}x = \mathcal{H}x$ for the case that $\mathcal{G}$ (or $\mathcal{H}$, respectively) is positive definite. For example, one could compute the Cholesky factorization $\mathcal{G} = LL^T$ and then consider the standard eigenvalue problem $\lambda I - L^{-1}AL^{-T}$ (see [11] and the references therein). However, there is no software available that takes advantage of the symmetry and possible spectral properties for the case that the generalized eigenvalue problem $\lambda\mathcal{G} - \mathcal{H}$ is indefinite [33], although these problems arise frequently in applications.

On the other hand, it is well known that the indefinite generalized Hermitian eigenvalue problem is related to the standard eigenvalue problem for a matrix that is self-adjoint with respect to an indefinite inner product. Indeed, if $\mathcal{G}$ is nonsingular, then $\lambda\mathcal{G} - \mathcal{H}$ is equivalent to the pencil $\lambda I - \mathcal{G}^{-1}\mathcal{H}$, where $\mathcal{G}^{-1}\mathcal{H}$ is self-adjoint with respect to the indefinite inner product induced by $\mathcal{G}$; i.e., $(\mathcal{G}^{-1}\mathcal{H})^*\mathcal{G} = \mathcal{G}(\mathcal{G}^{-1}\mathcal{H})$. This fact gives rise to the following basic idea: given a method for the solution of the standard eigenvalue problem for a matrix that is structured with respect to an indefinite inner product, try to generalize this method to Hermitian pencils.

In recent years, there has been interest in generalizing Jacobi's algorithm for the symmetric eigenvalue problem to other structured eigenvalue problems such as Hamiltonian or doubly structured eigenvalue problems (see [2, 3, 8]). Hamiltonian matrices are skew-adjoint with respect to the inner product induced by the matrix

$$J := J_n = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \in \mathbb{C}^{2n\times 2n}, \tag{1.4}$$

where I_n denotes the $n\times n$ identity matrix. Thus, a matrix $H \in \mathbb{C}^{2n\times 2n}$ is *Hamiltonian* if and only if $HJ + JH^* = 0$. Analogously, a matrix $S \in \mathbb{C}^{2n\times 2n}$ is called *skew-Hamiltonian* if and only if $SJ - JS^* = 0$. Following the basic idea just mentioned, it will be shown in this paper that Jacobi-like methods for Hamiltonian matrices can be generalized to the case of Hermitian pencils. We will focus on even-sized pencils. A generalization to odd-sized Hermitian pencils is possible, but this needs a more detailed discussion which is not presented here.

The interest in generalized Jacobi-like methods is due to several reasons. First, these methods are inherently parallelizable and backward stable if one restricts oneself to unitary transformation matrices. (For example, backward stability for the Jacobi-like methods proposed in [8] has been shown in [32].) Moreover, Jacobi's classical algorithm is more accurate than the QR-algorithm if a proper stopping criterion is used [5, 21] and has the advantage of converging very fast, if the matrix under consideration is already close to being diagonal. Thus, given a structured eigenvalue problem with a matrix already close to a condensed form that is to be computed, Jacobi-like algorithms are expected to converge much faster than methods that ignore this special property. Hence, Jacobi-like methods may be attractive for the solution of eigenvalue problems that depend on a parameter, e.g., H_∞ control problems [35], where one has to compute the eigenvalues of a Hamiltonian matrix $H(\gamma)$ depending continuously on a real parameter γ. Once a Hamiltonian Schur form for a matrix $H(\gamma_0)$ with some

specific value γ_0 has been computed, the corresponding transformations will transform matrices $H(\gamma)$, where γ is sufficiently close to the γ_0, to a form close to the Hamiltonian Schur form. It is then reasonable to use a Jacobi-like method for the solution of the eigenvalue problem with $H(\gamma)$.

The paper is organized as follows. After reviewing antitriangular forms for Hermitian pencils and relating them to the Hamiltonian Schur form of Hamiltonian matrices in section 2, we will introduce the algorithm JIGH4 (a Jacobi-like method that is based on the solution of 4×4 subproblems and that generalizes the algorithm of Bunse-Gerstner and Faßbender for Hamiltonian matrices [2]) in section 3. In section 4, we will propose the algorithm JIGH2 (a method based on the solution of 2×2 subproblems) that we will prove to be locally quadratically convergent in section 5. Since this algorithm may sometimes stagnate, we propose a slightly modified version called MJIGH2 in section 6 that is also locally quadratically convergent, but does not stagnate in practice. In section 7, we present numerical test results for the comparison of the methods JIGH4 and MJIGH2.

2. Antitriangular forms for Hermitian pencils. The Jacobi-like algorithms that will be presented in this paper are supposed to be structure-preserving algorithms. (The eigenvalues of Hermitian pencils occur in pairs $(\lambda, \bar{\lambda})$ (see, e.g., [31]), and we want to maintain this property.) The structure of Hermitian pencils is preserved under congruence transformations $\lambda\mathcal{G} - \mathcal{H} \mapsto P^*(\lambda\mathcal{G} - \mathcal{H})P$, where P is nonsingular, and since we are interested in restricting ourselves to unitary transformations for the sake of numerical stability, we will consider the problem of finding condensed forms for Hermitian pencils under simultaneous unitary similarity. The classical Schur form would be such that both $\mathcal{G}$ and $\mathcal{H}$ are diagonal. However, it is well known that a pair of Hermitian matrices is simultaneously unitarily diagonalizable if and only if the matrices commute. On the other hand, indefinite Hermitian pencils may have complex conjugate eigenvalues, and in this case a diagonal form cannot exist, even if we allow general congruence transformations instead of unitary ones. An alternative to diagonal forms are the so-called antitriangular forms that have been introduced in [23]. An $n \times n$-matrix $A = (a_{jk})$ is called *lower antitriangular* if $a_{jk} = 0$ for all j, k such that $j + k \leq n$.

DEFINITION 2.1. *We say that a Hermitian pencil $\lambda\mathcal{G}-\mathcal{H} \in \mathbb{C}^{n\times n}$ is in antitriangular form if both $\mathcal{G}$ and $\mathcal{H}$ are lower antitriangular.*

Clearly, antitriangular forms display the eigenvalues of the pencil and a nested set of deflating subspaces. Moreover, antitriangular forms for Hermitian pencils are related to Schur-like forms for skew-Hamiltonian/Hamiltonian pencils that have been discussed in [22], and thus, they are related to Hamiltonian Schur forms for Hamiltonian matrices. (A skew-Hamiltonian/Hamiltonian pencil is a pencil $\lambda S - H$ such that S is skew-Hamiltonian and H is Hamiltonian.) If $\lambda S - H$ is a skew-Hamiltonian/Hamiltonian pencil in Schur-like form, then $\lambda iJS - JH$ is a Hermitian pencil that is congruent to a pencil in antitriangular form:

$$\lambda S - H = \left[\begin{array}{cc} \diagdown & \square \\ & \diagdown \end{array}\right] \quad \Rightarrow \quad \lambda iJS - JH = \left[\begin{array}{cc} & \diagdown \\ \diagdown & \square \end{array}\right] \sim \left[\begin{array}{cc} \diagup & \\ \square & \diagup \end{array}\right].$$

(Here, $\sim$ denotes congruence.) From this point of view, antitriangular forms are the natural forms to look for if one is interested in obtaining condensed forms for indefinite Hermitian pencils under unitary transformations.

It is well known that the Hamiltonian Schur form for Hamiltonian matrices or,

analogously, the Schur-like form for skew-Hamiltonian/Hamiltonian pencils does not always exist if the matrix or pencil under consideration has purely imaginary eigenvalues. A similar observation can be made for antitriangular forms for Hermitian pencils. Here, real eigenvalues are the ones that might cause problems. A necessary and sufficient condition for the existence of antitriangular forms for Hermitian pencils was obtained in [23, Theorem 15 and Corollary 21]. We will not quote those results in full generality, but only the following important special case.

PROPOSITION 2.2. *Let $\lambda\mathcal{G} - \mathcal{H} \in \mathbb{C}^{n\times n}$ be a Hermitian pencil having no real eigenvalues if n is even, or exactly one real eigenvalue (counting multiplicities) if n is odd. Then there exists a unitary matrix $\mathcal{U} \in \mathbb{C}^{n\times n}$ such that $\lambda\mathcal{U}^*\mathcal{G}\mathcal{U} - \mathcal{U}^*\mathcal{H}\mathcal{U}$ is in antitriangular form.*

3. A method working on 4×4 subproblems (JIGH4). In this section, we generalize the Jacobi-like algorithm for Hamiltonian matrices by Bunse-Gerstner and Faßbender [2] to Hermitian pencils. We start with the following short survey.

3.1. A short survey on Jacobi methods. The idea of Jacobi's method [16] for the diagonalization of a symmetric matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is to successively apply similarity transformations with rotation matrices such that each one diagonalizes a particular 2×2 submatrix of A. In each step, one chooses a pivot pair of indices (k, l) and applies a similarity transformation with a rotation matrix $U = (u_{ij})$ that annihilates the entries (l, k) and (k, l) of A. Here, U coincides with the identity except for the elements $u_{kk} = u_{ll} = \cos\alpha$ and $-u_{kl} = u_{lk} = \sin\alpha$.

The *classical* Jacobi algorithm chooses indices (k, l) such that a_{kl}^2 is maximal in each step while *cyclic* Jacobi methods use fixed sequences of indices (k, l) such as

$$(k, l) = (1, 2), (1, 3), \ldots, (1, n), (2, 3), \ldots, (2, n), \ldots, (n - 1, n),$$

where every possible index pair is considered exactly once. Usually, the performance of $n(n-1)/2$ Jacobi steps is called a *sweep*, and hence, cyclic Jacobi methods consider every possible index pair once in a sweep. Under certain conditions both the classical Jacobi method and the cyclic Jacobi methods are convergent and their asymptotic convergence rate is quadratic (see [14, 17, 27, 34]).

Jacobi's method has been adapted to other classes of matrices; see, e.g., [2, 3, 6, 8, 10, 13, 26]. In particular, Stewart [29] and Eberlein [7] generalized Jacobi's method to the computation of the Schur decomposition of a general complex matrix. Similarly to Jacobi's original method, a pivot pair (k, l) is chosen and a similarity transformation with a rotation $U = (u_{ij})$ is applied that yields the Schur decompositions of the corresponding 2×2 subproblem. Again, U coincides with the identity except for the elements $u_{kk} = u_{ll} = \cos\alpha$, $u_{kl} = -e^{-i\theta}\sin\alpha$, and $u_{lk} = e^{i\theta}\sin\alpha$. While Stewart proposed to use only pivot elements from the lower subdiagonal, Eberlein proposed to allow all elements from the lower triangular part of the matrix as pivot elements. Moreover, the methods differ in the choice of rotation parameters. Charlier and van Dooren [4] generalized Stewart's method to the computation of the generalized Schur decomposition of matrix pencils and were able to prove global convergence under certain restrictions, whereas there is as yet no convergence proof for the method of Eberlein. It has been observed, however, that Eberlein's method converges faster than Stewart's method if the matrix under consideration is not close to being normal.

In 1990, Byers [3] adapted Stewart's method to the computation of the Hamiltonian Schur form of a Hamiltonian matrix A. In order to preserve the structure of Hamiltonian matrices in each step, the proposed algorithm considers 4×4 subproblems rather than 2×2 subproblems. However, it has been observed that convergence

may be very slow if the matrix under consideration is not close to a normal matrix. Sometimes, the method does not converge at all. One problem is that some of the 4×4 subproblems may have purely imaginary eigenvalues, i.e., eigenvalues with zero real part. In this case, the Hamiltonian Schur form for the subproblem need not exist. Hence, a reduction as proposed in the algorithm may be impossible. Subproblems with purely imaginary eigenvalues may occur even when the original matrix A has no purely imaginary eigenvalues.

In 1997, Bunse-Gerstner and Faßbender proposed a different Jacobi-like algorithm that generalizes Eberlein's method. Again, the algorithm considers 4×4 subproblems instead of 2×2 subproblems. As for Byers' method, subproblems with purely imaginary eigenvalues may occur such that a reduction in the current step is not necessarily possible. However, the authors were able to observe convergence for all their test problems and found that the performance of their method was superior to that of Byers' method when the Hamiltonian matrix under consideration was not normal. Unfortunately, there is no convergence proof for this method so far.

It is the aim of this section to generalize Jacobi-like methods for Hamiltonian matrices to the case of indefinite Hermitian pencils. Given the fact that the convergence behavior of Bunse-Gerstner and Faßbender's method is superior to that of Byers' method if the matrix under consideration is not normal (a pair of simultaneously unitarily antidiagonalizable Hermitian matrices could be interpreted as a Hermitian pencil corresponding to a normal Hamiltonian matrix), we will focus on the first method. Moreover, we will restrict ourselves to the case of even-sized pencils. Generalizations to the case of odd-sized pencils are possible but will involve the solution of some 3×3 subproblems. Details will not be presented in this paper. (Note that 3×3 Hermitian pencils always have at least one real eigenvalue. However, if the pencil has exactly one real eigenvalue, then the antitriangular form always exists (see Proposition 2.2). The real eigenvalue is then displayed in the middle of the antidiagonal.)

3.2. Solving the eigenvalue problem for 4×4 Hermitian pencils. The antitriangular form for 4×4 Hermitian pencils (provided that this form exists) can be computed by adapting the method for the computation of Schur-like forms for skew-Hamiltonian/Hamiltonian pencils developed in [22]. This requires an a priori knowledge of eigenvalues and eigenvectors. (In our MATLAB implementation, the routine `eig` has been used for the solution of the 4×4 eigenvalue problems.) If $\lambda\mathcal{G} - \mathcal{H}$ is a regular 4×4 Hermitian pencil we have to distinguish three different cases:

(i) $\lambda\mathcal{G} - \mathcal{H}$ has two pairs of complex conjugate eigenvalues;
(ii) $\lambda\mathcal{G} - \mathcal{H}$ has two real eigenvalues and a pair of complex conjugate eigenvalues;
(iii) $\lambda\mathcal{G} - \mathcal{H}$ has four real eigenvalues.

Here, possible infinite eigenvalues are considered to be real eigenvalues. If we are in case (i) or (ii), let v be an eigenvector associated with a nonreal eigenvalue λ_1 of $\lambda\mathcal{G} - \mathcal{H}$. Then $\mathcal{G}v$ (or $\mathcal{H}v$) and v are orthogonal because

$$\lambda_1 v^*\mathcal{G}v = v^*\mathcal{H}v = (v^*\mathcal{H}v)^* = \bar{\lambda}_1 v^*\mathcal{G}v.$$

Thus, there exists a unitary matrix $Q = [q1, q2, q3, q4]$ such that $q1 = v/\|v\|$ and $q4 = \mathcal{G}v/\|\mathcal{G}v\|$. We then obtain

$$Q^*(\lambda\mathcal{G} - \mathcal{H})Q = \lambda \begin{bmatrix} 0 & 0 & 0 & g_{14} \\ 0 & g_{22} & g_{23} & g_{24} \\ 0 & \overline{g}_{23} & g_{33} & g_{34} \\ \overline{g}_{14} & \overline{g}_{24} & \overline{g}_{34} & g_{44} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & h_{14} \\ 0 & h_{22} & h_{23} & h_{24} \\ 0 & \overline{h}_{23} & h_{33} & h_{34} \\ \overline{h}_{14} & \overline{h}_{24} & \overline{h}_{34} & h_{44} \end{bmatrix}. \tag{3.1}$$

The eigenvalues of this pencil are $\lambda_1 = \overline{h}_{14}/\overline{g}_{14}$ and $\bar{\lambda}_1$ together with the eigenvalues of the subpencil

$$\lambda \begin{bmatrix} g_{22} & g_{23} \\ \overline{g}_{23} & g_{33} \end{bmatrix} - \begin{bmatrix} h_{22} & h_{23} \\ \overline{h}_{23} & h_{33} \end{bmatrix}. \tag{3.2}$$

If we are in case (ii), then (3.2) has two real eigenvalues and we stop the reduction process. The form (3.1) will be called *partial antitriangular form*. If we are in case (i), then (3.2) has a pair of complex conjugate eigenvalues and we repeat the procedure above to transform this subpencil to antitriangular form. For our 4×4 pencil, this means that there exists a unitary matrix $\tilde{Q}$ such that

$$\tilde{Q}^*(\lambda\mathcal{G} - \mathcal{H})\tilde{Q} = \lambda \begin{bmatrix} 0 & 0 & 0 & g_{14} \\ 0 & 0 & \tilde{g}_{23} & \tilde{g}_{24} \\ 0 & \overline{\tilde{g}}_{23} & \tilde{g}_{33} & \tilde{g}_{34} \\ \overline{g}_{14} & \overline{\tilde{g}}_{24} & \overline{\tilde{g}}_{34} & g_{44} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & h_{14} \\ 0 & 0 & \tilde{h}_{23} & \tilde{h}_{24} \\ 0 & \overline{\tilde{h}}_{23} & \tilde{h}_{33} & \tilde{h}_{34} \\ \overline{h}_{14} & \overline{\tilde{h}}_{24} & \overline{\tilde{h}}_{34} & h_{44} \end{bmatrix}.$$

Clearly, there is a continuum of simultaneous unitary similarity transformations that bring $\lambda\mathcal{G}-\mathcal{H}$ to antitriangular form, if there are no real eigenvalues. Thus, the question arises, which one to choose, in order to guarantee best convergence properties of the corresponding Jacobi-like method. One could choose the matrix $\tilde{Q}$ that is nearest to the identity or such that the sum of the absolute values of the left lower triangular part of $\tilde{Q}$ is minimized. At this stage, we do not know which strategy is the best. The two strategies mentioned above work well for pencils that are already close to antitriangular form, but they seem to slow down the convergence process for random Hermitian pencils. In our numerical experiments, we observed fastest convergence by using the following strategy for choosing the eigenvectors used in the reduction procedure for case (i) and case (ii) above:

1. For the first reduction step (towards partial antitriangular form), consider only eigenvectors associated with eigenvalues with negative imaginary part and among those choose the normalized eigenvector that is closest to the first unit vector.

2. For the second reduction step (only in case (i)), choose the eigenvector associated with the eigenvalue with negative imaginary part.

Thus, we consider only eigenvectors associated with eigenvalues with negative imaginary parts. (Clearly, one can also consider an analogous strategy that chooses only eigenvectors associated with eigenvalues with positive imaginary part.)

3.3. The algorithm JIGH4. We have provided all ingredients to formulate the algorithm JIGH4 (Jacobi-like algorithm for the indefinite generalized Hermitian eigenvalue problem based on 4×4 subproblems). We use a *cyclic-by-row*-type ordering scheme of pivot indices, e.g., in the 8×8 case,

$$(1,2,7,8),(1,3,6,8),(1,4,5,8),(2,3,6,7),(2,4,5,7),(3,4,5,6),(1,2,7,8),\ldots.$$

Throughout the rest of this paper, if $X = (x_{ij}) \in \mathbb{C}^{2n\times 2n}$, let $X_{kl}^{(4)}$ denote the matrix

$$X_{kl}^{(4)} = \begin{bmatrix} x_{kk} & x_{kl} & x_{k,2n+1-l} & x_{k,2n+1-k} \\ x_{lk} & x_{ll} & x_{l,2n+1-l} & x_{l,2n+1-k} \\ x_{2n+1-l,k} & x_{2n+1-l,l} & x_{2n+1-l,2n+1-l} & x_{2n+1-l,2n+1-k} \\ x_{2n+1-k,k} & x_{2n+1-k,l} & x_{2n+1-k,2n+1-l} & x_{2n+1-k,2n+1-k} \end{bmatrix}.$$

Then algorithm JIGH4 takes the form as given below.

Algorithm JIGH4: Given a $2n \times 2n$ Hermitian pencil $\lambda\mathcal{G} - \mathcal{H}$ having no real eigenvalues, a stopping criterion, and a strategy for the solution of 4×4 subproblems, the algorithm computes the antitriangular form of $\lambda\mathcal{G} - \mathcal{H}$.

while stopping criterion not satisfied
  for $k = 1, \ldots, n-1$
    for $l = k+1, \ldots, n$
      if $\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)}$ is singular or has real eigenvalues only
        $Q = I_4$;
      elseif $\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)}$ has no real eigenvalues
        compute a unitary Q such that $Q^*(\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)})Q$ is in antitriangular form;
      else
        compute a unitary Q such that $Q^*(\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)})Q$ is in partial antitriangular form;
      end
      set $\tilde{Q} := I_{2n}$ and $\tilde{Q}_{kl}^{(4)} := Q$;
      set $\mathcal{G} := \tilde{Q}^*\mathcal{G}\tilde{Q}$ and $\mathcal{H} := \tilde{Q}^*\mathcal{H}\tilde{Q}$;
    end
  end
end

Following [2], a "sweep" denotes one performance of the "while"-loop, although this does not correspond to $n(n-1)/2$ single Jacobi-steps. Unfortunately, as for the Hamiltonian Jacobi-like method by Bunse-Gerstner and Faßbender, there is no convergence proof for JIGH4, but convergence can be observed in numerical experiments.

4. A method working on 2×2 subproblems (JIGH2). In this section, we discuss the algorithm JIGH2 (Jacobi-like algorithm for the indefinite generalized Hermitian eigenvalue problem based on 2×2 subproblems). The reason for considering 4×4 subproblems instead of 2×2 subproblems in JIGH4 was that we wanted to deal with subproblems which have a structure corresponding to the one of the original problem. For example, if we consider the 8×8 pencil sketched below, then the 4×4 subpencil indicated by the discs is the smallest Hermitian subpencil that contains the $(2,3)$-entry of the pencil and that may be used to transport "weight" from the upper antitriangular part of the pencil into the lower antitriangular part.

$$\lambda \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \circ & \circ & \cdot & \cdot & \circ & \bullet & \cdot \\ \cdot & \circ & \circ & \cdot & \cdot & \bullet & \bullet & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \circ & \bullet & \cdot & \cdot & \bullet & \bullet & \cdot \\ \cdot & \bullet & \bullet & \cdot & \cdot & \bullet & \bullet & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \circ & \circ & \cdot & \cdot & \circ & \bullet & \cdot \\ \cdot & \circ & \circ & \cdot & \cdot & \bullet & \bullet & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \circ & \bullet & \cdot & \cdot & \bullet & \bullet & \cdot \\ \cdot & \bullet & \bullet & \cdot & \cdot & \bullet & \bullet & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Thus, if we want to work on 2×2 subproblems, some of them have to be non-Hermitian. For the rest of the paper, if $X = (x_{ij}) \in \mathbb{C}^{2n\times 2n}$, let $X_{kl}^{(2)}$ denote the matrix

$$X_{kl}^{(2)} = \begin{bmatrix} x_{kl} & x_{k,2n+1-k} \\ x_{2n+1-l,l} & x_{2n+1-l,2n+1-k} \end{bmatrix}.$$

Suppose that we want to eliminate the (k,l)-entry (and thus, simultaneously, the (l,k)-entry) of the pencil $\lambda\mathcal{G}-\mathcal{H}$, where we assume $k<l\le n$. This entry is contained in the 2×2 subpencil $\lambda\mathcal{G}_{kl}^{(2)}-\mathcal{H}_{kl}^{(2)}$ which we assume to be regular. This subpencil is indicated by discs in the sketch (4.1) below for the 8×8 case with $(k,l)=(2,3)$.

(4.1)
$$\lambda\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \circ & \cdot & \cdot & \cdot & \bullet & \cdot \\ \cdot & + & \cdot & \cdot & \cdot & + & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot & \cdot & \bullet & \cdot \\ \cdot & + & \cdot & \cdot & \cdot & + & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \circ & \cdot & \cdot & \cdot & \bullet & \cdot \\ \cdot & + & \cdot & \cdot & \cdot & + & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \bullet & \cdot & \cdot & \cdot & \bullet & \cdot \\ \cdot & + & \cdot & \cdot & \cdot & + & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

(Analogously, the subpencil indicated by plus signs contains the $(3,2)$-entry, but for the moment, we ignore this subpencil.) Clearly, the subpencil $\lambda\mathcal{G}_{kl}^{(2)}-\mathcal{H}_{kl}^{(2)}$ is no longer Hermitian, but we can use it to transport "weight" from the upper antitriangular part of the pencil into the lower antitriangular part by applying an antitriangular version of the generalized Schur decomposition; i.e., we compute unitary matrices $U=(u_{ij})$ and $V=(v_{ij})$ such that $U(\lambda\tilde{\mathcal{G}}-\tilde{\mathcal{H}})V$ is in antitriangular form. Indeed, if $w=(w_1,\ w_2)^T$ is a normalized eigenvector of $\lambda\tilde{\mathcal{G}}-\tilde{\mathcal{H}}$ and if $x=(x_1,\ x_2)^T=\tilde{\mathcal{G}}w/\|\tilde{\mathcal{G}}w\|$ (or $x=(x_1,\ x_2)^T=\tilde{\mathcal{H}}w/\|\tilde{\mathcal{H}}w\|$ if $\tilde{\mathcal{G}}w=0$), then

$$U=\begin{bmatrix} -x_2 & x_1 \\ \overline{x}_1 & \overline{x}_2 \end{bmatrix},\quad V=\begin{bmatrix} w_1 & -\overline{w}_2 \\ w_2 & \overline{w}_1 \end{bmatrix}$$

are unitary and $U(\lambda\tilde{\mathcal{G}}-\tilde{\mathcal{H}})V$ is in antitriangular form.

In order to preserve the structure of the Hermitian pencil, we define $\tilde{Q}$ to be the matrix that differs from the identity I_{2n} only in the submatrix

(4.2)
$$\tilde{Q}_{kl}^{(4)}=Q=\begin{bmatrix} \overline{u}_{11} & 0 & \overline{u}_{21} & 0 \\ 0 & v_{11} & 0 & v_{12} \\ \overline{u}_{12} & 0 & \overline{u}_{22} & 0 \\ 0 & v_{21} & 0 & v_{22} \end{bmatrix}\quad\text{or}\quad \tilde{Q}_{kl}^{(4)}=Q=\begin{bmatrix} \overline{u}_{11} & \overline{u}_{21} & 0 & 0 \\ \overline{u}_{12} & \overline{u}_{22} & 0 & 0 \\ 0 & 0 & v_{11} & v_{12} \\ 0 & 0 & v_{21} & v_{22} \end{bmatrix},$$

if $l<n+1-l$ or $l>n+1-l$, respectively. (Note that the case $l=n+1-l$ cannot occur if n is even.) Then, we transform the pencil via $\tilde{Q}^*(\lambda\mathcal{G}-\mathcal{H})\tilde{Q}$. To illustrate the effect of this simultaneous similarity transformation, we again refer to the sketch in (4.1). Note that in the updated pencil, the subpencil indicated by the bold discs in (4.1) is just $U(\lambda\tilde{\mathcal{G}}-\tilde{\mathcal{H}})V$. Since we used a structure preserving transformation, the subpencil indicated by the plus signs is also transformed to antitriangular form. From this point of view, we again worked on a 4×4 subpencil, but the transformation was computed by solving an (unstructured) 2×2 problem only. We note that the unstructured subpencil $\lambda\tilde{\mathcal{G}}-\tilde{\mathcal{H}}$ can always be reduced to antitriangular form. This is different if we want to eliminate the (k,k)-entry of the Hermitian pencil, where $k\le n$. Then, we have to consider the 2×2 Hermitian subpencil $\lambda\mathcal{G}_{kk}^{(2)}-\mathcal{H}_{kk}^{(2)}$ and if this pencil has real eigenvalues, then an antitriangular form under simultaneous unitary similarity need not exist. (In this case, we do not transform the subpencil at all.) During the computation of the 2×2 antitriangular forms, there are two choices for the transformation matrices. As a strategy, we suggest choosing the transformation

matrices that are closest to the identity; i.e., in the computations, we start with the (normalized) eigenvector that is closest to the first unit vector.

Algorithm JIGH2: Given a stopping criterion and a strategy for the solution of 2×2 subproblems, the algorithm computes the antitriangular form of a $2n \times 2n$ Hermitian pencil $\lambda\mathcal{G} - \mathcal{H}$ having no real eigenvalues.

while stopping criterion not satisfied
  for $k = 1, \ldots, n$
    if $\lambda\mathcal{G}_{kk}^{(2)} - \mathcal{H}_{kk}^{(2)}$ is regular and has no real eigenvalues
      compute a unitary $Q \in \mathbb{C}^{2\times 2}$ such that $Q^*(\lambda\mathcal{G}_{kk}^{(2)} - \mathcal{H}_{kk}^{(2)})Q$ is in antitriangular form;
    else
      set $Q = I_2$;
    end
    set $\tilde{Q} := I_{2n}$ and $\tilde{Q}_{kk}^{(2)} := Q$;
    set $\mathcal{G} := \tilde{Q}^*\mathcal{G}\tilde{Q}$ and $\mathcal{H} := \tilde{Q}^*\mathcal{H}\tilde{Q}$;
    for $l = k+1, \ldots, 2n-k$
      if $\lambda\mathcal{G}_{kl}^{(2)} - \mathcal{H}_{kl}^{(2)}$ is regular
        compute unitary $U, V \in \mathbb{C}^{2\times 2}$ such that $U(\lambda G_{kl}^{(2)} - H_{kl}^{(2)})V$ is in antitriangular form and set Q as in (4.2);
        set $\tilde{Q} := I_{2n}$ and $\tilde{Q}_{kl}^{(4)} := Q$;
        set $\mathcal{G} := \tilde{Q}^*\mathcal{G}\tilde{Q}$ and $\mathcal{H} := \tilde{Q}^*\mathcal{H}\tilde{Q}$;
      end
    end
  end
end

Again, a "sweep" will denote one complete performance of the "while"-loop. Since JIGH2 works on 2×2 subproblems rather than on 4×4 problems, it can be considered a direct generalization of Eberlein's method [7] rather than a generalization of the method of Bunse-Gerstner and Faßbender [2].

5. Convergence properties of JIGH2. In this section, we consider the convergence properties of JIGH2. It is well known that Jacobi's classical algorithm is asymptotically quadratically convergent (see [17, 27]). The same is also known for several generalizations; see [15] for a general proof of local quadratic convergence of Jacobi-type methods. However, the results obtained there are based on the minimization of a particular smooth function. For the standard eigenvalue problem with a symmetric matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, this smooth function is the so-called *off-norm*, i.e., the sum over all squares of off-diagonal elements. The corresponding function for a Hermitian pencil $\lambda\mathcal{G} - \mathcal{H}$, $\mathcal{G} = (g_{ij})$, $\mathcal{H} = (h_{ij})$ would be

$$\sigma(\mathcal{G}, \mathcal{H}) := \sqrt{\sum_{i=1}^{2n-1} \sum_{j=1}^{i} \left(|g_{j,i+1-j}|^2 + |h_{j,i+1-j}|^2\right)}. \tag{5.1}$$

Experiments show that $\sigma(\mathcal{G}, \mathcal{H})$ is not necessarily decreasing, at least not at the beginning at the process. Therefore, and since it is not guaranteed that all occurring 2×2 subproblems can be solved, a detailed discussion of the convergence properties of JIGH2 is necessary. In the following, we will proof local quadratic convergence of

JIGH2 provided that for the solution of 2×2 subproblems the strategy is used that chooses the transformation matrices that are closest to the identity.

First, let us introduce some notation which is adopted from [28] (see also [30]). Given two pairs $(a_1, b_1), (a_2, b_2) \in (\mathbb{C} \times \mathbb{C}) \setminus \{(0,0)\}$, let

$$\operatorname{dif}\Big((a_1, b_1), (a_2, b_2)\Big) := \inf_{p,q:\max(|p|,|q|)=1} \max(pa_1 + qa_2,\ pb_1 + qb_2).$$

Then by [30, Theorem VI.1.11] we have that $\operatorname{dif}((a_1, b_1), (a_2, b_2)) > 0$ if and only if the spectra of the 1×1 pencils $\lambda a_1 - b_1$ and $\lambda a_2 - b_2$ are disjoint.

LEMMA 5.1. *Let $\varepsilon, \eta > 0$ and $\lambda\mathcal{G} - \mathcal{H} \in \mathbb{C}^{2\times 2}$, $\mathcal{G} = (g_{ij})$, $\mathcal{H} = (h_{ij})$ be such that $|g_{11}|, |h_{11}| < \varepsilon$, $|g_{22}|, |h_{22}| < \eta$, and $\varrho = \operatorname{dif}((g_{12}, h_{12}), (\overline{g}_{12}, \overline{h}_{12})) > 0$. If $4\eta\varepsilon/\varrho^2 < 1$, then $\lambda\mathcal{G} - \mathcal{H}$ has no real eigenvalues and, in particular, $\lambda\mathcal{G} - \mathcal{H}$ can be transformed to antitriangular form.*

Proof. With the assumptions above, $\lambda\mathcal{G} - \mathcal{H}$ has two distinct eigenvalues by [30, Theorem VI.2.13]. For $t \in \mathbb{R}$ consider the pencil

$$P(t) := \lambda \begin{bmatrix} tg_{11} & g_{12} \\ \overline{g}_{12} & g_{22} \end{bmatrix} - \begin{bmatrix} th_{11} & h_{12} \\ \overline{h}_{12} & h_{22} \end{bmatrix}.$$

Then $P(t)$ is a regular Hermitian pencil for all $t \in [0,1]$ (this follows by investigating its determinant) and $P(1) = \lambda\mathcal{G} - \mathcal{H}$. Since $\varrho > 0$, the pencil $P(0)$ has two complex conjugate eigenvalues. Assume that the eigenvalues of $P(1)$ are real. Consider

$$\tilde{t} := \inf_{t\in[0,1]} \{t :\ P(t) \text{ has real eigenvalues}\}.$$

Since the eigenvalues of regular pencils depend continuously on the entries of the pencil, we obtain that $P(\tilde{t})$ has a real eigenvalue with multiplicity two. Since $|\tilde{t}g_{11}|, |\tilde{t}h_{11}| < \varepsilon$, this contradicts Theorem VI.2.13. in [30]. (This theorem states, in particular, that the eigenvalues must be distinct.) Thus, $\lambda\mathcal{G} - \mathcal{H}$ cannot have real eigenvalues. ∎

Before we show local convergence of JIGH2, let us analyze what happens in one single Jacobi-step of the algorithm. Let us assume that we are in the μth sweep and that we perform the pth Jacobi-step in this sweep. Since any sweep consists of

$$s := \frac{n^2}{4} = \sum_{k=1}^{n/2} (2k-1)$$

Jacobi-steps, we are performing step number $\nu := \mu s + p$, currently working on the updated pencil $\lambda\mathcal{G}_\nu - \mathcal{H}_\nu$, $\mathcal{G}_\nu = (g_{ij}^{(\nu)})$, $\mathcal{H}_\nu = (h_{ij}^{(\nu)})$, where $\lambda\mathcal{G}_0 - \mathcal{H}_0 := \lambda\mathcal{G} - \mathcal{H}$. Assume, furthermore, that we want to eliminate the (k,l)-entry (and the (l,k)-entry if $k \neq l$) of the current pencil $\lambda\mathcal{G}_\nu - \mathcal{H}_\nu$, where $k \leq l \leq n$. Let

$$\delta_\nu := \max\Big\{|g_{ij}^{(\nu)}|, |h_{ij}^{(\nu)}| \ \Big|\ i < k \text{ or } (i = k \text{ and } i \leq j < l)\Big\}, \tag{5.2}$$

$$\varepsilon_\nu := \max\Big\{|g_{ij}^{(\nu)}|, |h_{ij}^{(\nu)}| \ \Big|\ i + j \leq n\Big\}, \tag{5.3}$$

$$\eta_\nu := \max\Big\{|g_{ij}^{(\nu)}|, |h_{ij}^{(\nu)}| \ \Big|\ i + j \geq n + 1\Big\}, \tag{5.4}$$

$$\varrho_\nu := \min_{i\neq j} \operatorname{dif}\Big((g_{i,n+1-i}^{(\nu)}, h_{i,n+1-i}^{(\nu)}), (g_{j,n+1-j}^{(\nu)}, h_{j,n+1-j}^{(\nu)})\Big), \tag{5.5}$$

where we assume $\varrho_\nu > 0$ and $4\varepsilon_\nu\eta_\nu/\varrho_\nu^2 < 1$. Thus, ε_ν and η_ν, respectively, are the largest absolute values of entries of $\mathcal{G}_\nu$ and $\mathcal{H}_\nu$ in the strict upper antitriangular part

or in the lower antitriangular part (including the antidiagonal), respectively, while δ_ν is the largest absolute value of entries in the strict upper antitriangular part that have already been considered (and possibly annihilated) once in the current sweep. In the following, we will distinguish two cases.

Case 1 ($k = l$). Then the current step of JIGH2 computes a unitary $Q \in \mathbb{C}^{2\times 2}$ such that

$$Q^* \left(\lambda \begin{bmatrix} g_{kk} & g_{k,n+1-k} \\ \overline{g}_{k,n+1-k} & g_{n+1-k,n+1-k} \end{bmatrix} - \begin{bmatrix} h_{kk} & h_{k,n+1-k} \\ \overline{h}_{k,n+1-k} & h_{n+1-k,n+1-k} \end{bmatrix} \right) Q$$

is in antitriangular form. Note that such a Q exists by Lemma 5.1. Clearly, Q can be taken of the form

$$Q = \begin{bmatrix} \cos x & -e^{-i\alpha} \sin x \\ e^{i\alpha} \sin x & \cos x \end{bmatrix}$$

for some $x, \alpha \in \mathbb{R}$. By Theorem VI.2.13 in [30], we then obtain $|\sin x| < 2\varepsilon_\nu/\varrho_\nu$. Note that only elements in the kth and $(n+1-k)$th rows and columns of the updated pencil $\lambda G_{\nu+1} - H_{\nu+1}$ have been changed. For $i < k$, we obtain

$$|g_{ik}^{(\nu+1)}| = |g_{ik}^{(\nu)} \cos x + g_{i,n+1-k}^{(\nu)} e^{i\alpha} \sin x| \leq |g_{ik}^{(\nu)}| + |g_{i,n+1-k}^{(\nu)}| \cdot |\sin x| < \delta_\nu \left(1 + 2\frac{\varepsilon_\nu}{\varrho_\nu} \right).$$

The same bound holds for $|h_{ik}^{(\nu+1)}|$. Similarly, we obtain for $k < i < n+1-k$ that

$$|g_{ik}^{(\nu+1)}|, |h_{ik}^{(\nu+1)}| < \varepsilon_\nu \left(1 + 2\frac{\eta_\nu}{\varrho_\nu} \right).$$

Moreover, we obtain

$$\begin{aligned} g_{n+1-k,k}^{(\nu+1)} &= -g_{kk}^{(\nu)} e^{i\alpha} \sin x \cos x + g_{n+1-k,k}^{(\nu)} \cos^2 x - g_{k,n+1-k}^{(\nu)} (e^{i\alpha} \sin x)^2 \\ &\quad + g_{n+1-k,n+1-k}^{(\nu)} e^{i\alpha} \sin x \cos x \\ &= g_{n+1-k,k}^{(\nu)} + \Delta_k \quad (\text{using } \cos^2 x = 1 - \sin^2 x), \end{aligned}$$

where

$$|\Delta_k| < \varepsilon_\nu \cdot 2\frac{\varepsilon_\nu}{\varrho_\nu} + \eta_\nu \cdot 4\frac{\varepsilon_\nu^2}{\varrho_\nu^2} + \eta_\nu \cdot 4\frac{\varepsilon_\nu^2}{\varrho_\nu^2} + \eta_\nu \cdot 2\frac{\varepsilon_\nu}{\varrho_\nu} = 2\left(\frac{\varepsilon_\nu^2}{\varrho_\nu} + 4\eta_\nu \frac{\varepsilon_\nu^2}{\varrho_\nu^2} + \eta_\nu \frac{\varepsilon_\nu}{\varrho_\nu} \right).$$

A corresponding result holds for $h_{n+1-k,k}^{(\nu+1)}$. Then the inequality

$$\operatorname{dif}((a_1+e_1, b_1+f_1), (a_2+e_2, b_2+f_2)) \geq \operatorname{dif}((a_1, b_1), (a_2, b_2)) - \max(|e_1|+|e_2|, |f_1|+|f_2|)$$

(see (2.19) in [30]) implies

$$\operatorname{dif}\Big(\left(g_{n+1-i,i}^{(\nu+1)}, h_{n+1-i,i}^{(\nu+1)} \right), \left(g_{n+1-j,j}^{(\nu+1)}, h_{n+1-j,j}^{(\nu+1)} \right) \Big) > \varrho_\nu - 4\left(\frac{\varepsilon_\nu^2}{\varrho_\nu} + 4\eta_\nu \frac{\varepsilon_\nu^2}{\varrho_\nu^2} + \eta_\nu \frac{\varepsilon_\nu}{\varrho_\nu} \right)$$

for all $i, j = 1, \ldots, n$, $i \neq j$.

Case 2 ($k \neq l$). Then the current step of JIGH2 computes unitary matrices $P, Q \in \mathbb{C}^{2\times 2}$ such that

$$P^* \left(\lambda \begin{bmatrix} g_{kl}^{(\nu)} & g_{k,n+1-k}^{(\nu)} \\ g_{n+1-l,l}^{(\nu)} & g_{n+1-l,n+1-k}^{(\nu)} \end{bmatrix} - \begin{bmatrix} h_{kl}^{(\nu)} & h_{k,n+1-k}^{(\nu)} \\ h_{n+1-l,l}^{(\nu)} & h_{n+1-l,n+1-k}^{(\nu)} \end{bmatrix} \right) Q$$

is in antitriangular form. Without loss of generality, we may assume that P and Q have the forms

$$P^* = \begin{bmatrix} \cos x & -e^{-i\alpha}\sin x \\ e^{i\alpha}\sin x & \cos x \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} \cos y & -e^{-i\beta}\sin y \\ e^{i\beta}\sin y & \cos y \end{bmatrix},$$

where $x, y, \alpha, \beta \in \mathbb{R}$. By Theorem VI.2.13 in [30], we have $|\sin x|, |\sin y| < 2\varepsilon_\nu/\varrho_\nu$. Only elements of the kth, lth, $(n+1-l)$th, and $(n+1-k)$th rows and columns of $G_{\nu+1}$ and $H_{\nu+1}$ have been changed. If $l < n+1-l$, then we obtain that

$$\begin{aligned} |g_{kk}^{(\nu+1)}| &= |g_{kk}^{(\nu+1)}\cos^2 x - g_{n+1-l,k}^{(\nu+1)} e^{-i\alpha}\cos x \sin x - g_{k,n+1-l}^{(\nu+1)} e^{i\alpha}\cos x \sin x \\ &\quad + g_{n+1-l,n+1-l}^{(\nu+1)}\sin^2 x| \\ &< \delta_\nu + \varepsilon_\nu \cdot 2\frac{\varepsilon_\nu}{\varrho_\nu} + \varepsilon_\nu \cdot 2\frac{\varepsilon_\nu}{\varrho_\nu} + \eta_\nu \cdot 4\frac{\varepsilon_\nu^2}{\varrho_\nu^2} = \delta_\nu + 4\frac{\varepsilon_\nu^2}{\varrho_\nu} + 4\frac{\varepsilon_\nu^2\eta_\nu}{\varrho_\nu^2}, \end{aligned}$$

and if, furthermore, $k < i < l$, then we obtain that

$$|g_{ik}^{(\nu+1)}| = |g_{ik}^{(\nu)}\cos y - g_{i,n+1-l}^{(\nu)} e^{i\beta}\sin y| < \delta_\nu + 2\frac{\varepsilon_\nu^2}{\varrho_\nu}.$$

By this and similar computations, we finally obtain for the case $l < n+1-l$ that $|g_{ij}^{(\nu+1)}|$ and $|h_{ij}^{(\nu+1)}|$ have the following upper bounds:

	$j=k$	$j=l$	$j=n+1-l$
$i=k$	$\delta_\nu + \frac{4\varepsilon_\nu^2}{\varrho_\nu} + \frac{4\varepsilon_\nu^2\eta_\nu}{\varrho_\nu^2}$	0	$\varepsilon_\nu\left(1 + 2\frac{\delta_\nu+\eta_\nu}{\varrho_\nu} + \frac{4\varepsilon_\nu^2}{\varrho_\nu^2}\right)$
$k<i<l$	$\delta_\nu + 2\frac{\varepsilon_\nu^2}{\varrho_\nu}$	$\varepsilon_\nu\left(1 + 2\frac{\eta_\nu}{\varrho_\nu}\right)$	$\varepsilon_\nu\left(1 + 2\frac{\delta_\nu}{\varrho_\nu}\right)$
$i=l$	0	$\varepsilon_\nu\left(1 + \frac{4\eta_\nu}{\varrho_\nu} + \frac{4\varepsilon_\nu\eta_\nu}{\varrho_\nu^2}\right)$	$--$
$l<i<n+1-l$	$\varepsilon_\nu\left(1 + 2\frac{\eta_\nu}{\varrho_\nu}\right)$	$\varepsilon_\nu\left(1 + 2\frac{\eta_\nu}{\varrho_\nu}\right)$	$--$
$i=n+1-l$	$\varepsilon_\nu\left(1 + 2\frac{\delta_\nu+\eta_\nu}{\varrho_\nu} + \frac{4\varepsilon_\nu^2}{\varrho_\nu^2}\right)$	$--$	$--$
$\substack{n+1-l \\ <i<n+1-k}$	$\varepsilon_\nu\left(1 + 2\frac{\eta_\nu}{\varrho_\nu}\right)$	$--$	$--$

Analogously, we obtain for the case $l > n+1-l$ that $|g_{ij}^{(\nu+1)}|$ and $|h_{ij}^{(\nu+1)}|$ have the following upper bounds:

	$j=k$	$j=n+1-l$	$j=l$
$i=k$	$\delta_\nu + 4\frac{\varepsilon_\nu\delta_\nu}{\varrho_\nu} + 4\frac{\varepsilon_\nu^3}{\varrho_\nu^2}$	$\delta_\nu + 2\frac{\varepsilon_\nu^2+\varepsilon_\nu\delta_\nu}{\varrho_\nu} + 4\frac{\varepsilon_\nu^2\delta_\nu}{\varrho_\nu^2}$	0
$k<i<n+1-l$	$\delta_\nu + 2\frac{\varepsilon_\nu^2}{\varrho_\nu}$	$\varepsilon_\nu\left(1 + 2\frac{\delta_\nu}{\varrho_\nu}\right)$	$\varepsilon_\nu\left(1 + 2\frac{\eta_\nu}{\varrho_\nu}\right)$
$i=n+1-l$	$\delta_\nu + 2\frac{\varepsilon_\nu^2+\varepsilon_\nu\delta_\nu}{\varrho_\nu} + 4\frac{\varepsilon_\nu^2\delta_\nu}{\varrho_\nu^2}$	$\varepsilon_\nu\left(1 + 4\frac{\delta_\nu}{\varrho_\nu} + 4\frac{\varepsilon_\nu\delta_\nu}{\varrho_\nu^2}\right)$	$--$
$n+1-l<i<l$	$\delta_\nu + 2\frac{\varepsilon_\nu^2}{\varrho_\nu}$	$\varepsilon_\nu\left(1 + 2\frac{\delta_\nu}{\varrho_\nu}\right)$	$--$
$i=l$	0	$--$	$--$
$l<i<n+1-k$	$\varepsilon_\nu\left(1 + 2\frac{\eta_\nu}{\varrho_\nu}\right)$	$--$	$--$

Moreover, we obtain for both cases $l < n+1-l$ and $l > n+1-l$ that

$$|g_{ij}^{(\nu+1)}|, |h_{ij}^{(\nu+1)}| < \delta_\nu\left(1 + 2\frac{\varepsilon_\nu}{\varrho_\nu}\right)$$

for $i < k$ and $j \le n+1-k$, and that

$$g^{(\nu+1)}_{n+1-k,k} = g^{(\nu)}_{n+1-k,k} + \Delta_1, \qquad g^{(\nu+1)}_{n+1-l,l} = g^{(\nu)}_{n+1-l,l} + \Delta_2,$$
$$h^{(\nu+1)}_{n+1-k,k} = h^{(\nu)}_{n+1-k,k} + \Delta_3, \qquad h^{(\nu+1)}_{n+1-l,l} = h^{(\nu)}_{n+1-l,l} + \Delta_4,$$

where

$$|\Delta_1|, |\Delta_2|, |\Delta_3|, |\Delta_4| < 2\left(\frac{\varepsilon_\nu^2}{\varrho_\nu} + 4\eta_\nu\frac{\varepsilon_\nu^2}{\varrho_\nu^2} + \eta_\nu\frac{\varepsilon_\nu}{\varrho_\nu}\right).$$

Thus, from the analysis of both Case 1 and 2 and using $\delta_\nu \le \varepsilon_\nu \le \eta_\nu$ and $4\varepsilon_\nu\eta_\nu/\varrho_\nu^2 < 1$, we find that

$$\delta_{\nu+1} = 0 \qquad \text{if } \nu \text{ is a multiple of } \frac{n^2}{4}, \tag{5.6}$$

$$\delta_{\nu+1} < \delta_\nu + 4\frac{\varepsilon_\nu^2}{\varrho_\nu} + 4\frac{\varepsilon_\nu^2\eta_\nu}{\varrho_\nu^2} \qquad \text{if } \nu \text{ is not a multiple of } \frac{n^2}{4}, \tag{5.7}$$

$$\varepsilon_{\nu+1} < \varepsilon_\nu\left(1 + 4\frac{\eta_\nu}{\varrho_\nu} + 4\frac{\varepsilon_\nu\eta_\nu}{\varrho_\nu^2}\right) < \varepsilon_\nu\left(2 + 4\frac{\eta_\nu}{\varrho_\nu}\right), \tag{5.8}$$

$$\varrho_{\nu+1} > \varrho_\nu - 4\left(\frac{\varepsilon_\nu^2}{\varrho_\nu} + 4\eta_\nu\frac{\varepsilon_\nu^2}{\varrho_\nu^2} + \eta_\nu\frac{\varepsilon_\nu}{\varrho_\nu}\right). \tag{5.9}$$

Using the above, we will now show that JIGH2 is locally convergent and that the asymptotic convergence rate is quadratic.

THEOREM 5.2. *Let $\lambda\mathcal{G} - \mathcal{H} \in \mathbb{C}^{n\times n}$ be a Hermitian pencil and for $\nu \in \mathbb{N}\cup\{0\}$, let δ_ν, ε_ν, η_ν, and ϱ_ν be defined as in* (5.2)–(5.5). *Moreover, let*

$$\eta := \max(\|\mathcal{G}\|_F, \|\mathcal{H}\|_F), \quad \varrho := \frac{\varrho_0}{2}, \quad \text{and} \quad \varepsilon := \varepsilon_0\left(2 + 4\frac{\eta}{\varrho}\right)^s, \tag{5.10}$$

where $s := \frac{n^2}{4}$. If $\varrho_0 > 0$ and if ε_0 is so small that

$$\frac{\varepsilon\eta}{\varrho^2} < \frac{1}{4}, \qquad \frac{\varepsilon^2}{\varrho} + 4\eta\frac{\varepsilon^2}{\varrho^2} + \eta\frac{\varepsilon}{\varrho} \le \frac{1}{4n^2}, \qquad \text{and} \quad 2n^2\varepsilon^2\left(\frac{1}{\varrho} + \frac{\eta}{\varrho^2}\right) \le \varepsilon_0, \tag{5.11}$$

then there exists a constant $C > 0$ such that $\varepsilon_{(\mu+1)s} < C\varepsilon_{\mu s}^2$ for all $\mu \in \mathbb{N}\cup\{0\}$.

Proof. From (5.8) and (5.9) we obtain that ε_ν (and ϱ_ν, respectively) may increase (or decrease, respectively) in each Jacobi step. We first show by induction that this increase (decrease, respectively) remains under control, i.e., that for $\mu \in \mathbb{N}\cup\{0\}$ and $p = 0, \dots, s$, we have that

$$\varepsilon_{\mu s} \le \frac{\varepsilon_0}{2^\mu}, \qquad \varrho_{\mu s} \ge \varrho_0 - \sum_{j=1}^{\mu}\frac{1}{2^{\mu+1}}\varrho_0 > \varrho, \quad \text{and} \tag{5.12}$$

$$\varepsilon_{\mu s+p} \le \varepsilon_{\mu s}\left(2 + 4\frac{\eta}{\varrho}\right)^p, \qquad \varrho_{\mu s+p} \ge \varrho_{\mu s} - p\frac{\varrho_0}{2^\mu n^2}. \tag{5.13}$$

$(\mu, p) = (0, 0)$: There is nothing to prove.

$(\mu,p) \Rightarrow (\mu,p+1)$: Let $p < s$. By the induction hypothesis for (μ,p) and $(\mu,0)$, we have that

$$\varepsilon_{\mu s+p} \le \varepsilon_{\mu s}\left(2+4\frac{\eta}{\varrho}\right)^p \le \frac{\varepsilon_0}{2^\mu}\left(2+4\frac{\eta}{\varrho}\right)^p \le \frac{\varepsilon}{2^\mu}$$

and
$$\varrho_{\mu s+p} \ge \varrho_{\mu s} - p\frac{\varrho_0}{2^\mu n^2} \ge \varrho_0 - \sum_{j=1}^{\mu}\frac{1}{2^{j+1}}\varrho_0 - \frac{n^2}{4}\frac{\varrho_0}{2^\mu n^2} = \varrho_0 - \sum_{j=1}^{\mu+1}\frac{1}{2^{j+1}}\varrho_0 > \varrho.$$

Then we obtain from (5.8) and (5.9) that

$$\begin{aligned}
\varepsilon_{\mu s+p+1} &< \varepsilon_{\mu s+p}\left(2+4\frac{\eta_{\mu s+p}}{\varrho_{\mu s+p}}\right) \le \varepsilon_{\mu s+p}\left(2+4\frac{\eta}{\varrho}\right) \le \varepsilon_{\mu s}\left(2+4\frac{\eta}{\varrho}\right)^{p+1};\\
\varrho_{\mu s+p+1} &> \varrho_{\mu s+p} - 4\left(\frac{\varepsilon^2_{\mu s+p}}{\varrho_{\mu s+p}} + 4\eta\frac{\varepsilon^2_{\mu s+p}}{\varrho^2_{\mu s+p}} + \eta\frac{\varepsilon_{\mu s+p}}{\varrho_{\mu s+p}}\right)\\
&\ge \varrho_{\mu s} - p\frac{\varrho_0}{2^\mu n^2} - 4\left(\frac{\varepsilon^2}{(2^\mu)^2\varrho} + 4\eta\frac{\varepsilon^2}{(2^\mu)^2\varrho^2} + \eta\frac{\varepsilon}{2^\mu\varrho}\right)\\
&\ge \varrho_{\mu s} - p\frac{\varrho_0}{2^\mu n^2} - \frac{4}{2^\mu}\left(\frac{\varepsilon^2}{\varrho} + 4\eta\frac{\varepsilon^2}{\varrho^2} + \eta\frac{\varepsilon}{\varrho}\right)\\
&\ge \varrho_{\mu s} - (p+1)\frac{\varrho_0}{2^\mu n^2} \qquad \Big(\text{by (5.11)}\Big).
\end{aligned}$$

$(\mu,p) \Rightarrow (\mu+1,0)$: For obtaining a bound for $\varepsilon_{(\mu+1)s}$, let us note that during the $(\mu+1)$st sweep each entry in the strict upper antitriangular part of the current pencil is set to zero at one step, and may then increase according to (5.7) during the rest of the sweep. Since $\varepsilon_{\mu s+p} \le \varepsilon/2^\mu$ and $\varrho_{\mu s+p} \ge \varrho$ for $p = 0,\dots,s$, we obtain by using (5.7) and (5.11) that

$$\varepsilon_{(\mu+1)s} < \frac{n^2}{4}\left(\frac{\varepsilon}{2^\mu}\right)^2\left(\frac{4}{\varrho}+4\frac{\eta}{\varrho^2}\right) \le \frac{1}{2}\frac{\varepsilon_0}{(2^\mu)^2} \le \frac{\varepsilon_0}{2^{\mu+1}}. \tag{5.14}$$

This concludes the proof of (5.12) and (5.13). In particular, (5.13) implies that

$$\varepsilon_{\mu s+p} \le \varepsilon_{\mu s}\left(2+4\frac{\eta}{\varrho}\right)^s.$$

Using this inequality instead of $\varepsilon_{\mu s+p} \le \varepsilon/2^\mu$, we obtain analogously to (5.14) that

$$\varepsilon_{(\mu+1)s} < n^2\varepsilon^2_{\mu s}(2+4\eta/\varrho)^{2s}\left(\frac{1}{\varrho}+\frac{\eta}{\varrho^2}\right) = C\varepsilon^2_{\mu s},$$

where C depends only on n, η, and ϱ. □

Unfortunately, JIGH2 does not converge globally as the following example shows:

$$\lambda\mathcal{G}-\mathcal{H} = \lambda\begin{bmatrix} 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\end{bmatrix} - \begin{bmatrix} 2 & 0 & 0 & i\\ 0 & 0 & 2i & 1\\ 0 & -2i & -4 & 0\\ -i & 1 & 0 & 1\end{bmatrix}.$$

This pencil has the spectrum $\{\pm\frac{1}{2}+\sqrt{7/4}i,\ \pm\frac{1}{2}-\sqrt{7/4}i\}$, and thus, it can be reduced to antitriangular form. However, JIGH2 stagnates, because the subpencil $\lambda\mathcal{G}^{(2)}_{11}-\mathcal{H}^{(2)}_{11}$ has real eigenvalues ±1.

6. A modified method (MJIGH2). The phenomenom of stagnation of JIGH2 that was noted in the previous section also occurs during numerical experiments. This problem can be solved by slightly modifying the algorithm JIGH2. Note that if during one sweep of JIGH2, we encounter a 2×2 subpencil with real eigenvalues (and thus, there is a chance of stagnation of JIGH2), then this 2×2 subpencil may be contained in a 4×4 subpencil that has (at least two) nonreal eigenvalues. In this case, performing Jacobi-steps as in JIGH4 may avoid stagnation of the method. This idea motivates the following modification of JIGH2 presented below that we will call MJIGH2 (modified Jacobi-like algorithm for the indefinite generalized Hermitian eigenvalue problem based on 2×2 subproblems).

Algorithm MJIGH2: Given a $2n \times 2n$ Hermitian pencil $\lambda\mathcal{G} - \mathcal{H}$ having no real eigenvalues, a stopping criterion, and strategies for the solution of 2×2 and 4×4 subproblems, the algorithm computes the antitriangular form of $\lambda\mathcal{G} - \mathcal{H}$.

while stopping criterion not satisfied
  for $k = 1, \dots, n$
    if $\lambda\mathcal{G}_{kk}^{(2)} - \mathcal{H}_{kk}^{(2)}$ is regular and has no real eigenvalues
      compute a unitary $Q \in \mathbb{C}^{2\times 2}$ such that $Q^*(\lambda\mathcal{G}_{kk}^{(2)} - \mathcal{H}_{kk}^{(2)})Q$ is in antitriangular form;
      set $\tilde{Q} = I_{2n}$ and $\tilde{Q}_{kk}^{(2)} := Q$;
      set $\mathcal{G} := \tilde{Q}^*\mathcal{G}\tilde{Q}$ and $\mathcal{H} := \tilde{Q}^*\mathcal{H}\tilde{Q}$;
      for $l = k+1, \dots, 2n-k$
        compute unitary $U, V \in \mathbb{C}^{2\times 2}$ such that $U(\lambda\mathcal{G}_{kl}^{(2)} - \mathcal{H}_{kl}^{(2)})V$ is in antitriangular form and set Q as in (4.2);
        set $\tilde{Q} := I_{2n}$ and $\tilde{Q}_{kl}^{(4)} := Q$;
        set $\mathcal{G} := \tilde{Q}^*\mathcal{G}\tilde{Q}$ and $\mathcal{H} := \tilde{Q}^*\mathcal{H}\tilde{Q}$;
      end
    else
      for $l = k+1, \dots, n$
        if $\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)}$ is singular or has real eigenvalues only
          $Q = I_4$;
        elseif $\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)}$ has no real eigenvalues
          compute a unitary Q such that $Q^*(\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)})Q$ is in antitriangular form;
        else
          compute a unitary Q such that $Q^*(\lambda\mathcal{G}_{kl}^{(4)} - \mathcal{H}_{kl}^{(4)})Q$ is in partial antitriangular form;
        end
        set $\tilde{Q} := I_{2n}$ and $\tilde{Q}_{kl}^{(4)} = Q$;
        set $\mathcal{G} := \tilde{Q}^*\mathcal{G}\tilde{Q}$ and $\mathcal{H} := \tilde{Q}^*\mathcal{H}\tilde{Q}$;
      end
    end
  end
end

Thus, whenever a Hermitian 2×2 subproblem having real eigenvalues occurs, we continue the second "for"-loop by solving 4×4 subproblems instead of 2×2 problems. As seen in section 5, no 2×2 Hermitian subproblems having only real

eigenvalues occur if $\lambda\mathcal{G} - \mathcal{H}$ is sufficiently close to being in antitriangular form, and thus, MJIGH2 reduces to JIGH2 in this case. This means that MJIGH2 has the same local convergence properties as JIGH2. In addition, convergence has been observed for all test examples in the current research.

7. Numerical experiments. For numerical tests, the algorithms JIGH4 and MJIGH2 were implemented in MATLAB Version 5.3 and run on a PC with a Pentium III processor (800 MHz). The relative machine precision was $\texttt{eps} = 2.2204 \times 10^{-16}$. The strategies used for the solution of the 4×4 and 2×2 subproblems were the ones explained in sections 3.2 and 4. Given a Hermitian pencil $\lambda\mathcal{G} - \mathcal{H} \in \mathbb{C}^{n\times n}$, $\mathcal{G} = (g_{ij})$, $\mathcal{H} = (h_{ij})$, with

$$\operatorname{norm}(\mathcal{G}, \mathcal{H}) := \sqrt{\|\mathcal{G}\|_F^2 + \|\mathcal{H}\|_F^2} = 1,$$

we chose as a stopping criterion

$$e(\mathcal{G}, \mathcal{H}) := \max\left\{ |g_{ij}|, |h_{ij}| \middle| i + j \leq n \right\} < 50\texttt{eps}. \tag{7.1}$$

It should be noted that this stopping criterion has been designed only for the numerical experiments performed in this paper. In practice, one should rather compare $|g_{ij}|, |h_{ij}|$ with $|g_{n+1-i,i}|$, $|h_{n+1-i,i}|$, $|g_{n+1-j,j}|$, $|h_{n+1-j,j}|$, $|g_{n+1-i,n+1-j}|$, and $|h_{n+1-i,n+1-j}|$, generalizing the componentwise error analysis for Jacobi's algorithm for positive definite matrices in [21]. This is a nontrivial problem that requires a detailed perturbation analysis. This topic is currently under investigation.

A first test was run for 50 randomly generated $2n \times 2n$ Hermitian pencils $\lambda\mathcal{G} - \mathcal{H}$ with $\operatorname{norm}(\mathcal{G}, \mathcal{H}) = 1$ having no eigenvalues with imaginary part of modulus smaller than 10^{-10}. Here, both $\mathcal{G}$ and $\mathcal{H}$ were constructed via the command

```
A=randn(2*n)+i*randn(2*n); A=A+A'.
```

Then, the eigenvalues of $\lambda\mathcal{G} - \mathcal{H}$ were computed. If the moduli of the imaginary parts of all eigenvalues were larger than or equal to 10^{-10}, then the pencil was normalized to one and used for the experiment.

The test was run over different dimensions $n = 3, 4, \ldots, 9, 10, 15, 20, 25, 30$. Convergence of both JIGH4 and MJIGH2 has been observed for all test problems. Figure 7.1 shows the average number of sweeps necessary for convergence. Note that except for the case $n = 3$, this number was always lower with MJIGH2 than with JIGH4.

Figure 7.2 shows the typical convergence behavior of both methods for the case $n = 10$, i.e., for a randomly generated 20×20 Hermitian pencil. At the beginning both methods show a similar behavior. In both cases, the convergence rate is not necessarily monotone, but, as soon as the pencil tends to being close to antitriangular form, the convergence rate of MJIGH2 becomes quadratic very fast, while the rate of JIGH4 is somewhere in between linear and quadratic, slowly approaching quadratic convergence towards the end of the process.

Thus, MJIGH2 converges faster than JIGH4. But, for a fair comparison, we also have to compare the cost per sweep of each method. Since an estimate of flop count is difficult to obtain (it depends on the nature of the eigenvalues of the 4×4 and 2×2 subpencils), the average flop count per sweep has been determined experimentally and is presented in the left graph of Figure 7.3 for both JIGH2 and JIGH4.

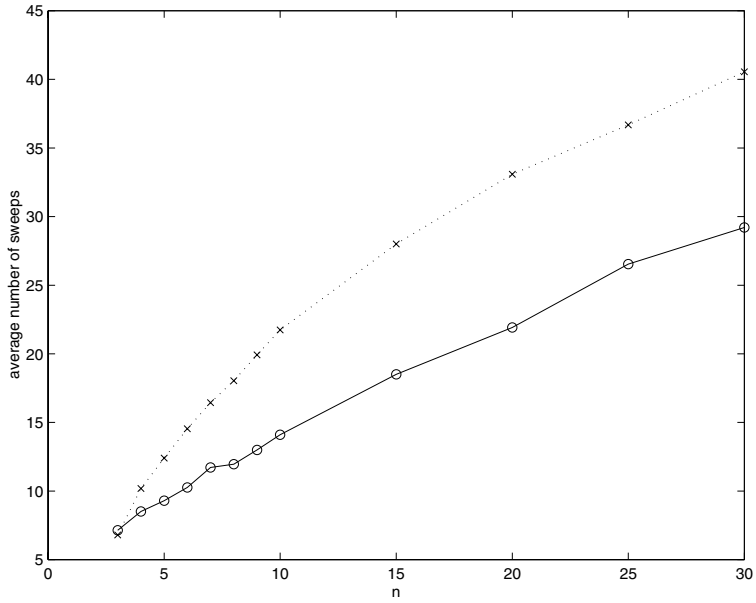


FIG. 7.1. 50 *randomly generated Hermitian pencils: MJIGH*2 (—), *JIGH*4 (···).

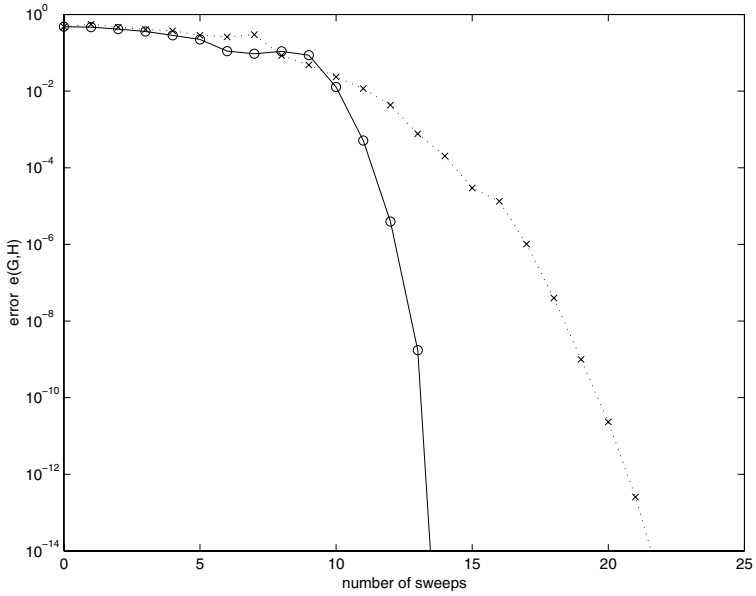


FIG. 7.2. *Typical convergence behavior: MJIGH*2 (—), *JIGH*4 (···).

The right graph of Figure 7.3 displays the average flop count per sweep for JIGH2 in percentage of the one for JIGH4. Clearly, the corresponding flop count for MJIGH2 is somewhere in between the ones for JIGH2 and JIGH4, due to the fact that some loops of MJIGH2 correspond to loops in JIGH4 if 2×2 Hermitian subpencils having real eigenvalues occur. However, the numerical experiments showed that this happens

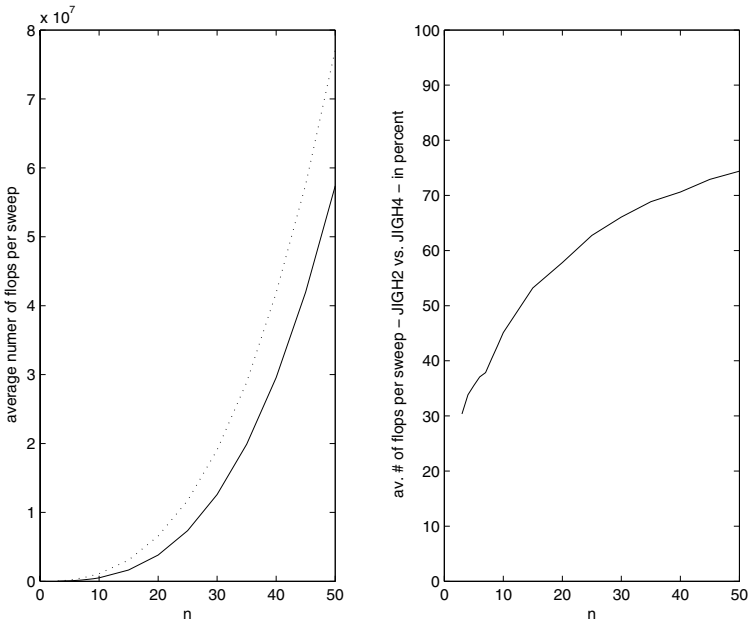


FIG. 7.3. *Flop count: JIGH*2 (—), *JIGH*4 ($\cdots$).

only a few times and only during the first few sweeps such that the average flop count per sweep for MJIGH2 is almost the same as the one for JIGH2.

In order to compare the performance of JIGH4 with that of the algorithm of Bunse-Gerstner and Faßbender for Hamiltonian matrices, a second test was run over 50 pencils of the form

$$\lambda\mathcal{G} - \mathcal{H} = \begin{bmatrix} Z_n & 0 \\ 0 & I_n \end{bmatrix} (i\lambda J_n - J_n H) \begin{bmatrix} Z_n & 0 \\ 0 & I_n \end{bmatrix}, \tag{7.2}$$

where Z_n is the $n \times n$ matrix with ones on the antidiagonal and zeros elsewhere and $H \in \mathbb{C}^{2n\times 2n}$ is a Hamiltonian matrix generated in the following way. First, we used the commands

```
A=randn(n)+i*randn(n);
B=randn(n)+i*randn(n);
C=randn(n)+i*randn(n);
H=[A B+B';C+C' -A'].
```

Then the eigenvalues of H were computed. If the moduli of the real parts of all eigenvalues were larger than or equal to 10^{-10}, the matrix was used for the experiment. (The same kind of matrix has been used for the numerical experiments in [2].) Finally, the pencil in (7.2) was normalized. Thus, Hermitian pencils of the form (7.2) can be interpreted as "Hamiltonian matrix"-pencils, and we expect JIGH4 to perform for them as the method of Bunse-Gerstner and Faßbender does for Hamiltonian matrices.

Figure 7.4 shows the average number of sweeps necessary for convergence for different dimensions $n = 3, 4, \ldots, 9, 10, 15, 20, 25, 30, 35, 40$. For the left graph, we used a stopping criterion corresponding to the one used in [2]; i.e., we chose the

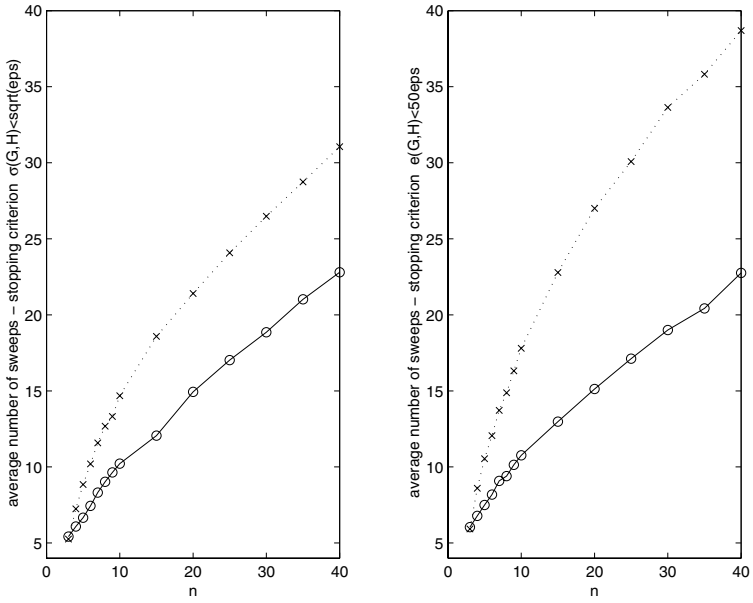


FIG. 7.4. 50 *random Hamiltonian matrix-pencils: MJIGH*2 (—), *JIGH*4 ($\cdots$).

criterion

$$\sigma(\mathcal{G},\mathcal{H}) = \sqrt{\sum_{i=1}^{2n-1}\sum_{j=1}^{i}(\|g_{j,i+1-j}\|^2 + \|h_{j,i+1-j}\|^2)} < \sqrt{\texttt{eps}}.$$

Indeed, the graph corresponding to the algorithm JIGH4 is quite similar to the graph obtained by Bunse-Gerstner and Faßbender for their experiments on Hamiltonian matrices. This shows that JIGH4 indeed performs in similar fashion as the algorithm by Bunse-Gerstner and Faßbender for Hamiltonian matrices, as expected.

On the other hand, the average number of sweeps was always lower with MJIGH2 than with JIGH4. This becomes even more evident in the right graph where we used the more rigid stopping criterion (7.1). Note that there is hardly any difference in the two graphs corresponding to MJIGH2. Once the stopping criterion (5.1) was satisfied with MJIGH2, then at most (if at all) one additional sweep was needed such that the stopping criterion (7.1) was also satisfied. This is due to the fact that the convergence rate of MJIGH2 was already quadratic when $\sigma(\mathcal{G},\mathcal{H}) \approx \sqrt{\texttt{eps}}$. In contrast, when the stopping criterion (5.1) was satisfied with JIGH4, then several additional sweeps (e.g., an average number of eight sweeps in the case $n = 40$) were needed until the stopping criterion (7.1) was also satisfied. Note that for both JIGH4 and MJIGH2 the average number of sweeps needed for convergence was always lower for Hamiltonian matrix pencils than for randomly generated Hermitian pencils. This is obviously caused by the rather simple structure of $\mathcal{G}$.

A third test was run over 50 Hermitian pencils close to being in antitriangular form. For this, we randomly generated Hermitian pencils $\lambda\mathcal{G} - \mathcal{H}$ having no real eigenvalues and with $\text{norm}(\lambda\mathcal{G} - \mathcal{H}) = 1$ as in the first test. Then, we used JIGH4 to reduce those pencils to antitriangular form and we randomly generated Hermitian pencils $\lambda\tilde{\mathcal{G}} - \tilde{\mathcal{H}}$ with $\text{norm}(\lambda\tilde{\mathcal{G}} - \tilde{\mathcal{H}}) = 0.01$. Then the test was run over 50 pencils of the

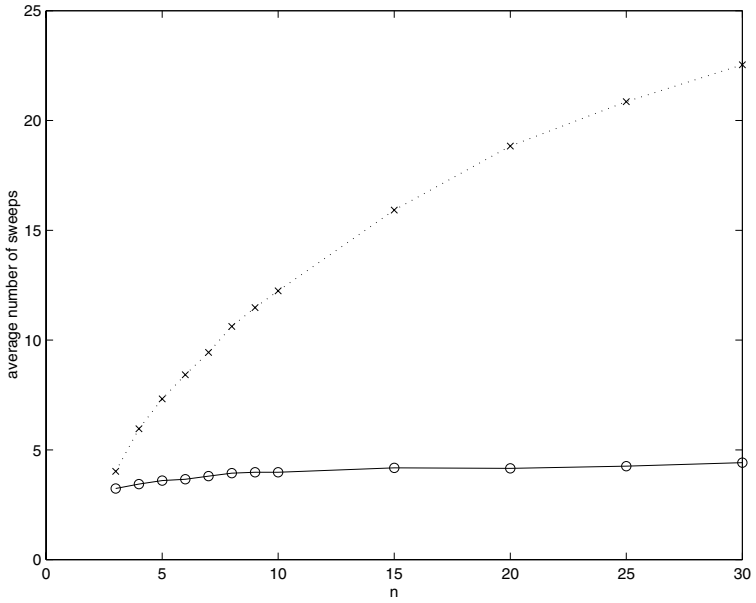


FIG. 7.5. 50 *Hermitian pencils close to antitriangular form: MJIGH*2 (—), *JIGH*4 ($\cdots$).

form $\lambda(\mathcal{G}+\tilde{\mathcal{G}})-(\mathcal{H}+\tilde{\mathcal{H}})$, where again we allowed only pencils having no eigenvalues with imaginary parts of moduli larger than or equal to 10^{-10}.

Figure 7.5 shows the average number of sweeps necessary for convergence. Note that this number increases drastically with JIGH4. For the dimension $n = 30$ an average of already more than 20 sweeps was necessary for convergence. Although this number is much lower than the corresponding one for randomly generated Hermitian pencils, it is much too high to be favorable, considering the fact that the pencil is already close to being in antitriangular form. On the other hand, the average number of sweeps necessary for convergence was almost constant between four and five with MJIGH2. Thus, the convergence rate of MJIGH2 was almost quadratic even from the beginning of the process. Hence, MJIGH2 exploits the special structure of Hermitian pencils that are close to antitriangular form much better than JIGH4 does.

8. Conclusions. We presented Jacobi-like methods for the solution of the indefinite generalized Hermitian eigenvalue problem. JIGH4 generalizes the method of Bunse-Gerstner and Faßbender for Hamiltonian matrices [2] while MJIGH2 is a slightly modified version of the locally quadratically convergent method JIGH2 that can be interpreted as a direct generalization of Eberlein's method [7] to the case of Hermitian pencils. Numerical experiments show that MJIGH2 is faster and less expensive than JIGH4. Moreover, MJIGH2 excels by the almost constant low number of sweeps needed for convergence if the Hermitian pencil under consideration is already close to being in antitriangular form. Hence, this algorithm may become attractive for the solution of such generalized Hermitian eigenvalue problems.

Several aspects have not been addressed in this paper. One one hand, we restricted our research to Hermitian pencils of even size, but we note that a generalization to the case of odd-sized pencils is possible. On the other hand, Bunse-Gerstner and Faßbender showed in [2] that their method allows parallel implementation. Clearly, this is

possible for JIGH4 as well, while for MJIGH2 further modifications and investigations are needed.

Acknowledgment. The author would like to thank V. Mehrmann, K. Veselić, and, in particular, H. Faßbender for many helpful discussions on this topic.

REFERENCES

[1] M. Borri and P. Mantegazza, *Efficient solution of quadratic eigenproblems arising in dynamic analysis of structures*, Comput. Methods Appl. Mech. Engrg., 12 (1977), pp. 19–31.
[2] A. Bunse-Gerstner and H. Faßbender, *A Jacobi-like method for solving algebraic Riccati equations on parallel computers*, IEEE Trans. Automat. Control, 42 (1997), pp. 1071–1084.
[3] R. Byers, *A Hamiltonian Jacobi algorithm*, IEEE Trans. Automat. Control, 35 (1990), pp. 566–570.
[4] J. P. Charlier and P. van Dooren, *A Jacobi-like algorithm for computing the generalized Schur form of a regular pencil*, J. Comput. Appl. Math., 27 (1989), pp. 17–36.
[5] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
[6] P. J. Eberlein, *A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 74–88.
[7] P. J. Eberlein, *On the Schur decomposition of a matrix for parallel computation*, IEEE Trans. Comput., 36 (1987), pp. 167–174.
[8] H. Faßbender, D. S. Mackey, and N. Mackey, *Hamiltonian and Jacobi come full circle: Jacobi algorithms for structured Hamiltonian eigenproblems*, Linear Algebra Appl., 332–334 (2001), pp. 37–80.
[9] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, Harcourt Brace Jovanovich, Publishers, New York, London, 1982.
[10] H. H. Goldstine and L. P. Horwitz, *A procedure for the diagonalization of normal matrices*, J. Assoc. Comput. Mach., 6 (1959), pp. 176–195.
[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.
[12] J.-S. Guo, W.-W. Lin, and C. S. Wang, *Numerical solutions for large sparse quadratic eigenvalue problems*, Linear Algebra Appl., 225 (1995), pp. 57–89.
[13] D. Hacon, *Jacobi's method for skew-symmetric matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 619–628.
[14] P. Henrici, *On the speed of convergence of cyclic and quasicyclic Jacobi methods for computing the eigenvalues of hermitian matrices*, J. Soc. Indust. Appl. Math, 6 (1958), pp. 144–162.
[15] K. Hüper, *A Calculus Approach to Matrix Eigenvalue Algorithms*, Habilitationsschrift, Universität Würzburg, Würzburg, Germany, 2002.
[16] C. G. J. Jacobi, *Über ein leichtes Verfahren, die in der Theorie der Säcularströmungen vorkommenden Gleichungen numerisch aufzulösen*, J. Reine Angew. Math., 30 (1846), pp. 51–95.
[17] H. P. M. van Kempen, *On the convergence of the classical Jacobi method for real symmetric matrices with non-distinct eigenvalues*, Numer. Math., 9 (1966), pp. 11–18.
[18] P. Lancaster, *Lambda-Matrices and Vibrating Systems*, International Series of Monographs in Pure and Applied Mathematics 94, Pergamon Press, Oxford, UK, 1966.
[19] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Clarendon Press, Oxford, UK, 1995.
[20] A. J. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, 24 (1979), pp. 913–921.
[21] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.
[22] C. Mehl, *Condensed forms for skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 454–476.
[23] C. Mehl, *Anti-triangular and anti-m-Hessenberg forms for Hermitian matrices and pencils*, Linear Algebra Appl., 317 (2000), pp. 143–176.
[24] V. Mehrmann, *Existence, uniqueness, and stability of solutions to singular linear quadratic optimal control problems*, Linear Algebra Appl., 121 (1989), pp. 291–331.
[25] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem, Theory, and Numerical Solution*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Heidelberg, Germany, 1991.

[26] M. H. C. Paardekooper, *An eigenvalue algorithm for skew-symmetric matrices*, Numer. Math., 17 (1971), pp. 189–202.
[27] A. Schönhage, *Zur quadratischen Konvergenz des Jacobi-Verfahrens*, Numer. Math., 6 (1964), pp. 410–412.
[28] G. W. Stewart, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.
[29] G. W. Stewart, *A Jacobi-like algorithm for computing the Schur decomposition of a non-Hermitian matrix*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 853–864.
[30] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[31] R. C. Thompson, *The characteristic polynomial of a principal subpencil of a Hermitian matrix pencil*, Linear Algebra Appl., 14 (1976), pp. 135–177.
[32] F. Tisseur, *Stability of structured Hamiltonian eigensolvers*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 103–125.
[33] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Review, 43 (2001), pp. 235–286.
[34] J. H. Wilkinson, *Note on the quadratic convergence of the cyclic Jacobi process*, Numer. Math, 4 (1962), pp. 296–300.
[35] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice–Hall, Upper Saddle River, NJ, 1995.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 986–994

A TOTALLY POSITIVE FACTORIZATION OF RECTANGULAR MATRICES BY THE NEVILLE ELIMINATION*

M. GASSÓ† AND JUAN R. TORREGROSA†

Abstract. An $n \times m$ real matrix A is a totally positive matrix if all its minors are nonnegative. The Neville elimination process is studied in relation to the existence of a totally positive factorization LS of a rectangular matrix. An LS factorization is obtained for a totally positive matrix, where L is a lower echelon form matrix, of size $n \times k$, and S is an upper echelon form matrix, of size $k \times m$, and both L and S are totally positive matrices.

Key words. totally positive matrix, Neville elimination, factorization LU

AMS subject classifications. 65F30, 15A23, 15A48

DOI. 10.1137/S0895479802415557

1. Introduction. An $n \times m$ real matrix A is called *totally positive* (*strictly totally positive*) if all its minors are nonnegative (positive). Totally positive (strictly totally positive) matrices will be referred to as TP (STP) matrices. When A is lower triangular (upper triangular) and we analyze the sign of its minors with indices of columns less than or equal to (greater than or equal to) indices of rows, we obtain the corresponding concepts ΔTP and ΔSTP. These matrices have become increasingly important in approximation theory, computer aided geometric design, and other fields [10]. For a comprehensive survey of this subject from an algebraic point of view, complete with historical references, see [1].

The existence of an LU factorization of a TP matrix A, with L and U TP matrices, had been studied by some authors. In [1] Ando proved that if A is a TP matrix, there exists a factorization $A = LU$ with L and U TP matrices. Cryer described in [2] a method to obtain the mentioned factorization when A is a square STP matrix, and proved in [3] that A is an $n \times n$ TP matrix if and only if A has an LU factorization such that L and U are TP.

Recently, the Neville elimination process has been used in order to obtain a totally positive LU factorization and QR factorization [7]. The essence of this elimination process is to produce zeros in a column of a matrix by adding to each row an appropriate multiple of the previous one. Eventual reorderings of the rows of the matrix may be necessary, as will be made clear in section 2.

Fiedler and Markham obtained, in [4] and [5], a factorization for nonsingular TP matrices that satisfy the properties CC (consecutive-column) and CR (consecutive-row), using the Neville elimination. From this process Gasca and Peña obtained, in [6] and [8], an LU factorization for matrices satisfying the WR (without row exchange) condition. These authors obtained a characterization of nonsingular TP matrices and STP matrices and proved that in these cases the upper echelon form matrix U obtained is also totally positive.

In this paper we are going to describe a variant of the Neville elimination process that we call "quasi-Neville," which allows us to obtain, for every rectangular TP

*Received by the editors October 2, 2002; accepted for publication (in revised form) by H. Woerdeman August 11, 2003; published electronically April 21, 2004. This research was supported by Spanish DGI grant BFM2001-0081-C03-02.

http://www.siam.org/journals/simax/25-4/41555.html

†Departamento de Matemática Aplicada, Universidad Politécnica de Valencia, Camino de Vera, 14, 46071 Valencia, Spain (mgasso@mat.upv.es, jrtorre@mat.upv.es).

matrix A of size $n \times m$, a factorization $A = LS$, where L is a lower echelon form matrix, of size $n \times p$, and S an upper echelon form matrix, of size $p \times m$, and both L and S are TP matrices.

2. Notation and preliminary results. In general, we shall use notation similar to that of [1]. Given $k, n \in N$, $k \leq n$, $Q_{k,n}$ will denote the totality of strictly increasing sequences of k natural numbers less than or equal to n:

$$\alpha = (\alpha_i)_{i=1}^{k} \in Q_{k,n} \quad \text{if } 1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_k \leq n.$$

Let A be an $n \times m$ real matrix, and let $\alpha \in Q_{k,n}$ and $\beta \in Q_{l,m}$, with $k \leq n$ and $l \leq m$. Let $A[\alpha|\beta]$ denote the $k \times l$ submatrix of A containing rows numbered by α and columns numbered by β.

We shall frequently use upper triangular $n \times m$ matrices S in *upper echelon form*, that is, $n \times m$ matrices $S = (s_{ij})$ such that

(1) if the kth row is zero ($k < n$), then the rows below it are zero;

(2) if s_{ij} is the first nonzero entry in the ith row, then $s_{hj} = 0$ $\forall h \geq i$, and if $s_{i'j'}$ is the first nonzero entry in the i'th row ($i < i' \leq n$), then $j' > j$.

On the other hand, we say that L is an $n \times m$ *lower echelon form matrix* if L^T is an upper echelon form matrix.

As we have mentioned in section 1, the Neville elimination is a procedure to create zeros in a matrix by means of adding to a given row any multiple of the previous one.

More precisely, we recall the Neville elimination process [6] for any $n \times m$ matrix $A = (a_{ij})$. Let $\bar{A}_1 := (\bar{a}^1_{ij})$ be such that $\bar{a}^1_{ij} = a_{ij}$, $1 \leq i \leq n$ and $1 \leq j \leq m$. If there are zeros in the first column of $\bar{A}_1$, we carry the corresponding rows down to the bottom in such a way that the relative order among them is the same as in $\bar{A}_1$. We denote the new matrix by A_1. If we have not carried any row down to the bottom, then $A_1 := \bar{A}_1$. In both cases, let $i_1 := 1$. The method consists of constructing a finite sequence of matrices A_k such that, for each A_k, the submatrix formed by its $k-1$ initial columns is an upper echelon form matrix. In fact, if $A_t = (a^t_{ij})$, then we introduce zeros in its tth column below the position (i_t, t), thus forming the $n \times m$ matrix

$$\bar{A}_{t+1} = \big(\bar{a}^{t+1}_{ij}\big),$$

where, for any j such that $1 \leq j \leq m$, we have

$$\begin{aligned}
&\bar{a}^{t+1}_{ij} := a^t_{ij}, && && i = 1, 2, \ldots, i_t,\\
&\bar{a}^{t+1}_{ij} := a^t_{ij} - \frac{a^t_{it}}{a^t_{i-1,t}} a^t_{i-1,t} && \text{if } a^t_{i-1,t} \neq 0, && i_t < i \leq n,\\
&\bar{a}^{t+1}_{ij} := a^t_{ij} && \text{if } a^t_{i-1,t} = 0, && i_t < i \leq n.
\end{aligned}$$

Observe that with our assumptions, $a^t_{i-1,t} = 0$ implies $a^t_{it} = 0$. Then we define

$$i_{t+1} := \begin{cases} i_t & \text{if } a^t_{i_t,t}\big(= \bar{a}^{t+1}_{i_t,t}\big) = 0,\\ i_t + 1 & \text{if } a^t_{i_t,t}\big(= \bar{a}^{t+1}_{i_t,t}\big) \neq 0. \end{cases}$$

If $\bar{A}_{t+1}$ has zeros in the $(t+1)$th column in the row i_{t+1} or below it, we will carry these rows down as we have done with $\bar{A}_1$. The matrix obtained in this way will be denoted by $A_{t+1} = (a^{t+1}_{ij})$. Of course, if there is no row that has been carried down, then $A_{t+1} := \bar{A}_{t+1}$. After a finite number of steps we get $\bar{A}_{\bar{t}-1}$, $A_{\bar{t}-1}$, and

$$\bar{A}_{\bar{t}} = U \quad (\bar{t} \leq m+1),$$

where U is an upper echelon form matrix. In this process the element

$$p_{ij} := a_{ij}^j, \quad 1 \le j \le m,\ i_j \le i \le n,$$

is called the (i,j) *pivot* of the Neville elimination of A, and the number

$$m_{ij} := \begin{cases} a_{ij}^j / a_{i-1,j}^j & \text{if } a_{i-1,j}^j \neq 0, \\ 0 & \text{if } a_{i-1,j}^j = 0, \end{cases} \qquad 1 \le j \le m,\ i_j < i \le n,$$

is the (i,j) *multiplier* of the Neville elimination of A. We observe that $m_{ij} = 0$ if and only if $a_{ij}^j = 0$.

When the Neville elimination of A can be carried out without row exchanges, $i_t = t\ \forall t$, and

$$A_{t+1} = \begin{bmatrix} a_{11}^1 & a_{12}^1 & \cdots & a_{1t}^1 & a_{1t+1}^1 & \cdots & a_{1m}^1 \\ 0 & a_{22}^2 & \cdots & a_{2t}^2 & a_{2t+1}^2 & \cdots & a_{2m}^2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & a_{tt}^t & a_{tt+1}^t & \cdots & a_{tm}^t \\ 0 & 0 & \cdots & 0 & a_{t+1t+1}^{t+1} & \cdots & a_{t+1m}^{t+1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & a_{nt+1}^{t+1} & \cdots & a_{nm}^{t+1} \end{bmatrix}.$$

The Neville elimination process, for an $n \times m$ matrix, can be matricially described by elementary and permutation matrices. We shall use notation similar to that in [8]. Let $E_{ij}(\alpha)$, $1 \le i \ne j \le n$, be the elementary triangular matrix whose (r,s) entry $1 \le r,\ s \le n$, is given by

$$\begin{cases} 1 & \text{if } r = s, \\ \alpha & \text{if } (r,s) = (i,j), \\ 0 & \text{elsewhere.} \end{cases}$$

We are interested in the matrices $E_{i+1,i}(\alpha)$, which for simplicity will be denoted by $E_{i+1}(\alpha)$.

For a matrix A satisfying the WR condition, the Neville elimination process can be written

$$\begin{aligned} &E_n(-m_{n,n-1}) \cdots \{E_3(-m_{32}) \cdots E_n(-m_{n,2})\} \\ &\quad \times \{E_2(-m_{21}) \cdots E_{n-1}(-m_{n-1,1}) E_n(-m_{n1})\} A = U, \end{aligned}$$

where U is an upper echelon form matrix of size $n \times m$.

When we apply the Neville elimination to a matrix A which does not satisfy the WR condition, it is necessary to use permutation matrices P_{ij}. We can see in [9] the matricial description of the Neville elimination process for matrices which does not satisfy the WR condition. In order to avoid using permutation matrices we introduce the following matrices which we use when the Neville elimination process produces a zero row.

DEFINITION 2.1. *The* reduced identity matrix *is the matrix obtained from the identity matrix by deleting one or more columns. We denote by* $I_n^{j_1,\ldots,j_k}$ *the matrix obtained from the* $n \times n$ *identity matrix by deleting the columns* $j_1, j_2, \ldots, j_k$.

It is easy to prove the following statements.

PROPOSITION 2.2. (a) *Every reduced identity matrix is a TP matrix.*

(b) *If A is an $n \times m$ matrix such that its rows $j_1, j_2, \ldots, j_k$ are zero rows, then $A = I_n^{j_1,\ldots,j_k} A_1$, where A_1 is the $(n-k) \times m$ matrix obtained from A by deleting its rows $j_1, j_2, \ldots, j_k$. Moreover, if A is a TP matrix, so is A_1.*

(c) *The left or right multiplication of a reduced identity matrix by elementary matrices E_i produces a lower echelon form matrix.*

3. Neville elimination and *TP* matrices.

THEOREM 3.1. *Let A be an $n \times m$ matrix satisfying the WR condition. A is a totally positive matrix if and only if A has an LU factorization by the Neville elimination with L and U TP matrices.*

Proof. If L and U are TP matrices, then $A = LU$ is a TP matrix (see [1]). Conversely, for a matrix A satisfying the WR condition, the Neville elimination process can be written

$$\begin{aligned} &E_n(-m_{n,n-1}) \cdots \{E_3(-m_{32}) \cdots E_n(-m_{n2})\} \\ &\quad \times \{E_2(-m_{21}) \cdots E_{n-1}(-m_{n-1,1}) E_n(-m_{n1})\} A = U, \end{aligned}$$

where U is an upper echelon form matrix of size $n \times m$. Equivalently, one has

$$F_{n-1} F_{n-2} \cdots F_1 A = U,$$

where F_i, $i = 1, 2, \ldots, n-1$, is the lower triangular matrix

$$F_i = \begin{bmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ & 0 & \ddots & & & & & & \\ & & \ddots & 1 & & & & & \\ & & & -m_{i+1,i} & 1 & & & & \\ & & & & -m_{i+2,i} & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & -m_{ni} & 1 \end{bmatrix}$$

$$= E_{i+1,i}(-m_{i+1,i}) \cdots E_{n,n-1}(-m_{ni}).$$

Then

$$A = F_1^{-1} F_2^{-1} \cdots F_{n-1}^{-1} U = LU,$$

where $F_i^{-1} = E_{n,n-1}(m_{ni}) \cdots E_{i+1,i}(m_{i+1,i})$. L is a TP matrix since it is a product of TP matrices. By applying Corollary 3.4 of [1] we can ensure that U is a TP matrix. ∎

From this theorem we can obtain the following results.

COROLLARY 3.2. *Let A be a TP matrix, of size $n \times m$, such that A or A^T satisfies the WR condition. Then A has an LU factorization, by the Neville elimination process, with L and U TP matrices.*

Example 3.3. The TP matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

does not satisfy the WR condition, but by applying the Neville elimination process to A^T we obtain

$$A^T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = LU.$$

Then $A = U^T L^T$, with U^T and L^T TP matrices.

COROLLARY 3.4. *Let A be an $n \times m$ matrix with* $\operatorname{rank}(A) = n$. *If A is a TP matrix, then A has an LU factorization, by the Neville elimination process, with L and U TP matrices.*

This result does not hold when the matrix A has full rank by columns, as we can see in the following example.

Example 3.5. Let A be the following TP matrix of size 6×5, with $\operatorname{rank}(A) = 5$:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 3 & 4 \\ 0 & 0 & 1 & 4 & 6 \end{bmatrix}.$$

The Neville elimination process allows us to obtain

$$E_{65}(-1/2)E_{43}(-1)E_{54}(-1)E_{65}(-1)A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0 & 1/2 \end{bmatrix}.$$

By using only elementary matrices $E_{i+1,i}(\alpha)$, we cannot obtain an upper echelon form matrix U. We observe that A does not satisfy the WR condition. In this case we will apply the quasi-Neville elimination process.

In the next section we are going to analyze the existence of a totally positive factorization for matrices which do not satisfy the WR condition.

4. Quasi-Neville elimination and *TP* matrices. As we have said in section 1, Ando proved in [1] that if A is a TP matrix, then there exists a factorization $A = LU$, with L and U TP matrices. We have observed that if we need permutation matrices in the Neville elimination process, the matrix L that this process provides is not, in general, a TP matrix.

Gasca and Peña in [6] introduced the condition N: *An $n \times m$ matrix A satisfies the condition N if, whenever we have carried some rows down to the bottom in the Neville elimination of A, those rows were zero rows, and the same condition is satisfied in the Neville elimination of U^T.* By this condition, it is possible to characterize TP matrices by the Neville elimination (see [6]).

THEOREM 4.1. *Let A be an $n \times m$ matrix. A is totally positive if and only if it satisfies the condition N and all the pivots are nonnegative.*

If reordering of the rows is necessary in the Neville elimination process of a TP matrix A, then, from the above theorem, we can ensure that the rows carried down to the bottom are zero rows. The *quasi-Neville elimination process* consists of leaving

the zero row in its position and continuing the elimination process with the matrix obtained from A by deleting the zero rows.

Example 4.2. Consider the following TP matrix of size 5×4:

$$A = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 3 & 2 & 3 & 3 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 4 & 4 \\ 1 & 1 & 5 & 6 \end{bmatrix}.$$

By applying, in this order, the elementary matrices $E_{54}(-1)$, $E_{43}(-1)$, $E_{32}(-1/3)$, and $E_{21}(-3/2)$, we obtain

$$\bar{A}_2 = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 0 & 1/2 & 3/2 & 3/2 \\ 0 & 1/3 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix},$$

and, by applying $E_{32}(-2/3)$, we obtain

$$\bar{A}_3 = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 0 & 1/2 & 3/2 & 3/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

Then

$$E_{32}(-2/3)E_{21}(-3/2)E_{32}(-1/3)E_{43}(-1)E_{54}(-1)A = \bar{A}_3.$$

We can observe that $\bar{A}_3 = I_5^3 A_3$, where

$$A_3 = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 0 & 1/2 & 3/2 & 3/2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

Finally, applying $E_{43}(-1/2)$ to matrix A_3 gives

$$S = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 0 & 1/2 & 3/2 & 3/2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Therefore, the matricial description of the quasi-Neville elimination process is

$$\begin{aligned} A &= E_{54}(1)E_{43}(1)E_{32}(1/3)E_{21}(3/2)E_{32}(2/3)\bar{A}_3 \\ &= E_{54}(1)E_{43}(1)E_{32}(1/3)E_{21}(3/2)E_{32}(2/3)I_5^3E_{43}(1/2)S \\ &= LS, \end{aligned}$$

where $L = E_{54}(1)E_{43}(1)E_{32}(1/3)E_{21}(3/2)E_{32}(2/3)I_5^3E_{43}(1/2)$ is a lower echelon form matrix. It is easy to see that L and S are TP matrices.

Let $A = (a_{ij})$ be a TP matrix of size $n \times m$. Let $\bar{A}_1 := (\bar{a}^1_{ij})$ be such that $\bar{a}^1_{ij} = a_{ij}$, $1 \leq i \leq n$ and $1 \leq j \leq m$. If there are zeros in the first column of $\bar{A}_1$, for example, in positions $(i_1, 1), (i_2, 1), \ldots, (i_k, 1)$, from Theorem 4.1 we can ensure that the rows $i_1, i_2, \ldots, i_k$ are zero rows. Let A_1 be the matrix, of size $(n-k) \times m$, obtained from $\bar{A}_1$ by deleting its rows $i_1, i_2, \ldots, i_k$. Then

$$A = \bar{A}_1 = I_n^{i_1, i_2, \ldots, i_k} A_1.$$

We observe that A_1 is a TP matrix. If there are no zeros in the first column of $\bar{A}_1$, then $A_1 := \bar{A}_1$.

We apply the Neville elimination process to create zeros in the first column of matrix A_1, and we obtain a new TP matrix $\bar{A}_2$ such that

$$\bar{A}_2 = E_{n-k}(-m_{n-k1}) \cdots E_2(-m_{21}) A_1,$$

where $E_i(\alpha)$ is the elementary matrix $E_{ii-1}(\alpha)$.

Now we apply this process to matrix $\bar{A}_2$. If there are zeros in the second column of $\bar{A}_2$, for example, in positions $(j_1, 2), (j_2, 2), \ldots, (j_l, 2)$, we obtain the TP matrix A_2 from $\bar{A}_2$ by deleting its rows $j_1, j_2, \ldots, j_l$. Then

$$\bar{A}_2 = I_{n-k}^{j_1, j_2, \ldots, j_l} A_2,$$

where A_2 has all nonzero rows. If there are no zeros in the second column of $\bar{A}_2$, then $A_2 := \bar{A}_2$. Therefore

$$A = I_n^{i_1, i_2, \ldots, i_k} E_2(m_{21}) \cdots E_{n-k}(m_{n-k1}) I_{n-k}^{j_1, j_2, \ldots, j_l} A_2.$$

After a finite number of steps we get $\bar{A}_{t-1}$, A_{t-1}, and $A_t = S$, where S is an upper echelon form TP matrix, of size $p \times m$. Then, by using the matrices F_i described in Theorem 3.1, we have

$$A = I_n^{i_1, i_2, \ldots, i_k} F_1^{-1} I_{n-k}^{j_1, j_2, \ldots, j_l} F_2^{-1} \cdots F_{m-1}^{-1} S = LS,$$

where the matrix $L = I_n^{i_1, i_2, \ldots, i_k} F_1^{-1} I_{n-k}^{j_1, j_2, \ldots, j_l} F_2^{-1} \cdots F_{m-1}^{-1}$ is a lower echelon form matrix product of TP matrices.

So we can establish the following results.

THEOREM 4.3. *Let A be a TP matrix of size $n \times m$. The quasi-Neville elimination process can be described as*

$$A = I_{n-1} F_{n-1} \cdots F_2 I_2 F_1 I_1 S,$$

where F_i, $i = 1, 2, \ldots, n-1$, is a lower triangular matrix product of elementary matrices; I_i, $i = 1, 2, \ldots, n-1$, is a reduced identity matrix; and S is an upper echelon form matrix.

THEOREM 4.4. *Let A be a matrix of size $n \times m$. A is a TP matrix if and only if $A = LS$, where L is a lower echelon form matrix, of size $n \times p$, and S is an upper echelon form matrix, of size $p \times m$, and both L and S are TP matrices.*

We can observe, following the construction of matrices L and S, that $\operatorname{rank}(A) = \operatorname{rank}(S) = p$. On the other hand, if A satisfies the WR condition, the decomposition LS coincides with the decomposition obtained in Theorem 3.1.

Example 4.5. Consider the TP matrix

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

Let A_1 be the matrix of size 4×4, obtained from A by deleting its second row. So $A = I_5^2 A_1$. Now, by applying in this order the elementary matrices $E_{43}(-1)$, $E_{32}(-1)$, and $E_{21}(-1)$ to matrix A_1, we obtain

$$\bar{A}_2 = E_{21}(-1)E_{32}(-1)E_{43}(-1)A_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Then $A = I_5^2 E_{43}(1)E_{32}(1)E_{21}(1)\bar{A}_2$.

Let A_2 be the matrix of size 3×4, obtained from $\bar{A}_2$ by deleting its third row. Then $\bar{A}_2 = I_4^3 A_2$. By applying the elementary matrix $E_{32}(-1/2)$ to matrix A_2, we obtain

$$\bar{A}_3 = E_{32}(-1/2)A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

So

$$A = I_5^2 E_{43}(1)E_{32}(1)E_{21}(1)I_4^3 E_{32}(1/2)S = LS,$$

where

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 3/2 & 1 \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and both L and S are TP matrices.

Acknowledgment. The authors are grateful to the referees for helpful comments.

REFERENCES

[1] T. Ando, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.
[2] C.W. Cryer, *The LU-factorization of totally positive matrices*, Linear Algebra Appl., 7 (1973), pp. 83–92.
[3] C.W. Cryer, *Some properties of totally positive matrices*, Linear Algebra Appl., 15 (1976), pp. 1–25.
[4] M. Fiedler and T.L. Markham, *Consecutive-column and -row properties of matrices and the Loewner-Neville factorization*, Linear Algebra Appl., 266 (1997), pp. 243–259.
[5] M. Fiedler and T.L. Markham, *A factorization of totally nonsingular matrices over a ring with identity*, Linear Algebra Appl., 304 (2000), pp. 161–171.

[6] M. Gasca and J.M. Peña, *Total positivity and Neville elimination*, Linear Algebra Appl., 165 (1992), pp. 25–44.
[7] M. Gasca and J.M. Peña, *Total positivity, QR factorization, and Neville elimination*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1132–1140.
[8] M. Gasca and J.M. Peña, *A matricial description of Neville elimination with applications to total positivity*, Linear Algebra Appl., 202 (1994), pp. 33–53.
[9] M. Gasso and J.R. Torregosa, *A PLU-factorization of rectangular matrices by the Neville elimination*, Linear Algebra Appl., 357 (2002), pp. 163–171.
[10] S. Karlin, *Total Positivity*, Stanford University Press, Stanford, CA, 1968.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 995–1020

ON INVERSE QUADRATIC EIGENVALUE PROBLEMS WITH PARTIALLY PRESCRIBED EIGENSTRUCTURE*

MOODY T. CHU†, YUEN-CHENG KUO‡, AND WEN-WEI LIN‡

Abstract. The inverse eigenvalue problem of constructing real and symmetric square matrices M, C, and K of size $n \times n$ for the quadratic pencil $Q(\lambda) = \lambda^2 M + \lambda C + K$ so that $Q(\lambda)$ has a prescribed subset of eigenvalues and eigenvectors is considered. This paper consists of two parts addressing two related but different problems.

The first part deals with the inverse problem where M and K are required to be positive definite and semidefinite, respectively. It is shown via construction that the inverse problem is solvable for any k, given complex conjugately closed pairs of distinct eigenvalues and linearly independent eigenvectors, provided $k \leq n$. The construction also allows additional optimization conditions to be built into the solution so as to better refine the approximate pencil. The eigenstructure of the resulting $Q(\lambda)$ is completely analyzed.

The second part deals with the inverse problem where M is a fixed positive definite matrix (and hence may be assumed to be the identity matrix I_n). It is shown via construction that the monic quadratic pencil $Q(\lambda) = \lambda^2 I_n + \lambda C + K$, with $n+1$ arbitrarily assigned complex conjugately closed pairs of distinct eigenvalues and column eigenvectors which span the space $\mathbb{C}^n$, always exists. Sufficient conditions under which this quadratic inverse eigenvalue problem is uniquely solvable are specified.

Key words. quadratic eigenvalue problem, inverse eigenvalue problem, partially prescribed spectrum, partial eigenstructure assignment

AMS subject classifications. 65F15, 15A22, 65H17, 93B55

DOI. 10.1137/S0895479803404484

1. Introduction.

Given $n \times n$ complex matrices M, C, and K, the task of finding scalars λ and nonzero vectors $\mathbf{x}$ satisfying

$$Q(\lambda)\mathbf{x} = 0, \tag{1.1}$$

where

$$Q(\lambda) := Q(\lambda; M, C, K) = \lambda^2 M + \lambda C + K, \tag{1.2}$$

is known as the *quadratic eigenvalue problem* (QEP). The scalars λ and the corresponding vectors $\mathbf{x}$ are called, respectively, eigenvalues and eigenvectors of the quadratic pencil $Q(\lambda)$. Together, $(\lambda, \mathbf{x})$ is called an *eigenpair* of $Q(\lambda)$. The QEP has received much attention because its formation has repeatedly arisen in many different disciplines, including applied mechanics, electrical oscillation, vibro-acoustics, fluid mechanics, and signal processing. In a recent treatise, Tisseur and Meerbergen [17] surveyed many applications, mathematical properties, and a variety of numerical techniques for the QEP. It is known that the QEP has $2n$ finite eigenvalues over the complex field, provided that the leading matrix coefficient M is nonsingular. The QEP arising in practice often entails some additional conditions on the matrices.

*Received by the editors May 6, 2003; accepted for publication (in revised form) by N. J. Higham August 18, 2003; published electronically June 4, 2004.

http://www.siam.org/journals/simax/25-4/40448.html

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (chu@math.ncsu.edu). The research of this author was supported in part by the National Science Foundation under grants DMS-9803759 and DMS-0073056.

‡Department of Mathematics, National Tsinghua University, Hsinchu, 300, Taiwan (d883207@oz.nthu.edu.tw, wwlin@am.nthu.edu.tw).

For example, if M, C, and K represent the mass, damping, and stiffness matrices, respectively, in a mass-spring system, then it is required that all matrices be real-valued and symmetric, and that M and K be positive definite and semidefinite, respectively. It is this class of constraints on the matrix coefficients in (1.2) that underlines our main contribution in this paper.

In mathematical modelling, we generally assume that there is a correspondence between the endogenous variables, that is, the internal parameters, and the exogenous variables, that is, the external behavior. In most of the applications involving (1.1), specifications of the underlying physical system are embedded in the matrix coefficients M, C, and K, while the resulting bearing of the system usually can be interpreted via its eigenvalues and eigenvectors. The process of analyzing and deriving the spectral information and, hence, inducing the dynamical behavior of a system from a priori known physical parameters such as mass, length, elasticity, inductance, and capacitance, is referred to as a *direct* problem. The *inverse* problem, in contrast, is to validate, determine, or estimate the parameters of the system according to its observed or expected behavior. The concern in the direct problem is to express the behavior in terms of the parameters, whereas in the inverse problem the concern is to express the parameters in terms of the behavior. The inverse problem is just as important as the direct problem in applications.

There has been a lot of interest in the inverse eigenvalue problem, including the notable pole assignment problem. Some general reviews and extensive bibliographies in this regard can be found, for example, in the first author's recent articles [3] and [4]. This paper concerns itself with the inverse problem of the QEP.

The term *inverse quadratic eigenvalue problem* (IQEP) adopted in the literature usually is for general matrix coefficients. In this paper we shall use it distinctively to stress the additional structure imposed upon the matrix coefficients. Two scenarios will be considered separately:

- Determine real, symmetric matrix coefficients M, C, and K with M positive definite and K positive semidefinite so that the resulting QEP has a prescribed set of k eigenpairs.
- Assume that the symmetric and positive definite leading matrix coefficient M is known and fixed. Then determine real and symmetric matrix coefficients C and K so that the resulting QEP has a prescribed set of k eigenpairs.

Other types of IQEPs have been studied under modified conditions. For instance, the IQEP studied by Ram and Elhay [13] is for symmetric tridiagonal coefficients, and, instead of prescribed eigenpairs, two sets of eigenvalues are given. In a series of articles, Starek and Inman [16] studied the IQEPs associated with nonproportional underdamped systems. Settings for some other mechanical applications can be found at the web site [14]. Our study in this paper stems from the speculation that the notion of the IQEP has the potential of leading to an important modification tool for model updating [5], model tuning, and model correction [1, 10, 15, 18], when compared with an analytical model. We will discuss this specific application in a separate paper.

We note that in several recent works, including those by Chu and Datta [2] and Nichols and Kautsky [12], as well as Datta, Elhay, Ram, and Sarkissian [6, 7], studies are undertaken toward a feedback design problem for a second-order control system. That consideration eventually leads to either a full or a partial eigenstructure assignment problem for the QEP. The proportional and derivative state feedback controller designated in these studies is capable of assigning specific eigenvalues and making the resulting system insensitive to perturbations. Nonetheless, these results cannot meet

the basic requirement that the quadratic pencil be symmetric.

In a large or complicated physical system, it is often impossible to obtain complete spectral information. Furthermore, quantities related to high frequency terms generally are susceptible to measurement errors due to the finite bandwidth of measuring devices. Spectral information, therefore, should not be used at its full extent. For these reasons, it might be more sensible to consider an IQEP where only a *portion* of the eigenvalues and eigenvectors is prescribed. A natural question to ask is how much eigeninformation is needed to ensure that an IQEP is solvable.

To facilitate the discussion, we shall describe the partial eigeninformation via the pair $(\Lambda, X) \in \mathbb{R}^{k\times k} \times \mathbb{R}^{n\times k}$ of matrices, where

$$\Lambda = \operatorname{diag}\{\lambda_1^{[2]}, \ldots, \lambda_\ell^{[2]}, \lambda_{2\ell+1}, \ldots, \lambda_k\} \tag{1.3}$$

with

$$\lambda_j^{[2]} = \begin{bmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{bmatrix} \in \mathbb{R}^{2\times 2}, \quad \beta_j \neq 0, \quad \text{for} \quad j = 1, \ldots, \ell, \tag{1.4}$$

and

$$X = [\mathbf{x}_{1R}, \mathbf{x}_{1I}, \ldots, \mathbf{x}_{\ell R}, \mathbf{x}_{\ell I}, \mathbf{x}_{2\ell+1}, \ldots, \mathbf{x}_k]. \tag{1.5}$$

The true eigenvalues and eigenvectors are readily identifiable via the transformation

$$R := \operatorname{diag}\left\{ \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}, \ldots, \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}, I_{k-2\ell} \right\}, \tag{1.6}$$

with $i = \sqrt{-1}$. That is, by defining

$$\tilde{\Lambda} = R^H \Lambda R = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_{2\ell-1}, \lambda_{2\ell}, \lambda_{2\ell+1}, \ldots, \lambda_k\} \in \mathbb{C}^{k\times k}, \tag{1.7}$$

$$\tilde{X} = XR = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{2\ell-1}, \mathbf{x}_{2\ell}, \mathbf{x}_{2\ell+1}, \ldots, \mathbf{x}_k] \in \mathbb{C}^{n\times k}, \tag{1.8}$$

respectively, the IQEP is concerned about finding a *real-valued* quadratic pencil $Q(\lambda)$ (with its matrix coefficients possessing a certain specified structure) so that $Q(\lambda_j)\mathbf{x}_j = 0$ for all $j = 1, \ldots, k$. The true (complex-valued) eigenvalues and eigenvectors of the desired quadratic pencil $Q(\lambda)$ can be induced from the pair (Λ, X) of real matrices. In this case, note that $\mathbf{x}_{2j-1} = \mathbf{x}_{jR} + i\mathbf{x}_{jI}$, $\mathbf{x}_{2j} = \mathbf{x}_{jR} - i\mathbf{x}_{jI}$, $\lambda_{2j-1} = \alpha_j + i\beta_j$, and $\lambda_{2j} = \alpha_j - i\beta_j$ for $j = 1, \ldots, \ell$, whereas $\mathbf{x}_j$ and λ_j are all real-valued for $j = 2\ell + 1, \ldots, k$. For convenience, we shall denote henceforth the set of diagonal elements of $\tilde{\Lambda}$, which is precisely the spectrum of Λ, by $\sigma(\Lambda)$. We shall call (Λ, X) an *eigeninformation pair* of the quadratic pencil $Q(\lambda)$.

The two types of IQEP considered in this paper can be formulated as follows.

Inverse standard quadratic eigenvalue problem (ISQEP). Given an eigeninformation pair (Λ, X), find real and symmetric matrices M, C, and K with M and K positive definite and semidefinite, respectively, so that the equation

$$MX\Lambda^2 + CX\Lambda + KX = 0 \tag{1.9}$$

is satisfied.

Inverse monic quadratic eigenvalue problem (IMQEP). Given an eigeninformation pair (Λ, X), find real and symmetric matrices C and K that satisfy the equation

$$X\Lambda^2 + CX\Lambda + KX = 0. \tag{1.10}$$

Before we move into further details, some remarks highlighting the fundamental differences between the two problems might help to capture the main points in the fairly involved mathematics later on.

1. In the IMQEP, it suffices to consider the monic quadratic pencil (1.10) for the more general case where the leading matrix coefficient M is positive definite and *fixed*. Since M is known, let $M = LL^\top$ denote the Cholesky decomposition of M. Then

$$Q(\lambda)\mathbf{x} = 0 \quad \Leftrightarrow \quad \widetilde{Q}(\lambda)(L^\top \mathbf{x}) = 0, \tag{1.11}$$

where

$$\widetilde{Q}(\lambda) := \lambda^2 I_n + \lambda L^{-1} C L^{-\top} + L^{-1} K L^{-\top}. \tag{1.12}$$

Thus, without loss of generality, we may assume that the given matrix M in the IMQEP is the $n \times n$ identity matrix I_n to begin with. It is not the case with the ISQEP. The leading matrix coefficient M in the ISQEP is part of the unknowns to be determined.

2. Note that the IMQEP requires only symmetry and nothing else of the two matrix coefficients C and K. The symmetry of C and K implies that there are in total $n(n+1)$ unknowns to be determined in the inverse problem. Since each eigenpair $(\lambda, \mathbf{x})$ characterizes a system of n equations, it is natural to conjecture that a monic quadratic pencil could be determined from any given $n+1$ eigenpairs that are closed under complex conjugation. One of our main contributions in this paper is to substantiate this conjecture after a necessary condition is satisfied. We offer a constructive proof in this paper showing that the solution for the IMQEP is in fact unique.
3. In contrast, the positive definiteness imposed on the ISQEP is much more complicated than a mere count of the numbers of the unknowns and equations. It turns out that the amount of eigeninformation cannot contain more than n eigenpairs. We show that, given any $k \le n$ distinct eigenvalues and linearly independent eigenvectors closed under complex conjugation, the ISQEP is always solvable, but the solution often is not unique. Furthermore, the remaining unspecified eigenstructure of the reconstructed quadratic pencil is in fact quite limited. In particular, at the upper end when $k = n$, that is, when the number of prescribed eigenpairs is equal to the dimension of the ambient space, every prescribed eigenvalue is a double eigenvalue and the remaining eigenstructure is completely fixed.
4. Though both problems are solved by constructive proofs, the mathematical techniques employed to derive the main results for the two problems are indispensably different. It appears counter to intuition that the IMQEP is much harder to analyze than the ISQEP.

It might be appropriate to attribute the first technique for solving the inverse problem of the QEP to a short exposition in the book [9, p. 173]. Unfortunately, the method derived from that discussion is not capable of producing symmetric C and K. Our contribution is innovative in four areas: First, we give a recipe for the construction of a solution to each of the two inverse problems. These recipes can be turned into numerical algorithms. Second, we specify necessary and sufficient conditions under which the IQEP is solvable. Third, we completely characterize the eigenstructure of the reconstructed quadratic pencil. Finally, we propose a way to refine the construction process so that the best approximation subject to some additional optimal conditions can be established.

2. Solving ISQEP. In this section we present a general theory elucidating how the ISQEP could be solved with the prescribed spectral information (Λ, X). Our proof is constructive. As a by-product, numerical algorithms can also be developed thence. Examples of numerical schemes and applications will be discussed in section 2.3. We shall assume henceforth, in the formulation of an ISQEP, that the given spectral information (Λ, X) is always in the form of (1.3) and (1.5).

2.1. Recipe of construction. Starting with the given pair of matrices (Λ, X), consider the null space $\mathcal{N}(\Omega)$ of the augmented matrix

$$\Omega := \begin{bmatrix} X^\top & \Lambda^\top X^\top \end{bmatrix} \in \mathbb{R}^{k\times 2n}.$$

Denote the dimension of $\mathcal{N}(\Omega)$ by m. If X has linearly independently columns (as we will assume later), then $m = 2n - k$. Note that $m \geq n$, if we have assumed $k \leq n$ (for the reason to be seen later) in the formulation of the ISQEP. Let the columns of the matrix

$$\begin{bmatrix} U^\top \\ V^\top \end{bmatrix} \in \mathbb{R}^{2n\times m}$$

with $U^\top, V^\top \in \mathbb{R}^{n\times m}$ denote *any* basis of the subspace $\mathcal{N}(\Omega)$. The equation

$$\begin{bmatrix} X^\top & \Lambda^\top X^\top \end{bmatrix} \begin{bmatrix} U^\top \\ V^\top \end{bmatrix} = 0 \tag{2.1}$$

holds. Define the quadratic pencil $Q(\lambda)$ by the matrix coefficients

$$M = V^\top V, \tag{2.2}$$

$$C = V^\top U + U^\top V, \tag{2.3}$$

$$K = U^\top U. \tag{2.4}$$

We claim that the above definitions are sufficient for constructing a solution to the ISQEP. The theory will be established in several steps.

THEOREM 2.1. *Given any pair of matrices* (Λ, X) *in the form of* (1.3) *and* (1.5)*, let* U *and* V *be an arbitrary solution to* (2.1)*. Then* (Λ, X) *is an eigenpair of the quadratic pencil* $Q(\lambda)$ *with matrix coefficients* M*,* C*, and* K *defined according to* (2.2)*,* (2.3)*, and* (2.4)*, respectively.*

Proof. Upon substitution, we see that

$$\begin{aligned} MX\Lambda^2 + CX\Lambda + KX &= V^\top VX\Lambda^2 + \left(V^\top U + U^\top V\right) X\Lambda + \left(U^\top U\right) X \\ &= V^\top \left(VX\Lambda + UX\right)\Lambda + U^\top \left(VX\Lambda + UX\right) = 0. \end{aligned}$$

The last equality is due to the properties of U and V in (2.1). ∎

By this construction, all matrix coefficients in $Q(\lambda)$ are obviously real and symmetric. Note also that both matrices M and K are positive semidefinite. However, it is not clear whether $Q(\lambda)$ is a trivial quadratic pencil. Toward that end, we claim that the assumption that X has full column rank is sufficient and necessary for the regularity of $Q(\lambda)$.

THEOREM 2.2. *The leading matrix coefficient* $M = V^\top V$ *is nonsingular, provided that* X *has full column rank. In this case, the resulting quadratic pencil* $Q(\lambda)$ *is regular; that is,* $\det(Q(\lambda))$ *is not identically zero.*

Proof. Suppose that $V^\top \in \mathbb{R}^{n\times m}$ is not of full row rank. There exists an orthogonal matrix $G \in \mathbb{R}^{m\times m}$ such that

$$V^\top G = \begin{bmatrix} V_1^\top & 0_{n\times m_2} \end{bmatrix},$$

where $V_1^\top \in \mathbb{R}^{n\times m_1}$ and $0_{n\times m_2}$ denotes the zero matrix of size $n \times m_2$. Note that $m_1 < n$ and $m_2 = m - m_1$. Postmultiply the same G to $U^\top$ and partition the product into

$$U^\top G = \begin{bmatrix} U_1^\top & U_2^\top \end{bmatrix},$$

where $U_1^\top \in \mathbb{R}^{n\times m_1}$ and $U_2^\top \in \mathbb{R}^{n\times m_2}$. Note that $m_2 > m - n$. On the other hand, we see from (2.1) that

$$X^\top U_2^\top = 0,$$

whereas the columns of $U_2^\top$ are necessarily linearly independent by construction. It follows that

$$n - k \ge m_2 > m - n,$$

which contradicts the fact that $m = 2n - k$. Thus, the matrix $V^\top$ must be of full row rank and then $M = V^\top V$ is nonsingular. ∎

THEOREM 2.3. *Suppose in a given pair of matrices (Λ, X) that all eigenvalues in Λ are distinct and that X is not of full column rank. Then the quadratic pencil $Q(\lambda)$ defined by* (2.2), (2.3), *and* (2.4) *is singular.*

Proof. It is easy to check that (2.1) remains true if Λ and X are replaced by $\tilde{\Lambda}$ and $\tilde{X}$ defined in (1.7) and (1.8), respectively. Let μ be an arbitrary complex number not in $\sigma(\Lambda)$. Observe that

$$\begin{bmatrix} \tilde{X}^\top & \tilde{\Lambda}^\top \tilde{X}^\top \end{bmatrix} \begin{bmatrix} I & -\mu I \\ 0 & I \end{bmatrix} \begin{bmatrix} I & \mu I \\ 0 & I \end{bmatrix} \begin{bmatrix} U^\top \\ V^\top \end{bmatrix} = 0.$$

It follows that

$$\begin{bmatrix} \tilde{X}^\top & (\tilde{\Lambda}^\top - \mu I)\tilde{X}^\top \end{bmatrix} \begin{bmatrix} \mu V^\top + U^\top \\ V^\top \end{bmatrix} = 0.$$

By assumption, $\tilde{X}$ is not of full column rank. We may therefore assume that for some $2 \le q \le k$,

$$\tilde{\mathbf{x}}_q = \sum_{j=1}^{q-1} r_j \tilde{\mathbf{x}}_j,$$

where not all r_j, $j = 1, \ldots, q-1$, are zero. Define

$$\Gamma := \begin{bmatrix} 1 & & & r_{1,q} & & 0 \\ & \ddots & & \vdots & & \\ & & \ddots & r_{q-1,q} & & \\ & & & 1 & & \\ & & & & \ddots & \\ & 0 & & & & 1 \end{bmatrix} \in \mathbb{C}^{k\times k},$$

with $r_{j,q} = -\frac{\lambda_q - \mu}{\lambda_j - \mu} r_j$, $j = 1, \ldots, q-1$. Clearly,

$$\Gamma^\top \begin{bmatrix} \tilde{X}^\top & (\tilde{\Lambda}^\top - \mu I)\tilde{X}^\top \end{bmatrix} \begin{bmatrix} \mu V^\top + U^\top \\ V^\top \end{bmatrix} = 0. \tag{2.5}$$

Notice that, by construction, the qth row of $\Gamma^\top(\tilde{\Lambda}^\top - \mu I)\tilde{X}^\top$ is zero. Let $y(\mu)^\top$ denote the qth row of $\Gamma^\top \tilde{X}^\top$, which cannot be identically zero because the spectrum of Λ has distinct elements. We thus see from (2.5) that

$$y(\mu)^\top \left(\mu V^\top + U^\top\right) = 0.$$

It follows that $y(\mu)^\top Q(\mu) = 0$. Since $\mu \in \mathbb{C}$ is arbitrary, $Q(\lambda)$ must be singular. ∎

We conclude this section with one important remark. The rank condition $k = n$ plays a pivotal role in ISQEP. It is the critical value for the regularity of the quadratic pencil $Q(\lambda)$ defined by the matrix coefficients (2.2), (2.3), and (2.4). In fact, it is clear now that corresponding to any given $\Lambda \in \mathbb{R}^{k\times k}$, $X \in \mathbb{R}^{n\times k}$ in the form of (1.3) and (1.5), the quadratic pencil $Q(\lambda)$ can always be factorized into the product

$$Q(\lambda) = \left(\lambda V^\top + U^\top\right)(\lambda V + U). \tag{2.6}$$

If $k > n$, then $\operatorname{rank}(\lambda V + U) \le 2n - k < n$; hence $\det(Q(\lambda)) \equiv 0$ for all λ. It is for this reason that we always assume that $k \le n$ in the formulation of an ISQEP.

2.2. Eigenstructure of $Q(\lambda)$. We have shown in the preceding section how to define the matrix coefficients so that the corresponding quadratic pencil possesses a prescribed set of k eigenvalues and eigenvectors. The ISQEP thereby is solved via construction. An interesting question to ask is how the unspecified eigenpair in the constructed pencil should look. In this section we examine the remaining eigenstructure of the quadratic pencil $Q(\lambda)$ created from our scheme.

THEOREM 2.4. *Let $(\Lambda, X) \in \mathbb{R}^{k\times k} \times \mathbb{R}^{n\times k}$ in the form of* (1.3) *and* (1.5) *denote the partial eigeninformation and $Q(\lambda)$ be the quadratic pencil defined by coefficients* (2.2), (2.3), *and* (2.4). *Assume that X has full column rank k.*

1. *If $k = n$, then $Q(\lambda)$ has double eigenvalue λ_j for each $\lambda_j \in \sigma(\Lambda)$.*
2. *If $k < n$, then $Q(\lambda)$ has double eigenvalue λ_j for each $\lambda_j \in \sigma(\Lambda)$. The remaining $2(n-k)$ eigenvalues of $Q(\lambda)$ are all complex conjugate with nonzero imaginary parts. In addition, if the matrices U and V in* (2.1) *are chosen from an orthogonal basis of the null space of Ω, then the remaining $2(n-k)$ eigenvalues are only $\pm i$ with corresponding eigenvectors $\mathbf{z} \pm i\mathbf{z}$, where $X^\top \mathbf{z} = 0$.*

Proof. The case $k = n$ is easy. The matrices $U^\top$ and $V^\top$ involved in (2.1) forming the null space of Ω are square matrices of size n. We also know from Theorem 2.2 that $V^\top$ is nonsingular. Observe that

$$V^{-1}U = -X\Lambda X^{-1}. \tag{2.7}$$

Using the factorization (2.6), we see that

$$\det(Q(\lambda)) = (\det(\lambda V + U))^2.$$

It is clear that $Q(\lambda)$ has double eigenvalue λ_j at every $\lambda_j \in \sigma(\Lambda)$.

We now consider the case when $k < n$. Since $X^\top \in \mathbb{R}^{k\times n}$ is of full row rank, there exists an orthogonal matrix $P_1 \in \mathbb{R}^{n\times n}$ such that

$$X^\top P_1^\top = \begin{bmatrix} X_{11}^\top & 0_{n\times(n-k)} \end{bmatrix}, \tag{2.8}$$

where $X_{11}^\top \in \mathbb{R}^{k\times k}$ is nonsingular. There also exists an orthogonal matrix $Q_1 \in \mathbb{R}^{m\times m}$ such that

$$P_1 V^\top Q_1 = \begin{bmatrix} V_{11}^\top & 0_{k\times(n-k)} & 0_{k\times(m-n)} \\ V_{21}^\top & \mathcal{A} & 0_{(n-k)\times(m-n)} \end{bmatrix} \in \mathbb{R}^{n\times m}, \tag{2.9}$$

with appropriate sizes for the other three submatrices. In particular, note that both $V_{11}^\top \in \mathbb{R}^{k\times k}$ and $\mathcal{A} \in \mathbb{R}^{(n-k)\times(n-k)}$ are nonsingular matrices because $V^\top$ is of full row rank by Theorem 2.2. From the fact that

$$\begin{bmatrix} X^\top & \Lambda^\top X^\top \end{bmatrix} \begin{bmatrix} P_1^\top & 0 \\ 0 & P_1^\top \end{bmatrix} \begin{bmatrix} P_1 & 0 \\ 0 & P_1 \end{bmatrix} \begin{bmatrix} U^\top \\ V^\top \end{bmatrix} Q_1 = 0, \tag{2.10}$$

we conclude that the structure of $P_1 U^\top Q_1$ must be of the form

$$P_1 U^\top Q_1 = \begin{bmatrix} U_{11}^\top & 0_{k\times(n-k)} & 0_{k\times(m-n)} \\ U_{21}^\top & \Delta & \mathcal{B} \end{bmatrix} \in \mathbb{R}^{n\times m}, \tag{2.11}$$

where $\mathcal{B} \in \mathbb{R}^{(n-k)\times(n-k)}$ is nonsingular. Because $\left[\begin{smallmatrix} U^\top \\ V^\top \end{smallmatrix}\right]$ is of full column rank, together with the fact that both $\mathcal{A}$ and $\mathcal{B}$ in (2.9) and (2.11) are nonsingular, it follows that $\left[\begin{smallmatrix} U_{11}^\top \\ V_{11}^\top \end{smallmatrix}\right]$ must be of full column rank. Note that $U_{11}^\top$ is nonsingular if and only if Λ has no zero eigenvalue. Using $V_{11}^\top$ as a pivot matrix to eliminate $V_{21}^\top$ in (2.9), we may claim that there exits a nonsingular matrix P_2 such that

$$\tilde{U}^\top := P_2 P_1 U^\top Q_1 = \begin{bmatrix} U_{11}^\top & 0 & 0 \\ \tilde{U}_{21}^\top & \Delta & \mathcal{B} \end{bmatrix},$$
$$\tilde{V}^\top := P_2 P_1 V^\top Q_1 = \begin{bmatrix} V_{11}^\top & 0 & 0 \\ 0 & \mathcal{A} & 0 \end{bmatrix}.$$

Compute the three matrices

$$\tilde{M} := \tilde{V}^\top \tilde{V} = \begin{bmatrix} V_{11}^\top V_{11} & 0 \\ 0 & \mathcal{A}\mathcal{A}^\top \end{bmatrix},$$
$$\tilde{C} := \tilde{U}^\top \tilde{V} + \tilde{V}^\top \tilde{U} = \begin{bmatrix} U_{11}^\top V_{11} + V_{11}^\top U_{11} & V_{11}^\top \tilde{U}_{21} \\ \tilde{U}_{21}^\top V_{11} & \mathcal{A}\Delta^\top + \Delta\mathcal{A}^\top \end{bmatrix},$$
$$\tilde{K} := \tilde{U}^\top \tilde{U} = \begin{bmatrix} U_{11}^\top U_{11} & U_{11}^\top \tilde{U}_{21} \\ \tilde{U}_{21}^\top U_{11} & \tilde{U}_{21}^\top \tilde{U}_{21} + \mathcal{B}\mathcal{B}^\top + \Delta\Delta^\top \end{bmatrix},$$

and define the quadratic pencil $\tilde{Q}(\lambda) := \lambda^2 \tilde{M} + \lambda \tilde{C} + \tilde{K}$. By construction, it is clear that $\tilde{Q}(\lambda) = (P_2 P_1) Q(\lambda) (P_2 P_1)^\top$. This congruence relation ensures that $\tilde{Q}(\lambda)$ preserves the same eigenvalue information as $Q(\lambda)$. Define

$$Q_{11}(\lambda) := \lambda^2 (V_{11}^\top V_{11}) + \lambda (V_{11}^\top U_{11} + U_{11}^\top V_{11}) + U_{11}^\top U_{11}, \tag{2.12}$$

$$P_3 := \begin{bmatrix} I & 0 \\ -\tilde{U}_{21}^\top (\lambda V_{11} + U_{11}) Q_{11}^{-1}(\lambda) & I \end{bmatrix}. \tag{2.13}$$

It is further seen that $\tilde{Q}(\lambda)$ can be factorized as

$$P_3 \begin{bmatrix} Q_{11}(\lambda) & 0 \\ 0 & (\lambda\mathcal{A} + \Delta)(\lambda\mathcal{A}^\top + \Delta^\top) + \mathcal{B}\mathcal{B}^\top \end{bmatrix} P_3^\top. \tag{2.14}$$

We thus have effectively decomposed the quadratic pencil $\tilde{Q}(\lambda)$ into two subpencils.

By construction, we see from (2.8), (2.10), and (2.12) that the quadratic subpencil $Q_{11}(\lambda)$ in (2.12) exactly solves the ISQEP with spectral data (Λ, X_{11}). For this problem, we have already proved in the first part that $Q_{11}(\lambda)$ must have double eigenvalue λ_j for each $\lambda_j \in \sigma(\Lambda)$. It remains only to check the eigenvalues for the subpencil $(\mu\mathcal{A}+\Delta)(\mu\mathcal{A}^\top+\Delta^\top)+\mathcal{B}\mathcal{B}^\top$. Recall that the matrix $\mathcal{B}$ in (2.11) is nonsingular. The matrix $(\mu\mathcal{A}+\Delta)(\mu\mathcal{A}^\top+\Delta^\top)+\mathcal{B}\mathcal{B}^\top$ is positive definite for every $\mu \in \mathbb{R}$. In particular, its determinant cannot be zero for any real μ. Therefore, the remaining eigenvalues of $Q(\lambda)$ all must be complex conjugate with nonzero imaginary parts.

If, in addition, the columns of $\left[\begin{smallmatrix} U^\top \\ V^\top \end{smallmatrix}\right]$ in (2.10) are orthogonal to begin with, then both $\mathcal{A}$ and $\mathcal{B}$ are $(n-k)\times(n-k)$ orthogonal matrices, and the submatrix Δ in (2.11) must be a zero matrix. By (2.14), the remaining eigenvalues of $Q(\lambda)$ can only be $\pm i$. Observe further that there exists a nonsingular $W \in \mathbb{R}^{k\times k}$ such that

$$\left[\begin{array}{cc} I & 0 \\ 0 & W \end{array}\right]\left[\begin{array}{cc} U & V \\ X^\top & \Lambda^\top X^\top \end{array}\right]\left[\begin{array}{cc} U^\top & X \\ V^\top & X\Lambda \end{array}\right]\left[\begin{array}{cc} I & 0 \\ 0 & W^\top \end{array}\right] = \left[\begin{array}{cc} I_{2n-k} & 0 \\ 0 & I_k \end{array}\right]. \tag{2.15}$$

It follows that

$$\begin{aligned} U^\top U + XW^\top W X^\top &= I_n, \\ U^\top V + XW^\top W\Lambda^\top X^\top &= 0, \\ V^\top U + X\Lambda W^\top W X^\top &= 0, \\ V^\top V + X\Lambda W^\top W\Lambda^\top X^\top &= I_n. \end{aligned} \tag{2.16}$$

For any $\mathbf{z}$ satisfying $X^\top\mathbf{z} = 0$, we see from the above equations that

$$\begin{aligned} U^\top U\mathbf{z} &= \mathbf{z}, \\ V^\top V\mathbf{z} &= \mathbf{z}, \\ U^\top V\mathbf{z} + V^\top U\mathbf{z} &= 0. \end{aligned}$$

This shows that $Q(\pm i)(\mathbf{z} \pm i\mathbf{z}) = 0$. ∎

Theorem 2.4 is significant on several fronts. First, if $k = n$, then *all* eigenvalues of $Q(\lambda)$ are counted. Second, if $k < n$ and if the basis of null space $\mathcal{N}(\Omega)$ is selected to be mutually orthogonal (as we normally would do by using, say, MATLAB), then again all eigenvalues of $Q(\lambda)$ are determined. In other words, we are *not* allowed to supplement any additional $n-k$ eigenpairs to simplify this ISQEP. The solution of our method for ISQEP is the most natural way for k $(< n)$ prescribed pairs of eigenvalues and eigenvectors. In section 2.3, we shall study how the nonorthogonal basis of $\mathcal{N}(\Omega)$ can help to improve the ISQEP approximation.

We can further calculate the geometric multiplicity of the double roots characterized in Theorem 2.5.

THEOREM 2.5. *Let* (Λ, X) *in the form of* (1.3) *and* (1.5) *denote the prescribed eigenpair of the quadratic pencil* $Q(\lambda)$ *defined earlier. Assume that* Λ *has distinct spectrum and* X *has full column rank. Then the following hold:*

1. *Each real-valued* $\lambda_j \in \sigma(\Lambda)$ *has an elementary divisor of degree* 2*; that is, the dimension of the null space* $\mathcal{N}(Q(\lambda_j))$ *is* 1.
2. *The dimension of* $\mathcal{N}(Q(\lambda_j))$ *of a complex-valued eigenvalue* $\lambda_j \in \sigma(\Lambda)$ *is generically* 1. *That is, pairs of matrices* (Λ, X) *of which a complex-valued eigenvalue has a linear elementary divisor form a measure zero set.*

Proof. Real-valued eigenvalues correspond to those $\lambda_j \in \sigma(\Lambda)$ with $j = 2\ell + 1, \ldots, k$. We have already seen in Theorem 2.1 that $Q(\lambda_j)\mathbf{x}_j = 0$, where $\mathbf{x}_j$ is the jth column of X. Suppose that the $\mathcal{N}(Q(\lambda_j))$ has dimension greater than 1. From (2.6), it must be that

$$\text{rank}\left(\lambda_j V^\top + U^\top\right) \le n - 2. \tag{2.17}$$

Rewrite (2.1) as

$$\begin{bmatrix} X^\top & \Lambda^\top X^\top \end{bmatrix} \begin{bmatrix} I & -\lambda_j I \\ 0 & I \end{bmatrix} \begin{bmatrix} I & \lambda_j I \\ 0 & I \end{bmatrix} \begin{bmatrix} U^\top \\ V^\top \end{bmatrix} = 0, \tag{2.18}$$

which is equivalent to

$$\begin{bmatrix} X^\top & (\Lambda^\top - \lambda_j I)X^\top \end{bmatrix} \begin{bmatrix} \lambda_j V^\top + U^\top \\ V^\top \end{bmatrix} = 0. \tag{2.19}$$

Note that, since Λ has distinct spectrum and $X^\top$ has full row rank,

$$\text{rank}\left((\Lambda^\top - \lambda_j I)X^\top\right) = k - 1$$

or, equivalently,

$$\dim\left(\mathcal{N}\left((\Lambda^\top - \lambda_j I)X^\top\right)\right) = n - k + 1. \tag{2.20}$$

On the other hand, there exists an orthogonal $G_j \in \mathbb{R}^{m\times m}$ such that

$$\begin{bmatrix} \lambda_j V^\top + U^\top \\ V^\top \end{bmatrix} G_j = \begin{bmatrix} U_{j1}^\top & 0 \\ V_{j1}^\top & V_{j2}^\top \end{bmatrix}, \tag{2.21}$$

where, due to (2.17), $V_{j2}^\top$ has at least $m - (n-2) = n - k + 2$ linearly independent columns. We then see from (2.19) that

$$(\Lambda^\top - \lambda_j I)X^\top V_{j2}^\top = 0,$$

implying that $\dim(\mathcal{N}((\Lambda^\top - \lambda_j I)X^\top)) \ge n - k + 2$. This contradicts (2.20).

To examine the complex-valued case, notice that (1.9) can be rewritten as

$$M(XR)(R^H\Lambda^2 R) + C(XR)(R^H\Lambda R) + KXR = 0,$$

where R is defined in (1.6). In particular, from (1.7) and (1.8), for $1 \le j \le 2\ell$, we have

$$Q(\lambda_j)\mathbf{x}_j = 0.$$

We first consider the case $k = n$. Two observations are due at the moment. First, the matrix V in the basis $\left[\begin{smallmatrix} U^\top \\ V^\top \end{smallmatrix}\right]$ for the null space $\mathcal{N}([X^\top, \Lambda^\top X^\top])$ can be an arbitrary nonsingular matrix. Second, if there exists another vector $\mathbf{z} \in \mathbb{C}^n$ independent of $\mathbf{x}_j$ such that $Q(\lambda_j)\mathbf{z} = 0$, we claim that for this kind of eigenvalue the matrix $(V^\top V)^{-1}$ must satisfy some kinds of algebraic varieties in $\mathbb{R}^{n\times n}$. Putting these two facts together, we conclude that any complex-valued eigenvalue having a linear elementary divisor must come from a set of measure zero.

To see the claim concerning the algebraic varieties for $(V^\top V)^{-1}$, we use (2.7) and (1.7) to rewrite $\lambda_j V + U$ as

$$\lambda_j V + U = VXR(\lambda_j I - \Lambda)R^H X^{-1},$$

and thus factorize $Q(\lambda_j)$ as

$$\begin{aligned} Q(\lambda_j) &= (\lambda_j V^\top + U^\top)(\lambda_j V + U) \\ &= X^{-\top}\bar{R}(\lambda_j I - \Lambda)R^\top X^\top V^\top VXR(\lambda_j I - \Lambda)R^H X^{-1}. \end{aligned} \tag{2.22}$$

If $Q(\lambda_j)\mathbf{z} = 0$, from (2.22) we have

$$R^\top X^\top V^\top VXR(\lambda_j I - \Lambda)R^H X^{-1}\mathbf{z} = \tau \mathbf{e}_j, \tag{2.23}$$

where $\mathbf{e}_j$ is the jth standard unit vector and τ is some scalar. Rewrite (2.23) as

$$(\lambda_j I - \Lambda)R^H X^{-1}\mathbf{z} = \tau R^H X^{-1}(V^\top V)^{-1}X^{-\top}\bar{R}\mathbf{e}_j.$$

Hence, a necessary condition for the existence of $\mathbf{z}$ is that $(V^\top V)^{-1}$ must satisfy the algebraic equation

$$\mathbf{e}_j^\top R^H X^{-1}(V^\top V)^{-1}X^{-\top}\bar{R}\mathbf{e}_j = 0. \tag{2.24}$$

We note in passing that the condition (2.24) for $(V^\top V)^{-1}$ is also sufficient since the above argument can be reversed to show the existence of a vector $\mathbf{z}$ in the null space of $Q(\lambda_j)$.

For the case $k < n$, a similar argument holds. Indeed, using the decompositions (2.12) and (2.14) given in Theorem 2.4, a sufficient and necessary condition for the existence of $\mathbf{z}$ is exactly the same as (2.24) where X and V are replaced by X_{11} and V_{11}, respectively. In either case, outside the algebraic variety, the elementary divisor of a generically prescribed complex eigenvalue therefore is of degree 2. ∎

To further demonstrate the subtlety of the second statement in Theorem 2.5, we make an interesting observation as follows.

COROLLARY 2.6. *Suppose in the given (Λ, X) that X has full column rank and that Λ has distinct spectrum. Assume further that $\{\pm i\} \subset \sigma(\Lambda)$. Construct the quadratic pencil $Q(\lambda)$ by taking an orthogonal basis $[U, V]^\top$ for the null space $\mathcal{N}(\Omega)$. Then the dimension of $\mathcal{N}(Q(\pm i))$ is 2. In other words, in this nongeneric case, both eigenvalues $\pm i$ have two linear elementary divisors.*

Proof. From (2.15), we have $W(X^\top X + \Lambda^\top X^\top X\Lambda)W^\top = I_n$. It follows that

$$W^\top W = (X^\top X + \Lambda^\top X^\top X\Lambda)^{-1}.$$

The last equation in (2.16) gives rise to

$$(V^\top V)^{-1} = (I - X\Lambda W^\top W\Lambda^\top X^\top)^{-1}.$$

Upon substitution, it holds that

$$\begin{aligned} X^{-1}(V^\top V)^{-1}X^{-\top} &= X^{-1}(I - X\Lambda W^\top W\Lambda^\top X^\top)^{-1}X^{-\top} \\ &= X^{-1}(I - X\Lambda(X^\top X + \Lambda^\top X^\top X\Lambda)^{-1}\Lambda^\top X^\top)^{-1}X^{-\top} \\ &= X^{-1}(I - X\Lambda X^{-1}(I + X^{-\top}\Lambda^\top X^\top X\Lambda X^{-1})^{-1}X^\top\Lambda^\top X^\top)^{-1}X^{-\top} \\ &= X^{-1}(I + X\Lambda X^{-1}X^{-\top}\Lambda^\top X^\top)X^{-\top} \\ &= X^{-1}X^{-\top} + \Lambda X^{-1}X^{-\top}\Lambda^\top, \end{aligned} \tag{2.25}$$

where the fourth equality is, after some algebraic manipulation, due to the Sherman–Morrison–Woodbury formula. Substituting (2.25) into (2.24) and assuming that j is the index that defines $\lambda_j = \pm i$, we find that

$$\begin{aligned}\mathbf{e}_j^\top R^H X^{-1}(V^\top V)^{-1}X^{-\top}\bar{R}\mathbf{e}_j &= \mathbf{e}_j^\top (R^H X^{-1}X^{-\top}\bar{R} + R^H \Lambda R R^H X^{-1}X^{-\top}\bar{R}R^\top \Lambda^\top \bar{R})\mathbf{e}_j \\ &= \mathbf{e}_j^\top (\tilde{X}^{-1}\tilde{X}^{-\top} + \tilde{\Lambda}\tilde{X}^{-1}\tilde{X}^{-\top}\tilde{\Lambda}^\top)\mathbf{e}_j = 0.\end{aligned}$$

The sufficient condition is met and, therefore, $\dim(\mathcal{N}(Q(\pm i))) = 2$. ∎

2.3. Numerical experiment. In this section we intend to highlight two main points by numerical examples. The first example demonstrates the eigenstructure of a solution to a typical ISQEP. From the discussion in the preceding sections, we already have a pretty good idea about how the eigenstructure should look. We now demonstrate numerically how the selection of U and V might affect the geometric multiplicity of the double eigenvalue. The second example has important meaning in applications. We demonstrate how some additional optimization constraints can be incorporated into the construction of a solution to ISQEP. These additional constraints are imposed by some logistic reasons with the hope of better approximating a real physical system. In this example, we also experiment with the effect of feeding various amounts of information on eigenvalues and eigenvectors to the construction. In particular, we compare the discrepancy between a given (analytic) quadratic pencil and the resulting ISQEP approximation by varying the values of k and the optimal constraints. All calculations are done by using MATLAB in its default (double) precision. For the ease of running text, however, we shall report all numerals in five digits only.

Example 1. Consider the ISQEP where the partial eigenstructure $(\Lambda, X) \in \mathbb{R}^{5\times 5} \times \mathbb{R}^{5\times 5}$ is randomly generated. Assume

$$X = \begin{bmatrix} -0.4132 & 5.2801 & 2.9437 & -6.6098 & -9.6715 \\ -4.3518 & 3.2758 & -5.1656 & 9.1024 & -9.1357 \\ -0.1336 & -4.0588 & 2.5321 & 3.3049 & -4.4715 \\ -5.1414 & 4.4003 & -2.2721 & 5.2872 & 6.9659 \\ 8.6146 & -4.0112 & -6.9380 & 1.4345 & -4.4708 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} -0.2168 & -4.3159 & 0 & 0 & 0 \\ 4.3159 & -0.2168 & 0 & 0 & 0 \\ 0 & 0 & 2.0675 & -0.9597 & 0 \\ 0 & 0 & 0.9597 & 2.0675 & 0 \\ 0 & 0 & 0 & 0 & -0.3064 \end{bmatrix}.$$

Choose a basis $\left[\begin{smallmatrix} U_1^\top \\ V_1^\top \end{smallmatrix}\right]$ for the null space $\mathcal{N}([X^\top \Lambda^\top X^\top])$, say,

$$U_1^\top = \begin{bmatrix} 0.26861 & 0.56448 & -0.08687 & 0.39491 & -0.24252 \\ 0.32690 & -0.24385 & 0.00804 & -0.32844 & 0.42471 \\ -0.33739 & 0.27725 & -0.15949 & -0.05883 & 0.58406 \\ -0.13374 & 0.43824 & 0.09638 & 0.28605 & 0.46936 \\ -0.42433 & 0.17867 & 0.69977 & -0.12829 & -0.16140 \end{bmatrix},$$

$$V_1^\top = \begin{bmatrix} 0.51817 & 0.09467 & 0.20341 & -0.04075 & 0.32693 \\ 0.25575 & 0.38674 & -0.09339 & -0.32830 & -0.22850 \\ 0.31749 & -0.02297 & 0.63841 & 0.01156 & 0.05987 \\ -0.02434 & -0.40196 & 0.09987 & 0.65755 & 0.09646 \\ 0.27184 & 0.02061 & -0.01859 & 0.30413 & -0.03669 \end{bmatrix},$$

and construct

$$Q_1(\lambda) = \lambda^2(V_1^\top V_1) + \lambda(V_1^\top U_1 + U_1^\top V_1) + (U_1^\top U_1).$$

This quadratic pencil has double eigenvalue λ_j for each $\lambda_j \in \sigma(\Lambda)$, according to our theory. Furthermore, we compute the singular values of each $Q(\lambda_j)$ and find that

$$\begin{aligned}
\mathrm{svd}(Q_1(-0.21683 \pm 4.3159i)) &= \{17.394, 15.039, 4.3974, 2.6136, 1.2483 \times 10^{-15}\},\\
\mathrm{svd}(Q_1(2.0675 \pm 0.95974i)) &= \{5.9380, 4.9789, 1.1788, 0.45926, 4.6449 \times 10^{-16}\},\\
\mathrm{svd}(Q_1(-0.30635)) &= \{1.0937, 1.0346, 0.89436, 0.18528, 3.8467 \times 10^{-17}\},
\end{aligned}$$

implying that the dimension of the null space $Q(\lambda_j)$ is precisely 1 for each $\lambda_j \in \sigma(\Lambda)$.

However, suppose we choose a special basis for $\mathcal{N}([X^\top \Lambda^\top X^\top])$ by

$$\begin{bmatrix} U_2^\top \\ V_2^\top \end{bmatrix} = \begin{bmatrix} U_1^\top V_1^{-\top} X^{-1} \\ X^{-1} \end{bmatrix}$$

and construct

$$Q_2(\lambda) = \lambda^2(V_2^\top V_2) + \lambda(V_2^\top U_2 + U_2^\top V_2) + (U_2^\top U_2).$$

We find that

$$\begin{aligned}
\mathrm{svd}(Q_2(-0.21683 \pm 4.3159i)) = \{&15.517,\ 0.12145,\ 0.07626,\ 3.4880 \times 10^{-15},\\
&7.9629 \times 10^{-16}\},\\
\mathrm{svd}(Q_2(2.0675 \pm 0.95974i)) = \{&21.064,\ 0.16325,\ 0.02540,\ 3.2321 \times 10^{-15},\\
&5.2233 \times 10^{-16}\},\\
\mathrm{svd}(Q_2(-0.30635)) = \{&20.995,\ 0.19733,\ 0.08264,\ 0.02977,\ 1.6927 \times 10^{-15}\}.
\end{aligned}$$

In this case, each of the four complex-valued eigenvalues of $\sigma(\Lambda)$ has linear elementary divisors.

Example 2. We can further exploit the freedom in the selection of the basis for the null space $\mathcal{N}(\Omega)$. In this example we first demonstrate a few ways to select the basis under some special circumstances. We then illustrate the effect of available eigeninformation on the construction.

To fix the idea, we first generate randomly a 10×10 symmetric quadratic pencil $\hat{Q}(\lambda) = \lambda^2\hat{M} + \lambda\hat{C} + \hat{K}$, where $\hat{M}$ and $\hat{K}$ are also positive definite, as an analytic model. We then compare the effect of k on its ISQEP approximations for $k = 1, \ldots, 10$. To save the space, we shall not report the data of these test matrices $\hat{M}$, $\hat{C}$, and $\hat{K}$ in this paper, but will make them available upon request. We merely report that the spectrum of $\hat{Q}(\lambda)$ turns out to be the following 10 pairs of complex-conjugate values:

$$\begin{aligned}
\{&-0.27589 \pm 1.8585i,\ -0.19201 \pm 1.5026i,\ -0.15147 \pm 1.0972i,\ -0.11832 \pm 0.54054i,\\
&-0.07890 \pm 1.3399i,\ -0.07785 \pm 0.76383i,\ -0.07716 \pm 0.86045i,\ -0.07254 \pm 1.1576i,\\
&-0.06276 \pm 0.97722i,\ -0.05868 \pm 0.18925i\}.
\end{aligned}$$

These eigenvalues are not arranged in any specific order. Without loss of generality, we shall *pretend* that the first five pairs in the above list are the partially described eigenvalues and that we wish to reconstruct the quadratic pencil. For $\ell = 1, \ldots, 5$ (and

hence $k = 2\ell$), denote these eigenvalues as $\alpha_\ell \pm i\beta_\ell$. Also, define partial eigenpairs $(\Lambda_{2\ell}, X_{2\ell})$ of $\hat{Q}(\lambda)$ according to (1.3) and (1.5); that is,

$$\Lambda_{2l} = \text{diag}\left\{ \begin{bmatrix} \alpha_1 & \beta_1 \\ -\beta_1 & \alpha_1 \end{bmatrix}, \ldots, \begin{bmatrix} \alpha_\ell & \beta_\ell \\ -\beta_\ell & \alpha_\ell \end{bmatrix} \right\}, \tag{2.26}$$

$$X_{2\ell} = [x_{1R}, x_{1I}, \ldots, x_{\ell R}, x_{\ell I}], \tag{2.27}$$

where $x_{\ell R} \pm i x_{\ell I}$ is the eigenvector of $\hat{Q}(\lambda)$ corresponding to $\alpha_\ell \pm i\beta_\ell$.

Let $\begin{bmatrix} U_\ell^\top \\ V_\ell^\top \end{bmatrix} \in \mathbb{R}^{2n\times(2n-2\ell)}$ be an orthogonal basis for $\mathcal{N}([X_{2\ell}^\top \Lambda_{2\ell}^\top X_{2\ell}^\top])$. We now introduce three ways to select a *new* basis for $\mathcal{N}([X_{2\ell}^\top \Lambda_{2\ell}^\top X_{2\ell}^\top])$, each of which is done for a different optimization purpose. The physical meaning of these optimal constraints will be explained at the end of this section.

Case 1. Suppose $\hat{K} = L_{\hat{K}} L_{\hat{K}}^\top$ and $\hat{M} = L_{\hat{M}} L_{\hat{M}}^\top$ are the Cholesky factorizations of $\hat{K}$ and $\hat{M}$ in the model pencil, respectively. Find a matrix $G_{\ell 1}^\top \in \mathbb{R}^{(2n-2\ell)\times(2n-2\ell)}$ by solving the sequence of least-square problems

$$\min \left\| \begin{bmatrix} U_\ell^\top \\ V_\ell^\top \end{bmatrix} G_{\ell 1}^\top(:, j) - \begin{bmatrix} L_{\hat{K}} & 0_{n-2\ell} \\ 0_{n-2\ell} & L_{\hat{M}} \end{bmatrix} (:, j) \right\|_2 \tag{2.28}$$

for each of its columns $G_{\ell 1}^\top(:, j)$, $j = 1, \ldots, 2n-2\ell$. For convenience, we have adopted here the MATLAB notation $(:, j)$ to denote the jth column of a matrix.

The solution of (2.28) is intended to not only solve the ISQEP, but also best approximate the original $\hat{K}$ and $\hat{M}$ in the sense that the quantity

$$\|U_\ell^\top G_{\ell 1}^\top G_{\ell 1} U_\ell - \hat{K}\|_F + \|V_\ell^\top G_{\ell 1}^\top G_{\ell 1} V_\ell - \hat{M}\|_F \tag{2.29}$$

is minimized among all possible $G_{\ell 1}^\top \in \mathbb{R}^{(2n-2\ell)\times(2n-2\ell)}$. Once such a matrix $G_{\ell 1}^\top$ is found, we compute the coefficient matrices according to our recipe, that is,

$$\begin{aligned} &M_{\ell 1} = V_\ell^\top G_{\ell 1}^\top G_{\ell 1} V_\ell, \quad K_{\ell 1} = U_\ell^\top G_{\ell 1}^\top G_{\ell 1} U_\ell, \\ &C_{\ell 1} = U_\ell^\top G_{\ell 1}^\top G_{\ell 1} V_\ell + V_\ell^\top G_{\ell 1}^\top G_{\ell 1} U_\ell, \end{aligned} \tag{2.30}$$

and define the quadratic pencil

$$Q_{\ell 1}(\lambda) = \lambda^2 M_{\ell 1} + \lambda C_{\ell 1} + K_{\ell 1}, \tag{2.31}$$

according to $\ell = 1, \ldots, 5$.

Case 2. We first transform $V_\ell^\top$ to $[V_{\ell 0}^\top, 0]$ by an orthogonal transformation. Then we find a matrix $G_{\ell 2}^\top \in \mathbb{R}^{(2n-2\ell)\times(2n-2\ell)}$ in the form

$$G_{\ell 2}^\top = \begin{bmatrix} E_{\ell 2}^\top & 0 \\ 0 & F_{\ell 2}^\top \end{bmatrix}, \tag{2.32}$$

where $E_{\ell 2}^\top = V_{\ell 0}^{-\top} L_{\hat{M}}$ and $F_{\ell 2}^\top$ is an arbitrary $(n-2\ell)\times(n-2\ell)$ orthogonal matrix.

Case 3. We transform $U_\ell^\top$ to $[U_{\ell 0}^\top, 0]$ by an orthogonal transformation. Then we find a matrix $G_{\ell 3}^\top \in \mathbb{R}^{(2n-2\ell)\times(2n-2\ell)}$ in the form

$$G_{\ell 3}^\top = \begin{bmatrix} E_{\ell 3}^\top & 0 \\ 0 & F_{\ell 3}^\top \end{bmatrix}, \tag{2.33}$$

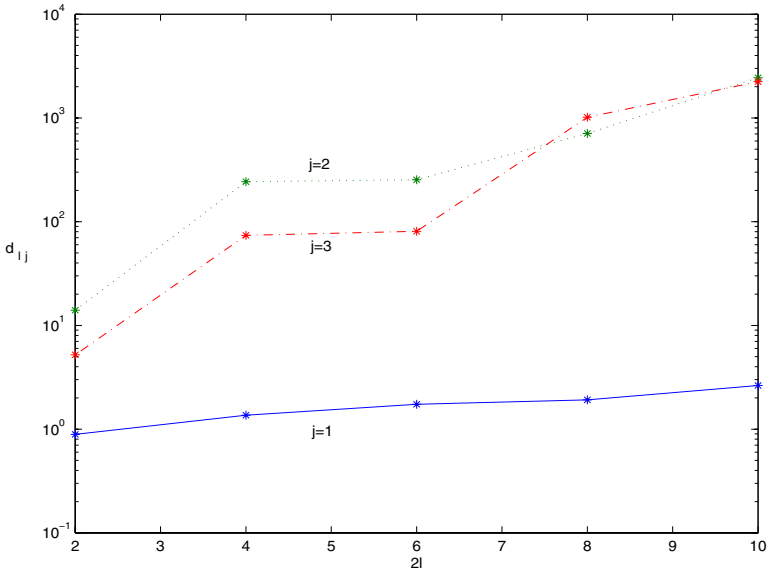


FIG. 2.1. *Errors of ISQEP approximations.*

where $E_{\ell 3}^\top = U_{\ell 0}^{-\top} L_{\hat{K}}$ and $F_{\ell 3}^\top$ is an arbitrary $(n-2\ell)\times(n-2\ell)$ orthogonal matrix.

The purpose of finding $G_{\ell 2}^\top$ and $G_{\ell 3}^\top$ in the form of (2.32) and (2.33) is to not only solve the ISQEP, but also best approximate the original $\hat{M}$ and $\hat{K}$, respectively, in the sense that

$$\|V_\ell^\top G_{\ell 2}^\top G_{\ell 2} V_\ell - \hat{M}\|_F \tag{2.34}$$

and

$$\|U_\ell^\top G_{\ell 3}^\top G_{\ell 3} U_\ell - \hat{K}\|_F \tag{2.35}$$

are minimized by $G_{\ell 2}^\top$ and $G_{\ell 3}^\top$, respectively. Once these matrices are found, we define quadratic pencils $Q_{\ell 2}(\lambda)$ and $Q_{\ell 3}(\lambda)$ in exactly the same way as we define $Q_{\ell 1}(\lambda)$.

It would be interesting to see how the reconstructed quadratic pencils for the ISQEP, with the above-mentioned optimization in mind, approximate the original pencil. Toward that end, we measure the total difference

$$d_{\ell j} = \|M_{\ell j} - \hat{M}\|_F + \|C_{\ell j} - \hat{C}\|_F + \|K_{\ell j} - \hat{K}\|_F \tag{2.36}$$

between the original pencil and the reconstructed pencil for each $j = 1, 2, 3$ and $\ell = 1, \ldots, 5$.

In Figure 2.1 we plot the *error* $d_{\ell j}$ between $\hat{Q}(\lambda)$ and $Q_{\ell j}(\lambda)$ for the various cases. Not surprisingly, we notice that the quadratic pencil $Q_{\ell 1}(\lambda)$ constructed from $G_{\ell 1}^\top$ is superior to the other two. What might be interesting to note is that in Case 1 the amount of eigeninformation available to the ISQEP does not seem to make any significance difference in the measurement of $d_{\ell 1}$. That is, all $d_{\ell 1}$ seem to be of the same order regardless of the value of ℓ. We think a reason for this is because $G_{\ell 1}^\top$ has somewhat more freedom to choose so that $M_{\ell 1}$ and $K_{\ell 1}$ better approximate $\hat{M}$ and $\hat{K}$, respectively.

In real application for vibrating systems, the stiffness matrix $\hat{K}$ and the mass matrix $\hat{M}$ of a mathematical model can usually be obtained by a finite element or finite difference method. It is the damping matrix $\hat{C}$ in such a system that is generally not known. If some partial eigenstructure can be measured by experiment, then the construction proposed in Case 1 might be a good way to recover the original system by best approximating the stiffness matrix and the mass matrix in the sense of minimizing (2.29).

3. Solving IMQEP. With the existence theory established in the preceding section for the ISQEP where $k = n$ plays a vital role in deciding whether the resulting quadratic pencil is singular, it is interesting to study in this section yet another scenario of the IQEP.

In the IMQEP, the leading matrix coefficient M is known and fixed and only symmetric C and K are to be determined. We have already suggested earlier by counting the cardinality of unknowns and equations that the number of prescribed eigenpairs could go up to $k = n + 1$. Since the prescribed eigenvectors now form a matrix X of size $n \times (n+1)$, by assuming that X is of full rank, there is at least one column in the given $n \times (n+1)$ matrix X depending linearly on the other columns. The following analysis is contingent on whether this linearly dependent column is real-valued or complex-valued. We separate the discussion into two cases. Either case shows a way to solve the IMQEP.

3.1. Real linearly dependent eigenvector. Assume that the linearly dependent column vector is real-valued. By rearranging the columns if necessary, we may assume without loss of generality that this vector is $\mathbf{x}_{n+1}$. It follows that the $n \times n$ submatrix

$$X_1 := [\mathbf{x}_1, \bar{\mathbf{x}}_1, \ldots, \mathbf{x}_{2\ell-1}, \bar{\mathbf{x}}_{2\ell-1}, \mathbf{x}_{2\ell+1}, \ldots, \mathbf{x}_n] \tag{3.1}$$

of $\tilde{X}$ defined in (1.8) is nonsingular. Let

$$\Lambda_1 := \mathrm{diag}\{\lambda_1, \bar{\lambda}_1, \ldots, \lambda_{2\ell-1}, \bar{\lambda}_{2\ell-1}, \lambda_{2\ell+1}, \ldots, \lambda_n\} \tag{3.2}$$

be the corresponding submatrix of $\tilde{\Lambda}$ defined in (1.7). Both matrices are closed under complex conjugation in the sense defined before.

Define

$$S := X_1 \Lambda_1 X_1^{-1}. \tag{3.3}$$

Note that, due to the complex conjugation, S is a real-valued $n \times n$ matrix. Define a quadratic pencil $Q(\lambda)$ via the factorization

$$Q(\lambda) := (\lambda I_n + S + C)(\lambda I_n - S), \tag{3.4}$$

where C is yet to be determined. Upon comparing the expression of (3.4) with (1.10), we see that

$$K = -(S + C)S. \tag{3.5}$$

The first criterion for solving the IMQEP is that both matrices C and K be real-valued and symmetric. Thus the undetermined real-valued matrix C must first satisfy the following two equations simultaneously:

$$\begin{cases} C^\top = C, \\ S^\top C - CS = S^2 - (S^\top)^2. \end{cases} \tag{3.6}$$

The following result provides a partial characterization of the matrix C we are looking for.

THEOREM 3.1. *The general solution to* (3.6) *is given by the formula*

$$C = -(S + S^\top) + \sum_{j=1}^{n} \gamma_j \mathbf{y}_j \mathbf{y}_j^\top, \tag{3.7}$$

where vectors $\mathbf{y}_j$, $j = 1, \ldots n$, *are the columns of the matrix*

$$Y_1 := X_1^{-\top} = [\mathbf{y}_1, \ldots, \mathbf{y}_{2\ell-1}, \mathbf{y}_{2\ell}, \mathbf{y}_{2\ell+1}, \ldots, \mathbf{y}_n], \tag{3.8}$$

and the scalars γ_j, $j = 1, \ldots, n$, *are arbitrary complex numbers.*

Proof. It is easy to see that $-(S + S^\top)$ is a particular solution of (3.6). The formula thus follows from an established result [11, section 12.5, Theorem 1]. ∎

It might be worth mentioning that the columns of Y_1 are also closed under complex conjugation and, hence, C is real-valued if and only if the corresponding coefficients γ_j are complex conjugate. It remains only to determine these combination coefficients in (3.7) so that the IMQEP is solved. Toward that end, observe first that

$$X_1 \Lambda_1^2 + C X_1 \Lambda_1 + K X_1 = 0, \tag{3.9}$$

regardless of how the scalars γ_j, $j = 1, \ldots, n$, are chosen. In other words, n pairs of the given data have already satisfied the spectral constraint in the IMQEP. We use the fact that the last pair $(\lambda_{n+1}, \mathbf{x}_{n+1}) \in \mathbb{R} \times \mathbb{R}^n$ in the given data must also be an eigenpair of $Q(\lambda)$ in (3.4) to determine the parameters γ_j, $j = 1, \ldots, n$.

Define

$$\mathbf{z} := (\lambda_{n+1} I - S)\mathbf{x}_{n+1} \in \mathbb{R}^n. \tag{3.10}$$

Plugging the eigenpair $(\lambda_{n+1}, \mathbf{x}_{n+1})$ into (3.4) and using (3.7), we obtain the equation

$$\lambda_{n+1}\mathbf{z} = S^\top \mathbf{z} - \sum_{j=1}^{n} \gamma_j \mathbf{y}_j \mathbf{y}_j^\top \mathbf{z},$$

which can be written as

$$-X_1^\top(\lambda_{n+1}\mathbf{z} - S^\top \mathbf{z}) = \operatorname{diag}\{\mathbf{y}_1^\top \mathbf{z}, \ldots, \mathbf{y}_n^\top \mathbf{z}\} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix}. \tag{3.11}$$

Obviously, values of $\gamma_j \mathbf{y}_j^\top \mathbf{z}$, $j = 1, \ldots, n$, are uniquely determined. However, the value of γ_j is unique only if

$$\mathbf{y}_j^\top \mathbf{z} = \mathbf{e}_j^\top X_1^{-1} \mathbf{z} \neq 0, \tag{3.12}$$

where $\mathbf{e}_j$ denotes the jth standard unit vector. In terms of the original data, the condition can be written equivalently as

$$\mathbf{e}_j^\top (\lambda_{n+1} I - \Lambda_1) X_1^{-1} \mathbf{x}_{n+1} \neq 0. \tag{3.13}$$

If we assume that the condition (3.12) holds for all $j = 1, \ldots, n$, then the last step in solving the IMQEP is to show that elements in the solution $\{\gamma_1, \ldots, \gamma_n\}$ to (3.1) are closed under complex conjugation in exactly the same way as columns of Y_1 are.

For convenience, we shall denote

$$\mathbf{r} := \lambda_{n+1}\mathbf{z} - S^\top \mathbf{z} \in \mathbb{R}^n. \tag{3.14}$$

The first 2ℓ elements in (3.11) are

$$-\mathbf{x}_{2j-1}^\top \mathbf{r} = \mathbf{y}_{2j-1}^\top \mathbf{z}\gamma_{2j-1}, \tag{3.15}$$

$$-\mathbf{x}_{2j}^\top \mathbf{r} = \mathbf{y}_{2j}^\top \mathbf{z}\gamma_{2j} \quad \text{for } j = 1, \ldots, \ell. \tag{3.16}$$

Recall $\mathbf{x}_{2j-1} = \bar{\mathbf{x}}_{2j}$ and $\mathbf{y}_{2j-1} = \bar{\mathbf{y}}_{2j}$ for $j = 1, \ldots, \ell$. Upon taking the conjugation of (3.15) and comparing with (3.16), we conclude that

$$\gamma_{2j} = \bar{\gamma}_{2j-1} \quad \text{for } j = 1, \ldots, \ell. \tag{3.17}$$

Similarly, $\gamma_k \in \mathbb{R}$ for $k = 2\ell+1, \ldots, n$. It is now finally proved that both C and K are indeed real-valued and symmetric. We summarize our first major result as follows.

THEOREM 3.2. *Let $(\tilde{\Lambda}, \tilde{X}) \in \mathbb{C}^{(n+1)\times(n+1)} \times \mathbb{C}^{n\times(n+1)}$ be given as in* (1.7) *and* (1.8). *Assume that one eigenvector, say, $\mathbf{x}_{n+1} \in \mathbb{R}^n$, depends linearly on the remaining eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which are linearly independent. If the condition* (3.13) *is satisfied for all $j = 1, \ldots, n$, then the IMQEP has a unique solution.*

We point out quickly that (3.11) is not necessarily consistent. In particular, a possible scenario is as follows.

COROLLARY 3.3. *Under the same assumptions as in Theorem* 3.2, *if $\mathbf{e}_j^\top X_1^{-1}\mathbf{z} = 0$ and $\mathbf{e}_j^\top(\lambda_{n+1}I - \Lambda_1^\top)X_1^\top \mathbf{z} \neq 0$ for some j, then the IMQEP has no solution.*

3.2. Complex linearly dependent eigenvector. Assume that the linearly dependent column vector is complex-valued. By rearranging the columns if necessary, we may assume without loss of generality that this vector is $\mathbf{x}_{2\ell}$. It follows that the $n \times n$ matrix

$$X_1 = [\underbrace{\mathbf{x}_1, \bar{\mathbf{x}}_1, \ldots, \mathbf{x}_{2\ell-3}, \bar{\mathbf{x}}_{2\ell-3}}_{\textit{complex-conjugated}}, \underbrace{\mathbf{x}_{2\ell+1}, \ldots, \mathbf{x}_{n+1}}_{\textit{real-valued}}, \quad \underbrace{\mathbf{x}_{2\ell-1}}_{\textit{complex-valued}} \quad]$$

is nonsingular. For convenience, we shall reindex the sequence of the above column vectors by successive integers. Without causing ambiguity, we shall use the same notation for the renumbered vectors. Specifically, we rewrite the above X_1 as

$$X_1 = [\underbrace{\mathbf{x}_1, \bar{\mathbf{x}}_1, \ldots, \mathbf{x}_{2m-1}, \bar{\mathbf{x}}_{2m-1}}_{\textit{complex-conjugated}}, \underbrace{\mathbf{x}_{2m+1}, \ldots, \mathbf{x}_{n-1}}_{\textit{real-valued}}, \quad \underbrace{\mathbf{x}_n}_{\textit{complex-valued}} \quad] \tag{3.18}$$

column by column but rename only the indices, and define the corresponding

$$\Lambda_1 = \text{diag}\{\lambda_1, \bar{\lambda}_1, \ldots, \lambda_{2m-1}, \bar{\lambda}_{2m-1}, \lambda_{2m+1}, \ldots, \lambda_{n-1}, \lambda_n\}. \tag{3.19}$$

We could further assume in (3.18) that

$$\bar{\mathbf{x}}_n \in \text{span}\{\mathbf{x}_1, \bar{\mathbf{x}}_1, \ldots, \mathbf{x}_{2m-1}, \bar{\mathbf{x}}_{2m-1}, \mathbf{x}_n\}, \tag{3.20}$$

since otherwise one of the real-valued eigenvectors (and this is possible only if $2m+1 < n$) must be linearly dependent and we would go back to the case in section 3.1. The following argument is analogous to that of section 3.1, but additional details need to be filled in.

Following (3.3) through (3.5) except that S is now complex-valued, we want to determine the matrix C in the factorization (3.4) and the corresponding K via several steps. We first require both C and K to be Hermitian. That is, the matrix C must satisfy the following equations:

$$\begin{cases} C^H = C \in \mathbb{C}^{n\times n}, \\ S^H C - CS = S^2 - (S^H)^2 \in \mathbb{C}^{n\times n}. \end{cases} \tag{3.21}$$

In contrast to Theorem 3.1, the characterization of C is a little bit more complicated.

THEOREM 3.4. *The general solution to* (3.21) *is given by the formula*

$$\begin{aligned} C = &-(S+S^H) + \gamma_1 \mathbf{y}_1\mathbf{y}_2^H + \gamma_2\mathbf{y}_2\mathbf{y}_1^H + \cdots + \gamma_{2m-1}\mathbf{y}_{2m-1}\mathbf{y}_{2m}^H + \gamma_{2m}\mathbf{y}_{2m}\mathbf{y}_{2m-1}^H \\ &+ \gamma_{2m+1}\mathbf{y}_{2m+1}\mathbf{y}_{2m+1}^H + \cdots + \gamma_{n-1}\mathbf{y}_{n-1}\mathbf{y}_{n-1}^H, \end{aligned} \tag{3.22}$$

where vectors $\mathbf{y}_i$, $i = 1, \ldots, n$, *are the columns of the matrix*

$$Y_1 := X_1^{-H} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{2m}, \mathbf{y}_{2m+1}, \ldots, \mathbf{y}_n]. \tag{3.23}$$

Proof. Again, the formula is similar to that in Theorem 3.1 using exactly the same established result [11, section 12.5, Theorem 1]. The slight complication is due to the fact that the first $2m$ eigenvalues of S and S^H coincide in a conjugated way and the last eigenvalues of S and S^H are distinct. ∎

Note that for $j = 1, \ldots, m$, $\mathbf{y}_{2j-1}$, $\mathbf{y}_{2j} = \bar{\mathbf{y}}_{2j-1}$ are the eigenvectors of S^H with eigenvalues to $\bar{\lambda}_{2j-1}$ and λ_{2j-1}, respectively. Likewise, for $k = 2m+1, \ldots, n-1$, $\mathbf{y}_k \in \mathbb{R}^n$ is the eigenvector of S^H corresponding to $\lambda_j \in \mathbb{R}$. Finally, $\mathbf{y}_n \in \mathbb{C}^n$ is the eigenvector of S^H corresponding to $\bar{\lambda}_n \in \mathbb{C}$.

By construction, we already know that (3.9) is satisfied with X_1 defined by (3.18) and C defined by (3.22). It remains to determine the coefficients $\gamma_1, \ldots, \gamma_{n-1}$ so that the deleted linearly dependent vector $\bar{\mathbf{x}}_n$ (the original $\mathbf{x}_{2\ell}$ before the reindexing) is also an eigenvector with eigenvalue $\bar{\lambda}_n$. Of course, we also need to make sure that the resulting C and K are real-valued ultimately.

Let

$$\mathbf{z} = (\bar{\lambda}_n I_n - S)\bar{\mathbf{x}}_n. \tag{3.24}$$

Substituting the eigenpair $(\bar{\lambda}_n, \bar{\mathbf{x}}_n)$ into (3.4) and using (3.22), we obtain

$$(\bar{\lambda}_n I_n - S^H)\mathbf{z} = -[\mathbf{y}_1, \ldots, \mathbf{y}_{2m}, \mathbf{y}_{2m+1}, \ldots, \mathbf{y}_{n-1}] \begin{bmatrix} \gamma_1\mathbf{y}_2^H\mathbf{z} \\ \gamma_2\mathbf{y}_1^H\mathbf{z} \\ \vdots \\ \gamma_{2m-1}\mathbf{y}_{2m}^H\mathbf{z} \\ \gamma_{2m}\mathbf{y}_{2m-1}^H\mathbf{z} \\ \gamma_{2m+1}\mathbf{y}_{2m+1}^H\mathbf{z} \\ \vdots \\ \gamma_{n-1}\mathbf{y}_{n-1}^H\mathbf{z} \end{bmatrix}. \tag{3.25}$$

With the assumption of (3.20), it is not difficult to see that

$$\mathbf{y}_j^H\mathbf{z} = 0 \quad \text{for } j = 2m+1, \ldots n-1. \tag{3.26}$$

The equation of (3.25) is equivalent to the equation

$$(\bar{\lambda}_n I_n - \Lambda_1^H) X_1^H \mathbf{z} = - \begin{bmatrix} \gamma_1 \mathbf{y}_2^H \mathbf{z} \\ \gamma_2 \mathbf{y}_1^H \mathbf{z} \\ \vdots \\ \gamma_{2m-1} \mathbf{y}_{2m}^H \mathbf{z} \\ \gamma_{2m} \mathbf{y}_{2m-1}^H \mathbf{z} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.27}$$

The left-hand side of (3.27) is completely known. It is now clear that the coefficients $\gamma_1, \ldots, \gamma_{2m}$ are uniquely determined if

$$\mathbf{y}_j^H \mathbf{z} \neq 0 \quad \text{for } j = 1, \ldots, 2m, \tag{3.28}$$

whereas the coefficients $\gamma_{2m+1}, \ldots, \gamma_{n-1}$ in (3.25) (and hence in (3.22)) can be arbitrary real numbers so long as the last $n - 2m$ equations in (3.27) are consistent; that is,

$$\mathbf{x}_j^H \mathbf{z} = 0 \quad \text{for } j = 2m+1, \ldots, n-1. \tag{3.29}$$

Assuming (3.28) and (3.29), we now show that the resulting matrix C in (3.22) is Hermitian. Toward that end, it suffices to show that $\gamma_{2j-1} = \bar{\gamma}_{2j}$ for $j = 1, \ldots, m$. Based on (3.26) and (3.29), we introduce the following two vectors for convenience:

$$\mathbf{p} := X_1^H \mathbf{z} = [p_1, \ldots, p_{2m}, 0, \ldots, 0, p_n]^T \in \mathbb{C}^n, \tag{3.30}$$

$$\mathbf{q} := X_1^{-1} \bar{\mathbf{x}}_n = [q_1, \ldots, q_{2m}, 0, \ldots, 0, q_n]^T \in \mathbb{C}^n. \tag{3.31}$$

For $j = 1, \ldots m$, the $(2j-1)$th and the $(2j)$th components of (3.27) are, respectively,

$$\begin{aligned} (\bar{\lambda}_n - \bar{\lambda}_{2j-1}) p_{2j-1} &= -\gamma_{2j-1} \mathbf{y}_{2j}^H \mathbf{z} = -\gamma_{2j-1} (\bar{\lambda}_n - \lambda_{2j}) q_{2j}, \\ (\bar{\lambda}_n - \bar{\lambda}_{2j}) p_{2j} &= -\gamma_{2j} \mathbf{y}_{2j-1}^H \mathbf{z} = -\gamma_{2j} (\bar{\lambda}_n - \lambda_{2j-1}) q_{2j-1}. \end{aligned}$$

Since $\lambda_{2j-1} = \bar{\lambda}_{2j}$, it follows that

$$p_{2j-1} = -\gamma_{2j-1} q_{2j}, \tag{3.32}$$

$$p_{2j} = -\gamma_{2j} q_{2j-1}. \tag{3.33}$$

On the other hand, observe that

$$\mathbf{z} = (\bar{\lambda}_n I_n - S) \bar{\mathbf{x}}_n = (\bar{\lambda}_n I_n - \bar{S} + \bar{S} - S) \bar{\mathbf{x}}_n = (\bar{S} - S) \bar{\mathbf{x}}_n \tag{3.34}$$

since $\bar{\mathbf{x}}_n$ is an eigenvector of $\bar{S}$. Observe also that

$$(\bar{S} - S) \mathbf{x}_j = 0 \quad \text{for } j = 1, \ldots, n-1 \tag{3.35}$$

because of the complex conjugation. It follows that

$$(\bar{S} - S) \bar{\mathbf{x}}_n = q_n (\bar{S} - S) \mathbf{x}_n, \tag{3.36}$$

$$\mathbf{x}_{2j-1}^H (\bar{S} - S) \bar{\mathbf{x}}_n = -\gamma_{2j-1} q_{2j}, \tag{3.37}$$

$$\mathbf{x}_{2j}^H (\bar{S} - S) \bar{\mathbf{x}}_n = -\gamma_{2j} q_{2j-1}. \tag{3.38}$$

Taking the conjugation of (3.37) and using (3.36), we obtain

$$-\bar{q}_n \bar{\mathbf{x}}_{2j-1}^H (\bar{S} - S)\bar{\mathbf{x}}_n = -\bar{\gamma}_{2j-1}\bar{q}_{2j}. \tag{3.39}$$

Comparing (3.38) and (3.39), since $\bar{\mathbf{x}}_{2j-1} = \mathbf{x}_{2j}$ for all $j = 1, \ldots, m$, we obtain a critical relationship that

$$\bar{q}_n \gamma_{2j} q_{2j-1} = -\bar{\gamma}_{2j-1}\bar{q}_{2j} \quad \text{for } j = 1, \ldots, m. \tag{3.40}$$

Now we are ready to show that $\gamma_{2j} = \bar{\gamma}_{2j-1}$, $j = 1, \ldots, m$. We rewrite $\mathbf{x}_n = \bar{X}_1 \bar{\mathbf{q}}$ from (3.31) as

$$\begin{aligned} \mathbf{x}_n &= \bar{q}_1 \bar{\mathbf{x}}_1 + \bar{q}_2 \bar{\mathbf{x}}_2 + \cdots + \bar{q}_{2m-1}\bar{\mathbf{x}}_{2m-1} + \bar{q}_{2m}\bar{\mathbf{x}}_{2m} + \bar{q}_n \bar{\mathbf{x}}_n \\ &= \bar{q}_1 \mathbf{x}_2 + \bar{q}_2 \mathbf{x}_1 + \cdots + \bar{q}_{2m-1}\mathbf{x}_{2m} + \bar{q}_{2m}\mathbf{x}_{2m-1} + \bar{q}_n \bar{\mathbf{x}}_n. \end{aligned} \tag{3.41}$$

Replacing the last term by

$$\bar{q}_n \bar{\mathbf{x}}_n = \bar{q}_n q_1 \mathbf{x}_1 + \bar{q}_n q_2 \mathbf{x}_2 + \cdots + \bar{q}_n q_{2m}\mathbf{x}_{2m} + |q_n|^2 \mathbf{x}_n,$$

we obtain the equality

$$\begin{aligned} \mathbf{x}_n = {} & (\bar{q}_n q_1 + \bar{q}_2)\mathbf{x}_1 + (\bar{q}_n q_2 + \bar{q}_1)\mathbf{x}_2 + \cdots + (\bar{q}_n q_{2m-1} + \bar{q}_{2m})\mathbf{x}_{2m-1} \\ & + (\bar{q}_n q_{2m} + \bar{q}_{2m-1})\mathbf{x}_{2m} + |q_n|^2 \mathbf{x}_n. \end{aligned}$$

Since $\{\mathbf{x}_1, \ldots, \mathbf{x}_{2m}, \mathbf{x}_n\}$ are linearly independent, it holds that

$$\bar{q}_n q_{2j-1} + \bar{q}_{2j} = 0 \quad \text{for } j = 1, \ldots, m. \tag{3.42}$$

Substituting (3.42) into (3.40), we have proved that $\gamma_{2j} = \bar{\gamma}_{2j-1}$ for $j = 1, \ldots, m$.

By now, we have completed the proof that the matrix C constructed using (3.27) is Hermitian. We are ready to state our second major result.

THEOREM 3.5. *Let* $(\tilde{\Lambda}, \tilde{X}) \in \mathbb{C}^{(n+1)\times(n+1)} \times \mathbb{C}^{n\times(n+1)}$ *be given as in* (1.7) *and* (1.8). *Assume that one eigenvector, say,* $\mathbf{x}_{2\ell} \in \mathbb{C}^n$, *depends linearly on the remaining eigenvectors which are linearly independent. Then suppose the following:*

1. *Suppose* $\ell = \frac{n+1}{2}$*; that is, suppose that there is no real-valued vector at all in* X. *If the condition* (3.28) *is satisfied for* $j = 1, \ldots, n-1$, *then the IMQEP has a unique solution.*
2. *Suppose* $\ell < \frac{n+1}{2}$ *and that* (3.20) *holds. If the condition* (3.28) *is satisfied for* $j = 1, \ldots, 2\ell - 2$ *and the condition* (3.29) *is satisfied for* $j = 2\ell + 1, \ldots, n+1$, *then the IMQEP has infinitely many solutions; otherwise it has no solution.*

Proof. Thus far, we have already shown that both matrices C and K can be constructed uniquely and are Hermitian. It remains only to show that C and K are real symmetric. It suffices to prove that $C = \bar{C}$ and $K = \bar{K}$.

Consider the IMQEP associated with the spectral data $(\bar{\tilde{\Lambda}}, \bar{\tilde{X}})$, the complex conjugate of the original data $(\tilde{\Lambda}, \tilde{X})$. Then the sufficient condition (3.28) for the problem associated with $(\tilde{\Lambda}, \tilde{X})$ applies equally well to the new problem associated with $(\bar{\tilde{\Lambda}}, \bar{\tilde{X}})$. A quadratic pencil therefore can be constructed to solve the new IMQEP. Indeed, by repeating the procedure of construction described above, it is not difficult to see that the constructed pencil for $(\bar{\tilde{\Lambda}}, \bar{\tilde{X}})$ is of the form

$$\widetilde{Q}(\lambda) = \lambda^2 I_n + \lambda \bar{C} + \bar{K}.$$

Since Λ and X are closed under complex conjugation, the spectral information $(\bar{\tilde{\Lambda}}, \bar{\tilde{X}})$ is actually a reshuffle of (Λ, X). As a matter of fact, these two IMQEPs are the same problem. In the first case where $\ell = \frac{n+1}{2}$, the solution is already unique. In the second case where $\ell < \frac{n+1}{2}$ and (3.20) holds, so long as the arbitrarily selected real coefficients $\gamma_{2m+1}, \ldots, \gamma_{n-1}$ remain fixed, the complex-conjugated coefficients $\gamma_1, \ldots, \gamma_{2m}$ are also uniquely determined. In either case, we must have that $C = \bar{C} = C^H$ and $K = \bar{K} = K^H$. ∎

3.3. Numerical examples. The argument presented in the proceeding section offers a constructive way to solve the IMQEP. In this section we use numerical examples to illustrate the two cases discussed above. For the ease of running text, we report all numbers in five significant digits only, though all calculations are carried out in full precision.

Example 3. To generate test data, we first randomly generate a 5×5 real symmetric quadratic pencil $Q(\lambda) = \lambda^2 I + \lambda C + K$ and compute its "exact" eigenpairs (Λ_e, X_e) numerically. We obtain that $\Lambda_e = \text{diag}\{\lambda_1, \ldots, \lambda_{10}\}$, $X_e = [\mathbf{x}_1, \ldots, \mathbf{x}_{10}]$ with $\lambda_1 = -0.31828 + 0.86754i = \bar{\lambda}_2$, $\lambda_3 = -0.95669 + 0.17379i = \bar{\lambda}_4$, $\lambda_5 = -4.4955$, $\lambda_6 = 1.5135$, $\lambda_7 = -0.24119 + 0.029864i = \bar{\lambda}_8$, $\lambda_9 = 0.91800$, $\lambda_{10} = -1.7359$, and the corresponding eigenvectors

$$\mathbf{x}_1 = \bar{\mathbf{x}}_2 = \begin{bmatrix} 15.159 - 11.123i \\ -77.470 - 14.809i \\ 2.1930 - 10.275i \\ 0.38210 + 16.329i \\ 57.042 + 18.419i \end{bmatrix}, \quad \mathbf{x}_3 = \bar{\mathbf{x}}_4 = \begin{bmatrix} 65.621 + 34.379i \\ 22.625 + 24.189i \\ -37.062 + 15.825i \\ -9.6496 + 14.401i \\ -0.61893 + 25.609i \end{bmatrix},$$

$$\mathbf{x}_5 = \begin{bmatrix} 2.2245 \\ 1.5893 \\ 2.1455 \\ 2.1752 \\ 1.6586 \end{bmatrix}, \quad \mathbf{x}_6 = \begin{bmatrix} 34.676 \\ -5.8995 \\ 37.801 \\ -66.071 \\ -6.6174 \end{bmatrix}, \quad \mathbf{x}_7 = \bar{\mathbf{x}}_8 = \begin{bmatrix} 35.257 - 0.31888i \\ -25.619 - 4.2156i \\ 98.914 - 1.0863i \\ -21.348 + 5.8290i \\ -97.711 - 1.0693i \end{bmatrix},$$

$$\mathbf{x}_9 = \begin{bmatrix} -97.828 \\ 10.879 \\ 100.00 \\ -4.3638 \\ 22.282 \end{bmatrix}, \quad \mathbf{x}_{10} = \begin{bmatrix} -1.3832 \\ 4.4564 \\ -1.1960 \\ -4.0934 \\ 5.7607 \end{bmatrix}.$$

Note that the above spectral data are not arranged in any specific order. According to our theory, any $n+1$ eigenpairs satisfying the specification of (3.1) and (3.2) and the sufficient condition (3.13) or (3.28), depending upon whether assumptions in sections 3.1 or 3.2 with $\ell = \frac{n+1}{2}$ are applicable, should ensure the full recovery of the original pencil.

Case 1. Suppose the prescribed partial eigeninformation is given by

$$(\tilde{\Lambda}, \tilde{X}) = \left(\text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\},\ [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6]\right).$$

It is easy to check that the real-valued eigenvector $\mathbf{x}_6$ depends linearly on the first five eigenvectors which are linearly independent. This fits the situation discussed in section 3.1 where we choose to work with

$$(\hat{\Lambda}_1, \hat{X}_1) = \left(\text{diag}\{\lambda_1, \bar{\lambda}_1, \lambda_3, \bar{\lambda}_3, \lambda_5\},\ [\mathbf{x}_1, \bar{\mathbf{x}}_1, \mathbf{x}_3, \bar{\mathbf{x}}_3, \mathbf{x}_5]\right).$$

TABLE 3.1

Eigenpairs	Residual $\|\hat{Q}_1(\lambda_j)\mathbf{x}_j\|_2$
$(\lambda_1, \mathbf{x}_1)$	2.2612e-015
$(\lambda_2, \mathbf{x}_2)$	2.2612e-015
$(\lambda_3, \mathbf{x}_3)$	2.9827e-015
$(\lambda_4, \mathbf{x}_4)$	2.9827e-015
$(\lambda_5, \mathbf{x}_5)$	2.0381e-015
$(\lambda_6, \mathbf{x}_6)$	1.8494e-014
$(\lambda_7, \mathbf{x}_7)$	7.9955e-014
$(\lambda_8, \mathbf{x}_8)$	7.9955e-014
$(\lambda_9, \mathbf{x}_9)$	4.4264e-014
$(\lambda_{10}, \mathbf{x}_{10})$	4.5495e-014

TABLE 3.2

$\|\hat{C} - C\|_2$	1.8977e-014
$\|\hat{K} - K\|_2$	7.3897e-014

We construct the unique real symmetric quadratic pencil

$$\hat{Q}(\lambda) = \lambda^2 I_5 + \lambda\hat{C} + \hat{K}$$

by the method described in the proof of Theorem 3.2. In Tables 3.1 and 3.2, we show the residual $\|\hat{Q}(\lambda_j)\mathbf{x}_j\|_2$, where $(\lambda_j, \mathbf{x}_j)$ are the computed eigenpairs of $Q(\lambda)$ for $j = 1, \ldots, 10$, as well as the difference $\|\hat{C} - C\|_2$ and $\|\hat{K} - K\|_2$, respectively.

Case 2. Suppose the prescribed spectral information is given by

$$(\tilde{\Lambda}, \tilde{X}) = \left(\text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_7, \lambda_8\},\ [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_7, \mathbf{x}_8]\right).$$

Note that all eigenvectors are complex-valued. This fits the situation discussed in section 3.2 with $\ell = \frac{n+1}{2}$, where we choose to work with

$$(\check{\Lambda}_1, \check{X}_1) = \left(\text{diag}\{\lambda_1, \bar{\lambda}_1, \lambda_3, \bar{\lambda}_3, \lambda_7\},\ [\mathbf{x}_1, \bar{\mathbf{x}}_1, \mathbf{x}_3, \bar{\mathbf{x}}_3, \mathbf{x}_7]\right).$$

We construct the unique real symmetric quadratic pencil

$$\check{Q}(\lambda) = \lambda^2 I_5 + \lambda\check{C} + \check{K}$$

by the method described in the proof of Theorem 3.5. In Tables 3.3 and 3.4, we show the residual $\|\check{Q}(\lambda_j)\mathbf{x}_j\|_2$ for $j = 1, \ldots, 10$, as well as the difference $\|\check{C} - C\|_2$ and $\|\check{K} - K\|_2$, respectively.

It can be checked that both cases above satisfy the sufficient conditions (3.13) and (3.28), respectively. The errors shown in the tables seem to be quite satisfactory.

Example 4. In the previous example we demonstrated two scenarios of prescribed spectral information that give rise to the same unique solution to the IMQEP. Now we demonstrate the second situation in Theorem 3.5 when both $\ell < \frac{n+1}{2}$ and (3.20) take place. Our theory asserts that there will be either infinitely many solutions to the IMQEP or no solution at all.

Consider the case where $n = 4$ and the prescribed eigenvalues are given by $\lambda_1 = 3.3068 + 8.1301i = \bar{\lambda}_2$, $\lambda_3 = 1.8702 + 2.7268i = \bar{\lambda}_4$, $\lambda_5 = 5.4385$ with corresponding eigenvectors

$$\mathbf{x}_1 = \bar{\mathbf{x}}_2 = \begin{bmatrix} 0 \\ 9.2963 + 1.5007i \\ 2.3695 + 1.9623i \\ 3.8789 + 1.0480i \end{bmatrix},\ \mathbf{x}_3 = \bar{\mathbf{x}}_4 = \begin{bmatrix} 0 \\ 6.5809 + 8.3476i \\ 4.9742 + 8.0904i \\ 1.1356 + 5.5542i \end{bmatrix},\ \mathbf{x}_5 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

TABLE 3.3

Eigenpairs	Residual $\|\check{Q}(\lambda_j)\mathbf{x}_j\|_2$
$(\lambda_1, \mathbf{x}_1)$	4.5422e-016
$(\lambda_2, \mathbf{x}_2)$	4.5422e-016
$(\lambda_3, \mathbf{x}_3)$	7.8025e-016
$(\lambda_4, \mathbf{x}_4)$	7.8025e-016
$(\lambda_5, \mathbf{x}_5)$	3.7137e-014
$(\lambda_6, \mathbf{x}_6)$	2.9549e-014
$(\lambda_7, \mathbf{x}_7)$	9.4143e-016
$(\lambda_8, \mathbf{x}_8)$	9.4143e-016
$(\lambda_9, \mathbf{x}_9)$	6.0018e-014
$(\lambda_{10}, \mathbf{x}_{10})$	4.6464e-014

TABLE 3.4

$\|\check{C} - C\|_2$	1.9222e-014
$\|\check{K} - K\|_2$	1.7951e-014

TABLE 3.5

	$\gamma_3 = 2.56$	$\gamma_3 = 40.6$	$\gamma_3 = 506$
Eigenpairs	Residual	Residual	Residual
$(\lambda_1, \mathbf{x}_1)$	2.9334e-011	2.9334e-011	2.9334e-011
$(\lambda_2, \mathbf{x}_2)$	2.9334e-011	2.9334e-011	2.9334e-011
$(\lambda_3, \mathbf{x}_3)$	7.8802e-011	7.8802e-011	7.8802e-011
$(\lambda_4, \mathbf{x}_4)$	7.8802e-011	7.8802e-011	7.8802e-011
$(\lambda_5, \mathbf{x}_5)$	1.7764e-015	2.8422e-014	4.5475e-013

respectively. It is obvious upon inspection that the linearly dependent vector in the above X must be a complex-valued vector. Let this linearly dependent vector be $\mathbf{x}_4$. Then the real symmetric quadratic pencil

$$Q(\lambda) = \lambda^2 I + \lambda C + K,$$

where $C = -(S + S^H) + \gamma_1 \mathbf{y}_1 \mathbf{y}_2^H + \gamma_2 \mathbf{y}_2 \mathbf{y}_1^H + \gamma_3 \mathbf{y}_3 \mathbf{y}_3^H$ and $K = -(S + C)S$, can be constructed with arbitrary $\gamma_3 \in \mathbb{R}$. In Table 3.5 we show the residual $\|Q(\lambda_j)\mathbf{x}_j\|_2$ for $j = 1, \ldots, 5$ with various values of γ_3.

Suppose we modify the first entries of the complex eigenvectors to

$$\mathbf{x}_1 = \bar{\mathbf{x}}_2 = \begin{bmatrix} 9.2963 + 1.5007i \\ 9.2963 + 1.5007i \\ 2.3695 + 1.9623i \\ 3.8789 + 1.0480i \end{bmatrix}, \ \mathbf{x}_3 = \bar{\mathbf{x}}_4 = \begin{bmatrix} 6.5809 + 8.3476i \\ 6.5809 + 8.3476i \\ 4.9742 + 8.0904i \\ 1.1356 + 5.5542i \end{bmatrix}, \ \mathbf{x}_5 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Still, we see that the linearly dependent vector in the corresponding X must be a complex-valued vector, say $\mathbf{x}_4$. However, we find that the condition (3.29) is not satisfied because

$$\mathbf{x}_3^H \mathbf{z} = \mathbf{x}_3^H (\bar{\lambda}_4 I_4 - S)\bar{\mathbf{x}}_4 = -115.54 + 600.67i \neq 0.$$

The system (3.27) being inconsistent, the real coefficient γ_3 in (3.22) is not solvable. We conclude that the prescribed vectors and the corresponding scalars λ_i, $i = 1, \ldots, 5$, indicated above cannot be part of the spectrum of any 4×4 real-valued, symmetric and monic quadratic pencil.

4. Conclusion. The quadratic eigenvalue problem arises in many important applications. Its inverse problem is equally important in practice. In a large or complicated system, often it is the case that only partial eigeninformation is available. To understand how a physical system modelled by a quadratic pencil should be modified with only partial eigeninformation in hand, it will be very helpful to first understand how the IQEP should be solved. Some general theory toward that end has been presented in this paper.

In the first part of this paper, we found that the ISQEP is solvable, provided that the number of given eigenpairs is less than or equal to the size of the matrices and that the given vectors are linearly independent. A simple recipe for constructing such a matrix was described, which can serve as the basis for numerical computation. We also found that the unspecified eigenstructure of the reconstructed quadratic pencil is quite limited in the sense discussed in section 2.2. We demonstrated three different ways for the construction that not only satisfied the spectral constraints but also best approximated the original analytical model in some least-squares sense.

In the second part of this paper, we established some general existence theory for the inverse problem when the leading matrix coefficient M is known and fixed. The procedure used in the proof can also provide a basis for numerical computation.

It should be noted that the stiffness matrix K is normally more complicated than the mass matrix M. The requirement of maintaining physical feasibility also imposes constraints on the stiffness matrix, making it less flexible and more difficult to construct. Thus, one usual way of formulating an inverse eigenvalue problem is to have the stiffness matrix K determined and fixed from the existing structure, known as the static constraints, and then to find the mass matrix M so that some desired natural frequencies are achieved. This is sometimes so desired even without the damping term C. By exchanging the roles of M and K, the discussion in this paper could be applied equally well to the IQEP formed with the aforementioned static constraints in mind.

The study made in this paper should have shed light on the long-standing question of how much a quadratic pencil could be updated, modified, or tuned if some of its eigenvalues and eigenvectors are to be kept invariant. Finally, we should point out that there are unfinished tasks in this study. Among these, sensitivity analysis in the case of a unique solution, robustness in the case of multiple solutions, and existence theory where M or K is specially structured are just a few interesting topics that remain to be further investigated.

REFERENCES

[1] J. Carvalho, B. N. Datta, W. W. Lin, and C. S. Wang, *Eigenvalue Embedding in a Quadratic Pencil Using Symmetric Low Rank Updates*, preprint, National Center for Theoretical Sciences, National Tsinghua University, Hsinchu, Taiwan, 2001.

[2] E. K.-W. Chu and B. N. Datta, *Numerically robust pole assignment for second-order systems*, Internat. J. Control, 64 (1996), pp. 1113–1127.

[3] M. T. Chu, *Inverse eigenvalue problems*, SIAM Rev., 40 (1998), pp. 1–39.

[4] M. T. Chu and G. H. Golub, *Structured inverse eigenvalue problems*, Acta Numer., 11 (2002), pp. 1–71.

[5] B. N. Datta, *Finite element model updating, eigenstructure assignment and eigenvalue embedding techniques for vibrating systems*, Mechanical Systems and Signal Processing, 16 (2002), pp. 83–96.

[6] B. N. Datta, S. Elhay, Y. M. Ram, and D. R. Sarkissian, *Partial eigenstructure assignment for the quadratic pencil*, J. Sound Vibration, 230 (2000), pp. 101–110.

[7] B. N. Datta and D. R. Sarkissian, *Theory and computations of some inverse eigenvalue problems for the quadratic pencil*, in Structured Matrices in Mathematics, Computer Science, and Engineering. I, Contemp. Math. 280, AMS, Providence, RI, 2001, pp. 221–240.
[8] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.
[9] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[10] W. R. Ferng, W. W. Lin, D. Pierce, and C. S. Wang, *Nonequivalence transformation of λ-matrix eigenproblems and model embedding approach to model tunning*, Numer. Linear Algebra Appl., 8 (2001), pp. 53–70.
[11] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.
[12] N. K. Nichols and J. Kautsky, *Robust eigenstructure assignment in quadratic matrix polynomials: Nonsingular case*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 77–102.
[13] Y. M. Ram and S. Elhay, *An inverse eigenvalue problem for the symmetric tridiagonal quadratic pencil with application to damped oscillatory systems*, SIAM J. Appl. Math., 56 (1996), pp. 232–244.
[14] Y. M. Ram, http://me.lsu.edu/~ram/PAPERS/publications.html.
[15] D. D. Sivan and Y. M. Ram, *Physical modifications to vibratory systems with assigned eigendata*, Trans. ASME J. Appl. Mech., 66 (1999), pp. 427–432.
[16] L. Starek and D. J. Inman, *Symmetric inverse eigenvalue vibration problem and its applications*, Mechanical Systems and Signal Processing, 15 (2001), pp. 11–29.
[17] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286; also available online at http://www.ma.man.ac.uk/~ftisseur.
[18] D. C. Zimmerman and M. Widengren, *Correcting finite element models using a symmetric eigenstructure assignment technique*, AIAA J., 28 (1990), pp. 1670–1676.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1021–1031

A NOTE ON MULTIPLICATIVE BACKWARD ERRORS OF ACCURATE SVD ALGORITHMS*

FROILÁN M. DOPICO† AND JULIO MORO†

Abstract. Multiplicative backward stability results are presented for two algorithms which compute the singular value decomposition of dense matrices. These algorithms are the classical one-sided Jacobi algorithm, with a stringent stopping criterion, and an algorithm which uses one-sided Jacobi to compute high accurate singular value decompositions of matrices given as rank-revealing factorizations. When multiplicative backward errors are small, the multiplicative perturbation theory for the singular value decomposition developed in the last decade can be applied to get high accuracy bounds on the errors of the computed singular values and vectors.

Key words. singular value decomposition, Jacobi algorithm, high relative accuracy, rank-revealing decompositions, mutiplicative perturbation theory

AMS subject classifications. 65F15, 65G50

DOI. 10.1137/S0895479803427005

1. Introduction.

The singular value decomposition (SVD) of a matrix $G \in \mathbb{R}^{m\times n}$ $(m \geq n)$ is the factorization $G = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times n}$ has orthonormal columns, $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is nonnegative and diagonal. The columns of U are the left singular vectors of G, the columns of V are the right singular vectors of G, and σ_i are the singular values of G. Given two nonsingular square matrices D_1 and D_2, the matrix D_1GD_2 is called a multiplicative perturbation of G. In the last decade, a perturbation theory bounding the differences between the singular values and vectors of G and D_1GD_2 has been developed [13, 18, 19, 17]. Let $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_n \geq 0$ be, respectively, the singular values of G and D_1GD_2 and u_i, v_i, $\tilde{u}_i, \tilde{v}_i$, $i = 1, \ldots, n$, be the corresponding pairs of left and right singular vectors. Let us denote by $\|\,.\,\|$ the usual Euclidean vector norm when the argument is a vector and the spectral, or two, matrix norm when the argument is a matrix. Then the multiplicative perturbation theory essentially bounds

$$\frac{|\sigma_i - \tilde{\sigma}_i|}{\sigma_i} \quad \text{and} \quad \max\{\|v_i - \tilde{v}_i\|, \|u_i - \tilde{u}_i\|\}\mathrm{relgap}_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\mathrm{relgap}_i = \min_{j\neq i} |\sigma_i - \tilde{\sigma}_j|/\sigma_i$, by a small integer constant times $\max\{\|I - D_1\|, \|I - D_2\|\}$ [13, 19]. Therefore, if D_1 and D_2 are close to the identity matrix, the relative differences between the singular values of G and D_1GD_2 are small, and the differences between the singular vectors multiplied by the relative gaps are also small. Obviously, (1) makes sense only if $\sigma_i \neq 0$. If $\sigma_i = 0$, then it is trivial that $\tilde{\sigma}_i = 0$ and it can be shown that the differences between the corresponding singular vectors are simply less than a small integer constant times $\max\{\|I - D_1\|, \|I - D_2\|\}$ [13, 19]. Notice that classical perturbation theory [22], valid for additive perturbations of the type $G + E$, bounds absolute differences between singular values, i.e., $|\sigma_i - \tilde{\sigma}_i| \leq \|E\|$,

*Received by the editors April 30, 2003; accepted for publication (in revised form) by I.C.F. Ipsen October 9, 2003; published electronically June 4, 2004. The research conducted for this paper was partially supported by the Ministerio de Ciencia y Tecnología of Spain through grant BFM-2000-0008.

http://www.siam.org/journals/simax/25-4/42700.html

†Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés, Spain (dopico@math.uc3m.es, jmoro@math.uc3m.es).

and the gaps appearing in the singular vector bounds are also absolute, i.e., $\text{gap}_i = \min_{j \neq i} |\sigma_i - \tilde{\sigma}_j|/\sigma_1$.

Multiplicative perturbation theory has been successfully used in proving that some algorithms are able to compute the SVD with high relative accuracy when applied to matrices with special structure. Here high relative accuracy means that the relative errors in the computed singular values are of order ϵ, with ϵ being the machine precision, and that the errors in the computed singular vectors are of order ϵ divided by the corresponding singular value relative gap, i.e., relgap_i. Well-known examples of matrices for which it is possible to compute high relative accuracy SVDs are bidiagonal matrices [5, 14]; matrices of the form $G = BD$, with D diagonal and B well-conditioned [6, 11, 20]; positive definite matrices of the form DAD, with D diagonal and A well-conditioned [6]; and matrices for which it is possible to compute accurately a rank-revealing decomposition [4]. This latter class contains the previous ones and many others (see also [3, 7, 8]). A technical remark is in order here: although the approach in [4] includes the case of bidiagonal matrices, since bidiagonal matrices are acyclic, the original approaches in [5, 14] are much faster and do not require one to compute a rank-revealing factorization.

There exists a relative perturbation theory for additive perturbations which gives structured bounds for the quantities appearing in (1); see [17] and references therein. This perturbation theory has been used to guarantee the high relative accuracy of the SVDs computed by some algorithms [6, 11, 20]. However, it was shown in [4] that multiplicative perturbation theory can also be used in these cases. Thus, at present, it seems that multiplicative perturbation theory has a wider applicability in the context of high relative accuracy computations of SVDs. In fact, multiplicative perturbation theory and some of its applications have already been presented in some recent text books [2, sections 5.2.1, 5.4.2, 5.4.3].

Although the accurate computation of the SVD is still a work in progress and, as a consequence, it is still too early to know which tools will be the most useful in future developments, there are sound reasons to support the prominent role of multiplicative perturbation theory: for instance, the simplicity of the bounds, or the simple way in which multiplicative perturbation bounds can be composed with each other.

In spite of the present importance of multiplicative perturbation theory, there is no theorem so far stating in multiplicative form the backward stability properties of high accuracy algorithms for the SVD, i.e., a theorem saying that the computed SVD of a matrix G is essentially the exact SVD of a nearby *multiplicative* perturbation of G. In the context of usual algorithms for SVD computations the usual backward stability result [1, section 4.9.1] states that the computed SVD of a matrix G is essentially the exact SVD of a nearby *additive* perturbation of G, i.e., a matrix $G + E$ with $\|E\| \leq p(m,n)\,\epsilon\,\|G\|$ and $p(m,n)$ a modestly growing function of m and n. Our goal in this note is to prove a very strong form of multiplicative backward stability for two algorithms which are able to compute SVDs with high relative accuracy in some important cases. The starting point will be the roundoff error analysis previously developed by other authors in [4, 6], and especially in [11]. The theorems we obtain have already been used in [9] and greatly simplify the way in which the error bounds for singular values and vectors are obtained in [6, 20, 4] just by using the multiplicative perturbation theory for the SVD. Moreover, we hope that the theorems we present will be useful in future error analyses of accurate SVD algorithms.

Finally, it is interesting to stress that the multiplicative backward error results we are going to present cannot be deduced from additive backward error results of

the form $G + E$ just by factoring out the inverse or pseudoinverse of G. This is obvious in the case of standard backward stability results [1, section 4.9.1], because the information about the perturbation is just $\|E\| \leq p(m,n)\,\epsilon\,\|G\|$. Therefore, if we write $G + E = G(I + G^{-1}E)$, the most we can assert on the magnitude of the multiplicative perturbation is $\|G^{-1}E\| \leq p(m,n)\,\epsilon\,\|G^{-1}\|\,\|G\|$, and the condition number $\|G^{-1}\|\,\|G\|$ can be very large. On the other hand, factoring out G in the additive backward error result appearing in [11, Proposition 3.13] for the one-sided Jacobi algorithm will play an essential role in our developments, but this is not the only thing to do. In fact, we will need to introduce multiplicative perturbations on *both* sides of the matrix. This is the reason why the stability results presented in [6, 11] are mixed forward-backward error results.

The paper is organized as follows: a multiplicative backward stability theorem is proved in section 2 for the one-sided Jacobi algorithm, and the same is done for Algorithm 3.1 of [4] in section 3. Finally, in section 4 we discuss a different version of one-sided Jacobi, which is usually faster although the error bounds are weaker.

Notation and model of arithmetic. In the statements of the subsequent theorems big-O notation will be used. Given a scalar quantity b, the meaning of $O(\epsilon b)$ is that $O(\epsilon b) = p(m,n)\,\epsilon\,b + O(\epsilon^2)$ with $p(m,n)$ a polynomial of low degree in the dimensions m, n of the problem.

The conventional error model for floating point arithmetic with guard-digit will be used:

$$\texttt{fl}(a \odot b) = (a \odot b)(1 + \delta),$$

where a and b are real floating point numbers, $\odot \in \{+, -, \times, /\}$, and $|\delta| \leq \epsilon$, where ϵ is the machine precision. Moreover, we assume that neither overflow nor underflow occur. For the sake of simplicity, we will commit a slight abuse of notation, denoting by $\texttt{fl}(expr)$ the computed result in finite precision of expression $expr$, instead of its rigorous meaning of the closest floating point number to $expr$.

2. Backward error of one-sided Jacobi SVD algorithm. One-sided Jacobi algorithms for the SVD [15, section 8.6.3] multiply a matrix by a sequence of Jacobi rotations, all of them acting on the same side. When the rotations are applied to the matrix from the left (right), the goal is to converge to a matrix with orthogonal rows (columns). These two different implementations of one-sided Jacobi will be called, respectively, left-handed and right-handed Jacobi. A detailed pseudocode for the right-handed Jacobi algorithm can be found in [6, Algorithm 4.1]. The left-handed version follows easily from the right-handed version applied to the transpose matrix.

A plain implementation of one-sided Jacobi yields an algorithm much slower than the SVD algorithms based on first bidiagonalizing the matrix. However, one-sided Jacobi has an important advantage: if the stopping criterion proposed in [6, Algorithm 4.1] is used, then the one-sided Jacobi algorithm is able to compute the SVD with high relative accuracy for matrices that are the product of a diagonal matrix (possibly with elements of widely varying magnitudes) and a well-conditioned matrix. To be more precise, let D be a diagonal matrix; then high relative accuracy is achieved for matrices of the type DB if B has full row rank and is well-conditioned, or BD if B has full column rank and is well-conditioned. This high relative accuracy was first proved in [6] under a minor proviso. A proof valid in general was presented in [11] (see also references therein) and [20]. In this latter proof it is essential that the Jacobi rotations are applied on the side opposite to the diagonal matrix D. At present,

fast and sophisticated versions of one-sided Jacobi algorithm are being developed by Drmač along the ideas of [12].

It is very important to remark that if one-sided Jacobi is implemented as in [6, Algorithm 4.1], then underflows appear frequently for very ill conditioned matrices, and the high relative accuracy in the computed SVD expected for matrices of the form DB or BD (see previous paragraph) is lost. To get results with high relative accuracy, whenever the singular values are inside the range of the arithmetic, the Jacobi rotations have to be carefully implemented according to the method developed in [10].

The next theorem proves that the one-sided Jacobi SVD algorithm on a square invertible matrix produces a small multiplicative backward error; i.e., the computed SVD is nearly the exact SVD of a close multiplicative perturbation of the original matrix. We restrict ourselves to square matrices because, in practice, for the nonsquare case a QR factorization is computed first, and then one-sided Jacobi is applied to the square factor R. This reduces the computational cost. The following notation will be used: the ith column (resp., row) of any matrix A is denoted by $A(:,i)$ (resp., $A(i,:)$), $\widetilde{A}$ denotes the last matrix in the sequence computed by the right-handed Jacobi process, and $\kappa(A)$ is the spectral condition number of A. This theorem is based on the error analysis presented in [11, Proposition 3.13] and shows that with a small additional effort a strong backward multiplicative result can be obtained.

THEOREM 2.1. *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix and let $\widehat{U}\widehat{\Sigma}\widehat{V}^T$ be the SVD computed in finite arithmetic with machine precision ϵ by the right-handed Jacobi SVD algorithm applied on A with stopping criterion*[1]

$$\max_{i\neq j} \mathtt{fl}\left(\frac{|\widetilde{A}(:,i)^T\widetilde{A}(:,j)|}{\|\widetilde{A}(:,i)\|\,\|\widetilde{A}(:,j)\|}\right) \leq n\,\epsilon \qquad \text{for} \quad i\neq j. \tag{2}$$

Then there exist matrices U', V', E_L, $E_R \in \mathbb{R}^{n\times n}$, such that U' and V' are orthogonal,

$$\begin{aligned} &\|U'-\widehat{U}\| = O(\epsilon), \quad \|V'-\widehat{V}\| = O(\epsilon),\\ &\|E_L\| = O(\epsilon), \quad \|E_R\| = O(\epsilon\kappa(A_N)), \end{aligned} \tag{3}$$

where $A_N = D_N^{-1}A$, with D_N a diagonal matrix with elements $(D_N)_{ii} = \|A(i,:)\|$, and

$$(I+E_L)A(I+E_R) = U'\widehat{\Sigma}V'^T. \tag{4}$$

Proof. It is known [11, Proposition 3.13] that, under the conditions above, the matrix $\widetilde{A}$ satisfying the stopping criterion (2) can be written as

$$\widetilde{A} = (A+\delta A)V'$$

for an orthogonal matrix V' with $\|V'-\widehat{V}\| = O(\epsilon)$ and δA such that

$$\|\delta A(i,:)\| \leq \epsilon_J\,\|A(i,:)\|, \qquad i=1,\ldots,n, \tag{5}$$

[1] A similar result holds with $n\epsilon$ replaced by any tolerance *tol* in criterion (2). In that case, $\|U'-\widehat{U}\| \leq n\,tol + O(\epsilon)$ and $\|E_L\| \leq n\,tol + O(\epsilon)$. Notice, however, that if the tolerance is larger than $O(\epsilon)$, then the computed left singular vectors will fail, in general, to be orthogonal up to $O(\epsilon)$.

for a certain $\epsilon_J = O(\epsilon)$ which depends on the sweeps required for convergence.[2] Hence,

$$\widetilde{A} = A(I + E_R)V' \tag{6}$$

for $E_R = A^{-1}\delta A$. If we now scale $A = D_N A_N$, so that A_N has rows of unit Euclidean length, the bound (5) implies

$$\|E_R\|_F \leq \|A_N^{-1}\|_F \, \|D_N^{-1}\delta A\|_F \leq \sqrt{n}\,\epsilon_J \, \|A_N^{-1}\|_F,$$

where $\|\cdot\|_F$ stands for the Frobenius norm.[3] Finally, since $\|A_N\|_F = \sqrt{n}$, it follows that the Frobenius norm of E_R, and consequently its spectral norm, is bounded by $\epsilon_J\,\kappa_F(A_N) = O(\epsilon\kappa(A_N))$.

On the other hand, recall that if we denote by $\widetilde{\Sigma}$ the diagonal matrix whose ith diagonal entry is the Euclidean norm of the ith column of $\widetilde{A}$, then $\widehat{\Sigma}$ and $\widehat{U}$ are computed as $\widehat{\Sigma} = \mathtt{fl}(\widetilde{\Sigma})$ and $\widehat{U} = \mathtt{fl}(\widetilde{A}\widehat{\Sigma}^{-1})$. Notice that each element $\widehat{u}_{ij}$ of $\widehat{U}$ can be written as $\widehat{u}_{ij} = (\widetilde{A}_{ij}/\widehat{\Sigma}_{jj})(1+\epsilon_{ij})$ with $|\epsilon_{ij}| \leq \epsilon$. Let U be the matrix such that $\widetilde{A} = U\,\widehat{\Sigma}$. Then (6) implies that

$$U\widehat{\Sigma}(V')^T = A(I + E_R)$$

with $\|U - \widehat{U}\|_F \leq \epsilon\|U\|_F$. It remains only to show, using the stopping criterion, that there is an orthogonal matrix U' such that

$$U = (I + E_L)^{-1}U'$$

with $\|E_L\| = O(\epsilon)$ and $\|U' - \widehat{U}\| = O(\epsilon)$.

It follows from condition (2) that each off-diagonal element of U^TU is bounded in absolute value by $cn\epsilon + O(\epsilon^2)$, with c a small integer constant. The diagonal elements of U^TU, on the other hand, are $1 + \alpha_{ii}$ with $|\alpha_{ii}| \leq cn\epsilon + O(\epsilon^2)$. Thus, $\|U^TU - I\|_F \leq cn^2\epsilon + O(\epsilon^2)$. If $U = W_L(I + \delta\Sigma)W_R^T$ is the SVD of U, then $\|\delta\Sigma\|_F \leq cn^2\epsilon + O(\epsilon^2)$. Denoting $U' = W_LW_R^T$, it follows that $U = (I + \delta U)U'$, where U' is orthogonal and $\|\delta U\|_F = \|\delta\Sigma\|_F$.

Defining $E_L = (I + \delta U)^{-1} - I$, we obtain that $\|E_L\|_F = \|\delta U\|_F + O(\|\delta U\|_F^2) \leq cn^2\epsilon + O(\epsilon^2)$.

Finally, $\|\widehat{U} - U'\|_F \leq \|\widehat{U} - U\|_F + \|U - U'\|_F$, but $\|U - U'\|_F = \|\delta U\|_F \leq cn^2\epsilon + O(\epsilon^2)$, and $\|\widehat{U} - U\|_F \leq \epsilon\|U\|_F \leq \sqrt{n}\epsilon + O(\epsilon^2)$. ∎

As explained in the introduction, applying multiplicative perturbation results to (4) yields relative error bounds on the singular values of order $O(\epsilon\kappa(A_N))$ and of order $O(\epsilon\kappa(A_N))$ divided by the relative gaps in the singular vectors. Thus, the magnitude of $\kappa(A_N)$ gives the relative accuracy of the computed SVD. In this respect, recall that $\kappa(A_N) \leq \sqrt{n}\min\kappa(D\,A)$, with D any diagonal matrix [21].

3. Backward error of a SVD algorithm for rank-revealing decompositions. A *rank-revealing decomposition* (RRD) [4] of $G \in \mathbb{R}^{m\times n}$, $m \geq n$, is a factorization $G = XDY^T$ with $D \in \mathbb{R}^{r\times r}$ diagonal and nonsingular and $X \in \mathbb{R}^{m\times r}$, $Y \in$

[2] Admittedly, it is not fully true that $\epsilon_J = O(\epsilon)$ with the meaning we have given to $O(\epsilon)$ in the *Notation*, since a dependence in the number of steps required for the convergence of the algorithm is hidden in the constant of the $O(\epsilon)$ (see [11, Proposition 3.13]). However, extensive numerical experience indicates that this dependence is polynomic in the dimensions of the problem.

[3] One can also show that $\|E_R\| \leq \sqrt{n}\,\epsilon_J\,\|A_N^{-1}\|$ in the spectral norm.

$\mathbb{R}^{n\times r}$, where both matrices X, Y have full column rank and are well-conditioned (notice that this implies $r = \mathrm{rank}(G)$). One of the most important contributions of Demmel et al. in [4] is developing algorithms which compute high relative accuracy SVDs for any matrix such that an RRD can be computed with enough accuracy. The accuracy required in the computed $\widehat{X}, \widehat{D}$ and $\widehat{Y}$ is the following (see [4, Theorem 2.1]):

1. each entry of D has small relative error,

$$|D_{ii} - \widehat{D}_{ii}| \leq O(\epsilon)|D_{ii}|, \tag{7}$$

2. $\widehat{X}$ and $\widehat{Y}$ have small norm errors,

$$\|X - \widehat{X}\| = O(\epsilon)\|X\| \quad \text{and} \quad \|Y - \widehat{Y}\| = O(\epsilon)\|Y\|. \tag{8}$$

Once the RRD is computed, Algorithms 3.1 or 3.2 in [4] can be used to compute the SVD with high relative accuracy. Both algorithms have as inputs the three factors, X, D and Y, of an RRD. The error bounds for the computed SVD presented in [4] for Algorithm 3.2 are better than those proved for Algorithm 3.1. However, the authors of [4] strongly recommend the use of Algorithm 3.1. The reasons are that Algorithm 3.1 is faster and that no significant difference in accuracy is observed in practice.

In this section we prove that Algorithm 3.1 in [4] produces a small backward multiplicative error when executed in finite precision arithmetic. This result is based on the proof of Theorem 3.1 in [4, section 3.2.1] and greatly clarifies the way in which the error bounds for the computed singular values and vectors are obtained in [4]. The error analysis done in [4] is backward multiplicative up to the one-sided Jacobi step of Algorithm 3.1 in [4]. From this point on the analysis is made in the forward sense and becomes quite involved. The crucial ingredient to get a multiplicative backward error result like Theorem 3.1 below is Theorem 2.1 for one-sided Jacobi proved in section 2.

Algorithm 1 below is the version of Algorithm 3.1 in [4] we analyze. We stress that the inputs for Algorithm 1 are the three matrices $X \in \mathbb{R}^{m\times r}$, $D \in \mathbb{R}^{r\times r}$, $Y \in \mathbb{R}^{n\times r}$ of a RRD. Moreover, the QR and LQ factorizations appearing in the Algorithm are economy size or reduced factorizations, i.e., if $C = QR$ is a $n \times r$ matrix ($n > r$), then Q is a $n \times r$ matrix with orthonormal columns.

Algorithm 1.

Input: rank-revealing decomposition, X, D, Y, of $G = XDY^T \in \mathbb{R}^{m\times n}$.
Output: singular value decomposition $U\Sigma V^T$ of G.

1. *Compute a QR decomposition with column pivoting, $XD = QRP$, of XD.*
2. *Compute the product $W = RPY^T$ using conventional matrix multiplication.*
3. *Compute a LQ decomposition $W = L_\omega Q_\omega^T$ of W.*
4. *Compute an SVD $L_\omega = U_\omega \Sigma V_\omega^T$ of L_ω using right-handed Jacobi.*
5. *Compute the products $U = QU_\omega$ and $V = Q_\omega V_\omega$. Strassen's method may be used.*

We should point out that this implementation differs from the one presented in [4]: here the Jacobi step is split in two stages, steps 3 and 4. This is recommended in [4, section 3.3] to reduce the computational cost of the one-sided Jacobi step, the most expensive one in the whole algorithm. This saving is clear if the rank r is less than n. In the case $r = n$, W is square and, at first glance, the computation of the

LQ factorization of W would increase the cost because right-handed Jacobi does not make any use of the triangular form of L_ω. However, if the LQ factorization of W is done with row pivoting, then numerical experience shows that more than one sweep is saved in right-handed Jacobi. This is enough to compensate the cost of the LQ factorization and makes step 3 of Algorithm 1 still interesting. Anyway, the reader can check that skipping step 3 above does not affect the error bounds in Theorem 3.1.

THEOREM 3.1. *Algorithm* 1 *produces a small multiplicative backward error; i.e., if* $\widehat{U}\widehat{\Sigma}\widehat{V}^T$ *is the SVD computed by the algorithm in finite arithmetic with machine precision* ϵ*, then there exist matrices* $U' \in \mathbb{R}^{m\times r}$, $V' \in \mathbb{R}^{n\times r}$, $E \in \mathbb{R}^{m\times m}$, $F \in \mathbb{R}^{n\times n}$ *such that* U' *and* V' *have orthonormal columns,*

$$\begin{aligned} \|U' - \widehat{U}\| = O(\epsilon), \quad \|V' - \widehat{V}\| = O(\epsilon), \\ \|E\| = O(\epsilon\kappa(X)), \quad \|F\| = O(\epsilon\kappa(R')\kappa(Y)), \end{aligned} \tag{9}$$

where R' *is the best conditioned row diagonal scaling of the triangular matrix* R *appearing in step* 1 *of Algorithm* 1 *and*

$$(I+E)G(I+F) = U'\widehat{\Sigma}V'^T. \tag{10}$$

Remark 1. It is proved in [4] that $\kappa(R')$ is at most of order $O(n^{3/2}\kappa(X))$, but in practice extensive numerical tests show that $\kappa(R')$ behaves as $O(n)$ [4, 9]. One can get rid of the factor $\kappa(R')$ at the price of using the more costly Algorithm 3.2 of [4]. The proof of this follows closely the proof of Theorem 3.1.

Proof. Since we will use results in [4, section 3.2.1], we need to match our notation with that of [4]: the matrices Q, W, R (and R') appearing in the proof, which are the computed ones, are named in the proof *without* hats. The rest of the computed matrices are denoted, as elsewhere in this paper, with their hats on.

It is shown in [4, p. 34] that, after step 2 of Algorithm 1, the matrix Q computed in step 1 and the matrix W computed in step 2 are such that

$$(I+E_1)G(I+F_1) = QW \tag{11}$$

for square matrices E_1, F_1 with

$$\|E_1\| = O(\epsilon\,\kappa(X)), \qquad \|F_1\| = O(\epsilon\,\kappa(R')\,\kappa(Y)).$$

Although the columns of the computed Q are not exactly orthonormal, it is well known [16, p. 360] that there exists a matrix Q' with orthonormal columns such that

$$Q = Q' + E_q = (I + E_q(Q')^T)Q', \tag{12}$$

with $\|E_q\| = O(\epsilon)$. Thus, (11) becomes $(I+E_1')G(I+F_1) = Q'W$, with $\|E_1'\| = O(\epsilon\,\kappa(X))$.

The LQ factorization of W in step 3 of Algorithm 1 is equivalent to computing a QR factorization of $W^T \in \mathbb{R}^{n\times r}$. The usual additive backward error analysis of the QR factorization, applied columnwise [16, p. 360], ensures that the computed $\widehat{L}_\omega$ satisfies

$$\widehat{L}_\omega(Q'_\omega)^T = (W + E_\omega),$$

where $Q'_\omega \in \mathbb{R}^{n\times r}$ is a matrix with orthonormal columns satisfying $\|Q'_\omega - \widehat{Q}_\omega\| = O(\epsilon)$ for the computed $\widehat{Q}_\omega$. The backward error E_ω satisfies the rowwise bound

$$\|E_\omega(i,:)\| = O(\varepsilon)\|W(i,:)\|, \qquad i = 1,\ldots,r. \tag{13}$$

If we write $W + E_\omega = W(I + W^\dagger E_\omega)$ multiplicatively, with $W^\dagger$ the pseudoinverse of W, then

$$W = \widehat{L}_\omega (Q'_\omega)^T (I + W^\dagger E_\omega)^{-1}.$$

Now, let $R' = (D')^{-1}R$ be the best conditioned row scaling of the triangular matrix R computed in step 1. In order to bound $\|W^\dagger E_\omega\|$, we define $Z = (D')^{-1}W$ and $E_z = (D')^{-1}E_\omega$. The equations (13) imply $\|E_z\| = O(\epsilon)\|Z\|$, and since both D' and Z have full rank, we obtain

$$\|W^\dagger E_\omega\| = \|Z^\dagger E_z\| = O(\epsilon)\kappa(Z) = O(\epsilon\kappa(R')\kappa(Y)).$$

The last equality above is a consequence of the first equation in [4, p. 34], which implies $\|(D')^{-1}\delta W\| = O(\epsilon)\|R'\|\,\|Y\|$ for the error δW in the matrix multiplication of step 2 of Algorithm 1. Therefore, since $Z = R'PY^T - (D')^{-1}\delta W$, we arrive at $\kappa(Z) \leq \kappa(R')\kappa(Y)(1 + O(\epsilon)\kappa(R')\kappa(Y))$.

Thus, upon completion of step 3 of Algorithm 1, we have

$$(I + E_2)G(I + F_2) = Q'\,\widehat{L}_\omega (Q'_\omega)^T \tag{14}$$

with $E_2 = E'_1$, $I + F_2 = (I + F_1)(I + W^\dagger E_\omega)$ and $\|F_2\| = O(\epsilon\kappa(R')\kappa(Y))$.

Now, Theorem 2.1 applied to step 4 ensures the existence of $r \times r$ matrices $\overline{U}'$, $\overline{V}'$, E_L, E_R with $\overline{U}'$, $\overline{V}'$ orthogonal,

$$\begin{aligned} &\|\overline{U}' - \widehat{U}_\omega\| = O(\epsilon), \qquad \|\overline{V}' - \widehat{V}_\omega\| = O(\epsilon) \\ &\|E_L\| \leq O(\epsilon), \qquad \|E_R\| \leq O(\epsilon\kappa((D')^{-1}\widehat{L}_\omega)), \end{aligned} \tag{15}$$

and

$$\widehat{L}_\omega = (I + E_L)\overline{U}'\widehat{\Sigma}(\overline{V}')^T(I + E_R), \tag{16}$$

where $\widehat{U}_\omega\widehat{\Sigma}\widehat{V}_\omega^T$ is the SVD computed by the right-handed Jacobi SVD algorithm on $\widehat{L}_\omega$. In (16), $(I + E_L)$ and $(I + E_R)$ appear in a different side than in (4). It is easy to see that this does not change the first order error bounds. Notice that we have replaced the unit row scaling of $\widehat{L}_\omega$ with the scaling given by $(D')^{-1}$. We can do this because the condition number of the former matrix is not larger than a factor $\sqrt{r}$ times the condition number of the latter [21]. Note also that $\kappa((D')^{-1}\widehat{L}_\omega) = \kappa((D')^{-1}\widehat{L}_\omega\,(Q'_\omega)^T) = \kappa(Z + E_z) = \kappa(Z)(1 + O(\epsilon)\kappa(Z))$. Hence,

$$\|E_R\| = O(\epsilon\kappa(R')\kappa(Y)).$$

Substituting (16) into (14) leads to

$$(I + E_3)G(I + F_3) = Q'\overline{U}'\widehat{\Sigma}(\overline{V}')^T(Q'_\omega)^T,$$

where $I + E_3 = (I + \widetilde{E}_L)^{-1}(I + E_2)$ and $I + F_3 = (I + F_2)(I + \widetilde{E}_R)^{-1}$ for $\widetilde{E}_L = Q'E_L(Q')^T$ and $\widetilde{E}_R = Q'_\omega E_R(Q'_\omega)^T$. Clearly, $\|E_3\| = O(\epsilon\kappa(X))$ and $\|F_3\| = O(\epsilon\kappa(R')\kappa(Y))$.

Finally, it only remains to show that $\widehat{U} = \texttt{fl}(Q\widehat{U}_\omega)$ and $\widehat{V} = \texttt{fl}(\widehat{Q}_\omega\widehat{V}_\omega)$ differ from $Q'\overline{U}'$ and $Q'_\omega\overline{V}'$ by $O(\epsilon)$. We show it for $\widehat{U}$; the argument for $\widehat{V}$ is analogous. Using (12) and (15), we obtain $Q\widehat{U}_\omega = Q'\overline{U}' + O(\epsilon)$. Moreover, the standard error

analysis for matrix multiplication implies that $\|\widehat{U} - Q\widehat{U}_\omega\|_F \leq r^2\epsilon + O(\epsilon^2)$. The proof is concluded by observing that $\|\widehat{U} - Q'\overline{U}'\|_F \leq \|\widehat{U} - Q\widehat{U}_\omega\|_F + \|Q\widehat{U}_\omega - Q'\overline{U}'\|_F$. ∎

Multiplicative perturbation theory for the SVD applied to (10) yields relative error bounds of order $O(\epsilon\,\kappa(R')\,\max(\kappa(X),\kappa(Y)))$ on the singular values and of order $O(\epsilon\kappa(R')\max(\kappa(X),\kappa(Y)))$ divided by the relative gaps on the singular vectors. These are the bounds previously obtained in [4, Theorem 3.1]. The backward multiplicative error (10) in Theorem 3.1 for Algorithm 1 can be easily combined with the backward multiplicative error coming from computing a RRD, with errors (7), (8), to produce an overall multiplicative backward error similar to (10) [9, section 2.1]. Other more general forward errors in the computation of a RRD can be managed in a similar way.

4. The left-handed version. The backward error analysis in section 3 has been performed assuming that right-handed Jacobi is employed in step 4 of Algorithm 1. However, it has been observed that Algorithm 1 with *left-handed* Jacobi on L_ω is usually much faster. For instance, for rank-revealing decompositions coming from quasi-Cauchy matrices, the following differences in computational cost (using double precision arithmetic) have been reported in [3, p. 572]: 50 Jacobi sweeps if right-handed Jacobi is used in step 4 and no more than 8 sweeps (4.6 on average) for the left-handed version. In the numerical experiments presented in [9, section 6.2] for random 100×100 matrices in RRD form, the average number of sweeps in the right version doubles the number of sweeps in the left version.[4] A heuristic reason of this significant difference in computational cost is that the rows of L_ω are usually closer to being orthogonal than its columns; thus left-handed Jacobi is expected to converge faster (see [20, p. 988] for a more detailed explanation of the advantages of one version of one-sided Jacobi over the other depending on the scaling). These discrepancies in speed make it interesting to undertake a brief analysis of the multiplicative backward stability properties of Algorithm 1 using left-handed Jacobi in step 4. Before we begin, it should be noted that all these remarks may be modified by future improvements in one-sided Jacobi SVD algorithms. According to numerical tests conducted using a preliminary version of the fast and sophisticated right-handed Jacobi routine which is being developed by Drmač, right-handed Jacobi could be much faster than the usual plain implementation of left-handed Jacobi.

The error bounds for left-handed Jacobi on an invertible matrix $A \in \mathbb{R}^{n\times n}$ remain as in Theorem 2.1, at the prize of replacing the $O(\epsilon\kappa(A_N))$ with $O(\epsilon\gamma)$, where

$$\gamma = \max_{i=0,1,\ldots,q} \kappa(B_i). \tag{17}$$

Here, each B_i is the diagonal scaling with unit rows of the matrix $A_i = D_i\,B_i$ ($A_0 = A$) resulting from the action of the ith finite precision rotation along the process of left-handed Jacobi, and A_q is the first iterate satisfying the stopping criterion

$$\max_{i\neq j} \mathtt{fl}\left(\frac{|A_q(i,:)A_q(j,:)^T|}{\|A_q(i,:)\|\,\|A_q(j,:)\|}\right) \leq n\,\epsilon \qquad \text{for} \quad i \neq j. \tag{18}$$

To explain the origin of the additional factor γ, notice that, according to [6, Theorem 4.1], if A_i (resp., A_{i+1}) is the matrix obtained after the ith (resp., $(i+1)$th) finite

[4] Both in [3] and in [9] Algorithm 1 runs on square matrices and has been implemented without step 3. If step 3 is done with row pivoting, then right-handed Jacobi can improve its speed by more than one sweep, but this is not enough to wipe out the differences with the left-handed version.

precision rotation, then A_{i+1} can be written as

$$A_{i+1} = R_{i+1}(A_i + \delta A_i),$$

where R_{i+1} is an exact rotation and the backward error δA_i is such that $\|\delta B_i\| \leq 72\epsilon + O(\epsilon^2)$ for the row scaling $\delta A_i = D_i\, \delta B_i$, where D_i is the diagonal matrix with the row norms of A_i on the diagonal. Hence,

$$A_{i+1} = R_{i+1} A_i (I + E_i)$$

with $\|E_i\| = \|A_i^{-1}\delta A_i\| = \|B_i^{-1}\delta B_i\| \leq (72\epsilon + O(\epsilon^2))\kappa(B_i)$. Notice that replacing $\|B_i^{-1}\|$ with $\kappa(B_i)$ increases the bound at most by a factor $\sqrt{n}$.

Repeating the argument for all q rotations up to convergence, one obtains

$$A_q = (\widetilde{U}')^T A(I + \widetilde{E})$$

for an exact orthogonal matrix $\widetilde{U}'$ and a matrix $\widetilde{E}$ such that $\|\widetilde{E}\| \leq (72\epsilon + O(\epsilon^2))q\gamma$, with γ given by (17). The constant q in the previous error bound is pessimistic, and in fact with a finer implementation of left-handed Jacobi q can be replaced by $(s-1)p$, where s is the number of sweeps up to convergence, each of them implemented in p parallell steps [11].

Using the stopping criterion as in the end of the proof of Theorem 2.1 shows that if $\widehat{U}\widehat{\Sigma}\widehat{V}^T$ is the SVD computed by left-handed Jacobi on A with stopping criterion (18), then

$$A(I + \widetilde{E}_R) = \widetilde{U}'\widehat{\Sigma}\widetilde{V}'^T$$

for orthogonal matrices $\widetilde{U}'$, $\widetilde{V}'$ within a distance $O(\epsilon)$ of $\widehat{U}$, $\widehat{V}$, and

$$\|\widetilde{E}_R\| \leq 72\epsilon q\gamma + cn^2\epsilon + O(\epsilon^2) = O(\epsilon\gamma).$$

This last bound makes explicit the proviso needed in [6] to guarantee that one-sided Jacobi is able to compute the SVD with high relative accuracy for matrices of the form DB, where D is diagonal and B is well-conditioned: γ cannot be much larger than $\kappa(B)$.

Plugging these backward errors into the proof of Theorem 3.1, we obtain for the left-handed version of Algorithm 1 (i.e., the one using left-handed Jacobi in step 4) the backward error bound

$$(I + \widetilde{E})G(I + \widetilde{F}) = U'\widehat{\Sigma}V'^T,$$

where, as in Theorem 3.1, U' and V' have orthonormal columns,

$$\|U' - \widehat{U}\| = O(\epsilon), \quad \|V' - \widehat{V}\| = O(\epsilon)$$

for the computed matrices $\widehat{U}$, $\widehat{\Sigma}$, $\widehat{V}$, and the backward errors satisfy

$$\|\widetilde{E}\| = O(\epsilon\kappa(X)), \quad \|\widetilde{F}\| = O(\epsilon \max\{\gamma, \kappa(R')\kappa(Y)\}),$$

with γ being the constant defined in (17) for left-handed Jacobi on the matrix $\widehat{L}_\omega$ computed in step 3 of Algorithm 1. Therefore, the error bounds for this left-handed version of Algorithm 1 are larger than those for the right-handed one. Only if γ is of the order $O(\kappa(R')\kappa(Y))$ the same accuracy will be achieved. It is claimed in [6] that there is strong numerical evidence of $\gamma/\kappa(B_0) \approx 1$. This has also been observed in the numerical experiments done in [9]. Hence, it seems that the increase in speed of the left-handed version is not penalized by a loss of accuracy.

Acknowledgment. The authors thank Prof. Zlatko Drmač for providing the source code for his state-of-the-art implementation of the one-sided Jacobi SVD routine, and also for many illuminating discussions on one-sided Jacobi SVD algorithms.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK User's Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[3] J. Demmel, *Accurate singular value decompositions of structured matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 562–580.

[4] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[5] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 873–912.

[6] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[7] J. Demmel and P. Koev, *Accurate SVDs of polynomial Vandermonde matrices involving orthonormal polynomials*, Linear Algebra Appl., to appear.

[8] J. Demmel and P. Koev, *Accurate SVDs of weakly diagonally dominant M-matrices*, Numer. Math., to appear.

[9] F. M. Dopico, J. M. Molera, and J. Moro, *An orthogonal high relative accuracy algorithm for the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 25 (2003), pp 301–351.

[10] Z. Drmač, *Implementation of Jacobi rotations for accurate singular value computation in floating-point arithmetic*, SIAM J. Sci. Comput., 18 (1997), pp. 1200–1222.

[11] Z. Drmač, *Accurate computation of the product-induced singular value decomposition with applications*, SIAM J. Numer. Anal., 35 (1998), pp. 1969–1994.

[12] Z. Drmač, *A posteriori computation of the singular vectors in a preconditioned Jacobi SVD algorithm*, IMA J. Numer. Anal., 19 (1999), pp. 191–213.

[13] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[14] K. Fernando and B. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.

[16] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[17] I. Ipsen, *Relative perturbation results for matrix eigenvalues and singular values*, Acta Numerica (1998), pp. 151–201.

[18] R.-C. Li, *Relative perturbation theory:* I. *Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.

[19] R.-C. Li, *Relative perturbation theory:* II. *Eigenspace and singular subspace variations*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 471–492.

[20] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.

[21] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

[22] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1032–1045

CONNECTIONS BETWEEN REALIZATIONS OF A TRANSFER FUNCTION*

K.-H. FÖRSTER† AND B. NAGY‡

Abstract. It is one of the basic facts of linear time-invariant systems theory that any two minimal (canonical) realizations are connected in the best possible way: by system similarity. We study five different types of possible connections between two arbitrary realizations of a transfer function, and are interested in questions of existence (sufficient and/or necessary conditions), uniqueness, and description of all (or of a possibly large class of) connecting operators or pairs of operators. In the case of the existence of nonnegative realizations we seek nonnegative connecting pairs or operators.

Key words. linear time-invariant systems, controllability, observability, intertwining operators, generalized inverses

AMS subject classifications. 15A09, 93B05, 93B07, 93B17

DOI. 10.1137/S0895479802410013

1. Introduction. Let $\mathbf{F}$ denote a field. Let W be a (strictly proper) rational matrix function mapping the input space $\mathbf{F}^b$ into the output space $\mathbf{F}^c$, and let (C_j, A_j, B_j) $(j = 1, 2)$ be two realizations of W; i.e., let

$$W(z) = C_j(zI - A_j)^{-1}B_j \equiv C_j(z - A_j)^{-1}B_j \quad (j = 1, 2).$$

We shall assume that the (finite) quadratic matrices A_j have orders n_j (which are not less than the McMillan degree n_0 of the rational matrix function W). In the classical theory of linear time-invariant finite dimensional systems W is the transfer function of the system, and any realization of order (by this we mean the dimension of A_j) n_0 is a minimal or canonical realization of W. One of the basic facts of this theory (cf., e.g., the already classical works by Kalman, Falb, and Arbib [14] and Kalman [13]) is that for every pair of minimal realizations of W there is a unique invertible matrix $T : \mathbf{F}^{n_0} \to \mathbf{F}^{n_0}$ establishing system similarity between the two minimal realizations (the exact definition will be reproduced in Definition 2.1).

System similarity is clearly the best possible connection between two realizations of W (having necessarily the same orders). The action of the general linear group of a given order determines the orbit of a given realization, and natural and important questions related to this action have been studied thoroughly (cf., e.g., [1], [2], [4], [5]). The purpose of this paper is different. We consider (see Definition 2.1) at least four weaker *types of connections* between two realizations of a transfer function: strong, end, weak connectedness by an operator (matrix) T, and weak connectedness by a pair (T, S). Also, in some cases we shall be interested in the *properties of the connecting* (we could also say: intertwining) *operators*.

Probably the most important variant of this notion is *strong* connectedness, and it has implicitly played a useful role in several more or less classical situations.

*Received by the editors June 19, 2002; accepted for publication (in revised form) by U. Helmke September 9, 2003; published electronically June 4, 2004.

http://www.siam.org/journals/simax/25-4/41001.html

†Department of Mathematics, Technical University of Berlin, Sekr. MA 6-4, Straße des 17. Juni 136, D-10623 Berlin, Germany (foerster@math.tu-berlin.de).

‡Department of Analysis, Institute of Mathematics, Budapest University of Technology and Economics, H-1521 Budapest, Hungary (bnagy@math.bme.hu). The work of this author was supported by Hungarian National Scientific Grant OTKA T-030042.

For example, let (C, A, B) be an arbitrary realization of W, and let $K(C, A)$ denote the unobservable subspace of this realization; i.e., let

$$K(C, A) := \cap_{k=0}^{\infty} \ker(CA^k).$$

Let X denote the state space of this realization (i.e., the space of A), let $X_2 := X/K(C, A)$ be the corresponding factor (quotient) space, and let $P : X \to X_2$ be the canonical projection. Then the induced map $A_2 : X_2 \to X_2$ is well defined; further, there exist operators C_2, B_2 such that the "*observable factor realization*" (C_2, A_2, B_2) is also a realization of W, and the projection P strongly connects the original realization to the factor realization in the sense of Definition 2.1 (cf. [18, p. 131]):

$$\begin{array}{ccccccc} & \overset{C}{\swarrow} & X & \overset{A}{\longleftarrow} & X & \overset{B}{\nwarrow} & \\ \mathbf{F}^c & & \downarrow P & & \downarrow P & & \mathbf{F}^b. \\ & \underset{C_2}{\nwarrow} & X_2 & \underset{A_2}{\longleftarrow} & X_2 & \underset{B_2}{\swarrow} & \end{array} \tag{1.1}$$

For applications of this construction see also [18, pp. 52–54, p. 281].

The problem of when an *arbitrary* realization (C, A, B) of a transfer function W is strongly connected to some (and then every) *minimal* realization (C_0, A_0, B_0) of W has been solved completely (but not using this terminology) by the present authors in [7, Proposition 4.1]: it is the case exactly when the state space X is a (not necessarily direct) sum

$$X = K(C, A) + J(A, B),$$

where $J(A, B) := \operatorname{span}(\cup_{k=0}^{\infty} A^k B)$ is the smallest A-invariant subspace of X containing $\operatorname{im} B$, the range space of B. For the applications of this result for the theory (and practice) of nonnegative realizations of a transfer function we refer the reader to [7] and [6].

Further, recall that Byrnes and Hurt [2, p. 89] call (in a different terminology) realization 1 *a simulation* of realization 2 if (in our terminology) 1 is strongly connected to 2 (in the direction $1 \to 2$).

Note that considerations and constructions very close to the problem of strong connectedness are contained, e.g., in the papers [12] and [3]. The first one gives a very good motivation why the notions of *inclusion, contraction, expansion, restriction, aggregation* (of systems or, more exactly, realizations) are useful and important in engineering and economical applications, and [3] is also a good review of recent developments using these concepts. Strong connectedness in our sense is a generalization of the notions of restriction and aggregation (between realizations).

We want to fix some *terminology and notation.* For a transfer function W we shall say that W maps $\mathbf{F}^b$ into $\mathbf{F}^c$ or write, equivalently, $W : \mathbf{F}^b \to \mathbf{F}^c$. Similar notation will be used in section 3 when the ground field $\mathbf{F}$ is the field $\mathbf{R}$ of the reals. We shall consider only finite dimensional realizations. For any realization (C, A, B) of W the dimension of the space of A (the state space) is the *order* of the realization. The order of any minimal (order) realization is the *McMillan degree* of W, and here will usually be denoted by n_0. Speaking about different realizations, for short we shall use the expressions realization 1, realization 2, etc., whenever no misunderstanding is possible.

We shall have no reason for considering special bases in any finite dimensional space, so we always shall work with fixed bases. Accordingly, we shall use the words

operator and matrix interchangeably. When working with dual operators and dual realizations, we shall use the dual bases. For the properties of the used generalized inverses we refer to [16]. Note that for a given right inverse T^{-R} of T we obtain a *corresponding left inverse of* T^* by defining

$$(T^*)^{-L} := (T^{-R})^*,$$

and conversely. $\operatorname{im} T = \operatorname{colsp} T$, $\ker T$, $\operatorname{rowsp} T$ will denote the range space, kernel, and row space of the operator (matrix) T, respectively. For the Kronecker (tensor) product $\otimes$ of two matrices and the notation $\operatorname{vec} T$ (stacking the elements of the matrix T in lexicographical order into a single column vector) see, e.g., [11]. The set of $n \times k$ matrices is denoted by $M(n \times k)$, and the transpose of the matrix A by A^t.

For the basics on positive (i.e., nonnegative) realizations of a transfer function we refer the reader, e.g., to [7] or [6].

2. Connections between realizations

DEFINITION 2.1. *Let (C_j, A_j, B_j) be two realizations of the (strictly proper) rational matrix function W mapping $\mathbf{F}^b$ into $\mathbf{F}^c$ and having the orders*

$$n_j := \dim A_j \geq n_0 \quad (j = 1, 2),$$

where n_0 is the order of any minimal realization of W. Assume that there exist two linear operators $T, S : \mathbf{F}^{n_1} \to \mathbf{F}^{n_2}$ such that (at least) one of the following properties 1–5 *holds for every nonnegative integer k by referring to the diagram(s) below:*

$$\begin{array}{ccccccc} & {\scriptstyle C_1} & & {\scriptstyle A_1^k} & & {\scriptstyle B_1} & \\ & \swarrow & \mathbf{F}^{n_1} & \longleftarrow & \mathbf{F}^{n_1} & \nwarrow & \\ \mathbf{F}^c & & \downarrow T & & \downarrow S & & \mathbf{F}^b. \\ & \nwarrow & \mathbf{F}^{n_2} & \longleftarrow & \mathbf{F}^{n_2} & \swarrow & \\ & {\scriptstyle C_2} & & {\scriptstyle A_2^k} & & {\scriptstyle B_2} & \end{array} \tag{2.1}$$

The realizations (with subscripts 1 *and* 2 *(in the indicated order) are called*

1. *system similar (by T) if and only if $T = S$, the (two-sided) inverse operator T^{-1} exists (hence $n_1 = n_2$), and the diagrams above commute (in the usual sense),*

2. *strongly connected by T if and only if $T = S$, and the diagrams above commute,*

3. *end connected by T if and only if $T = S$, and the extremal triangles in the diagram above commute in the usual sense,*

4. *weakly connected by T or, equivalently, by the pair (T,T) if and only if $T = S$, and the diagrams above commute between the extreme spaces in the sense that all four ways (paths) between $\mathbf{F}^b$ and $\mathbf{F}^c$ yield the same operators,*

5. *weakly connected by the pair (T,S) if and only if the diagrams above commute between the extreme spaces as in property* 4.

Equivalently, we shall also say that realization 1 *is weakly connected to realization* 2 *by T, etc. Finally, in some cases we may and shall omit the qualifiers "by T," etc.*

Remark 2.2. Clearly, each property above implies the lower ones in the list. It can be shown by examples that the converse is not valid for any pair of the first four properties. We shall prove that each pair of realizations of any given transfer function has property 4. Nevertheless, property 5 can be of considerable significance if the classes of considered realizations or/and of connecting operators is restricted.

LEMMA 2.3. *Assume that the realizations* 1 *and* 2 *have orders n_j $(j = 1, 2)$, let $n := \max[n_1, n_2]$, and define the controllability and the observability operators*

$$G_j := \begin{pmatrix} B_j & A_j B_j & \cdots & A_j^{n-1} B_j \end{pmatrix} : \mathbf{F}^{bn} \to \mathbf{F}^{n_j}, \tag{2.2}$$

$$M_j := \begin{pmatrix} C_j \\ C_jA_j \\ \vdots \\ C_jA_j^{n-1} \end{pmatrix} : \mathbf{F}^{n_j} \to \mathbf{F}^{cn} \tag{2.3}$$

$(j = 1, 2)$, *respectively.*

(1) *Assume that realization* 1 *is controllable. Then the controllability operator* G_1 *has a right inverse* G_1^{-R}, *and for every such right inverse the corresponding operator* T *(i.e., the pair* (T, T)*) defined by*

$$T = T(G_1^{-R}) := G_2G_1^{-R}$$

weakly connects realization 1 *to* 2. *Further, the diagrams* (2.1) *(with* $S := T$*) commute for every nonnegative integer* k *between each space* $\mathbf{F}^{n_1}$ *and* $\mathbf{F}^c$*; hence the left-hand side triangle commutes in the usual sense.*

(2) *Assume that realization* 2 *is observable. Then the observability operator* M_2 *has a left inverse* M_2^{-L}. *These left inverses are exactly the dual (adjoint) operators of corresponding right inverses* $[G^*]_2^{-R}$ *of the controllability operator* $[G^*]_2$ *for the dual realization* (B_2^*, A_2^*, C_2^*). *Define*

$$U : \mathbf{F}^{n_2} \to \mathbf{F}^{n_1}, \qquad U = U\left([G^*]_2^{-R}\right) := [G^*]_1[G^*]_2^{-R}.$$

Then U *connects realization* 2^* *to realization* 1^* *exactly as* T *did realization* 1 *to realization* 2 *in part* (1), *and (with the corresponding left inverse* M_2^{-L}*) we have*

$$U^* = M_2^{-L}M_1.$$

Proof. (1) The assumptions imply that $\operatorname{im} G_1$ is the entire space $\mathbf{C}^{n_1}$; hence the stated right inverses exist. From the equality of the corresponding Markov coefficients in both realizations we obtain

$$C_2A_2^kT = C_2A_2^kG_2G_1^{-R} = C_1A_1^kG_1G_1^{-R} = C_1A_1^k$$

for every nonnegative integer k. For $k = 0$ we have obtained the stated commutativity of the left-hand side triangle, and this implies for every nonnegative integer k that $C_2TA_1^k = C_1A_1^k$, i.e., the commutativity *between the spaces* $\mathbf{F}^{n_1}$ *and* $\mathbf{F}^c$. Further,

$$M_2TG_1 = M_1G_1 = M_2G_2;$$

hence $M_2(TG_1 - G_2) = 0$. Postmultiplication of the column block matrix M_2 by the first block column of $TG_1 - G_2$, i.e., by $TB_1 - B_2$, yields

$$C_2A_2^k(TB_1 - B_2) = 0 \quad (k = 0, 1, 2, \dots).$$

Hence the diagrams commute between the extreme spaces for every nonnegative integer k, and the proof is complete.

(2) It is well known that a realization is observable if and only if its dual is controllable. Hence for the dual diagrams we can apply part (1) and obtain that the diagrams

$$\begin{array}{ccccccc} & \overset{C_1^*}{\nearrow} & \mathbf{F}^{n_1} & \overset{A_1^{*k}}{\longrightarrow} & \mathbf{F}^{n_1} & \overset{B_1^*}{\searrow} & \\ \mathbf{F}^c & & \uparrow U & & \uparrow U & & \mathbf{F}^b \\ & \underset{C_2^*}{\searrow} & \mathbf{F}^{n_2} & \underset{A_2^{*k}}{\longrightarrow} & \mathbf{F}^{n_2} & \underset{B_2^*}{\nearrow} & \end{array}$$

commute for every nonnegative integer k between any space $\mathbf{F}^{n_2}$ and $\mathbf{F}^b$, and between the extreme spaces, whereas the right-hand side triangle commutes in the usual sense. Returning again to realizations 1 and 2 (by reversing the diagrams above), we make use of the equalities

$$[G^*]_j^* = \begin{pmatrix} C_j^* & \cdots & (A_j^*)^{n-1} C_j{}^* \end{pmatrix}^* = M_j \quad (j = 1, 2).$$

Hence

$$[G^*]_2^{-R} = [M_2^*]^{-R} = [M_2^{-L}]^*.$$

Here, in the last equality, the corresponding right and left inverses are understood. Hence we obtain the formula for U^*. □

COROLLARY 2.4. *Let (C, A, B) be an arbitrary realization of a rational matrix function W such that $n = \dim A \geq n_0$ (the McMillan degree of W), and let (C_0, A_0, B_0) be any minimal realization of W. Then there are two operators U and T such that for every nonnegative integer k the following diagrams commute between the extreme spaces*:

$$\begin{array}{ccccccc} & \overset{C}{\swarrow} & \mathbf{F}^n & \overset{A^k}{\longleftarrow} & \mathbf{F}^n & \overset{B}{\nwarrow} & \\ \mathbf{F}^c & & \downarrow U & & \downarrow U & & \mathbf{F}^b, \\ & \underset{C_0}{\nwarrow} & \mathbf{F}^{n_0} & \underset{A_0^k}{\longleftarrow} & \mathbf{F}^{n_0} & \underset{B_0}{\swarrow} & \end{array}$$

and the second diagram is

$$\begin{array}{ccccccc} & \overset{C}{\swarrow} & \mathbf{F}^n & \overset{A^k}{\longleftarrow} & \mathbf{F}^n & \overset{B}{\nwarrow} & \\ \mathbf{F}^c & & \uparrow T & & \uparrow T & & \mathbf{F}^b. \\ & \underset{C_0}{\nwarrow} & \mathbf{F}^{n_0} & \underset{A_0^k}{\longleftarrow} & \mathbf{F}^{n_0} & \underset{B_0}{\swarrow} & \end{array}$$

Moreover, for every nonnegative integer k the first diagrams commute between $\mathbf{F}^b$ and each space $\mathbf{F}^{n_0}$, and the second diagrams commute between each space $\mathbf{F}^{n_0}$ and $\mathbf{F}^c$. Hence in the first diagram the right-hand side triangle, and in the second diagram the left-hand side triangle are properly commutative.

Proof. The statements are straightforward consequences of Lemma 2.3. We check only one of them: we have seen that for the dual diagrams, writing S here instead of the operator U there,

$$B_0^* A_0^{*k} = B^* A^{*k} S \quad (k = 0, 1, 2, \dots)$$

hold. Taking the duals, we obtain

$$A_0^k B_0 = S^* A^k B \quad (k = 0, 1, 2, \dots).$$

We now define our operator U by $U := S^*$ and obtain similarly the other stated commutativity properties. □

THEOREM 2.5. *Assume that (C_j, A_j, B_j) $(j = 1, 2)$ are two realizations of the same rational matrix function W with orders $n_j \geq n_0$. Then there is an operator V weakly connecting realization 1 to realization 2.*

Proof. Let (C_0, A_0, B_0) be any minimal realization of the matrix function W. By Corollary 2.4, there are two operators U and T such that the upper and the lower halves of the following diagram commute between the extreme spaces, and the lower

left and upper right triangles are properly commutative (for each nonnegative integer k):

$$\begin{array}{ccccc} & \overset{C_1}{\swarrow} & \mathbf{F}^{n_1} & \overset{A_1^k}{\longleftarrow} & \mathbf{F}^{n_1} & \overset{B_1}{\nwarrow} & \\ & & \downarrow U & & \downarrow U & & \\ \mathbf{F}^c & \overset{}{\longleftarrow} & \mathbf{F}^{n_0} & \overset{}{\longleftarrow} & \mathbf{F}^{n_0} & \overset{}{\longleftarrow} & \mathbf{F}^b. \\ & \underset{C_0}{} & \downarrow T & \underset{A_0^k}{} & \downarrow T & \underset{B_0}{} & \\ & \underset{C_2}{\nwarrow} & \mathbf{F}^{n_2} & \underset{A_2^k}{\longleftarrow} & \mathbf{F}^{n_2} & \underset{B_2}{\swarrow} & \end{array}$$

Define $V := TU$. Making use of the commutativity of the mentioned triangles, we obtain

$$C_2VA_1^kB_1 = C_2TUA_1^kB_1 = C_0UA_1^kB_1 = C_1A_1^kB_1,$$

$$C_2A_2^kVB_1 = C_2A_2^kTUB_1 = C_2A_2^kTB_0 = C_2A_2^kB_2.$$

In both lines above in the last equalities we have used the commutativity between the extreme spaces in the upper and in the lower halves of the diagram. □

Recall that for any linear operator $L : X \to Y$ acting between finite dimensional spaces the so-called *pseudoinverse(s)* $L^- : Y \to X$ always exist and are defined by the property

$$LL^-L = L.$$

THEOREM 2.6. *Let (C_j, A_j, B_j) $(j = 1, 2)$ be two realizations of the same transfer function W mapping $\mathbf{F}^b$ into $\mathbf{F}^c$ and having the orders*

$$n_1 := \dim A_1, \qquad n_2 := \dim A_2 \geq n_0,$$

where n_0 is the McMillan degree of W. Then to every pair of pseudoinverses of the controllability operator G_1 and the observability operator M_2 there correspond two linear operators $(T, S) : \mathbf{F}^{n_1} \to \mathbf{F}^{n_2}$ defined by

$$S := M_2^-M_1, \qquad T := G_2G_1^-$$

such that for every nonnegative integer k the diagrams (2.1) *commute between $\mathbf{F}^b$ and $\mathbf{F}^c$; i.e., realization* 1 *is weakly connected to realization* 2 *by the pair (T, S).*

Proof. Let $n := \max[n_1, n_2]$. Consider the observability operators (2.3) and the controllability operators (2.2). For any fixed pair (G_1^-, M_2^-) define the linear operators (T, S) as above. Since $M_1B_1 = M_2B_2$, we obtain that

$$M_2SB_1 = M_2M_2^-M_1B_1 = M_2M_2^-M_2B_2 = M_2B_2.$$

In view of the definition of M_2 we have proved that the lowest and the right-hand side ways $(C_2A_2^kSB_1)$ in the diagrams commute for $k = 0, 1, \ldots, n-1$. Since $n \geq n_j$, for the larger values of k this commutativity follows from the Cayley–Hamilton theorem. Since $C_1G_1 = C_2G_2$, we also obtain

$$C_2TG_1 = C_2G_2G_1^-G_1 = C_1G_1G_1^-G_1 = C_1G_1.$$

The definition of the operator G_1 shows that the uppermost and the left-hand side ways $(C_2TA_1^kB_1)$ in the diagrams commute for $k = 0, 1, \ldots, n-1$, and the Cayley–Hamilton theorem yields this for the larger values of k. It is well known that the

uppermost and the lowest ways in the diagrams commute; hence so do all four ways for every nonnegative integer k. □

Next we shall give a characterization of the *strong connectedness* of an arbitrary pair of realizations in terms of the solvability of a system of linear equations.

THEOREM 2.7. *Realization* 1 *is strongly connected to realization* 2 *by the matrix* T *if and only if the vector* $\operatorname{vec} T$ *solves the following system of linear equations:*

$$\begin{pmatrix} I_{n_1} \otimes C_2 \\ I_{n_1} \otimes A_2 - A_1^t \otimes I_{n_2} \\ B_1^t \otimes I_{n_2} \end{pmatrix} \operatorname{vec} T = \begin{pmatrix} \operatorname{vec} C_1 \\ 0 \\ \operatorname{vec} B_2 \end{pmatrix}.$$

(For the Kronecker product $\otimes$ *and the* $\operatorname{vec} T$ *notation see, e.g.,* [11, pp. 242ff.]. *Further,* I *denote identities of the given dimensions.)*

Proof. The matrix T strongly connects realization 1 to 2 if and only if T satisfies all the following equations:

$$C_2T = C_1, \qquad A_2T - TA_1 = 0, \qquad TB_1 = B_2.$$

According to [11, p. 255], this holds if and only if

$$(I \otimes C_2) \operatorname{vec} T = \operatorname{vec} C_1, \quad (I \otimes A_2 - A_1^t \otimes I) \operatorname{vec} T = 0, \quad (B_1^t \otimes I) \operatorname{vec} T = \operatorname{vec} B_2,$$

where the identities I have the dimensions indicated in the statement of the theorem. The assertions then follow. □

THEOREM 2.8. *Consider two realizations of the same transfer function.*

(1) *If realization* 1 *is strongly connected to realization* 2*, then for the controllability and observability operators of Lemma* 2.3 *we have*

$$\operatorname{colsp} M_1 \subset \operatorname{colsp} M_2, \qquad \operatorname{rowsp} G_2 \subset \operatorname{rowsp} G_1. \tag{2.4}$$

(2) *If* (2.4) *holds and either realization* 1 *is controllable or realization* 2 *is observable, then* 1 *is strongly connected to* 2.

Proof. (1) Assume that realization 1 is strongly connected to 2 by T. Then T solves the system of matrix equations

$$M_2T = M_1, \qquad TG_1 = G_2;$$

hence (2.4) is satisfied.

(2) Assume that (2.4) holds. Then there are matrices R, S such that

$$M_2R = M_1, \qquad SG_1 = G_2.$$

Further, we clearly have $M_1G_1 = M_2G_2$. Therefore, by [16, Theorem 2.3.3], there is a matrix T satisfying

$$M_2T = M_1, \qquad TG_1 = G_2.$$

It follows that $C_2T = C_1$, $TB_1 = B_2$, and also

$$A_2^kTB_1 = A_2^kB_2 = TA_1^kB_1 \quad (k = 0, 1, \dots, n-1).$$

Since n is not less than the degree of the characteristic polynomial of A_i $(i = 1, 2)$, the above equality holds for every nonnegative integer k. Then for every $j = 0, 1, \dots, n-1$ we have

$$(A_2T - TA_1)A_1^jB_1 = A_2TA_1^jB_1 - TA_1^{j+1}B_1 = (A_2^{j+1}T - TA_1^{j+1})B_1 = 0.$$

Assume now that realization 1 is controllable (the proof assuming that realization 2 is observable is completely similar). By our assumption, then $A_2T = TA_1$; thus T connects realization 1 to 2. □

A necessary and sufficient condition for the *strong connectedness* of *certain* pairs of realizations was given in [7, Proposition 4.1]. We cite it here (without a proof) using our terminology.

PROPOSITION [7]. *An arbitrary realization (C, A, B) is strongly connected to any minimal realization (C_0, A_0, B_0) of the same transfer function W if and only if*

$$J(A, B) + K(C, A) = \mathbf{F}^n,$$

where n is the dimension of the space of A, the sum is not necessarily direct, and (cf. the introduction) $J(A, B)$ and $K(C, A)$ denote the smallest A-invariant subspace containing the range space of B and the largest A-invariant subspace in the kernel of C, respectively.

The following equivalent characterization may also be useful.

PROPOSITION 2.9. *An arbitrary realization (C, A, B) of order n is strongly connected to any minimal realization (C_0, A_0, B_0) of the same transfer function W if and only if one of the following two equivalent statements holds:*

(1) $\dim K(C, A) = n - n_0$,

(2) $\dim K(C, A) \geq n - n_0$,

where n_0 denotes the McMillan degree of the transfer function.

Proof. (1) clearly implies (2). Assume (2), and consider one canonical tri-invariant decomposition (a variant of the Kalman decomposition, cf. also [8, pp. 214–216]) of the realization (C, A, B), i.e., a direct sum decomposition $\mathbf{F}^n = L \oplus M \oplus N$, with respect to which

$$C = \begin{pmatrix} 0 & C_2 & C_3 \end{pmatrix}, \qquad A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ B_2 \\ 0 \end{pmatrix}.$$

Then the direct sum $L \oplus M$ yields the subspace $K(C, A) + J(A, B)$. The first subspace L is known to be $K(C, A)$, in our case with dimension $\geq n - n_0$. The second subspace M of the direct sum is known to have dimension n_0. Hence the subspace $K(C, A) + J(A, B) = K(C, A) \oplus M$ has dimension $\geq n$; i.e., the (equivalent) statements of Proposition [7] hold. Assuming the latter, the tri-invariant decomposition

$$\mathbf{F}^n = L \oplus M \oplus N = [K(C, A) + J(A, B)] \oplus N$$

shows that the third direct summand N is $\{0\}$, and we know that the second (M) must have dimension n_0. Hence the dimension of the first direct summand, i.e., of $K(C, A)$, is $n - n_0$, so (1) follows. □

By duality, we now obtain the following.

PROPOSITION 2.10. *A minimal realization (C_0, A_0, B_0) is strongly connected to a realization (C, A, B) of order n of the same rational matrix function W if and only if (with the above notation)*

$$J(A, B) \cap K(C, A) = \{0\}.$$

Proof. A linear operator T strongly connects the given minimal realization to the other one if and only if the dual operator T^* strongly connects the dual realization

(B^*, A^*, C^*) to the dual (minimal) realization (B_0^*, A_0^*, C_0^*). By the cited Proposition [7], this is the case if and only if

$$J(A^*, C^*) + K(B^*, A^*) = \mathbf{F}^n.$$

The two subspaces on the left-hand side are the annihilators of $K(C, A)$ and of $J(A, B)$, respectively (cf. [8, sections 2.7 and 2.8]). Hence, with the notation $^\perp$ for the annihilator, the above equality is the same as

$$K(C, A)^\perp + J(A, B)^\perp = \mathbf{F}^n.$$

This holds if and only if the assertion of the proposition holds. □

The same duality considerations or the canonical tri-invariant decomposition yield the following proposition.

PROPOSITION 2.11. *A minimal realization (C_0, A_0, B_0) is strongly connected to a realization (C, A, B) of the same rational matrix function W if and only if one of the following two equivalent statements holds:*

(1) $\dim J(A, B) = n_0$,

(2) $\dim J(A, B) \le n_0$.

The following corollary contains the "if statements" of Propositions [7] and 2.10. On the other hand, its proof is an evident combination of them.

COROLLARY 2.12. *Assume that (C_j, A_j, B_j) $(j = 1, 2)$ are two realizations of orders n_j of the same rational matrix function W and that*

$$J(A_1, B_1) + K(C_1, A_1) = \mathbf{F}^{n_1}, \qquad J(A_2, B_2) \cap K(C_2, A_2) = \{0\}.$$

Then realization 1 *is strongly connected to realization* 2.

The assumptions of the following theorem are stronger than those of the corollary above. Accordingly, the implication is also stronger (uniqueness), and we shall prove it independently of the propositions above, with the help of Lemma 2.3.

THEOREM 2.13. *Apply the notation of Lemma* 2.3.

(1) *If realization* 1 *is controllable and realization* 2 *is observable, then realizations* 1 *and* 2 *are strongly connected by a unique operator T defined by*

$$T := G_2 G_1^{-R} = M_2^{-L} M_1.$$

Here any right inverse or any left inverse yields the same operator T.

(2) *If realization* 1 *is controllable and realization* 2 *is observable, then the operator T defined above is invertible if and only if both realizations are minimal.*

Proof. (1) Since $M_1 G_1 = M_2 G_2$, multiplying by *any* left inverse M_2^{-L} from the left, and by *any* right inverse G_1^{-R} from the right, we see that

$$M_2^{-L} M_1 = G_2 G_1^{-R}.$$

Hence the operator T of Lemma 2.3 is now independent of the particular right inverse, the operator U^* is now independent of the particular left inverse, and we have

$$T = G_2 G_1^{-R} = M_2^{-L} M_1 = U^*.$$

Hence we see that in the diagrams the (left and the right) triangles commute in the usual sense. Further, the equality of the Markov parameters for both realizations yields

$$M_2 A_2 G_2 = M_1 A_1 G_1 = M_1 G_1 G_1^{-R} A_1 G_1 = M_2 G_2 G_1^{-R} A_1 G_1 = M_2 T A_1 G_1.$$

Premultiplying by M_2^{-L} and postmultiplying by G_1^{-R}, we obtain $A_2T = TA_1$. Hence T strongly connects the realizations 1 and 2. On the other hand, if S is any operator doing the same, then $SA_1^k = A_2^kS$ for every nonnegative integer k. Multiplying on the right by B_1, we obtain $SA_1^kB_1 = A_2^kSB_1 = A_2^kB_2$. Hence $S = G_2G_1^{-R} = T$; i.e., the connecting operator T is unique.

(2) If the operator T is invertible, then the realizations 1 and 2 are system similar or, equivalently, T^{-1} strongly connects 2 to 1. Hence for every nonnegative integer k we obtain $C_2A_2^k = C_1A_1^kT^{-1}$. It follows that $M_2 = M_1T^{-1}$. Therefore $\ker M_1 = (0)$.

Similarly, for every nonnegative integer k the commutative diagram shows that $A_1^kB_1 = T^{-1}A_2^kB_2$. Hence $G_1 = T^{-1}G_2$. Therefore $\operatorname{im} G_2$ is the entire space $\mathbf{F}^{n_2} = \mathbf{F}^{n_1}$, for T is an isomorphism. Hence both realizations are controllable and observable; i.e., they are minimal.

Conversely, assume that both realizations are minimal, and T strongly connects 1 to 2. Part (1) yields then the (unique) form of T. It can be shown as above that the operator T satisfies

$$M_1^{-L}M_2T = I, \qquad TG_1G_2^{-R} = I$$

for any indicated left and right inverses, respectively (which exist by the minimality assumptions). Hence T is invertible, and the proof is complete. □

Next we shall study *end connectedness* of realizations of a transfer function.

LEMMA 2.14. *The realization* 1 *is end connected to realization* 2 *of a given transfer function if and only if (with the standard notation)*

$$\operatorname{colsp} C_1 \subset \operatorname{colsp} C_2, \qquad \operatorname{rowsp} B_2 \subset \operatorname{rowsp} B_1.$$

Proof. Realization 1 is end connected to 2 if and only if there is a matrix T solving the system of matrix equations

$$C_2T = C_1, \qquad TB_1 = B_2.$$

By [16, Theorem 2.3.3], this holds if and only if $C_2B_2 = C_1B_1$ and the equations

$$C_2T = C_1, \qquad SB_1 = B_2$$

are consistent, i.e., have solutions (T, S). The first condition is fulfilled in our case (the common value is the first Markov parameter M_0), and the existence of the pair (T, S) means exactly the stated conditions on the column and row spaces, respectively. □

THEOREM 2.15. *Let $W : \mathbf{F}^b \to \mathbf{F}^c$ be a transfer function. The following properties are equivalent:*

(1) *each pair of its realizations is end connected (in both directions),*

(2) *for some minimal realization (C, A, B) we have*

$$c = \dim \operatorname{colsp} C, \qquad \dim \operatorname{rowsp} B = b; \tag{2.5}$$

i.e., C is surjective and B is injective,

(3) *for every realization* (2.5) *holds.*

Proof. First assume (1). By Lemma 2.14, the transfer function W has the property (1) if and only if for any two realizations (C_j, A_j, B_j) $(j = 1, 2)$ we have

$$\operatorname{colsp} C_1 = \operatorname{colsp} C_2, \qquad \operatorname{rowsp} B_1 = \operatorname{rowsp} B_2. \tag{2.6}$$

In particular, this is valid if one of them is the (minimal) realization (C, A, B). We clearly have

$$\dim \operatorname{colsp} C \le c, \qquad \dim \operatorname{rowsp} B \le b.$$

Assume that, e.g., $\dim \operatorname{colsp} C < c$, and let v denote a column vector with c components such that $v \notin \operatorname{colsp} C$. Define the block matrices

$$\hat{C}_1 := \begin{pmatrix} v & C \end{pmatrix}, \qquad \hat{A}_1 := \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}, \qquad \hat{B}_1 := \begin{pmatrix} 0 \\ B \end{pmatrix}.$$

It is clear that $(\hat{C}_1, \hat{A}_1, \hat{B}_1)$ is also a realization of the transfer function W, and $\dim \operatorname{colsp} \hat{C}_1 > \dim \operatorname{colsp} C$. In a similar way we see that $\dim \operatorname{rowsp} B < b$ implies the existence of a realization $(\hat{C}_2, \hat{A}_2, \hat{B}_2)$ such that $\dim \operatorname{rowsp} \hat{B}_2 > \dim \operatorname{rowsp} B$. Hence (1) implies (2). It is easily seen that (2) implies (3). Finally, (3) clearly implies (2.6). By Lemma 2.14, we obtain (1). □

3. Connections between nonnegative realizations of a transfer function. In the following result we describe exactly when a *nonnegative weakly connecting pair* exists in the case when all the matrices C_j, A_j, B_j are nonnegative (the values of c and b can be arbitrary). Hence all the occurring matrices may be assumed to be over the real field $\mathbf{R}$. $\operatorname{cone}[M]$ will denote the cone generated by the columns of a matrix M.

THEOREM 3.1. *Let (C_j, A_j, B_j) $(j = 1, 2)$ be two (entrywise) nonnegative realizations of the same (not identically zero) rational matrix function W mapping $\mathbf{R}^b$ into $\mathbf{R}^c$ and having the orders*

$$n_j = \dim A_j \ge n_0 \quad (j = 1, 2),$$

where n_0 is the McMillan degree of W. Consider the diagram (2.1) *for every $k = 0, 1, \ldots$ with $\mathbf{R}$ replacing $\mathbf{F}$ everywhere.*

A nonnegative matrix T as in (2.1)*, making the uppermost and the left-hand side ways commute between the extreme spaces, exists if and only if*

$$\operatorname{cone}[C_1 + K] \subset \operatorname{cone}[C_2]$$

for some matrix K such that the column space of G_1 is orthogonal to the row space of K. Further, a nonnegative matrix S as in (2.1)*, making the lowest and the right-hand side ways commute between the extreme spaces, exists if and only if*

$$\operatorname{cone}[(B_2 + L)^t] \subset \operatorname{cone}[B_1^t]$$

for some matrix L such that the column space of L is orthogonal to the row space of M_2.

If either $c = 1$ or $b = 1$, then there exists an (entrywise) nonnegative matrix T or $S : \mathbf{R}^{n_1} \to \mathbf{R}^{n_2}$ (on the "corresponding side") such that for every nonnegative integer k the above diagram commutes in the sense used above. Hence, if $c = b = 1$ (scalar-valued rational function or, equivalently, single-input single-output [SISO] system), then there is a pair (T, S) as above such that both matrices are nonnegative.

Remark 3.2. Any pair (T, S) as above *weakly connects the realization* 1 *to* 2. Note that examples show that in the case of a nonnegative system for which $c = 2$, $b = 1$ (single-input multiple-output system), or $c = 1$, $b = 2$ a *nonnegative* weakly connecting pair (in either direction!) may *not* exist.

Proof. Let $n := \max[n_1, n_2]$, and define the controllability and observability operators $G_j : \mathbf{R}^{nb} \to \mathbf{R}^{n_j}$ and $M_j : \mathbf{R}^{n_j} \to \mathbf{R}^{nc}$, respectively, as before.

A nonnegative T exists if and only if $[C_2T - C_1]G_1 = 0$ for some nonnegative matrix T. Define $K := C_2T - C_1$. Then the above condition becomes $KG_1 = 0$, which means exactly the stated orthogonality condition. The nonnegativity of T and the condition $C_2T = C_1 + K$ together are equivalent to the above cone condition.

Similarly, a nonnegative S exists if and only if $M_2[SB_1 - B_2] = 0$ for some nonnegative matrix S. Define $L := SB_1 - B_2$. Then we can proceed as in the preceding paragraph.

In the special case $b = 1$ the matrices B_j $(j = 1, 2)$ are nonnegative column vectors, and the nonvanishing condition for W implies that at least one entry of B_j is positive. Hence the second cone condition (with $L = 0$) is in the simple form

$$\operatorname{cone}[B_j^t] = \mathbf{R}_+ \quad (j = 1, 2)$$

satisfied, which proves the penultimate part of the theorem: the condition $b = 1$ implies the existence of a nonnegative weakly connecting (right) matrix S, independently of the value of c. A similar proof of the existence of a nonnegative weakly connecting (left) matrix T holds for the special case $c = 1$, independently of the value of b. ∎

Remark 3.3. Note that if we *require* either $C_2T = C_1$ or $SB_1 = B_2$, then the conditions for the existence of a nonnegative T or S, respectively, will have a very simple form (with $K = 0$ or $L = 0$).

Actually, in the SISO case we can even prove the existence of a stronger type of nonnegative connection.

THEOREM 3.4. *Assume that under the conditions of Theorem* 3.1 *we have* $c = 1$, $b = 1$ *(SISO nonnegative case). Then realization* 1 *is end connected to realization* 2 *by an (entrywise) nonnegative matrix* T.

Proof. The assertion means that there is a nonnegative $T \in M(n_2, n_1)$ such that

$$C_2T = C_1, \qquad TB_1 = B_2.$$

This system of equation is equivalent to the linear system (for the entries of $\operatorname{vec} T$)

$$\begin{pmatrix} \operatorname{vec} B_2 \\ \operatorname{vec} C_1 \end{pmatrix} = \begin{pmatrix} B_1^t \otimes I_{n_2} \\ I_{n_1} \otimes C_2 \end{pmatrix} \operatorname{vec} T.$$

Denote the qth entry of $\operatorname{vec} T$ by t_q, and the ith entry of the row (column) matrix C_j (B_j) by c_{ji} (b_{ji}), respectively. Then the above condition is equivalent to the existence of nonnegative numbers t_q $(q = 1, \dots, n_1n_2)$ such that

$$\begin{pmatrix} b_{21} \\ b_{22} \\ \vdots \\ b_{2n_2} \\ c_{11} \\ c_{12} \\ \vdots \\ c_{1n_1} \end{pmatrix} = t_1 \begin{pmatrix} b_{11} \\ 0 \\ \vdots \\ 0 \\ c_{21} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + t_{n_2} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ b_{11} \\ c_{2n_2} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + t_{(n_1-1)n_2+1} \begin{pmatrix} b_{1n_1} \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \\ c_{21} \end{pmatrix} + \cdots + t_{n_1n_2} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ b_{1n_1} \\ 0 \\ 0 \\ \vdots \\ c_{2n_2} \end{pmatrix}.$$

This is the description of the constraints of a linear programming problem in the so-called standard form. Let L denote the (right-hand side) coefficient matrix of the

above system of linear equations; then $L \in M(n_1 + n_2, n_1 n_2)$. By the well-known theorem of Farkas (see, e.g., [15, p. 34]), a (nonnegative) solution T exists if and only if for any column vector $v \in M(n_1 + n_2, 1)$ satisfying the inequality $L^t v \geq 0$ it follows that the scalar product sv of the left-hand side (column) vector s of the system with the vector v is nonnegative. Introduce the notation

$$v = (y_1, \ldots, y_{n_2}, z_1, \ldots, z_{n_1})^t.$$

$L^t v \geq 0$ holds if and only if for every $j = 1, \ldots, n_2$, $k = 1, \ldots, n_1$ we have

$$b_{1k} y_j + c_{2j} z_k \geq 0. \tag{3.1}$$

Multiplying by the nonnegative $b_{2j} c_{1k}$ and summing we obtain

$$\left(\sum_{k=1}^{n_1} b_{1k} c_{1k} \right) \left(\sum_{j=1}^{n_2} b_{2j} y_j \right) + \left(\sum_{j=1}^{n_2} b_{2j} c_{2j} \right) \left(\sum_{k=1}^{n_1} c_{1k} z_k \right) \geq 0.$$

The first factors in the terms are the scalar products $C_1 B_1$ and $C_2 B_2$, respectively, and they are equal to the Markov parameter w_0 of the (common) transfer function. If $w_0 > 0$, this implies

$$sv = \sum_{j=1}^{n_2} b_{2j} y_j + \sum_{k=1}^{n_1} c_{1k} z_k \geq 0.$$

By the Farkas theorem then, there is a nonnegative solution matrix T.

Consider now the case $w_0 = 0$. Then

$$\sum_{k=1}^{n_1} b_{1k} c_{1k} = w_0 = 0 = \sum_{j=1}^{n_2} b_{2j} c_{2j}.$$

If for some $j \in \{1, \ldots, n_2\}$ we have $b_{2j} > 0$, then the nonnegativity of the realizations implies $c_{2j} = 0$. Hence, for every $k = 1, \ldots, n_1$, (3.1) implies $b_{1k} y_j \geq 0$. By assumption, the transfer function is not identically 0; hence some b_{1k} is positive. Therefore $y_j \geq 0$. The other possibility is $b_{2j} = 0$ for all j. In any case we obtain

$$b_{2j} y_j \geq 0 \quad (j = 1, \ldots, n_2).$$

Similarly, if for some $k \in \{1, \ldots, n_1\}$ we have $c_{1k} > 0$, then $w_0 = 0$ implies $b_{1k} = 0$. Hence (3.1) implies $c_{2j} z_k \geq 0$ for every $j = 1, \ldots, n_2$. Since there is a positive c_{2j}, we obtain $z_k \geq 0$. The other possibility being $c_{1k} = 0$, we see that

$$c_{1k} z_k \geq 0 \quad (k = 1, \ldots, n_1).$$

Summing up the terms, we obtain for the case $w_0 = 0$ also $sv \geq 0$. The proof is complete. □

Acknowledgment. The authors are indebted to the referees for valuable suggestions that have improved the presentation of the paper.

REFERENCES

[1] C.I. Byrnes and M.A. Gauger, *Decidability criteria for the similarity problem, with applications to the moduli of linear dynamical systems*, Adv. Math., 25 (1977), pp. 59–90.
[2] C.I. Byrnes and N.E. Hurt, *On the moduli of linear dynamical systems*, in Studies in Analysis, Adv. in Math. Suppl. Studies 4, G.-C. Rota, ed., Academic Press, New York, 1979, pp. 83–122.
[3] L. Bakule, J. Rodellar, and J.M. Rossell, *Structure of expansion-contraction matrices in the inclusion principle for dynamic systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1136–1155.
[4] P. Falb, *Methods of Algebraic Geometry in Control Theory: Part* I, Birkhäuser, Boston, 1990.
[5] P. Falb, *Methods of Algebraic Geometry in Control Theory: Part* II, Birkhäuser, Boston, 1999.
[6] L. Farina and S. Rinaldi, *Positive Linear Systems*, Wiley-Interscience, New York, 2000.
[7] K.-H. Foerster and B. Nagy, *On nonnegative realizations of rational matrix functions and nonnegative input-output systems*, Oper. Theory Adv. Appl., 103 (1998), pp. 89–104.
[8] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, Wiley-Interscience, New York, 1986.
[9] P.R. Halmos, *Finite-Dimensional Vector Spaces*, Princeton University Press, Princeton, NJ, 1948.
[10] J. v.d. Hof, *System Theory and System Identification of Compartmental Systems*, Thesis, Rijksuniversiteit Groningen, The Netherlands, 1996.
[11] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[12] M. Ikeda, D.D. Siljak, and D.E. White, *An inclusion principle for dynamic systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 244–249.
[13] R.E. Kalman, *Lectures on Controllability and Observability*, C.I.M.E. Summer School 1968, Edizioni Cremonese, Roma, 1969.
[14] R.E. Kalman, P.L. Falb, and M.A. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
[15] V.G. Karmanov, *Mathematical Programming*, Mir, Moscow, 1989.
[16] C.R. Rao and S.K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.
[17] E.D. Sontag, *Mathematical Control Theory*, Springer-Verlag, New York, 1990.
[18] W.M. Wonham, *Linear Multivariable Control*, Springer-Verlag, New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1046–1073

REGULARIZATION OF LINEAR DISCRETE-TIME PERIODIC DESCRIPTOR SYSTEMS BY DERIVATIVE AND PROPORTIONAL STATE FEEDBACK*

YUEN-CHENG KUO†, WEN-WEI LIN†, AND SHU-FANG XU‡

Abstract. In this paper, we consider the regularization problem for the linear time-varying discrete-time periodic descriptor systems by derivative and proportional state feedback controls. Sufficient conditions are given under which derivative and proportional state feedback controls can be constructed so that the periodic closed-loop systems are regular and of index at most one. The construction procedures used to establish the theory are based on orthogonal and elementary matrix transformations and can, therefore, be developed to a numerically efficient algorithm. The problem of finite pole assignment of periodic descriptor systems is also studied.

Key words. linear periodic descriptor systems, regularization, derivative and proportional feedback

AMS subject classifications. 93B10, 93B40, 93B52, 93B55, 93C55, 65F22

DOI. 10.1137/S0895479802412632

1. Introduction.

We consider linear time-varying discrete-time periodic descriptor systems of the form

$$\begin{aligned} E_j x_{j+1} &= A_j x_j + B_j u_j, \\ y_j &= C_j x_j, \end{aligned} \tag{1.1}$$

where x_0 is given and the matrices $E_j,\ A_j \in \mathbb{R}^{n\times n}$, $B_j \in \mathbb{R}^{n\times m}$ $(m \le n)$, $C_j \in \mathbb{R}^{k\times n}$ are periodic with period $p \ge 1$, that is, $E_j = E_{j+p}$, $A_j = A_{j+p}$, $B_j = B_{j+p}$, and $C_j = C_{j+p}$ for all j. Throughout this paper we assume that the control matrices B_j are all of full column rank and the matrices E_j are allowed to be singular.

The number of contributions on linear time-varying discrete-time periodic systems has been increasing in recent times; see, for example, [5, 15, 20, 21, 22, 24, 25, 28, 29, 30, 31, 32] and references therein. This increasing interest in such systems is motivated by the large variety of processes that can be modelled through linear discrete-time periodic systems (e.g., multirate sampled-data systems, chemical processes, periodically time-varying filters and networks, seasonal phenomena, and so on [1, 2, 4, 16, 26, 27, 33]). The dynamics of linear discrete-time periodic descriptor systems (1.1) depend critically on the regularity and the eigenstructure of the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^{p}$ which form the homogeneous systems of (1.1), i.e.,

$$E_j x_{j+1} = A_j x_j. \tag{1.2}$$

The matrix pairs $\{(E_j, A_j)\}_{j=1}^{p}$ are called to be regular if $\det[C((\alpha_j, \beta_j)_{j=1}^{p})] \not\equiv 0$,

*Received by the editors August 5, 2002; accepted for publication (in revised form) by V. Mehrmann September 11, 2003; published electronically July 14, 2004. This research was supported in part by the National Center for Theoretical Sciences in Taiwan and the Natural Science Foundation of China under grant 19971007.
http://www.siam.org/journals/simax/25-4/41263.html

†Department of Mathematics, National Tsinghua University, Hsinchu, 300, Taiwan (d883207@oz.nthu.edu.tw, wwlin@am.nthu.edu.tw).

‡School of Mathematical Sciences, Peking University, Beijing, 100871, China (xsf@pku.edu.cn).

where

$$C((\alpha_j,\beta_j)_{j=1}^p) \equiv \begin{bmatrix} \alpha_1E_1 & 0 & \cdots & 0 & -\beta_1A_1 \\ -\beta_2A_2 & \alpha_2E_2 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & 0 \\ 0 & & 0 & -\beta_pA_p & \alpha_pE_p \end{bmatrix}, \tag{1.3}$$

in which α_j, β_j are complex variables for $j = 1, \ldots, p$.

DEFINITION 1.1. *Let $\{(E_j, A_j)\}_{j=1}^p$ be $n \times n$ regular matrix pairs. If there are complex numbers $\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_p$ with*

$$\left(\prod_{j=1}^p \alpha_j, \prod_{j=1}^p \beta_j\right) \equiv (\pi_\alpha, \pi_\beta) \neq (0,0) \tag{1.4}$$

satisfying $\det[C((\alpha_j,\beta_j)_{j=1}^p)] = 0$, *then we say that (π_α, π_β) is an eigenvalue pair of $\{(E_j, A_j)\}_{j=1}^p$.*

Note that if (π_α, π_β) is an eigenvalue of $\{(E_j, A_j)\}_{j=1}^p$, then (π_α, π_β) and $(\tau\pi_\alpha, \tau\pi_\beta)$ represent the same eigenvalue pair for any nonzero τ. If $\pi_\beta \neq 0$, then $\lambda = \pi_\alpha/\pi_\beta$ is a finite eigenvalue; otherwise $(\pi_\alpha, 0)$ is an infinite eigenvalue. The set of all eigenvalue pairs of $\{(E_j, A_j)\}_{j=1}^p$ is denoted by $\sigma\left(\{(E_j, A_j)\}_{j=1}^p\right)$.

It is easily seen that the determinant of $C((\alpha_j,\beta_j)_{j=1}^p)$ is a homogeneous polynomial in (π_α, π_β) of degree n which is of the form

$$\sum_{k=0}^n c_k \pi_\alpha^k \pi_\beta^{n-k}, \tag{1.5}$$

where $c_0, \ldots, c_n$ are complex numbers uniquely determined by $\{(E_j, A_j)\}_{j=1}^p$. For the regular matrix pairs $\{(E_j, A_j)\}_{j=1}^p$ this implies that at least one of the c_k's is nonzero, and hence we see from Definition 1.1 that there are exact n eigenvalue pairs (counting multiplicity) for $\{(E_j, A_j)\}_{j=1}^p$.

It was shown in [29] that the solvability of (1.2) is equivalent to the condition that the pencil

$$\alpha\mathcal{E} - \beta\mathcal{A} := \begin{bmatrix} \alpha E_1 & 0 & \cdots & 0 & -\beta A_1 \\ -\beta A_2 & \alpha E_2 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & 0 \\ 0 & & 0 & -\beta A_p & \alpha E_p \end{bmatrix} \tag{1.6}$$

is regular, i.e., $\det(\alpha\mathcal{E} - \beta\mathcal{A}) \not\equiv 0$. From (1.5) it is easy to check that

$$\sigma\left(\{(E_j, A_j)\}_{j=1}^p\right) = \left\{(\alpha^p, \beta^p) \mid \det(\alpha\mathcal{E} - \beta\mathcal{A}) = 0\right\}. \tag{1.7}$$

Hence, from (1.7) the solvability condition of (1.2) becomes the regularity of the matrix pairs $\{(E_j, A_j)\}_{j=1}^p$.

In order to alter the dynamics of the periodic descriptor systems (1.1), it is usually to use proportional state feedback to modify the matrices A_j, that is, the control

vectors are taken to be $u_j = F_j x_j + v_j$ for $j = 1, \ldots, p$. The closed-loop matrix pairs then become

$$\{(E_j, A_j + B_j F_j)\}_{j=1}^{p}. \tag{1.8}$$

Similarly, if we interchange the role of E_j and A_j, then we can also use derivative state feedback to modify the matrices E_j. The closed-loop matrix pairs become

$$\{(E_j + B_j G_j, A_j)\}_{j=1}^{p}, \tag{1.9}$$

where the control vectors are taken to be $u_j = -G_j x_{j+1} + v_j$, $j = 1, \ldots, p$. If a full state feedback of the form $u_j = -G_j x_{j+1} + F_j x_j + v_j$ is considered (see [6] for the case of $p = 1$), then it leads to periodic descriptor systems with the periodic matrix pairs of the form

$$\{(E_j + B_j G_j, A_j + B_j F_j)\}_{j=1}^{p}. \tag{1.10}$$

For the case of period $p = 1$, one has the time-invariant case $E_j = E$, $A_j = A$, $B_j = B$, $C_j = C$. It is well known that for a regular matrix pair (E, A) (i.e., $\det(\alpha E - \beta A) \neq 0$ for some $(\alpha, \beta) \in \mathbb{C}^2$) there exist nonsingular matrices P and Q which transform E and A into the Kronecker canonical form [17]:

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \qquad PAQ = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}.$$

Here J is a Jordan matrix corresponding to the finite eigenvalues of (E, A) and N is a nilpotent Jordan matrix corresponding to the infinite eigenvalues. The index of the matrix pair is the index of nilpotency of the nilpotent matrix N, i.e., (E, A) is of index ν, denoted by $\nu = \operatorname{ind}_\infty(E, A)$, if $N^{\nu-1} \neq 0$ and $N^\nu = 0$. By convention, if E is nonsingular, the pencil is said to be of index zero. If a matrix pair is regular and of index at most 1, the corresponding time-invariant continuous system

$$E\frac{dx}{dt} = Ax(t) + Bu(t)$$

has a unique solution for all admissible controls $u(t)$ with consistent initial conditions. In theory, such a system can be separated into purely dynamical and purely algebraic parts, and moreover, the algebraic part can be eliminated to give a reduced-order standard system. If the index is larger than 1, however, impulses can arise in the response of the system and the system can lose causality if the control is not sufficiently smooth [18]. Therefore, it is desirable to use a feedback control that ensures that the closed-loop system is regular and of index at most one, and furthermore, has the required finite poles. In the last few years, there has been an increasing interest in developing numerical algorithms for the regularization and the finite pole assignment of descriptor time-invariant systems by proportional and derivative feedback. See, for example, [6, 7, 8, 9, 10, 11, 12, 13, 23] and references therein.

In this paper, we focus on the following regularization and pole assignment problems: For given periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^{p}$, we first construct periodic derivative and proportional matrices G_j and F_j such that the periodic matrix pairs $\{(E_j + B_j G_j,\ A_j + B_j F_j)\}_{j=1}^{p}$ of the periodic closed-loop systems are regular and of

index at most one (see the definitions in the next section). Then we construct periodic feedback matrices G_j^f and F_j^f such that the periodic closed-loop systems not only are regular and have the required finite poles, but also have index at most one. To the best of our knowledge, for the case of period $p \geq 2$, these problems have not been investigated much in the literature.

Our contribution in this paper is threefold. First, in Theorem 2.5 we give an equivalent condition for the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^p$ to be regular and of index at most one. Second, in Theorems 3.1 and 4.3 we specify sufficient conditions under which derivative and proportional state feedback can be constructed so that the periodic closed-loop systems are regular and of index at most one. Third, in Theorem 5.1, we give the solvability condition for the finite pole assignment problem of the periodic matrix triples. The main proofs given in this paper can provide a numerically method for constructing the required feedback matrices, which is based on orthogonal and elementary matrix transformations.

This paper is organized as follows. In section 2 we introduce some notations and definitions, and give some preliminary results. In section 3 we present a canonical form under matrix transformations. In section 4, we use this canonical form to construct derivative and proportional feedback so that the periodic closed-loop systems are regular and of index at most one. The problem of finite pole assignment with derivative and proportional feedback is presented in section 5.

2. Preliminaries. In this section we introduce some notations and definitions, and give some preliminary results. Throughout this paper we use the following notations. For any given periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$ we use, alternatively, the script notations

$$\widetilde{\mathcal{E}}_j \equiv \widetilde{\mathcal{E}}(E_j, \ldots, E_{j+p-1}; A_{j+1}, \ldots, A_{j+p-1})$$

$$:= \begin{bmatrix} E_j & 0 & 0 & \cdots & \cdots & 0 \\ -A_{j+1} & E_{j+1} & 0 & & & 0 \\ 0 & -A_{j+2} & E_{j+2} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & E_{j+p-2} & 0 \\ 0 & \cdots & \cdots & 0 & -A_{j+p-1} & E_{j+p-1} \end{bmatrix}, \tag{2.1a}$$

$$\widetilde{\mathcal{A}}_j \equiv \widetilde{\mathcal{A}}(A_j) := \begin{bmatrix} 0_{n\times(p-1)n} & A_j \\ 0_{(p-1)n\times(p-1)n} & 0_{(p-1)n\times n} \end{bmatrix}, \tag{2.1b}$$

$$\mathcal{B}_j \equiv \mathcal{B}(B_j, \ldots, B_{j+p-1}) := \operatorname{diag}(B_j, \ldots, B_{j+p-1}). \tag{2.1c}$$

We also denote the null space of a matrix M by $\mathcal{N}(M)$, and use $\mathcal{S}_\infty(M)$ to denote a full rank matrix whose columns span the null space $\mathcal{N}(M)$. The indices "j" for all periodic coefficient matrices are chosen in $\{1, \ldots, p\}$ modulo p without ambiguity.

In terms of the above notations, we now characterize the regular periodic matrix pairs as follows.

LEMMA 2.1. *The following statements are equivalent.*

(1) *The periodic matrix pairs* $\{(E_j, A_j)\}_{j=1}^p$ *are regular.*
(2) *The matrix pair* $(\mathcal{E}, \mathcal{A})$ *in* (1.6) is regular, i.e., $\det(\alpha\mathcal{E} - \beta\mathcal{A}) \not\equiv 0$.
(3) The matrix pair $(\tilde{\mathcal{E}}_j, \tilde{\mathcal{A}}_j)$ is regular, for some $j \in \{1, \ldots, p\}$.
(4) The matrix pairs $(\tilde{\mathcal{E}}_j, \tilde{\mathcal{A}}_j)$ are regular, for all $j = 1, \ldots, p$.

Proof. By periodic Schur decomposition theorem [5] there exist unitary matrices Q_j and Z_j such that

$$Q_j E_j Z_j = \begin{bmatrix} b_{11}^j & \cdots & b_{1n}^j \\ & \ddots & \vdots \\ & & b_{nn}^j \end{bmatrix}, \qquad Q_j A_j Z_{j-1} = \begin{bmatrix} a_{11}^j & \cdots & a_{1n}^j \\ & \ddots & \vdots \\ & & a_{nn}^j \end{bmatrix},$$

where $Z_0 = Z_p$, from which we can derive that

$$\det[C\big((\alpha_j, \beta_j)_{j=1}^p\big)] = \prod_{i=1}^n \Big(\prod_{j=1}^p \alpha_j b_{ii}^j - \prod_{j=1}^p \beta_j a_{ii}^j \Big), \tag{2.2a}$$

$$\det(\alpha \mathcal{E} - \beta \mathcal{A}) = \prod_{i=1}^n \Big(\alpha^p \prod_{j=1}^p b_{ii}^j - \beta^p \prod_{j=1}^p a_{ii}^j \Big), \tag{2.2b}$$

$$\det(\alpha \widetilde{\mathcal{E}}_j - \beta \widetilde{\mathcal{A}}_j) = \prod_{i=1}^n \Big(\alpha^p \prod_{k=j}^{j+p-1} b_{ii}^k - \beta \alpha^{p-1} \prod_{k=j}^{j+p-1} a_{ii}^k \Big), \tag{2.2c}$$

for $j = 1, \ldots, p$.

It is easily seen that any equation in (2.2) which is not identical to zero, i.e., $(\prod_{j=1}^p b_{ii}^j, \prod_{j=1}^p a_{ii}^j) \neq (0, 0)$, for $i = 1, \ldots, n$, implies that the other equations in (2.2) are also not identical to zero. This completes the proof. ∎

In a similar fashion to the Kronecker canonical form for a regular matrix pair, we can transform regular periodic matrix pairs into periodic Kronecker canonical forms.

LEMMA 2.2. *Suppose that the periodic matrix pairs* $\{(E_j, A_j)\}_{j=1}^p$ *in systems* (1.1) *are regular. Then there exist nonsingular matrices* X_j *and* Y_j, $j = 1, \ldots, p$, *such that*

$$X_j E_j Y_j = \begin{bmatrix} I & 0 \\ 0 & E_j^0 \end{bmatrix}, \quad X_j A_j Y_{j-1} = \begin{bmatrix} A_j^f & 0 \\ 0 & I \end{bmatrix}, \tag{2.3}$$

where $Y_0 = Y_p$, $A_{j+p-1}^f A_{j+p-2}^f \cdots A_j^f \equiv J_j$, $(j = 1, \ldots, p)$ *is a Jordan matrix corresponding to the finite eigenvalues of* $\{(E_j, A_j)\}_{j=1}^p$ *and* $E_j^0 E_{j+1}^0 \cdots E_{j+p-1}^0 \equiv N_j$, $(j = 1, \ldots, p)$ *is a nilpotent Jordan matrix corresponding to the infinite eigenvalues of* $\{(E_j, A_j)\}_{j=1}^p$.

Proof. By periodic Schur decomposition theorem and the reordering of eigenvalues [5, 20] there are unitary matrices Q_j, P_j, $j = 1, \ldots, p$ so that

$$Q_j E_j P_j = \begin{bmatrix} E_{j,1} & E_{j,3} \\ 0 & E_{j,2} \end{bmatrix}, \quad Q_j A_j P_{j-1} = \begin{bmatrix} A_{j,1} & A_{j,3} \\ 0 & A_{j,2} \end{bmatrix} \tag{2.4}$$

are upper triangular, and moreover $E_{j,1}$ and $A_{j,2}$ are nonsingular and all diagonal elements of $E_{j,2} E_{j+1,2} \cdots E_{j+p-1,2}$ are zero for $j = 1, 2, \ldots, p$. We then let

$$\begin{bmatrix} E_{j,1}^{-1} & 0 \\ 0 & A_{j,2}^{-1} \end{bmatrix} \begin{bmatrix} E_{j,1} & E_{j,3} \\ 0 & E_{j,2} \end{bmatrix} = \begin{bmatrix} I & \hat{E}_{j,3} \\ 0 & \hat{E}_j^0 \end{bmatrix}, \tag{2.5a}$$

$$\begin{bmatrix} E_{j,1}^{-1} & 0 \\ 0 & A_{j,3}^{-1} \end{bmatrix} \begin{bmatrix} A_{j,1} & A_{j,3} \\ 0 & A_{j,2} \end{bmatrix} = \begin{bmatrix} \hat{A}_j^f & \hat{A}_{j,3} \\ 0 & I \end{bmatrix}. \tag{2.5b}$$

Next, we prove that there exists periodic matrices U_j and V_j, $j = 1, 2, \ldots, p$ such that

$$\begin{bmatrix} I & U_j \\ 0 & I \end{bmatrix} \begin{bmatrix} I & \hat{E}_{j,3} \\ 0 & \hat{E}_j^0 \end{bmatrix} \begin{bmatrix} I & V_j \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \hat{E}_j^0 \end{bmatrix}, \tag{2.6a}$$

$$\begin{bmatrix} I & U_j \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{A}_j^f & \hat{A}_{j,3} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & V_{j-1} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \hat{A}_j^f & 0 \\ 0 & I \end{bmatrix}. \tag{2.6b}$$

Comparing the both sides of (2.6a) and (2.6b) we have

$$V_j + U_j \hat{E}_j^0 + \hat{E}_{j,3} = 0, \tag{2.7a}$$

$$\hat{A}_j^f V_{j-1} + U_j + \hat{A}_{j,3} = 0 \tag{2.7b}$$

for $j = 1, 2, \ldots, p$, where $V_0 = V_p$. Eliminating U_j in (2.7) we get

$$V_j = \hat{A}_j^f V_{j-1} \hat{E}_j^0 + \hat{A}_{j,3} \hat{E}_j^0 - \hat{E}_{j,3} \tag{2.8}$$

for $j = 1, 2, \ldots, p$, from which we obtain

$$V_p = \left(\hat{A}_p^f \hat{A}_{p-1}^f \cdots \hat{A}_1^f \right) V_p \left(\hat{E}_1^0 \hat{E}_2^0 \cdots \hat{E}_p^0 \right) + D_p, \tag{2.9}$$

where

$$\begin{aligned} D_p = (\hat{A}_{p,3} \hat{E}_p^0 - \hat{E}_{p,3}) + \hat{A}_p^f (\hat{A}_{p-1,3} \hat{E}_{p-1}^0 - \hat{E}_{p-1,3}) \hat{E}_p^0 + \cdots + \\ \left(\hat{A}_p^f \hat{A}_{p-1}^f \cdots \hat{A}_2^f \right) \left(\hat{A}_{1,3} \hat{E}_1^0 - \hat{E}_{1,3} \right) \left(\hat{E}_2^0 \hat{E}_3^0 \cdots \hat{E}_p^0 \right). \end{aligned}$$

Notice that $(\hat{E}_1^0 \hat{E}_2^0 \cdots \hat{E}_p^0)$ is an upper triangular matrix with all diagonal elements zero, we can uniquely determine the matrix V_p from (2.9). Then, from (2.8) and (2.6b) we can uniquely determine V_j for $j = 1, 2, \ldots, p-1$, and U_j for $j = 1, 2, \ldots, p$, respectively.

Finally, by the well-known Jordan decomposition theorem we know that there exist nonsingular matrices G_j, Z_j, $j = 1, \ldots, p$ such that

$$G_j^{-1} \left(\hat{A}_{j+p-1}^f \hat{A}_{j+p-2}^f \cdots \hat{A}_j^f \right) G_j = J_j \text{ (Jordan form)}, \tag{2.10a}$$

$$Z_j^{-1} \left(\hat{E}_j^0 \hat{E}_{j+1}^0 \cdots \hat{E}_{j+p-1}^0 \right) Z_j = N_j \text{ (nilpotent Jordan form)}. \tag{2.10b}$$

Now let

$$\begin{aligned} X_j &:= \begin{bmatrix} G_{j+1}^{-1} & 0 \\ 0 & Z_j^{-1} \end{bmatrix} \begin{bmatrix} I & U_j \\ 0 & I \end{bmatrix} \begin{bmatrix} E_{j,1}^{-1} & 0 \\ 0 & A_{j,2}^{-1} \end{bmatrix} Q_j, \\ Y_j &:= P_j \begin{bmatrix} I & V_j \\ 0 & I \end{bmatrix} \begin{bmatrix} G_{j+1} & 0 \\ 0 & Z_{j+1} \end{bmatrix}, \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} E_j^0 &:= Z_j^{-1} \hat{E}_j^0 Z_{j+1}, \\ A_j^f &:= G_{j+1}^{-1} \hat{A}_j^f G_j. \end{aligned} \tag{2.12}$$

Then from (2.4)–(2.12) we have

$$X_j E_j Y_j = \begin{bmatrix} I & 0 \\ 0 & E_j^0 \end{bmatrix}, \quad X_j A_j Y_{j-1} = \begin{bmatrix} A_j^f & 0 \\ 0 & I \end{bmatrix}$$

with $\prod_{k=j+p-1}^{j} A_k^f = J_j$, a Jordan matrix, and $\prod_{k=j}^{j+p-1} E_k^0 = N_j$, a nilpotent Jordan matrix, for $j = 1, \ldots, p$. ∎

As an application of this lemma, let us consider the periodic system (1.1) with $u_j = 0$, i.e., the free periodic system

$$E_j x_{j+1} = A_j x_j. \tag{2.13}$$

Using Lemma 2.2 we can reduce the system (2.13) into a forward and a backward part:

$$x_{j+1}^f = A_j^f x_j^f, \quad A_{j+p}^f = A_j^f, \tag{2.14}$$

$$E_j^0 x_{j+1}^b = x_j^b, \quad E_{j+p}^0 = E_j^0, \tag{2.15}$$

where

$$x_j = Y_{j-1} \begin{bmatrix} x_j^f \\ x_j^b \end{bmatrix},$$

provided the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^p$ are regular, then we obtain from (2.14) and (2.15) the set of p subsampled systems:

$$x_{j+(i+1)p}^f = J_j x_{j+ip}^f, \quad i = 0, 1, 2, \ldots$$

$$N_j x_{j+(i+1)p}^b = x_{j+ip}^b, \quad i = 0, 1, 2, \ldots$$

for $j = 1, 2, \ldots, p$, which are time invariant. This shows that the dynamical properties of the system (2.13) depend critically on the eigenstructure of the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^p$. Especially, if $N_j = 0$ for $j = 1, 2, \ldots, p$, then $x_j^b = 0$ for all $j = 0, 1, \ldots$, and so in such case the system (2.13) is reduced into a reduced-order standard periodic system (2.14).

By Lemma 2.2, we can characterize the nilpotency of the regular periodic matrix pairs by the index of $(\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j)$.

LEMMA 2.3. *Assume that the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^p$ are regular and have the periodic Kronecker canonical forms as shown in* (2.3)*, then the nilpotency of the nilpotent matrix $E_j^0 \cdots E_{j+p-1}^0 \equiv N_j$ (i.e., $N_j^{\nu_j - 1} \neq 0$ and $N_j^{\nu_j} = 0$) is just equal to $\operatorname{ind}_\infty(\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j)$, which denotes the index of $(\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j)$, for $j = 1, \ldots, p$.*

Proof. Let

$$\mathcal{X}_j = \operatorname{diag}(X_j, \ldots, X_{j+p-1}), \qquad \mathcal{Y}_j = \operatorname{diag}(Y_j, \ldots, Y_{j+p-1}),$$

where X_j and Y_j are defined in Lemma 2.2, and we define $X_{k+p} = X_k$ and $Y_{k+p} = Y_k$ for all k. Then it follows from (2.3) that

$$\mathcal{X}_j \widetilde{\mathcal{E}}_j \mathcal{Y}_j = \begin{bmatrix} \hat{E}_j & 0 & 0 & \cdots & \cdots & 0 \\ -\hat{A}_{j+1} & \hat{E}_{j+1} & 0 & & & 0 \\ 0 & -\hat{A}_{j+2} & \hat{E}_{j+2} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \hat{E}_{j+p-2} & 0 \\ 0 & \cdots & \cdots & 0 & -\hat{A}_{j+p-1} & \hat{E}_{j+p-1} \end{bmatrix}, \tag{2.16}$$

$$\mathcal{X}_j\widetilde{\mathcal{A}}_j\mathcal{Y}_j = \widetilde{\mathcal{A}}(A_j) := \begin{bmatrix} 0_{n\times(p-1)n} & \hat{A}_j \\ 0_{(p-1)n\times(p-1)n} & 0_{(p-1)n\times n} \end{bmatrix}, \tag{2.17}$$

where

$$\hat{E}_j = X_jE_jY_j = \begin{bmatrix} I & 0 \\ 0 & E_j^0 \end{bmatrix}, \qquad \hat{A}_j = X_jA_jY_{j-1} = \begin{bmatrix} A_j^f & 0 \\ 0 & I \end{bmatrix}.$$

Notice the special structure of (2.16) and (2.17). Using the elementary row transformations, we can find nonsingular matrices $\mathcal{R}_j$ such that

$$\mathcal{K}(\widetilde{\mathcal{E}}_j) := \mathcal{R}_j\mathcal{X}_j\widetilde{\mathcal{E}}_j\mathcal{Y}_j = \left[\begin{array}{c|cc} I_{(p-1)n\times(p-1)n} & 0 & * \\ \hline & I & 0 \\ & 0 & E_j^0\cdots E_{j+p-1}^0 \end{array}\right], \tag{2.18}$$

$$\mathcal{K}(\widetilde{\mathcal{A}}_j) := \mathcal{R}_j\mathcal{X}_j\widetilde{\mathcal{A}}_j\mathcal{Y}_j = \left[\begin{array}{c|cc} 0_{(p-1)n\times(p-1)n} & * & 0 \\ \hline & A_{j+p-1}^f\cdots A_j^f & 0 \\ & 0 & I \end{array}\right], \tag{2.19}$$

which implies that

$$\begin{aligned} \operatorname{ind}_\infty\left(\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j\right) &= \operatorname{ind}_\infty\left(\mathcal{K}(\widetilde{\mathcal{E}}_j), \mathcal{K}(\widetilde{\mathcal{A}}_j)\right) \\ &= \operatorname{ind}_\infty\left(\begin{bmatrix} I & 0 \\ 0 & E_j^0\cdots E_{j+p-1}^0 \end{bmatrix}, \begin{bmatrix} A_{j+p-1}^f\cdots A_j^f & 0 \\ 0 & I \end{bmatrix}\right), \end{aligned}$$

and hence, the nilpotency ν_j of the nilpotent matrix $N_j \equiv E_j^0\cdots E_{j+p-1}^0$ is equal to $\operatorname{ind}_\infty(\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j)$, for $j = 1, \ldots, p$. □

According to the result of Lemma 2.3 the indexes of the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^p$ can be defined as follows.

DEFINITION 2.1. *The indexes of regular periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^p$ are defined by*

$$\nu_j = \operatorname{ind}_\infty\left(\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j\right), \qquad j = 1, 2, \ldots, p. \tag{2.20}$$

If $\nu_j \le 1$ for all $j = 1, \ldots, p$, i.e., E_j are all nonsingular or $N_j = 0$, for all j, then the periodic matrix pairs are said to be of index at most one.

Remark. (i) It is worthwhile to point out that the indexes ν_j for regular periodic matrix pairs are not necessarily equal. For example, the periodic matrix pairs $\{(E_j, A_j)\}_{j=1}^2$ with $A_j = I_2$, $j = 1, 2$ and

$$E_1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad E_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

have indexes $\nu_1 = 1$ and $\nu_2 = 2$.

(ii) As shown in the preceding part of this paper, the monodromy matrices

$$J_j = \prod_{k=j+p-1}^{j} A_k^f \ \text{ and } \ N_j = \prod_{k=j}^{j+p-1} E_k^0, \quad j = 1, \ldots, p,$$

play an important role in the representation of solutions of (1.1). From Lemma 2.3 it is reasonable to define the indexes of $\{(E_j, A_j)\}_{j=1}^p$ by (2.20). Note that the indexes of the enlarged cyclic forms as in (1.6) are not appropriate to define the indexes of $\{(E_j, A_j)\}_{j=1}^p$. To see this, let us consider the above given data again. A short calculation gives rise to

$$\mathrm{ind}_\infty\left(\begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix}, \begin{bmatrix} 0 & A_1 \\ A_2 & 0 \end{bmatrix}\right) = \mathrm{ind}_\infty\left(\begin{bmatrix} E_2 & 0 \\ 0 & E_1 \end{bmatrix}, \begin{bmatrix} 0 & A_2 \\ A_1 & 0 \end{bmatrix}\right) = 3,$$

which is neither equal to the nilpotency of E_1E_2 nor to the nilpotency of E_2E_1.

From (2.18) and (2.19), we immediately get the following result.

COROLLARY 2.4. *If periodic matrix pairs* $\{(E_j, A_j)\}_{j=1}^p$ *are regular and of index at most* 1, *then* $\mathrm{rank}\widetilde{\mathcal{E}}_j$ *is independent of* j. *Moreover, the number of finite eigenvalues of* $\{(E_j, A_j)\}_{j=1}^p$ *is equal to* $\gamma - (p-1)n$, *where* $\gamma = \mathrm{rank}\widetilde{\mathcal{E}}_j$.

We note that if regular periodic matrix pairs have some higher indexes, then, generally speaking, $\mathrm{rank}\widetilde{\mathcal{E}}_j$ is dependent on j. This can be illustrated by the simple example. Let $p = 3$, $A_j = I_3$, and

$$E_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad E_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad E_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

It is easy to verify that $\mathrm{rank}\widetilde{\mathcal{E}}_1 = 6$ and $\mathrm{rank}\widetilde{\mathcal{E}}_2 = \mathrm{rank}\widetilde{\mathcal{E}}_3 = 7$.

According to the result of Lemma 2.3 and Definition 2.1 the following equivalent condition follows by Lemma 1 of [6] (see also [21]) immediately.

THEOREM 2.5. *The periodic matrix pairs* $\{(E_j, A_j)\}_{j=1}^p$ *are regular and of index at most* 1 *if and only if*

$$\mathrm{rank}[\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j\mathcal{S}_\infty(\widetilde{\mathcal{E}}_j)] = pn \quad \textit{for } j = 1, 2, \ldots, p. \tag{2.21}$$

For the linear time-invariant descriptor systems $Ex_{k+1} = Ax_k + Bu_k$, the condition

$$\mathrm{rank}[\lambda E - A, B] = n \quad \forall\ \lambda \in \mathbb{C} \quad \text{and} \quad \mathrm{rank}[E, A\mathcal{S}_\infty(E), B] = n \tag{2.22}$$

give sufficient conditions for the solvability of regularization and pole assignment problems [6, 21]. By Lemma 2.3 and (2.22) it is motivated to give conditions for investigating the regularization problem and pole assignment problem of the linear time-varying periodic descriptor systems (1.1).

DEFINITION 2.2. *The periodic matrix triples* $\{(E_j, A_j, B_j)\}_{j=1}^p$ *satisfy conditions* (C1) *and* (C2) *if*

$$\text{(C1):} \qquad \mathrm{rank}[\lambda\mathcal{E} - \mathcal{A}, \mathcal{B}] = pn \quad \forall\ \lambda \in \mathbb{C}; \tag{2.23}$$

$$\text{(C2):} \qquad \mathrm{rank}[\widetilde{\mathcal{E}}_j, \widetilde{\mathcal{A}}_j\mathcal{S}_\infty(\widetilde{\mathcal{E}}_j), \mathcal{B}_j] = pn \quad \textit{for } j = 1, \ldots, p. \tag{2.24}$$

Here $\mathcal{E}, \mathcal{A}$ and $\mathcal{B} \equiv B_1$ are given in (1.6) and (2.1), respectively.

Remark. A natural question can be asked here: can we extend the condition $\mathrm{rank}[E, A\mathcal{S}_\infty(E), B] = n$ directly to the enlarged cyclic triples $(\mathcal{E}, \mathcal{A}, \mathcal{B})$ by

$$(\overline{\text{C2}}): \qquad \mathrm{rank}[\mathcal{E}, \mathcal{A}\mathcal{S}_\infty(\mathcal{E}), \mathcal{B}] = pn,$$

or equivalently,

$$\operatorname{rank}[E_j, A_j\mathcal{S}_\infty(E_{j-1}), B_j] = n, \quad j = 1, \ldots, p.$$

In fact, ($\overline{\text{C2}}$) is sufficient for (C2). From the reduction in Lemma 5 of [6] and ($\overline{\text{C2}}$) we can w.l.o.g. suppose that the periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$ have the following forms:

$$E_j = \begin{bmatrix} E_{11}^j & 0 & 0 \\ E_{21}^j & E_{22}^j & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_j = \begin{bmatrix} A_{11}^j & A_{12}^j & A_{13}^j \\ A_{21}^j & A_{22}^j & A_{23}^j \\ A_{31}^j & A_{32}^j & A_{33}^j \end{bmatrix}, \quad B_j = \begin{bmatrix} 0 \\ B_2^j \\ 0 \end{bmatrix},$$

with a compatible partitioning, where E_{11}^j, B_2^j are nonsingular, E_{22}^j and $(A_{33}^j)^\top$ are of full column rank for $j = 1, \ldots, p$. It is easy to check that $\operatorname{rank}[\tilde{\mathcal{E}}_j, \tilde{\mathcal{A}}_j\mathcal{S}_\infty(\tilde{\mathcal{E}}_j), \mathcal{B}_j] = pn$, for $j = 1, \ldots, p$. However, ($\overline{\text{C2}}$) is not a necessary condition for (C2). For example, if we let $E_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right], E_2 = A_1 = A_2 = \mathrm{I}_2, B_1 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right], B_2 = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$. One can check that $\{(E_j, A_j, B_j)\}_{j=1}^2$ satisfy (C2), but not ($\overline{\text{C2}}$). Therefore, (C2) is weaker than ($\overline{\text{C2}}$). In our main theorem (Theorem 4.3) we will show that (C2) implies the periodic feedback closed-loop systems are regular and of index at most one.

The following two lemmas are simple but useful for the proof of the main result in sections 4 and 5.

LEMMA 2.6. [6] *Let* (E, A, B) *satisfy* (C1) *or* (C2) *with* $p = 1$, *where* $E, A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$. *Then*

(i) (QEP, QAP, QBV) *satisfies* (C1) *or* (C2) *for any nonsingular* P, Q, *and* V;
(ii) $(E + BG, A + BF, B)$ *satisfies* (C1) *or* (C2) *for any* G *and* $F \in \mathbb{R}^{m\times n}$ *with* $\mathcal{N}(E) \subset \mathcal{N}(E + BG)$;
(iii) $\operatorname{rank}[E, A\overline{\mathcal{S}}, B] = n$ *for any matrix in the form* $\overline{\mathcal{S}} = [\mathcal{S}_\infty(E), R]$, *where* $R \in \mathbb{R}^{n\times l}$.

LEMMA 2.7. *Assume that the periodic matrix triples* $\{(E_j, A_j, B_j)\}_{j=1}^p$ *satisfy* (C1) *or* (C2). *Then*

(i) $\{(Q_jE_jP_j,\ Q_jA_jP_{j-1},\ Q_jB_jV_j)\}_{j=1}^p$ *satisfy* (C1) *or* (C2) *for any nonsingular matrices* $P_j, Q_j \in \mathbb{R}^{n\times n}$, *and* $V_j \in \mathbb{R}^{m\times m}$;
(ii) $\{(E_j+B_jG_j, A_j+B_jF_j, B_j)\}_{j=1}^p$ *satisfy* (C1) *or* (C2) *for any matrices* $G_j, F_j \in \mathbb{R}^{m\times n}$ *with* $\mathcal{N}(\widetilde{\mathcal{E}}_j) \subset \mathcal{N}(\widetilde{\mathcal{E}}(E_j^1, \ldots, E_{j+p-1}^1; A_{j+1}^1, \ldots, A_{j+p-1}^1))$, *where* $E_j^1 = E_j + B_jG_j$ *and* $A_j^1 = A_j + B_jF_j$.

Proof. By Lemma 2.6(i) we get (i). From Lemma 2.6(ii), (iii), and (2.24), (ii) follows immediately. ∎

3. Canonical forms of $\{(E_j, A_j, B_j)\}_{j=1}^p$. In this section we present an algorithm to reduce the periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$ into canonical forms by using orthogonal and elementary transformations. In the next section we show how to exploit these canonical forms to construct the required regularizing feedback.

Before describing the algorithm we introduce some convenient notations. We denote by $\mathcal{M}(m,n)$, $\mathcal{O}(n)$, $\mathcal{L}(n)$, and $\mathcal{R}(n)$ the sets of $m\times n$ matrices, $n\times n$ orthogonal, lower triangular, and upper triangular matrices, respectively. If $m = n$, we simplify $\mathcal{M}(n) := \mathcal{M}(m,n)$. Let T be a row or column transformation which is applied to a submatrix of a given matrix. Then we use $\overline{T}$ to denote the natural extension of T to be applied to the whole matrix. For example, let

$$C = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$$

with $C_{ii} \in \mathcal{M}(n_i)$, $i = 1,2,3$. Let $Q_2 \in \mathcal{O}(n_2)$ such that $Q_2 C_{22}$ is upper triangular and

$$R_1 = \begin{bmatrix} I_{n_1} & -C_{11}^{-1} C_{13} \\ 0 & I_{n_3} \end{bmatrix},$$

which is the transformation to eliminate C_{13} by C_{11}, i.e., $[C_{11}, C_{13}]R_1 = [C_{11}, 0\,]$. Then we have

$$\overline{Q}_2 = \begin{bmatrix} I_{n_1} & & \\ & Q_2 & \\ & & I_{n_3} \end{bmatrix}, \qquad \overline{R}_1 = \begin{bmatrix} I_{n_1} & 0 & -C_{11}^{-1} C_{13} \\ 0 & I_{n_2} & 0 \\ 0 & 0 & I_{n_3} \end{bmatrix}.$$

Algorithm 3.1.

Input: periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$ with $E_j, A_j \in \mathcal{M}(n)$ and $B_j \in \mathcal{M}(n,m)$ satisfying that for $j = 1, \ldots, p$, $\operatorname{rank}(B_j) = m$ and

(3.1)

$$\operatorname{rank} \left[\begin{array}{ccccc|ccccc} -A_j & E_j & & & & B_j & & & & \\ & -A_{j+1} & E_{j+1} & & & & B_{j+1} & & & \\ & & \ddots & \ddots & & & & \ddots & & \\ & & & -A_{j+p-2} & E_{j+p-2} & & & & B_{j+p-2} \end{array}\right] = (p-1)n.$$

Output: nonsingular matrices $Q_j,\ P_j \in \mathcal{M}(n)$, feedback matrices $G_j,\ F_j \in \mathcal{M}(m,n)$, and canonical forms

(3.2)

$$Q_j(E_j + B_j G_j) P_j = \left[\begin{array}{c|c} 0_{m\times(n-l_j)} & 0_{m\times l_j} \\ \hline & \begin{array}{c} 0_{n_0^{j+1}\times l_j} \\ \hline E_{11}^j \\ & \ddots \\ & & E_{pp}^j \end{array} \end{array}\right], \quad Q_j B_j = \left[\begin{array}{c} B_{11}^j \\ \hline 0 \\ \vdots \\ 0 \end{array}\right],$$

(3.3)

$$Q_j(A_j+B_jF_j)P_{j-1}=\left[\begin{array}{cccc|cccc} \multicolumn{4}{c|}{0_{m\times(n-l_{j-1})}} & \multicolumn{4}{c}{0_{m\times l_{j-1}}} \\ \hline A^j_{L(11)} & & & & A^j_{R(11)} & & & \\ A^j_{L(21)} & A^j_{L(22)} & & & A^j_{R(21)} & A^j_{R(22)} & & \\ \vdots & & \ddots & & \vdots & & \ddots & \\ A^j_{L(p1)} & \cdots & \cdots & A^j_{L(pp)} & A^j_{R(p1)} & \cdots & \cdots & A^j_{R(pp)} \end{array}\right]$$

for $j=1,\dots,p$, where

(i) $B^j_{11}\in\mathcal{R}(m)$ nonsingular,

(ii) $E^j_{kk}\in\mathcal{M}(n^{j+1}_k)$ nonsingular for $k=1,2,\dots,p-1$,

(iii) $E^j_{pp}\in\mathcal{R}(\hat{l}_j)$ nonsingular,

(iv) $A^j_{L(kk)}=\left[\begin{smallmatrix}0\\ L^j_{kk}\end{smallmatrix}\right]\in\mathcal{M}(n^{j+1}_{k-1},m^j_k)$ with $L^j_{kk}\in\mathcal{L}(m^j_k)$ nonsingular for $k=1,2\dots,p-1$,

(v) $A^j_{R(kk)}=\left[\begin{smallmatrix}R^j_{kk}\\ 0\end{smallmatrix}\right]\in\mathcal{M}(n^{j+1}_{k-1},n^j_k)$ with $R^j_{kk}\in\mathcal{R}(n^j_k)$ nonsingular for $k=1,2\dots,p-1$,

(vi) $A^j_{L(pp)}$ and $A^j_{R(pp)}$ are $(\hat{l}_j+n^{j+1}_{p-1})\times(n-l_{j-1}-\sum_{k=1}^{p-1}m^j_k)$ and $(\hat{l}_j+n^{j+1}_{p-1})\times\hat{l}_{j-1}$ matrices, respectively.

Here l_j, n^j_k, m^j_k, and $\hat{l}_j$ are nonnegative integers, which are determined by

(3.4) $$l_j=\operatorname{rank}[\,E_j,\,B_j\,]-m,\qquad n^{j+1}_0=n-m-l_j,$$

(3.5)

$$\hat{l}_j=l_j-\sum_{k=1}^{p-1}n^{j+1}_k,$$

$$n^j_k=kn+m+l_{j-1}-\sum_{i=1}^{k-1}n^j_i$$
$$-\operatorname{rank}\left[\begin{array}{ccccc|ccccc} E_{j-1} & & & & & B_{j-1} & & & & \\ -A_j & E_j & & & & & B_j & & & \\ & \ddots & \ddots & & & & & \ddots & & \\ & & -A_{j+k-1} & E_{j+k-1} & & & & & B_{j+k-1} & \end{array}\right],$$

(3.6) $$k=1,2,\dots,p-1,$$

(3.7) $$m^j_k=n^{j+1}_{k-1}-n^j_k,\quad k=1,2,\dots,p-1.$$

Initialization Step_0:

For $j=1,\dots,p$
set $Q_j=P_j=I_n$,
endfor j;
For $j=1,\dots,p$,

(I.1) find $Q^0_j\in\mathcal{O}(n)$ such that $Q^0_jB_j=[B^j_{11}0]$ with $B^j_{11}\in\mathcal{R}(m)$ nonsingular, $Q_j:=Q^0_jQ_j$,

(I.2) partition Q_jE_j as

$$\left[\begin{array}{c}E^j_a\\ \hline E^j_b\end{array}\right]\begin{array}{l}\}\,m\\ \}\,n-m\end{array}:=Q_jE_j,$$

(I.3) find $Q_b^j \in \mathcal{O}(n-m)$ and $P_b^j \in \mathcal{O}(n)$ such that

$$Q_b^j E_b^j P_b^j = \begin{bmatrix} 0 & 0 \\ 0 & \hat{E}_{11}^j \end{bmatrix} := \hat{E}_{00}^j$$

with $\hat{E}_{11}^j \in \mathcal{R}(l_j)$ nonsingular and $l_j = \operatorname{rank}[E_j, B_j] - m$, and update

$$E_a^j := E_a^j P_b^j, \quad Q_j := \overline{Q}_b^j Q_j, \quad P_j := P_j P_b^j,$$

(I.4) set $A_j^0 := \overline{Q}_b^j A_j, \quad A_{j+1}^0 := A_{j+1} P_b^j,$

endfor j;

For $j = 1, \ldots, p,$

(I.5) partition A_j^0 as

$$\left[\begin{array}{c} A_a^j \\ \hline A_b^j \end{array}\right] \begin{matrix} \} m \\ \} n-m \end{matrix} := A_j^0,$$

and partition A_b^j as

$$\begin{array}{c} \left[\begin{array}{c|c} \hat{A}_{L(11)}^j & \hat{A}_{R(11)}^j \end{array}\right] := A_b^j, \\ \begin{array}{cc} n-l_{j-1} & l_{j-1} \end{array} \end{array}$$

(I.6) set

$$\begin{aligned} G_j &:= -(B_{11}^j)^{-1} E_a^j P_j^{-1}, & F_j &:= -(B_{11}^j)^{-1} A_a^j P_{j-1}^{-1}, \\ E_{00}^j &:= 0_{(n-m-l_j)\times(n-l_j)}, & n_0^{j+1} &:= n-m-l_j, \end{aligned}$$

endfor j.

Induction Step_k:

For $k = 1, \ldots, p-1,$

For $j = 1, 2, \ldots, p,$

(K.1) **if** $k \geq 2$, **then for** $i = 1, \ldots, k-1,$ partition $\hat{A}_{L(k,i)}^j$ and $\hat{A}_{R(k,i)}^j$, respectively, as

$$\left[\begin{array}{c} A_{L(k,i)}^j \\ \hline \hat{A}_{L(k+1,i)}^j \end{array}\right] \} n_{k-1}^{j+1} := \hat{A}_{L(k,i)}^j, \quad \left[\begin{array}{c} A_{R(k,i)}^j \\ \hline \hat{A}_{R(k+1,i)}^j \end{array}\right] \} n_{k-1}^{j+1} := \hat{A}_{R(k,i)}^j,$$

endfor i; **endif**;

(K.2) partition $[\hat{A}_{L(k,k)}^j | \hat{A}_{R(k,k)}^j]$ as

$$\left[\begin{array}{c|c} \Phi_{k,1}^j & \Phi_{k,2}^j \\ \hline \Phi_{k,3}^j & \Phi_{k,4}^j \end{array}\right] \} n_{k-1}^{j+1} := \left[\begin{array}{c|c} \hat{A}_{L(k,k)}^j & \hat{A}_{R(k,k)}^j \end{array}\right];$$

endfor j;

For $j = 1, 2, \ldots, p,$

(K.3) find $U_k^j \in \mathcal{O}(n_{k-1}^{j+1})$ and $V_k^{j-1} \in \mathcal{O}\left(n - l_{j-1} - \sum_{i=1}^{k-1} m_i^j\right)$ such that

$$U_k^j \Phi_{k,1}^j V_k^{j-1} = \left[\begin{array}{c|c} A_{L(k,k)}^j & 0 \end{array}\right],$$

where $A_{L(k,k)}^j = \begin{bmatrix} 0 \\ L_{kk}^j \end{bmatrix}$ with $L_{kk}^j \in \mathcal{L}(m_k^j)$ nonsingular, update

$$\Phi_{k,2}^j := U_k^j \Phi_{k,2}^j, \quad \Phi_{k,3}^j := \Phi_{k,3}^j V_k^{j-1}, \quad E_{k-1,k-1}^j := U_k^j E_{k-1,k-1}^j,$$
$$Q_j := \overline{U}_k^j Q_j, \quad P_{j-1} := P_{j-1} \overline{V}_k^{j-1},$$

and **if** $k \geq 2$, **then for** $i = 1, \ldots, k-1$, update

$$A_{L(k,i)}^j := U_k^j A_{L(k,i)}^j, \quad A_{R(k,i)}^j := U_k^j A_{R(k,i)}^j,$$

endfor i; **endif**;

(K.4) partition $\Phi_{k,2}^j$ as $\left[\begin{array}{c} \Phi_{k,2a}^j \\ \hline \Phi_{k,2b}^j \end{array}\right] \begin{matrix} \} m_k^j \\ \\ \end{matrix} := \Phi_{k,2}^j$,

(K.5) find an elementary transformation T_k^{j-1} to eliminate $\Phi_{k,2b}^j$ by L_{kk}^j and update

$$P_{j-1} := P_{j-1} \overline{T}_k^{j-1} \text{ and } \left[\begin{array}{c|c} \Phi_{k,3}^j & \Phi_{k,4}^j \end{array}\right] := \left[\begin{array}{c|c} \Phi_{k,3}^j & \Phi_{k,4}^j \end{array}\right] \overline{T}_k^{j-1},$$

(K.6) find $V_{k,a}^{j-1} \in \mathcal{O}(l_{j-1} - \sum_{i=1}^{k-1} n_i^j)$ such that $\Phi_{k,2a}^j V_{k,a}^{j-1} = \left[R_{kk}^j \; 0\right]$ with $R_{kk}^j \in \mathcal{R}(n_k^j)$ nonsingular, set $A_{R(k,k)}^j := \begin{bmatrix} R_{kk}^j \\ 0 \end{bmatrix}$, and update $\Phi_{k,4}^j := \Phi_{k,4}^j V_{k,a}^{j-1}$, $P_{j-1} := P_{j-1} \overline{V}_{k,a}^{j-1}$,

(K.7) find $Q_k^{j-1} \in \mathcal{O}(l_{j-1} - \sum_{i=1}^{k-1} n_i^j)$ such that $\hat{E}_{kk}^{j-1} := Q_k^{j-1}(\hat{E}_{kk}^{j-1} V_{k,a}^{j-1})$ is upper triangular, update $\left[\begin{array}{c|c} \Phi_{k,3}^{j-1} & \Phi_{k,4}^{j-1} \end{array}\right] := Q_k^{j-1} \left[\begin{array}{c|c} \Phi_{k,3}^{j-1} & \Phi_{k,4}^{j-1} \end{array}\right]$, and **if** $k \geq 2$, **then for** $i = 1, 2, \ldots, k-1$, update

$$\hat{A}_{L(k+1,i)}^{j-1} := Q_k^{j-1} \hat{A}_{L(k+1,i)}^{j-1}, \quad \hat{A}_{R(k+1,i)}^{j-1} := Q_k^{j-1} \hat{A}_{R(k+1,i)}^{j-1},$$

endfor i; **endif**;

(K.8) partition $\Phi_{k,3}^j$ and $\Phi_{k,4}^j$, respectively, as

$$\underset{m_k^j}{\left[\hat{A}_{L(k+1,k)}^j \middle| \hat{A}_{L(k+1,k+1)}^j\right]} := \Phi_{k,3}^j, \qquad \underset{n_k^j}{\left[\hat{A}_{R(k+1,k)}^j \middle| \hat{A}_{R(k+1,k+1)}^j\right]} := \Phi_{k,4}^j,$$

endfor j;

For $j = 1, 2, \ldots, p$,

(K.9) partition $\hat{E}_{k,k}^j$ as

$$\left[\begin{array}{c|c} E_{k,k}^j & \hat{E}_{k,k+1}^j \\ \hline 0 & \hat{E}_{k+1,k+1}^j \end{array}\right] \begin{matrix} \} n_k^{j+1} \\ \\ n_k^{j+1} \end{matrix} := \hat{E}_{k,k}^j,$$

(K.10) find an elementary transformation S^j_{k+1} to eliminate $\hat{E}^j_{k,k+1}$ by $\hat{E}^j_{k+1,k+1}$, update $Q_j := \overline{S}^j_{k+1} Q_j$, and **for** $i = 1, \ldots, k+1$, update

$$\hat{A}^j_{L(k+1,i)} := S^j_{k+1} \hat{A}^j_{L(k+1,i)}, \quad \hat{A}^j_{R(k+1,i)} := S^j_{k+1} \hat{A}^j_{R(k+1,i)},$$

endfor i

endfor j;

endfor k.

For $j = 1, \ldots, p$,

set $E^j_{p,p} = \hat{E}^j_{p,p}$,

for $i = 1, \ldots, p$,

set $A^j_{L(p,i)} := \hat{A}^j_{L(p,i)}, \qquad A^j_{R(p,i)} := \hat{A}^j_{R(p,i)}$,

endfor i

endfor j.

In Figure 3.1 we illustrate the canonical forms of E_{j-1}, A_j, and E_j computed by Algorithm 3.1 for the case of $p = 4$:

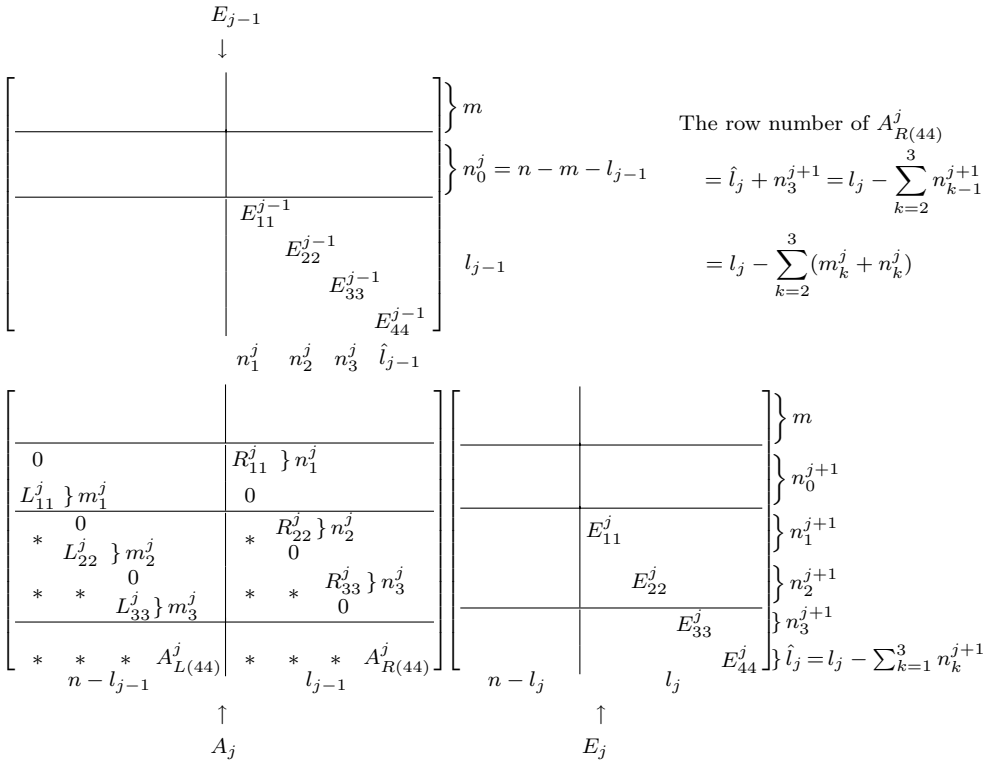


FIG. 3.1. *The canonical forms for* $p = 4$.

THEOREM 3.1. *If the periodic matrix triples* $\{(E_j, A_j, B_j)\}_{j=1}^p$ *satisfy* $\operatorname{rank}(B_j) = m$ *and conditions* (3.1)*, then the properties* (i)–(vi) *of outputs computed by Algorithm* 3.1 *hold and are completely determined by the relations given by* (3.4)–(3.7).

Remark. Note that (3.4)–(3.7) show that the sizes of submatrices E^j_{kk}, $A^j_{L(kk)}$, and $A^j_{R(kk)}$ $(k = 1, \ldots, p)$ computed by Algorithm 3.1 are uniquely determined by the original periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$.

Proof of Theorem 3.1. I. The numbers $l_j = \operatorname{rank}[E_j, B_j] - m$ and $n_0^{j+1} = n - m - l_j$ in (3.4) are obtained by (I.3) and (I.6) of Algorithm 3.1 immediately. The number $\hat{l}_j$ in (3.5) is just the size of E^j_{pp}.

The proof of (3.6). Algorithm 3.1 computes Q_j, P_j, F_j and G_j, $j = 1, \ldots, p$ such that

(3.8)
$$E_j^1 := Q_j^1(E_j + B_jG_j)P_j = \left[\begin{array}{c|c} 0_{m\times(n-l_j)} & 0_{m\times l_j} \\ \hline & \begin{array}{c} 0_{n_0^{j+1}\times l_j} \\ \hline E_b^j \end{array} \end{array}\right], \; Q_jB_j = \begin{bmatrix} B_{11}^j \\ 0 \\ \vdots \\ 0 \end{bmatrix} := B_j^1,$$

(3.9)
$$A_j^1 := Q_j(A_j + B_jF_j)P_{j-1} = \left[\begin{array}{c|c} 0_{m\times(n-l_j)} & 0_{m\times l_j} \\ \hline A_L^j & A_R^j \end{array}\right],$$

where $E_b^j := \operatorname{diag}\{E_{11}^j, \ldots, E_{p-1,p-1}^j, E_{pp}^j\}$,

$$A_L^j := \begin{bmatrix} A_{L(11)}^j & & 0 \\ \vdots & \ddots & \\ A_{L(p1)}^j & \cdots & A_{L(pp)}^j \end{bmatrix}, \quad A_R^j := \begin{bmatrix} A_{R(11)}^j & & 0 \\ \vdots & \ddots & \\ A_{R(p1)}^j & \cdots & A_{R(pp)}^j \end{bmatrix}$$

as in (3.2) and (3.3), respectively. From (I.3), (K.7), (K.9), and (K.10) it follows that $E^j_{kk} \in \mathcal{M}(n_k^{j+1})$ nonsingular, $k = 1, \ldots, p$, and E^j_{pp} is upper triangular. By (K.3) and (K.6), respectively, we have that $A^j_{L(k,k)}$ has the form $A^j_{L(kk)} = \left[\begin{smallmatrix} 0 \\ L^j_{kk} \end{smallmatrix}\right]$ with $L^j_{kk} \in \mathcal{L}(m_k^j)$ nonsingular, and $A^j_{R(kk)} = \left[\begin{smallmatrix} R^j_{kk} \\ 0 \end{smallmatrix}\right]$ with $R^j_{kk} \in \mathcal{R}(n_k^j)$ nonsingular for $k = 1, \ldots, p-1$. (We will prove that $R^j_{kk} \in \mathcal{R}(n_k^j)$ is nonsingular later!)

Consider the matrix

$$\mathcal{C}_{j,\ell}^1 = \left[\begin{array}{ccccc|ccccc} E_{j-1}^1 & & & & & B_{j-1}^1 & & & & \\ -A_j^1 & E_j^1 & & & & & B_j^1 & & & \\ & \ddots & \ddots & & & & & \ddots & & \\ & & -A_{j+\ell-2}^1 & E_{j+\ell-2}^1 & & & & & B_{j+\ell-2}^1 & \end{array}\right] \tag{3.10}$$

for j fixed and $\ell = 2, \ldots, p$.

Noting the special structures of E^1_{j+i}, A^1_{j+i}, and B^1_{j+i}, for each given ℓ we can use L^{j+i}_{kk} and R^{j+i}_{kk} $(i = \ell - 2, \ldots, 0,\ k = 1, 2, \ldots, \ell - i - 1)$ as pivots to eliminate the nonzero blocks in the same column of $A^{j+i}_{L(kk)}$ and $A^{j+i}_{R(kk)}$, respectively, of the $\mathcal{C}^1_{j,\ell}$ in (3.10) by row transformations, and finally we get the following forms, for $\ell = 2, \ldots, p$,

$$\mathcal{C}_{j,\ell}^\ell = \left[\begin{array}{ccccc|ccccc} E_{j-1}^\ell & & & & & B_{j-1}^1 & & & & \\ -A_j^\ell & E_j^\ell & & & & & B_j^1 & & & \\ & \ddots & \ddots & & & & & \ddots & & \\ & & -A_{j+\ell-2}^\ell & E_{j+\ell-2}^\ell & & & & & B_{j+\ell-2}^1 & \end{array}\right], \tag{3.11}$$

where

(3.12)
$$E^{\ell}_{j+i} = \left[\begin{array}{c|c} 0_{m\times(n-l_{j+i})} & 0_{m\times l_{j+i}} \\ \hline & \begin{array}{c} 0_{n_0^{j+i+1}\times l_{j+i}} \\ \hline E^{\ell,j+i}_b \end{array} \end{array}\right], \quad A^{\ell}_{j+i} = \left[\begin{array}{c|c} 0_{m\times(n-l_{j+i-1})} & 0_{m\times l_{j+i-1}} \\ \hline A^{\ell,j+i}_L & A^{\ell,j+i}_R \end{array}\right],$$

in which

$$E^{\ell,j+i}_b := \operatorname{diag}\{0,\dots,0,E^{j+i}_{\ell-i-1,\ell-i-1},\dots,E^{j+i}_{p,p}\} \tag{3.13}$$

for $i=-1,0,1,\dots,\ell-2$ and

$$A^{\ell,j+i}_L := \begin{bmatrix} A^{j+i}_{L(11)} & & & & & & \\ 0 & \ddots & & & & & \\ \vdots & & A^{j+i}_{L(\ell-i-1,\ell-i-1)} & & & & \\ \vdots & & 0 & A^{j+i}_{L(\ell-i,\ell-i)} & & & \\ \vdots & & \vdots & \vdots & \ddots & & \\ 0 & \cdots & 0 & A^{j+i}_{L(p,\ell-i)} & \cdots & A^{j+i}_{L(pp)} \end{bmatrix}, \tag{3.14}$$

$$A^{\ell,j+i}_R := \begin{bmatrix} A^{j+i}_{R(11)} & & & & & & \\ 0 & \ddots & & & & & \\ \vdots & & A^{j+i}_{R(\ell-i-1,\ell-i-1)} & & & & \\ \vdots & & 0 & A^{j+i}_{R(\ell-i,\ell-i)} & & & \\ \vdots & & \vdots & \vdots & \ddots & & \\ 0 & \cdots & 0 & A^{j+i}_{R(p,\ell-i)} & \cdots & A^{j+i}_{R(pp)} \end{bmatrix} \tag{3.15}$$

for $i=0,1,\dots,\ell-2$.

From the rank conditions (3.1) and (3.12)–(3.15) it is easy to derive the rank of the matrix $\mathcal{C}^{\ell}_{j,\ell}$ in (3.11), and therefore the rank of $\mathcal{C}^{1}_{j,\ell}$ in (3.10) is equal to

$$(\ell-1)n+m+l_{j-1}-\sum_{i=1}^{\ell-1} n^j_i. \tag{3.16}$$

By Lemma 2.6 it follows from (3.8), (3.9), and (3.16) that

$$\begin{aligned} n^j_{\ell-1} &= (\ell-1)n+m+l_{j-1}-\sum_{i=1}^{\ell-2} n^j_i \\ &\quad - \operatorname{rank}\left[\begin{array}{ccccc|ccccc} E_{j-1} & & & & & B_{j-1} & & & & \\ -A_j & E_j & & & & & B_j & & & \\ & \ddots & \ddots & & & & & \ddots & & \\ & & -A_{j+\ell-2} & E_{j+\ell-2} & & & & & B_{j+\ell-2} & \end{array}\right]. \end{aligned} \tag{3.17}$$

If we let $\ell - 1 = k$ in (3.17) for $k = 1, \ldots, p-1$, then (3.17) shows that (3.6) holds.

The proof of (3.7). From the rank conditions (3.1) we have that the submatrices $[-A_j^p, E_j^p, B_j^1]$ of $\mathcal{C}_{j,p}^p$ in (3.11) with $l = p$ are of full row rank, i.e., $\operatorname{rank}[-A_j^p, E_j^p, B_j^1] = n$. From the special structures of $E_b^{p,j}$, $A_L^{p,j}$, and $A_R^{p,j}$ it follows that $R_{kk}^j \in \mathcal{R}(n_k^j)$ nonsingular, for $k = 1, \ldots, p-1$ and $j = 1, \ldots, p$.

II. The properties (i)–(v) of the outputs of Algorithm 3.1 hold by (3.4), (3.5), and (3.7) immediately. To prove (vi), we denote by $r(A) = m$ and $c(A) = n$ the row and column numbers of $A \in \mathcal{M}(m, n)$, respectively. From the properties (ii), (iv), and (v) it follows that

$$\begin{aligned} r\left(A_{L(pp)}^j\right) = r\left(A_{R(pp)}^j\right) &= l_j - \sum_{k=2}^{p-1}(m_k^j + n_k^j) \\ &= l_j - \sum_{k=2}^{p-1} n_{k-1}^{j+1} \quad \text{(by (3.7))} \\ &= n_{p-1}^{j+1} + \left(l_j - \sum_{k=1}^{p-1} n_k^{j+1}\right) = n_{p-1}^{j+1} + \hat{l}_j \quad \text{(by (3.5))} \\ &= r\left(\begin{bmatrix} E_{p-1,p-1}^j & 0 \\ 0 & E_{pp}^j \end{bmatrix}\right). \end{aligned}$$

Similarly, from the properties (ii), (iv), and (v) we also have

$$c\left(A_{L(pp)}^j\right) = n - l_{j-1} - \sum_{k=1}^{p-1} m_k^j, \qquad c\left(A_{R(pp)}^j\right) = \hat{l}_{j-1} = l_{j-1} - \sum_{k=1}^{p-1} n_k^j. \qquad \square$$

4. Regularization of $\{(E_j, A_j, B_j)\}_{j=1}^p$. In this section we will use the canonical forms of $\{(E_j, A_j, B_j)\}_{j=1}^p$ computed by Algorithm 3.1 to construct derivative and proportional feedback controls so that the closed-loop systems are regular and of index at most one.

By Theorem 3.1 there are Q_j, P_j, G_j, and F_j such that

$$E_j^1 := Q_j(E_j + B_j G_j)P_j, \quad A_j^1 := Q_j(E_j + B_j F_j)P_{j-1}, \quad B_j^1 := Q_j B_j$$

have the canonical forms as in (3.2) and (3.3). For convenience, by Lemma 2.7, hereafter, we suppose w.l.o.g. that $\{(E_j, A_j, B_j)\}_{j=1}^p$ are of the canonical forms as in (3.2) and (3.3), and moreover, we assume that $\{(E_j, A_j, B_j)\}_{j=1}^p$ satisfy (C2) in (2.24).

Since the sizes of submatrices of (3.2) and (3.3) play an important role in the regularization of the periodic matrix triples, in the following lemma we will prove under condition (C2) that $c(A_{L(pp)}^j) \geq r(E_{p-1,p-1}^j)$, $j = 1, \ldots, p$.

LEMMA 4.1. *Suppose the periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$ satisfy* (C2). *Then it holds that*

$$n_{p-1}^{j+1} \leq n - l_{j-1} - \sum_{i=1}^{p-1} m_i^j, \quad j = 1, \ldots, p, \tag{4.1}$$

i.e., $r\big(E_{p-1,p-1}^j\big) \leq c\big(A_{L(pp)}^j\big)$ (*see also Figure* 3.1). *Moreover, let*

$$\delta_j := n - l_j - \sum_{i=1}^{p-1} m_i^{j+1} - n_{p-1}^{j+2} \geq 0, \tag{4.2}$$

where $\delta_0 = \delta_p$. Then it holds that

$$\sum_{j=1}^{p} \delta_j = pm. \tag{4.3}$$

Proof. Partition $\mathcal{S}_\infty(\widetilde{\mathcal{E}}_j)$ into

$$\mathcal{S}_\infty(\widetilde{\mathcal{E}}_j) = \begin{bmatrix} S_j^T, & S_{j+1}^T, \dots, & S_{j+p-1}^T \\ n & n & n \end{bmatrix}^T \tag{4.4}$$

and rewrite the equation $\widetilde{\mathcal{E}}_j \mathcal{S}_\infty(\widetilde{\mathcal{E}}_j) = 0$ in the forms

$$\begin{cases} E_j S_j = 0, \\ E_{j+\ell} S_{j+\ell} = A_{j+\ell} S_{j+\ell-1}, \quad \ell = 1, \dots, p-1. \end{cases} \tag{4.5}$$

Partitioning $S_{j+\ell-1}$ compatibly with $A_{j+\ell}$ by

$$S_{j+\ell-1} = \left[(S_{L(1)}^{j+\ell-1})^T, \dots, \quad (S_{L(p)}^{j+\ell-1})^T, \quad (S_{R(1)}^{j+\ell-1})^T, \dots, \quad (S_{R(p)}^{j+\ell-1})^T \right]^T \tag{4.6}$$

and comparing both sides of $E_{j+p-1} S_{j+p-1} = A_{j+p-1} S_{j+p-2}$ in (4.5) with $\ell = p-1$ we have

$$S_{L(1)}^{j+p-2} = 0, \quad S_{R(1)}^{j+p-2} = 0, \tag{4.7}$$

$$E_{11}^{j+p-1} S_{R(1)}^{j+p-1} = A_{L(22)}^{j+p-1} S_{L(2)}^{j+p-2} + A_{R(22)}^{j+p-1} S_{R(2)}^{j+p-2}. \tag{4.8}$$

Using (4.7) and comparing both sides of $E_{j+p-2} S_{j+p-2} = A_{j+p-2} S_{j+p-3}$ of (4.5) with $\ell = p-2$ we get

$$\begin{gathered} S_{L(1)}^{j+p-3} = 0, \quad S_{R(1)}^{j+p-3} = 0, \\ S_{L(2)}^{j+p-3} = 0, \quad S_{R(2)}^{j+p-3} = 0, \\ E_{22}^{j+p-2} S_{R(2)}^{j+p-2} = A_{L(33)}^{j+p-2} S_{L(3)}^{j+p-3} + A_{R(33)}^{j+p-3} S_{R(3)}^{j+p-3}. \end{gathered}$$

In such a way, in general, we have for each $\ell = 2, \dots, p-1$, that

$$S_{L(i)}^{j+p-\ell} = 0, \quad S_{R(i)}^{j+p-\ell} = 0, \quad i = 1, \dots, \ell-1, \tag{4.9}$$

$$E_{\ell-1,\ell-1}^{j+p-\ell+1} S_{R(\ell-1)}^{j+p-\ell+1} = A_{L(\ell\ell)}^{j+p-\ell+1} S_{L(\ell)}^{j+p-\ell} + A_{R(\ell\ell)}^{j+p-\ell+1} S_{R(\ell)}^{j+p-\ell}. \tag{4.10}$$

Finally, using (4.9) and comparing both sides of $E_j S_j = 0$ and $E_{j+1} S_{j+1} = A_{j+1} S_j$ in (4.5) with $\ell = 1$ we get

$$S_{L(i)}^{j} = S_{R(i)}^{j} = 0, \ i = 1, \dots, p-1, \ S_{R(p)}^{j} = 0, \tag{4.11}$$

$$E_{p-1,p-1}^{j+1} S_{R(p-1)}^{j+1} = A_{L(pp),a}^{j+1} S_{L(p)}^{j}, \tag{4.12}$$

where $A_{L(pp)}^{j+1}$ is partitioned into

$$A_{L(pp)}^{j+1} = \begin{bmatrix} A_{L(pp),a}^{j+1} \\ \hline A_{L(pp),b}^{j+1} \end{bmatrix} \begin{matrix} \} \, n_{p-1}^{j+2} \\ \} \, \hat{l}_{j+1} \end{matrix}. \tag{4.13}$$

On the other hand, it follows from (C2) and (4.4) that

$$\operatorname{rank}[E_j, A_j S_{j+p-1}, B_j] = n. \tag{4.14}$$

Note that the matrices E_j, A_j, and B_j are assumed to have the special structures as shown in (3.2) and (3.3). Consequently, from (4.14) we have

$$\operatorname{rank}(R_{11}^j S_{R(1)}^{j+p-1}) = n_1^j. \tag{4.15}$$

This, together with R_{11}^j nonsingular, shows that $S_{R(1)}^{j+p-1}$ has full row rank.

Now we rewrite (4.8) as

$$E_{11}^{j+p-1} S_{R(1)}^{j+p-1} = \begin{bmatrix} R_{22}^{j+p-1} S_{R(2)}^{j+p-2} \\ L_{22}^{j+p-1} S_{L(2)}^{j+p-2} \end{bmatrix},$$

which implies that $S_{R(2)}^{j+p-2}$ must have full row rank because E_{11}^{j+p-1} and R_{22}^{j+p-1} are nonsingular and $S_{R(1)}^{j+p-1}$ is of full row rank. Continuing this process, by (4.10) we can derive step by step that $S_{R(\ell)}^{j+p-\ell}$ must have full row rank for $\ell = 2, \ldots, p-1$. Finally, it follows from (4.12) and $S_{R(p-1)}^{j+1}$ of full row rank that $A_{L(pp),a}^{j+1}$ in (4.13) must have full row rank, and hence it must have

$$n_{p-1}^{j+2} \le n - l_j - \sum_{i=1}^{p-1} m_i^{j+1}.$$

Therefore, (4.1) holds, for $j = 1, \ldots, p$.

Using the equality $n_{k-1}^{j+1} = n_k^j + m_k^j$ in (3.7) it is easy to verify that

$$\sum_{j=1}^{p} \left(\sum_{i=1}^{p-1} m_i^{j+1} + n_{p-1}^{j+2} \right) = \sum_{j=1}^{p} n_0^j.$$

This, together with $n_0^j = n - m - l_{j-1}$ in (3.4), implies

$$\sum_{i=1}^{p} \delta_j = pm. \qquad \square$$

Lemma 4.1 shows that the integers $\{\delta_j\}_{j=1}^p$ in (4.2) are nonnegative and satisfy (4.3). We now use $\{\delta_j\}_{j=1}^p$ starting with a nonnegative integer r_1, recursively, to construct a sequence $\{r_j, s_j\}_{j=1}^p$ by

$$\begin{aligned} s_{j+1} &= \delta_j - r_j, \quad j = 1, \ldots, p, \\ r_{j+1} &= m - s_{j+1}, \quad j = 1, \ldots, p-1, \end{aligned} \tag{4.16}$$

where $s_1 = s_{p+1}$. Under certain condition of r_1, we will show that the integers $\{r_j, s_j\}_{j=1}^p$ defined by (4.16) are all nonnegative, which can determine the number of finite eigenvalues of periodic regularizing closed-loop systems. Let

$$L := \min_{1 \le j \le p} \left(\sum_{\ell=1}^{j-1} \delta_\ell - (j-1)m \right), \quad U := \max_{1 \le j \le p} \left(\sum_{\ell=1}^{j-1} \delta_\ell - (j-1)m \right), \tag{4.17}$$

where δ_ℓ are given by (4.2). Then the following lemma holds.

LEMMA 4.2. *If L and U defined by* (4.17) *satisfy*

$$U \le L + m, \tag{4.18}$$

and there is a nonnegative integer r_1 such that $U \le r_1 \le L + m$, then the integers of sequence $\{r_j, s_j\}_{j=1}^{p}$ defined by (4.16) *are all nonnegative and satisfy $0 \le r_j, s_j \le m$ for $j = 1, \ldots, p$.*

Proof. Since the nonnegative integer r_1 satisfies $U \le r_1 \le L+m$, from (4.17) and (4.18) we have

$$\sum_{\ell=1}^{j-1} \delta_\ell - (j-1)m \le r_1 \le \sum_{\ell=1}^{j-1} \delta_\ell - (j-2)m \tag{4.19}$$

for $j = 1, \ldots, p$. From (4.19) we recursively get

$$m \ge r_1 \ge 0, \quad s_{j+1} := \delta_j - r_j \ge 0, \quad r_{j+1} := m - s_{j+1} \ge 0, \tag{4.20}$$

for $j = 1, \ldots, p-1$. Furthermore, from (4.16) and (4.3) of Lemma 4.1 we have

$$s_1 := s_{p+1} = \delta_p - r_p = pm - \sum_{\ell=1}^{p-1} \delta_\ell - r_p = m - r_1 \ge 0. \tag{4.21}$$

This, together with (4.20), gives rise to

$$0 \le r_j,\, s_j \le m, \quad j = 1, \ldots, p. \qquad \square$$

Since the submatrix $A^{j+1}_{L(pp),a}$ of $A^{j+1}_{L(pp)}$ in (4.13) is of full row rank, there is an orthogonal matrix P^j_p such that

$$A^{j+1}_{L(pp)} P^j_p = \begin{bmatrix} \Delta^{j+1}_{p,1} & 0 \\ \Delta^{j+1}_{p,3} & \Delta^{j+1}_{p,2} \end{bmatrix}, \tag{4.22}$$

where $\Delta^{j+1}_{p,1} \in \mathcal{L}(n^{j+2}_{p-1})$ nonsingular. The Lemma 4.2 ensures that the integers s_j and r_j, $j = 1, \ldots, p$, defined by (4.16) are all nonnegative and satisfy $0 \le r_j, s_j \le m$ provided $r_1 \in [U, L+m]$. Now, for any nonnegative integer r_1 with $U \le r_1 \le L + m$ we define

(4.23)

$$G_j = (B^j_{11})^{-1} \underset{d_j}{[\,0,} \quad \underset{\delta_j}{K_j,} \quad \underset{l_j}{0\,]} P_j^T, \qquad F_j = (B^j_{11})^{-1} \underset{d_{j-1}}{[\,0,} \quad \underset{\delta_{j-1}}{H_j,} \quad \underset{l_{j-1}}{0\,]} P_{j-1}^T,$$

where

$$K_j = \begin{bmatrix} 0 & 0 \\ 0 & I_{r_j} \end{bmatrix} \in \mathcal{M}(m, \delta_j), \quad H_j = \begin{bmatrix} I_{s_j} & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{M}(m, \delta_{j-1}),$$

$$P_j = \operatorname{diag}\left\{ I_{m_1^{j+1} + \cdots + m_{p-1}^{j+1}},\ P^j_p,\ I_{l_j} \right\}, \qquad d_j = n - \delta_j - l_j.$$

Then we have

(4.24)

$$E^1_j := (E_j + B_j G_j) P_j = \left[\begin{array}{c|c|c} 0_{m \times (n - l_j - \delta_j)} & K_j & 0_{m \times l_j} \\ \hline & & 0_{n_0^{j+1} \times l_j} \\ & & \hline E^j_{11} \\ & & \qquad \ddots \\ & & \qquad\qquad E^j_{pp} \end{array}\right],$$

$$A_j^1 := (A_j + B_j F_j) P_{j-1}$$

(4.25)

$$= \left[\begin{array}{cccc|c|cccc} \multicolumn{4}{c|}{0_{m\times(n-l_{j-1}-\delta_{j-1})}} & H_j & \multicolumn{4}{c}{0_{m\times l_{j-1}}} \\ \hline A^j_{L(11)} & & & & & A^j_{R(11)} & & & \\ \vdots & \ddots & & & & \vdots & \ddots & & \\ A^j_{L(p-1,1)} & \cdots & A^j_{L(p-1,p-1)} & & & \vdots & & \ddots & \\ A^j_{L(p,1)} & \cdots & A^j_{L(p,p-1)} & \begin{bmatrix}\Delta^j_{p,1}\\ \Delta^j_{p,3}\end{bmatrix} & \begin{bmatrix}0\\ \Delta^j_{p,2}\end{bmatrix} & A^j_{R(p,1)} & \cdots & \cdots & A^j_{R(pp)} \end{array}\right].$$

We now prove our main theorem.

THEOREM 4.3. *Suppose the periodic matrix triples* $\{(E_j, A_j, B_j)\}_{j=1}^p$ *satisfy* (C2) *and* $U \le L+m$, *where* U *and* L *are given by* (4.17). *Then for any nonnegative integer* r_1 *with* $U \le r_1 \le L+m$, *the feedback matrices* G_j, F_j *constructed by* (4.23) *make the periodic closed-loop systems* $\{(E_j+B_jG_j, A_j+B_jF_j)\}_{j=1}^p$ *regular and have the index at most one. Moreover, the number of finite eigenvalues of* $\{(E_j+B_jG_j, A_j+B_jF_j)\}_{j=1}^p$ *is equal to*

$$f := r_j + \hat{l}_j = r_j + l_j - \sum_{k=1}^{p-1} n_k^{j+1}$$

for any $j \in \{1, \dots, p\}$.

Proof. Let

$$E_j^0 := E_j + B_jG_j, \quad A_j^0 := A_j + B_jF_j, \quad j = 1, \dots, p, \tag{4.26}$$

and let

$$\begin{aligned} \widetilde{\mathcal{E}}_j^0 &:= \widetilde{\mathcal{E}}(E_j^0, \dots, E_{j+p-1}^0, A_{j+1}^0, \dots, A_{j+p-1}^0), \widetilde{\mathcal{A}}_j^0 := \widetilde{\mathcal{A}}(A_j^0), \\ \widetilde{\mathcal{E}}_j^1 &:= \widetilde{\mathcal{E}}(E_j^1, \dots, E_{j+p-1}^1, A_{j+1}^1, \dots, A_{j+p-1}^1), \widetilde{\mathcal{A}}_j^1 := \widetilde{\mathcal{A}}(A_j^1) \end{aligned} \tag{4.27}$$

for $j = 1, \dots, p$. Since

$$\operatorname{rank}[\widetilde{\mathcal{E}}_j^0 \mathcal{P}_j,\ (\widetilde{\mathcal{A}}_j^0 \mathcal{P}_j)\mathcal{P}_j^T \mathcal{S}_\infty(\widetilde{\mathcal{E}}_j^0)] = \operatorname{rank}[\widetilde{\mathcal{E}}_j^1,\ \widetilde{\mathcal{A}}_j^1 \mathcal{S}_\infty(\widetilde{\mathcal{E}}_j^1)],$$

where $\mathcal{P}_j = \operatorname{diag}\{P_1, \dots, P_{j+p-1}\}$, by Theorem 2.5, in order to prove Theorem 4.3, it is sufficient to prove that

$$\operatorname{rank}[\widetilde{\mathcal{E}}_j^1, \widetilde{\mathcal{A}}_j^1 \mathcal{S}_\infty(\widetilde{\mathcal{E}}_j^1)] = pn, \quad j = 1, \dots, p. \tag{4.28}$$

For simplicity, here we only prove (4.28) for $j = 1$, but the others can be shown in a similar way. For $j = 1$, we first construct a basis for the null space $\mathcal{N}(\widetilde{\mathcal{E}}_1^1)$ of the forms

$$\mathcal{S}_\infty(\widetilde{\mathcal{E}}_1^1) \equiv \mathcal{S}_1 = \left[S_1^T, S_2^T, \dots, S_p^T\right]^T \tag{4.29}$$

with

$$S_k = [0_{n\times(n-l_p-r_p)}, S_{k1}, \dots, S_{kk}, 0, \dots, 0], \quad k = 1, \dots, p-1, \tag{4.30}$$

and

$$S_p = \left[\begin{array}{c|c|c|c} I_{n-l_p-r_p} & & & \\ 0 & & & \\ \vdots & S_{p1} & \cdots & S_{p,p-1} \\ 0 & & & \end{array}\right], \tag{4.31}$$

where

(4.32)

$$S_{kk} = \begin{array}{c} \left[\begin{array}{cc} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ S^k_{L(p)} & 0 \\ \hline 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{array}\right] \\ \begin{array}{cc} n^{k+2}_{p+2} & g^k_1 \end{array} \end{array} \begin{array}{l} m^{k+1}_1 \\ \vdots \\ m^{k+1}_{p-1} \\ n - l_k - \sum_{i=1}^{p-1} m^{k+1}_i \\ n^{k+1}_1 \\ \vdots \\ n^{k+1}_{p-1} \\ l_k - \sum_{i=1}^{p-1} n^{k+1}_i \end{array}, \qquad S^k_{L(p)} = \begin{bmatrix} I_{n^{k+2}_{p-1}} \\ 0 \end{bmatrix}, \quad g^k_1 = n^1_k - n^{k+2}_{p-1},$$

(4.33)

$$S_{k+i,k} = \begin{array}{c} \left[\begin{array}{ccc} 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ 0 & S^{k+i}_{L(p-i)} & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ S^{k+i}_{R(p-i)} & 0 & 0 \\ * & 0 & 0 \\ \vdots & \vdots & \vdots \\ * & 0 & 0 \end{array}\right] \\ \begin{array}{ccc} n^{k+i+1}_{p-i} & n^{k+i+1}_{p-i} & g^k_{i+1} \end{array} \end{array} \begin{array}{l} m^{k+i+1}_1 \\ \vdots \\ m^{k+i+1}_{p-i-1} \\ m^{k+i+1}_{p-i} \\ m^{k+i+1}_{p-i+1} \\ \vdots \\ n - l_{k+i} - \sum_{\nu=1}^{p-1} m^{k+i+1}_\nu \\ n^{k+i+1}_1 \\ \vdots \\ n^{k+i+1}_{p-i-1} \\ n^{k+i+1}_{p-i} \\ n^{k+i+1}_{p-i+1} \\ \vdots \\ l_{k+i} - \sum_{\nu=1}^{p-1} n^{k+i+1}_\nu \end{array}, \qquad \begin{array}{l} S^{k+i}_{L(p-i)} = I_{m^{k+i+1}_{p-i}}, \\ g^k_{i+1} = n^1_k - n^{k+i+2}_{p-i-1}, \\ k = 1, 2, \ldots, p-1, \\ i = 1, 2, \ldots, p-k-1, \end{array}$$

and

$$S_{p,p-i} = \big[0, \ldots, 0, \underset{i's}{|\, 0, \ldots, 0,} (S^p_{R(p-i)})^T, *, \ldots, *\big]^T, \quad i = 1, \ldots, p-1, \tag{4.34}$$

in which the submatrices $S^{k+i}_{R(p-i)}$, $k = 1, \ldots, p-1$, $i = 1, \ldots, p-k-1$, in (4.33) and $S^p_{R(p-i)}$, $i = 1, \ldots, p-1$, in (4.34) are to be determined.

From (4.29), (4.24), and (4.25) we have, for $k = 1, \ldots, p-1$,

$$E_{k+1}^1 \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline 0 \\ \vdots \\ 0 \\ S_{R(p-1)}^{k+1} \\ * \end{array}\right] = A_{k+1}^1 \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ S_{L(p)}^{k} \\ \hline 0 \\ \vdots \\ 0 \end{array}\right], \tag{4.35}$$

especially

$$E_{p-1,p-1}^{k+1} S_{R(p-1)}^{k+1} = \Delta_{p,1}^{k+1};$$

thus, $S_{R(p-1)}^{k+1} \in \mathcal{M}(n_{p-1}^{k+2})$ and "$*$" below it are uniquely determined with $S_{R(p-1)}^{k+1}$ nonsingular, since $E_{p-1,p-1}^{k+1}$, $E_{p,p}^{k+1}$ and $\Delta_{p,1}^{k+1}$ are nonsingular. Especially, taking $k = p-1$ in (4.35) we have that the matrix $S_{p,p-1}$ in (4.34) is uniquely determined with $S_{R(p-1)}^{p}$ nonsingular. Again, from (4.29), (4.24), and (4.25) we have, for $k = 1, \ldots, p-2$,

$$E_{k+2}^1 \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline 0 \\ \vdots \\ 0 \\ S_{R(p-2)}^{k+2} \\ * \\ * \end{array}\right] = A_{k+2}^1 \left[\begin{array}{cc} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & S_{L(p-1)}^{k+1} \\ 0 & 0 \\ \hline 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ S_{R(p-1)}^{k+1} & 0 \\ * & 0 \end{array}\right], \tag{4.36}$$

especially

$$E_{p-2,p-2}^{k+2} S_{R(p-2)}^{k+2} = \begin{bmatrix} R_{p-1,p-2}^{k+2} S_{R(p-1)}^{k+1} & 0 \\ 0 & L_{p-1,p-1}^{k+2} S_{L(p-1)}^{k+1} \end{bmatrix}.$$

This, together with $R_{p-1,p-1}^{k+1}$, $L_{p-1,p-1}^{k+1}$, and $S_{L(p-1)}^{k+1} = I_{m_{p-1}^{k+2}}$ all nonsingular, implies that the matrix $S_{R(p-2)}^{k+2} \in \mathcal{M}(n_{p-2}^{k+3})$ and all "$*$" below it are uniquely determined with $S_{R(p-2)}^{k+2}$ nonsingular. Especially, taking $k = p-2$ in (4.36) we have the matrix $S_{p,p-2}$ in (4.34) is also uniquely determined with $S_{R(p-2)}^{p}$ nonsingular.

In general, for $i = 2, \ldots, p-1$, $k = 1, \ldots, p-i$, comparing the both sides of

$$E^1_{k+i}\left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline 0 \\ \vdots \\ 0 \\ S^{k+i}_{R(p-i)} \\ * \\ \vdots \\ * \end{array}\right]\Bigg\}i's = A^1_{k+i}\left[\begin{array}{cc} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & S^{k+i+1}_{L(p-i-1)} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \hline 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ S^{k+i+1}_{R(p-i-1)} & 0 \\ * & 0 \\ \vdots & \vdots \\ * & 0 \end{array}\right]\Bigg\}(i-1)'s \tag{4.37}$$

step by step, the matrices $S^{k+i}_{R(p-i)} \in \mathcal{M}(n^{k+i+1}_{p-i})$ and all "$*$" below it are uniquely determined with $S^{k+i}_{R(p-i)}$ nonsingular. Furthermore, the matrix $S_{p,p-i}$ in (4.34) is then uniquely determined with $S^p_{R(p-i)}$ nonsingular.

From (4.32) and (4.35)–(4.37) we have, for each $k = 1, \ldots, p-1$, that

$$E^1_k S_{kk} = 0 \ \text{ and } \ E^1_{k+i} S_{k+i,k} = A^1_{k+i} S_{k+i-1,k}, \quad i = 1, \ldots, p-k, \tag{4.38}$$

and therefore

$$\widetilde{\mathcal{E}}^1_1 \mathcal{S}_1 = 0. \tag{4.39}$$

Moreover, a short calculation gives rise to

(4.40)

$$A^1_1 S_p = \left[\begin{array}{ccccc|c|ccc} \multicolumn{5}{c|}{0_{m\times(n-l_p-r_p)}} & \begin{bmatrix} I_{s_1} \\ 0 \end{bmatrix} & \multicolumn{3}{c}{0_{m\times(l_p-\hat{l}_p)}} \\ \hline A^1_{L(11)} & & & & & & \hat{A}^1_{R(11)} & & \\ \vdots & \ddots & & & & & \vdots & \ddots & \\ A^1_{L(p-1,1)} & \cdots & A^1_{L(p-1,p-1)} & & & & * & \cdots & \hat{A}^1_{R(p-1,p-1)} \\ A^1_{L(p1)} & \cdots & A^1_{L(p,p-1)} & \begin{bmatrix} \Delta^1_{p,1} \\ \Delta^1_{p,3} \end{bmatrix} & & \begin{bmatrix} 0 \\ * \end{bmatrix} & * & \cdots & * \end{array}\right],$$

where $\hat{A}^1_{R(kk)} = A^1_{R(kk)} S^p_{R(k)} = \left[\begin{smallmatrix} R^1_{kk} S^p_{R(k)} \\ 0 \end{smallmatrix}\right]$, $k = 1, \ldots, p-1$, with $R^1_{kk} S^p_{R(k)}$ nonsingular. From (4.40), and the special structures of E^1_k and A^1_k, $k = 1, \ldots, p$, as in (4.24) and (4.25), similarly to the proof of (3.16) we can derive that

$$\operatorname{rank}[\widetilde{\mathcal{E}}^1_1,\ \widetilde{\mathcal{A}}^1_1 \mathcal{S}_1] = pn.$$

This, together with $\mathcal{S}_1 = \mathcal{S}_\infty(\widetilde{\mathcal{E}}_1^1)$, implies that (4.28) holds for $j = 1$. For $j = 2, \ldots, p$, (4.28) can also be proved in a similar way. Furthermore, by Corollary 2.4 the periodic matrix pairs $\{(E_j + B_jG_j, A_j + B_jF_j)\}_{j=1}^p$ have $\gamma - (p-1)n$ finite eigenvalues, where

$$\gamma \equiv \gamma_j = \operatorname{rank}(\widetilde{\mathcal{E}}_j^1)$$

for any $j \in \{1, \ldots, p\}$. From (3.14), (3.15), (4.24), (4.25), and (4.16) one can easily see that $\gamma = r_j + \hat{l}_j + (p-1)n$; therefore, $\{(E_j + B_jG_j, A_j + B_jF_j)\}_{j=1}^p$ have $r_j + \hat{l}_j = r_j + l_j - \sum_{k=1}^{p-1} n_k^{j+1}$ finite eigenvalues for any $j \in \{1, \ldots, p\}$. □

Remark. (i) From Theorem 4.3 we see that the number of finite eigenvalues of $\{(E_j + B_jG_j, A_j + B_jF_j)\}_{j=1}^p$ is equal to $\hat{l}_1 + r_1$, where r_1 is any given nonnegative integer with $U \le r_1 \le L + m$. By Lemma 4.2 the integers $\{r_j, s_j\}_{j=1}^p$ defined by (4.16) are all nonnegative and satisfy $0 \le r_j, s_j \le m$ for $j = 1, \ldots, p$. If $r_1 = U$ and there is some $r_k = 0$, then at the time k we can only use the proportional feedback control $u_k = F_kx_k$ to regularize the periodic systems. If $r_1 = L + m$ and there are some $s_k = 0$, then at the time k we can only use the derivative feedback control $u_k = G_kx_{k+1}$ to regularize the periodic systems.

(ii) For the case of $p = 1$, Theorem 4.3 can be simplified to the result that for any given integer r_1 with $0 \le r_1 \le m$, i.e., $\ell_1 \le r_1 + \ell_1 \le m + \ell_1 = \operatorname{rank}([E, B])$ (here $\hat{\ell}_1 = \ell_1$), there exist matrices F, $G \in \mathcal{M}(m, n)$ such that $(E + BG, A + BF)$ is regular and of index at most one, and $\operatorname{rank}(E + BG) = r_1 + \ell_1 =$ the number of finite eigenvalues, which is equivalent to Theorem 6 of the main results of [6].

5. Pole assignment of periodic descriptor systems. In this section, we will study solvability for pole assignment problem of the resulting periodic regular descriptor systems in section 4. The problem of pole assignment is stated as follows.

Problem I. Given periodic matrix triples $\{(E_j, A_j, B_j)\}_{j=1}^p$ and a set

$$\mathcal{L} = \{(\pi_{\alpha_1}, \pi_{\beta_1}), \ldots, (\pi_{\alpha_n}, \pi_{\beta_n})\},$$

closed under conjugation, where $(\pi_{\alpha_i}, \pi_{\beta_i}) \in \mathbb{C}^2$ for $i = 1, \ldots, n$, find $G_j, F_j \in \mathcal{M}(m, n), j = 1, \ldots, p$, such that the periodic matrix pairs $\{(E_j + B_jG_j, A_j + B_jF_j)\}_{j=1}^p$ are regular and have all eigenvalue pairs in $\mathcal{L}$.

In practice, the number of infinite poles to be prescribed is limited and it is not desirable to assign finite poles to infinite positions. Thus, we construct the periodic feedback matrices G_j and $F_j, j = 1, \ldots, p$, such that the periodic closed-loop systems not only are regular and have the required finite poles, but also have index at most one. We have the following result for finite pole assignment.

THEOREM 5.1. *If the periodic matrix triples* $\{(E_j, A_j, B_j)\}_{j=1}^p$ *satisfy* (C1) *and* (C2) *and* $U \le L + m$, *where* U, L *are given by* (4.17), *then for any arbitrary set* $\mathcal{L}$ *of* γ *self-conjugate finite poles* $(\pi_{\alpha_i}, \pi_{\beta_i}), \pi_{\beta_i} \ne 0, i = 1, \ldots, f$, *and* $n - f$ *infinite poles* $(\pi_{\alpha_i}, 0), i = f + 1, \ldots, n$, *where* $U + \hat{l}_1 \le f \le L + m + \hat{l}_1$, *there exist periodic feedback matrices* G_j *and* $F_j, j = 1, \ldots, p$ *solving the pole assignment problem, Problem* I, *such that* $\{(E_j + B_jG_j, A_j + B_jF_j)\}_{j=1}^p$ *are regular and of index at most one.*

Proof. By Theorem 4.3 there exist G_j and $F_j^1, j = 1, \ldots, p$, such that the periodic closed-loop systems $\{(E_j + B_jG_j, A_j + B_jF_j^1)\}_{j=1}^p$ are regular and of index at most one, and

$$\operatorname{rank}[\widetilde{\mathcal{E}}(E_1 + B_1G_1, \ldots, E_p + B_pG_p; A_2 + B_2F_2^1, \ldots, A_p + B_pF_p^1)] = f + (p-1)n,$$

where $U + \hat{l}_1 \le f \le L + m + \hat{l}_1$. Moreover, by Lemma 2.7 the periodic matrix triples $\{(E_j + B_jG_j, A_j + B_jF_j^1)\}_{j=1}^p$ satisfy (C1) and (C2). From the results of [3] and

[22], it follows that there exist $F_j^2, j = 1, \ldots, p$, which assign f finite poles to these periodic systems while preserving precisely the remaining $n-f$ infinite poles invariant and such that the periodic closed-loop systems $\{(E_j + B_j G_j, A_j + B_j F_j)\}_{j=1}^p$, with $F_j = F_j^1 + F_j^2, j = 1, \ldots, p$, are regular and of index at most one. □

Remark. As for Remark (ii) of Theorem 4.3, in the case $p = 1$, Theorem 5.1 can also be reduced to the result of Theorem 14 of [6].

6. Conclusion. In this paper, we construct derivative and proportional state feedback controls so that the periodic closed-loop systems not only are regular and have the required finite poles, but also have index at most one. This property ensures the solvability of the resulting periodic closed-loop systems of the dynamic-algebraic equation. For the time-invariant case our main theorems can be simplified to the main results of [6]. The construction procedures are based on orthogonal and elementary transformations which can be used to develop an algorithm implementing in a numerically efficient way.

In practice it is expected that the periodic regularizing closed-loop systems are well-conditioned in the sense that the reduction to the periodic canonical forms is computationally reliable. How to develop a computational algorithm which optimizes the conditioning of the periodic regularizing closed-loop systems is currently still under investigation.

REFERENCES

[1] M. C. Berg, N. Amit, and J. D. Powell, *Multirate digital control system design*, IEEE Trans. Automat. Control, 33 (1988), pp. 1139–1150.

[2] S. Bittanti, *Deterministic and stochastic linear periodic systems*, in Time Series and Linear Systems, S. Bittanti, ed., Springer-Verlag, Berlin, 1986, pp. 141–182.

[3] S. Bittanti and P. Bolzern, *Discrete-time linear periodic systems: Gramian and modal criteria for reachability and controllability*, Internat. J. Control, 41 (1985), pp. 909–928.

[4] S. Bittanti, P. Colaneri, and G. de Nicolao, *The difference periodic Riccati equation for the periodic prediction problem*, IEEE Trans. Automat. Control, 33 (1988), pp. 706–712.

[5] A. Bojanczyk, G. Golub, and P. Van Dooren, *The periodic Schure decomposition. Algorithms and applications*, in Proceedings of the SPIE Conference, San Diego, 1770, 1992, pp. 31–42.

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

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

[8] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Feedback design for regularizing descriptor systems*, Linear Algebra Appl., 299 (1999), pp. 119–151.

[9] R. Byers, T. Greerts, and V. Mehrmann, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35 (1997), pp. 462–479.

[10] R. Byers, P. Kunkel, and V. Mehrmann, *Regularization of linear descriptor systems with variable coefficients*, SIAM J. Control Optim., 35 (1997), pp. 117–133.

[11] D. L. Chu, H. C. Chan, and D. W. C. Ho, *Regularization of singular systems by derivative and proportional output feedback*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 21–38.

[12] D. L. Chu, V. Mehrmann, and N. K. Nichols, *Minimum norm regularization of descriptor systems by mixed output feedback*, Linear Algebra Appl., 296 (1999), pp. 39–77.

[13] D. L. Chu and V. Mehrmann, *Disturbance decoupling for descriptor systems by state feedback*, SIAM J. Control Optim., 38 (2000), pp. 1830–1858.

[14] E. R. Ferrara, *Frequency-domain implementations of periodically time-varying filters*, IEEE Trans. Acoust. Speech Signal Process, 33 (1985), pp. 833–892.

[15] D. S. Flamm and A. J. Laub, *A new shift-invariant representation of periodic linear systems*, Systems Control Lett., 17 (1991), pp. 9–14.

[16] B. Francis and T. Georgiou, *Stability theory for linear time-invariant plants with periodic digital controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 820–832.

[17] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.
[18] T. Geerts, *Solvability conditions, consistency, and weak consistency for linear differential-algebraic equations and time-invariant linear systems: The general case*, Linear Algebra Appl., 181 (1993), pp. 111–130.
[19] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[20] J. J. Hench and A. J. Laub, *Numerical solution of the discrete-time periodic Riccati equation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1197–1210.
[21] J. Kautsky, N. K. Nichols, and E. K.-W. Chu, *Robust pole assignment in singular control systems*, Linear Algebra Appl., 121 (1989), pp. 9–37.
[22] M. Kono, *Eigenvalue assignment in periodic discrete-time system*, Internat. J. Control, 32 (1980), pp. 149–158.
[23] P. Kunkel, V. Mehrmann, and W. Rath, *Analysis and numerical solution of control problems in desciptor form*, Math. Control Signals Systems, 14 (2001), pp. 29–61.
[24] W. W. Lin, P. Van Dooren, and Q. F. Xu, *Equivalent Characterizations of Periodic Invariant Subspaces*, NCTS-TR-0215, National Tsing Hua Unv., Taiwan, 2002.
[25] W. W. Lin and J. G. Sun, *Perturbation analysis for the eigenproblem of periodic matrix pairs*, Linear Algebra Appl., 337 (2002), pp. 157–187.
[26] M. L. Liou and Y. L. Kuo, *Exact analysis of switched capacitor circuits with arbitrary inputs*, IEEE Trans. Circuits and Systems, 26 (1979), pp. 213–223.
[27] J. A. Richards, *Analysis of Periodically Time-Varying Systems*, Springer-Verlag, Berlin, 1983.
[28] J. Sreedhar and P. Van Dooren, *Forward/backward decomposition of periodic descriptor systems*, in Proc. 1997 ECC, Brussels, Belgium, paper FR-A-L7.
[29] J. Sreedhar and P. Van Dooren, *Periodic descriptor systems: Solvability and conditionability*, IEEE Trans. Automat. Control, 44 (1999), pp. 310–313.
[30] P. Van Dooren and J. Sreedhar, *When is a periodic discrete-time system equivalent to a time-invariant one?*, Linear Algebra Appl., 212/213 (1994), pp. 131–151.
[31] A. Varga, *Balancing related methods for minimal realization of periodic systems*, Systems Control Lett., 36 (1999), pp. 339–349.
[32] A. Varga, *Robust and minimum norm pole assignment with periodic state feedback*, IEEE Trans. Automat. Control, 45 (2000), pp. 1017–1022.
[33] J. Vlach, K. Singhal, and M. Vlach, *Computer oriented formulation of equations and analysis of switched-capacitor networks*, IEEE Trans. Circuits and Systems, 31 (1984), pp. 735–765.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1074–1109

CONVERGENCE OF RESTARTED KRYLOV SUBSPACES TO INVARIANT SUBSPACES*

CHRISTOPHER BEATTIE†, MARK EMBREE‡, AND JOHN ROSSI†

Abstract. The performance of Krylov subspace eigenvalue algorithms for large matrices can be measured by the angle between a desired invariant subspace and the Krylov subspace. We develop general bounds for this convergence that include the effects of polynomial restarting and impose no restrictions concerning the diagonalizability of the matrix or its degree of nonnormality. Associated with a desired set of eigenvalues is a maximum "reachable invariant subspace" that can be developed from the given starting vector. Convergence for this distinguished subspace is bounded in terms involving a polynomial approximation problem. Elementary results from potential theory lead to convergence rate estimates and suggest restarting strategies based on optimal approximation points (e.g., Leja or Chebyshev points); exact shifts are evaluated within this framework. Computational examples illustrate the utility of these results. Origins of superlinear effects are also described.

Key words. Krylov subspace methods, Arnoldi algorithm, Lanczos algorithm, polynomial restarts, invariant subspaces, eigenvalues, pseudospectra, perturbation theory, potential theory, Zolotarev-type polynomial approximation problems

AMS subject classifications. 15A18, 15A42, 31A15, 41A25, 65F15

DOI. 10.1137/S0895479801398608

1. Setting.

Let $\mathbf{A}$ be an $n \times n$ complex matrix with $N \le n$ distinct eigenvalues $\{\lambda_j\}_{j=1}^N$ with corresponding eigenvectors $\{\mathbf{u}_j\}_{j=1}^N$. (We do not label multiple eigenvalues separately and make no assertion regarding the uniqueness of the $\mathbf{u}_j$.) Each distinct eigenvalue λ_j has geometric multiplicity n_j and algebraic multiplicity m_j (so that $1 \le n_j \le m_j$ and $\sum_{j=1}^N m_j = n$). We aim to compute an invariant subspace associated with L of these eigenvalues, which for brevity we call the *good* eigenvalues, labeled $\{\lambda_1, \lambda_2, \ldots, \lambda_L\}$. We intend to use a Krylov subspace algorithm to approximate this invariant subspace, possibly with the aid of restarts as described below. The remaining $N - L$ eigenvalues, the *bad* eigenvalues, are not of interest and we wish to avoid excessive expense involved in inadvertently calculating the subspaces associated with them.

The class of algorithms considered here draws eigenvector approximations from Krylov subspaces generated by the starting vector $\mathbf{v}_1 \in \mathbb{C}^n$,

$$\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1) = \operatorname{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}^{\ell-1}\mathbf{v}_1\}.$$

Such algorithms, including the Arnoldi and biorthogonal Lanczos methods reviewed in section 1.1, differ in their mechanisms for generating a basis for $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ and selecting approximate eigenvectors from this Krylov subspace. Though these approximate eigenvectors are obvious objects of study, their convergence can be greatly complicated by eigenvalue multiplicity and defectiveness; see [21]. The bounds developed in

*Received by the editors November 21, 2001; accepted for publication (in revised form) by Z. Strakoš June 9, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/simax/25-4/39860.html

†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061–0123 (beattie@math.vt.edu, rossi@math.vt.edu).

‡Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK. Current address: Department of Computational and Applied Mathematics, Rice University, 6100 Main Street—MS 134, Houston, TX 77005–1892 (embree@caam.rice.edu). The research of this author was supported in part by UK Engineering and Physical Sciences Research Council Grant GR/M12414.

the following sections avoid these difficulties by instead studying convergence of the Krylov subspace to an invariant subspace associated with the good eigenvalues as the dimension of the Krylov subspace is increased.

Given two subspaces, $\mathcal{W}$ and $\mathcal{V}$ of $\mathbb{C}^n$, the extent to which $\mathcal{V}$ approximates $\mathcal{W}$ is measured (asymmetrically) by the *containment gap* (or just *gap*), defined as

$$\delta(\mathcal{W},\mathcal{V}) = \sup_{\mathbf{x}\in\mathcal{W}} \inf_{\mathbf{y}\in\mathcal{V}} \frac{\|\mathbf{y}-\mathbf{x}\|}{\|\mathbf{x}\|} = \sin(\vartheta_{\max}).$$

Here $\vartheta_{\max}$ is the largest canonical angle between $\mathcal{W}$ and a "closest" subspace $\widehat{\mathcal{V}}$ of $\mathcal{V}$ having dimension equal to $\dim \mathcal{W}$. (Throughout, $\|\cdot\|$ denotes the vector 2-norm and the matrix norm it induces.) Notice that if $\dim \mathcal{V} < \dim \mathcal{W}$, then $\delta(\mathcal{W},\mathcal{V}) = 1$, while $\delta(\mathcal{W},\mathcal{V}) = 0$ if and only if $\mathcal{W} \subseteq \mathcal{V}$. The gap can be expressed directly as the norm of a composition of projections: If $\mathbf{\Pi}_{\mathcal{W}}$ and $\mathbf{\Pi}_{\mathcal{V}}$ denote orthogonal projections onto $\mathcal{W}$ and $\mathcal{V}$, respectively, then $\delta(\mathcal{W},\mathcal{V}) = \|(\mathbf{I}-\mathbf{\Pi}_{\mathcal{V}})\mathbf{\Pi}_{\mathcal{W}}\|$ (see, e.g., Chatelin [7, sect. 1.4]).

The objective of this paper then is to measure the gap between Krylov subspaces and an m-dimensional invariant subspace $\mathcal{U}$ of $\mathbf{A}$ associated with the good eigenvalues. We explore how quickly $\delta(\mathcal{U},\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1))$ can be driven to zero as ℓ is increased, reflecting the speed of convergence, and how this behavior is influenced by the distribution of eigenvalues and nonnormality of $\mathbf{A}$. Note that $\delta(\mathcal{U},\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)) = 1$ when $\ell < m$. For $\ell \geq m$, our bounds ultimately take the form

$$\delta(\mathcal{U},\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)) \leq C_0\, C_1\, C_2 \min_{\phi\in\mathcal{P}_{\ell-m}} \frac{\max\{|\phi(z)| : z\in\Omega_{\text{bad}}\}}{\min\{|\phi(z)| : z\in\Omega_{\text{good}}\}}, \tag{1.1}$$

where $\mathcal{P}_\ell$ is the set of degree-ℓ polynomials, and Ω_{good} and Ω_{bad} are disjoint compact subsets of $\mathbb{C}$ containing the good and bad eigenvalues, respectively. The constant C_0 reflects nonnormal coupling between good and bad invariant subspaces, while C_2 reflects nonnormality within those two subspaces. The constant C_1 principally describes the effect of starting vector bias, though it, too, is influenced by nonnormality. In section 2 we identify the subspace $\mathcal{U}$, which in common situations will be the entire invariant subspace of $\mathbf{A}$ associated with the good eigenvalues, but will be smaller when $\mathbf{A}$ is derogatory or the starting vector $\mathbf{v}_1$ is deficient. The basic bound (1.1) is derived in section 3. Section 4 addresses the polynomial approximation problem embedded in (1.1), describing those factors that determine linear convergence rates or that lead to superlinear effects. Section 5 analyzes the constants C_1 and C_2, and section 6 provides computational examples illustrating the bounds.

Since it becomes prohibitively expensive to construct and store a good basis for $\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)$ when the dimension of $\mathbf{A}$ is large, practical algorithms typically limit the maximum dimension of the Krylov subspace to some $p \ll n$. If satisfactory estimates cannot be extracted from $\mathcal{K}_p(\mathbf{A},\mathbf{v}_1)$, then the algorithm is *restarted* by replacing $\mathbf{v}_1$ with some new $\mathbf{v} \in \mathcal{K}_p(\mathbf{A},\mathbf{v}_1)$ that is, one hopes, enriched in the component lying in the subspace $\mathcal{U}$. Since this $\mathbf{v}$ is chosen from the Krylov subspace, we can write $\mathbf{v} = \psi(\mathbf{A})\mathbf{v}_1$ for some polynomial ψ with $\deg(\psi) < p$. Our bounds also apply to this situation, and ideas from potential theory, outlined in section 4, motivate particular choices for the polynomial ψ.

The results presented here complement and extend earlier convergence theory, beginning with Saad's bound on the gap between a single eigenvector and the Krylov subspace for a matrix with simple eigenvalues [32]. Jia generalized this result to invariant subspaces associated with a single eigenvalue of a defective matrix, but

these bounds involve the Jordan form of $\mathbf{A}$ and derivatives of approximating polynomials [20]. Simoncini uses pseudospectra to describe block-Arnoldi convergence for defective matrices [37]. Interpreting restarted algorithms in terms of subspace iteration, Lehoucq developed an invariant subspace convergence theory incorporating results from Watkins and Elsner [25]. Calvetti, Reichel, and Sorensen studied single eigenvector convergence for Hermitian matrices using elements of potential theory [6]. A key feature of our approach is its applicability to general invariant subspaces, which may be better conditioned than individual eigenvectors (see, e.g., [39, Chap. V]). Notably, we estimate convergence rates for defective matrices without introducing any special choice of basis and without requiring knowledge of the Jordan form or any related similarity transformation.

Finally, we note that other measures of convergence may be more appealing in certain situations. Alternatives include Ritz values [20, 24], although convergence behavior can be obscure for matrices that are defective (or nearly so). The subspace residual is computationally attractive because it doesn't require a priori knowledge of the good invariant subspace. This measure can be related to gap convergence [17, 38].

1.1. Algorithmic context. Suppose $\mathbf{V}$ is an $n \times n$ unitary matrix that reduces $\mathbf{A}$ to upper Hessenberg form; i.e., $\mathbf{V}^*\mathbf{A}\mathbf{V} = \mathbf{H}$ for some upper Hessenberg matrix, $\mathbf{H}$. For any index $1 \le \ell \le n$, let $\mathbf{H}_\ell$ denote the ℓth principal submatrix of $\mathbf{H}$:

$$\mathbf{H}_\ell = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1\ell} \\ \beta_2 & h_{22} & \cdots & h_{2\ell} \\ & \ddots & \ddots & \vdots \\ & & \beta_\ell & h_{\ell\ell} \end{bmatrix} \in \mathbb{C}^{\ell\times\ell}.$$

The Arnoldi method [2, 32] builds up the matrices $\mathbf{H}$ and $\mathbf{V}$ one column at a time starting with the unit vector $\mathbf{v}_1 \in \mathbb{C}^n$, although the process is typically stopped well before completion, with $\ell \ll n$. The algorithm only accesses $\mathbf{A}$ through matrix-vector products, making this approach attractive when $\mathbf{A}$ is large and sparse.

Different choices for $\mathbf{v}_1$ produce distinct outcomes for $\mathbf{H}_\ell$. The defining recurrence may be derived from the fundamental relation

$$\mathbf{A}\mathbf{V}_\ell = \mathbf{V}_\ell\mathbf{H}_\ell + \beta_{\ell+1}\mathbf{v}_{\ell+1}\mathbf{e}_\ell^*,$$

where $\mathbf{e}_\ell$ is the ℓth column of the $\ell \times \ell$ identity matrix. The ℓth column of $\mathbf{H}_\ell$ is determined so as to force $\mathbf{v}_{\ell+1}$ to be orthogonal to the columns of $\mathbf{V}_\ell$, and $\beta_{\ell+1}$ then is determined so that $\|\mathbf{v}_{\ell+1}\| = 1$. Provided $\mathbf{H}_\ell$ is unreduced, the columns of $\mathbf{V}_\ell$ constitute an orthonormal basis for the order-ℓ Krylov subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1) = \operatorname{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}^{\ell-1}\mathbf{v}_1\}$. Since $\mathbf{V}_\ell^*\mathbf{A}\mathbf{V}_\ell = \mathbf{H}_\ell$, the matrix $\mathbf{H}_\ell$ is a Ritz–Galerkin approximation of $\mathbf{A}$ on this subspace, as described by Saad [33]. The eigenvalues of $\mathbf{H}_\ell$ are called *Ritz values* and will, in many circumstances, be reasonable approximations to some of the eigenvalues of $\mathbf{A}$. An eigenvector of $\mathbf{H}_\ell$ associated with a given Ritz value θ_j can be used to construct an eigenvector approximation for $\mathbf{A}$. Indeed, if $\mathbf{H}_\ell\mathbf{y}_j = \theta_j\mathbf{y}_j$, then the *Ritz vector* $\widehat{\mathbf{u}}_j = \mathbf{V}_\ell\mathbf{y}_j$ yields the residual

$$\|\mathbf{A}\widehat{\mathbf{u}}_j - \theta_j\widehat{\mathbf{u}}_j\| = |\beta_{\ell+1}|\,|\mathbf{e}_\ell^*\mathbf{y}_j|.$$

When $|\beta_{\ell+1}| \ll 1$, the columns of $\mathbf{V}_\ell$ nearly span an invariant subspace of $\mathbf{A}$. Small residuals more often arise from negligible trailing entries of the vector $\mathbf{y}_j$, indicating the most recent Krylov direction contributed negligibly to the Ritz vector $\widehat{\mathbf{u}}_j$.

Biorthogonal Lanczos methods have similar characteristics despite important differences both in conception and implementation; see, e.g., [4]. In particular, different bases for $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ are generated, and the associated Ritz values can differ considerably from those produced by the Arnoldi algorithm, even though the projection subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ remains the same.

Our focus here avoids the complications of Ritz value convergence and remains fixed on how well a good invariant subspace $\mathcal{U}$ is captured by $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$, without regard to how a basis for $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)$ has been generated.

1.2. Polynomial restarts. The first p steps of the Arnoldi or biorthogonal Lanczos recurrence require p matrix-vector products of the form $\mathbf{A}\mathbf{v}_k$, plus $\mathcal{O}(np^2)$ floating point operations for (bi)orthogonalization. For very large n and very sparse $\mathbf{A}$ (say, with a maximum number of nonzero entries per row very much smaller than n), the cost of orthogonalization will rapidly dominate as p grows. *Polynomial restarting* is one general approach to alleviate this prohibitive expense. At the end of $p+1$ steps of the recurrence, one selects some "best" vector $\mathbf{v}_1^+ \in \mathcal{K}_{p+1}(\mathbf{A}, \mathbf{v}_1)$ and restarts the recurrence from the beginning using $\mathbf{v}_1^+$. Different restart strategies differ essentially in how they attempt to condense progress made in the last $p+1$ steps into the vector $\mathbf{v}_1^+$. Since any vector in $\mathcal{K}_{p+1}(\mathbf{A}, \mathbf{v}_1)$ can be represented as $\psi_p(\mathbf{A})\mathbf{v}_1$ for some polynomial ψ_p of degree p or less, a restart of this type can be expressed as

$$\mathbf{v}_1^+ \leftarrow \psi_p(\mathbf{A})\mathbf{v}_1. \tag{1.2}$$

If subsequent restarts occur (relabeling $\mathbf{v}_1^+$ as $\mathbf{v}_1^{(1)}$), then

$$\begin{aligned}
\mathbf{v}_1^{(1)} &\leftarrow \psi_p^{[1]}(\mathbf{A})\mathbf{v}_1 && \text{(first restart)},\\
\mathbf{v}_1^{(2)} &\leftarrow \psi_p^{[2]}(\mathbf{A})\mathbf{v}_1^{(1)} && \text{(second restart)},\\
&\vdots\\
\mathbf{v}_1^{(\nu)} &\leftarrow \psi_p^{[\nu]}(\mathbf{A})\mathbf{v}_1^{(\nu-1)} && (\nu\text{th restart}).
\end{aligned}$$

We collect the effect of the restarts into a single aggregate polynomial of degree νp:

$$\mathbf{v}_1^{(\nu)} \leftarrow \Psi_{\nu p}(\mathbf{A})\mathbf{v}_1, \tag{1.3}$$

where $\Psi_{\nu p}(\lambda) = \prod_{k=1}^{\nu} \psi_p^{[k]}(\lambda)$ is called the *filter polynomial*.

Evidently, the restart vectors should retain and amplify components of the good invariant subspace while damping and eventually purging components of the bad invariant subspace. One obvious way of encouraging such a trend is to choose the polynomial $\Psi_{\nu p}(\lambda)$ to be as large as possible when evaluated on the good eigenvalues while being as small as possible on the bad eigenvalues. If the bad eigenvalues are situated within a known compact set Ω_{bad} (not containing any good eigenvalues), Chebyshev polynomials associated with Ω_{bad} are often a reasonable choice. When integrated with the Arnoldi algorithm, this results in the Arnoldi–Chebyshev method [34] (cf. [18]).

This Chebyshev strategy requires either a priori or adaptively generated knowledge of Ω_{bad}, a drawback. Sorensen identified an alternative approach, called *exact shifts*, that has proved extremely successful in practice. The filter polynomial $\Psi_{\nu p}$ is automatically constructed using Ritz eigenvalue estimates. Before each new restart of the Arnoldi method, one computes the eigenvalues of $\mathbf{H}_\ell$ and sorts the resulting $\ell = k + p$ Ritz values into two disjoint sets S_{good} and S_{bad}. The p Ritz values

in the set S_{bad} are used to define the restart polynomial $\psi_p(\lambda) = \prod_{j=k+1}^{k+p}(\lambda - \theta_j)$. Morgan discovered a remarkable consequence of this restart strategy: The updated Krylov subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^+)$, generated by the new starting vector $\mathbf{v}_1^+$ in (1.2) using exact shifts, satisfies $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^+) = \text{span}\{\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2, \dots, \widehat{\mathbf{u}}_k, \mathbf{A}\widehat{\mathbf{u}}_j, \mathbf{A}^2\widehat{\mathbf{u}}_j, \dots, \mathbf{A}^p\widehat{\mathbf{u}}_j\}$ for each index $j = 1, 2, \dots, k$ [27]. Thus, Sorensen's exact shifts will provide, in the stage following a restart, a subspace containing every possible Krylov subspace of dimension p that could be obtained with a starting vector that was a linear combination of the good Ritz vectors (cf. [32]). Furthermore, Sorensen showed how to apply shifts implicitly, regenerating the Krylov subspace $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^+)$ with only p matrix-vector products in a numerically stable way. Analogous features can be verified for the restarted biorthogonal Lanczos method using exact shifts to build polynomial filters. Such a strategy has been explored in [16, 9].

Assume now that an Arnoldi or biorthogonal Lanczos process has proceeded ℓ steps past the last of ν restarts, each of which (for the sake of simplicity) has the same order p. In the jth restart ($1 \le j \le \nu$), we use shifts $\{\mu_{jk}\}_{k=1}^p$. Define

$$\Psi_{\nu p}(\lambda) = \prod_{j=1}^{\nu} \prod_{k=1}^{p} (\lambda - \mu_{jk})$$

to be the aggregate restart polynomial after ν restarts. An iteration without restarts will have $p = \nu = 0$ and $\Psi_{\nu p}(\lambda) = 1$.

Let $\mathcal{K}_\tau(\mathbf{A}, \mathbf{v}_1^{(\nu)})$ denote the Krylov subspace of order τ generated by the starting vector $\mathbf{v}_1^{(\nu)}$ that is obtained after ν restarts. The following basic result follows immediately from the observation that $\mathbf{v}_1^{(\nu)} = \Psi_{\nu p}(\mathbf{A})\mathbf{v}_1$.

LEMMA 1.1. *For all* $\tau \ge 0$, $\mathcal{K}_\tau(\mathbf{A}, \mathbf{v}_1^{(\nu)}) = \Psi_{\nu p}(\mathbf{A})\, \mathcal{K}_\tau(\mathbf{A}, \mathbf{v}_1)$.

2. Reachable invariant subspaces. If the good eigenvalues are all simple, then the associated invariant subspace is uniquely determined as the span of good eigenvectors. However, if some of these eigenvalues are multiple, there could be a variety of associated invariant subspaces. Nonetheless, single-vector Krylov eigenvalue algorithms with polynomial restarts are capable of revealing only *one* of the many possible invariant subspaces for any given initial vector. Before developing convergence bounds, we first characterize this distinguished invariant subspace precisely.

Let $\mathcal{M}$ be the cyclic subspace generated by the initial starting vector $\mathbf{v}_1$,

$$\mathcal{M} = \text{span}\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \mathbf{A}^2\mathbf{v}_1, \dots\}.$$

$\mathcal{M}$ is evidently an invariant subspace of $\mathbf{A}$ and $s \equiv \dim(\mathcal{M}) \le n$. Since any invariant subspace of $\mathbf{A}$ that contains $\mathbf{v}_1$ must also contain $\mathbf{A}^\tau\mathbf{v}_1$, $\mathcal{M}$ is the *smallest* invariant subspace of $\mathbf{A}$ that contains $\mathbf{v}_1$. The s vectors of the Krylov sequence $\{\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \dots, \mathbf{A}^{s-1}\mathbf{v}_1\}$ are linearly independent, and thus constitute a basis for $\mathcal{M}$.

Recall that a linear transformation is *nonderogatory* if each eigenvalue has geometric multiplicity equal to 1; i.e., each distinct eigenvalue has precisely one eigenvector associated with it, determined up to scaling.

Define $\mathbf{A}|_{\mathcal{M}}$ to be the restriction of $\mathbf{A}$ to $\mathcal{M}$. The following result is well known; see, e.g., [1], [13, Chap. VII].

LEMMA 2.1. $\mathbf{A}|_{\mathcal{M}}$ *is nonderogatory, and* $\mathcal{K}_\tau(\mathbf{A}, \mathbf{v}_1^{(\nu)}) = \mathcal{K}_\tau(\mathbf{A}|_{\mathcal{M}}, \mathbf{v}_1^{(\nu)}) \subset \mathcal{M}$.

Define α_j to be the *ascent* (or *index*) of the eigenvalue λ_j, i.e., the minimum positive integer α such that $\text{Ker}\,(\mathbf{A}-\lambda_j)^\alpha = \text{Ker}\,(\mathbf{A}-\lambda_j)^{\alpha+1}$. This α_j is the maximum dimension of the n_j different Jordan blocks associated with λ_j, and $\text{Ker}\,(\mathbf{A} - \lambda_j)^{\alpha_j}$ then is the span of all generalized eigenvectors associated with λ_j.

The spectral projection onto each subspace $\mathrm{Ker}\,(\mathbf{A}-\lambda_j)^{\alpha_j}$ can be constructed in the following coordinate-free manner; see, e.g., [23, sect. I.5.3]. For each eigenvalue λ_j, $1 \le j \le N$, let Γ_j be some positively oriented Jordan curve in $\mathbb{C}$ containing λ_j in its interior and all other eigenvalues in its exterior. The spectral projection is defined as

$$\mathbf{P}_j \equiv \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_j} (z-\mathbf{A})^{-1}\, dz.$$

$\mathbf{P}_j$ is a projection onto the span of all generalized eigenvectors associated with λ_j. In particular, $\mathbf{P}_j\mathbf{v}_1$ will be a generalized eigenvector associated with λ_j and will generate a cyclic subspace $\mathcal{K}_{\alpha_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1) \subseteq \mathrm{Ker}\,(\mathbf{A}-\lambda_j)^{\alpha_j}$. Let $\widehat{\alpha}_j$ be the minimum index $\widehat{\alpha}$ so that $\mathcal{K}_{\widehat{\alpha}}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1) = \mathcal{K}_{\widehat{\alpha}+1}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$. This $\widehat{\alpha}_j$ is called the *ascent with respect to* $\mathbf{v}_1$ of the eigenvalue λ_j. Notice that $1 \le \widehat{\alpha}_j \le \alpha_j$ and $\mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$ is the *smallest* invariant subspace of $\mathbf{A}$ that contains $\mathbf{P}_j\mathbf{v}_1$. Furthermore, $\mathbf{P}_j\mathbf{v}_1$ is a generalized eigenvector of grade $\widehat{\alpha}_j$ associated with λ_j and $\widehat{\alpha}_j < \alpha_j$ only if $\mathbf{v}_1$ is deficient in all generalized eigenvectors of maximal grade α_j associated with λ_j.

Define spectral projections $\mathbf{P}_{\mathrm{good}}$ and $\mathbf{P}_{\mathrm{bad}}$ having ranges that are the maximal invariant subspaces associated with the good and bad eigenvalues, respectively, as

$$\mathbf{P}_{\mathrm{good}} = \sum_{j=1}^{L} \mathbf{P}_j \quad \text{and} \quad \mathbf{P}_{\mathrm{bad}} = \sum_{j=L+1}^{N} \mathbf{P}_j.$$

Note that $\mathbf{P}_{\mathrm{good}} + \mathbf{P}_{\mathrm{bad}} = \mathbf{I}$.

The following result in Lemma 2.2 characterizes $\mathcal{M}$. The first statement, included for comparison, is well known; the second is also understood, though we are unaware of its explicit appearance in the literature. Related issues are discussed in [1], [13, Chap. VII].

LEMMA 2.2. $\mathbb{C}^n = \oplus_{j=1}^{N} \mathrm{Ker}(\mathbf{A}-\lambda_j)^{\alpha_j}$ *with* $\sum_{j=1}^{N} \alpha_j \le n$, *and*
$\mathcal{M} = \oplus_{j=1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$ *with* $\sum_{j=1}^{N} \widehat{\alpha}_j = \dim \mathcal{M}$.

Proof. Since $\sum_{j=1}^{N} \mathbf{P}_j = \mathbf{I}$, any $\mathbf{x} \in \mathbb{C}^n$ can be written as $\mathbf{x} = \mathbf{I}\mathbf{x} = \sum_{j=1}^{N} \mathbf{P}_j\mathbf{x}$, which shows that $\mathbb{C}^n \subseteq \oplus_{j=1}^{N} \mathrm{Ker}(\mathbf{A}-\lambda_j)^{\alpha_j}$. The reverse inclusion is trivial.

For the second statement, use $\sum_{j=1}^{N} \mathbf{P}_j = \mathbf{I}$ to get, for any integer $\tau > 0$,

$$\mathbf{v}_1 = \sum_{j=1}^{N} \mathbf{P}_j\mathbf{v}_1, \quad \mathbf{A}\mathbf{v}_1 = \sum_{j=1}^{N} \mathbf{A}\mathbf{P}_j\mathbf{v}_1, \quad \ldots, \quad \mathbf{A}^{\tau}\mathbf{v}_1 = \sum_{j=1}^{N} \mathbf{A}^{\tau}\mathbf{P}_j\mathbf{v}_1.$$

Thus, for each integer $\tau > 0$, $\mathcal{K}_{\tau}(\mathbf{A}, \mathbf{v}_1) \subseteq \oplus_{j=1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$, and, in particular, for τ sufficiently large this yields $\mathcal{M} \subseteq \oplus_{j=1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$.

To show the reverse inclusion, note that for every $j = 1, \ldots, N$, there is a polynomial p_j such that $p_j(\mathbf{A}) = \mathbf{P}_j$. (This polynomial interpolates at eigenvalues: $p_j(\lambda_j) = 1$, p_j has $\alpha_j - 1$ zero derivatives at λ_j, and $p_j(\lambda_k) = 0$ for $\lambda_k \ne \lambda_j$; see, e.g., [19, sect. 6.1].) Thus for any $\mathbf{x} \in \oplus_{j=1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$, one can write

$$\mathbf{x} = \sum_{j=1}^{N} g_j(\mathbf{A})\mathbf{P}_j\mathbf{v}_1 = \sum_{j=1}^{N} g_j(\mathbf{A})p_j(\mathbf{A})\mathbf{v}_1 \in \mathcal{M}$$

for polynomials g_j with degree not exceeding $\widehat{\alpha}_j - 1$. Thus $\oplus_{j=1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1) \subseteq \mathcal{M}$, and so $\mathcal{M} = \oplus_{j=1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j\mathbf{v}_1)$. ∎

Let $\mathcal{X}_{\text{good}}$ and $\mathcal{X}_{\text{bad}}$ be the invariant subspaces of $\mathbf{A}$ associated with the good and bad eigenvalues, respectively. Then define $\mathcal{U}_{\text{good}} \equiv \mathcal{M} \cap \mathcal{X}_{\text{good}}$ and $\mathcal{U}_{\text{bad}} \equiv \mathcal{M} \cap \mathcal{X}_{\text{bad}}$. The following lemma develops a representation for $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$; it shows that $\mathcal{U}_{\text{good}}$ is the *maximum reachable invariant subspace* associated with the good eigenvalues that can be obtained from a Krylov subspace algorithm started with $\mathbf{v}_1$. "Maximum reachable invariant subspace" means that any invariant subspace $\mathcal{U}$ associated with the good eigenvalues and strictly larger than $\mathcal{U}_{\text{good}}$ is *unreachable*: The angle between $\mathcal{U}$ and any computable subspace generated from $\mathbf{v}_1$ is bounded away from zero independent of ℓ, p, ν, and choice of filter shifts $\{\mu_{jk}\}$.

LEMMA 2.3.

$$\mathcal{U}_{\text{good}} = \oplus_{j=1}^{L} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j \mathbf{v}_1), \qquad \mathcal{U}_{\text{bad}} = \oplus_{j=L+1}^{N} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j \mathbf{v}_1),$$

$$\dim \mathcal{U}_{\text{good}} = \sum_{j=1}^{L} \widehat{\alpha}_j \equiv m, \qquad \textit{and} \qquad \dim \mathcal{U}_{\text{bad}} = \sum_{j=L+1}^{N} \widehat{\alpha}_j = s - m.$$

Furthermore, for any subspace $\mathcal{U}$ of $\mathcal{X}_{\text{good}}$ that properly contains $\mathcal{U}_{\text{good}}$, i.e., $\mathcal{U}_{\text{good}} \subset \mathcal{U} \subseteq \mathcal{X}_{\text{good}}$, convergence in gap cannot occur. For all integers $\ell \geq 1$,

$$\delta(\mathcal{U}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \geq \frac{1}{\|\mathbf{P}_{\text{good}}\|} > 0.$$

Proof. Since $\mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j \mathbf{v}_1) \subseteq \operatorname{Ker}(\mathbf{A} - \lambda_j)^{\alpha_j}$, Lemma 2.2 leads to $\mathcal{M} \cap \mathcal{X}_{\text{good}} = \oplus_{j=1}^{L} \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j \mathbf{v}_1)$. Furthermore, $\dim \mathcal{K}_{\widehat{\alpha}_j}(\mathbf{A}, \mathbf{P}_j \mathbf{v}_1) = \widehat{\alpha}_j$ implies that $\dim \mathcal{U}_{\text{good}} = m$ as defined above. The analogous results for $\mathcal{U}_{\text{bad}}$ follow similarly.

Note that $\mathcal{X}_{\text{bad}} = \oplus_{j=L+1}^{N} \operatorname{Ker}(\mathbf{A} - \lambda_j)^{\alpha_j}$ so, for all $\ell \geq 0$,

$$\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)}) \subseteq \mathcal{M} \subseteq \mathcal{U}_{\text{good}} \oplus \mathcal{X}_{\text{bad}}.$$

Thus any $\mathbf{v} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})$ can be decomposed as $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ for some $\mathbf{w}_1 \in \mathcal{U}_{\text{good}}$ and $\mathbf{w}_2 \in \mathcal{X}_{\text{bad}}$. When $\mathcal{U}_{\text{good}}$ is a proper subspace of $\mathcal{U}$, there exists an $\widehat{\mathbf{x}} \in \mathcal{U}$ so that $\widehat{\mathbf{x}} \perp \mathcal{U}_{\text{good}}$ and $\|\widehat{\mathbf{x}}\| = 1$. Note that $\|\widehat{\mathbf{x}} - \mathbf{w}_1\| \geq \|\widehat{\mathbf{x}}\| = 1$. Now,

$$\begin{aligned}
\min_{\mathbf{v} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})} \|\mathbf{v} - \widehat{\mathbf{x}}\| &\geq \min_{\substack{\mathbf{w}_1 \in \mathcal{U}_{\text{good}} \\ \mathbf{w}_2 \in \mathcal{X}_{\text{bad}}}} \|\mathbf{w}_1 + \mathbf{w}_2 - \widehat{\mathbf{x}}\| \\
&\geq \min_{\substack{\mathbf{w}_1 \in \mathcal{U}_{\text{good}} \\ \mathbf{w}_2 \in \mathcal{X}_{\text{bad}}}} \frac{\|\mathbf{w}_2 - (\widehat{\mathbf{x}} - \mathbf{w}_1)\|}{\|\widehat{\mathbf{x}} - \mathbf{w}_1\|} \geq \min_{\substack{\mathbf{y} \in \mathcal{X}_{\text{good}} \\ \mathbf{w}_2 \in \mathcal{X}_{\text{bad}}}} \frac{\|\mathbf{w}_2 - \mathbf{y}\|}{\|\mathbf{y}\|} \\
&\geq \left(\max_{\substack{\mathbf{y} \in \mathcal{X}_{\text{good}} \\ \mathbf{w}_2 \in \mathcal{X}_{\text{bad}}}} \frac{\|\mathbf{P}_{\text{good}}(\mathbf{w}_2 - \mathbf{y})\|}{\|\mathbf{w}_2 - \mathbf{y}\|} \right)^{-1} = \frac{1}{\|\mathbf{P}_{\text{good}}\|}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\delta(\mathcal{U}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) &= \max_{\mathbf{x} \in \mathcal{U}} \min_{\mathbf{v} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})} \frac{\|\mathbf{v} - \mathbf{x}\|}{\|\mathbf{x}\|} \\
&\geq \min_{\mathbf{v} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})} \|\mathbf{v} - \widehat{\mathbf{x}}\| \geq \frac{1}{\|\mathbf{P}_{\text{good}}\|}. \qquad \square
\end{aligned}$$

This means that we have no hope of capturing any invariant subspace that contains a (generalized) eigenspace associated with multiple Jordan blocks—at least when using

a single vector iteration in exact arithmetic. On the other hand, convergence can occur to the good invariant subspace $\mathcal{U}_{\text{good}}$, with a rate that depends on properties of $\mathbf{A}$, $\mathbf{v}_1$, and the choice of filter shifts $\{\mu_{jk}\}$, as we shall see.

Almost every vector in an invariant subspace is a generalized eigenvector of maximal grade and so almost every starting vector will capture maximally defective Jordan blocks. While easily acknowledged, this fact can have perplexing consequences for the casual Arnoldi or biorthogonal Lanczos user, since eigenvectors of other Jordan blocks may be unexpectedly "washed out." Suppose $\mathbf{A}$ is defined as

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

$\mathbf{A}$ is in Jordan canonical form with the single eigenvalue $\lambda = 1$. Let $\mathbf{e}_j$ denote the jth column of the 5×5 identity matrix. Then $\mathbf{e}_2$ and $\mathbf{e}_5$ are eigenvectors of $\mathbf{A}$, $\mathbf{e}_1$ and $\mathbf{e}_4$ are generalized eigenvectors of grade 1 associated with the 2×2 and 3×3 Jordan blocks, and $\mathbf{e}_5$ is a generalized eigenvector of grade 2 associated with the 3×3 block.

For arbitrary $\beta \in \mathbb{C}$, the vector $\mathbf{v}_1 = [1\ \beta\ 1\ 1\ 1\ ]^{\mathrm{T}}$ generates a cyclic subspace spanned by the first three vectors in the Krylov sequence: $\mathbf{v}_1$, $\mathbf{A}\mathbf{v}_1$, and $\mathbf{A}^2\mathbf{v}_1$. By choosing $|\beta|$ to be large, we can give the starting vector $\mathbf{v}_1$ an arbitrarily large component in the direction of $\mathbf{e}_2$, the eigenvector associated with the 2×2 Jordan block.

Defining $\mathbf{M} = \left[\mathbf{v}_1,\, \mathbf{A}\mathbf{v}_1,\, \mathbf{A}^2\mathbf{v}_1\right]$ and $\widehat{\mathbf{H}} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & -3 \\ 0 & 1 & 3 \end{bmatrix}$, a simple calculation reveals $\mathbf{A}\mathbf{M} = \mathbf{M}\widehat{\mathbf{H}}$. The Jordan form of $\widehat{\mathbf{H}}$ is easy to calculate as follows:

$$\mathbf{R}^{-1}\widehat{\mathbf{H}}\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad \text{where } \mathbf{R} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.1}$$

The cyclic subspace generated by the single vector $\mathbf{v}_1$ has captured a *three*-dimensional invariant subspace, associated with the maximally defective 3×3 Jordan block. But this subspace is not the expected span$\{\mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5\}$. Using the change of basis defined by $\mathbf{R}$ in (2.1), one may calculate $\mathbf{A}(\mathbf{M}\mathbf{R}) = (\mathbf{M}\mathbf{R})(\mathbf{R}^{-1}\widehat{\mathbf{H}}\mathbf{R})$, which is

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

Note that $\mathbf{e}_5$ alone is revealed as the eigenvector associated with the eigenvalue 1; $\mathbf{e}_2$ has been washed out in spite of $\mathbf{v}_1$ having an arbitrarily large component in that direction. Indeed the eigenvector $\mathbf{e}_2$ (and so any subspace containing it) is unreachable from *any* starting vector $\mathbf{v}_1$ for which $\mathbf{e}_3^*\mathbf{v}_1 \neq 0$. In this example, $\mathbf{v}_1$ itself emerges as a generalized eigenvector of grade 2. Note that *every* vector $\mathbf{v}$ in $\mathbb{C}^5$ with $\mathbf{e}_3^*\mathbf{v} \neq 0$ is a generalized eigenvector of grade 2 associated with the eigenvalue 1.

We close this section with a computational example that both confirms the gap stagnation lower bound for derogatory matrices given in Lemma 2.3 and illustrates

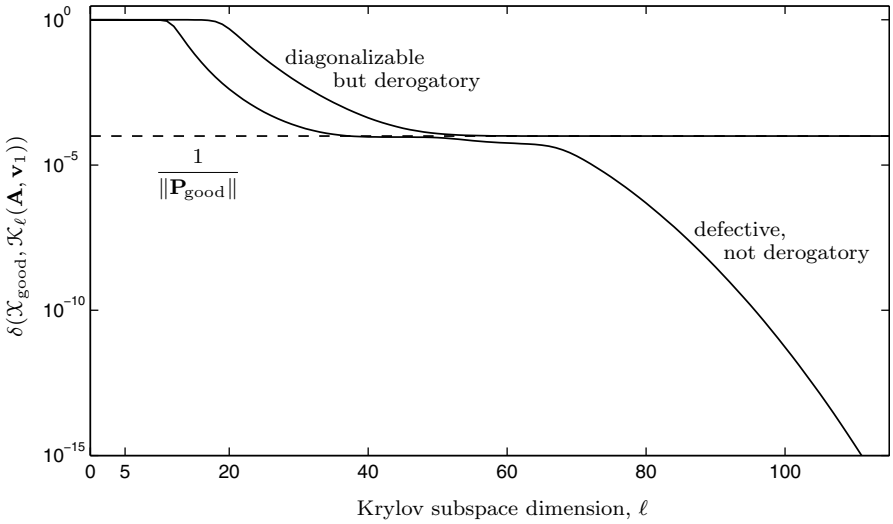


FIG. 2.1. *The Krylov subspace can never capture $\mathcal{X}_{\text{good}}$ when this subspace is associated with a derogatory eigenvalue; convergence is possible, however, when the associated eigenvalues are defective but not derogatory, as described by Lemma* 2.3.

other convergence properties explored in future sections. Consider two matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, each of dimension $n = 150$ with eigenvalues spaced uniformly in the interval $[0, 1]$. In both cases, all the eigenvalues are simple except for the single good eigenvalue $\lambda = 1$, which has algebraic multiplicity 5. In the first case, the geometric multiplicity also equals 5, so the matrix is diagonalizable but derogatory. In the second case, there is only one eigenvector associated with $\lambda = 1$, so it is defective but not derogatory. Both matrices are constructed so that $\|\mathbf{P}_{\text{good}}\| \approx 10^4$. Figure 2.1 illustrates the gap convergence for the Krylov subspace to the invariant subspace $\mathcal{X}_{\text{good}}$ associated with $\lambda = 1$. The starting vector $\mathbf{v}_1$ has $1/\sqrt{n}$ in each component; no restarting is used here. Convergence cannot begin until the fifth iteration, when the Krylov subspace dimension matches the dimension of $\mathcal{X}_{\text{good}}$. This initial period of stagnation is followed by a sublinear phase of convergence leading to a second stagnation period. This is the end of the story for the derogatory case, but for the nonderogatory case, the second stagnation period is transient and the convergence rate eventually settles toward a nearly linear rate. In fact, this rate improves slightly over the final iterations shown here, yielding so-called superlinear convergence, the subject of section 4.3. These convergence phases resemble those observed for the GMRES iteration, as described by Nevanlinna [28].

3. Basic estimates. Since all reachable subspaces are contained in $\mathcal{M}$ and $\mathbf{A}|_{\mathcal{M}}$ is nonderogatory, henceforth we assume without loss of generality that $\mathbf{A}$ itself is nonderogatory so that $n = \dim \mathcal{M}$, and $\mathbf{v}_1$ is not deficient in any generalized eigenvector of maximal grade. To summarize the current situation, $\mathbf{A}$ is an $n \times n$ matrix with $N \leq n$ distinct eigenvalues, $\{\lambda_j\}_{j=1}^N$, each having geometric multiplicity 1 and algebraic multiplicity m_j, so that $\sum_{j=1}^N m_j = n$. We seek L $(1 \leq L < N)$ of these eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_L\}$ (the "good" eigenvalues) together with the corresponding (maximal) invariant subspace $\mathcal{U}_{\text{good}}$ of dimension $m = \sum_{j=1}^L m_j$, which is now the net algebraic multiplicity of good eigenvalues since $\mathbf{A}$ is nonderogatory.

We begin by establishing two lemmas that are used to develop a bound for the gap in terms of a polynomial approximation problem in the subsequent theorems.

LEMMA 3.1. *Given $\mathcal{U}, \mathcal{V} \subseteq \mathbb{C}^n$, suppose $\widehat{\mathbf{u}} \in \mathcal{U}$ ($\|\widehat{\mathbf{u}}\| = 1$) and $\widehat{\mathbf{v}} \in \mathcal{V}$ satisfy*

$$\delta(\mathcal{U}, \mathcal{V}) = \max_{\mathbf{u} \in \mathcal{U}} \min_{\mathbf{v} \in \mathcal{V}} \frac{\|\mathbf{u} - \mathbf{v}\|}{\|\mathbf{u}\|} = \|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\|.$$

Then $\widehat{\mathbf{u}} - \widehat{\mathbf{v}} \perp \mathcal{V}$ and $\widehat{\mathbf{u}} - \widehat{\mathbf{v}} - \delta(\mathcal{U}, \mathcal{V})^2 \widehat{\mathbf{u}} \perp \mathcal{U}$.

Proof. The first assertion is a fundamental property of least squares approximation. To show the second, consider an arbitrary unit vector $\mathbf{u} \in \mathcal{U}$ and take $\varepsilon > 0$. Letting $\mathbf{\Pi}_{\mathcal{V}}$ denote the orthogonal projection onto $\mathcal{V}$, the optimality of $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ implies

$$\|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\|^2 \geq \frac{\|(\mathbf{I} - \mathbf{\Pi}_{\mathcal{V}})(\widehat{\mathbf{u}} + \varepsilon \mathbf{u})\|^2}{\|\widehat{\mathbf{u}} + \varepsilon \mathbf{u}\|^2}.$$

Expanding this inequality, noting $\widehat{\mathbf{v}} = \mathbf{\Pi}_{\mathcal{V}} \widehat{\mathbf{u}}$, and using the first assertion gives

$$\delta(\mathcal{U}, \mathcal{V})^2 (1 + 2\varepsilon \operatorname{Re}(\widehat{\mathbf{u}}^* \mathbf{u}) + \varepsilon^2) \geq \delta(\mathcal{U}, \mathcal{V})^2 + 2\varepsilon \operatorname{Re}((\widehat{\mathbf{u}} - \widehat{\mathbf{v}})^* \mathbf{u}) + \varepsilon^2 \|(\mathbf{I} - \mathbf{\Pi}_{\mathcal{V}})\mathbf{u}\|^2.$$

Collecting terms quadratic in ε on the left-hand side,

$$\varepsilon^2 (\delta(\mathcal{U}, \mathcal{V})^2 - \|(\mathbf{I} - \mathbf{\Pi}_{\mathcal{V}})\mathbf{u}\|^2) \geq 2\varepsilon \operatorname{Re}((\widehat{\mathbf{u}} - \widehat{\mathbf{v}} - \delta(\mathcal{U}, \mathcal{V})^2 \widehat{\mathbf{u}})^* \mathbf{u}).$$

Note that the left-hand side must be nonnegative. Repeating the above argument with $\mathbf{u}$ multiplied by a complex scalar of unit modulus, we can replace the right-hand side with $2\varepsilon\, |(\widehat{\mathbf{u}} - \widehat{\mathbf{v}} - \delta(\mathcal{U}, \mathcal{V})^2 \widehat{\mathbf{u}})^* \mathbf{u}|$. Thus for any unit vector $\widehat{\mathbf{u}} \in \mathcal{U}$,

$$\varepsilon\, (\delta(\mathcal{U}, \mathcal{V})^2 - \|(\mathbf{I} - \mathbf{\Pi}_{\mathcal{V}})\mathbf{u}\|^2) \geq 2\, |(\widehat{\mathbf{u}} - \widehat{\mathbf{v}} - \delta(\mathcal{U}, \mathcal{V})^2 \widehat{\mathbf{u}})^* \mathbf{u}| \geq 0.$$

Taking $\varepsilon \to 0$, we conclude that $\widehat{\mathbf{u}} - \widehat{\mathbf{v}} - \delta(\mathcal{U}, \mathcal{V})^2 \widehat{\mathbf{u}}$ is orthogonal to every $\mathbf{u} \in \mathcal{U}$. ∎

As the gap between subspaces closes ($\delta(\mathcal{U}, \mathcal{V}) \to 0$), $\widehat{\mathbf{u}} - \widehat{\mathbf{v}}$ becomes "almost" orthogonal to $\mathcal{U}$ in the sense that the projection of $\widehat{\mathbf{u}} - \widehat{\mathbf{v}}$ onto $\mathcal{U}$ has norm $\delta(\mathcal{U}, \mathcal{V})^2$.

LEMMA 3.2. *Let $\mathcal{P}_{m-1}$ denote the space of polynomials of degree $m-1$ or less. The mapping $\imath \colon \mathcal{P}_{m-1} \to \mathcal{U}_{\text{good}}$ defined by*

$$\imath(\psi) = \psi(\mathbf{A}) \mathbf{P}_{\text{good}} \mathbf{v}_1 \tag{3.1}$$

is an isomorphism between $\mathcal{P}_{m-1}$ and $\mathcal{U}_{\text{good}}$. Furthermore, there exist positive constants c_1 and c_2 so that

$$c_1 \|\psi\|_{\mathcal{P}_{m-1}} \leq \|\psi(\mathbf{A}) \mathbf{P}_{\text{good}} \mathbf{v}_1\| \leq c_2 \|\psi\|_{\mathcal{P}_{m-1}}, \tag{3.2}$$

uniformly for all $\psi \in \mathcal{P}_{m-1}$ for any fixed norm $\|\cdot\|_{\mathcal{P}_{m-1}}$ defined on the space $\mathcal{P}_{m-1}$.

Proof. $\imath$ is clearly linear. To see that $\imath$ maps $\mathcal{P}_{m-1}$ onto $\mathcal{U}_{\text{good}}$, observe that for any given $\mathbf{y} \in \mathcal{U}_{\text{good}}$, there exist polynomials $\{g_j(\lambda)\}_{j=1}^{L}$ with $\deg(g_j) \leq m_j - 1$ such that

$$\mathbf{y} = \sum_{j=1}^{L} g_j(\mathbf{A}) \mathbf{P}_j \mathbf{v}_1.$$

The L polynomials $\{g_j\}_{j=1}^{L}$ provide L separate "slices" of a single polynomial that can be recovered by (generalized) Hermite interpolation. Let ψ be a polynomial interpolant that interpolates g_j and its derivatives at λ_j:

$$\psi^{(k)}(\lambda_j) = g_j^{(k)}(\lambda_j)$$

for $k = 0, 1, \ldots, m_j - 1$ and $j = 1, 2, \ldots, L$. Theorem VIII.3.16 of [11] leads us first to observe that $\psi(\mathbf{A})\mathbf{P}_j = g_j(\mathbf{A})\mathbf{P}_j$ for each $j = 1, \ldots, L$. Then since $\deg(\psi) \le \sum_{j=1}^{L} m_j - 1 = m - 1$, we have from (3.1) that

$$\mathbf{y} = \sum_{j=1}^{L} \psi(\mathbf{A})\mathbf{P}_j\mathbf{v}_1 = \psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1 = \imath(\psi).$$

Since $\dim(\mathcal{P}_{m-1}) = \dim(\mathcal{U}_{\text{good}})$, $\text{nullity}(\imath) = 0$ and $\imath$ is bijective from $\mathcal{P}_{m-1}$ to $\mathcal{U}_{\text{good}}$. The last statement is an immediate consequence of the fact that linear bijections are bounded linear transformations with bounded inverses. ∎

THEOREM 3.3. *Suppose that* $\mathbf{A}$ *and* $\mathbf{v}_1$ *satisfy the assumptions of this section, and that none of the filter shifts* $\{\mu_{jk}\}$ *coincides with any of the good eigenvalues* $\{\lambda_j\}_{j=1}^{L}$. *For all indices* $\ell \ge m$, *the gap between the good invariant subspace,* $\mathcal{U}_{\text{good}}$, *and the Krylov subspace of order* ℓ, $\mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})$, *generated from the* ν*-fold restarted vector,* $\mathbf{v}_1^{(\nu)}$, *satisfies*

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \le C_0 \max_{\psi \in \mathcal{P}_{m-1}} \min_{\phi \in \mathcal{P}_{\ell-m}} \frac{\|\phi(\mathbf{A})\psi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\phi(\mathbf{A})\psi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|},$$

where $C_0 \equiv 1$ *if* $\mathcal{U}_{\text{good}} \perp \mathcal{U}_{\text{bad}}$ *and* $C_0 \equiv \sqrt{2}$ *otherwise.*

Proof. First, suppose $\mathcal{U}_{\text{good}} \perp \mathcal{U}_{\text{bad}}$. This implies that $\mathbf{P}_{\text{good}}$ and $\mathbf{P}_{\text{bad}}$ are orthogonal projections, $\mathcal{U}_{\text{good}}$ is an invariant subspace for both $\Psi_{\nu p}(\mathbf{A})$ and $[\Psi_{\nu p}(\mathbf{A})]^*$, and, as we will see, that $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) < 1$. Indeed, suppose instead that $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) = 1$. Then there is a vector $\widehat{\mathbf{u}} \in \mathcal{U}_{\text{good}}$ with $\|\widehat{\mathbf{u}}\| = 1$ such that $\widehat{\mathbf{u}} \perp \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})$. Define $\widehat{\mathbf{y}} \equiv [\Psi_{\nu p}(\mathbf{A})]^*\widehat{\mathbf{u}} \in \mathcal{U}_{\text{good}}$, and note that by Lemma 3.2, there exists a polynomial $\widehat{\psi} \in \mathcal{P}_{m-1}$ such that $\widehat{\mathbf{y}} = \widehat{\psi}(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1$. Now, for each $j = 1, 2, \ldots, \ell$, we have

$$\begin{aligned} 0 = \langle \widehat{\mathbf{u}},\ \mathbf{A}^{j-1}\mathbf{v}_1^{(\nu)} \rangle &= \langle \widehat{\mathbf{u}},\ \mathbf{A}^{j-1}\Psi_{\nu p}(\mathbf{A})\mathbf{v}_1 \rangle \\ &= \langle \widehat{\mathbf{y}},\ \mathbf{A}^{j-1}\mathbf{P}_{\text{good}}\mathbf{v}_1 \rangle \\ &= \langle \widehat{\psi}(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1,\ \mathbf{A}^{j-1}\mathbf{P}_{\text{good}}\mathbf{v}_1 \rangle. \end{aligned}$$

Since $\ell \ge m$, this implies first that $\|\widehat{\psi}(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\| = 0$ and then $\widehat{\mathbf{u}} = \mathbf{0}$. (Recall that $[\Psi_{\nu p}(\mathbf{A})]^*$ is bijective on $\mathcal{U}_{\text{good}}$ since $\Psi_{\nu p}$ has no roots in common with good eigenvalues.) But $\widehat{\mathbf{u}}$ was given to be a unit vector, so it must be that $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) < 1$.

There are optimal vectors $\widehat{\mathbf{v}} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})$ and $\widehat{\mathbf{x}} \in \mathcal{U}_{\text{good}}$ with $\|\widehat{\mathbf{x}}\| = 1$ so that

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) = \max_{\mathbf{x} \in \mathcal{U}_{\text{good}}} \min_{\mathbf{v} \in \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})} \frac{\|\mathbf{v} - \mathbf{x}\|}{\|\mathbf{x}\|} = \|\widehat{\mathbf{v}} - \widehat{\mathbf{x}}\|. \tag{3.3}$$

Since $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) < 1$, it must be that $\widehat{\mathbf{v}} \neq \mathbf{0}$. Furthermore, optimality for $\widehat{\mathbf{v}}$ means $\widehat{\mathbf{v}} - \widehat{\mathbf{x}} \perp \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})$ (viz., Lemma 3.1) and, in particular, $\widehat{\mathbf{v}}^*(\widehat{\mathbf{v}} - \widehat{\mathbf{x}}) = 0$. So, $\widehat{\mathbf{v}} \neq \mathbf{0}$ implies $\widehat{\mathbf{v}} \notin \mathcal{U}_{\text{bad}}$. There is a polynomial $\pi_{\ell-1} \in \mathcal{P}_{\ell-1}$ such that

$$\widehat{\mathbf{v}} = \pi_{\ell-1}(\mathbf{A})\mathbf{v}_1^{(\nu)} = \pi_{\ell-1}(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{v}_1.$$

Define $\mathcal{Q} = \mathcal{U}_{\text{good}} \cap \text{Ker}(\pi_{\ell-1}(\mathbf{A}))$ and let $\widehat{q}$ be the minimum (monic) annihilating polynomial for $\mathcal{Q}$.[1] Evidently, $\pi_{\ell-1}$ must contain $\widehat{q}$ as a factor.

[1] That is, $\widehat{q}$ is the minimum degree monic polynomial such that $\widehat{q}(\mathbf{A})\mathbf{r} = \mathbf{0}$ for all $\mathbf{r} \in \mathcal{Q}$.

Since $\widehat{\mathbf{v}} \notin \mathcal{U}_{\text{bad}}$, $\pi_{\ell-1}$ cannot be an annihilating polynomial for $\mathcal{U}_{\text{good}}$, so $\mathcal{Q} \neq \mathcal{U}_{\text{good}}$ and $\deg(\widehat{q}) \leq m-1$. One may factor $\pi_{\ell-1}$ as the product of a polynomial, ϕ, of degree $\ell - m$ and a polynomial, q, of degree $m-1$ containing $\widehat{q}$ as a factor,

$$\pi_{\ell-1}(\lambda) = \phi(\lambda) q(\lambda).$$

Observing that $\mathcal{U}_{\text{good}}$ is invariant for both $\phi(\mathbf{A})$ and $\phi(\mathbf{A})^*$, we may decompose $\widehat{\mathbf{x}}$ as $\widehat{\mathbf{x}} = \phi(\mathbf{A})\widehat{\mathbf{y}} + \mathbf{n}$ for some $\widehat{\mathbf{y}} \in \mathcal{U}_{\text{good}}$ and some $\mathbf{n} \in \operatorname{Ker}(\phi(\mathbf{A})^*) \cap \mathcal{U}_{\text{good}}$. Notice that $\widehat{\mathbf{v}}^*\phi(\mathbf{A})\widehat{\mathbf{y}} = \widehat{\mathbf{v}}^*\widehat{\mathbf{x}} = \widehat{\mathbf{v}}^*\widehat{\mathbf{v}} > 0$, so $\phi(\mathbf{A})\widehat{\mathbf{y}} \neq \mathbf{0}$. However, we'll see that it must happen that $\mathbf{n} = \mathbf{0}$. Indeed, Lemma 3.1 shows that if $\mathbf{z} \in \mathcal{U}_{\text{good}}$ is orthogonal to $\widehat{\mathbf{x}}$, $\widehat{\mathbf{x}}^*\mathbf{z} = 0$, then $\widehat{\mathbf{v}}^*\mathbf{z} = 0$ as well. In particular, for $\mathbf{z} = \|\mathbf{n}\|^2\phi(\mathbf{A})\widehat{\mathbf{y}} - \|\phi(\mathbf{A})\widehat{\mathbf{y}}\|^2\mathbf{n}$ we have $\widehat{\mathbf{x}}^*\mathbf{z} = 0$. Since $\operatorname{Ker}\phi(\mathbf{A})^* = \operatorname{Ran}\phi(\mathbf{A})^\perp$ implies $\widehat{\mathbf{v}}^*\mathbf{n} = 0$, we have

$$0 = \widehat{\mathbf{v}}^*\mathbf{z} = \|\mathbf{n}\|^2\widehat{\mathbf{v}}^*\phi(\mathbf{A})\widehat{\mathbf{y}}.$$

We have already seen that $\widehat{\mathbf{v}}^*\phi(\mathbf{A})\widehat{\mathbf{y}} > 0$, and so $\mathbf{n} = \mathbf{0}$. Thus we can safely exclude from the maximization in (3.3) all $\mathbf{x} \in \mathcal{U}_{\text{good}}$ except for those vectors having the special form $\mathbf{x} = \phi(\mathbf{A})\mathbf{y}$ for $\mathbf{y} \in \mathcal{U}_{\text{good}}$ and ϕ as defined above.

We can now begin our process of bounding the gap. Note that

$$\begin{aligned}
\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) &= \max_{\mathbf{x}\in\mathcal{U}_{\text{good}}} \min_{\mathbf{v}\in\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1^{(\nu)})} \frac{\|\mathbf{v}-\mathbf{x}\|}{\|\mathbf{x}\|} \\
&= \max_{\mathbf{x}\in\mathcal{U}_{\text{good}}} \min_{\phi\in\mathcal{P}_{\ell-m}} \min_{q\in\mathcal{P}_{m-1}} \frac{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})q(\mathbf{A})\mathbf{v}_1 - \mathbf{x}\|}{\|\mathbf{x}\|} \\
&= \max_{\mathbf{y}\in\mathcal{U}_{\text{good}}} \min_{\phi\in\mathcal{P}_{\ell-m}} \min_{q\in\mathcal{P}_{m-1}} \frac{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})[q(\mathbf{A})\mathbf{v}_1 - \mathbf{y}]\|}{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{y}\|}, \qquad (3.4)
\end{aligned}$$

where we are able to justify the substitution $\mathbf{x} = \Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{y}$ since $\Psi_{\nu p}(\mathbf{A})$ is an invertible map of $\mathcal{U}_{\text{good}}$ to itself.

Now by Lemma 3.2, $\mathbf{y} \in \mathcal{U}_{\text{good}}$ can be represented as $\mathbf{y} = \psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1$ for some $\psi \in \mathcal{P}_{m-1}$. Since $\mathbf{I} = \mathbf{P}_{\text{bad}} + \mathbf{P}_{\text{good}}$, one finds

$$\psi(\mathbf{A})\mathbf{v}_1 - \mathbf{y} = \psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1.$$

Continuing with (3.4), assign $q \equiv \psi \in \mathcal{P}_{m-1}$. Then

$$\begin{aligned}
\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) &\leq \max_{\substack{\mathbf{y}\in\mathcal{U}_{\text{good}} \\ (\mathbf{y}=\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1)}} \min_{\phi\in\mathcal{P}_{\ell-m}} \frac{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})[\psi(\mathbf{A})\mathbf{v}_1 - \mathbf{y}]\|}{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{y}\|} \\
&= \max_{\psi\in\mathcal{P}_{m-1}} \min_{\phi\in\mathcal{P}_{\ell-m}} \frac{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|},
\end{aligned}$$

as required, concluding the proof when $\mathcal{U}_{\text{good}} \perp \mathcal{U}_{\text{bad}}$.

In case $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$ are *not* orthogonal subspaces, we introduce a new inner product on $\mathbb{C}^n$ with respect to which they are orthogonal. For any $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$, define

$$\langle \mathbf{u}, \mathbf{v} \rangle_* \equiv \langle \mathbf{P}_{\text{good}}\mathbf{u}, \mathbf{P}_{\text{good}}\mathbf{v} \rangle + \langle \mathbf{P}_{\text{bad}}\mathbf{u}, \mathbf{P}_{\text{bad}}\mathbf{v} \rangle,$$

and define the gap with respect to the new norm $\|\cdot\|_* = \sqrt{\langle \cdot, \cdot \rangle_*}$ to be

$$\delta_*(\mathcal{W}, \mathcal{V}) = \sup_{\mathbf{x}\in\mathcal{W}} \inf_{\mathbf{y}\in\mathcal{V}} \frac{\|\mathbf{y}-\mathbf{x}\|_*}{\|\mathbf{x}\|_*}.$$

Notice that for any vector $\mathbf{w} \in \mathbb{C}^n$,

$$\|\mathbf{w}\|^2 = \|\mathbf{P}_{\text{good}}\mathbf{w} + \mathbf{P}_{\text{bad}}\mathbf{w}\|^2 \le 2\left(\|\mathbf{P}_{\text{good}}\mathbf{w}\|^2 + \|\mathbf{P}_{\text{bad}}\mathbf{w}\|^2\right) = 2\|\mathbf{w}\|_*^2,$$
$$\|\mathbf{P}_{\text{good}}\mathbf{w}\|_* = \|\mathbf{P}_{\text{good}}\mathbf{w}\|, \quad \text{and} \quad \|\mathbf{P}_{\text{bad}}\mathbf{w}\|_* = \|\mathbf{P}_{\text{bad}}\mathbf{w}\|.$$

In particular, for any $\mathbf{x} \in \mathcal{U}_{\text{good}}$ and $\mathbf{y} \in \mathbb{C}^n$ these relationships directly imply

$$\frac{\|\mathbf{y}-\mathbf{x}\|}{\|\mathbf{x}\|} \le \sqrt{2}\,\frac{\|\mathbf{y}-\mathbf{x}\|_*}{\|\mathbf{x}\|_*},$$

and so $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \le \sqrt{2}\ \delta_*(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)}))$. Since $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$ are orthogonal in this new inner product, we can apply the previous argument to conclude[2]

$$\begin{aligned}\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) &\le \sqrt{2}\ \max_{\psi\in\mathcal{P}_{m-1}}\ \min_{\phi\in\mathcal{P}_{\ell-m}}\ \frac{\|\phi(\mathbf{A})\psi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|_*}{\|\phi(\mathbf{A})\psi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|_*} \\ &= \sqrt{2}\ \max_{\psi\in\mathcal{P}_{m-1}}\ \min_{\phi\in\mathcal{P}_{\ell-m}}\ \frac{\|\phi(\mathbf{A})\psi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\phi(\mathbf{A})\psi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|}. \qquad \square\end{aligned}$$

If $\mathbf{N}$ is a square matrix with an invariant subspace $\mathcal{V}$, define

$$\|\mathbf{N}\|_{\mathcal{V}} \equiv \max_{\mathbf{v}\in\mathcal{V}} \frac{\|\mathbf{N}\mathbf{v}\|}{\|\mathbf{v}\|} = \|\mathbf{N}\mathbf{\Pi}_{\mathcal{V}}\|,$$

where $\mathbf{\Pi}_{\mathcal{V}}$ here denotes the orthogonal projection onto $\mathcal{V}$.

THEOREM 3.4. *Suppose* $\mathbf{A}$, $\mathbf{v}_1$, *and the shifts* $\{\mu_{jk}\}$ *satisfy the conditions of Theorem* 3.3. *Then for* $\ell \ge m$,

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \le C_0\, C_1 \min_{\phi\in\mathcal{P}_{\ell-m}} \|[\phi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})]^{-1}\|_{\mathcal{U}_{\text{good}}}\, \|\phi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\|_{\mathcal{U}_{\text{bad}}},$$

where C_0 *is as defined in Theorem* 3.3 *and*

$$C_1 \equiv \max_{\psi\in\mathcal{P}_{m-1}} \frac{\|\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} \tag{3.5}$$

is a constant independent of ℓ, ν, p, *or the filter shifts* $\{\mu_{jk}\}$.

Proof. Let $\mathbf{\Pi}_{\text{good}}$ and $\mathbf{\Pi}_{\text{bad}}$ denote the orthogonal projections onto $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$, respectively. Then

$$\begin{aligned}\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{P}_{\text{bad}}\psi(\mathbf{A})\mathbf{v}_1\| &= \|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{\Pi}_{\text{bad}}\mathbf{P}_{\text{bad}}\psi(\mathbf{A})\mathbf{v}_1\| \\ &\le \|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{\Pi}_{\text{bad}}\|\,\|\mathbf{P}_{\text{bad}}\psi(\mathbf{A})\mathbf{v}_1\|,\end{aligned}$$

and, assuming for the moment that $\phi(\mathbf{A})$ is invertible,

$$\begin{aligned}\|\mathbf{P}_{\text{good}}\psi(\mathbf{A})\mathbf{v}_1\| &= \|[\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})]^{-1}\mathbf{\Pi}_{\text{good}}\mathbf{P}_{\text{good}}\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\psi(\mathbf{A})\mathbf{v}_1\| \\ &\le \|[\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})]^{-1}\mathbf{\Pi}_{\text{good}}\|\,\|\mathbf{P}_{\text{good}}\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\psi(\mathbf{A})\mathbf{v}_1\|.\end{aligned}$$

[2] A more precise value for C_0 can be found as

$$1 \le C_0 = \sqrt{\frac{2\,\|\mathbf{I}-2\mathbf{P}_{\text{good}}\|^2}{1+\|\mathbf{I}-2\mathbf{P}_{\text{good}}\|^2}} \le \sqrt{2};$$

however, the marginal improvement in the final bound would not appear to merit the substantial complexity added.

Hence,

$$\frac{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{P}_{\text{bad}}\psi(\mathbf{A})\mathbf{v}_1\|}{\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\mathbf{P}_{\text{good}}\psi(\mathbf{A})\mathbf{v}_1\|} \leq \|[\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})]^{-1}\|_{\mathcal{U}_{\text{good}}}\|\Psi_{\nu p}(\mathbf{A})\phi(\mathbf{A})\|_{\mathcal{U}_{\text{bad}}}\frac{\|\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|}.$$

Minimizing with respect to ϕ and maximizing with respect to ψ yields the conclusion provided the expression for C_1 is finite. This is assured since, as an immediate consequence of (3.2), $\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\| = 0$ can occur only when $\psi = 0$. □

It is instructive to consider the situation where we seek only a single good eigenvalue, λ_1, which is simple. In this case $m = \dim \mathcal{U}_{\text{good}} = 1$; the conclusion of Theorem 3.3 may be stated as

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \leq C_0\, C_1 \min_{\phi\in\mathcal{P}_{\ell-1}} \frac{\|\phi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\mathbf{w}\|}{|\phi(\lambda_1)\Psi_{\nu p}(\lambda_1)|},$$

where $\mathbf{w} = \mathbf{P}_{\text{bad}}\mathbf{v}_1/\|\mathbf{P}_{\text{bad}}\mathbf{v}_1\|$ and $C_1 = \|\mathbf{P}_{\text{bad}}\mathbf{v}_1\|/\|\mathbf{P}_{\text{good}}\mathbf{v}_1\|$. Elementary geometric considerations yield the following alternate expression for C_1:

$$C_1 = \sqrt{\left(\frac{1}{\|\mathbf{P}_{\text{good}}\|}\frac{\sin\Theta(\mathcal{U}_{\text{good}}, \mathbf{v}_1)}{\cos\Theta(\mathcal{U}_{\text{bad}}^{\perp}, \mathbf{v}_1)}\right)^2 + \left(1 - \frac{1}{\|\mathbf{P}_{\text{good}}\|}\frac{\cos\Theta(\mathcal{U}_{\text{good}}, \mathbf{v}_1)}{\cos\Theta(\mathcal{U}_{\text{bad}}^{\perp}, \mathbf{v}_1)}\right)^2},$$

where $\Theta(\mathcal{U}_{\text{good}}, \mathbf{v}_1)$ and $\Theta(\mathcal{U}_{\text{bad}}^{\perp}, \mathbf{v}_1)$ are the smallest angles between $\mathbf{v}_1$ and the one-dimensional subspaces $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}^{\perp}$, respectively. This special case is stated as Proposition 2.1 of [18];[3] see also Saad's single eigenvalue convergence theory [32].

Our next step is to reduce the conclusion of Theorem 3.4 to an approximation problem in the complex plane. Let $\mathcal{U}$ be an invariant subspace of $\mathbf{A}$ associated with a compact subset $\Omega \subset \mathbb{C}$ (that is, Ω contains only those eigenvalues of $\mathbf{A}$ associated with $\mathcal{U}$ and no others). Define $\kappa(\Omega)$ as the smallest constant for which the inequality

$$\|f(\mathbf{A})\|_{\mathcal{U}} \leq \kappa(\Omega) \max_{z\in\Omega} |f(z)| \tag{3.6}$$

holds uniformly over all $f \in H(\Omega)$, where $H(\Omega)$ denotes the functions analytic on Ω.[4]

Evidently, the value of the constant $\kappa(\Omega)$ depends on the particular choice of Ω (a set containing, in any case, those eigenvalues of $\mathbf{A}$ associated with $\mathcal{U}$). The following properties of $\kappa(\Omega)$ are shared by the generalized Kreiss constant $\widetilde{\mathcal{K}}(\Omega)$ of Toh and Trefethen [41] (defined for $\mathcal{U} = \mathbb{C}^n$). $\kappa(\Omega)$ is monotone decreasing with respect to set inclusion on Ω. Indeed, if $\Omega_1 \subseteq \Omega_2$, then for each function f analytic on Ω_2,

$$\frac{\|f(\mathbf{A})\|_{\mathcal{U}}}{\max\{|f(z)| : z \in \Omega_1\}} \geq \frac{\|f(\mathbf{A})\|_{\mathcal{U}}}{\max\{|f(z)| : z \in \Omega_2\}}.$$

Thus, $\Omega_1 \subseteq \Omega_2$ implies $\kappa(\Omega_1) \geq \kappa(\Omega_2)$.

Since the constant functions are always among the available analytic functions on Ω, $\kappa(\Omega) \geq 1$. If $\mathbf{A}$ is normal, $\kappa(\Omega) = 1$. Indeed, if $\mathbf{A}$ is normal and Σ denotes the set of eigenvalues of $\mathbf{A}$ associated with the invariant subspace $\mathcal{U}$, then

$$1 \leq \kappa(\Omega) = \sup_{f\in H(\Omega)} \frac{\|f(\mathbf{A})\|_{\mathcal{U}}}{\max\{|f(z)| : z \in \Omega\}} = \sup_{f\in H(\Omega)} \frac{\max\{|f(\lambda)| : \lambda \in \Sigma\}}{\max\{|f(z)| : z \in \Omega\}} \leq 1.$$

[3] [18] contains an error amounting to the tacit assumption that $\mathbf{P}_{\text{good}}$ is an orthogonal projection, which is true only if $\mathcal{U}_{\text{good}} \perp \mathcal{U}_{\text{bad}}$. Thus the results coincide only in this special case (note $C_0 = 1$).

[4] For given $k \geq 1$, the sets Ω that (i) contain all eigenvalues of $\mathbf{A}$ and (ii) satisfy $\kappa(\Omega) \leq k$ are called *k-spectral sets* and figure prominently in dilation theory of operators [29].

If any eigenvalue associated with the invariant subspace $\mathcal{U}$ is defective, then some choices of Ω will not yield a finite value for $\kappa(\Omega)$. For example, let $\mathbf{A} = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ and take $\mathcal{U} = \mathbb{C}^2$ as an invariant subspace associated with the defective eigenvalue $\lambda = 0$. If Ω consists of the single point $\{0\}$ and $f(z) = z$, then evidently $\|f(\mathbf{A})\|_{\mathcal{U}} = 1$ but $\max_{z\in\Omega} |f(z)| = 0$. So, no finite value of $\kappa(\Omega)$ is possible (see [31, p. 440]). More generally, if Ω is the spectrum of a defective matrix $\mathbf{A}$, then the monic polynomial consisting of a single linear factor for each distinct eigenvalue of $\mathbf{A}$ is zero on Ω but cannot annihilate $\mathbf{A}$, as it has lower degree than the minimum polynomial of $\mathbf{A}$.

We now use κ to adapt Theorem 3.4 into a more approachable approximation problem. In particular, if Ω_{good} is a compact subset of $\mathbb{C}$ containing all the good eigenvalues of $\mathbf{A}$ but none of the bad, then

$$\begin{aligned}\|[\phi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})]^{-1}\|_{\mathcal{U}_{\text{good}}} &\le \kappa(\Omega_{\text{good}}) \max\{|[\phi(z)\Psi_{\nu p}(z)]^{-1}| : z \in \Omega_{\text{good}}\} \\ &= \frac{\kappa(\Omega_{\text{good}})}{\min\{|\phi(z)\Psi_{\nu p}(z)| : z \in \Omega_{\text{good}}\}}.\end{aligned}$$

Applying a similar bound to $\|\phi(\mathbf{A})\Psi_{\nu p}(\mathbf{A})\|_{\mathcal{U}_{\text{bad}}}$, we obtain the following result, the centerpiece of our development.

THEOREM 3.5. *Suppose* $\mathbf{A}$ *and* $\mathbf{v}_1$ *satisfy the conditions of Theorem* 3.3. *Let* Ω_{good} *and* Ω_{bad} *be disjoint compact subsets of* $\mathbb{C}$ *that contain, respectively, the good and bad eigenvalues of* $\mathbf{A}$*, and suppose that none of the filter shifts* $\{\mu_{jk}\}$ *lies in* Ω_{good}*. Then, for* $\ell \ge m$*,*

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \le C_0\, C_1\, C_2 \min_{\phi\in\mathcal{P}_{\ell-m}} \frac{\max\{|\Psi_{\nu p}(z)\phi(z)| : z \in \Omega_{\text{bad}}\}}{\min\{|\Psi_{\nu p}(z)\phi(z)| : z \in \Omega_{\text{good}}\}},$$

where C_0 *and* C_1 *are the constants introduced in Theorems* 3.3 *and* 3.4*, respectively, and* $C_2 \equiv \kappa(\Omega_{\text{good}})\,\kappa(\Omega_{\text{bad}})$*.*

Evidently, Theorem 3.5 can be implemented with a variety of choices for Ω_{good} and Ω_{bad}, which affects both the polynomial approximation problem and the constant C_2 (considered in section 5.3). The polynomial approximation problem, classified as "Zolotarev-type," is discussed in detail in the next section. Similar problems arise in calculating optimal ADI parameters [26].

4. The polynomial approximation problem. Theorem 3.5 suggests the gap between a Krylov subspace and an invariant subspace will converge to zero at a rate determined by how small polynomials of increasing degree can become on Ω_{bad} while maintaining a minimal uniform magnitude on Ω_{good}. How can this manifest as a linear convergence rate? Consider the ansatz

$$\min_{\phi\in\mathcal{P}_{\ell^*}} \frac{\max\{\,|\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{\,|\phi(z)| : z \in \Omega_{\text{good}}\}} = r^{\ell^*},$$

for some $0 < r \le 1$. Pick a fixed $\phi \in \mathcal{P}_{\ell^*}$, say, with exact degree ℓ^*. Then

$$\log\left(\frac{\max\{\,|\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{\,|\phi(z)| : z \in \Omega_{\text{good}}\}}\right) \ge\ \ell^* \log(r). \tag{4.1}$$

Introducing $U_\phi(z, \Omega_{\text{bad}}) \equiv \frac{1}{\ell^*} \log\left(\frac{|\phi(z)|}{\max\{\,|\phi(w)| : w\in\Omega_{\text{bad}}\}}\right)$, (4.1) is equivalent to

$$\min_{z\in\Omega_{\text{good}}} U_\phi(z, \Omega_{\text{bad}}) \le -\log(r).$$

Evidently, the size of r will be related to how large $U_\phi(z, \Omega_{\text{bad}})$ can be made uniformly throughout Ω_{good}; larger U_ϕ values allow smaller r (faster rates). $U_\phi(z, \Omega_{\text{bad}})$ has the following properties:

- $U_\phi(z, \Omega_{\text{bad}})$ is harmonic at z where $\phi(z) \neq 0$;
- $U_\phi(z, \Omega_{\text{bad}}) = \log|z| + c + o(1)$ for a finite constant c as $|z| \to \infty$;
- $U_\phi(z, \Omega_{\text{bad}}) \leq 0$ for all $z \in \partial\Omega_{\text{bad}}$.

Potential theory provides a natural setting for studying such approximation problems. It is central to the analysis of iterative methods for solving linear systems (see, e.g., [26] for ADI methods and [10, 28] for Krylov subspace methods), and has been used by Calvetti, Reichel, and Sorensen to analyze the Hermitian Lanczos algorithm with restarts [6]. We apply similar techniques here to study $U_\phi(z, \Omega_{\text{bad}})$.

4.1. Potential theory background. Let $\mathcal{D} \subset \mathbb{C}$ be a compact set whose complement, $\mathbb{C} \setminus \mathcal{D}$, is a connected Dirichlet region.[5] The Green's function of $\mathbb{C} \setminus \mathcal{D}$ with pole at infinity is defined as that function, $g[z, \mathcal{D}]$, that satisfies the following properties:

(i) g is harmonic in $\mathbb{C} \setminus \mathcal{D}$;
(ii) $\lim_{z\to\infty} g[z, \mathcal{D}] = \log|z| + \text{finite constant}$;
(iii) $\lim_{z\to\widehat{z}} g[z, \mathcal{D}] = 0$ for all $\widehat{z} \in \partial\mathcal{D}$;
(iv) $g[z, \mathcal{D}] > 0$ for all $z \in \mathbb{C} \setminus \mathcal{D}$.

Note that property (iv) can be deduced from (i), (ii), the fact that (ii) implies that $g > 0$ for all sufficiently large $|z|$, and the maximum principle for harmonic functions. The maximum principle also shows that $g[z, \mathcal{D}]$ is the only function satisfying (i)–(iv).

Example 4.1. If $\mathbb{C} \setminus \mathcal{D}$ is simply connected, one is assured (from the Riemann mapping theorem; see, e.g., [8, sect. VII.4]) of the existence of a function $F(z)$ that maps $\mathbb{C} \setminus \mathcal{D}$ conformally onto the exterior of the closed unit disk $\mathbb{C} \setminus \mathcal{B}_1 = \{z : |z| > 1\}$ such that $F(\infty) = \infty$. Such an F must behave asymptotically as $\alpha z + \mathcal{O}(1)$ as $|z| \to \infty$ for some constant α, since it must remain one-to-one in any neighborhood of ∞. Since $\log|z|$ is harmonic for any $z \neq 0$, one may check that $u(z) = \log|F(z)|$ is also harmonic in z wherever $F(z) \neq 0$, $u(\infty) = \infty$, and $u(z) \to 0$ as $|z| \to 1$ from $\mathbb{C} \setminus \mathcal{D}$. Thus, $\log|F(z)|$ is the Green's function with pole at infinity for $\mathbb{C} \setminus \mathcal{D}$. Evidently, $\lim_{|z|\to\infty} u(z) - \log|z| \to \log|\alpha|$. Notice that $\log|z|$ itself is the Green's function with pole at infinity for $\mathbb{C} \setminus \mathcal{B}_1$. ∎

Even for more complicated compact sets $\mathcal{D}$, the condition that $g[z, \mathcal{D}]$ is harmonic everywhere outside $\mathcal{D}$ with a pole at ∞ restricts the rate of growth of $g[z, \mathcal{D}]$ near ∞. Loosely speaking, as $|z|$ becomes very large, the compact set $\mathcal{D}$ becomes less and less distinguishable from a disk centered at 0 (say, with radius γ), and so $g[z, \mathcal{D}]$ becomes less and less distinguishable from $g[z, \mathcal{B}_\gamma] = \log|z/\gamma| = \log|z| - \log\gamma$, which is the Green's function with pole at infinity for $\mathbb{C} \setminus \mathcal{B}_\gamma = \{z : |z| > \gamma\}$. Indeed, from property (ii), $g[z, \mathcal{D}]$ has growth at infinity satisfying

$$\lim_{|z|\to\infty} g[z, \mathcal{D}] - \log|z| = -\log\gamma \tag{4.2}$$

for some constant $\gamma > 0$ known as the *logarithmic capacity* of the set $\mathcal{D}$. This γ can be thought of as the effective radius of $\mathcal{D}$ in the sense we've just described.

Example 4.2. Suppose $\Phi_\ell(z)$ is a monic polynomial of degree ℓ and let

$$\mathcal{D}_\varepsilon(\Phi_\ell) = \{z \in \mathbb{C} \ : |\Phi_\ell(z)| \leq \varepsilon\}$$

[5] See [8, sect. X.4]. For our purposes here, this can be taken to mean a set having a piecewise smooth boundary with no isolated points; the effect of isolated points is addressed in section 4.3.

be a family of regions whose boundaries are the ε-lemniscates of $\Phi_\ell(z)$. $\mathcal{D}_\varepsilon(\Phi_\ell)$ is compact for each $\varepsilon > 0$, though it need not be a connected region. With an easy calculation one may verify that $\mathcal{D}_\varepsilon(\Phi_\ell)$ has the Green's function (cf. [36, p. 164])

$$g[z, \mathcal{D}_\varepsilon(\Phi_\ell)] = \frac{1}{\ell} \log\left(\frac{|\Phi_\ell(z)|}{\varepsilon}\right). \qquad \square$$

Equipped with the Green's function $g[z, \mathcal{D}]$, we return to the analysis of the function $U_\phi(z, \mathcal{D})$ describing the error in our approximation problem. The following result is a simplified version of the Bernstein–Walsh lemma (see [36, sect. III.2]).

PROPOSITION 4.3. *Let $\mathcal{D}$ be a compact set with piecewise smooth boundary $\partial\mathcal{D}$. Suppose u is harmonic outside $\mathcal{D}$ and that $u(z) \le 0$ for $z \in \partial\mathcal{D}$. If $u(z) = \log|z| + c + o(1)$ for some constant c as $|z| \to \infty$, then $u(z) \le g[z, \mathcal{D}]$. In particular, if $\phi(z)$ is any polynomial of degree ℓ, then for each $z \in \mathbb{C} \setminus \mathcal{D}$*

$$U_\phi(z, \mathcal{D}) = \frac{1}{\ell} \log\left(\frac{|\phi(z)|}{\max\{|\phi(w)| : w \in \mathcal{D}\}}\right) \le g[z, \mathcal{D}]. \tag{4.3}$$

For certain special choices of $\mathcal{D} = \Omega_{\text{bad}}$, the polynomial approximation problem of Theorem 3.5 can be solved exactly.

THEOREM 4.4. *Suppose $\Phi_{\ell^*}(z)$ is a monic polynomial of degree ℓ^*. Let $\Omega_{\text{bad}} = \mathcal{D}_\varepsilon(\Phi_{\ell^*})$ be an associated ε-lemniscatic set as defined in Example* 4.2 *and suppose Ω_{good} is a compact subset of $\mathbb{C}$ such that $\Omega_{\text{good}} \cap \mathcal{D}_\varepsilon(\Phi_{\ell^*}) = \emptyset$. Then*

$$\min_{\phi \in \mathcal{P}_{\ell^*}} \frac{\max\{\,|\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{\,|\phi(z)| : z \in \Omega_{\text{good}}\}} = \frac{\varepsilon}{\min\{\,|\Phi_{\ell^*}(z)| : z \in \Omega_{\text{good}}\}}.$$

Proof. Using the Green's function for $\mathcal{D}_\varepsilon(\Phi_{\ell^*})$ described in Example 4.2, we can rearrange (4.3) to show that for any $\phi \in \mathcal{P}_{\ell^*}$,

$$\frac{|\phi(z)|}{\max\{|\phi(w)| : w \in \mathcal{D}_\varepsilon(\Phi_{\ell^*})\}} \le \frac{|\Phi_{\ell^*}(z)|}{\varepsilon}$$

holds for all $z \in \Omega_{\text{good}}$. Equality is attained for every $z \in \mathbb{C}$ whenever $\phi = \Phi_{\ell^*}$. Minimizing over $z \in \Omega_{\text{good}}$ and then maximizing over $\phi \in \mathcal{P}_{\ell^*}$ yields

$$\max_{\phi \in \mathcal{P}_{\ell^*}} \frac{\min\{|\phi(z)| : z \in \Omega_{\text{good}}\}}{\max\{|\phi(w)| : w \in \mathcal{D}_\varepsilon(\Phi_{\ell^*})\}} \le \frac{\min\{|\Phi_{\ell^*}(z)| : z \in \Omega_{\text{good}}\}}{\varepsilon}. \tag{4.4}$$

In fact, equality must hold in (4.4) since $\phi = \Phi_{\ell^*}$ is included in the class of functions over which the maximization occurs. The conclusion then follows by taking the reciprocal of both sides. $\square$

More general choices for $\mathcal{D} = \Omega_{\text{bad}}$ will not typically yield exactly solvable polynomial approximation problems, at least for fixed (finite) polynomial degree. However, the following asymptotic result holds as the polynomial degree increases.

THEOREM 4.5. *Let Ω_{bad} and Ω_{good} be two disjoint compact sets in the complex plane such that $\mathbb{C} \setminus \Omega_{\text{bad}}$ is a Dirichlet region. Then*

$$\lim_{\ell^* \to \infty} \min_{\phi \in \mathcal{P}_{\ell^*}} \left(\frac{\max\{|\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\phi(z)| : z \in \Omega_{\text{good}}\}}\right)^{1/\ell^*} = e^{-\min\{g[z, \Omega_{\text{bad}}] \,:\, z \in \Omega_{\text{good}}\}}, \tag{4.5}$$

where $g[z, \Omega_{\text{bad}}]$ is the Green's function of $\mathbb{C} \setminus \Omega_{\text{bad}}$ with pole at infinity.

Proof. The theorem is proved in [26, p. 236], where the left-hand side of (4.5) is referred to as the $(\ell^*, 0)$ Zolotarev number. We give here a brief indication of the proof to support later discussion. Inequality (4.3) can be manipulated to yield

$$\left(\frac{|\phi_{\ell^*}(z)|}{\max\{|\phi_{\ell^*}(w)| : w \in \Omega_{\rm bad}\}}\right)^{1/\ell^*} \leq e^{g[z,\Omega_{\rm bad}]},$$

which in turn implies

$$\left(\frac{\max\{|\phi_{\ell^*}(w)| : w \in \Omega_{\rm bad}\}}{\min\{|\phi_{\ell^*}(z)| : z \in \Omega_{\rm good}\}}\right)^{1/\ell^*} \geq e^{-\min\{g[z,\Omega_{\rm bad}] : z\in\Omega_{\rm good}\}}.$$

Furthermore, one may construct polynomials L_k that have as their zeros points distributed on the boundary $\partial\Omega_{\rm bad}$, the *Leja points* $\{\mu_1, \mu_2, \ldots, \mu_k\}$, defined recursively so that

$$\mu_{k+1} = \arg\max\left\{\prod_{j=1}^{k} |z - \mu_j| : z \in \Omega_{\rm bad}\right\};$$

see [36, sect. V.1]. This sequence of *Leja polynomials* satisfies asymptotic optimality,

$$\lim_{k\to\infty}\left(\frac{|L_k(z)|}{\max\{|L_k(w)| : w \in \Omega_{\rm bad}\}}\right)^{1/k} = e^{g[z,\Omega_{\rm bad}]} \tag{4.6}$$

for each $z \in \mathbb{C}\backslash\Omega_{\rm bad}$. Convergence is uniform on compact subsets of $\mathbb{C}\backslash\Omega_{\rm bad}$. Thus we can reverse the order of the limit with respect to polynomial degree and minimization with respect to $z \in \Omega_{\rm good}$, then take reciprocals to find

$$\lim_{k\to\infty}\left(\frac{\max\{|L_k(w)| : w \in \Omega_{\rm bad}\}}{\min\{|L_k(z)| : z \in \Omega_{\rm good}\}}\right)^{1/k} = e^{-\min\{g[z,\Omega_{\rm bad}] : z\in\Omega_{\rm good}\}}. \tag{4.7}$$

Since

$$\begin{aligned}\left(\frac{\max\{|L_{\ell^*}(w)| : w \in \Omega_{\rm bad}\}}{\min\{|L_{\ell^*}(z)| : z \in \Omega_{\rm good}\}}\right)^{1/\ell^*} &\geq \min_{\phi\in\mathcal{P}_{\ell^*}}\left(\frac{\max\{|\phi(w)| : w \in \Omega_{\rm bad}\}}{\min\{|\phi(z)| : z \in \Omega_{\rm good}\}}\right)^{1/\ell^*}\\ &\geq e^{-\min\{g[z,\Omega_{\rm bad}] : z\in\Omega_{\rm good}\}},\end{aligned}$$

equality must hold throughout and thus (4.5) holds. ∎

In the context of Example 4.1, where $F(z)$ was a conformal map taking the exterior of $\Omega_{\rm bad}$ to the exterior of the closed unit disk with $F(\infty) = \infty$, Theorem 4.5 reduces to (cf. [10, Thm. 2])

$$\lim_{\ell^*\to\infty}\min_{\phi\in\mathcal{P}_{\ell^*}}\left(\frac{\max\{|\phi(w)| : w \in \Omega_{\rm bad}\}}{\min\{|\phi(z)| : z \in \Omega_{\rm good}\}}\right)^{1/\ell^*} = \max_{z\in\Omega_{\rm good}}\frac{1}{|F(z)|}.$$

4.2. Effective restart strategies. The usual goal in constructing a restart strategy is to limit the size of the Krylov subspace (restricting the maximum degree of the polynomial ϕ) without degrading the asymptotic convergence rate. Demonstrating equality in (4.5) pivoted on the construction of an optimal family of polynomials—in this case, Leja polynomials. There are other possibilities, however. Fekete polynomials are the usual choice for the construction in Theorem 4.5; see [36, sect. III.1]. Chebyshev polynomials and Faber polynomials offer familiar alternatives. (For Hermi-

tian matrices, a practical Leja shift strategy has been developed by Baglama, Calvetti, and Reichel [3] and Calvetti, Reichel, and Sorenson [6]. Heuveline and Sadkane advocate numerical conformal mapping to determine Faber polynomials for restarting non-Hermitian iterations [18].) Once some optimal family of polynomials is known that solves (4.5), effective restart strategies become evident.

THEOREM 4.6. *Let Ω_{good} and Ω_{bad} be two disjoint compact sets in the complex plane containing, respectively, the good and bad eigenvalues of $\mathbf{A}$, and such that $\mathbb{C} \setminus \Omega_{\text{bad}}$ is a Dirichlet region. Suppose that $\Psi_{\nu p}(z)$ is the aggregate restart polynomial representing ν restarts each of order p.*

(a) *If polynomial restarts are performed using roots of optimal polynomials for Ω_{bad} (i.e., $\Psi_{\nu p}(z)$ are optimal polynomials of degree νp), then*

$$\lim_{\nu\to\infty} \min_{\phi\in\mathcal{P}_{\ell^*}} \left(\frac{\max\{|\Psi_{\nu p}(w)\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\Psi_{\nu p}(z)\phi(z)| : z \in \Omega_{\text{good}}\}} \right)^{\frac{1}{\nu p+\ell^*}} = e^{-\min\{g[z,\Omega_{\text{bad}}]:z\in\Omega_{\text{good}}\}}, \tag{4.8}$$

where $g[z, \Omega_{\text{bad}}]$ is the Green's function of Ω_{bad} with pole at infinity.

(b) *If the boundary of Ω_{bad} is a lemniscate of $\Psi_{\nu p}\Phi_{\ell^*}$,*

$$\Omega_{\text{bad}} = \mathcal{D}_\varepsilon(\Psi_{\nu p}\Phi_{\ell^*}) = \{z \in \mathbb{C} : |\Psi_{\nu p}(z)\Phi_{\ell^*}(z)| \le \varepsilon\},$$

for some degree-ℓ^* monic polynomial Φ_{ℓ^*} and some $\varepsilon > 0$, then

$$\min_{\phi\in\mathcal{P}_{\ell^*}} \frac{\max\{|\Psi_{\nu p}(w)\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\Psi_{\nu p}(z)\phi(z)| : z \in \Omega_{\text{good}}\}} = \frac{\varepsilon}{\min\{|\Psi_{\nu p}(z)\Phi_{\ell^*}(z)| : z \in \Omega_{\text{good}}\}}.$$

Proof. Part (b) follows immediately from Theorem 4.4. Part (a) can be seen by observing that since $\Psi_{\nu p}(z)$ is an asymptotically optimal family for Ω_{bad},

$$\begin{aligned} \frac{\max\{|\Psi_{\nu p}(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\Psi_{\nu p}(z)| : z \in \Omega_{\text{good}}\}} &\ge \min_{\phi\in\mathcal{P}_{\ell^*}} \left(\frac{\max\{|\Psi_{\nu p}(w)\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\Psi_{\nu p}(z)\phi(z)| : z \in \Omega_{\text{good}}\}} \right) \\ &\ge \left(e^{-\min\{g[z,\Omega_{\text{bad}}] \,:\, z\in\Omega_{\text{good}}\}} \right)^{\nu p+\ell^*}. \end{aligned}$$

Now fixing p and ℓ^*, the conclusion follows from (4.7) by following the subsequence generated by $\nu = 1, 2, \ldots$. ∎

Recall that the desired effect of the restart polynomial is to retain the rapid convergence rate of the full (unrestarted) Krylov subspace without requiring the dimension ℓ^* to grow without bound. We have seen here that restarting with optimal polynomials for Ω_{bad} recovers the expected linear convergence rate for Ω_{bad} (presuming one can identify this set, not a trivial matter in practice). Still, the unrestarted process may take advantage of the discrete nature of the spectrum, accelerating convergence beyond the expected linear rate. Designing a restart strategy that yields similar behavior is more elaborate.

4.3. Superlinear effects from assimilation of bad eigenvalues. In a variety of situations, the gap appears to converge superlinearly. True superlinear convergence is an asymptotic phenomenon that has a nontrivial meaning only for nonterminating iterations. Thus one must be cautious about describing superlinear effects relating to (unrestarted) Krylov subspaces, since $\mathcal{U}_{\text{good}}$ is eventually completely captured by the Krylov subspace as discussed in section 2. Here our point of view follows that of [46, 48], showing the estimated gap may be bounded by a family of linearly converging

processes exhibiting increasingly rapid linear rates. The next result mimics the Ritz value bounds for Hermitian matrices developed by van der Sluis and van der Vorst [47, sect. 6.6]. We assume here that Ω_{bad} consists of the union of s discrete points, potentially with some additional Dirichlet region. That is, some bad eigenvalues (typically those closest to the good eigenvalues, or distant outliers) are treated as discrete points, while any leftovers are collected in the Dirichlet region.

THEOREM 4.7. *Let Ω_{good} and Ω_{bad} be disjoint compact subsets of $\mathbb{C}$ and suppose Ω_{bad} contains s isolated points, $z_1,\ z_2,\ \ldots, z_s$. Define a sequence of $s+1$ nested subsets as $\Omega_k = \Omega_{k+1} \cup \{z_k\}$ for $k = 1, \ldots, s$ with $\Omega_1 \equiv \Omega_{\text{bad}}$, so that each set $\Omega_k \supset \Omega_{k+1} \neq \emptyset$ differs from adjacent sets in the sequence by single points. Define also the associated diameters*

$$\mathsf{e}_k \equiv \max\{|w - z_k| : w \in \Omega_k\} \quad \textit{and} \quad \mathsf{d}_k \equiv \min\{|z - z_k| : z \in \Omega_{\text{good}}\}.$$

Then for $r = 1, \ldots, s$ and each $\ell^* > r$,

$$\min_{\phi \in \mathcal{P}_{\ell^*}} \frac{\max\{|\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\phi(z)| : z \in \Omega_{\text{good}}\}} \leq \left(\prod_{j=1}^{r} \frac{\mathsf{e}_j}{\mathsf{d}_j}\right) \min_{\phi \in \mathcal{P}_{\ell^*-r}} \frac{\max\{|\phi(w)| : w \in \Omega_{r+1}\}}{\min\{|\phi(z)| : z \in \Omega_{\text{good}}\}}.$$

Proof. Fix an integer $k \geq 1$ and observe that

$$\begin{aligned}
\min_{\phi \in \mathcal{P}_{\ell^*}} \frac{\max_{w \in \Omega_k} |\phi(w)|}{\min_{z \in \Omega_{\text{good}}} |\phi(z)|} &\leq \min_{\phi \in \mathcal{P}_{\ell^*-1}} \frac{\max_{w \in \Omega_k} |(w - z_k)\phi(w)|}{\min_{z \in \Omega_{\text{good}}} |(z - z_k)\phi(z)|} \\
&= \min_{\phi \in \mathcal{P}_{\ell^*-1}} \frac{\max_{w \in \Omega_{k+1}} |(w - z_k)\phi(w)|}{\min_{z \in \Omega_{\text{good}}} |(z - z_k)\phi(z)|} \\
&\leq \frac{\mathsf{e}_k}{\mathsf{d}_k} \min_{\phi \in \mathcal{P}_{\ell^*-1}} \frac{\max_{w \in \Omega_{k+1}} |\phi(w)|}{\min_{z \in \Omega_{\text{good}}} |\phi(z)|}.
\end{aligned}$$

The conclusion follows by applying the argument repeatedly for $k = 1, 2, \ldots, r$. ∎

Asymptotically, the discrete points in Ω_{bad} have no effect on the convergence rate.

COROLLARY 4.8. *In the notation of Theorem 4.7, suppose Ω_{s+1} is a Dirichlet region. Then*

$$\lim_{\ell^* \to \infty} \min_{\phi \in \mathcal{P}_{\ell^*}} \left(\frac{\max\{|\phi(w)| : w \in \Omega_{\text{bad}}\}}{\min\{|\phi(z)| : z \in \Omega_{\text{good}}\}}\right)^{1/\ell^*} \leq e^{-\min\{g[z, \Omega_{s+1}] \,:\, z \in \Omega_{\text{good}}\}},$$

where $g[z, \Omega_{s+1}]$ is the Green's function with pole at infinity associated with $\mathbb{C} \setminus \Omega_{s+1}$.

Proof. The result follows by applying the asymptotic approach of Theorem 4.5 to the result of Theorem 4.7 for $r = s$. ∎

To demonstrate such superlinear effects, we consider a parameterized diagonal matrix $\mathbf{A}_\alpha$ having 100 bad eigenvalues spaced uniformly in the unit interval $[-1-\alpha, -\alpha]$ and 4 good eigenvalues uniformly spaced in $[0, 1]$. Figure 4.1 illustrates convergence of the gap $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}_\alpha, \mathbf{v}_1))$ for $\alpha = 0.1$, 0.01, 0.05, and 0.001, always with the starting vector $\mathbf{v}_1$ having $1/\sqrt{n}$ in each component ($n = 104$). Above each convergence curve are bounds from Theorem 3.5 and Theorem 4.7. (The calculation of C_1 is addressed in section 5.1.) For the superlinear bounds, take Ω_{bad} to be the set of bad eigenvalues and set Ω_r to be Ω_{bad} less the $r-1$ rightmost bad eigenvalues. We approximate the optimal polynomial in Theorem 4.7 by Chebyshev polynomials for $\Omega_{r+1}^{\text{conv}}$ (see [35, sect. IV.4.1] for details). Notice the envelope produced by the aggregated linear rates creates a superlinear convergence effect to an extent determined

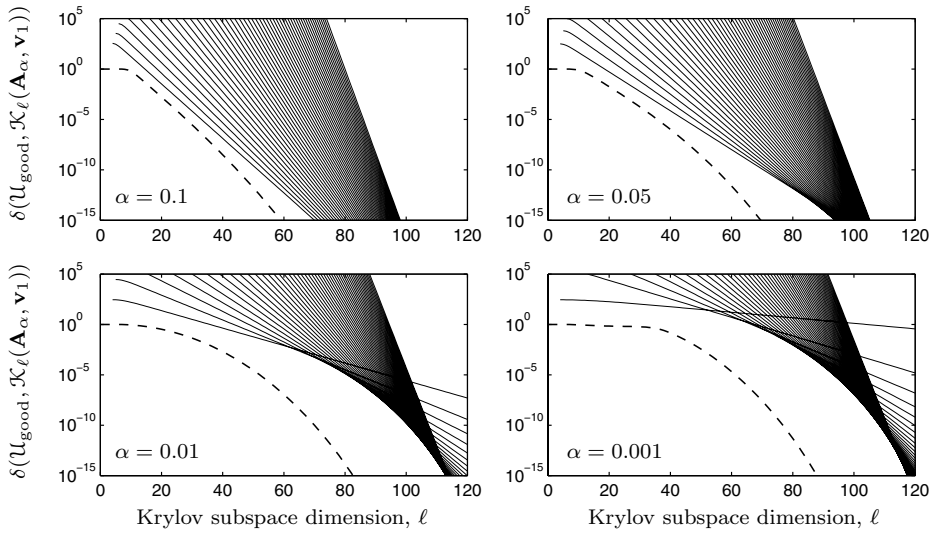


FIG. 4.1. *Aggregate linear rates produce a superlinear effect. Observed gap convergence (broken line) and aggregate bounds (solid lines) computed using Theorems* 3.5 *and* 4.7 *for* Ω_r *with* $r = 1, \ldots, 50$.

by the "granularity" of bad eigenvalues as viewed from the nearest good eigenvalue. Greater granularity (smaller α) causes poor initial rates due to nearby bad eigenvalues, which rapidly dissipate as these eigenvalues are assimilated, yielding to improved rates determined by more remote bad eigenvalues. The same phenomenon is observed in section 6.4 for a Markov chain eigenvalue problem. But assimilation of nearby bad eigenvalues is not the only mechanism for superlinear convergence. In section 5.3, we describe how nonnormality can also give rise to such behavior, illustrated experimentally in section 6.2.

5. Analysis of constants. This section contains a more detailed discussion of the constants C_1 and C_2 that arise in the convergence bounds given in Theorems 3.4 and 3.5. The magnitude of these constants controls the predicted start of the linear phase of convergence: larger constants suggest delayed linear convergence. Thus we seek an appreciation of those matrix and starting vector properties that lead to more or less favorable convergence bounds.

5.1. Bounding C_1. Notice that

$$C_1 = \max_{\psi \in \mathcal{P}_{m-1}} \frac{\|\psi(\mathbf{A})\mathbf{P}_{\rm bad}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\rm good}\mathbf{v}_1\|} = \max_{\mathbf{v} \in \mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)} \frac{\|\mathbf{P}_{\rm bad}\mathbf{v}\|}{\|\mathbf{P}_{\rm good}\mathbf{v}\|} = \max_{\mathbf{x} \in \mathbb{C}^m} \frac{\|\mathbf{P}_{\rm bad}\mathbf{V}_m\mathbf{x}\|}{\|\mathbf{P}_{\rm good}\mathbf{V}_m\mathbf{x}\|},$$

where the columns of $\mathbf{V}_m$ form a basis for $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$. This last expression for C_1 is simply the largest generalized singular value of the pair of matrices $\mathbf{P}_{\rm bad}\mathbf{V}_m$ and $\mathbf{P}_{\rm good}\mathbf{V}_m$ (see, e.g., [14, sect. 8.7.3]). This is how we determine C_1 for our examples.

The dependence of C_1 on the starting vector $\mathbf{v}_1$ is critical. If $\mathbf{v}_1$ is biased against $\mathcal{U}_{\rm good}$, then C_1 will be large and our bounds predict a delay in convergence. Likewise, a good starting vector accelerates convergence as expected.[6] We investigate this

[6]Though our bounds explicitly incorporate restart effects into the polynomial approximation problem, an alternative approach could instead handle restarts via the constant C_1, which we expect to shrink as restarts enrich the starting vector in $\mathcal{U}_{\rm good}$.

behavior with an illustrative example, but first give bounds for C_1 that relate its magnitude to the orientation of $\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1)$ relative to $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$.

PROPOSITION 5.1. *Under the conditions of Theorem* 3.4,

$$\frac{1}{\|\mathbf{P}_{\text{good}}\|} \frac{\delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}})}{\delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{bad}})} \leq C_1 \leq \frac{\|\mathbf{P}_{\text{good}}\| \, \delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}})}{1 - \|\mathbf{P}_{\text{good}}\| \, \delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}})},$$

where the second inequality holds provided $\|\mathbf{P}_{\text{good}}\| \, \delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}}) < 1$.

Proof. If $\mathbf{\Pi}_{\text{good}}$ denotes the orthogonal projection onto $\mathcal{U}_{\text{good}}$, then $\mathbf{I} - \mathbf{\Pi}_{\text{good}} = (\mathbf{I} - \mathbf{\Pi}_{\text{good}})(\mathbf{I} - \mathbf{P}_{\text{good}})$, and so

$$\|(\mathbf{I} - \mathbf{\Pi}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\| \leq \|(\mathbf{I} - \mathbf{P}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\| = \|\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|.$$

Thus,

$$\begin{aligned}
\delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}}) &= \max_{\psi \in \mathcal{P}_{m-1}} \min_{\mathbf{u} \in \mathcal{U}_{\text{good}}} \frac{\|\mathbf{u} - \psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \\
&= \max_{\psi \in \mathcal{P}_{m-1}} \frac{\|(\mathbf{I} - \mathbf{\Pi}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \\
&= \max_{\psi \in \mathcal{P}_{m-1}} \frac{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \, \frac{\|(\mathbf{I} - \mathbf{\Pi}_{\text{good}})(\mathbf{I} - \mathbf{P}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} \\
&\leq \max_{\psi \in \mathcal{P}_{m-1}} \frac{\|(\mathbf{I} - \mathbf{P}_{\text{bad}})(\mathbf{I} - \mathbf{\Pi}_{\text{bad}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \, \frac{\|\mathbf{P}_{\text{bad}}\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} \\
&\leq \|\mathbf{I} - \mathbf{P}_{\text{bad}}\| \max_{\psi \in \mathcal{P}_{m-1}} \frac{\|(\mathbf{I} - \mathbf{\Pi}_{\text{bad}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \, \frac{\|\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} \\
&\leq \|\mathbf{P}_{\text{good}}\| \, \delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{bad}}) \, C_1.
\end{aligned}$$

This gives the first inequality. For the second, note that for any $\psi \in \mathcal{P}_{m-1}$,

$$\begin{aligned}
\frac{\|\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} &= \frac{\|(\mathbf{I} - \mathbf{P}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \, \frac{\|\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} \\
&= \frac{\|(\mathbf{I} - \mathbf{P}_{\text{good}})(\mathbf{I} - \mathbf{\Pi}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \, \frac{\|\psi(\mathbf{A})(\mathbf{P}_{\text{good}} + \mathbf{P}_{\text{bad}})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|} \\
&\leq \|\mathbf{I} - \mathbf{P}_{\text{good}}\| \, \frac{\|(\mathbf{I} - \mathbf{\Pi}_{\text{good}})\psi(\mathbf{A})\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{v}_1\|} \left(1 + \frac{\|\psi(\mathbf{A})\mathbf{P}_{\text{bad}}\mathbf{v}_1\|}{\|\psi(\mathbf{A})\mathbf{P}_{\text{good}}\mathbf{v}_1\|}\right).
\end{aligned}$$

(A more frugal inequality leads to a sharper but rather intricate upper bound for C_1.) Maximizing over $\psi \in \mathcal{P}_{m-1}$ and noting that $\|\mathbf{I} - \mathbf{P}_{\text{good}}\| = \|\mathbf{P}_{\text{good}}\|$ [22] yields

$$C_1 \leq \|\mathbf{P}_{\text{good}}\| \, \delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}})(1 + C_1).$$

When $\|\mathbf{P}_{\text{good}}\| \, \delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}}) < 1$, this expression can be rearranged to give the desired upper bound. ∎

The bounds given in Proposition 5.1 can be disparate when $\|\mathbf{P}_{\text{good}}\|$ is large or $\delta(\mathcal{K}_m(\mathbf{A}, \mathbf{v}_1), \mathcal{U}_{\text{good}})$ is close to one. To obtain alternative lower bounds, approximate the maximizing polynomial ψ in (3.5). Some intuitively appealing choices for the roots of $\psi \in \mathcal{P}_{m-1}$ include the Ritz values or harmonic Ritz values generated from $\mathcal{K}_{m-1}(\mathbf{A}, \mathbf{P}_{\text{good}}\mathbf{v}_1)$. (This is motivated by the fact that taking ψ to be a degree-m polynomial with the m Ritz values from $\mathcal{K}_m(\mathbf{A}, \mathbf{P}_{\text{good}}\mathbf{v}_1)$ as roots would zero the denominator of the expression (3.5) for C_1.)

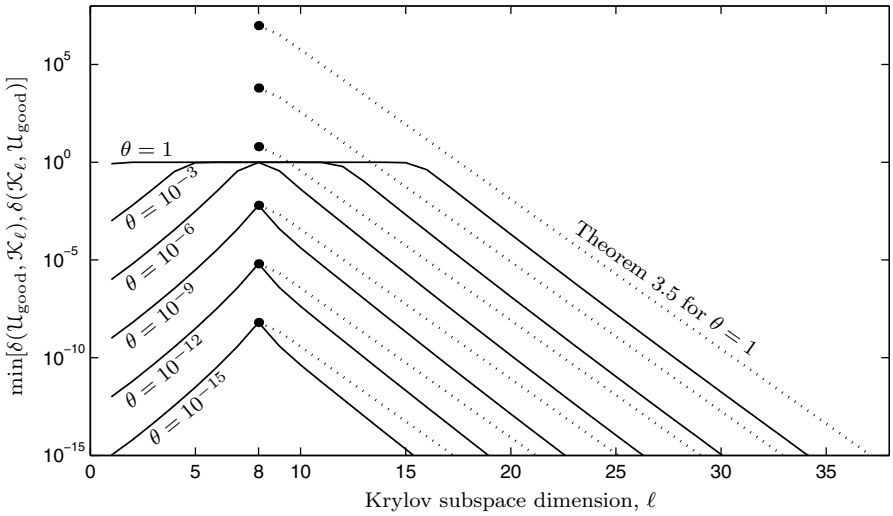


FIG. 5.1. *The effect of a biased starting vector on gap convergence. The solid lines denote the computed gap convergence curves for starting vectors* $\mathbf{v}_1$ *that form angles of* θ *radians with* $\mathcal{U}_{\rm good}$. *The dotted lines show the bound derived from Theorem* 3.5 *for each value of* θ. *The black dots denote the values of* C_1. *In the vertical axis label,* $\mathcal{K}_\ell$ *is a shorthand for* $\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)$.

5.2. An illustration of starting vector influence. Consider a Hermitian matrix $\mathbf{A} \in \mathbb{C}^{128\times 128}$ with eight good eigenvalues uniformly distributed in the interval $[1,2]$. The remaining eigenvalues uniformly fill the interval $[-1,0]$. Since $\mathbf{A}$ is normal, the constants C_0 and C_2 are trivial, $C_0 = C_2 = 1$. Theorem 3.5 thus bounds gap convergence as the product of the constant C_1, which depends on the starting vector, and a polynomial approximation problem, which is independent of it. Taking $\Omega_{\rm bad} = [-1,0]$ and $\Omega_{\rm good} = [1,2]$, Theorem 4.5 yields an asymptotic convergence factor of $3-\sqrt{2} \approx 0.1716$, an expedient rate due to the good separation of $\Omega_{\rm good}$ from $\Omega_{\rm bad}$. To study the role of C_1, we construct six different starting vectors $\mathbf{v}_1$ that form angles of $\theta = 10^{-15}, 10^{-12}, 10^{-9}, 10^{-6}, 10^{-3}, 1$ radians with $\mathcal{U}_{\rm good}$. (Each starting vector has equal components in each unwanted eigenvector direction.) Figure 5.1 shows the result of this experiment. The gap convergence curves are solid lines; the dotted lines show bounds from Theorem 3.5. For the finite-degree polynomial approximation problem in Theorem 3.5, we use Chebyshev polynomials for $\Omega_{\rm bad} = [-1,0]$. (Since $\delta(\mathcal{U}_{\rm good},\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1)) = 1$ when $\ell < m = \dim \mathcal{U}_{\rm good} = 8$, we show the complementary measure $\delta(\mathcal{K}_\ell(\mathbf{A},\mathbf{v}_1),\mathcal{U}_{\rm good})$ for the first seven iterations.) As predicted by our bounds, the asymptotic convergence rate appears largely independent of the orientation of $\mathbf{v}_1$. Interestingly, even a considerable starting vector bias toward $\mathcal{U}_{\rm good}$ yields only a modest improvement in convergence, which may appear even less significant for problems with slower convergence rates.

5.3. Bounding C_2. In contrast to C_1, which was strongly linked to the orientation of the starting vector $\mathbf{v}_1$ with respect to the good invariant subspace, the constant C_2 has a somewhat more diffuse interpretation. C_2 captures the effect of the nonnormality of $\mathbf{A}$, yet ambiguity in the selection of $\Omega_{\rm good}$ and $\Omega_{\rm bad}$ injects wide variability to the values C_2 can achieve. Generally speaking, choosing the sets $\Omega_{\rm good}$ and $\Omega_{\rm bad}$ to be overly large yields a small constant C_2 at the expense of a slow convergence rate for the polynomial approximation problem. Shrinking these sets increases

the constant but improves the predicted convergence rate. The smallest possible sets that can be chosen for Ω_{good} and Ω_{bad} are the sets of good and bad eigenvalues, respectively. If $\mathbf{A}$ is diagonalizable, it is possible to pose the approximation problem over these discrete point sets, at the expense of a potentially large C_2 term arising from eigenvector conditioning.

LEMMA 5.2. *Suppose Σ is a subset of the spectrum of $\mathbf{A}$ consisting only of nondefective eigenvalues, and let $\mathcal{U}$ denote the maximal invariant subspace associated with eigenvalues in Σ. If the columns of $\mathbf{X}$ are eigenvectors of $\mathbf{A}$ forming a basis for $\mathcal{U}$, then*

$$\kappa(\Sigma) \leq \text{cond}_2(\mathbf{X}).$$

(The condition number $\text{cond}_2(\cdot)$ is the ratio of the maximum to the minimum nonzero singular value.)

Proof. Observe that $\mathbf{\Pi} \equiv \mathbf{X}\,(\mathbf{X}^*\mathbf{X})^{-1}\,\mathbf{X}^*$ defines an orthogonal projection onto $\mathcal{U}$, and suppose $\mathbf{\Lambda}$ is a diagonal matrix with entries in Σ such that $\mathbf{AX} = \mathbf{X\Lambda}$. Then for any function f that is analytic on Σ, $f(\mathbf{A})\mathbf{X} = \mathbf{X}f(\mathbf{\Lambda})$, and

$$\begin{aligned}
\|f(\mathbf{A})\|_{\mathcal{U}} &= \|f(\mathbf{A})\mathbf{X}\,(\mathbf{X}^*\mathbf{X})^{-1}\,\mathbf{X}^*\| \\
&= \|\mathbf{X}f(\mathbf{\Lambda})\,(\mathbf{X}^*\mathbf{X})^{-1}\,\mathbf{X}^*\| \\
&\leq \|\mathbf{X}\|\,\|\,(\mathbf{X}^*\mathbf{X})^{-1}\,\mathbf{X}^*\|\,\|f(\mathbf{\Lambda})\| \\
&= \text{cond}_2(\mathbf{X}) \max_{\lambda\in\Sigma} |f(\lambda)|. \qquad \square
\end{aligned}$$

Now if Ω_{good} and Ω_{bad} in Theorem 3.5 are precisely the sets of good and bad eigenvalues of $\mathbf{A}$, respectively, Lemma 5.2 leads to a bound on C_2.

FIRST COROLLARY TO THEOREM 3.5. *To the conditions of Theorem 3.5, add the assumption that $\mathbf{A}$ is diagonalizable,*

$$\mathbf{A}[\mathbf{X}_{\text{good}}, \mathbf{X}_{\text{bad}}] = [\mathbf{X}_{\text{good}}, \mathbf{X}_{\text{bad}}]\,\text{diag}(\mathbf{\Lambda}_{\text{good}},\ \mathbf{\Lambda}_{\text{bad}}).$$

Then

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \leq C_0\, C_1\, \widehat{C}_2 \min_{\phi\in\mathcal{P}_{\ell^*}} \frac{\max_{j=L+1,\ldots,N} |\phi(\lambda_j)\Psi_{\nu p}(\lambda_j)|}{\min_{k=1,\ldots,L} |\phi(\lambda_k)\Psi_{\nu p}(\lambda_k)|}, \tag{5.1}$$

where C_0 and C_1 are as defined in Theorems 3.3 and 3.4 and

$$\widehat{C}_2 \equiv \text{cond}_2(\mathbf{X}_{\text{good}})\,\text{cond}_2(\mathbf{X}_{\text{bad}}).$$

When $\mathbf{A}$ is far from normal, the constant $\widehat{C}_2$ will typically be large; it grows infinite as $\mathbf{A}$ tends toward a defective matrix. However, such extreme situations are not necessarily associated with severe degradation in convergence behavior, and so the bound (5.1) will be most appropriate when $\mathbf{A}$ is either normal or nearly so.

Nonnormality can complicate invariant subspace computation in a variety of ways. The good eigenvalues can be individually ill-conditioned, with $\text{cond}_2(\mathbf{X}_{\text{good}}) \gg 1$, while the associated invariant subspace is perfectly conditioned. In other cases, one may find the good eigenvalues are well-conditioned, while the bad eigenvalues are highly nonnormal (as when $\text{cond}_2(\mathbf{X}_{\text{bad}}) \gg \text{cond}_2(\mathbf{X}_{\text{good}}) \approx 1$).[7] In either case, the

[7] This is the case for the Markov chain example described in section 6.4. Trefethen describes another example, the Gauss–Seidel iteration matrix for the centered difference discretization of the second derivative [43, Ex. 10].

good invariant subspace may still have physical significance, and we would like to understand how this ill-conditioning affects the rate at which we can compute it.

Since nonnormal matrices are of special interest, consideration of pseudospectra yields a natural approach that often can provide sharper, more descriptive convergence bounds. Recall that the *ε-pseudospectrum* [42, 43] is the set

$$\Lambda_\varepsilon(\mathbf{A}) \equiv \{z \in \mathbb{C} : \|(z-\mathbf{A})^{-1}\| \geq \varepsilon^{-1}\},$$

or, equivalently, $\Lambda_\varepsilon(\mathbf{A}) = \{z \in \Lambda(\mathbf{A}+\mathbf{E}) : \|\mathbf{E}\| \leq \varepsilon\}$, where $\Lambda(\mathbf{M})$ denotes the set of eigenvalues of a matrix $\mathbf{M}$.

For a fixed ε, $\Lambda_\varepsilon(\mathbf{A})$ is a closed set in the complex plane consisting of the union of no more than N connected sets, each of which must contain at least one eigenvalue. As $\varepsilon \to 0$, $\Lambda_\varepsilon(\mathbf{A})$ tends to N disjoint disks (whose radii depend on eigenvalue conditioning and defectiveness) centered at and shrinking around the N distinct eigenvalues.

LEMMA 5.3. *Let $\mathcal{U}$ be an invariant subspace of $\mathbf{A}$ and suppose Σ is the set of eigenvalues associated with $\mathcal{U}$.*

(a) *Let Ω be a set containing Σ but no eigenvalues of $\mathbf{A}$ outside Σ, and suppose the boundary $\partial\Omega$ is the finite union of positively oriented Jordan curves. Then*

$$\kappa(\Omega) \leq \frac{1}{2\pi} \int_{\partial\Omega} \|(z-\mathbf{A})^{-1}\|_{\mathcal{U}} \, |dz|. \tag{5.2}$$

(b) *Let Σ_ε contain the union of those connected components of $\Lambda_\varepsilon(\mathbf{A})$ that include $\lambda \in \Sigma$, and suppose further that Σ_ε contains no eigenvalues outside of Σ and its boundary $\partial\Sigma_\varepsilon$ is the finite union of positively oriented Jordan curves. Then*

$$\kappa(\Sigma_\varepsilon) \leq \frac{\mathcal{L}(\partial\Sigma_\varepsilon)}{2\pi\varepsilon}, \tag{5.3}$$

where $\mathcal{L}(\partial\Sigma_\varepsilon)$ is the length of the boundary of Σ_ε.

Proof. For part (a), let $\mathbf{\Pi}$ be the orthogonal projector onto the given invariant subspace $\mathcal{U}$ and let $\mathbf{P}$ be the spectral projector for $\mathbf{A}$ associated with $\mathcal{U}$. For any function f analytic on Ω, $\|f(\mathbf{A})\|_{\mathcal{U}} = \|f(\mathbf{A})\mathbf{\Pi}\| = \|f(\mathbf{A})\mathbf{P}\mathbf{\Pi}\| \leq \|f(\mathbf{A})\mathbf{P}\|$. Now,

$$f(\mathbf{A})\mathbf{P} = \frac{1}{2\pi \mathrm{i}} \int_{\partial\Omega} f(z)(z-\mathbf{A})^{-1} \, dz.$$

Thus for any vector $\mathbf{x} \in \mathcal{U}$,

$$\begin{aligned} \|f(\mathbf{A})\mathbf{x}\| &\leq \frac{1}{2\pi} \int_{\partial\Omega} |f(z)| \, \|(z-\mathbf{A})^{-1}\mathbf{x}\| \, |dz| \\ &\leq \left(\frac{1}{2\pi} \int_{\partial\Omega} \|(z-\mathbf{A})^{-1}\|_{\mathcal{U}} \, |dz| \right) \max_{z\in\partial\Omega} |f(z)| \, \|\mathbf{x}\|. \end{aligned}$$

But since f is analytic on Ω, $\max_{z\in\partial\Omega} |f(z)| = \max_{z\in\Omega} |f(z)|$. Part (b) follows from (a) by assigning $\Omega = \Sigma_\varepsilon$. ∎

Pseudospectral bounds were developed by Trefethen to bound the GMRES residual norm [42], and Simoncini has used a similar approach to analyze block-Arnoldi convergence [37]. In the single eigenvector case, her Theorem 3.1 closely resembles our (5.6) below. (Lemma 5.3 could easily be sharpened to instead involve $\Lambda_\varepsilon(\mathbf{U}^*\mathbf{A}\mathbf{U})$, where the columns of $\mathbf{U}$ form an orthonormal basis for $\mathcal{U}_{\text{good}}$; note that $\Lambda_\varepsilon(\mathbf{U}^*\mathbf{A}\mathbf{U}) \subseteq \Lambda_\varepsilon(\mathbf{A})$ [40].)

The pseudospectral approach leads to a robust alternative to the eigenvector-based bound (5.1).[8] Suppose ε is sufficiently small that the components of the ε-pseudospectrum enclosing the good eigenvalues are disjoint from those components enclosing the bad eigenvalues. $\Lambda_\varepsilon(\mathbf{A})$ can then be contained in the two disjoint sets $\Sigma_\varepsilon^{\text{good}}$ and $\Sigma_\varepsilon^{\text{bad}}$, leading to an alternative bound.

SECOND COROLLARY TO THEOREM 3.5. *Assume the conditions of Theorem* 3.5 *and suppose that $\varepsilon > 0$ is sufficiently small that $\Sigma_\varepsilon^{\text{good}} \cap \Sigma_\varepsilon^{\text{bad}} = \emptyset$. Then, provided $\Psi_{\nu p}(z)$ has no roots in $\Sigma_\varepsilon^{\text{good}}$, and the boundaries of $\Sigma_\varepsilon^{\text{good}}$ and $\Sigma_\varepsilon^{\text{bad}}$ are finite unions of positively oriented Jordan curves,*

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \le C_0\, C_1\, \widetilde{C}_2(\varepsilon) \min_{\phi \in \mathcal{P}_{\ell^*}} \frac{\max\{|\phi(z)\Psi_{\nu p}(z)| : z \in \Sigma_\varepsilon^{\text{bad}}\}}{\min\{|\phi(z)\Psi_{\nu p}(z)| : z \in \Sigma_\varepsilon^{\text{good}}\}}, \tag{5.4}$$

where C_0 and C_1 are as defined in Theorems 3.3 *and* 3.4, *and*

$$\widetilde{C}_2(\varepsilon) \equiv \frac{\mathcal{L}(\partial\Sigma_\varepsilon^{\text{good}})\, \mathcal{L}(\partial\Sigma_\varepsilon^{\text{bad}})}{4\pi^2\varepsilon^2}. \tag{5.5}$$

$\mathcal{L}(\partial\Sigma_\varepsilon^{\text{good}})$ and $\mathcal{L}(\partial\Sigma_\varepsilon^{\text{bad}})$ are the boundary lengths of $\Sigma_\varepsilon^{\text{good}}$ and $\Sigma_\varepsilon^{\text{bad}}$, respectively.

This pseudospectral bound holds for a range of ε-values, providing a natural mechanism for adjusting the sets Ω_{good} and Ω_{bad}. As ε gets smaller, $\widetilde{C}_2(\varepsilon)$ generally increases, but the convergence rate induced by the polynomial approximation problem improves, since the sets on which the approximation problem is posed recede from one another. For the most descriptive convergence bound, take the envelope of individual bounds corresponding to a variety of ε-values; see Figures 6.1 and 6.3. Of course, the bound (5.4) is only meaningful when ε is sufficiently small that $\Sigma_\varepsilon^{\text{good}} \cap \Sigma_\varepsilon^{\text{bad}} = \emptyset$. The need to take ε particularly small to satisfy this condition may signal an ill-conditioned problem; consider enlarging the set of good eigenvalues.

In some situations, one may wish to use different values of ε for the good and bad pseudospectra, in which case (5.4) changes in the obvious way. Furthermore, when the good eigenvalues are normal (i.e., one can take $\text{cond}_2(\mathbf{X}_{\text{good}}) = 1$), it is best to combine the pseudospectra and eigenvector approaches to obtain

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1^{(\nu)})) \le \frac{C_0\, C_1\, \mathcal{L}(\Sigma_\varepsilon^{\text{bad}})}{2\pi\varepsilon} \min_{\phi \in \mathcal{P}_{\ell^*}} \frac{\max\{|\phi(z)\Psi_{\nu p}(z)| : z \in \Sigma_\varepsilon^{\text{bad}}\}}{\min_{k=1,\ldots,L} |\phi(\lambda_k)\Psi_{\nu p}(\lambda_k)|}. \tag{5.6}$$

We close this section by pointing out one nonnormal situation where the eigenvector-based bound (5.1) can be dramatically superior to the pseudospectral bound (5.4). Suppose for simplicity that $\dim \mathcal{U}_{\text{good}} = \dim \mathcal{U}_{\text{bad}}$ with $\mathcal{U}_{\text{good}} \approx \mathcal{U}_{\text{bad}}$ for some diagonalizable $\mathbf{A}$. It is possible for the basis vectors in $\mathbf{X}_{\text{good}}$ and $\mathbf{X}_{\text{bad}}$ to be perfectly conditioned on their own, but terribly conditioned if taken together, e.g.,

$$\mathbf{X}_{\text{good}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{X}_{\text{bad}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \gamma & 0 \\ 0 & \gamma \end{bmatrix},$$

with $0 < |\gamma| \ll 1$. This results in $\widehat{C}_2 = 1$ but $\widetilde{C}_2(\varepsilon) \gg 1$ for usefully small values of ε. (This can be remedied by considering the pseudospectra of $\mathbf{A}$ orthogonally projected

[8] Note that Greenbaum has demonstrated how more clever use of eigenvector information can sometimes be superior to estimating integrals of the resolvent norm [15].

onto $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$.) What is happening here? The more alike $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$ are, the more prominent their general orientation is in the Krylov subspace, possibly resulting in an initial period of rapid sublinear convergence. Discriminating the fine difference between $\mathcal{U}_{\text{good}}$ and $\mathcal{U}_{\text{bad}}$ may still be challenging.

6. Some examples. How well does the machinery constructed in the previous sections work? Here we demonstrate our bounds for a variety of examples. These test problems are contrived to illustrate the effects we have described as cleanly as possible. Eigenvalue problems from applications inevitably involve more complicated spectral structure.

6.1. Influence of nonnormality on predicted rates. We begin with two examples involving nondiagonalizable matrices where pseudospectral convergence bounds can be used to good effect. (While the examples in this subsection and the next are defective, we emphasize that the pseudospectral bound can also be useful for diagonalizable matrices with large values of $\widehat{C}_2$.) Define

$$\mathbf{A} = \begin{bmatrix} \mathbf{D}_{\text{good}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{58}(-1) \end{bmatrix}, \tag{6.1}$$

where $\mathbf{D}_{\text{good}}$ is a 6×6 diagonal matrix containing good eigenvalues uniformly distributed in $[1, 2]$, and $\mathbf{J}_{58}(-1)$ is a Jordan block of dimension 58 with the bad eigenvalue $\lambda = -1$ on the main diagonal and 1's on the first superdiagonal. Note that $\mathcal{U}_{\text{good}} \perp \mathcal{U}_{\text{bad}}$, so $C_0 = 1$. Since the good eigenvalues are normal, we apply the hybrid pseudospectral bound (5.6). The ε-pseudospectrum of a direct sum of matrices is the union of the ε-pseudospectra of each component matrix [45], so we need focus only on the pseudospectra of the Jordan block, which are circular disks for all $\varepsilon > 0$ [30]; see Figure 6.1. It follows that $\widetilde{C}_2(\varepsilon) = r_\varepsilon/\varepsilon$, where r_ε is the radius of $\Sigma_\varepsilon^{\text{bad}} = \Lambda_\varepsilon(\mathbf{J}_{58}(-1))$, determined numerically. For $\phi \in \mathcal{P}_{\ell^*}$ we take the Chebyshev polynomial for $\Sigma_\varepsilon^{\text{bad}}$, $\phi(z) = (z+1)^{\ell^*}$. For all ε such that $r_\varepsilon < 2$, (5.6) gives

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le \frac{C_1 r_\varepsilon}{\varepsilon} \left(\frac{r_\varepsilon}{2}\right)^{\ell^*}, \tag{6.2}$$

where we have used the fact that $|\phi(\lambda)| \ge 2$ for all good eigenvalues λ. The convergence curve and corresponding bounds are shown in Figure 6.1 for the starting vector $\mathbf{v}_1$ with $1/\sqrt{n}$ in each component; no restarting is performed. Interestingly, for small values of ε the bound (5.6) accurately captures the finite termination that must occur when $\ell = n = 64$, a trait exhibited by pseudospectral bounds in other contexts.

Our second example is the same, except the good eigenvalues are now replaced with a Jordan block,

$$\mathbf{A} = \begin{bmatrix} \mathbf{J}_6(\frac{3}{2}) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{58}(-1) \end{bmatrix}, \tag{6.3}$$

where $\mathbf{J}_6(\frac{3}{2})$ is a 6×6 Jordan block with $\frac{3}{2}$ on the main diagonal and 1's on the first superdiagonal; $\mathbf{J}_{58}(-1)$ is as before. Again note that $\mathcal{U}_{\text{good}} \perp \mathcal{U}_{\text{bad}}$, implying $C_0 = 1$. Since both the good and bad eigenvalues are defective, apply the pseudospectral bound (5.4). Recalling that the pseudospectra of Jordan blocks are circular disks, let $r_\varepsilon^{\text{bad}}$ and $r_\varepsilon^{\text{good}}$ denote the radii of $\Sigma_\varepsilon^{\text{bad}} = \Lambda_\varepsilon(\mathbf{J}_{58}(-1))$ and $\Sigma_\varepsilon^{\text{good}} = \Lambda_\varepsilon(\mathbf{J}_6(\frac{3}{2}))$, respectively; see the left plot of Figure 6.2. The Second Corollary to Theorem 3.5

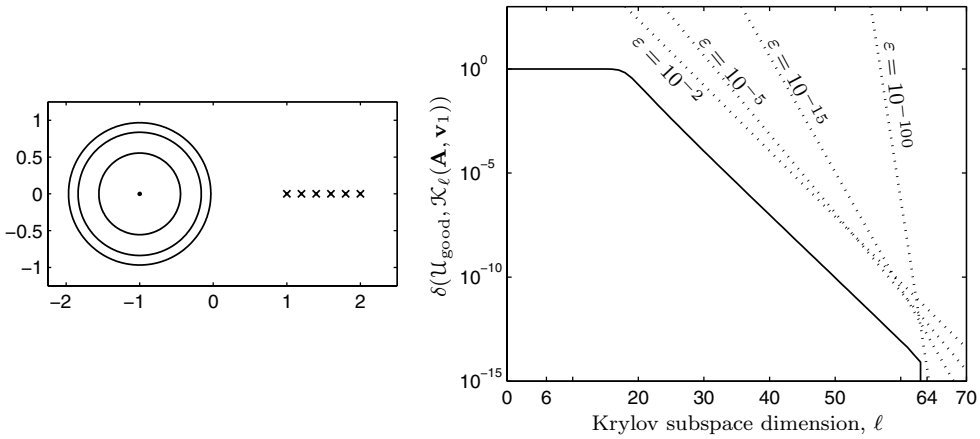


FIG. 6.1. *On the left, good eigenvalues* (×) *and pseudospectral boundaries* $\partial\Sigma_\varepsilon^{\text{bad}}$ *for* $\varepsilon = 10^{-2}$, 10^{-5}, 10^{-15}, *and* 10^{-100}, *where* $\mathbf{A}$ *is given by* (6.1). *(The bad eigenvalue* (·) *is obscured by the* $\varepsilon = 10^{-100}$ *boundary.) On the right, gap convergence (solid line) together with the bound* (6.2) *(dotted lines) for each of the pseudospectral curves shown on the left. For small values of* ε, (6.2) *captures the finite termination that must occur at the* 64*th iteration.*

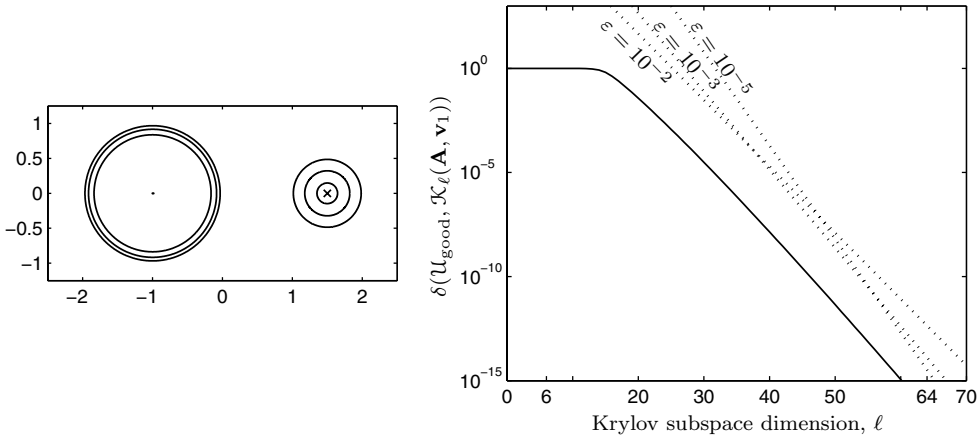


FIG. 6.2. *On the left, bad eigenvalue* (·), *good eigenvalue* (×), *and pseudospectral boundaries* $\partial\Sigma_\varepsilon^{\text{bad}}$ *and* $\partial\Sigma_\varepsilon^{\text{good}}$ *for* $\mathbf{A}$ *given by* (6.3) *and* $\varepsilon = 10^{-2}$, 10^{-3}, *and* 10^{-5}. *On the right, gap convergence (solid line) with the bound* (6.4) *(dotted lines) for the three* ε *values used in the left plot.*

holds whenever $r_\varepsilon^{\text{bad}} + r_\varepsilon^{\text{good}} < \frac{5}{2}$. For such ε, $\widetilde{C}_2(\varepsilon) = r_\varepsilon^{\text{bad}}\, r_\varepsilon^{\text{good}}/\varepsilon^2$ and

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le C_1 \frac{r_\varepsilon^{\text{bad}}\, r_\varepsilon^{\text{good}}}{\varepsilon^2} \left(\frac{r_\varepsilon^{\text{bad}}}{\frac{5}{2} - r_\varepsilon^{\text{good}}} \right)^{\ell^*}, \tag{6.4}$$

where again we have taken for $\phi \in \mathcal{P}_{\ell^*}$ the Chebyshev polynomial for $\Sigma_\varepsilon^{\text{bad}}$, $\phi(z) = (z+1)^{\ell^*}$. The convergence curve and corresponding bounds are shown in Figure 6.2 for the starting vector $\mathbf{v}_1$ with $1/\sqrt{n}$ in each component; no restarting is performed.

6.2. Superlinear effects due to nonnormality. Our final example of pseudospectral bounds addresses the matrix

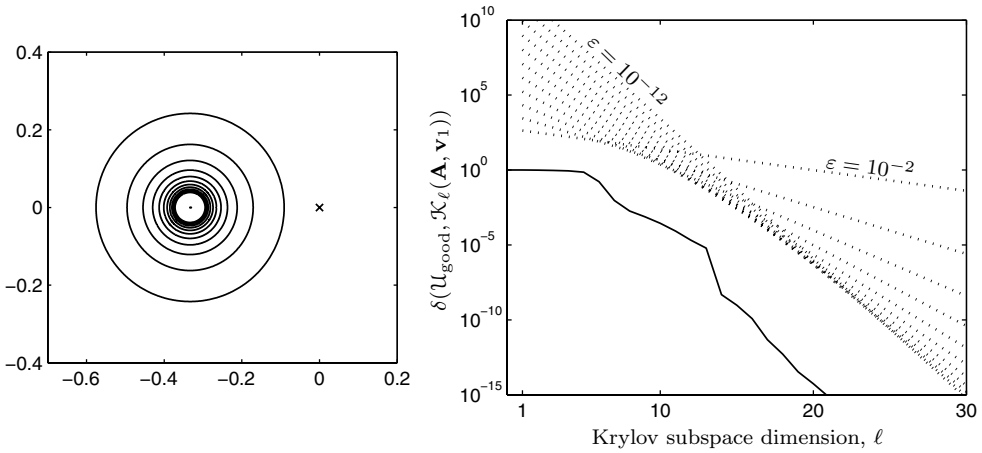


FIG. 6.3. *On the left, bad eigenvalue* (·)*, good eigenvalue* (×)*, and pseudospectral boundaries* $\partial\Sigma_\varepsilon^{\text{bad}}$ *for* $\mathbf{A}$ *given by* (6.5) *and* $\varepsilon = 10^{-2}, \ldots, 10^{-12}$*. On the right, gap convergence (solid line) with the bound* (6.4) *(dotted lines) for the eleven* ε *values shown in the left plot.*

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{bmatrix}, \tag{6.5}$$

where there is a single good eigenvalue $\lambda = 0$ (with multiplicity 1) and a bad eigenvalue $\lambda = -\frac{1}{3}$ associated with the 63×63 bidiagonal matrix $\mathbf{F}$, which has $-\frac{1}{3}$ in the main diagonal entries and $1/j$ in the $(j, j+1)$ entry of the superdiagonal. Like the Jordan blocks described before, the pseudospectra of $\mathbf{F}$ are circular disks [30], but the radii of these disks shrink much more rapidly as ε decreases than observed for the Jordan block. As a result, the convergence rate steadily improves as ε gets smaller; this is compensated by growing $\widetilde{C}_2(\varepsilon)$ values. Taking $\phi(z) = (z + \frac{1}{3})^{\ell^*}$, we obtain

$$\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1)) \le \frac{C_1 r_\varepsilon}{\varepsilon} \, (3r_\varepsilon)^{\ell^*}, \tag{6.6}$$

provided $r_\varepsilon < \frac{1}{3}$, where r_ε is the radius of $\Sigma_\varepsilon^{\text{bad}}$. Figure 6.3 shows the spectrum of $\mathbf{A}$ and pseudospectra of $\mathbf{F}$. As ε gets smaller, the bound (6.6) traces out an envelope that predicts early stagnation followed by improving linear convergence rates. This is "superlinear" convergence, but of a different nature from that described in section 4.3. Figure 6.3 shows these bounds along with the gap convergence curve for a vector $\mathbf{v}_1$ with real entries drawn from the standard normal distribution. Pseudospectral bounds for GMRES exhibit similar superlinear behavior for matrices like $\mathbf{F}$ [10, 12]. Although all the examples here have used defective matrices, these bounds are also appropriate for diagonalizable matrices with a large eigenvector condition number.

6.3. Shift selection for restarted algorithms. The results of section 4 indicate that effective restart strategies can be constructed using optimal polynomials associated with sets containing the bad eigenvalues. In this section, we give some examples of how choices for $\Psi_{\nu p}$ based on partial information (or misinformation) about bad eigenvalue location affect the observed convergence rates and illustrate how well our bounds can predict this.

Consider the 200×200 upper triangular matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{D}_{\text{good}} & \mathbf{C} \\ \mathbf{0} & \mathbf{D}_{\text{bad}} \end{bmatrix},$$

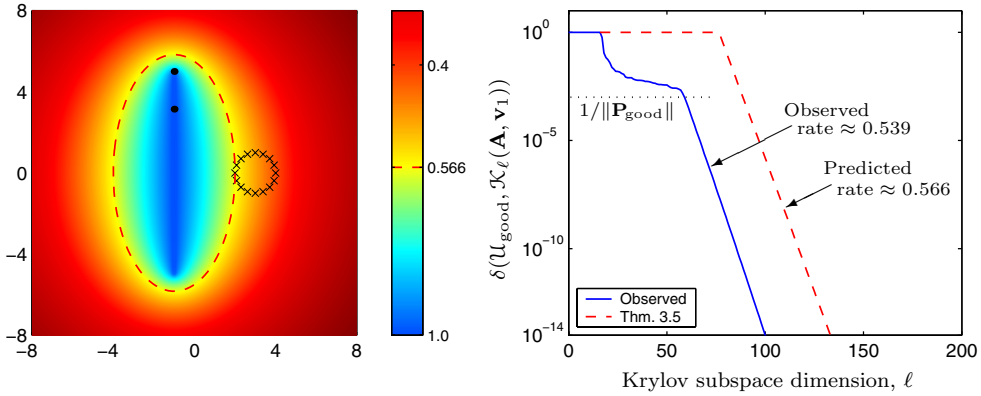


FIG. 6.4. *Unrestarted subspace. On the left, good and bad eigenvalues are shown in the "potential field" generated by the bad eigenvalues. The colorbar is calibrated to show effective convergence rates for different components of* $\mathcal{U}_{\text{good}}$. *The right plot shows the observed gap history (solid line) together with a bound (dashed line) derived from the First Corollary to Theorem* 3.5.

where $\mathbf{D}_{\text{good}}$ is a 16×16 diagonal matrix of good eigenvalues, distributed uniformly around the circle in the complex plane centered at 3 with radius 1; $\mathbf{D}_{\text{bad}}$ is a diagonal matrix containing the bad eigenvalues distributed uniformly along the line segment (designated $\mathcal{I}_{\text{bad}}$) parallel to the imaginary axis connecting the points $-1 \pm 5\text{i}$; $\mathbf{C}$ is a full (row) rank matrix scaled so that $\|\mathbf{P}_{\text{good}}\| \approx 1000$. The starting vector, $\mathbf{v}_1$, has normally distributed random complex entries. (The same $\mathbf{v}_1$ was used for all experiments shown in this subsection.)

Figure 6.4 compares the predicted and observed convergence curves for the unrestarted iteration, where the Krylov subspace grows without bound. The left plot displays the equipotentials of $g[z, \mathcal{I}_{\text{bad}}]$—the physical analog is the potential field generated by a continuous (line) charge distribution spread over $\mathcal{I}_{\text{bad}}$. The color bar is calibrated to show $\exp(-g[z, \mathcal{I}_{\text{bad}}])$, giving the predicted convergence rates at locations in the complex plane if good eigenvalues were present there. In particular, the lowest equipotential contour passing through a good eigenvalue is shown; it leads via (4.5) to a predicted convergence rate of ≈ 0.566. The right plot shows the iteration history of $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}, \mathbf{v}_1))$ versus the iteration index ℓ. After an early sublinear surge that flattens out near $1/\|\mathbf{P}_{\text{good}}\|$, an observed linear rate of ≈ 0.539 emerges. In separate experiments (not shown), we have varied the magnitude of $\|\mathbf{C}\|$ (in effect changing $\|\mathbf{P}_{\text{good}}\|$) and have observed variations in the sublinear stagnation level roughly proportional to $1/\|\mathbf{P}_{\text{good}}\|$, consistent with the discussion surrounding Figure 2.1. The convergence bound is derived from the First Corollary to Theorem 3.5, using for ϕ Chebyshev polynomials for $\mathcal{I}_{\text{bad}}$. (For all experiments in this subsection, $C_0 = \sqrt{2}$, $C_1 \approx 4.4325 \times 10^{11}$, $\widehat{C}_2 \approx 1.2439 \times 10^3$.)

Figure 6.5 shows results for polynomial restarts using fast Leja points [3] associated with $\mathcal{I}_{\text{bad}}$. These appear as a dense line of white dots atop the black band of bad eigenvalues. The base dimension is 20 and restarts are each of order 5. (The Krylov subspace dimension never exceeds 25.) The left plot displays the effective potential, $g[z, \Omega_{\text{bad}}]$, generated by 180 fast Leja points—Ω_{bad} is the smallest polynomial lemniscate generated by the aggregate filter polynomial that contains all bad eigenvalues. The lowest equipotential contour passing through a good eigenvalue is shown; it leads via (4.5) and Example 4.2 to a predicted convergence rate of ≈ 0.576. The bound on the right was obtained from the First Corollary to Theorem 3.5, using

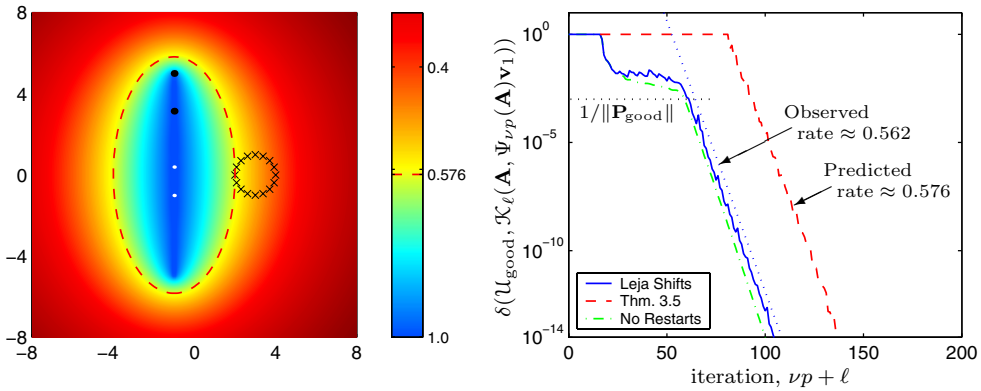


FIG. 6.5. *Polynomial restarts at fast Leja points of* $\mathcal{I}_{\text{bad}}$ *(band of closely spaced white dots). The base dimension is* 20 *and restarts are each of degree* $p = 5$ *(so the subspace dimension never exceeds* 25*).*

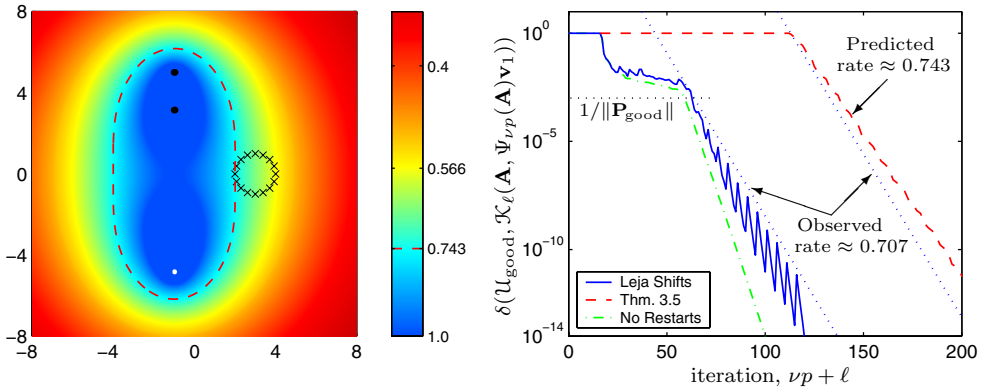


FIG. 6.6. *Polynomial restarts with fast Leja points (twin bands of closely spaced white dots) for two subintervals covering only* 60% *of the bad eigenvalues. The subspace dimensions are as in Figure* 6.5*.*

Chebyshev polynomials for $\mathcal{I}_{\text{bad}}$ up to the base dimension, then including the shift polynomials.

Figures 6.6 and 6.7 show the effect of poorer choices for the filter shifts. Suppose we mistakenly believe the bad eigenvalues to be concentrated toward the ends of the interval $\mathcal{I}_{\text{bad}}$ and choose filter shifts accordingly grouped in two subintervals that omit the central portion of $\mathcal{I}_{\text{bad}}$ (which we believe to be devoid of bad eigenvalues). We use fast Leja points again but this time for pairs of disjoint intervals that in fact cover only 60% and 20%, respectively, of the bad eigenvalues. These are asymptotically optimal filter shifts for misguided guesses of the bad eigenvalue distribution. Ω_{bad} is again the smallest polynomial lemniscate generated by 180 fast Leja points that contains all bad eigenvalues. Here it takes on a more pronounced dumb-bell appearance, reflecting the absence of zeros from the middle of $\mathcal{I}_{\text{bad}}$. As before, the base dimension is 20 and restarts are each of order 5. The convergence rate is seen to deteriorate to ≈ 0.707 and ≈ 0.807, respectively, and is predicted to within an accuracy of roughly 3%–5.2%. By comparing the equipotential contours of Figures 6.4 and 6.5 with those of Figures 6.6 and 6.7, notice the filter shifts in the latter cases create a potential significantly different from what either the bad eigenvalues or optimal filter shifts would generate.

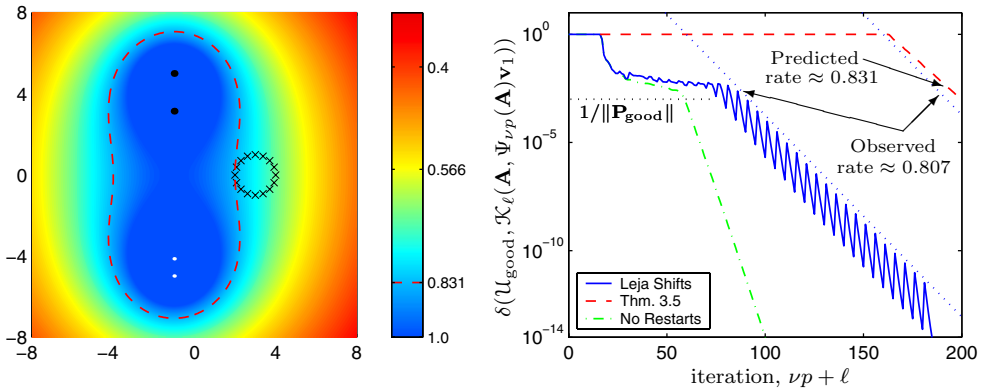


FIG. 6.7. *Polynomial restarts with fast Leja points (twin bands of closely spaced white dots) for two subintervals covering only* 20% *of the bad eigenvalues. The subspace dimensions are as in Figure* 6.5.

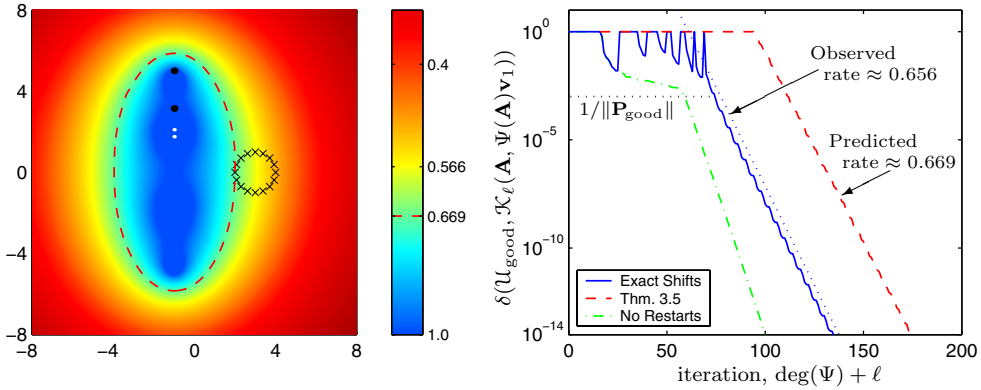


FIG. 6.8. *Polynomial restarts using exact shifts (white dots) determined by choosing Ritz values with real part smaller than* 1. *The subspace dimension never exceeds* 20.

Figure 6.8 shows the result of using Sorensen's exact shifts. The subspace dimension is limited to be no larger than 20, and a Ritz value is used as a shift if it has real part smaller than 1. (The early convergence plateaus occur when the subspace is compressed to have dimension smaller than the number of good eigenvalues.) The potential plot on the left is based on 180 exact shifts. Although these shifts fall outside the convex hull of the bad eigenvalues, they effectively recover the potential generated by those eigenvalues. The convergence rate is predicted to within 2% of the observed rate. The use of exact shifts yields a convergence rate within 25% of the rate for the unrestarted iteration (Figure 6.4) at a lower computational cost and without requiring a priori localization of bad eigenvalues to determine optimal shifts (as in Figure 6.5 for good localization and Figures 6.6 and 6.7 for poor localization).

6.4. Markov chain example. We close by examining a more realistic eigenvalue problem, taking $\mathbf{A}$ to be the transition matrix for a Markov chain that describes a random walk on a triangular lattice. See Saad [35, sect. II.5.1] for details of this example, a common test problem for iterative eigenvalue algorithms. Since all the rows of a transition matrix sum to 1, $\mathbf{A}$ must have an eigenvalue $\lambda = 1$, and the

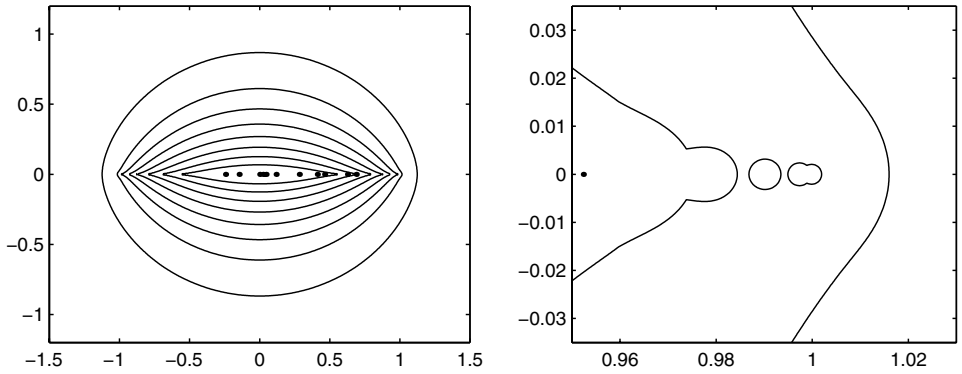


FIG. 6.9. *Eigenvalues and pseudospectra for the random walk transition matrix for a triangular lattice with* 1275 *nodes. The left plot shows the spectrum and boundaries of ε-pseudospectra for $\varepsilon = 10^{-1}, \ldots, 10^{-8}$. The right plot zooms around $\lambda = 1$, indicating ε-pseudospectra for $\varepsilon = 10^{-2}, 10^{-3}$.*

Perron–Frobenius theorem assures this eigenvalue is simple (see, e.g., [5, Thm. 1.4]). The *left* eigenvector corresponding to $\lambda = 1$ determines a stationary distribution of the Markov chain, so we are interested in the convergence of $\delta(\mathcal{U}_{\text{good}}, \mathcal{K}_\ell(\mathbf{A}^*, \mathbf{v}_1))$, where $\mathcal{U}_{\text{good}}$ is the invariant subspace of $\mathbf{A}^*$ for $\lambda = 1$. Here we consider a lattice with a base and height of 50 nodes, yielding a transition matrix of dimension $n = 1275$. This matrix exhibits a significant degree of nonnormality, mostly associated with ill-conditioned eigenvalues far from $\lambda = 1$, as one can infer from the pseudospectra illustrated in Figure 6.9. Unlike the previous examples in this section, the good eigenvalue is quite close to bad eigenvalues, as highlighted by the close-up on the right of Figure 6.9.

The eigenvalues of $\mathbf{A}$ appear to be real with $\lambda = 0$ having algebraic and geometric multiplicity 25. (Though we formally stipulate that $\mathbf{A}$ be nonderogatory in section 3, our proofs require only that the good eigenvalues be nonderogatory.) The bound (5.1) based on the conditioning of the matrices of good and bad eigenvectors is simplest to evaluate. We have $C_0 = \sqrt{2}$, and compute $\widehat{C}_2 \approx 3.546 \times 10^9$; for a particular starting vector with normally distributed real random entries, $C_1 \approx 9.933$. Labeling the eigenvalues from right to left, the polynomial approximation problem in (5.1) reduces in this single eigenvector case to a minimax approximation on $\Lambda_{\text{bad}} = \{\lambda_2, \ldots, \lambda_n\}$ subject to normalization at $\lambda_1 = 1$. Bounding this approximation problem using Chebyshev polynomials on $[\lambda_n, \lambda_2]$ gives a pessimistic result, as can be seen in the convergence plot in Figure 6.10. The superlinear bounds of Theorem 4.7 yield a marked improvement. In the language of Theorem 4.7, we take $\Omega_k = \{\lambda_j\}_{j=k+1}^n$ and reduce to an approximation problem over Ω_{r+1} for $r = 1, \ldots, 10$, for which we use Chebyshev polynomials on $[\lambda_n, \lambda_r]$. An even better bound is obtained by treating Λ_{bad} completely as a discrete point set. One approachable way of doing this is to take $\Lambda_{\text{good}} = \{\lambda_1\}$ and note that

$$\min_{\phi \in \mathcal{P}_{\ell^*}} \frac{\max\{|\phi(\lambda)| : \lambda \in \Lambda_{\text{bad}}\}}{\min\{|\phi(\lambda)| : \lambda \in \Lambda_{\text{good}}\}} = \min_{\substack{\phi \in \mathcal{P}_{\ell^*} \\ \phi(\lambda_1)=1}} \max_{\lambda \in \Lambda_{\text{bad}}} |\phi(\lambda)| \leq \min_{\substack{\phi \in \mathcal{P}_{\ell^*} \\ \phi(0)=1}} \|\phi(\mathbf{S})\mathbf{r}\|, \tag{6.7}$$

where $\mathbf{S} = \text{diag}(\lambda_2 - \lambda_1, \ldots, \lambda_n - \lambda_1)$ and $\mathbf{r} = [1, 1, \ldots, 1]^{\text{T}}$. The last term of (6.7) can be computed as the residual norm of the GMRES algorithm applied to $\mathbf{S}$ with initial residual $\mathbf{r}$; this is no more than a factor of $\sqrt{n}$ worse than the first term in (6.7). The

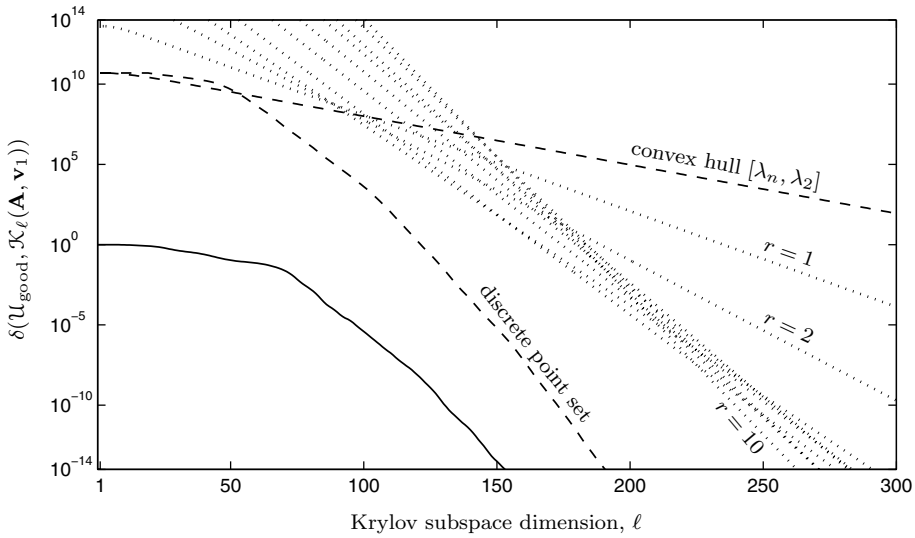


FIG. 6.10. *Gap convergence for the random walk example,* $n = 1275$ *(solid line). The dashed lines represent the bound* (5.1)*. The best result is obtained when the bad eigenvalues are treated as a discrete point set for the approximation problem, while a slower rate is predicted when the bad eigenvalues are treated as an interval. The dotted lines utilize the superlinear bounds of Theorem* 4.7 *for* $r = 1, \ldots, 10$.

resultant bound is shown in Figure 6.10. Alternatively, the minimax problem on the left-hand side of (6.7) could be solved directly via a linear program.

Acknowledgments. We thank Dan Sorensen for many constructive comments during the early stages of this work, and Joe Ball, Nick Trefethen, and Henk van der Vorst for helpful discussions. The pseudospectral computations in section 6 were based on software developed by Trefethen and Wright [44, 49].

REFERENCES

[1] M. Arioli, V. Pták, and Z. Strakoš, *Krylov sequences of maximal length and convergence of GMRES*, BIT, 38 (1998), pp. 636–643.
[2] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[3] J. Baglama, D. Calvetti, and L. Reichel, *Fast Leja points*, Electron. Trans. Numer. Anal., 7 (1998), pp. 124–140.
[4] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.
[5] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.
[6] D. Calvetti, L. Reichel, and D. C. Sorensen, *An implicitly restarted Lanczos method for large symmetric eigenvalue problems*, Electron. Trans. Numer. Anal., 2 (1994), pp. 1–21.
[7] F. Chatelin, *Eigenvalues of Matrices*, Wiley, Chichester, UK, 1993.
[8] J. B. Conway, *Functions of One Complex Variable*, 2nd ed., Springer-Verlag, New York, 1978.
[9] G. De Samblanx and A. Bultheel, *Nested Lanczos: Implicitly restarting an unsymmetric Lanczos algorithm*, Numer. Algorithms, 18 (1998), pp. 31–50.
[10] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.
[11] N. Dunford and J. Schwartz, *Linear Operators, Part* I: *General Theory*, Wiley, New York, 1971.

[12] M. Embree, *Convergence of Krylov Subspace Methods for Nonnormal Matrices*, D.Phil. Thesis, Oxford University, Oxford, UK, 2000.
[13] F. R. Gantmacher, *The Theory of Matrices, Vol.* 1, 2nd ed., Chelsea, New York, 1959.
[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins, Baltimore, MD, 1996.
[15] A. Greenbaum, *Using the Cauchy Integral Formula and the Partial Fractions Decomposition of the Resolvent to Estimate* $\|f(A)\|$, manuscript, University of Washington, Seattle, WA, 2000.
[16] E. J. Grimme, D. C. Sorensen, and P. van Dooren, *Model reduction of state space systems via an implicitly restarted Lanczos method*, Numer. Algorithms, 12 (1995), pp. 1–31.
[17] M. Haviv and Y. Ritov, *Bounds on the error of an approximate invariant subspace for non-self-adjoint matrices*, Numer. Math., 67 (1994), pp. 491–500.
[18] V. Heuveline and M. Sadkane, *Arnoldi-Faber method for large non Hermitian eigenvalue problems*, Electron. Trans. Numer. Anal., 5 (1997), pp. 62–76.
[19] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[20] Z. Jia, *The convergence of generalized Lanczos methods for large unsymmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 843–862.
[21] Z. Jia and G. W. Stewart, *An analysis of the Rayleigh–Ritz method for approximating eigenspaces*, Math. Comp., 70 (2001), pp. 637–647.
[22] T. Kato, *Estimation of iterated matrices, with application to the von Neumann condition*, Numer. Math., 2 (1960), pp. 22–29.
[23] T. Kato, *Perturbation Theory for Linear Operators*, corrected 2nd ed., Springer-Verlag, Berlin, 1980.
[24] L. Knizhnerman, *Error bounds for the Arnoldi method: A set of extreme eigenpairs*, Linear Algebra Appl., 296 (1999), pp. 191–211.
[25] R. B. Lehoucq, *Implicitly restarted Arnoldi methods and subspace iteration*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 551–562.
[26] A. L. Levin and E. B. Saff, *Optimal ray sequences of rational functions connected with the Zolotarev problem*, Constr. Approx., 10 (1994), pp. 235–273.
[27] R. B. Morgan, *On restarting the Arnoldi method for large nonsymmetric eigenvalue problems*, Math. Comp., 65 (1996), pp. 1213–1230.
[28] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, Switzerland, 1993.
[29] V. I. Paulsen, *Completely Bounded Maps and Dilations*, Longman Scientific and Technical, Harlow, UK, 1986.
[30] L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162/164 (1992), pp. 153–185.
[31] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.
[32] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.
[33] Y. Saad, *Projection methods for solving large sparse eigenvalue problems*, in Matrix Pencils: Proceedings, Lecture Notes in Math. 973, B. Kågström and A. Ruhe, eds., Springer-Verlag, Berlin, 1983, pp. 121–144.
[34] Y. Saad, *Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems*, Math. Comp., 42 (1984), pp. 567–588.
[35] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.
[36] E. B. Saff and V. Totik, *Logarithmic Potentials with External Fields*, Springer-Verlag, Berlin, 1997.
[37] V. Simoncini, *Ritz and pseudo-Ritz values using matrix polynomials*, Linear Algebra Appl., 241/243 (1996), pp. 787–801.
[38] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[39] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[40] K.-C. Toh and L. N. Trefethen, *Calculation of pseudospectra by the Arnoldi iteration*, SIAM J. Sci. Comput., 17 (1996), pp. 1–15.
[41] K.-C. Toh and L. N. Trefethen, *The Kreiss matrix theorem on a general complex domain*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 145–165.
[42] L. N. Trefethen, *Approximation theory and numerical linear algebra*, in Algorithms for Approximation II, J. C. Mason and M. G. Cox, eds., Chapman and Hall, London, 1990.
[43] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, Essex, UK, 1992, pp. 234–266.

[44] L. N. Trefethen, *Computation of pseudospectra*, in Acta Numerica 8, Cambridge University Press, Cambridge, UK, 1999, pp. 247–295.
[45] L. N. Trefethen, *Spectra and pseudospectra: The behavior of non-normal matrices and operators*, in The Graduate Student's Guide to Numerical Analysis '98, M. Ainsworth, J. Levesley, and M. Marletta, eds., Springer-Verlag, Berlin, 1999, pp. 217–250.
[46] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.
[47] A. van der Sluis and H. A. van der Vorst, *The convergence behavior of Ritz values in the presence of close eigenvalues*, Linear Algebra Appl., 88/89 (1987), pp. 651–694.
[48] H. A. van der Vorst and C. Vuik, *The superlinear convergence behaviour of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.
[49] T. G. Wright, *MATLAB Pseudospectra GUI*, 2000–2001, http://www.comlab.ox.ac.uk/pseudospectra/psagui.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1110–1123

ON THE SPECTRA OF CERTAIN MATRICES GENERATED BY INVOLUTORY AUTOMORPHISMS*

ARIEH ISERLES† AND ANTONELLA ZANNA‡

Abstract. Let $A = P + K$ be an $n \times n$ complex matrix with $P = \frac{1}{2}(A - HAH)$ and $K = \frac{1}{2}(A + HAH)$, H being a unitary involution. Having characterized all unitary involutions, we investigate the spectral structure of P and K and, in particular, characterize the eigenvalues of K as zeros of a rational function, and prove that, for normal A, $\sigma(K)$ resides in the convex hull of $\sigma(A)$. We also demonstrate that this need not be true when A is not normal.

Key words. inner automorphism, involution, Lie algebra, Lie triple system

AMS subject classification. 15A18

DOI. 10.1137/S0895479803423780

1. Introduction.

The goal of this paper is to explore a number of issues in matrix analysis which have arisen in research at the interface of geometric numerical integration and computational linear algebra. Although the results of this paper stand alone and they do not require any elaboration of these issues, the latter are important in motivating and setting the backdrop for our work, hence we commence by reviewing them briefly.

Let $\mathfrak{g}$ be a matrix Lie algebra. The approximation of $\exp A$, where $A \in \mathfrak{g}$, is a central step in most numerical methods for the solution of differential equations evolving in Lie groups [5]. The purpose of such "Lie-group solvers" is to propagate the solution within the Lie group $\mathcal{G}$, say, whose Lie algebra is $\mathfrak{g}$. Therefore, it is of critical importance that the approximate exponential resides in $\mathcal{G}$ whenever the argument lives in $\mathfrak{g}$. Unfortunately, many standard techniques to approximate the exponential fail to respect the structure for some Lie groups. In particular, all such methods fail to map an arbitrary element of $\mathfrak{sl}(n)$ to $\mathrm{SL}(n)$ for $n \geq 3$ [7]. This has motivated a new breed of methods, designed to respect Lie-group structure: [1, 2] and, in particular, [8] and [6]. The latter two publications are based on a factorization of $\exp A$ for $A \in \mathfrak{g}$ using *generalized polar decomposition (GPD).* Thus, let κ be a Lie-algebra automorphism, hence a linear operator such that $\kappa([A, B]) = [\kappa(A), \kappa(B)]$, $A, B \in \mathfrak{g}$. In addition, we assume that it is an involution, i.e., $\kappa(\kappa(A)) = A$ for all $A \in \mathfrak{g}$. In that case it is possible, as originally proposed in [8], to represent the Lie algebra $\mathfrak{g}$ as a direct sum of two linear spaces,

$$\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k},$$

where

$$\mathfrak{p} = \{X \in \mathfrak{g} \,:\, \kappa(X) = -X\}$$

*Received by the editors March 4, 2003; published electronically (in revised form) by M. Chu October 24, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/simax/25-4/42378.html

†Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK (A.Iserles@damtp.cam.ac.uk).

‡Department of Informatics, University of Bergen, Høyteknologisenteret, N-5020 Bergen, Norway (Antonella.Zanna@ii.uib.no).

is a Lie triple system (i.e., $[X,[Y,Z]] \in \mathfrak{p}$ for all $X,Y,Z \in \mathfrak{p}$), while

$$\mathfrak{k} = \{X \in \mathfrak{g} : \kappa(X) = X\}$$

is a subalgebra of $\mathfrak{g}$. Specifically, we let $P = \frac{1}{2}[A - \kappa(A)] \in \mathfrak{p}$, $K = \frac{1}{2}[A + \kappa(A)] \in \mathfrak{k}$. It is possible to prove that there exist functions $X(t)$ and $Y(t)$, $t \geq 0$, evolving in $\mathfrak{p}$ and $\mathfrak{k}$, respectively, such that

$$\exp(tA) = \exp(X(t)) \exp(Y(t)).$$

The functions X and Y can be evaluated as a linear combination of commutators in the free algebra generated by $\{P, K\}$,

$$\begin{aligned} X(t) &= tP - \tfrac{1}{2}t^2[P,K] - \tfrac{1}{6}t^3[K,[P,K]] + \tfrac{1}{24}t^4([P,[P,[P,K]]] \\ &\qquad - [K,[K,[P,K]]]) + \mathcal{O}\big(t^5\big), \\ Y(t) &= tK - \tfrac{1}{12}t^3[P,[P,K]] + \mathcal{O}\big(t^5\big). \end{aligned}$$

In a practical algorithm, the series above are truncated and the calculation is accompanied by a number of linear-algebraic techniques which, while being of no direct relevance to the theme of the present paper, are critically important in reducing computational complexity [6].

Suppose that, in addition, the dimension of $\mathfrak{p}$ is small, so that it is easy to compute $\exp(X(t))$ exactly, while $\dim \mathfrak{k} < \dim \mathfrak{g}$. The idea is to iterate this procedure, representing $\mathfrak{k}$ as a direct sum of a Lie triple system and a subalgebra using another involutory automorphism and so on. The ultimate outcome is a representation

$$\exp(tA) = \exp(X_1(t)) \exp(X_2(t)) \cdots \exp(X_m(t)), \tag{1.1}$$

where each $\exp(X_k(t))$ can be evaluated with relative ease.

Suppose that A is a "nice" matrix: the real parts of its eigenvalues are relatively small. Even cursory examination of the numerical stability of the representation (1.1) (or its approximation) demonstrates that it is important for the matrix

$$K = K(A) = \tfrac{1}{2}[A + \kappa(A)] \tag{1.2}$$

to share this "niceness"; otherwise we might be generating very large matrices in the course of computation, thereby losing accuracy in finite arithmetic. This brings us to the central theme of this paper, the connection between the spectra of A and those of P and K.

In section 2 we characterize all involutory inner automorphisms in $\mathrm{U}(n)$. Let $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r \in \mathbb{C}^n$, where $r \leq n$, be vectors of unit length, orthogonal to each other. Then

$$\kappa(A) = HAH, \qquad H = I - 2\sum_{k=1}^{r} \boldsymbol{u}_k \boldsymbol{u}_k^*, \tag{1.3}$$

is a unitary involutory automorphism. Moreover, every involutory, unitary inner automorphism of $\mathrm{M}_n[\mathbb{C}]$ can be represented in the form (1.3) for some orthogonal vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r$. Thus, such automorphisms are a natural generalization of a similarity transformation by the familiar Householder reflection. Furthermore, we show in section 2 an interesting correspondence between the free group generated by

the involutory matrices $I - 2\boldsymbol{u}_k\boldsymbol{u}_k^*$, $k = 1, 2, \ldots, n$, and the additive group $\mathbb{Z}_n^2$ of binary n-tuples, with addition defined *modulo* 2.

Section 3 is devoted to our main result. We prove that, for a *normal* matrix A, the eigenvalues of K reside in the convex hull of the eigenvalues of A. This confirms that if A is "nice," so is K: in particular, if A is a stable matrix, this feature is shared by K.

Finally, in section 4 we demonstrate that normalcy of A is crucial. Once we consider general matrices $A \in \mathrm{M}_n[\mathbb{C}]$, the result of section 3 is no longer valid and the eigenvalues of K might well reside outside the convex hull of $\sigma(A)$.

2. Involutory automorphisms in U(n). An *automorphism* in the general linear algebra $\mathrm{M}_n[\mathbb{C}]$ is a linear function $\kappa : \mathrm{M}_n[\mathbb{C}] \to \mathrm{M}_n[\mathbb{C}]$ such that $\kappa([A, B]) = [\kappa(A), \kappa(B)]$ for all $A, B \in \mathrm{M}_n[\mathbb{C}]$. κ is an *inner* automorphism if it is of the form $\kappa(A) = HAH^{-1}$ for some nonsingular matrix $H \in \mathrm{M}_n[\mathbb{C}]$ and an *involution* if $\kappa(\kappa(A)) = A$ for all $A \in \mathrm{M}_n[\mathbb{C}]$. Except for the most ubiquitous example of an automorphism, $\kappa(A) = A^{-\top}$, virtually all other automorphisms of interest are inner. This applies in particular to the automorphisms in [6, 8], which, in addition, need be involutory, so that GPD exists.

Note that an involutory inner automorphism is of the form $\kappa(A) = HAH$, where H is itself an involutory matrix, $H^2 = I$. The computation of $\kappa(A)$ thus involves two products with the matrix H and applications to geometric integration at the root of the current work necessitate repeated computation of an automorphism in each time step. For reasons of numerical stability we thus restrict the discussion to unitary matrices H, since they enjoy the best-possible conditioning, and this is our practice in what follows. Therefore, the main focus of this section is on the unitary involutory inner automorphism

$$\kappa(A) = HAH, \quad A \in \mathrm{M}_n[\mathbb{C}], \quad \text{where} \quad H \in \mathrm{U}(n), \ \ H^2 = I. \tag{2.1}$$

THEOREM 2.1. *Let* $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_s\}$ *be an orthonormal basis of a subspace of* $\mathbb{C}^n$, *where* $s \in \{0, 1, \ldots, n\}$. *Then*

$$H = I - 2\sum_{k=1}^{s} \boldsymbol{u}_k\boldsymbol{u}_k^* \tag{2.2}$$

is an involution in $\mathrm{U}(n)$. *Moreover, every involution in* $\mathrm{U}(n)$ *can be represented in the form* (2.2).

Proof. It is straightforward to prove that any H defined by (2.2) is unitary and Hermitian, hence a unitary involution. This proves the first, trivial statement of the theorem.

Next, let us assume that $H \in \mathrm{U}(n)$ is an involution, therefore H is Hermitian. This implies that the matrix $I - H$ is also Hermitian. Suppose that the eigenvalues and normalized eigenvectors of $I - H$ are $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$ and $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n \in \mathbb{C}^n$, respectively. It follows that

$$I - H = \sum_{k=1}^{n} \alpha_k \boldsymbol{u}_k\boldsymbol{u}_k^*.$$

Exploiting orthogonality of eigenvectors, we deduce that

$$H^2 = I - 2\sum_{k=1}^{n} \alpha_k \boldsymbol{u}_k\boldsymbol{u}_k^* + \sum_{k=1}^{n}\sum_{l=1}^{n} \alpha_k\alpha_l(\boldsymbol{u}_k^*\boldsymbol{u}_l)\boldsymbol{u}_k\boldsymbol{u}_l^* = I - \sum_{k=1}^{n}(\alpha_k - 2)\alpha_k \boldsymbol{u}_k\boldsymbol{u}_k^*.$$

Since $H^2 = I$ and the matrices $\boldsymbol{u}_k\boldsymbol{u}_k^*$ are linearly independent, it follows that $\alpha_k \in \{0, 2\}$. Without loss of generality, we may assume that $\alpha_1 = \cdots = \alpha_s = 2$, $\alpha_{s+1} = \cdots = \alpha_n = 0$, and this results in the representation (2.2) and completes the proof of the theorem. □

Note that, once we have characterized all unitary involutions in $\mathrm{U}(n)$, we immediately obtain from (2.1) a characterization of all unitary involutory inner automorphisms of $\mathrm{M}_n[\mathbb{C}]$.

Let $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ be a unitary basis of $\mathbb{C}^n$, $\|\boldsymbol{u}_k\| = 1$. Given a vector $\boldsymbol{\theta} \in \mathbb{Z}_2^n$, we set

$$G_{\boldsymbol{\theta}} = I - 2\sum_{k=1}^{n} \theta_k \boldsymbol{u}_k\boldsymbol{u}_k^*.$$

Note that, by the last theorem, $G_{\boldsymbol{\theta}}$ is a unitary involution. It is trivial to verify that

$$G_{\boldsymbol{\theta}}G_{\boldsymbol{\varphi}} = G_{\boldsymbol{\theta}+\boldsymbol{\varphi} \bmod 2}.$$

Therefore,

$$\boldsymbol{G} = \{G_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in \mathbb{Z}_2^n\}$$

is an Abelian multiplicative group, isomorphic to $\mathbb{Z}_2^n$. Another easy observation is that the space of all unitary involutions (2.2), for all $r = 0, 1, \ldots, n$, is a free group generated by $G_{\boldsymbol{e}_k}$, $k = 1, 2, \ldots, n$. Needless to say, $r = 0$ yields the identity matrix and it is easy to prove that $G_{1,1,\ldots,1} = -I$. The latter is a consequence of the well-known identity

$$\sum_{k=1}^{n} \boldsymbol{u}_k\boldsymbol{u}_k^* = I,$$

known in quantum chemistry as *resolution of identity.*

3. The spectrum of a normal matrix $\boldsymbol{K}$. The purpose of this section is to study the eigenvalues $\mu_1, \mu_2, \ldots, \mu_n$ and eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ of the matrix

$$K = \tfrac{1}{2}[A + \kappa(A)], \tag{3.1}$$

where $A \in \mathrm{M}_n[\mathbb{C}]$ and κ is a unitary involutory inner automorphism, $\kappa(A) = HAH$.

Since $H \in \mathrm{U}(n)$, it has a full set of unitary eigenvectors, therefore $H = QDQ^*$, where $Q \in \mathrm{U}(n)$ and D is diagonal. Moreover, since $H^2 = I$, necessarily, without loss of generality,

$$D = \begin{bmatrix} I_{r\times r} & O_{r\times s} \\ O_{s\times r} & -I_{s\times s} \end{bmatrix},$$

where $r + s = n$. Letting $\tilde{A} = Q^*AQ$, $\tilde{K} = Q^*KQ$, we deduce at once from (3.1) that

$$\tilde{K} = \tfrac{1}{2}(\tilde{A} + D\tilde{A}D) = \begin{bmatrix} \tilde{A}_{1,1} & O \\ O & \tilde{A}_{2,2} \end{bmatrix}, \qquad \text{where} \qquad \tilde{A} = \begin{bmatrix} \tilde{A}_{1,1} & \tilde{A}_{1,2} \\ \tilde{A}_{2,1} & \tilde{A}_{2,2} \end{bmatrix}$$

and $\tilde{A}_{1,1} \in \mathrm{M}_r[\mathbb{C}]$, $\tilde{A}_{2,2} \in \mathrm{M}_s[\mathbb{C}]$. We thus conclude that

$$\sigma(K) = \sigma(\tilde{K}) = \sigma(\tilde{A}_{1,1}) \cup \sigma(\tilde{A}_{2,2}). \tag{3.2}$$

Moreover, if $\tilde{K}\tilde{\boldsymbol{v}} = \mu\tilde{\boldsymbol{v}}$, then

$$\tilde{\boldsymbol{v}} = \begin{bmatrix} \tilde{\boldsymbol{v}}_1 \\ \tilde{\boldsymbol{v}}_2 \end{bmatrix}, \qquad \tilde{\boldsymbol{v}}_1 \in \mathbb{C}^r, \ \ \tilde{\boldsymbol{v}}_2 \in \mathbb{C}^s$$

and either

$$\tilde{A}_{1,1}\tilde{\boldsymbol{v}}_1 = \mu\tilde{\boldsymbol{v}}_1, \qquad \tilde{\boldsymbol{v}}_2 = \boldsymbol{0}$$

or

$$\tilde{\boldsymbol{v}}_1 = \boldsymbol{0}, \qquad \tilde{A}_{2,2}\tilde{\boldsymbol{v}}_2 = \mu\tilde{\boldsymbol{v}}_2.$$

PROPOSITION 3.1. *Let $\boldsymbol{v}$ be an eigenvector of K. Then it is also an eigenvector of H.*

Proof. Let $\tilde{\boldsymbol{v}} = Q^*\boldsymbol{v}$, $\boldsymbol{w} = H\boldsymbol{v}$ and $\tilde{\boldsymbol{w}} = Q^*\boldsymbol{w}$. Note that

$$K\boldsymbol{v} = \mu\boldsymbol{v} \qquad \Rightarrow \qquad (A + HAH)\boldsymbol{v} = 2\mu\boldsymbol{v} \qquad \Rightarrow \qquad (HA + AH)\boldsymbol{v} = 2\mu\boldsymbol{w};$$

therefore $K\boldsymbol{w} = \mu\boldsymbol{w}$. Thus, $\boldsymbol{w} \neq \boldsymbol{0}$ is also an eigenvector of K corresponding to the same eigenvalue μ.

We have

$$\tilde{\boldsymbol{w}} = Q^*(QDQ^*)Q\tilde{\boldsymbol{v}} = D\tilde{\boldsymbol{v}} = \begin{bmatrix} \tilde{\boldsymbol{v}}_1 \\ -\tilde{\boldsymbol{v}}_2 \end{bmatrix};$$

hence

$$\boldsymbol{w} = Q\begin{bmatrix} \tilde{\boldsymbol{v}}_1 \\ -\tilde{\boldsymbol{v}}_2 \end{bmatrix}.$$

Recall that either $\tilde{\boldsymbol{v}}_1$ or $\tilde{\boldsymbol{v}}_2$ is a zero vector. If $\tilde{\boldsymbol{v}}_2 = \boldsymbol{0}$, then

$$H\boldsymbol{v} = \boldsymbol{w} = Q\begin{bmatrix} \tilde{\boldsymbol{v}}_1 \\ \tilde{\boldsymbol{v}}_2 \end{bmatrix} = Q\tilde{\boldsymbol{v}} = \boldsymbol{v},$$

while, by the same token, if $\tilde{\boldsymbol{v}}_1 = \boldsymbol{0}$, then

$$H\boldsymbol{v} = -\boldsymbol{v}.$$

This concludes the proof. ∎

An alternative formulation of Proposition 3.1 is that the eigenvectors of K reside in one of the linear spaces

$$\mathcal{K}_n = \{\boldsymbol{x} \in \mathbb{C}^n \ : \ H\boldsymbol{x} = \boldsymbol{x}\} \qquad \text{or} \qquad \mathcal{P}_n = \{\boldsymbol{x} \in \mathbb{C}^n \ : \ H\boldsymbol{x} = -\boldsymbol{x}\}.$$

Without loss of generality we assume that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathcal{K}_n$ and $\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_n \in \mathcal{P}_n$. Of course, $\mathcal{K}_n \oplus \mathcal{P}_n = \mathbb{C}^n$.

Recalling the characterization (2.2) of unitary involutions, we denote

$$\mathcal{U} = \text{Span}\,\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_s\}$$

(the double use of the integer variable s is, as will be apparent soon, not a variable overload). Since

$$H\boldsymbol{v} = \boldsymbol{v} - 2\sum_{k=1}^{s}(\boldsymbol{u}_k^*\boldsymbol{v})\boldsymbol{u}_k,$$

we deduce that

$$\begin{aligned} \boldsymbol{v} \in \mathcal{P}_n \quad &\Rightarrow \quad \boldsymbol{v} = \sum_{k=1}^{s} (\boldsymbol{u}_k^* \boldsymbol{v}) \boldsymbol{u}_k \quad &\Rightarrow \quad \mathcal{P}_n = \mathcal{U}, \\ \boldsymbol{v} \in \mathcal{K} \quad &\Rightarrow \quad \boldsymbol{u}_k^* \boldsymbol{v} = 0, \;\; k = 1, 2, \ldots, s \quad &\Rightarrow \quad \mathcal{K}_n = \mathcal{U}^\perp. \end{aligned}$$

So far we have allowed an arbitrary $A \in \mathrm{M}_n[\mathbb{C}]$. In the remainder of this section we stipulate that A is *normal*, hence $A^*A = AA^*$ and its normalized eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ form a unitary basis of $\mathbb{C}^n$. We denote the eigenvalues of A by $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively.

THEOREM 3.2. *Let $A \in \mathrm{M}_n[\mathbb{C}]$ be a normal matrix, $H \in \mathrm{U}(n)$ be an involution and $K = \frac{1}{2}(A + HAH)$. Then*

$$\sigma(K) \subset \operatorname{conv} \sigma(A), \tag{3.3}$$

where $\sigma(B)$ is the spectrum of the matrix B, while $\operatorname{conv} \Omega$ *is the (closed) convex hull of $\Omega \in \mathbb{C}$.*

Proof. Let $\sigma(K) = \{\mu_1, \mu_2, \ldots, \mu_n\}$ and let $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ be the corresponding normalized eigenvectors. For clarity, whenever possible we dispense with a subscript and let $K\boldsymbol{v} = \mu\boldsymbol{v}$. Recall that either $\boldsymbol{v} \in \mathcal{K}_n$ or $\boldsymbol{v} = \mathcal{P}_n$.

If $\boldsymbol{v} \in \mathcal{K}_n$, then $H\boldsymbol{v} = \boldsymbol{v}$; hence

$$\mu\boldsymbol{v} = K\boldsymbol{v} = \tfrac{1}{2}(A\boldsymbol{v} + HA\boldsymbol{v}),$$

and therefore

$$\mu = \tfrac{1}{2}\boldsymbol{v}^*(A\boldsymbol{v} + HA\boldsymbol{v}) = \boldsymbol{v}^* A\boldsymbol{v}.$$

The conclusion is true also for $\boldsymbol{v} \in \mathcal{P}_n$, since the sign of the second term changes twice. Therefore

$$\boldsymbol{v}^* A\boldsymbol{v} = \mu. \tag{3.4}$$

Since A is normal, it is true that

$$A = \sum_{k=1}^{n} \lambda_k \boldsymbol{w}_k \boldsymbol{w}_k^*.$$

Therefore

$$\mu = \boldsymbol{v}^* A\boldsymbol{v} = \sum_{k=1}^{n} \lambda_k |\boldsymbol{v}^* \boldsymbol{w}_k|^2$$

and we deduce that μ is a convex linear combination of $\lambda_1, \lambda_2, \ldots, \lambda_n$. Consequently, $\mu \in \operatorname{conv} \sigma(A)$ and (3.3) is valid. ∎

The implications of Theorem 3.2 are clear: if A is a stable matrix, then so is K and the ℓ_2 logarithmic norm of K, an appropriate measure of its stability, cannot exceed that of A. Figure 3.1 displays the eigenvalues (and their convex hull) of a 100×100 normal matrix A and of the corresponding matrix K, with $s = 5$. Both A and H have been randomly generated, using matrices from Gaussian unitary ensemble. Evidently, the eigenvalues of K are consistent with Theorem 3.2. Moreover, they appear to be a moderate perturbation of the eigenvalues of A. We believe that this is, in general, the

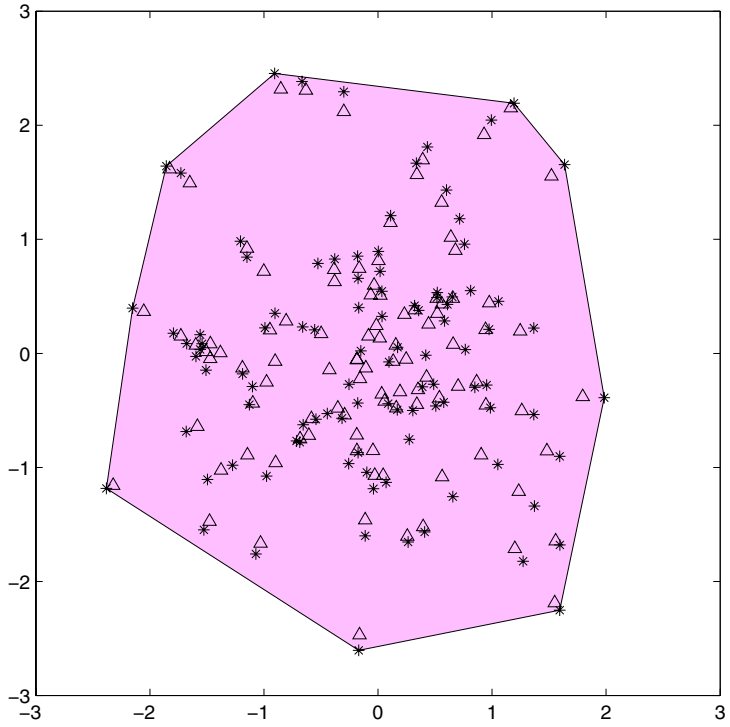


FIG. 3.1. *The eigenvalues of A (denoted by asterisks) and of K (denoted by triangles) for $n = 100$ and $s = 5$.*

case when either s or $n - s$ is significantly smaller than n. Thus, Figure 3.2 depicts the eigenvalues of the same 100×100 matrix A but different K, generated randomly using $s = 50$. Another structural detail is apparent, it recurs in other computational experiments and is further illustrated in Figure 3.3: the eigenvalues of K "shrink" toward the center of the convex hull. It is unclear by this stage whether this behavior is a stochastic artifact or whether $\sigma(K)$ can be always confined to significantly smaller geometric structure inside $\mathrm{conv}\sigma(A)$.

A measure of support for our observation that, for small s, $\sigma(K)$ is, in general, a moderate perturbation of $\sigma(A)$ is provided by the following result.

LEMMA 3.3. *Assume that A is a normal matrix. Then,*

$$\|A - K\|_{\mathsf{F}} \leq \frac{\sqrt{2s}}{2} \mathrm{diam}\, \mathrm{conv}\sigma(A), \tag{3.5}$$

where $\mathrm{diam}\, \Omega$ *is the diameter of the set* $\Omega \subset \mathbb{C}$.

Proof. For simplicity sake, let us denote

$$H = I - 2UU^*,$$

where $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_s]$.

We commence by noting that $A - K = P$, where $P = \frac{1}{2}(A - HAH)$ has been defined in section 1. Moreover,

$$P = UU^*A + AUU^* - 2UU^*AUU^* = U[A^*U - UU^*A^*U]^* + [AU - UU^*AU]U^*;$$

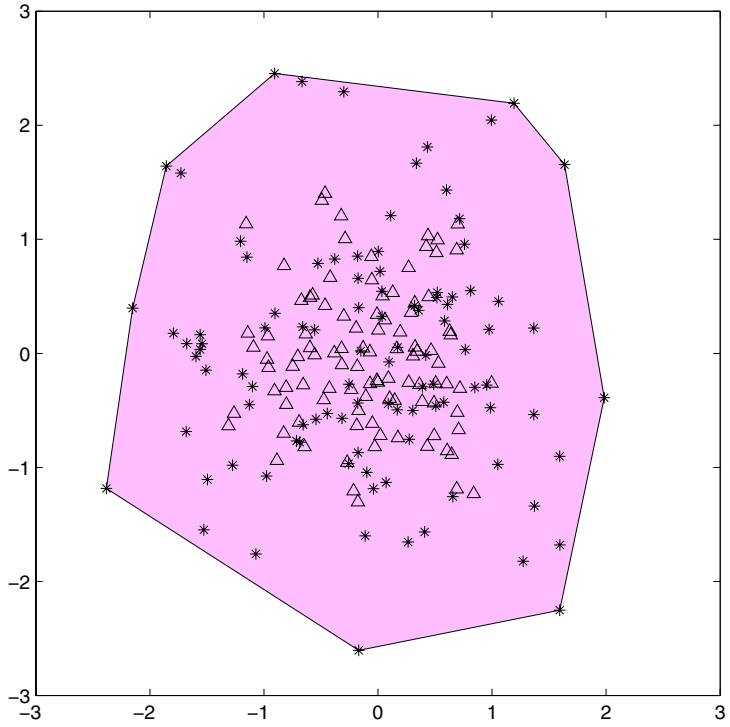


FIG. 3.2. *The eigenvalues of A (denoted by asterisks) and of K (denoted by triangles) for $n = 100$ and $s = 50$.*

therefore P is a rank-$2s$ matrix with (at most) $2s$ nonzero eigenvalues, $\kappa_1, \kappa_2, \ldots, \kappa_{2s}$. We deduce that

$$\|A - K\|_{\mathsf{F}}^2 = \sum_{i=1}^{2s} |\kappa_i|^2.$$

It is easy to verify that, if κ_i is an eigenvalue of P corresponding to the eigenvector $\boldsymbol{y}_i$, then

$$P(H\boldsymbol{y}_i) = -HP\boldsymbol{y}_i = -H\kappa_i\boldsymbol{y}_i = -\kappa_i(H\boldsymbol{y}_i);$$

that is, $-\kappa_i$ is also an eigenvalue of P corresponding to the eigenvector $H\boldsymbol{y}_i$. Hence, assuming that the eigenvalues of P are labelled so that $\kappa_{i+s} = -\kappa_i$, we obtain

$$\|A - K\|_{\mathsf{F}}^2 = 2\sum_{i=1}^{s} |\kappa_i|^2 \leq 2s \max_{i=1,\ldots,s} |\kappa_i|^2 = 2s\,[\rho(P)]^2,$$

where $\rho(\,\cdot\,)$ denotes the spectral radius. Since the spectrum of P is symmetric with respect to the origin, we deduce that $\rho(P) = \max_{i=1,\ldots,s} |\kappa_i| = \frac{1}{2}$ diam conv$\sigma(P)$.

Recall that, if $B, C \in \mathrm{M}_n[\mathbb{C}]$ are normal, with $\sigma(B) = \{\beta_i : i = 1, \ldots, n\}$ and $\sigma(C) = \{\gamma_i : i = 1, \ldots, n\}$, then

$$\sigma(B + C) \subseteq \mathrm{conv}\{(\beta_i + \gamma_j) : i, j = 1, \ldots, n\}$$

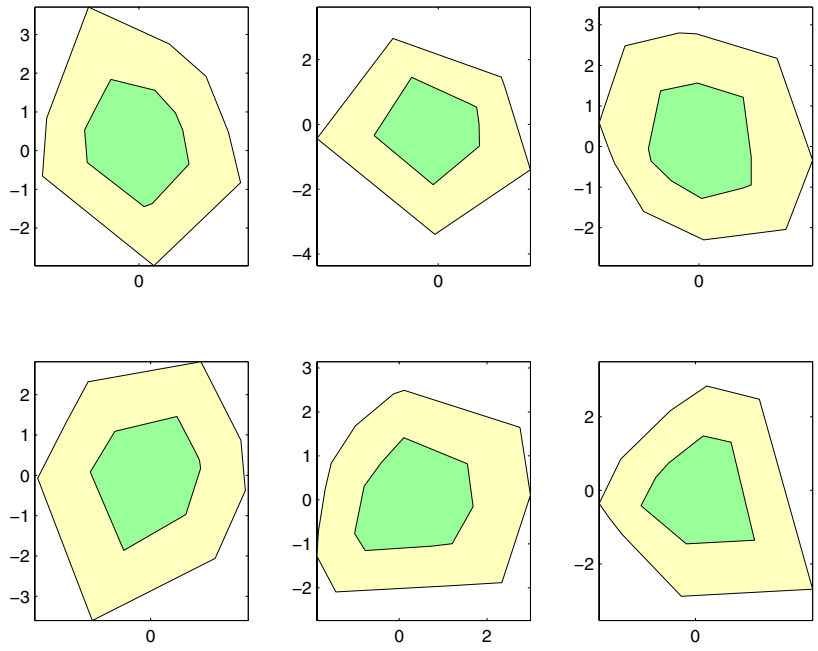


FIG. 3.3. *The convex hulls of $\sigma(A)$ and $\sigma(K)$ for six random matrices (A being normal) with $n = 200$ and $s = 100$.*

[4]. Since A and HAH are both normal and, HAH being similar to A, share the same eigenvalues λ_i, $i = 1, \ldots, n$, we have

$$\sigma(P) = \sigma[\tfrac{1}{2}(A - HAH)] \subseteq \tfrac{1}{2}\mathrm{conv}\{(\lambda_i - \lambda_j) : i, j = 1, \ldots, n\}.$$

We observe that the set $\mathrm{conv}\{(\lambda_i - \lambda_j) : i, j = 1, \ldots, n\}$ has a point symmetry at the origin: if $(\lambda_i - \lambda_j)$ is a vertex of the convex hull, so is $-(\lambda_i - \lambda_j)$. Therefore,

$$\mathrm{diam}\ \mathrm{conv}\{(\lambda_i - \lambda_j) : i, j = 1, \ldots, n\} = 2 \max_{i,j=1,\ldots,n} |\lambda_i - \lambda_j| = 2\mathrm{diam}\ \mathrm{conv}\sigma(A).$$

Thus,

$$\rho(P) = \tfrac{1}{2}\mathrm{diam}\ \mathrm{conv}\sigma(P) \le \tfrac{1}{2}\mathrm{diam}\ \mathrm{conv}\sigma(A)$$

and we conclude that

$$\|A - K\|_{\mathsf{F}}^2 \le \frac{s}{2}\mathrm{diam}\ \mathrm{conv}\sigma(A),$$

from which (3.5) follows by taking square roots. ∎

It is worthwhile to mention that (3.5) is sharp in the case $s = 1$. Letting $H = I - 2\boldsymbol{u}\boldsymbol{u}^*$, a simple calculation reveals that P has rank two and that its eigenvalues are

$$\kappa_{1,2} = \pm\sqrt{\boldsymbol{u}^* A^2 \boldsymbol{u} - (\boldsymbol{u}^* A \boldsymbol{u})^2}.$$

Hence,

$$\|A - K\|_{\mathsf{F}}^2 = \|P\|_{\mathsf{F}}^2 = |\kappa_1|^2 + |\kappa_2|^2 = 2\rho(P)^2.$$

Assume that the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A and the corresponding normalized eigenvectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are ordered so that

$$|\lambda_1 - \lambda_n| = \max_{k,l=1,\ldots,n} |\lambda_k - \lambda_l|,$$

and take $\boldsymbol{u} = \frac{\sqrt{2}}{2}(\boldsymbol{x}_1 + \boldsymbol{x}_n)$. It is readily observed that $\boldsymbol{u}^* A^k \boldsymbol{u} = \frac{1}{2}(\lambda_1^k + \lambda_n^k)$, $k \geq 0$; hence

$$\rho(P) = \max\{|\kappa_1|, |\kappa_2|\} = \frac{1}{2}|\lambda_1 - \lambda_n| = \frac{1}{2}\text{diam conv}\sigma(A),$$

from which (3.5) follows as an equality.

The implication of (3.3) for $s = 1$ is that the eigenvalues of A and K can be ordered so that, *on average,*

$$|\lambda_k - \mu_k| = \mathcal{O}\left(n^{-1}\right), \qquad k = 1, 2, \ldots. \tag{3.6}$$

This can be extended to $s \geq 2$, as long as s is small in comparison with n, since P is always of rank $2s$.

Although (3.6) is only a probabilistic statement, not an absolute estimate, it goes some way toward explaining the phenomenon that we have observed in Figure 3.1.

Another interesting connection between $\sigma(A)$ and $\sigma(K)$ is highlighted in our next result.

LEMMA 3.4. *Let $\boldsymbol{v} \in \mathcal{K}_n$ be an eigenvector of K with eigenvalue μ. Then either $\mu \in \sigma(A)$ or it is a zero of the rational function*

$$\psi(x) = \sum_{k=1}^{n} \frac{|\zeta_k|^2}{\lambda_k - x}, \tag{3.7}$$

where

$$\boldsymbol{z} = \sum_{k=1}^{s} \beta_k \boldsymbol{u}_k = \sum_{l=1}^{n} \zeta_l \boldsymbol{x}_l,$$

$\beta_k = \boldsymbol{u}_k^* A\boldsymbol{v}$, $k = 1, 2, \ldots, s$, and $\boldsymbol{x}_l$, $l = 1, 2, \ldots, n$ are the eigenvectors of A.

Proof. Suppose first that $A\boldsymbol{v} \in \mathcal{K}_n$. Then $H\boldsymbol{v} = \boldsymbol{v}$, $HA\boldsymbol{v} = A\boldsymbol{v}$, $K\boldsymbol{v} = \mu\boldsymbol{v}$, and (3.1) imply that $A\boldsymbol{v} = \mu\boldsymbol{v}$, hence $\mu \in \sigma(A)$. Let us turn our attention to the other case, namely $A\boldsymbol{v} \cap \mathcal{P}_n \neq \{\boldsymbol{0}\}$. Since $\mathcal{K}_n = \mathcal{U}^\perp$, we choose an arbitrary unitary basis of $\mathcal{U}^\perp$, namely $\{\boldsymbol{u}_{s+1}, \boldsymbol{u}_{s+2}, \ldots, \boldsymbol{u}_n\}$. Thus,

$$\boldsymbol{v} = \sum_{k=s+1}^{n} \alpha_k \boldsymbol{u}_k, \qquad \alpha_k = \boldsymbol{u}_k^* \boldsymbol{v}, \quad k = s+1, s+2, \ldots, n, \quad \|\boldsymbol{\alpha}\| = 1. \tag{3.8}$$

It follows at once from the definition of K that

$$\mu\boldsymbol{v} = K\boldsymbol{v} = \tfrac{1}{2}(A + HAH)\boldsymbol{v} = \tfrac{1}{2}(A\boldsymbol{v} + HA\boldsymbol{v}) = A\boldsymbol{v} - \sum_{l=1}^{s} (\boldsymbol{u}_l^* A\boldsymbol{v})\boldsymbol{u}_l.$$

Therefore

$$\boldsymbol{v} = (A - \mu I)^{-1} \sum_{l=1}^{s} (\boldsymbol{u}_l^* A\boldsymbol{v})\boldsymbol{u}_l.$$

Comparison with (3.8) yields

$$\sum_{l=s+1}^{n} \alpha_l \boldsymbol{u}_l = (A-\mu I)^{-1} \sum_{l=1}^{s} \beta_l \boldsymbol{u}_l,$$

where $\beta_l = \boldsymbol{u}_l^* A \boldsymbol{v}$. Multiplying with $\boldsymbol{u}_k^*$, $k = 1, 2, \ldots, s$, on the left, $\boldsymbol{v} \in \mathcal{U}^\perp$ implies that

$$0 = \boldsymbol{u}_k^* \boldsymbol{v} = \sum_{l=1}^{s} \beta_l \boldsymbol{u}_k^* (A-\mu I)^{-1} \boldsymbol{u}_l, \qquad k = 1, 2, \ldots, s.$$

Letting $\Phi_{k,l} = \boldsymbol{u}_k^* (A-\mu I)^{-1} \boldsymbol{u}_l$, $k, l = 1, 2, \ldots, s$, we thus deduce $\Phi \boldsymbol{\beta} = \boldsymbol{0}$, consequently $\boldsymbol{\beta}^* \Phi \boldsymbol{\beta} = 0$. Written in longhand, this is equivalent to

$$\sum_{k=1}^{s} \sum_{l=1}^{s} \bar{\beta}_k \beta_l \boldsymbol{u}_k^* (A-\mu I)^{-1} \boldsymbol{u}_l = 0.$$

Therefore

$$\boldsymbol{z}^* (A-\mu I)^{-1} \boldsymbol{z} = 0, \qquad \text{where} \qquad \boldsymbol{z} = \sum_{k=1}^{s} \beta_k \boldsymbol{u}_k.$$

If $\boldsymbol{z} = \boldsymbol{0}$, then $A\boldsymbol{v} \in \mathcal{K}_n$, a possibility that we have already ruled out. Therefore $\boldsymbol{z} \neq \boldsymbol{0}$.

We expand $\boldsymbol{z}$ in the eigenvectors of A,

$$\boldsymbol{z} = \sum_{k=1}^{n} \zeta_k \boldsymbol{x}_k.$$

Therefore

$$0 = \boldsymbol{z}^* (A-\mu I)^{-1} \boldsymbol{z} = \sum_{k=1}^{n} \frac{|\zeta_k|^2}{\lambda_k - \mu},$$

and this proves (3.7). ∎

Lemma 3.4 comes into its own when A is Hermitian, when also, trivially, K is Hermitian, in which case we recover some known results in linear algebra, namely the *Weyl theorem* on eigenvalues of a sum of Hermitian matrices and the *interlacing eigenvalues theorem for bordered matrices* [3].

Let us assume that $s = 1$, thus, that H is a Householder reflection. Ordering the eigenvalues of A as $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and those of K as $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$, we already know from Theorem 3.2 that $\mu_1, \mu_2, \ldots, \mu_n \in [\lambda_1, \lambda_n]$. Assume further that the spectrum of A is distinct, $\sigma(A) \cap \sigma(K) = \emptyset$, and that all the ζ_ks in Lemma 3.4 are nonzero—in other words, that $\boldsymbol{u}_1$ is not orthogonal to an eigenvector of A.

The function ψ from (3.7), being a rational function of type $(n-1)/n$, has exactly $n-1$ real zeros for distinct λ_ks. Since $\dim \mathcal{K}_n = n-1$, it follows that *all* its zeros are also eigenvalues of K. It is a trivial observation, though, that ψ changes sign in every interval of the form $(\lambda_k, \lambda_{k+1})$, $k = 1, 2, \ldots, n-1$. Therefore, there must be at least one μ_l in each interval of this form and a trivial counting argument demonstrates that

there must be a single μ_l in each interval, except for one interval that encloses two μ_ls. In other words, there exists $p \in \{1, 2, \ldots, n-1\}$ such that

$$\begin{aligned} &\mu_k \in (\lambda_k, \lambda_{k+1}), \qquad k = 1, 2, \ldots, p, \\ &\mu_k \in (\lambda_{k-1}, \lambda_k), \qquad k = p+1, p+2, \ldots, n. \end{aligned} \tag{3.9}$$

This "almost interlace" property is illustrated in Figure 3.4.

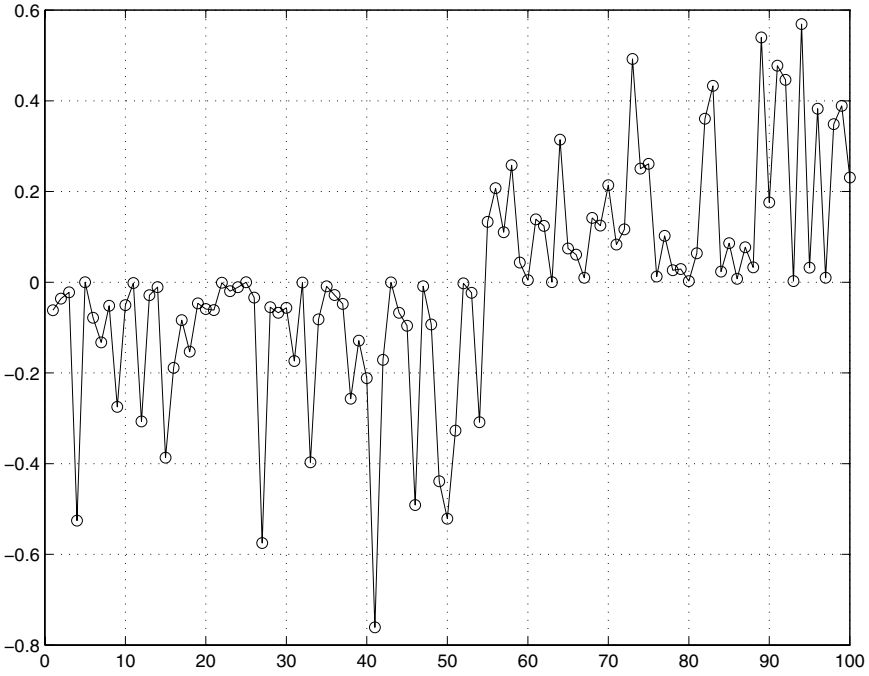


FIG. 3.4. *The sequence* $\boldsymbol{\lambda} - \boldsymbol{\mu}$ *for a random* 100×100 *Hermitian matrix* A.

The restrictive conditions (distinct eigenvalues, $\sigma(A) \cap \sigma(K) = \emptyset$ and $\boldsymbol{u}_1^\top \boldsymbol{x}_k \neq 0$) can be all removed by a limiting argument, except that open intervals in (3.9) need be replaced by closed intervals.

The situation is more complicated for $s \geq 2$. Thus, for example, for $s = 2$ similar argument implies that there must be a μ_l, corresponding to a $\boldsymbol{v} \in \mathcal{K}_n$, in exactly $n-2$ intervals of the form $(\lambda_k, \lambda_{k+1})$. Moreover, eigenvectors in $\mathcal{P}_n$ contribute two extra μ_ks, which can reside anywhere in $[\lambda_1, \lambda_n]$.

4. The nonnormal case. Since $K\boldsymbol{v} = \mu\boldsymbol{v}$, $\|\boldsymbol{v}\| = 1$, implies in the proof of Theorem 3.2 that $\mu = \boldsymbol{v}^* A \boldsymbol{v}$, we deduce that $\mu \in \mathrm{W}(A)$, where $\mathrm{W}(B)$ is the *numerical range* or *field of values* of a matrix $B \in \mathrm{M}_n[\mathbb{C}]$ [4], and that $\sigma(K)$ lies in a closed ball of radius $\|A\|$. This, however, falls short of establishing a connection between $\sigma(K)$ and $\sigma(A)$ for a general (rather than normal) $A \in \mathrm{M}_n[\mathbb{C}]$. In this section we demonstrate that normalcy, although used just once in the proof of Theorem 3.2, is not an artifact of the method of proof; it is possible to find nonnormal matrices A for which $\sigma(K) \not\subset \mathrm{conv}\sigma(A)$.

Specifically, we let

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & 0 & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} = \sum_{k=1}^{n-1} \boldsymbol{e}_k \boldsymbol{e}_{k+1}^{\top} \in \mathrm{M}_n[\mathbb{R}]. \tag{4.1}$$

Our first observation is that $\|A\| = 1$, therefore $|\mu_k| \le 1$, $k = 1, 2, \ldots, n$. (We retain the notation from section 3.)

Let $H = I - 2\boldsymbol{u}\boldsymbol{u}^*$, $\|\boldsymbol{u}\| = 1$, therefore $s = 1$. We choose $\boldsymbol{v} \in \mathcal{K}_n$ and suppose first that $A\boldsymbol{v} \in \mathcal{K}_n$. As apparent from the proof of Theorem 3.2, this implies $A\boldsymbol{v} = \mu\boldsymbol{v}$, therefore $\mu = 0$ and $\boldsymbol{v} = \pm\boldsymbol{e}_1$. Since $H\boldsymbol{e}_1 = \boldsymbol{e}_1 - 2\bar{u}_1\boldsymbol{u}$, this takes place if and only if $u_1 = 0$. As soon as we rule this out, μ is necessarily a zero of the function

$$\psi(x) = \boldsymbol{u}^*(A - xI)^{-1}\boldsymbol{u},$$

which we have defined in the proof of Lemma 3.4: this is true regardless of A being normal. Nilpotency of A implies that

$$\psi(x) = -\sum_{k=0}^{n-1} (\boldsymbol{u}^* A^k \boldsymbol{u}) x^{-k-1}. \tag{4.2}$$

Thus, we seek zeros away from the origin of the function

$$\varphi(x) = -x^n\psi(x) = \sum_{k=0}^{n-1} (\boldsymbol{u}^* A^k \boldsymbol{u}) x^{n-k-1}.$$

To concentrate on a specific example, let us choose $\boldsymbol{u} = n^{-1/2}\boldsymbol{1}$. Therefore, $\boldsymbol{u}^{\top} A^k \boldsymbol{u} = (n-k)/n$ and

$$\varphi(x) = \frac{1}{n} \sum_{k=0}^{n-1} (k+1) x^k$$

yields

$$n(1-x)^2\varphi(x) = nx^{n+1} - (n+1)x^n + 1. \tag{4.3}$$

Let $r\mathrm{e}^{\mathrm{i}\theta}$ be a zero of φ, $r \in (0, 1]$. Then (4.3) implies that

$$r^{2n} = \frac{1}{|n+1-nr\mathrm{e}^{\mathrm{i}\theta}|^2} = \frac{1}{(n+1)^2 - 2n(n+1)r\cos\theta + n^2r^2} \ge \frac{1}{[(1+r)n+1]^2}.$$

Therefore,

$$r \ge \frac{1}{[(1+r)n+1]^{1/n}} \ge \frac{1}{(2n+1)^{1/n}}.$$

Since the one eigenvalue not covered by this analysis is $\boldsymbol{u}^* A\boldsymbol{u} = 1 - 1/n$, it follows that

$$\sigma(K) \subset \{z \in \mathbb{C} : (2n+1)^{-1/n} \le |z| \le 1 - 1/n\}.$$

As a matter of fact, it is possible to prove, with extra effort, that $\rho(K) = 1 - 1/n$. Yet, this is not necessary to the observation that $\sigma(K)$ extends well outside $\mathrm{conv}\sigma(A) = \{0\}$.

The case $\boldsymbol{u} = n^{-1/2}\mathbf{1}$ is generic in the following sense. Whenever $s = 1$, necessarily $\boldsymbol{u}^* A \boldsymbol{u} \in \sigma(K)$, since $\boldsymbol{u}$ spans the one-dimensional linear space $\mathcal{P}_n$. Therefore, unless $\boldsymbol{u}^* A \boldsymbol{u} = \sum_{k=1}^{n-1} \bar{u}_k u_{k+1} = 0$, it is true that $\sigma(K)$ contains points outside the origin and the inclusion (3.3) does not hold.

Acknowledgments. A number of colleagues were generous in their advice on the subject matter of this paper, and we are happy to acknowledge discussions with Brad Baxter (Birkbeck College, University of London), John Butcher (University of Auckland), and Ernst Hairer (Université de Genevé).

REFERENCES

[1] E. Celledoni and A. Iserles, *Approximating the exponential from a Lie algebra to a Lie group*, Math. Comp., 69 (2000), pp. 1457–1480.

[2] E. Celledoni and A. Iserles, *Methods for the approximation of the matrix exponential in a Lie-algebraic setting*, IMA J. Numer. Anal., 21 (2001), pp. 463–488.

[3] R. A. Horn and C. R. Horn, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[4] R. A. Horn and C. R. Horn, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[5] A. Iserles, H. Munthe-Kaas, S. P. Norsett, and A. Zanna, *Lie-group methods*, Acta Numer., 9 (2000), pp. 215–365.

[6] A. Iserles and A. Zanna, *Efficient Computation of the Matrix Exponential by Generalized Polar Decompositions*, Tech. Rep. NA2002/09, University of Cambridge, 2002.

[7] F. Kang and Z.-J. Shang, *Volume-preserving algorithms or source-free dynamical systems*, Numer. Math., 1 (1995), pp. 451–463.

[8] H. Munthe-Kaas, G. R. W. Quispel, and A. Zanna, *Generalized polar decompositions on Lie groups with involutive automorphisms*, Found. Comput. Math., 1 (2001), pp. 297–324.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1124–1139

CONVEX NONCOMMUTATIVE POLYNOMIALS HAVE DEGREE TWO OR LESS*

J. WILLIAM HELTON† AND SCOTT MCCULLOUGH‡

Abstract. A polynomial p (with real coefficients) in noncommutative variables is matrix convex provided

$$p(tX + (1-t)Y) \le tp(X) + (1-t)p(Y)$$

for all $0 \le t \le 1$ and for all tuples $X = (X_1, \dots, X_g)$ and $Y = (Y_1, \dots, Y_g)$ of symmetric matrices on a common finite dimensional vector space of a sufficiently large dimension (depending upon p). The main result of this paper is that every matrix convex polynomial has degree two or less. More generally, the polynomial p has degree at most two if convexity holds only for all matrices X and Y in an "open set." An analogous result for nonsymmetric variables is also obtained.

Matrix convexity is an important consideration in engineering system theory. This motivated our work, and our results suggest that matrix convexity in conjunction with a type of "system scalability" produces surprisingly heavy constraints.

Key words. matrix convex, linear matrix inequality, noncommutative polynomial

AMS subject classifications. 26B25, 47A56, 47A63, 47L25

DOI. 10.1137/S0895479803421999

1. Introduction.

Let $x = \{x_1, \dots, x_g\}$ denote noncommuting indeterminates and let $\mathcal{N}(x)$ denote the set of polynomials in the indeterminates x. For example,

$$p = x_1 x_2^3 + x_2^3 x_1 + x_3 x_1 x_2 + x_2 x_1 x_3$$

is a symmetric polynomial in $\mathcal{N}(x)$.

A symmetric polynomial p is *matrix convex* if for each positive integer n, each pair of tuples $X = (X_1, \dots, X_g)$ and $Y = (Y_1, \dots, Y_g)$ of symmetric $n \times n$ matrices, and each $0 \le t \le 1$,

$$p(tX + (1-t)Y) \le tp(X) + (1-t)p(Y), \tag{1.1}$$

where for an $n \times n$ matrix A, the notation $A \ge 0$ means A is positive semidefinite; i.e., A is symmetric and $\langle Ax, x\rangle \ge 0$ for all vectors x. Even in one variable, convexity in the noncommutative setting differs from convexity in the commuting case because here Y need not commute with X. For example, to see $p = x^4$ is not matrix convex, let

$$X = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix} \text{ and } Y = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

and compute

$$\frac{1}{2}X^4 + \frac{1}{2}Y^4 - \left(\frac{1}{2}X + \frac{1}{2}Y\right)^4 = \begin{pmatrix} 164 & 120 \\ 120 & 84 \end{pmatrix},$$

*Received by the editors January 27, 2003; accepted for publication (in revised form) by R. Bhatia July 8, 2003; published electronically July 14, 2004.
http://www.siam.org/journals/simax/25-4/42199.html

†Department of Mathematics, University of California, San Diego, CA 92093 (helton@osiris.ucsd.edu, http://www.ucsd.edu). The research of this author was partially supported by the NSF, DARPA, and Ford Motor Company.

‡Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (sam@math.ufl.edu). The research of this author was partially supported by NSF grant DMS-0140112.

which is not positive semidefinite. On the other hand, to verify that x^2 is a matrix convex polynomial, observe that

$$tX^2 + (1-t)Y^2 - (tX + (1-t)Y)^2 \\ = t(1-t)(X^2 - XY - YX + Y^2) = t(1-t)(X-Y)^2 \geq 0.$$

Our main theorem, Theorem 3.1, says (in several contexts) that

any noncommutative polynomial which is "matrix convex" on an "open set" has degree two or less.

Historical background for this result appears in section 8.2. The paper begins with the formal setup and definitions including that of open set (see section 2). After stating Theorem 3.1 we prove the theorem for symmetric variables X in two special cases, first when the polynomial is matrix convex everywhere and second when the polynomial is "matrix convex on the polydisk," since these are both important special cases and their proofs illustrate the general approach. The everywhere positive case is taken up in section 4. Section 5 contains a key lemma and the proof of the main result in the case that the polynomial is matrix convex on the polydisk. The proof of the general case for both symmetric and nonsymmetric variables is presented in section 6. As an aside we mention, in section 6.3, alternative proofs which yield partial results. A refinement of the main result which connects the work with linear matrix inequalities (LMIs) is discussed in section 7. The paper concludes with section 8, which indicates engineering motivation.

Here is the idea of the proof. A noncommutative polynomial p has a calculus second directional derivative q which is also a polynomial with degree the same as that of p, unless p has degree less than or equal to one, in which case $q = 0$. Our working definition of matrix convex, as discussed in section 2.4, corresponds to the second directional derivative q of p being a "matrix positive" polynomial. Earlier results [M01], [H02], and [MPprept] say that matrix positive noncommutative polynomials are all sums of squares.[1] We compute that, if the degree of p exceeds two, then q has terms which preclude it from being a sum of squares. This settles the matrix convex everywhere case.

Convexity on an open set corresponds to positivity of the second derivative q on that set, but now q is not likely to be a sum of squares. In this case, we apply a type of noncommutative Positivstellensatz from [CHSY]. In the symmetric variable and "positive on the polydisk" case, the Positivstellensatz of [HMPprept] suffices.

2. Definitions. We shall now give formal definitions at appropriate levels of generality.

2.1. Noncommutative polynomials. Of interest are two classes of noncommutative variables $x = \{x_1, \ldots, x_g\}$. In the first the x_j are symmetric and in the second they are free of relations. (So far in the introduction we have discussed only the symmetric situation.) In both cases, the definition of a convex polynomial requires g new noncommutative variables $\{h_1, \ldots, h_g\}$ either symmetric or free in correspondence with the nature of x. Now we give more details.

Let $\mathcal{F}(x)$ denote the free semigroup on the noncommutative generators $x = \{x_1, \ldots, x_g\}$. In common language, $\mathcal{F}(x)$ is the semigroup of words in $x_1, \ldots, x_g$. Note that the empty word $\emptyset$ is the identity in $\mathcal{F}(x)$.

[1] This is in contrast with the situation in the commutative case emanating from Hilbert's work and his 17th problem; see [R00] for results and a survey and [PV99] for a general closely related Positivstellensatz.

Let $\mathcal{N}(x)$ denote the polynomials, over the field of real numbers $\mathbb{R}$, in the noncommuting generators $x = \{x_1, \dots, x_g\}$. Thus $\mathcal{N}(x)$ is the free $\mathbb{R}$-algebra on x. As a vector space, $\mathcal{N}(x)$ consists of real linear combinations of words w from $\mathcal{F}(x)$. Concretely, a $p \in \mathcal{N}(x)$ is an expression of the form

$$p = \sum_{w \in \mathcal{F}(x)} p_w w, \tag{2.1}$$

where the sum is finite and each $p_w \in \mathbb{R}$. The algebra $\mathcal{N}(x)$ has a natural involution ^{T}, which behaves in the following way. Given a word $w = x_{j_1} x_{j_2} \cdots x_{j_n}$ from $\mathcal{F}(x)$ viewed as an element of $\mathcal{N}(x)$, the involution applied to w is

$$w^T = x_{j_n} \cdots x_{j_2} x_{j_1}.$$

In general, given p as in (2.1), $p^T = \sum p_w w^T$. A polynomial p *in* $\mathcal{N}(x)$ *is symmetric* provided $p^T = p$.

Define $\mathcal{F}(x)[h]$ and $\mathcal{N}(x)[h]$ by analogy with $\mathcal{F}(x)$ and $\mathcal{N}(x)$ as the free semigroup and free $\mathbb{R}$-algebra in the $2g$ variables $\{x, h\} = \{x_1, \dots, x_g, h_1, \dots, h_g\}$, respectively. While $\mathcal{F}(x)[h]$ and $\mathcal{N}(x)[h]$ are the same as $\mathcal{F}(x)$ and $\mathcal{N}(x)$ with g replaced by $2g$, in what follows the variables x and h will play a somewhat different role. Often we will write $\mathcal{N}$ (resp., $\mathcal{F}$) instead of either $\mathcal{N}(x)$ or $\mathcal{N}(x)[h]$ (resp., $\mathcal{F}(x)$ or $\mathcal{F}(x)[h]$).

Let $\mathcal{F}_*(x)$ and $\mathcal{N}_*(x)$ denote the free semigroup and free $\mathbb{R}$-algebra on the $2g$ variables $\{x, x^T\} = \{x_1, \dots, x_g, x_1^T, \dots, x_g^T\}$. The involution in this setting is determined by $x_j \mapsto x_j^T$ and $x_j^T \mapsto x_j$, so that if w is a word in $\{x, x^T\}$, say, $w = z_1 \cdots z_n$, then

$$w^T = z_n^T \cdots z_1^T.$$

Here $z_j \in \{x, x^T\}$. The involution extends from $\mathcal{F}_*(x)$ to $\mathcal{N}_*(x)$ in the canonical way.

Finally, the notation $\mathcal{F}_*(x)[h]$ and $\mathcal{N}_*(x)[h]$ will denote the free semigroup and free $\mathbb{R}$-algebra on the $4g$ generators

$$\{x_1, \dots, x_g, x_1^T, \dots, x_g^T, h_1, \dots, h_g, h_1^T, \dots, h_g^T\}$$

with involution defined by analogy with $\mathcal{F}_*(x)$ and $\mathcal{N}_*(x)$. Often we will write $\mathcal{N}_*$ (resp., $\mathcal{F}_*$) instead of either $\mathcal{N}_*(x)$ or $\mathcal{N}_*(x)[h]$ (resp., $\mathcal{F}_*(x)$ or $\mathcal{F}_*(x)[h]$).

2.2. Matrix noncommutative polynomials. Given a finite index set $\mathcal{J}$ and a set $\mathcal{S}$, let $M_{\mathcal{J}}(\mathcal{S})$ denote the matrices with entries from $\mathcal{S}$ indexed by $\mathcal{J}$. Thus, an $M \in M_{\mathcal{J}}(\mathcal{S})$ has the form $M = (M_{j,\ell})_{j,\ell \in \mathcal{J}}$ for some $M_{j,\ell} \in \mathcal{S}$. In the case $\mathcal{J} = \{1, \dots, n\}$, the set $M_{\mathcal{J}}(\mathcal{S})$ is simply $M_n(\mathcal{S})$, the $n \times n$ matrices with entries from $\mathcal{S}$. Similarly, view $\mathcal{S}^{\mathcal{J}}$ as (column) vectors indexed by $\mathcal{J}$. For instance, when $\mathcal{J} = \{1, \dots, n\}$, we find $\mathcal{S}^{\mathcal{J}} = \mathcal{S}^n$ is the set of n-vectors with entries from $\mathcal{S}$.

If $\mathcal{J}$ is a finite subset of $\mathcal{F}$ and $\mathcal{S} = \mathcal{N}$, then $M_{\mathcal{J}}(\mathcal{N})$ is an algebra with involution

$$M^T = (M_{v,w})_{v,w \in \mathcal{J}}^T = (M_{w,v}^T)_{v,w \in \mathcal{J}}.$$

Further, given $V \in \mathcal{N}^{\mathcal{J}}$ and $M \in M_{\mathcal{J}}(\mathcal{N})$, the definition

$$V^T M V = \sum_{u,w \in \mathcal{J}} V_u^T M_{u,w} V_w$$

is unavoidable. Elements of $M_{\mathcal{J}}(\mathcal{N})$ are naturally identified with noncommutative matrix-valued polynomials by writing $p \in M_{\mathcal{J}}(\mathcal{N})$ as

$$p = \sum_{w \in \mathcal{F}} p_w w \tag{2.2}$$

just as in (2.1), but now where $p_w \in M_{\mathcal{J}}(\mathbb{R})$. With this notation, the involution is given by

$$p^T = \sum_{w \in \mathcal{F}} p_w^T w^T.$$

A matrix-valued noncommutative polynomial of degree one is a linear pencil. Explicitly, in the $\mathcal{N}(x)$ case, a linear pencil Λ has the form

$$\Lambda = \Lambda_0 + \sum_1^g \Lambda_j x_j,$$

where $\Lambda_j \in M_n(\mathbb{R})$ for some n (or more generally, the Λ_j are operators on a Hilbert space).

2.3. Substituting matrices for indeterminates. Often we shall be interested in evaluating a polynomial p in $\mathcal{N}(x)$ at a tuple of bounded symmetric operators $X = (X_1, \ldots, X_g)$ on a common real Hilbert space $\mathcal{H}$. Define $X^{\emptyset} = I$, the identity operator on $\mathcal{H}$; given a word $w \in \mathcal{F}(x)$ different from the empty word, $w = x_{j_1} x_{j_2} \cdots x_{j_n}$, let

$$X^w = X_{j_1} X_{j_2} \cdots X_{j_n};$$

and given p as in (2.1), define $p(X) = \sum p_w X^w$. Note that the involution on $\mathcal{N}$ is compatible with the transpose operation on operators on real Hilbert space,

$$p(X)^T = p^T(X),$$

where $p(X)^T$ denotes the transpose of the operator $p(X)$ (with respect to the native inner product). Often the Hilbert space is $\mathbb{R}^n$ and so the operators X_j are real symmetric $n \times n$ matrices and $p(X)^T$ is just the usual transpose of the $n \times n$ matrix $p(X)$.

Let $\mathcal{B}(\mathcal{H})$ denote the bounded linear operators on $\mathcal{H}$. A fixed tuple $X = (X_1, \ldots, X_g)$ of symmetric elements of $\mathcal{B}(\mathcal{H})$ determines an algebra homomorphism $\mathcal{N}(x) \longrightarrow \mathcal{B}(\mathcal{H})$ which preserves ^{T} by evaluation, $p \mapsto p(X)$. This evaluation mapping extends to matrix polynomials $M_{\mathcal{J}}(\mathcal{N}(x)) \longrightarrow M_{\mathcal{J}}(\mathcal{B}(\mathcal{H}))$ by defining, for a p in $M_{\mathcal{J}}(\mathcal{N}(x))$ with entries $p_{j,\ell}$, the matrix $p(X)$ as the matrix with entries $p_{j,\ell}(X)$. In other words, we apply the evaluation map entrywise. Note that $M_{\mathcal{J}}(\mathcal{B}(\mathcal{H}))$ is naturally identified with $\mathcal{B}(\oplus_{\mathcal{J}}\mathcal{H})$ and that, in the notation of (2.2),

$$p(X) = \sum p_w \otimes X^w,$$

where the coefficients are matrices. If $X = (X_1, \ldots, X_g)$ and $H = (H_1, \ldots, H_g)$ are tuples of symmetric operators on $\mathcal{H}$, then the evaluation homomorphism defined by

$$p(x, h) = p(x)[h] \mapsto p(X, H) = p(X)[H]$$

acts as a mapping $\mathcal{N}(x)[h] \longrightarrow \mathcal{B}(\mathcal{H})$.

In the $\mathcal{N}_*(x)$ case, evaluation is allowed at arbitrary tuples $X = (X_1, \dots, X_g)$ of operators on a common real Hilbert space $\mathcal{H}$, where now X_j^T is substituted for x_j^T. Evaluation of $p \in M_{\mathcal{J}}(\mathcal{N}_*(x))$ or $p \in M_{\mathcal{J}}(\mathcal{N}_*(x)[h])$ at tuples X or (X, H) is defined also.

LEMMA 2.1. *Given d, there exists a real Hilbert space $\mathcal{K}$ of dimension $\sum_0^{2d} g^j$ and a tuple $Y = (Y_1, \dots, Y_g)$ of symmetric operators on $\mathcal{K}$ such that if $p \in \mathcal{N}(x)$ has degree at most d and if $p(Y) = 0$, then $p = 0$.*

Similarly, there exists a Hilbert space $\mathcal{K}$ of dimension $\sum_0^{2d}(2g)^j$ and a tuple of operators $Y = (Y_1, \dots, Y_g)$ on $\mathcal{K}$ such that if $p \in \mathcal{N}_*(x)$ has degree at most d and if $p(Y) = 0$, then $p = 0$.

We will have use of the following variant of Lemma 2.1, which uses only that for each p there is a Y (perhaps depending upon p) in Lemma 2.1. Let $\mathcal{B}^{\mathrm{sym}}(\mathcal{H})^g$ denote g-tuples $X = (X_1, \dots, X_g)$ of symmetric operators on $\mathcal{H}$. Let $\mathcal{B}(\mathcal{H})^g$ denote all g-tuples of operators on $\mathcal{H}$. In the case $\mathcal{H} = \mathbb{R}^n$, we write $(M_n^{\mathrm{sym}})^g$ and M_n^g in place of $\mathcal{B}^{\mathrm{sym}}(\mathbb{R}^n)^g$ and $\mathcal{B}(\mathbb{R}^n)^g$.

LEMMA 2.2. *Given d, there exists a Hilbert space $\mathcal{K}$ of dimension $\sum_0^{2d} g^j$ such that if G is an open subset of $\mathcal{B}^{sym}(\mathcal{K})^g$, if $p \in \mathcal{N}(x)$ has degree at most d, and if $p(X) = 0$ for all $X \in G$, then $p = 0$.*

Similarly, there exists a Hilbert space $\mathcal{K}$ of dimension $\sum_0^{2d}(2g)^j$ such that if G is an open subset of $\mathcal{B}(\mathcal{K})^g$, if $p \in \mathcal{N}_*(x)$ has degree at most d, and if $p(X) = 0$ for all $X \in G$, then $p = 0$.

Proof. Choose a $Z \in G$ and let $h, k \in \mathcal{K}$ be given. Define the old-fashioned polynomial on $t \in \mathbb{R}$,

$$s(t) = \langle p((1-t)Z + tY)h, k\rangle,$$

where Y is the tuple from Lemma 2.1 and $(1-t)Z + tY$ is the tuple

$$((1-t)Z_1 + tY_1, \dots, (1-t)Z_g + tY_g).$$

Since G is open and $p(X) = 0$ for $X \in G$, $s(t) = 0$ for small t. Since s is a polynomial, $s = 0$, and hence substituting $t = 1$ gives $\langle p(Y)h, k\rangle = 0$. Thus, $p(Y) = 0$. ∎

2.4. Matrix convexity and positivity. A polynomial $q \in \mathcal{N}(x)$ is matrix positive if $q(X) \geq 0$ for all tuples $X = (X_1, \dots, X_g)$ of symmetric operators on finite dimensional Hilbert space. Matrix positive for q in either $\mathcal{N}(x)[h]$, $\mathcal{N}_*(x)$, or $\mathcal{N}_*(x)[h]$ is defined in a similar fashion.

Matrix positive polynomials are sums of squares.

THEOREM 2.3. *Given d, there exists a Hilbert space $\mathcal{K}$ of dimension $N(d) = \sum_0^d g^j$ such that if $q \in \mathcal{N}(x)$, the degree of q is at most d, and $q(X) \geq 0$ for all tuples $X = (X_1, \dots, X_g)$ on $\mathcal{K}$, then there exists $r_j \in \mathcal{N}(x)$, $1 \leq j \leq N(d)$, such that $q = \sum r_j^T r_j$.*

Similarly, there exists a Hilbert space $\mathcal{K}$ of dimension $N(d) = \sum_0^d (2g)^j$ such that if $q \in \mathcal{N}_*(x)$, q has degree at most d, and $q(X) \geq 0$ for all tuples $X = (X_1, \dots, X_g)$ on $\mathcal{K}$, then there exists $r_j \in \mathcal{N}_*(x)$, $1 \leq j \leq N(d)$, such that $q = \sum r_j^T r_j$.

Versions of this sum-of-squares result can be found in [H02], [M01], and [MPprept].

2.4.1. Matrix convexity. Matrix convexity can be formulated in terms of the second derivative and positivity, just as in the case of a real variable. Given a polynomial $p \in \mathcal{N}(x)$,

$$r(x)[h] := p(x+h) - p(x)$$

is a polynomial in $\mathcal{N}(x)[h]$. Define the Hessian q of p to be the part of $r(x)[h]$ which is homogeneous of degree two in h. Alternatively, the Hessian is the second directional derivative of p,

$$q(x)[h] := \frac{d^2 p(x+th)}{dt^2}|_{t=0}.$$

For example, $p = x_1^2 x_2$ has Hessian

$$q(x)[h] = h_1^2 x_2 + h_1 x_1 h_2 + x_1 h_1 h_2.$$

If $q \neq 0$, that is, if p has degree greater than or equal to two, then the degree of q equals the degree of p.

THEOREM 2.4 (see [HMer98]). *A polynomial $p \in \mathcal{N}$ is matrix convex if and only if its Hessian $q(x)[h]$ is matrix positive.*

A polynomial $p \in \mathcal{N}_*(x)$ is matrix convex if (1.1) holds for all tuples X and Y whether symmetric or not. The Hessian of p is again the homogeneous-of-degree-two-in-h part of $p(x+h) - p(x)$. For instance, the Hessian of $p(x) = xx^T x$ is $xh^T h + hx^T h + hh^T x$. Theorem 2.4 is true with $\mathcal{N}$ replaced by $\mathcal{N}_*$.

2.4.2. Positivity domains. Let $M_\infty(\mathcal{N})$ and $M_\infty(\mathcal{N}_*)$ denote the unions $\cup_{n=1}^\infty M_n$ $(\mathcal{N}(x))$ and $\cup_{n=1}^\infty M_n(\mathcal{N}_*(x))$, respectively. Fix a subset $\mathcal{P}$ of $M_\infty(\mathcal{N})$ or $M_\infty(\mathcal{N}_*)$. The case that $\mathcal{P}$ consists of symmetric polynomials is of primary interest, but we will have occasion to consider more general collections. Given a real Hilbert space $\mathcal{H}$, let $\mathcal{D}_\mathcal{P}(\mathcal{H})$ denote the tuples $X = (X_1, \dots, X_g)$ such that each X_j is an operator on $\mathcal{H}$ and $p(X) \geq 0$ for each $p \in \mathcal{P}$. In the $\mathcal{N}$ case each X_j is, of course, assumed symmetric.

The *positivity domain of* $\mathcal{P}$, denoted $\mathcal{D}_\mathcal{P}$, is the collection of tuples X such that $X \in \mathcal{D}_\mathcal{P}(\mathcal{H})$ for some $\mathcal{H}$. The fact that $\mathcal{D}_\mathcal{P}$ is not actually a set presents no logical difficulties and typically it may be assumed that the Hilbert spaces are separable and even finite dimensional.

2.4.3. Matrix convexity on a positivity domain. Given a collection $\mathcal{P} \subset M_\infty(\mathcal{N}(x))$ with corresponding positivity domain $\mathcal{D}_\mathcal{P}$, a polynomial $q \in \mathcal{N}(x)[h]$ is *matrix positive on* $\mathcal{D}_\mathcal{P}$ if $q(X)[H]$ is positive semidefinite for all tuples $X = (X_1, \dots, X_g)$ and $H = (H_1, \dots, H_g)$ of symmetric operators on a common Hilbert space such that $X \in \mathcal{D}_\mathcal{P}$. The polynomial $p \in \mathcal{N}(x)$ is *matrix convex on* $\mathcal{D}_\mathcal{P}$ provided its Hessian is matrix positive on $\mathcal{D}_\mathcal{P}$. When $\mathcal{D}_\mathcal{P}$ is all matrices, for example, if $\mathcal{P}$ consists of the polynomial 1, then matrix convexity on $\mathcal{D}_\mathcal{P}$ is the same as matrix convexity.

Matrix convex on a positivity domain is defined in the $\mathcal{N}_*$ case in the expected manner.

2.4.4. The openness condition.

DEFINITION 2.5 (openness property). *The positivity domain $\mathcal{D}_\mathcal{P}$ has the* openness property *provided that there is an integer n_0 with the property that when $n > n_0$, the set of matrices $\mathcal{D}_\mathcal{P} \cap M_n$ is equal to the closure of the interior of $\mathcal{D}_\mathcal{P} \cap M_n$. Often we say such a $\mathcal{D}_\mathcal{P}$ is an* open positivity domain.

3. The main theorem.

THEOREM 3.1. *If a noncommutative symmetric polynomial p is matrix convex on some positivity domain which satisfies the openness condition, then p has degree two or less. Here either $p \in \mathcal{N}(x)$ or $p \in \mathcal{N}_*(x)$ with matrix convex interpreted accordingly.*

4. Proof of Theorem 3.1 for everywhere convex polynomials. We first treat the special case of Theorem 3.1 in which $p \in \mathcal{N}$ is matrix positive everywhere, since it is easy and serves as a guide to part of the proof of Theorem 3.1.

PROPOSITION 4.1. *If a noncommutative symmetric polynomial p in symmetric variables is matrix convex everywhere, then p has degree two or less. That is, if $p \in \mathcal{N}(x)$ is matrix convex (everywhere), then the degree of p is at most two.*

Given $p \in \mathcal{N}$,

$$p = \sum_w p_w w,$$

we say p *contains* the word u or u *appears* in p if $p_u \neq 0$.

Proof. Let $q(x)[h]$ denote the second directional derivative of p in direction h. It is a symmetric polynomial which is homogeneous of degree two in h. By Theorem 2.4 the polynomial p is matrix convex if and only if q is matrix positive. Thus, by Theorem 2.3, q is a sum of squares so that there exist an m and polynomials r_j in x and h such that q has the form

$$q = \sum_{j=1}^{m} r_j^T r_j. \tag{4.1}$$

Write each r_j as

$$r_j = \sum_{w \in \mathcal{F}(x)[h]} r_j(w) w,$$

where all but finitely many of the coefficients $r_j(w) \in \mathbb{R}$ are 0.

We begin our analysis of the r_j by showing that each r_j has degree in h no greater than one. For a polynomial $r \in \mathcal{N}(x)[h]$, let $\deg_h(r)$ denote the degree of r in h and $\deg_x(r)$ denote the degree of r in x. Let

$$d_h = \max\{\deg_h(r_j) : j\},$$

let

$$d_x = \max\{\deg_x(w) : \text{ there exists } j \text{ so that } r_j \text{ contains } w \text{ and } \deg_h(w) = d_h\},$$

and let

$$\mathcal{S}_{d_x,d_h} := \{w : r_j \text{ contains } w \text{ for some } j,\ \deg_h(w) = d_h, \text{ and } \deg_x(w) = d_x\}.$$

The portion of q homogeneous of degree $2d_h$ in h and $2d_x$ in x is

$$Q = \sum_{\{j=1,\dots,m,\ v,w \in \mathcal{S}_{d_x,d_h}\}} r_j(v) r_j(w) v^T w.$$

Since, for $v_j, w_j \in \mathcal{S}_{d_x,d_h}$, the equality $v_1^T w_1 = v_2^T w_2$ can occur if and only if $v_1 = v_2$ and $w_1 = w_2$, we see that $Q \neq 0$ and thus $\deg_h(q) = 2d_h$. Since q has degree two in h, we obtain $2d_h = 2$, so $d_h = 1$.

Now we turn to bounding the total degree of q. The asymptotics of a matrix positive q dictate that it have even degree. Accordingly, denote the degree of q by

$2N$. Recall that $2N$ is also the degree of p, since we may assume degree $p \geq 3$, or the corollary is proved. Thus the polynomial p contains a term of the form

$$t := x_{\ell_1} x_{\ell_2} x_{\ell_3} \cdots x_{\ell_{2N}}. \tag{4.2}$$

The second derivative of t in the direction h contains a term of the form

$$\mu := h_{\ell_1} h_{\ell_2} x_{\ell_3} \cdots x_{\ell_{2N}}$$

and consequently $q(x)[h]$ contains the term μ. Thus, at least one of the products $r_{j_0}^T r_{j_0}$ must contain μ. Use now that r_{j_0} has degree at most one in h to conclude that r_{j_0} must contain the term $h_{\ell_2} x_{\ell_3} \cdots x_{\ell_{2N}}$ and therefore the polynomial r_{j_0} has (total) degree at least $2N-1$.

Next observe canceling the terms of largest (total) degree in $\sum r_j^T r_j$ is impossible, so each r_j is a polynomial of degree half of the degree of q or less. That is, $\deg(r_j) \leq N$ for each j, including r_{j_0}. It follows that $N \leq 1$. ∎

5. Gram representations. In this section we lay groundwork for proving Theorem 3.1 and prove a special case which illustrates the general idea.

5.1. A Gram representation for a polynomial. The analogue of the sum-of-squares representation (4.1) used in the proof of Corollary 4.1 required for the proof in the general case is a Gram representation for a polynomial $q(x)[h] = q(x,h)$ which is homogeneous of degree two in h and matrix positive on a positivity domain. We discuss the case of symmetric variables. The case of nonsymmetric variables is similar but notationally more complicated.

Since q is homogeneous of degree two in h, it may be written as

$$q(x,h) = V(x)[h]^T M(x) V(x)[h] = V^T M V, \tag{5.1}$$

where the *border* vector $V(x)[h]$ is linear in h and has the form

$$V(x)[h] := \begin{pmatrix} V^1(x)[h_1] \\ \vdots \\ V^k(x)[h_k] \end{pmatrix}, \text{ where } V^j(x)[h_j] = \begin{pmatrix} h_j m_1^j(x) \\ h_j m_2^j(x) \\ \vdots \\ h_j m_{\ell_j}^j(x) \end{pmatrix}. \tag{5.2}$$

The m_r^j are monomials in x, and the matrix M is symmetric and its entries are noncommutative polynomials in x. The following lemma says we may (and we will) take V to have the property that for each fixed j all of the $m_r^j(x)$ are distinct monomials.

LEMMA 5.1. *There is a $V^T M V$ representation (5.1) for $q(x)[h]$ in which for each fixed j all of the $m_i^j(x)$ are distinct monomials. Here "distinct" precludes one monomial being a scalar multiple of another.*

Proof. One can represent $q(x)[h]$ as in (5.1) with m_i^j being monomials. Clearly the only issue is whether two of these monomials are collinear. The proof that collinearity in $V(x)[h]$ is removable can be done with induction where the key induction step goes as follows. Suppose we have a q with the representation

$$q(x)[h] = \begin{pmatrix} m \\ \alpha m \\ n \end{pmatrix}^T \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \begin{pmatrix} m \\ \alpha m \\ n \end{pmatrix}$$

with α a real number and m and n noncollinear monomials. Write q as

$$\begin{aligned} q(x)[h] = \; & m^T(p_{11} + \alpha^2 p_{22} + \alpha p_{21} + \alpha p_{12})m \\ & + m^T(p_{13} + \alpha p_{23})n + n^T(p_{31} + \alpha p_{32})m + n^T p_{33} n, \end{aligned}$$

which leads to the representation

$$q(x)[h] = \begin{pmatrix} m \\ \\ n \end{pmatrix}^T \begin{pmatrix} p_{11} + \alpha^2 p_{22} + \alpha p_{21} + \alpha p_{12} & p_{13} + \alpha p_{23} \\ \\ p_{31} + \alpha p_{32} & p_{33} \end{pmatrix} \begin{pmatrix} m \\ \\ n \end{pmatrix},$$

which has linearly independent borders. □

Here "distinct" precludes one monomial being a scalar multiple of another. It will be convenient at times to index the polynomial entries of the matrix M by the monomials $\mathcal{J}$ in $V(x)[h]$ as in subsection 2.2. In this way M has entries $M_{h_j m_\ell^j, h_{j'} m_{\ell'}^{j'}}$.

For convenience, arrange

$$m_1^j < m_2^j < \cdots < m_{\ell_j}^j$$

in, say, graded lexicographic order (that is, low degree is less than high degree and after that dictionary order breaks ties). Also we assume that each monomial is essential to representing q; that is, no proper subset of $\{m_1^j, \ldots, m_{\ell_j}^j\}$ produces such a representation of q. In particular, no row (or column) of M is identically zero. Such a "Gram" representation always exists and, along with a surprising positivity property, is proved in [CHSY]. This will be recalled formally later; see Theorem 6.1 (Theorem 8.3 of [CHSY]) stated near the end of the proof. See also [HMPprept] for a result which is more general in certain directions.

In the next subsection we prove a property of M unique in the fact that it represents q, the Hessian of a polynomial.

5.2. The degree of *q* versus positivity of its representer. The following key lemma is presented for symmetric as well as nonsymmetric variables, since this does not complicate notation.

LEMMA 5.2. *Let p be a symmetric polynomial in either $\mathcal{N}(x)$ or $\mathcal{N}_*(x)$. Suppose the Hessian $q(x,h)$ of p (which is in either $\mathcal{N}(x)[h]$ or $\mathcal{N}_*(x)[h]$ depending upon p and is homogeneous of degree two in h) is represented by $V^T M V$ as in* (5.1). *If the degree of q in x and h together exceeds two, then there is an integer n_0 such that $M(X)$ is not positive semidefinite on any open set of tuples X of matrices of dimension n greater than or equal to n_0. In fact, if d is the degree of M in x and h jointly, then n_0 can be chosen to be equal to either $\sum_0^d g^j$ or $\sum_0^d (2g)^j$ in the $\mathcal{N}$ and $\mathcal{N}_*$ cases, respectively.*

Proof. First we treat p with general nonsymmetric x and h.

Let N denote the degree[2] of p; then p must contain a term of one of the following forms:

$$t := x_i^T x_j m \quad \text{or} \quad t := x_i x_j m$$

or

$$t := x_i x_j^T m \quad \text{or} \quad t := x_i^T x_j^T m,$$

[2]Convexity is assumed on a region only, so we cannot use asymptotic arguments to conclude immediately that N is even.

where m is a monomial of degree $N-2$ in x, x^T. We work through in detail what happens when a t of the form $t = x_i^T x_j m$ appears in p, since the other cases go similarly. In the second directional derivative $q(x,h)$ of p, a term of the form

$$\mu := h_i^T h_j m \tag{5.3}$$

appears. Thus, in the j part, V^j, of the border vector V, the monomial $h_j m$ appears. The monomial $m_{\ell_j}^j$ has largest degree (in x as it is a monomial in x only) of those monomials in V^j and thus the degree of $m_{\ell_j}^j$ is at least $N-2$.

Suppose $M_{h_j m_{\ell_j}^j, h_j m_{\ell_j}^j} \neq 0$. In this case

$$q_{\ell_j} := m_{\ell_j}^{j^T} h_j^T M_{h_j m_{\ell_j}^j, h_j m_{\ell_j}^j} h_j m_{\ell_j}^j$$

is a nonzero polynomial which is part of q and which cannot be canceled by any other part of q by the nature of the $q = V^T M V$ representation and the fact that $m_{\ell_j}^j$ is largest in the monomial ordering. (The key property here is the distinctness of the terms in V which prevents the $h_j m_\ell^j$ from repeating.) It follows that the degree of q is at least twice the degree of $h_j m_{\ell_j}^j$, and hence the degree of q is at least $2(N-1)$. On the other hand, the degree of q is N. Hence, $2(N-1) \leq \deg(q) \leq N$ and it follows that $N \leq 2$. Thus p and q have degree no greater than two.

From the preceding paragraph, if q has degree exceeding two, then

$$M_{h_j m_{\ell_j}^j, h_j m_{\ell_j}^j} = 0.$$

Fix $n \geq \sum_0^d (2g)^j$ and let $\mathcal{O} = \{X \in (M_n)^g : M(X) \geq 0\}$. For each $X \in \mathcal{O}$ and monomial w appearing in V, we see that the entries $M_{h_j m_{\ell_j}^j, w}(X)$ of the matrix $M(X)$ are zero. This is because $M(X)$ is positive semidefinite and the diagonal entry $M_{h_j m_{\ell_j}^j, h_j m_{\ell_j}^j}(X)$ is zero. If $\mathcal{O}$ contains an open set, then, from Lemma 2.2, each $M_{h_j m_{\ell_j}^j, w} = 0$, which contradicts our standing assumption that $h_j m_{\ell_j}^j$ is actually needed to represent q. Thus, $\mathcal{O}$ contains no open set and this is the conclusion of the lemma. Thus we have proved the lemma for the $\mathcal{N}_*$ case when p contains $t = x_i^T x_j m$. If p contains $t := x_i x_j m$ or $t := x_i x_j^T m$ or $t := x_i^T x_j^T m$, times an irrelevant scalar multiple, the proof proceeds exactly as before with $\mu := h_i h_j m$, respectively, $\mu := h_i h_j^T m$, or, respectively, $\mu := h_i^T h_j^T m$, replacing $\mu = h_i^T h_j m$. The lemma is proved for the $\mathcal{N}_*$ case.

The proof for the case with symmetric variables x, h is a minor variation of the proof we just gave. □

5.3. Proof of a special case. Theorem 3.1 for polynomials in $\mathcal{N}$ and special $\mathcal{D}_\mathcal{P}$ follows from Lemma 5.2 and either the main result of [HMPprept] or specialization of Theorem 8.3 of [CHSY] about rational functions to polynomials. The main value of presenting this case is that the proof is short yet informative.

THEOREM 5.3. *If $p \in \mathcal{N}(x)$ is matrix convex on the collection $\mathcal{D}$ of all tuples $X = (X_1, \ldots, X_g)$ of symmetric operators acting on a common Hilbert space with each X_j a contraction, that is, $\|X_j\| \leq 1$, then p has degree at most two. We emphasize that the conclusion holds whenever p is matrix convex on a positivity domain $\mathcal{D} = \mathcal{D}_\mathcal{P}$ which contains all tuples of symmetric contractions.*

Proof. The hypothesis on p implies that its Hessian q satisfies $q(X)[H] \succeq 0$ for all tuples $X = (X_1, \ldots, X_g)$ of symmetric contractions and all tuples $H = (H_1, \ldots, H_g)$ of symmetric operators (all on the same Hilbert space). As a special case of the main result of [HMPprept], it follows that q has a representation $q = V^T M V$ as in (5.1) with $M(X) \succeq 0$ for all tuples $X = (X_1, \ldots, X_g)$ of symmetric contractions. Lemma 5.2 implies q has degree at most two. Since $\deg(p) = \deg(q)$, we conclude that the degree of p is at most two. □

Note that this is Theorem 3.1 for polynomials in $\mathcal{N}$ except here we have a special type of set, a polydisk, which satisfies the openness condition. It is tempting to conclude that Theorem 3.1 follows immediately from this by scaling and translating the unit polydisk. However, in our noncommutative setting, translation is permissible only by a multiple of the identity.

The restriction to $\mathcal{D}_\mathcal{P}$ consisting of contractions is occasioned by use of [HMPprept]. However, as we soon see, the substitution of a key result from [CHSY] permits the extension of the result to any positivity domain which satisfies the openness condition.

6. Proof of Theorem 3.1. Our proof of Theorem 3.1 for matrix convex polynomials in either $\mathcal{N}$ or $\mathcal{N}_*$ and general positivity domains $\mathcal{D}_\mathcal{P}$ requires Theorem 8.3 of [CHSY] which analyzes, very generally, positivity of the M in $V^T M V$ representations.

6.1. Background. Theorem 8.3 in [CHSY] actually was stated at a sufficient level of generality for the case at hand. The statement requires considerable notation which explains why we did not do this earlier. The first subsection follows the layout of [CHSY] and describes the general structure. The statement of Theorem 8.3 in [CHSY] is in the second subsection.

6.1.1. $V(x)[h]$ for the general case. In a slight change of notation, we now consider quadratic functions in the tuple of variables h, some of which are constrained to be symmetric and some not.

Define h as

$$h := \{h_{-N}, \ldots, h_{-1}, h_1, \ldots, h_N, h_{N+1}, \ldots, h_r, h_{r+1}, \ldots, h_k\}, \tag{6.1}$$

where $\{h_j\}_{j=r+1}^k$ are constrained to be symmetric and $h_j = h_{-j}^T$ for $j = 1, \ldots, N$. That is, separate h into three different parts as follows: the first part[3] $\{h_j\}_{j=-N}^N$ has the pairwise restriction that $h_{-j} = h_j^T$ for $j = 1, \ldots, N$; the second part $\{h_j\}_{j=N+1}^r$ has no restriction; and the third part $\{h_j\}_{j=r+1}^N$ has each h_j constrained to be symmetric. Let $\mathcal{I}$ denote the integers between $-N$ and k except for 0. Thus, $\mathcal{I}$ is the index set for the h_j which are the entries of h.

Any noncommutative symmetric quadratic $q(x)[h]$ can be put in the form

$$V(x)[h]^T M_q V(x)[h],$$

where M_q is a rational function in x which can be taken to be a polynomial in x in the case that q is a polynomial, and where the border $V(x)[h]$ has the form

$$V(x)[h] := \begin{pmatrix} V^{mix}(x)[h] \\ V^{pure}(x)[h] \\ V^{sym}(x)[h] \end{pmatrix}, \tag{6.2}$$

[3]The integer 0 is not included in the index set $j = -N, \ldots, N$ of the first part, but for simplicity of notation we do not make this explicit, since it is clear from context.

with $V^{mix}(x)[h]$, $V^{pure}(x)[h]$, and $V^{sym}(x)[h]$ defined as follows:

$$V^{mix}(x)[h] = \begin{pmatrix} h_{-N}m_1^{-N}(x) \\ \vdots \\ h_{-N}m_{\ell_{-N}}^{-N}(x) \\ \vdots \\ h_{-1}m_1^{-1}(x) \\ \vdots \\ h_{-1}m_{\ell_{-1}}^{-1}(x) \\ h_1 m_1^1(x) \\ \vdots \\ h_1 m_{\ell_1}^1(x) \\ \vdots \\ h_N m_1^N(x) \\ \vdots \\ h_N m_{\ell_N}^h(x) \end{pmatrix}, \qquad V^{pure}(x)[h] = \begin{pmatrix} h_{N+1}m_1^{N+1}(x) \\ \vdots \\ h_{N+1}m_{\ell_{N+1}}^{N+1}(x) \\ \vdots \\ h_r m_1^r(x) \\ \vdots \\ h_r m_{\ell_r}^r(x) \end{pmatrix},$$

$$V^{sym}(x)[h] = \begin{pmatrix} h_{r+1}m_1^{r+1}(x) \\ \vdots \\ h_{r+1}m_{\ell_{r+1}}^{r+1}(x) \\ \vdots \\ h_k m_1^k(x) \\ \vdots \\ h_k m_{\ell_k}^k(x) \end{pmatrix}.$$

In order to illustrate the above definitions, we give a simple example of a quadratic function and its border vector representation. Let the quadratic function $q(x)[h]$ be given by $q(x)[h] = h_1^T\rho_1(x)h_1 + h_1\rho_2(x)h_1^T + h_2\rho_3(x)h_2^T + h_3^T\rho_4(x)h_3 + h_4\rho_5(x)h_4$, where h_1, h_2, and h_3 are not symmetric and $h_4 = h_4^T$. The ρ_j are rational functions in x. For this quadratic $q(x)[h]$, the border vector has the following structure:

$$V(x)[h] = \begin{pmatrix} h_1 \\ h_1^T \end{pmatrix} \text{ Mixed} \\ \begin{pmatrix} h_2^T \\ h_3 \end{pmatrix} \text{ Pure} \\ \begin{pmatrix} h_4 \end{pmatrix} \text{ Symmetric}.$$

Note that this representation of $q(x)[h]$ might require simple relabeling of variables. For example, if $q(x)[\{h,k\}] = h^TA(x)h + kB(x)k^T$, then $h_1 = h$, $h_2 = k^T$, and

$$V(x)[h] = V^{pure}(x)[h] = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}. \tag{6.3}$$

Allowing simple relabeling of variables increases the scope of such representations to include all cases.

6.1.2. Positive quadratic functions: Theorem 8.3 of [CHSY]. The main result of Theorem 8.3 of [CHSY] for a noncommutative rational function $q(x)[h]$ which is quadratic in h when specialized to polynomials gives the following theorem.

THEOREM 6.1 (Theorem 8.3 of [CHSY]).

Assumptions: *Consider a noncommutative polynomial $q(x)[h]$ which is a quadratic in the variables h and a set of polynomials $\mathcal{P}$ and its positivity domain $\mathcal{D}_{\mathcal{P}}$. Write $q(x)[h]$ in the form $q(x)[h] = V(x)[h]^T M(x)V(x)[h]$. Suppose that the following two conditions hold:*

i. *the positivity domain $\mathcal{D}_{\mathcal{P}}$ satisfies the openness property for some big enough n_0;*
ii. *the border vector $V(x)[h]$ of the quadratic function $q(x)[h]$ has for each fixed j distinct monomials m_i^j, $i = 1, 2, \dots \ell_j$.*

Conclusion: *The following statements are equivalent:*

a. *$q(X)[H]$ is a positive semidefinite matrix for each pair of tuples of matrices X and H for which $X \in \mathcal{D}_{\mathcal{P}}$;*
b. *$M(X) \succeq 0$ for all X in $\mathcal{D}_{\mathcal{P}}$.*

6.2. Proof for the general case. Now we finish the proof of Theorem 3.1. Choose a representation $V^T MV$ for q, the Hessian of p, where M is a matrix with entries which are polynomial in x. We wish to apply Theorem 6.1 so we must check its hypotheses (i) and (ii). Hypothesis (i) follows immediately from the fact that Theorem 3.1 requires p to be matrix convex (hence q to be matrix positive) on an open positivity domain. Hypothesis (ii) follows immediately from the fact that a representing M exists and from Lemma 5.1 which says that such a representation $V^T MV$ can always be replaced by one with distinct monomials in the border V. Theorem 6.1 implies that $M(X) \succeq 0$ for all tuples X, either symmetric or general as the case may be. An application of Lemma 5.2 just as in the proof of Theorem 5.3 completes the proof. □

6.3. Alternate proofs. We make a few remarks about the possibility of alternate proofs.

First, directly proving Theorem 3.1 for $f(s) = s^n$, where s is a single variable ($g = 1$), is easy and well known [A79], [RS79]. More generally, suppose that $g = 1$ and that p has degree n and is matrix convex everywhere. Then $\lim_{t\to\infty} \frac{1}{t^n} p(ts) = s^n$ is matrix convex. Thus $n = 0, 1$, or 2. Note that matrix convexity on an open set is not strong enough to accommodate this asymptotic argument, but, although we do not include it, it is possible to give elementary proofs for various open sets.

Next consider a polynomial p in $g > 1$ variables which is matrix convex everywhere. Make a linear change of (collapsing of) variables $Ly = x$, where L is any $g \times 1$ matrix with real entries. Then $k(y) := p(Ly)$ is a matrix convex polynomial in one variable and so has degree less than or equal to two. However, the fact that each such k has degree at most two does not necessarily imply that p has degree two. For example, if p has the property that whenever all variables x_i and x_j commute, then $p = 0$, and $k = 0$, since $(Ly)_i, (Ly)_j$ commute. Thus any polynomial which has the form

$$\sum_j l_j c_j r_j, \tag{6.4}$$

where c_j is the commutator of two polynomials, has the "$k = 0$" property. Conversely, if p has the $k = 0$ property, then p has a representation as in (6.4). Thus there are many polynomials which the one-variable result says nothing about.

7. Representing quadratic polynomials as LMIs. The following corollary of Theorem 3.1 gives a little more detail.

COROLLARY 7.1. *A matrix convex noncommutative symmetric polynomial p as in Theorem* 3.1 *can be written as*

$$p(x) = c_0 + \Lambda_0(x) + \sum_{j=1}^{N} \Lambda_j(x)^T \Lambda_j(x),$$

where $\Lambda_0, \ldots, \Lambda_N$ are linear in x and c_0 is a constant.

Proof. Convexity and Theorem 3.1 tell us that p has degree two or less. Set $\phi(x) := p(x) - c_0 - \Lambda_0(x)$, where $c_0 + \Lambda_0(x)$ is the affine linear part of p. The polynomial ϕ is a homogeneous quadratic by construction. Thus the Hessian of ϕ in direction h, which is of course homogeneous quadratic, equals $\phi(h)$. Matrix convexity says that this Hessian is matrix positive, so ϕ is matrix positive. Every matrix positive noncommutative polynomial is a sum of squares; see [H02], [M01], [MPprept]. Thus ϕ is a sum of squares,

$$\phi = \sum_{j=1}^{N} \Lambda_j(x)^T \Lambda_j(x).$$

Each of the Λ_j have degree at most one in x, as ϕ has degree two in x and since it is impossible to cancel highest degree terms in this sum-of-squares representation for ϕ. □

Remark. If q is concave, so that $p = -q$ is convex, and is represented as in Corollary 7.1, then the linear pencil

$$L(x) := \begin{pmatrix} c_0 + \Lambda_0(x) & \Lambda(x)^T \\ \Lambda(x) & -I \end{pmatrix} \tag{7.1}$$

has the same negativity domain as q, where

$$\Lambda(x) := \begin{pmatrix} \Lambda_1(x) \\ \Lambda_2(x) \\ \vdots \\ \Lambda_N(x) \end{pmatrix}.$$

This is because q is a Schur complement of $-L$ and

$$p(x) = c_0 + \Lambda_0(x) + \sum_{j=1}^{N} \Lambda_j(x)^T \Lambda_j(x) \leq 0$$

implies $c_0 + \Lambda_0(x) \leq 0$.

Those familiar with linear matrix inequalities (LMIs) see immediately that $L(x) \leq 0$ is an LMI. Thus Corollary 7.1 associates any matrix convex polynomial with an LMI. This and a variety of examples suggest to the authors that problems which correspond to concave or convex rational functions can be "converted" to equivalent LMI problems. Our speculation is bound up with the issue of convex positivity domains $\mathcal{D}_{\mathcal{P}}$, an issue not addressed in this paper (since our focus has been on noncommutative polynomials). To prove something along the lines we suggest will require vast machinery beyond that constructed here.

8. History and engineering motivation. We begin with motivation for our convexity results and then turn to history.

8.1. Engineering motivation. Motivation for this paper comes from engineering system theory. One of the main practical advances in the 1990s was methodology for converting many linear systems problems directly to matrix inequalities. See, for example, [SIG97] and [GN99], which give collections of fairly recent results along these lines.

These methods are well behaved numerically (up to matrices of modest size) provided the inequalities are convex in some sense. Further system problems where the statement of the problem does not explicitly mention system size (as is true with most classical textbook problems of control theory) typically convert to matrix inequalities where the variables are matrices. The key point is that statements which are made for these matrices must hold for matrices of any size. That is, all of the formulae in these problems scale automatically with system size (the system dimension is not explicitly mentioned). We informally call these *dimensionless* or *scalable problems*; see [H02m]. Dimensionless problems typically produce collections of noncommutative rational functions.

Thus a key issue is to analyze matrix convexity of collections of noncommutative rational functions. While this article treats only the special case of a single polynomial, the result is so strong that one suspects that even at great levels of generality noncommutative convex situations are rare and very rigid.

The author's impression (vastly incomplete, since there are thousands of engineering matrix inequality papers) of the systems literature is that whenever a dimensionless problem converts to a "convex problem," possibly by change of variables, it converts to an LMI. This is how convexity is acquired and proved in practice. The (vague) speculation in the remark in section 7, that any matrix convex problem is "associated" with some LMI, implies that matrix convexity is not fundamentally less restrictive than are LMIs for dimensionless problems.

8.2. History. Matrix convex functions have been studied since the 1930s as in the very early papers by [K36], [BS55] and followed closely after the groundbreaking work of Löwner [L34]. The focus of work until the 1990s, when engineering became an influence, was on functions of one (matrix) variable. Functions such as logs and fractional powers were studied and the closest result to the one for polynomials in this paper is described in Theorem 8.1.

THEOREM 8.1. *The function $f(X) = X^r$ on positive definite symmetric matrices X is matrix convex if $1 \le r \le 2$ or $-1 \le r \le 0$ and matrix concave if $0 \le r \le 1$.*

Theorem 8.1 is due to Ando [A79]. Conversely, Shorrock and Rizvi [RS79] show that for other values of r, the function f is neither convex nor concave. We have not seen the early derivative consequence of this—that a monic polynomial in one variable is matrix convex if and only if its degree is less than or equal to two.

More recent advances on matrix convexity are summarized in [LM00], which proves at considerable generality the matrix convexity of Schur complements. Also the special type of matrix convex structure, LMIs, recently popular with engineers, was discussed above.

REFERENCES

[A79] T. ANDO, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26 (1979), pp. 203–241.

[BS55] J. BENDAT AND S. SHERMAN, *Monotone and convex operator functions*, Trans. Amer. Math. Soc., 79 (1955), pp. 58–71.

[CHSY] J. F. Camino, J. W. Helton, R. E. Skelton, and J. Ye, *Matrix inequalities: A symbolic procedure to determine convexity automatically*, Integral Equations Operator Theory, 46 (2003), pp. 399–454.

[GN99] L. El-Ghaoui and S. Niculescu, eds., *Advances in Linear Matrix Inequality Methods in Control*, Adv. Des. Control 2, SIAM, Philadelphia, 1999.

[H02] J. W. Helton, *"Positive" noncommutative polynomials are sums of squares*, Ann. of Math. (2), 156 (2002), pp. 675–694.

[H02m] J. W. Helton, *Manipulating matrix inequalities automatically*, in Mathematical Systems Theory in Biology, Communications, Computation, and Finance (Notre Dame, IN, 2002), IMA Vol. Math. Appl. 134, Springer-Verlag, New York, 2003, pp. 237–256.

[HMer98] J. W. Helton and O. Merino, *Sufficient conditions for optimization of matrix functions*, in IEEE Conference on Decision and Control, Tampa, FL, IEEE Control Systems Society, Piscataway, NJ, 1998, pp. 1–5.

[HMPprept] J. W. Helton, S. McCullough, and M. Putinar, *Matrix representations for positive non-commutative polynomials*, Positivity, to appear; also available from http://www.math.ufl.edu/∼sam/xandh.dvi(ps)(pdf).

[K36] F. Kraus, *Über Konvexe Matrixfunktionen*, Math. Z., 41 (1936), pp. 18–42.

[LM00] C.-K. Li and R. Mathias, *Extremal characterizations of the Schur complement and resulting inequalities*, SIAM Rev., 42 (2000), pp. 233–246.

[L34] C. Löwner, *Über Monotone Matrixfunktionen*, Math. Z., 38 (1934), pp. 177–216.

[M01] S. McCullough, *Factorization of operator-valued polynomials in several noncommuting variables*, Linear Algebra Appl., 326 (2001), pp. 193–203.

[MPprept] S. McCullough and M. Putinar, *Non-commutative sums of squares*, Pacific J. Math., to appear; also available from http://www.math.ufl.edu/∼sam/sos.dvi(ps)(pdf).

[PV99] M. Putinar and F.-H. Vasilescu, *Solving moment problems by dimensional extension*, Ann. of Math. (2), 149 (1999), pp. 1087–1107.

[R00] B. Reznick, *Some concrete aspects of Hilbert's* 17*th problem*, in Real Algebraic Geometry and Ordered Structures (Baton Rouge, LA, 1996), Contemp. Math. 253, C. Delzell and J. Madden, eds., AMS, Providence, RI, 2000, pp. 251–272.

[RS79] M. H. Rizvi and R. W. Shorrock, *A note on matrix-convexity*, Canad. J. Statist., 7 (1979), pp. 39–41.

[SIG97] R. E. Skelton, T. Iwasaki, and K. M. Grigoriadis, *A Unified Algebraic Approach to Linear Control Design*, Taylor & Francis, London, 1997.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1140–1162

H-UNITARY AND LORENTZ MATRICES: A REVIEW*

YIK-HOI AU-YEUNG†, CHI-KWONG LI‡, AND LEIBA RODMAN‡

Dedicated to Professor Yung-Chow Wong, Emeritus Professor of The University of Hong Kong, on his 90th birthday

Abstract. Many properties of H-unitary and Lorentz matrices are derived using elementary methods. Complex matrices that are unitary with respect to the indefinite inner product induced by an invertible Hermitian matrix H are called H-unitary, and real matrices that are orthogonal with respect to the indefinite inner product induced by an invertible real symmetric matrix are called Lorentz. The focus is on the analogues of singular value and CS (cos – sin) decompositions for general H-unitary and Lorentz matrices, and on the analogues of Jordan form, in a suitable basis with certain orthonormality properties, for diagonalizable H-unitary and Lorentz matrices. Several applications are given, including connected components of Lorentz similarity orbits, products of matrices that are simultaneously positive definite and H-unitary, products of reflections, and stability and robust stability.

Key words. Lorentz matrices, indefinite inner product

AMS subject classification. 15A63

DOI. 10.1137/S0895479803421896

1. Introduction.

Let $M_n = M_n(\mathbb{F})$ be the algebra of $n \times n$ matrices with entries in the field $\mathbb{F} = \mathbb{C}$, the complex numbers, or $\mathbb{F} = \mathbb{R}$, the real numbers. If $H \in M_n$ is an invertible Hermitian (symmetric in the real case) matrix, a matrix $A \in M_n$ is called *H-unitary* if $A^*HA = H$.

The literature on the subject is voluminous, starting with the invention of non-Euclidean geometry in the 19th century; in the 20th century, besides being of considerable theoretical mathematical interest, studies of H-unitary matrices were motivated by applications in physics, in particular, in relativity theory, and later by applications in electrical engineering, where functions with values in the group of H-unitary matrices play a significant role. Without attempting to give a literature guide on the subject, which would take us too far afield, we indicate an early influential source [1] and books [19], [9], [2], and [17, Chapter 2], where H-unitary matrices are treated in considerable depth from the point of view of the theory of matrices, in contrast with the point of view of Lie group theory in many other sources. For applications of H-unitary valued functions in engineering and interpolation, see, e.g., the books [16] and [3] and for an exposition from the point of view of numerical methods see the recent review [14].

In this paper we present several canonical forms of H-unitary matrices and demonstrate some of their applications. The exposition is kept purposely on an elementary level, but at the same time is self-contained (with few exceptions), to make the article accessible to a large audience. Thus, occasionally results are stated and proved not in

*Received by the editors April 10, 2003; accepted for publication (in revised form) by M. L. Overton August 4, 2003; published electronically July 14, 2004.

http://www.siam.org/journals/simax/25-4/42189.html

†P. O. Box 16065, Stanford, CA 94309 (tauyeun@muse.sfusd.edu). This author is currently an honorary research fellow with the University of Hong Kong.

‡Department of Mathematics, College of William and Mary, P. O. Box 8795, Williamsburg, VA 23187-8795 (ckli@math.wm.edu, lxrodm@math.wm.edu). The research of the second author was partially supported by NSF grant DMS-0071944. The research of the third author was partially supported by NSF grant DMS-9988579.

the most generally known form. Many results in this paper are known, in which case we provide short transparent proofs. Hopefully, this will give a gentle introduction on the subject to beginners and researchers in fields other than matrix theory.

To avoid the well-known cases of unitary or real orthogonal matrices, we assume throughout that H is indefinite. In our discussion, we often assume that $H = J := I_p \oplus -I_q$ for some positive integers p and q with $p+q = n$. There is no harm in doing so because of the following observation.

Observation 1.1. If $S \in M_n$ is invertible, then $A \in M_n$ is H-unitary if and only if $S^{-1}AS$ is S^*HS-unitary.

In the real case, a matrix A is often called *Lorentz* if it is J-unitary. We will use the terminology "J-unitary" instead of "Lorentz" for convenience.

The following notation will be used in the paper.

$M_{p\times q} = M_{p\times q}(\mathbb{F})$: the $\mathbb{F}$-vector space of $p \times q$ matrices with entries in $\mathbb{F}$;

A^*: the conjugate transpose of $A \in M_{p\times q}$; it reduces to the transpose A^t of A in the real case;

$\mathrm{Spec}\,(A)$: the spectrum of a matrix A;

$\mathrm{diag}\,(X_1, \ldots, X_r) = X_1 \oplus \cdots \oplus X_r$: the block diagonal matrix with diagonal blocks $X_1, \ldots, X_r$ (in the given order);

$\sqrt{A}$: the unique positive definite square root of a positive definite matrix A;

I_p: the $p \times p$ identity matrix;

$[x, y] = y^*Hx$: the indefinite inner product induced by H;

$\mathcal{U}_{\mathbb{F}}^H$: the group of all H-unitary matrices with entries in $\mathbb{F}$.

On several occasions we will use the identification of the complex field as a subalgebra of real 2×2 matrices:

$$x + iy \in \mathbb{C},\ x, y \in \mathbb{R} \quad \longleftrightarrow \quad \begin{pmatrix} x & y \\ -y & x \end{pmatrix} \in M_2(\mathbb{R}). \tag{1.1}$$

2. CS decomposition. In this section we let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, and $J = I_p \oplus -I_q$. Let $\mathcal{U}_n$ be the unitary group in M_n, and let $\mathcal{U}(p, q)$ be the group of matrices $U_1 \oplus U_2$ such that $U_1 \in \mathcal{U}_p$ and $U_2 \in \mathcal{U}_q$.

Observation 2.1. A matrix $A \in M_n$ is J-unitary if and only if UAV is J-unitary for any/all $U, V \in \mathcal{U}(p, q)$.

The following lemma is useful (its verification is straightforward).

LEMMA 2.2. *A matrix* $\begin{pmatrix} \sqrt{I_p + MM^*} & M^* \\ M^* & \sqrt{I_q + M^*M} \end{pmatrix}$ *is J-unitary, as well as positive definite, for every $p \times q$ matrix M.*

For any (usual) unitary matrix $A \in M_n$, there are matrices $X, Y \in \mathcal{U}(p, q)$ such that

$$XAY = I_r \oplus \begin{pmatrix} C & S \\ -S^t & C \end{pmatrix} \oplus I_s,$$

where $C, S \in M_{p-r}$ are diagonal matrices with positive diagonal entries satisfying $C^2 + S^2 = I_{p-r}$. This is known as the CS ($\cos - \sin$) decomposition of A, see, e.g., [13, p. 78]. We have the following analogous CS ($\cosh - \sinh$) decomposition theorem for a J-unitary matrix.

THEOREM 2.3. *A matrix $A \in M_n$ is J-unitary if and only if there exist $X, Y \in \mathcal{U}(p, q)$ and a $p \times q$ matrix $D = [d_{ij}]$, where $d_{11} \geq \cdots \geq d_{mm} > 0$ for some $m \leq \min\{p, q\}$ and all other entries of D are zero, such that*

$$XAY = \begin{pmatrix} \sqrt{I_p + DD^t} & D \\ D^t & \sqrt{I_q + D^tD} \end{pmatrix}. \tag{2.1}$$

Moreover, the matrix D is uniquely determined by A.

Proof. Let $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ such that $A_{11} \in M_p$ and $A_{22} \in M_q$. Suppose $U_1, V_1 \in \mathcal{U}_p$ and $U_2, V_2 \in \mathcal{U}_q$ are such that

$$U_1 A_{12} V_2 = D_1 = \begin{pmatrix} \tilde{D}_1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad U_2 A_{21} V_1 = D_2 = \begin{pmatrix} \tilde{D}_2 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\tilde{D}_1$ and $\tilde{D}_2$ are diagonal matrices with positive diagonal entries arranged in nonincreasing order. Let $U = U_1 \oplus U_2, V = V_1 \oplus V_2 \in \mathcal{U}(p,q)$. Then

$$B = UAV = \begin{pmatrix} R & D_1 \\ D_2 & S \end{pmatrix}.$$

Since $B^* J B = J$, we see that

$$R^*R - D_2^*D_2 = I_p, \quad D_1^*D_1 - S^*S = -I_q, \quad R^*D_1 - D_2^*S = 0.$$

Since $B^* J B J = I_n$, we have $BJB^* = J$, and hence

$$RR^* - D_1D_1^* = I_p, \quad D_2D_2^* - SS^* = -I_q, \quad RD_2^* - D_1S^* = 0.$$

Note that $R^*R = I_p + D_2^*D_2$ and $RR^* = I_p + D_1D_1^*$ have the same eigenvalues, and thus,

$$\tilde{D}_1 = \tilde{D}_2 = d_1 I_{m_1} \oplus \cdots \oplus d_r I_{m_r}$$

for some $d_1 > d_2 > \cdots > d_r > 0$ and positive integers $m_1, \ldots, m_r$. Furthermore, $RR^* = I_p + D_1D_1^*$ implies that R has orthogonal rows with lengths that equal the singular values of R, arranged in nonincreasing order; $R^*R = I_p + D_2^*D_2$ implies that R has orthogonal columns with lengths that equal the singular values of R, arranged in nonincreasing order. As a result,

$$R = \sqrt{1+d_1^2}X_1 \oplus \cdots \oplus \sqrt{1+d_r^2}X_r \oplus X_{r+1},$$

where $X_j \in \mathcal{U}_{m_j}$ for $j = 1, \ldots, r$ and $X_{r+1} \in \mathcal{U}_{p-m}$ with $m = m_1 + \cdots + m_r$. Similarly, one can show that

$$S = \sqrt{1+d_1^2}Y_1 \oplus \cdots \oplus \sqrt{1+d_r^2}Y_r \oplus Y_{r+1},$$

where $Y_j \in \mathcal{U}_{m_j}$ for $j = 1, \ldots, r$ and $Y_{r+1} \in \mathcal{U}_{q-m}$. Suppose

$$Z = X_1 \oplus \cdots \oplus X_r \oplus X_{r+1} \oplus Y_1 \oplus \cdots \oplus Y_r \oplus Y_{r+1} \in \mathcal{U}(p,q),$$

$X = Z^*U$ and $Y = V$. Then XAY has the asserted form.

The uniqueness of D follows from (2.1), because $\sqrt{1+d_{jj}^2} \pm d_{jj}$, $j = 1, \ldots, m$, are the singular values of A different from 1. ∎

A different proof (using the exchange operator, in the terminology of [14]) of Theorem 2.3 is given in [14]. The proof of the above theorem uses only the elementary facts: (a) every rectangular matrix has a singular value decomposition, (b) XY and YX have the same eigenvalues for any $X, Y \in M_n$, (c) $Z \in M_n$ has orthogonal rows and columns with lengths arranged in nonincreasing size if and only if Z is a direct sum of multiples of unitary (if $\mathbb{F} = \mathbb{C}$) or real orthogonal (if $\mathbb{F} = \mathbb{R}$) matrices. Yet, we can deduce other known canonical forms, which have been obtained by more sophisticated techniques involving Lie theory, functions and power series of matrices, etc.

THEOREM 2.4. *Let $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right) \in M_n$ be such that $A_{11} \in M_p$ and $A_{22} \in M_q$. Then A is J-unitary if and only if any one of the following four conditions holds true.*

(a) *There are* $U \in \mathcal{U}(p,q)$ *and a* $p \times q$ *matrix* M *such that*

$$UA = \begin{pmatrix} \sqrt{I_p + MM^*} & M \\ M^* & \sqrt{I_q + M^*M} \end{pmatrix}.$$

(b) *There are* $U \in \mathcal{U}(p,q)$ *and a* $p \times q$ *matrix* L *such that*

$$UA = (I_p \oplus iI_q) \exp\left(i \begin{pmatrix} 0_p & L \\ -L^* & 0_q \end{pmatrix} \right) (I_p \oplus iI_q)^* = \exp \begin{pmatrix} 0_p & L \\ L^* & 0_q \end{pmatrix}.$$

(c) *There are* $U \in \mathcal{U}(p,q)$ *and a* $p \times q$ *matrix* K *with all singular values less than* 1 *such that*

$$UA = \begin{pmatrix} I_p & K \\ K^* & I_q \end{pmatrix} \begin{pmatrix} I_p & -K \\ -K^* & I_q \end{pmatrix}^{-1}.$$

(d) *Setting* $X = A_{11}(I_p + A_{21}^*A_{21})^{-1/2}$ *and* $Y = A_{22}(I_q + A_{12}^*A_{12})^{-1/2}$, *we have*

$$X \in \mathcal{U}_p, \quad Y \in \mathcal{U}_q, \quad \text{and} \quad X^*A_{12} = A_{21}^*Y.$$

Clearly, one can write analogous conditions with U on the right, with the same right-hand sides as in (a), (b), and (c), and get special forms for AU. We omit the statements.

Note that the formula in (a) is a *polar decomposition* of A. It follows in particular that both factors in the polar decomposition of a J-unitary matrix are also J-unitary, a well-known fact in Lie theory. Thus, the matrices U and M in (a) are determined uniquely. Similarly, the matrices on the right sides in conditions (b) and (c) are different representations of the positive definite part of A, and are also uniquely determined.

Proof. By Theorem 2.3, $A \in M_n$ is J-unitary if and only if there are $X, Y \in \mathcal{U}(p,q)$ satisfying (2.1). Setting $U = YX$, we get the equivalent condition (a) in view of Lemma 2.2.

To prove the equivalent condition (b), suppose A is J-unitary, and XAY has the form (2.1). Note that

$$\begin{pmatrix} \sqrt{1+d_j^2} & d_j \\ d_j & \sqrt{1+d_j^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \exp\left(i \begin{pmatrix} 0 & \ell_j \\ -\ell_j & 0 \end{pmatrix} \right) \begin{pmatrix} 1 & 0 \\ 0 & -i \end{pmatrix},$$

where $\sinh \ell_j = d_j$. Hence,

$$XAY = (I_p \oplus iI_q) \exp\left(i \begin{pmatrix} 0 & \tilde{L} \\ -\tilde{L}^t & 0 \end{pmatrix} \right) (I_p \oplus iI_q)^* = \exp \begin{pmatrix} 0_p & \tilde{L} \\ \tilde{L}^t & 0_q \end{pmatrix},$$

where $\tilde{L}$ is the $p \times q$ matrix with ℓ_j at the (j,j) entry whenever $d_j > 0$, and zeros elsewhere. Let $U = YX$ and

$$\begin{pmatrix} 0 & L \\ -L^* & 0 \end{pmatrix} = Y \begin{pmatrix} 0 & \tilde{L} \\ -\tilde{L}^t & 0 \end{pmatrix} Y^*.$$

We get condition (b). Conversely, suppose (b) holds. Consider a singular value decomposition of $L = V_1^* \tilde{L} V_2$, where $V_1 \in \mathcal{U}_p$ and $V_2 \in \mathcal{U}_q$. Let $V = V_1 \oplus V_2 \in \mathcal{U}(p,q)$. Then

$$(VU)AV^* = \exp \begin{pmatrix} 0 & \tilde{L} \\ \tilde{L}^t & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{I_p + DD^t} & D \\ D^t & \sqrt{I_q + D^tD} \end{pmatrix}$$

has the form (2.1) with $X = VU$ and $Y = V^*$.

Next, we turn to the equivalent condition (c). Suppose A is J-unitary, and XAY has the form (2.1). Note that

$$\begin{pmatrix} \sqrt{1+d_j^2} & d_j \\ d_j & \sqrt{1+d_j^2} \end{pmatrix} = \begin{pmatrix} 1 & k_j \\ k_j & 1 \end{pmatrix} \begin{pmatrix} 1 & -k_j \\ -k_j & 1 \end{pmatrix}^{-1},$$

where $k_j \in (0,1)$, satisfying $2k_j/(1-k_j^2) = d_j$. Hence,

$$XAY = \begin{pmatrix} I_p & \tilde{K} \\ \tilde{K}^t & I_q \end{pmatrix} \begin{pmatrix} I_p & -\tilde{K} \\ -\tilde{K}^t & I_q \end{pmatrix}^{-1},$$

where $\tilde{K}$ is the $p \times q$ matrix with k_j at the (j,j) entry whenever $d_j > 0$, and zeros elsewhere. Let $U = YX$ and

$$\begin{pmatrix} 0 & K \\ K^* & 0 \end{pmatrix} = Y \begin{pmatrix} 0 & \tilde{K} \\ \tilde{K}^t & 0 \end{pmatrix} Y^*.$$

We get condition (b). Conversely, suppose (c) holds. Putting $M = 2(I_p - KK^*)^{-1}K$, we see that (c) implies (a). Thus, A is J-unitary.

Finally, we consider the equivalent condition (d). Suppose A is J-unitary and condition (a) holds with $U = U_1 \oplus U_2$, where $U_1 \in \mathcal{U}_p$ and $U_2 \in \mathcal{U}_q$; i.e.,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} U_1^* & 0 \\ 0 & U_2^* \end{pmatrix} \begin{pmatrix} \sqrt{I_p + MM^*} & M \\ M^* & \sqrt{I_q + M^*M} \end{pmatrix}.$$

Then

$$X = A_{11}(I_p + A_{21}^*A_{21})^{-1/2} = U_1^*\sqrt{I_p + MM^*}(I_p + MM^*)^{-1/2} = U_1^*,$$

$$Y = A_{22}(I_q + A_{12}^*A_{12})^{-1/2} = U_2^*\sqrt{I_q + M^*M}(I_p + M^*M)^{-1/2} = U_2^*,$$

and

$$X^*A_{12} = U_1(U_1^*M) = M = (MU_2)U_2^* = A_{21}^*Y.$$

Thus, condition (d) holds. Conversely, suppose condition (d) holds. Putting $M = X^*A_{12}$, we see that condition (a) holds with $U = X^* \oplus Y^* \in \mathcal{U}(p,q)$. So, A is J-unitary. ∎

Recall (see [22]) that the rows of $(X_0, Y_0) \in M_{r\times p} \times M_{r\times q}$ (respectively, $(X_1 \;\; Y_1) \in M_{r\times p} \times M_{r\times q}$) are *initial vectors* (respectively, *final vectors*) of the J-unitary matrix A if

$$(X_0 \,|\, Y_0)A = (X_1 \,|\, Y_1).$$

There is an interest in determining a J-unitary matrix in terms of its initial and final vectors; see [22]. In this connection, we can use Theorem 2.4 (c) to get the following corollary.

COROLLARY 2.5. *Suppose A is a J-unitary matrix expressed as in Theorem* 2.4 (c) *with $U = U_1 \oplus U_2$. Let $(X_0 \,|\, Y_0), (X_1 \,|\, Y_1) \in M_{r\times p} \times M_{r\times q}$ be initial vectors and final vectors of A, respectively. Then*

$$X_0U_1^*(I_p - KK^*) = X_1(I_p + KK^*) - 2Y_1K^*$$

and

$$Y_0U_2^*(I_q - K^*K) = Y_1(I_q + K^*K) - 2X_1K.$$

In particular, if $r = p$ *and* $\det(X_0) \neq 0$*, then*

$$X_0^{-1}[X_1(I_p + KK^*) - 2Y_1K^*]$$

is a constant matrix, i.e., independent of $(X_0 \,|\, Y_0)$ *and* $(X_1 \,|\, Y_1)$*. Similarly, if* $r = q$ *and* $\det(Y_0) \neq 0$*, then*

$$Y_0^{-1}[Y_1(I_q + K^*K) - 2X_1K]$$

is a constant matrix.

By the canonical form in Theorem 2.4 (a), we have the following.

COROLLARY 2.6. *The group of* J*-unitary matrices is homeomorphic to* $\mathcal{U}_p \times \mathcal{U}_q \times \mathbb{F}^{pq}$*. In the real case, it is a real analytic manifold consisting of four arcwise connected components, and the identity component is locally isomorphic to* $\mathbb{R}^{p(p-1)/2} \times \mathbb{R}^{q(q-1)/2} \times \mathbb{R}^{pq}$*. In the complex case, it is a real analytic manifold consisting of one arcwise connected component which is locally isomorphic to* $\mathbb{R}^{p^2} \times \mathbb{R}^{q^2} \times \mathbb{R}^{2pq}$*.*

COROLLARY 2.7. *The group* $\mathcal{U}(p,q)$ *is a maximal bounded subgroup of* $\mathcal{U}_{\mathbb{F}}^J$*.*

Proof. The group $\mathcal{U}(p,q)$ is clearly bounded. Let $\mathcal{G}$ be a subgroup of $\mathcal{U}_{\mathbb{F}}^J$ that strictly contains $\mathcal{U}(p,q)$, and let $A \in \mathcal{G} \setminus \mathcal{U}(p,q)$. Then in the representation (2.1) of A, we clearly have $D \neq 0$, and

$$\begin{pmatrix} \sqrt{I + DD^t} & D \\ D^t & \sqrt{I + D^tD} \end{pmatrix} \in \mathcal{G}.$$

But since $D \neq 0$, the cyclic subgroup generated by $\left(\begin{smallmatrix} \sqrt{I+DD^t} & D \\ D^t & \sqrt{I+D^tD} \end{smallmatrix}\right)$ is not bounded. ∎

In connection with Corollary 2.7 observe that there exist bounded cyclic subgroups of $\mathcal{U}_{\mathbb{F}}^J$ which are not contained in $\mathcal{U}(p,q)$ (see Theorem 4.9).

Suppose $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right) \in M_n$ is a J-unitary matrix with $A_{11} \in M_p$ and $A_{22} \in M_q$. By Theorem 2.3, A_{11} and A_{22} are invertible. For $\mathbb{F} = \mathbb{R}$, define

$$\sigma_+(A) = \begin{cases} 1 & \text{if } \det A_{11} > 0, \\ -1 & \text{if } \det A_{11} < 0, \end{cases} \qquad \sigma_-(A) = \begin{cases} 1 & \text{if } \det A_{22} > 0, \\ -1 & \text{if } \det A_{22} < 0. \end{cases} \tag{2.2}$$

We can use Theorem 2.4 (a) to deduce the following corollary (see, e.g., [8]).

COROLLARY 2.8 ($\mathbb{F} = \mathbb{R}$). *For any* J*-unitary matrices* $A, B \in M_n(\mathbb{R})$*,*

$$\sigma_+(A)\sigma_+(B) = \sigma_+(AB) \qquad \text{and} \qquad \sigma_-(A)\sigma_-(B) = \sigma_-(AB).$$

Proof. From Theorem 2.4 (a),

$$A = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \sqrt{I_p + MM^t} & M \\ M^t & \sqrt{I_q + M^tM} \end{pmatrix},$$

$$B = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} \begin{pmatrix} \sqrt{I_p + NN^t} & N \\ N^t & \sqrt{I_q + N^tN} \end{pmatrix},$$

for some real $p \times q$ matrices M and N. Thus, $\sigma_+(A) = \det U_1$, $\sigma_+(B) = \det V_1$, and

$$\sigma_+(AB) = \text{sign}\,\{\det(U_1\sqrt{I_p + MM^t}V_1\sqrt{I_p + NN^t} + U_1MV_1N^t)\}. \tag{2.3}$$

Let $M_x = xM$, $N_x = xN$, $0 \le x \le 1$. By the comment before the corollary and Theorem 2.4 (a), the matrix

$$W_x := U_1\sqrt{I_p + M_xM_x^t}V_1\sqrt{I_p + N_xN_x^t} + U_1M_xV_2N_x^t$$

is invertible for every $x \in [0, 1]$. Therefore, the sign of the determinant of W_x does not depend on x, and we have

$$\begin{aligned}\text{sign}\,\{\det W_1\} = \text{sign}\,\{\det W_0\} &= \text{sign}\,\{\det(U_1V_1)\} \\ &= \text{sign}\,\{\det U_1\}\ \text{sign}\,\{\det V_1\} = \sigma_+(A)\sigma_+(B).\end{aligned}$$

In view of (2.3) the result for σ_+ follows. The proof for σ_- is similar. ∎

We conclude this section with the following well-known result that H-unitary matrices can be obtained by applying linear fractional transforms to H-skewadjoint matrices. A matrix $K \in M_n$ is called *H-skewadjoint* if $[Kx, y] = -[x, Ky]$ for every $x, y, \in \mathbb{F}^n$, i.e., $HK = -K^*H$. Here we just assume that $H \in M_n$ is an invertible Hermitian (symmetric in the real case) matrix.

PROPOSITION 2.9. *Suppose A is H-unitary, and $\mu, \xi \in \mathbb{F}$ satisfy $|\mu| = 1$ with $\det(A - \mu I) \neq 0$ and $-\bar{\xi} \neq \xi$ if $\mathbb{F} = \mathbb{C}$. Then the matrix $K = (\xi A + \mu\bar{\xi}I)(A - \mu I)^{-1}$ is H-skewadjoint such that $\det(K - \xi I) \neq 0$. Conversely, suppose $K \in M_n$ is H-skewadjoint, and $\mu, \xi \in \mathbb{F}$ satisfy $|\mu| = 1$, $\det(K - \xi I) \neq 0$, and $-\bar{\xi} \neq \xi$ if $\mathbb{F} = \mathbb{C}$. Then $A = \mu(K + \bar{\xi}I)(K - \xi I)^{-1}$ is H-unitary such that $\det(A - \mu I) \neq 0$.*

For a proof see, e.g., [4, pp. 38–39] or [9]; the proposition is also easy to verify directly using algebraic manipulations.

3. Diagonalizable H-unitary matrices. In this section we assume that $H = H^* \in M_n(\mathbb{F})$ is indefinite and invertible but not necessarily equal to J, as in the previous section.

Evidently, A is H-unitary if and only if $S^{-1}AS$ is S^*HS-unitary, for any invertible matrix S. In the complex case, a canonical form under this transformation is described in [11] and [12] (see also [9]); other canonical forms in both real and complex cases are given in [20]. We will not present these forms in full generality and consider in sections 3.1 and 3.2 only the diagonalizable cases, which will suffice for the applications presented in later sections.

Let $J_j(\lambda)$ denote the upper triangular $j \times j$ Jordan block with eigenvalue λ. In the real case, we let $J_{2k}(\lambda \pm i\mu)$ be the almost upper triangular $2k \times 2k$ real Jordan block with a pair of nonreal complex conjugate eigenvalues $\lambda \pm i\mu$ (here λ and μ are real and $\mu \neq 0$) as follows:

$$J_{2k}(\lambda \pm i\mu) = \begin{pmatrix} J_2(\lambda \pm i\mu) & I_2 & 0_2 & \dots & 0_2 \\ 0_2 & J_2(\lambda \pm i\mu) & I_2 & \dots & 0_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_2 & 0_2 & 0_2 & \dots & J_2(\lambda \pm i\mu) \end{pmatrix},$$

$$J_2(\lambda \pm i\mu) = \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}.$$

We also use the following notation: G_j is the $j \times j$ matrix with 1's on the top-right, bottom-left diagonal, and zeros in all other positions.

PROPOSITION 3.1 ($\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$). *A matrix $A \in M_n(\mathbb{F})$ is H-unitary for some invertible Hermitian matrix $H \in M_n(\mathbb{F})$ if and only if A is invertible and similar (over $\mathbb{F}$) to $(A^{-1})^*$, and the following condition holds in the real case: Each Jordan block of eigenvalue 1 having even size (if it exists) appears an even number of times in the Jordan form of A, and each Jordan block of eigenvalue -1 having even size (if it exists) appears an even number of times in the Jordan form of A.*

Proof. Consider the complex case first. "Only if" is clear: $A^*HA = H$ implies $(A^{-1})^* = HAH^{-1}$. For the "if" part, observe that without loss of generality (using the transformation $(A, H) \mapsto (S^{-1}AS, S^*HS)$ for a suitable invertible S), we may assume that A is in the Jordan form. Considering separately every Jordan block of A with a unimodular eigenvalue, and collecting together every pair of Jordan blocks of equal size with eigenvalues λ and μ such that $\lambda\overline{\mu} = 1$, we reduce the proof to two cases:

(i) $A = J_j(\lambda)$, $|\lambda| = 1$;

(ii) $A = J_j(\lambda) \oplus J_j(\mu)$, $\lambda\overline{\mu} = 1$, $|\lambda| \neq 1$.

In case (ii), by making a similarity transformation, we may assume that $A = J_j(\lambda) \oplus (\overline{J_j(\lambda)})^{-1}$; then a calculation shows that A is G_{2j}-unitary. In case (i), by making a similarity transformation, assume (cf. [9, section 2.3])

$$A = \lambda \begin{pmatrix} 1 & 2i & 2i^2 & \dots & 2i^{j-1} \\ 0 & 1 & 2i & \dots & 2i^{j-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix};$$

then A is G_j-unitary.

Now let $\mathbb{F} = \mathbb{R}$. Consider the "if" part. As in the complex case, we may assume that A is in the real Jordan form (see [10, Chapter 12]), and furthermore that A has one of the following four forms:

(a) $A = J_m(1)$ or $A = J_m(-1)$, where m is odd;

(b) $A = J_{2k}(\lambda \pm i\mu)$, where $\lambda^2 + \mu^2 = 1$, $\mu > 0$;

(c) $A = J_k(\lambda) \oplus J_k(\lambda^{-1})$, where λ is real and $|\lambda| \geq 1$, and in cases $\lambda = \pm 1$ the size k is even;

(d) $A = J_k(\lambda \pm i\mu) \oplus J_k(\lambda' \pm i\mu')$, where k is even, $\lambda^2 + \mu^2 > 1$ and $\lambda' + i\mu' = (\lambda - i\mu)^{-1}$.

In the cases (c) and (d) A is similar to a matrix of the form $\left(\begin{smallmatrix} B & 0 \\ 0 & (B^t)^{-1} \end{smallmatrix}\right)$, where $B \in M_k(\mathbb{R})$, and the matrix $\left(\begin{smallmatrix} B & 0 \\ 0 & (B^t)^{-1} \end{smallmatrix}\right)$ is $\left(\begin{smallmatrix} 0 & I_k \\ I_k & 0 \end{smallmatrix}\right)$-unitary, as one verifies easily. The case (b) is reduced to the already proven complex case by using the identification (1.1).

Consider the case (a). The case when $A = J_m(-1)$ is easily reduced, by replacing A with $-A$, and by making a similarity transformation, to the case when $A = J_m(1)$; thus, we assume $A = J_m(1)$. The matrix $K = (I - A)(I + A)^{-1}$ has the Jordan form $J_m(0)$ and by Proposition 2.9 is H-skewadjoint if and only if A is H-unitary. Thus, it suffices to find an invertible real symmetric H such that $J_m(0)$ is H-skewadjoint, i.e., $HJ_m(0) = -(J_m(0))^t H$. One such H is given by $H = [h_{j,k}]_{j,k=1}^m \in M_m(\mathbb{R})$ with the entries $h_{j,m+1-j} = (-1)^{j+1}$, $j = 1, \dots, m$, and all other entries being zero.

We now prove the "only if" part in the real case. Let $A \in M_n(\mathbb{R})$ be H-unitary, and assume first that one, but not both, of the numbers 1 and -1 is an eigenvalue of A (if $1, -1 \notin \operatorname{Spec}(A)$, we are done.) Say, $-1 \notin \operatorname{Spec}(A)$. By Proposition 2.9, the matrix $K = (I - A)(I + A)^{-1}$ is H-skewadjoint. Since the derivative of the

function $f(z) = \frac{1-z}{1+z}$ is $f'(z) = \frac{-2}{(1+z)^2}$, which is nonzero for $z \in \mathrm{Spec}\,(A)$, the calculus of functions of the matrix A (which can be found in many graduate texts on linear algebra; see, for example, [18] or [15, Chapter 6]) shows that the Jordan blocks of K with eigenvalue 0 of K have sizes equal to the sizes of corresponding Jordan blocks of A with eigenvalue 1. Now $K = H^{-1}(HK)$ is a product of an invertible symmetric matrix H^{-1} and a skewsymmetric matrix HK, and therefore every nilpotent Jordan block of even size in the Jordan form of K appears an even number of times (see [21, Lemma 2.2]). Hence the same property holds for the Jordan blocks of A corresponding to the eigenvalue 1.

We leave aside the more difficult case when both 1 and -1 are eigenvalues of A. This case can be dealt with using the proof of [6, Theorem 9, section I.5], upon replacing there the operation of transposition $A \mapsto A^t$ by the operation of H-adjoint: $A \mapsto H^{-1}A^tH$, and making use of the already mentioned fact that every nilpotent Jordan block of even size in the Jordan form of an H-skewadjoint matrix appears an even number of times. All arguments in the proof go through, and we omit the details. □

Thus, in the complex case, if λ is an eigenvalue of an H-unitary matrix A, then so is $\overline{\lambda}^{-1}$, with the same algebraic and geometric multiplicities as λ; a similar statement applies to pairs of complex conjugate eigenvalues of real H-unitary matrices.

3.1. The complex case. We assume $\mathbb{F} = \mathbb{C}$ in this subsection. Denote by $\mathcal{R}_\lambda(A) = \mathrm{Ker}\,(A - \lambda I)^n$ the root subspace corresponding to the eigenvalue λ of an $n \times n$ matrix A. We need orthogonality properties of the root subspaces and certain eigenvectors.

LEMMA 3.2. *Let A be H-unitary.*

(a) *If $v \in \mathcal{R}_\lambda(A)$, $w \in \mathcal{R}_\mu(A)$, where $\lambda\overline{\mu} \neq 1$, then v and w are H-orthogonal:*

$$[v, w] = 0. \tag{3.1}$$

(b) *If x is an eigenvector of A corresponding to the eigenvalue λ, and if* either $|\lambda| \neq 1$ or $|\lambda| = 1$ *and $(A - \lambda I)y = x$ for some vector y, then $[x, x] = 0$.*

Proof. (a) We have

$$(A - \lambda I)^p v = 0, \qquad (A - \mu I)^q w = 0, \tag{3.2}$$

for some positive integers p and q. We prove (3.1) by induction on $p + q$ (see [11, Lemma 3]). If $p = q = 1$, then $(A - \lambda I)v = (A - \mu I)w = 0$, and therefore

$$\lambda\overline{\mu}[v, w] = [\lambda v, \mu w] = [Av, Aw] = [v, w],$$

which implies (in view of $\lambda\overline{\mu} \neq 1$) that $[v, w] = 0$. Assume now (3.2) holds, and assume that $[v', w'] = 0$ for v', w' satisfying (3.2) with smaller values of $p + q$. We let $v' = (A - \lambda I)v$, $w' = (A - \mu I)w$, and then

$$\begin{aligned}
\lambda\overline{\mu}[v, w] &= [Av - v', Aw - w'] \\
&= [Av, Aw] - [v', Aw] - [Av, w'] + [v', w'] \\
&= [Av, Aw] - [v', w'] - [v', \mu w] - [v', w'] - [\lambda v, w'] + [v', w'] \\
&= [Av, Aw] = [v, w],
\end{aligned}$$

where the first equality on the fourth line follows by the induction hypothesis. So, the desired conclusion $[v, w] = 0$ is obtained.

Part (b) under the hypothesis that $|\lambda| \neq 1$ follows from (a) (take $\mu = \lambda$). Assume now

$$(A - \lambda I)x = 0, \quad (A - \lambda I)y = x, \quad x \neq 0, \quad |\lambda| = 1.$$

Arguing by contradiction, suppose that $[x, x] \neq 0$. Then, adding to y a suitable multiple of x, we may assume without loss of generality that $[y, x] = 0$. Now

$$\begin{aligned}
[x, x] &= y^*(A - \lambda I)^* H(A - \lambda I)y \\
&= y^*(A^*HA - \overline{\lambda}HA - \lambda A^*H + H)y \quad (\text{using } A^*HA = H) \\
&= y^*(H - \overline{\lambda}HA - \lambda A^*H + H)y \\
&= -\overline{\lambda}y^*H(A - \lambda I)y - \lambda y^*(A - \lambda I)^*Hy \\
&= -\overline{\lambda}y^*Hx - \lambda x^*Hy \\
&= 0,
\end{aligned}$$

a contradiction. ∎

THEOREM 3.3. *A diagonalizable matrix $A \in M_n(\mathbb{C})$ is H-unitary if and only if there exists an invertible matrix S such that $(S^{-1}AS, S^*HS)$ equals*

$$\left(U_1 \oplus \ldots \oplus U_m \oplus U_{m+1} \oplus \ldots \oplus U_{m+q}, \epsilon_1 \oplus \ldots \oplus \epsilon_m \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right), \tag{3.3}$$

where $\epsilon_j = \pm 1$ $(j = 1, \ldots, m)$, the complex numbers U_j for $j = 1, \ldots, m$ are unimodular, and the 2×2 matrices U_j for $j = m+1, \ldots, m+q$ are of the form $U_j = \left(\begin{smallmatrix} \lambda_j & 0 \\ 0 & (\overline{\lambda_j})^{-1} \end{smallmatrix}\right)$, $|\lambda_j| \neq 1$.

Moreover, the representation of an H-unitary matrix A as in the right-hand side of (3.3) *is unique up to a simultaneous permutation of pairs (U_j, ϵ_j), $j = 1, \ldots, m$, and up to a permutation of blocks $U_{m+1}, \ldots, U_{m+q}$.*

Proof. The part "if" being obvious, consider the "only if" part. In view of Lemma 3.2, we need to consider only the case when A has either only one (possibly of high multiplicity) unimodular eigenvalue λ, or only one pair $\lambda, (\overline{\lambda})^{-1}$ of nonunimodular eigenvalues (again, possibly of high multiplicity). Since A is diagonalizable, in the first case $A = \lambda I$, and using a congruence transformation $H \mapsto S^*HS$ we put H in the diagonal form, as required. In the second case, we may assume $A = \lambda I \oplus (\overline{\lambda})^{-1}I$, and (by Lemma 3.2) $H = \left(\begin{smallmatrix} 0 & Q \\ Q^* & 0 \end{smallmatrix}\right)$ for some (necessarily invertible) matrix Q. A transformation

$$\begin{pmatrix} 0 & Q \\ Q^* & 0 \end{pmatrix} \mapsto \begin{pmatrix} I & 0 \\ 0 & (Q^{-1})^* \end{pmatrix} \begin{pmatrix} 0 & Q \\ Q^* & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & Q^{-1} \end{pmatrix} = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$$

shows that A is $\left(\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right)$-unitary, and a simultaneous permutation of rows and columns in A and in $\left(\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right)$ yields the desired form. ∎

3.2. The real case. In this subsection $\mathbb{F} = \mathbb{R}$. We say that a matrix $A \in M_n(\mathbb{R})$ is diagonalizable if A is similar to a diagonal matrix (over the complex field). Thus, $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$is diagonalizable, and $\operatorname{Spec}(A) = \{i, -i\}$. If $\lambda \in \operatorname{Spec}(A)$ is real, we let $\mathcal{R}_\lambda(A) = \operatorname{Ker}(A - \lambda I)^n \subseteq \mathbb{R}^n$. If $\lambda \pm \mu i$ is a pair of nonreal complex conjugate eigenvalues of A, we let

$$\mathcal{R}_{\lambda \pm \mu i}(A) = \operatorname{Ker}\left(A^2 - 2\lambda A + (\lambda^2 + \mu^2)I\right)^n \subseteq \mathbb{R}^n.$$

Then we have a direct sum (see, e.g., [10, section 12.2])

$$\mathbb{R}^n = \sum_{j=1}^{k} \mathcal{R}_{\lambda_j}(A) + \sum_{j=1}^{\ell} \mathcal{R}_{\lambda_j \pm i\mu_j}(A),$$

where $\lambda_1, \dots, \lambda_k$ are all distinct real eigenvalues of A (if any), and $\lambda_1 \pm i\mu_1, \dots, \lambda_\ell \pm i\mu_\ell$ are all distinct pairs of nonreal complex conjugate eigenvalues of A (if any), where it is assumed that $\mu_j > 0$.

If A is H-unitary, then by Proposition 3.1 the eigenvalues of A can be collected into sets of the following four structures (for a particular A, some of these sets may be absent):

(i) $\lambda = \pm 1 \in \operatorname{Spec}(A)$;

(ii) $\{\lambda, \overline{\lambda}\} \subseteq \operatorname{Spec}(A)$, where $|\lambda| = 1$ and the imaginary part of λ is positive;

(iii) $\{\lambda, \lambda^{-1}\} \subseteq \operatorname{Spec}(A)$, where $\lambda \in \mathbb{R}$, $|\lambda| > 1$;

(iv) $\{\lambda, \overline{\lambda}, \lambda^{-1}, \overline{\lambda}^{-1}\} \subseteq \operatorname{Spec}(A)$, where λ has positive imaginary part and $|\lambda| > 1$.

According to these four structures, we let

$$\mathcal{RR}_\lambda(A) := \begin{cases} \mathcal{R}_\lambda(A) & \text{if } \lambda = 1 \text{ or } \lambda = -1, \\ \mathcal{R}_{\lambda,\overline{\lambda}}(A) & \text{if } |\lambda| = 1 \text{ and the imaginary part of } \lambda \text{ is positive}, \\ \mathcal{R}_\lambda(A) \dot{+} \mathcal{R}_{\lambda^{-1}}(A) & \text{if } \lambda \in \mathbb{R}, \ |\lambda| > 1, \\ \mathcal{R}_{\lambda,\overline{\lambda}} \dot{+} \mathcal{R}_{\lambda^{-1},\overline{\lambda}^{-1}}(A) & \text{if } \lambda \text{ has positive imaginary part and } |\lambda| > 1. \end{cases}$$

With this notation, we can now state an orthogonality result analogous to Lemma 3.2.

LEMMA 3.4. *Let A be H-unitary.*

(a) *If $v \in \mathcal{RR}_\lambda(A)$, $w \in \mathcal{RR}_\mu(A)$, where $\lambda \neq \mu$, then v and w are H-orthogonal:*

$$[v, w] = 0. \tag{3.4}$$

(b) *If $x \in \mathbb{R}^n$ is an eigenvector corresponding to a real eigenvalue λ of A, and* either $|\lambda| \neq 1$ or $\lambda = \pm 1$ *and there exists $y \in \mathbb{R}^n$ such that $(A - \lambda I)y = x$, then $[x, x] = 0$.*

(c) *If $\lambda = \mu + i\nu$ is a nonreal eigenvalue of A with positive imaginary part ν and with $|\lambda| \neq 1$, and if $(A^2 - 2\mu A + (\mu^2 + \nu^2)I)x = 0$, then $[x, x] = 0$.*

(d) *If $\lambda = \mu + i\nu$ is a nonreal eigenvalue of A with positive imaginary part ν and with $|\lambda| = 1$, and if*

$$(A^2 - 2\mu A + I)x = 0, \quad (A^2 - 2\mu A + I)y = x \tag{3.5}$$

for some $y \in \mathbb{R}^n$, then $[x, x] = 0$.

Proof. Part (a) follows from Lemma 3.2 by considering a complexification of A, i.e., considering A as a complex matrix representing a linear transformation in $\mathbb{C}^n$. The same complexification takes care of statement (c). Part (b) is proved in exactly the same way as part (b) of Lemma 3.2.

It remains to prove part (d). Assume (3.5) holds, and arguing by contradiction, suppose $[x, x] \neq 0$. Let

$$y_N = y + \alpha x + \beta A x, \tag{3.6}$$

where $\alpha, \beta \in \mathbb{R}$ are chosen so that

$$[y_N, x] = [Ay_N, x] = 0. \tag{3.7}$$

This choice of α and β is possible. Indeed, (3.7) amounts to the following system of linear equations for α and β:

$$\alpha[x,x] + \beta[Ax,x] = -[y,x],$$

$$\alpha[Ax,x] + \beta[A^2x,x] = \alpha[Ax,x] + \beta(2\mu[Ax,x] - [x,x]) = -[Ay,x].$$

The determinant of the system is

$$-[Ax,x]^2 - [x,x]^2 + 2\mu[x,x][Ax,x],$$

which is negative since $[x,x] \neq 0$ and $-1 < \mu < 1$. Clearly, $(A^2 - 2\mu A + I)y_N = x$, and using (3.7), we obtain

$$\begin{aligned}
[x,x] &= [(A^2 - 2\mu A + I)y_N, x] = [A^2 y_N, x] = [A^2 y_N, (A^2 - 2\mu A + I)y_N] \\
&= [A^2 y_N, A^2 y_N] - 2\mu[Ay_N, y_N] + [A^2 y_N, y_N] = \text{ (because } [Ay_N, Ay_N] = [y_N, y_N]) \\
&= [y_N, y_N] - 2\mu[Ay_N, y_N] + [x + 2\mu Ay_N - y_N, y_N] \\
&= [y_N, y_N] - 2\mu[Ay_N, y_N] + [x, y_N] + 2\mu[Ay_N, y_N] - [y_N, y_N] = 0,
\end{aligned}$$

a contradiction to our supposition. □

THEOREM 3.5. *Let H be a real symmetric invertible $n \times n$ matrix. A diagonalizable matrix $A \in M_n(\mathbb{R})$ is H-unitary if and only if there exists an invertible matrix $S \in M_n(\mathbb{R})$ such that $S^{-1}AS$ equals*

$$U_0 \oplus U_1 \oplus \ldots \oplus U_q \oplus U_{q+1} \oplus \ldots \oplus U_{q+r} \oplus U_{q+r+1} \oplus \ldots \oplus U_{q+r+s}, \tag{3.8}$$

and $S^t H S$ equals

$$H_0 \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \epsilon_{q+1} I_2 \oplus \ldots \oplus \epsilon_{q+r} I_2 \oplus \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix}, \tag{3.9}$$

where the constituents of (3.8) *and* (3.9) *are as follows:*

(i) *$U_0, H_0 \in M_m(\mathbb{R})$ are diagonal orthogonal matrices;*

(ii) *For $j = 1, \ldots, q$, the 2×2 matrices U_j are of the form $U_j = \left(\begin{smallmatrix} \lambda_j & 0 \\ 0 & \lambda_j^{-1} \end{smallmatrix}\right)$, where $\lambda_j \in \mathbb{R}$, $|\lambda_j| > 1$;*

(iii) *For $j = q+1, \ldots, q+r$, the 2×2 matrices U_j are of the form $U_j = \left(\begin{smallmatrix} \lambda_j & \mu_j \\ -\mu_j & \lambda_j \end{smallmatrix}\right)$, where $\lambda_j^2 + \mu_j^2 = 1$ and $\mu_j > 0$, and the ϵ_j's are ± 1;*

(iv) *For $j = q+r+1, \ldots, q+r+s$, the 4×4 matrices U_j are of the form*

$$U_j = \begin{pmatrix} \lambda_j & \mu_j & 0 & 0 \\ -\mu_j & \lambda_j & 0 & 0 \\ 0 & 0 & \lambda_j' & \mu_j' \\ 0 & 0 & -\mu_j' & \lambda_j' \end{pmatrix}, \qquad \lambda_j^2 + \mu_j^2 > 1, \quad \mu_j > 0,$$

$$\lambda_j' + \mu_j' i = (\lambda_j - \mu_j i)^{-1}.$$

One or more of types (i)–(iv) *may be absent in* (3.8) *and* (3.9).

Moreover, the representation of an H-unitary matrix A as in (3.8), (3.9) *is unique up to a simultaneous permutation of constituent pairs.*

Proof. We prove the (nontrivial) "only if" part. By Lemma 3.4, we may assume that one of the following four cases (a)–(d) happens: (a) $\mathrm{Spec}\,(A) = 1$ or $\mathrm{Spec}\,(A) = -1$; (b) $\mathrm{Spec}\,(A) = \{\lambda, \lambda^{-1}\}$, $|\lambda| > 1$, λ real; (c) $\mathrm{Spec}\,(A) = \{\lambda \pm i\mu\}$, $\lambda^2 + \mu^2 = 1$, $\mu > 0$; (d) $\mathrm{Spec}\,(A) = \{\lambda \pm i\mu, (\overline{\lambda \pm i\mu})^{-1}\}$, $\lambda^2 + \mu^2 > 1$, $\mu > 0$. In the cases (a) and (b), one argues as in the proof of Theorem 3.3. Consider the case (c). Applying the transformation $A \mapsto S^{-1}AS$, $H \mapsto S^tHS$, where S is a real invertible matrix, we can assume that A is in the real Jordan form; i.e., since A is diagonalizable,

$$A = \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix} \in M_{2m}(\mathbb{R}).$$

Partition H: $H = [H_{j,k}]_{j,k=1}^m$, where $H_{j,k}$ is 2×2. It turns out that (since A is H-unitary) $H_{j,k} = \left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$, where a, b are real numbers (which depend on j and k). Indeed, fix j and k, and let $H_{j,k} = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, $a, b, c, d \in \mathbb{R}$. Equation

$$\begin{pmatrix} \lambda & -\mu \\ \mu & \lambda \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

may be rewritten as a system of four homogeneous linear equations with unknowns a, b, c, d as follows:

$$\begin{pmatrix} \lambda^2 - 1 & -\mu\lambda & -\mu\lambda & \mu^2 \\ \mu\lambda & \lambda^2 - 1 & -\mu^2 & -\mu\lambda \\ \mu\lambda & -\mu^2 & \lambda^2 - 1 & -\mu\lambda \\ \mu^2 & \mu\lambda & \mu\lambda & \lambda^2 - 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = 0. \tag{3.10}$$

It is easy to see, using $\lambda^2 + \mu^2 = 1$, that the general solution of (3.10) is $\{(a, b, c, d)^t : a = d,\ b = -c\}$, and hence H has the required form. Now the proof of Theorem 3.5 in the case (c) reduces to the complex case via the identification (1.1).

Finally, consider the case (d). As in the proof of case (c), the proof of Theorem 3.5 in case (d) boils down to the following claim. Let λ and μ be real numbers such that $\lambda^2 + \mu^2 > 1$ and $\mu > 0$, let $\lambda' + i\mu' = (\lambda - i\mu)^{-1}$, and assume that

$$\begin{pmatrix} \lambda & -\mu & 0 & 0 \\ \mu & \lambda & 0 & 0 \\ 0 & 0 & \lambda' & -\mu' \\ 0 & 0 & \mu' & \lambda' \end{pmatrix} H \begin{pmatrix} \lambda & \mu & 0 & 0 \\ -\mu & \lambda & 0 & 0 \\ 0 & 0 & \lambda' & \mu' \\ 0 & 0 & -\mu' & \lambda' \end{pmatrix} = H, \tag{3.11}$$

where

$$H = \begin{pmatrix} 0 & 0 & a & b \\ 0 & 0 & c & d \\ e & f & 0 & 0 \\ g & h & 0 & 0 \end{pmatrix} \in M_4(\mathbb{R}); \tag{3.12}$$

then, in fact,

$$H = \begin{pmatrix} 0 & 0 & a & b \\ 0 & 0 & -b & a \\ e & f & 0 & 0 \\ -f & e & 0 & 0 \end{pmatrix}. \tag{3.13}$$

(It follows from Lemma 3.4 that the 2×2 top left and bottom right corners of H in (3.12) are zeros.) Rewrite (3.11) as a system of linear equations

$$\begin{pmatrix} \lambda\lambda'-1 & -\mu'\lambda & -\mu\lambda' & \mu\mu' \\ \mu'\lambda & \lambda\lambda'-1 & -\mu\mu' & -\mu\lambda' \\ \mu\lambda' & -\mu\mu' & \lambda\lambda'-1 & -\mu'\lambda \\ \mu\mu' & \mu'\lambda & \mu\lambda' & \lambda\lambda'-1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = 0, \tag{3.14}$$

and an analogous system for e, f, g, h. Since $\lambda' + i\mu' = (\lambda - i\mu)^{-1}$, we have $\mu\lambda' = \mu'\lambda$ and $\lambda\lambda' + \mu\mu' = 1$, and our claim follows easily. □

4. Applications. In this section we present several applications and consequences of the canonical forms given in sections 2 and 3.

4.1. Connected components of H-unitary similarity orbit. Since the group of H-unitary matrices is connected in the complex case, the H-unitary orbit

$$\mathcal{U}_{\mathbb{F}}^{H}(A) = \{U^{-1}AU : U \in \mathcal{U}_{\mathbb{F}}^{H}\}$$

is also (arcwise) connected. In the real case, this need not be true. Using Theorem 3.5, we sort out the number of connected components in the H-unitary orbit of an H-unitary diagonalizable matrix A, in the real case.

We assume from now on in this section that $\mathbb{F} = \mathbb{R}$. Let $\mathcal{U}_{\mathbb{R},0}^{H}$ be the connected component of $\mathcal{U}_{\mathbb{R}}^{H}$ containing the identity. Since by Corollaries 2.6 and 2.8 the factor group $\mathcal{U}_{\mathbb{R}}^{H}/\mathcal{U}_{\mathbb{R},0}^{H}$ is isomorphic to $\{1,-1\} \times \{1,-1\}$, the H-unitary orbit may have one, two, or four connected components. The proof of the following lemma is obvious.

LEMMA 4.1. *Let $A \in M_n(\mathbb{R})$. The orbit $\mathcal{U}_{\mathbb{R}}^{H}(A)$ has one, two, or four connected components if and only if the group*

$$\{U \in \mathcal{U}_{\mathbb{R}}^{H} : AU = UA\} \tag{4.1}$$

intersects all connected components of $\mathcal{U}_{\mathbb{R}}^{H}$, intersects only two connected components of $\mathcal{U}_{\mathbb{R}}^{H}$, or is contained in $\mathcal{U}_{\mathbb{R},0}^{H}$, respectively.

LEMMA 4.2. *The orbit $\mathcal{U}_{\mathbb{R}}^{H}(A)$ and the orbit $\mathcal{U}_{\mathbb{R}}^{S^tHS}(S^{-1}AS)$ have the same number of connected components, for every invertible $S \in M_n(\mathbb{R})$.*

Proof. Notice that $\mathcal{U}_{\mathbb{R}}^{S^tHS}(S^{-1}AS) = S^{-1}\left(\mathcal{U}_{\mathbb{R}}^{H}(A)\right)S$.

LEMMA 4.3. *If A is diagonalizable, H-unitary, and has no real eigenvalues, then $\mathcal{U}_{\mathbb{R}}^{H}(A)$ has four connected components.*

Proof. By Lemma 4.1 we have to prove that the group (4.1) is connected, and by Lemma 4.2 and Theorem 3.5 we may assume that A and H are given by

$$A = U_{q+1} \oplus \ldots \oplus U_{q+r} \oplus U_{q+r+1} \oplus \ldots \oplus U_{q+r+s} \in M_{2m}(\mathbb{R}),$$

$$H = \epsilon_{q+1}I_2 \oplus \ldots \oplus \epsilon_{q+r}I_2 \oplus \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix},$$

where the U_j's and ϵ_j's are as in Theorem 3.5. Consider the map

$$X \in M_m(\mathbb{C}) \;\mapsto\; \phi(X) \in M_{2m}(\mathbb{R}),$$

defined entrywise by $\phi(x+iy) = \left(\begin{smallmatrix} x & y \\ -y & x \end{smallmatrix}\right)$, $x, y \in \mathbb{R}$. We obviously have $A = \phi(\widehat{A})$, $H = \phi(\widehat{H})$, where $\widehat{A}$ is $\widehat{H}$-unitary. Since every real matrix S commuting with A has

the form $S = [S_{j,k}]$, with the 2×2 blocks $S_{j,k} = \left(\begin{smallmatrix} \alpha_{j,k} & \beta_{j,k} \\ -\beta_{j,k} & \alpha_{j,k} \end{smallmatrix}\right)$, $\alpha_{j,k}, \beta_{j,k} \in \mathbb{R}$ ([10, Theorem 12.4.2]), we have

$$\begin{aligned} &\phi(\{\widehat{U} \in M_m(\mathbb{C}) : \widehat{A}\widehat{U} = \widehat{U}\widehat{A},\ \widehat{U}^*\widehat{H}\widehat{U} = \widehat{H}\}) \\ &\quad = \{U \in M_{2m}(\mathbb{R}) : AU = UA,\ \ U^tHU = H\}. \end{aligned}$$

Now $\widehat{A}$ is diagonalizable, and therefore the canonical form of Theorem 3.3, together with the connectedness of the H-unitary group in the complex case, guarantees that the group

$$\{\widehat{U} \in M_m(\mathbb{C}) : \widehat{A}\widehat{U} = \widehat{U}\widehat{A},\ \ \widehat{U}^*\widehat{H}\widehat{U} = \widehat{H}\}$$

is connected. Since ϕ is continuous, the group $\{U \in M_{2m}(\mathbb{R}) : AU = UA,\ \ U^tHU = H\}$ is connected as well. ∎

LEMMA 4.4. *If A is diagonalizable, H-unitary, and all its eigenvalues are real and different from ± 1, then $\mathcal{U}^H_{\mathbb{R}}(A)$ has two connected components:*

$$\{U^{-1}AU : U \in \mathcal{U}^H_{\mathbb{R}},\ \det U = 1\} \quad \textit{and} \quad \{U^{-1}AU : U \in \mathcal{U}^H_{\mathbb{R}},\ \det U = -1\}.$$

Proof. We have to prove that the group (4.1) intersects exactly two components of the H-unitary group, and all elements of (4.1) have determinant 1. In view of Lemma 4.2 and Theorem 3.5, we may assume that

$$A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}, \quad m \text{ times}, \quad |\lambda| > 1, \quad \lambda \in \mathbb{R},$$

$$H = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad m \text{ times}.$$

It will be convenient to apply a simultaneous row and column permutation to represent A and H in the form

$$A = \begin{pmatrix} \lambda I_m & 0 \\ 0 & \lambda^{-1} I_m \end{pmatrix}, \quad H = \begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix}.$$

Now a matrix A belongs to (4.1) if and only if $A = \left(\begin{smallmatrix} U & 0 \\ 0 & (U^t)^{-1} \end{smallmatrix}\right)$, where $U \in M_m(\mathbb{R})$ is invertible. Clearly, $\det(A) = 1$. To see that the group (4.1) intersects the component of the H-unitary group defined by $\sigma_+ = \sigma_- = -1$ (see (2.2)), we transform A and H again as follows:

$$\frac{1}{2}\begin{pmatrix} I & I \\ I & -I \end{pmatrix} A \begin{pmatrix} I & I \\ I & -I \end{pmatrix} = \frac{1}{2}\begin{pmatrix} U + (U^t)^{-1} & U - (U^t)^{-1} \\ U - (U^t)^{-1} & U + (U^t)^{-1} \end{pmatrix},$$

$$\begin{pmatrix} I & I \\ I & -I \end{pmatrix} H \begin{pmatrix} I & I \\ I & -I \end{pmatrix} = \begin{pmatrix} 2I & 0 \\ 0 & -2I \end{pmatrix}.$$

Clearly, there exist invertible $U \in M_m(\mathbb{R})$ such that $\det(U + (U^t)^{-1}) < 0$. ∎

The proof of the following theorem is obtained by using the preceding lemmas, and arguing analogously in the case when eigenvalues ± 1 are present.

THEOREM 4.5. *Let $A \in M_n(\mathbb{R})$ be H-unitary and diagonalizable, where $H \in M_n(\mathbb{R})$ is symmetric and invertible (recall the standing assumption that H is indefinite). Then the orbit $\mathcal{U}_{\mathbb{R}}^H(A)$ has four connected components if and only if A has no real eigenvalues, and has two connected components*

$$\{U^{-1}AU : U \in \mathcal{U}_{\mathbb{R}}^H,\ \det U = 1\} \quad \text{and} \quad \{U^{-1}AU : U \in \mathcal{U}_{\mathbb{R}}^H,\ \det U = -1\}$$

if and only if A has real eigenvalues but all of them are different from ± 1.

Assume now that 1 or -1 (or both) belongs to $\operatorname{Spec}(A)$. If the quadratic form x^tHx, $x \in \operatorname{Ker}(A^2 - I)$, is indefinite, then $\mathcal{U}_{\mathbb{R}}^H(A)$ is connected. If the quadratic form x^tHx, $x \in \operatorname{Ker}(A^2 - I)$, is (positive or negative) definite, then, in fact,

$$\mathcal{U}_{\mathbb{R}}^H(A) = \{U^{-1}AU : U \in \mathcal{U}_{\mathbb{R}}^H,\ \det U = 1\},$$

and the orbit $\mathcal{U}_{\mathbb{R}}^H(A)$ is connected if A has real eigenvalues different from ± 1, and has two connected components otherwise.

4.2. Products of positive definite J-unitary matrices. Let $J = I_p \oplus -I_q$. Consider the problem of characterizing those J-unitary matrices that can be written as the product

$$\begin{pmatrix} \sqrt{I_p + XX^*} & X \\ X^* & \sqrt{I_q + X^*X} \end{pmatrix} \begin{pmatrix} \sqrt{I_p + YY^*} & N \\ Y^* & \sqrt{I_q + Y^*Y} \end{pmatrix}, \tag{4.2}$$

with $X, Y \in M_{p\times q}(\mathbb{F})$. A related question has been considered by van Wyk [22] for the case $H = [-1] \oplus I_3$.

THEOREM 4.6. *$\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. The following statements are equivalent for a J-unitary matrix A:*

(a) *A is J-unitarily similar to a matrix of the form (4.2); i.e., $A = U^{-1}BU$ for some J-unitary U and some B of the form (4.2).*

(b) *A is J-unitarily similar to a matrix of the form*

$$\begin{pmatrix} \sqrt{I_p + CC^t} & C \\ C^t & \sqrt{I_q + C^tC} \end{pmatrix},$$

where $C = [c_{ij}]$ with $c_{11} \geq \cdots \geq c_{ss} > 0$ for some $s \leq \min\{p, q\}$ and all other entries of C are zero.

(c) *The eigenvalues of A are positive and semisimple, i.e., no Jordan blocks of size bigger than 1 in the Jordan form of A.*

(d) *A is of the form (4.2).*

Proof. (b) $\Rightarrow$ (a) is obvious, whereas (a) $\Rightarrow$ (c) follows because (4.2) is a product of two positive definite matrices, and every product of two positive definite matrices has positive and semisimple eigenvalues. Assume (c) holds. By Theorems 3.3 and 3.5, we have

$$S^{-1}AS = I_r \oplus U_{m+1} \oplus \ldots \oplus U_{m+s}, \quad S^*JS = I_{r_+} \oplus -I_{r_-} \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \oplus \ldots \oplus \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

for some invertible $S \in M_n(\mathbb{F})$, where $U_j = \left(\begin{smallmatrix} \lambda_j & 0 \\ 0 & \lambda_j^{-1} \end{smallmatrix}\right)$, $\lambda_j \in \mathbb{R}$, $|\lambda_j| > 1$. Applying a suitable matrix transformation T, we obtain $T^*(S^*JS)T = J$ and

$$T^{-1}(S^{-1}AS)T = \begin{pmatrix} D_1 & C \\ C^t & D_2 \end{pmatrix}, \tag{4.3}$$

where

$$D_1 = \operatorname{diag}\left(\frac{1}{2}(\lambda_1+\lambda_1^{-1}),\dots,\frac{1}{2}(\lambda_s+\lambda_s^{-1}),\ \overbrace{1,\dots,1}^{r_+ \text{ times}}\right),$$

$$D_2 = \operatorname{diag}\left(\frac{1}{2}(\lambda_1+\lambda_1^{-1}),\dots,\frac{1}{2}(\lambda_s+\lambda_s^{-1}),\ \overbrace{1,\dots,1}^{r_- \text{ times}}\right),$$

and $C=[c_{ij}]$ with $c_{jj}=\frac{1}{2}(\lambda_j-\lambda_j^{-1})$ for $j=1,\dots,s$, and all other c_{ij} equal to zero. Thus, the right-hand side of (4.3) has the form as required in (b).

Finally, suppose (a) holds true. Let $A=U^{-1}CDU$, where C and D are the two matrices as in (4.2), and U is J-unitary. Then

$$A = U^{-1}C(U^{-1})^*U^*DU.$$

Now both $C_1=U^{-1}C(U^{-1})^*$ and $D_1=U^*DU$ are J-unitary and positive definite. By the uniqueness of polar decomposition of invertible matrices, and applying Theorem 2.4 (a) with A replaced with C_1 and with D_1, we see that $A=C_1D_1$ is of the form (4.2). It is obvious that (d) implies (a). ∎

One can check that for a given J-unitary matrix A with positive and semisimple eigenvalues, a representation in the form (4.2) is not unique, as observed in [22].

4.3. Products of reflections. In $M_n(\mathbb{R})$, a matrix of the form $T_v=I-2vv^t$ is called a reflection, and $T_v(x)=x-2(v^tx)v$ is just a reflection of the vector x about the plane $v^\perp$. We can extend the definition using the indefinite inner product $[x,y]=y^tHx$, and define

$$T_v = I-2vv^tH/(v^tHv), \quad \text{where} \quad v^tHv\neq 0. \tag{4.4}$$

One readily checks that T_v is an H-unitary matrix such that $T_v(x)=x-2[x,v]v/[v,v]$. Assuming that $H=J=I_p\oplus -I_q$, if v is in the linear span of two basic vectors e_i and e_j, then we say that T_v is an *elementary reflection*. Thus, an elementary reflection is a direct sum of a 2×2 matrix and I_{n-2} (or a diagonal matrix in the degenerate case) with determinant -1. We have the following result.

THEOREM 4.7. *A matrix $A\in M_n(\mathbb{R})$ is J-unitary if and only if it is a product of at most $f(p,q)=p(p-1)+q(q-1)+\min\{p,q\}+4$ elementary reflections.*

Proof. The ($\Leftarrow$) is clear. Conversely, suppose A is J-unitary. By Theorem 2.3, there exist $X=X_1\oplus X_2$ and $Y=Y_1\oplus Y_2$ with $X_1,Y_1\in\mathcal{U}_p$ and $X_2,Y_2\in\mathcal{U}_q$ such that

$$XAY = \begin{pmatrix} -\sqrt{I_p+DD^t} & D \\ -D^t & \sqrt{I_q+D^tD}\end{pmatrix}, \tag{4.5}$$

with D as in Theorem 2.3. First, we show that X_1 is a product of no more than $p(p-1)/2+1$ elementary reflections. This can be proved by simple inductive arguments as follows. Suppose $X_1=[u_1|\cdots|u_p]$. By elementary considerations or by [13, p. 226], there are elementary reflections $T_1,\dots,T_{p-1}$ such that $T_1\cdots T_{p-1}u_1=e_1$. Hence $T_1\cdots T_{p-1}X_1=[1]\oplus\tilde{X}_1$. Repeat the arguments to $\tilde{X}_1$ and so on until we get a 2×2

matrix, which is either an elementary reflection or the product of two elementary reflections. Thus, we can write X_1 as a product of no more than

$$[(p-1)+(p-2)+\cdots+1]+1 = p(p-1)/2+1$$

elementary reflections. We can apply similar arguments to X_2, Y_1, Y_2, and conclude that each of the matrices X and Y can be written as the product of at most $p(p-1)/2+q(q-1)/2+2$ elementary reflections. Finally, we deal with XAY. Let $E_{ij} = e_i e_j^t \in M_n(\mathbb{R})$ for $i,j \in \{1,\ldots,n\}$. Evidently, $XAY = B_1 \cdots B_m$, where

$$B_j = \sqrt{1+d_j^2}(-E_{jj}+E_{p+j,p+j}) + d_j(E_{j,p+j}-E_{p+j,j}) + \sum_{k\neq i,j} E_{kk},\ j=1,\ldots,m.$$

To show that B_j is a matrix of the form (4.4), we only need to deal with the 2×2 matrix

$$\begin{pmatrix} -\sqrt{1+d^2} & d \\ -d & \sqrt{1+d^2} \end{pmatrix} \quad \text{with } d \geq 0;$$

here $J = [1] \oplus [-1]$. To this end, let $f(\theta) = 2\sec\theta\tan\theta$ with $\theta \in [0,\pi/2)$. Then f maps $[0,\pi/2)$ to $[0,\infty)$. So, there exists $\theta \in [0,\pi/2)$ such that $f(\theta) = d$. Let $v = (\sec\theta, \tan\theta)^t$. Then $v^t J v = 1$, and

$$R = I_2 - 2vv^tJ = \begin{pmatrix} -1-2\tan^2\theta & 2\sec\theta\tan\theta \\ -2\sec\theta\tan\theta & 1+2\tan^2\theta \end{pmatrix} = \begin{pmatrix} -\sqrt{1+d^2} & d \\ -d & \sqrt{1+d^2} \end{pmatrix};$$

here we have used $1+2\tan^2\theta = \sqrt{1+d^2} = \sqrt{1+(2\sec\theta\tan\theta)^2}$. The result follows. ∎

If one uses (general) reflections instead of elementary reflections, then the number of reflections needed to represent every J-unitary matrix as a product of reflections can be considerably improved (cf. Theorem 4.7): Every J-unitary matrix, in the real as well as in the complex case, can be written as a product of no more than n reflections, a result that goes back to [5].

4.4. Stability and robust stability of J-unitary matrices. In the complex case the results of this section are given in [9] (see also references therein).

In applications, one often needs conditions for powers of a matrix to be bounded. We say that a matrix $A \in M_n(\mathbb{F})$ is *forward stable* if the set $\{A^m\}_{m=0}^{\infty}$ is bounded, and is *backward stable* if A is invertible and the set $\{A^m\}_{m=-\infty}^{0}$ is bounded.

THEOREM 4.8 ($\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$). *The following statements are equivalent for an H-unitary matrix A:*

(a) *A is forward stable;*

(b) *A is backward stable;*

(c) *A is diagonalizable and has only unimodular eigenvalues.*

Proof. It is well-known (and easy to see from the Jordan form of A) that (c) is equivalent to both forward and backward stability of A, whereas (a) (resp., (b)) is equivalent to A having all its eigenvalues inside the closed unit circle (resp., outside the open unit circle), with unimodular eigenvalues, if any, being semisimple; i.e., their geometric multiplicity coincides with their algebraic multiplicity. (This remark applies to any $A \in M_n(\mathbb{F})$, not necessarily H-unitary.) It remains to observe that if $\lambda \in \sigma(A)$, then $\overline{\lambda}^{-1} \in \sigma(A)$, and so (a) and (b) are equivalent for H-unitary matrices. ∎

In view of Theorem 4.8, we say that an H-unitary matrix is *stable* if it is backward or forward stable. For A H-unitary, we say that A is *robustly stable* if there is $\epsilon > 0$ such that every G-unitary matrix B is stable, provided G is Hermitian and

$$\|G - H\| + \|B - A\| < \epsilon.$$

Here $\|\cdot\|$ is any fixed norm in $M_n(\mathbb{F})$. Note that by taking ϵ sufficiently small, the invertibility of G is guaranteed.

THEOREM 4.9 ($\mathbb{F} = \mathbb{C}$). *An H-unitary matrix A is robustly stable if and only if A is diagonalizable with only unimodular eigenvalues and every eigenvector is H-definite:*

$$Ax = \lambda x, \quad x \neq 0 \quad \Longrightarrow \quad x^*Hx \neq 0. \tag{4.6}$$

Proof. By Theorem 4.8, we can assume to start with that A is diagonalizable with only unimodular eigenvalues. By Theorem 3.3 we may further assume that

$$A = U_1 \oplus \ldots \oplus U_m, \quad H = \epsilon_1 \oplus \ldots \oplus \epsilon_m, \tag{4.7}$$

where U_j and ϵ_j are as in (3.3).

Assume first that (4.6) does not hold. Then there exist indices $j \neq k$ such that $U_j = U_k = \lambda$ and $\epsilon_j \neq \epsilon_k$. For notational convenience assume $j = 1$, $k = 2$, and $\epsilon_1 = 1$. Let q be any complex number different from $-\lambda$. Then a straightforward computation shows that

$$A(q) := \begin{pmatrix} \frac{1}{2}(\lambda + q) + \frac{1}{2}(\overline{\lambda + q})^{-1} & \frac{1}{2}(\lambda + q) - \frac{1}{2}(\overline{\lambda + q})^{-1} \\ \frac{1}{2}(\lambda + q) - \frac{1}{2}(\overline{\lambda + q})^{-1} & \frac{1}{2}(\lambda + q) + \frac{1}{2}(\overline{\lambda + q})^{-1} \end{pmatrix} \oplus U_3 \ldots \oplus U_m \tag{4.8}$$

is H-unitary, as close to A as we wish (if q is sufficiently close to zero), and has non-unimodular eigenvalues $\lambda + q$, $\overline{\lambda + q}^{-1}$ (if q is chosen so that $|\lambda + q| \neq 1$). For such a choice of q, the matrix $A(q)$ cannot be stable. This proves the "only if" part.

To prove the "if" part, assume that (4.6) holds true. Let $\lambda_1, \ldots, \lambda_k$ be all the distinct eigenvalues of A, and let $\delta > 0$ be so small that each disk $\{z \in \mathbb{C} : |z - \lambda_j| \leq \delta\}$ does not contain any eigenvalues of A besides λ_j.

To continue the proof, we need the well-known notion of the gap between subspaces. If $\mathcal{M}, \mathcal{N}$ are subspaces in $\mathbb{C}^n$, the *gap* $\operatorname{gap}(\mathcal{M}, \mathcal{N})$ is defined as $\|P_{\mathcal{M}} - P_{\mathcal{N}}\|_{\text{op}}$, where $P_{\mathcal{M}}$ (resp., $P_{\mathcal{N}}$) is the orthogonal projection onto $\mathcal{M}$ (resp., $\mathcal{N}$), and $\|\cdot\|_{\text{op}}$ is the operator norm (i.e., the largest singular value). We refer the reader to [10] for many basic properties of the gap. Returning to our proof, we need the following property (see [10, section 15.2]):

$$\begin{aligned} &\forall\ \epsilon_2 > 0\ \exists\ \epsilon_1 > 0 \ \text{ such that } \ \|B - A\| < \epsilon_1 \\ &\qquad \Longrightarrow \ \max_{j=1}^{k} \left(\operatorname{gap}\left(\mathcal{R}_{\Omega_j}(B), \mathcal{R}_{\lambda_j}(A)\right)\right) < \epsilon_2. \end{aligned} \tag{4.9}$$

Here $\mathcal{R}_{\Omega_j}(B)$ is the sum of all root subspaces of B corresponding to the eigenvalues of B in the disk $\Omega_j := \{z \in \mathbb{C} : |z - \lambda_j| \leq \delta\}$. Taking $\epsilon_2 < 1$ we guarantee that for every j, the dimensions of $\mathcal{R}_{\Omega_j}(B)$ and of $\mathcal{R}_{\lambda_j}(A)$ coincide, and in particular, B cannot have eigenvalues outside of $\cup_{j=1}^{k} \Omega_j$.

On the other hand, since H is definite on each $\mathcal{R}_{\lambda_j}(A)$, and since the property of being definite is preserved under sufficiently small perturbations of H and sufficiently

small perturbations of $\mathcal{R}_{\lambda_j}(A)$ (with respect to the gap), there exists $\epsilon_3 > 0$ (which depends on H and A only) such that a Hermitian matrix G is invertible and definite on each $\mathcal{R}_{\Omega_j}(B)$ $(j = 1, \dots, k)$ as long as $\|G - H\| < \epsilon_3$ and gap $(\mathcal{R}_{\Omega_j}(B), \mathcal{R}_{\lambda_j}(A)) < \epsilon_3$. Take $\epsilon_2 = \min\{1, \epsilon_3\}$ in (4.9); as a result, letting $\epsilon = \min\{\epsilon_3, \epsilon_1\}$, we obtain that G is definite on each $\mathcal{R}_{\Omega_j}(B)$ provided that

$$\|G - H\| + \|B - A\| < \epsilon.$$

If B is, in addition, G-unitary, then by Lemma 3.2 (b), B is diagonalizable with only unimodular eigenvalues; i.e., B is stable. ∎

We consider now robustly stable H-unitary matrices in the real case.

THEOREM 4.10. $\mathbb{F} = \mathbb{R}$. *An H-unitary matrix A is robustly stable if and only if A is diagonalizable with only unimodular eigenvalues and the following conditions hold: For eigenvalues ± 1 of A (if any),*

$$Ax = \pm x, \quad x \in \mathbb{R}^n \setminus \{0\} \quad \Longrightarrow \quad x^t H x \neq 0; \tag{4.10}$$

for every pair of complex conjugate eigenvalues $\mu \pm i\nu$, $\mu^2 + \nu^2 = 1$, of A (if any),

$$(A^2 - 2\mu A + I)x = 0, \quad x \in \mathbb{R}^n \setminus \{0\} \quad \Longrightarrow \quad x^t H x \neq 0. \tag{4.11}$$

Proof. The "if" part follows from the complex result (Theorem 4.9). For the "only if" part, we will prove that if A is diagonalizable with only unimodular eigenvalues and at least one of the conditions (4.10) and (4.11) does not hold, then there exists a real H-unitary matrix B as close as we wish to A which is not stable. We may assume, using the canonical form of Theorem 3.5, that either $A = \pm I_2$, $H = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, or

$$A = \begin{pmatrix} \mu & \nu & 0 & 0 \\ -\nu & \mu & 0 & 0 \\ 0 & 0 & \mu & \nu \\ 0 & 0 & -\nu & \mu \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad \mu^2 + \nu^2 = 1, \quad \nu > 0.$$

In the former case, use

$$B = \begin{pmatrix} \frac{1}{2}(\pm 1 + q) + \frac{1}{2}(\pm 1 + q)^{-1} & \frac{1}{2}(\pm 1 + q) - \frac{1}{2}(\pm 1 + q)^{-1} \\ \frac{1}{2}(\pm 1 + q) - \frac{1}{2}(\pm 1 + q)^{-1} & \frac{1}{2}(\pm 1 + q) + \frac{1}{2}(\pm 1 + q)^{-1} \end{pmatrix}, \quad q \in \mathbb{R} \text{ close to zero.}$$

In the latter case, use (4.8) with $\lambda = \mu + i\nu$, and take advantage of the identification (1.1). ∎

4.5. Stability and robust stability of differential equations. The results of the preceding section have immediate applications to systems of differential equations. Consider the system

$$E\frac{dx}{dt} = iH(t)x, \quad t \in \mathbb{R}, \tag{4.12}$$

where $H(t)$ is a given piecewise continuous function that takes values in the set of $n \times n$ Hermitian matrices, E is a fixed (constant) invertible $n \times n$ Hermitian matrix, and $x(t)$ is a $\mathbb{C}^n$-valued function of t to be found. We assume in addition that $H(t)$ is periodic with a period $\omega \neq 0$: $H(t + \omega) = H(t)$ for all $t \in \mathbb{R}$.

The *matrizant* $X(t)$ of equation (4.12) is defined as the unique $n \times n$ matrix valued solution of the initial value problem

$$E\frac{dX}{dt} = iH(t)X, \quad X(0) = I. \tag{4.13}$$

If $X(t)$ is the matrizant, then differentiating the function X^*EX with respect to t, and using (4.13) and the property of E and H being Hermitian, we obtain

$$\frac{d}{dt}(X(t)^*EX(t)) = \frac{dX^*}{dt}EX + X^*E\frac{dX}{dt} = -iX^*HE^{-1}EX + X^*(iH)X = 0; \tag{4.14}$$

thus X^*EX is constant. Evaluating X^*EX at $t = 0$, we obtain $X(t)^*EX(t) = E$ for all $t \in \mathbb{R}$; in other words, the matrizant is E-unitary valued. Furthermore, since $H(t)$ is periodic with period ω, it is easy to see (because of the uniqueness of the solution of the initial value problem) that $X(t+\omega) = X(t)X(\omega)$, $t \in \mathbb{R}$, and by repeatedly applying this equality we obtain $X(t+m\omega) = X(t)(X(\omega))^m$, m any integer. Therefore, the equation (4.12) is *forward stable*; i.e., all solutions are bounded when $t \to +\infty$, precisely when the set $\{X(\omega)^m\}_{m=0}^{\infty}$ is bounded, and the equation (4.12) is *backward stable*; i.e., all solutions are bounded when $t \to -\infty$, precisely when the set $\{X(\omega)^m\}_{m=-\infty}^{0}$ is bounded. Recalling Theorem 4.8, we have the following theorem.

THEOREM 4.11. *The following conditions are equivalent:*

(a) *Equation* (4.12) *is forward stable.*
(b) *Equation* (4.12) *is backward stable.*
(c) *The matrix $X(\omega)$, where $X(t)$ is the matrizant, is diagonalizable and has only unimodular eigenvalues.*

Thus, we say that (4.12) is *stable* if it is backward or forward stable. We say that (4.12) is *robustly stable* if there exists $\epsilon > 0$ (which depends on E and $H(t)$ only) such that every system

$$\widetilde{E}\frac{dx}{dt} = i\widetilde{H}(t)x, \quad t \in \mathbb{R},$$

is stable provided that the Hermitian valued ω-periodic piecewise continuous function $\widetilde{H}(t)$ and the constant Hermitian matrix $\widetilde{E}$ are such that

$$\|\widetilde{E} - E\| + \max\{\|\widetilde{H}(t) - H(t)\| : 0 \le t < \omega\} < \epsilon.$$

Using the continuous dependence of the solutions of (4.12) on the data E and $H(t)$ (see, e.g., [9, section II.1.1] for details), Theorem 4.9 yields the following theorem.

THEOREM 4.12. *Equation* (4.12) *is robustly stable if and only if the matrix $X(\omega)$ is diagonalizable, has only unimodular eigenvalues, and every eigenvector is E-definite.*

Theorems 4.11 and 4.12 (with $\widetilde{E} = E$) are given in [9]. The book also contains more advanced material concerning stability of (4.12), as well as references to the original literature. In particular, connected components of robustly stable systems (4.12) are described in [9]; in the real skewsymmetric case the study of connected components of robustly stable periodic systems goes back to [7].

There are complete analogues of Theorems 4.11 and 4.12 in the real case, in which case the system of differential equations is

$$\begin{aligned} &E\frac{dx}{dt} = H(t)x, \quad t \in \mathbb{R}, \quad E = E^t \in M_n(\mathbb{R}) \text{ invertible}, \\ &H(t)^t = -H(t) \in M_n(\mathbb{R}). \end{aligned} \tag{4.15}$$

We assume in addition that $H(t)$ is periodic with period $\omega \neq 0$. The matrizant $X(t)$ is defined again as the solution of the initial value problem $E\frac{dX}{dt} = H(t)X(t)$, $X(0) = I$. As in (4.14) one obtains that $X(t)^tEX(t)$ is constant; hence $X(t)$ is E-unitary valued.

The definitions of stability (forward or backward, which turn out to be the same) and of robust stability of (4.15) are analogous to those given above for the complex case, with only real perturbations allowed for the robust stability. We have from Theorems 4.8 and 4.10 the following theorem.

THEOREM 4.13. *Equation* (4.15) *is stable if and only if the real matrix* $X(\omega)$*, where* $X(t)$ *is the matrizant of* (4.15)*, is diagonalizable and has only unimodular eigenvalues. Equation* (4.15) *is robustly stable if and only if it is stable, and, in addition, every eigenvector of* $X(\omega)$ *corresponding to an eigenvalue* ± 1 *(if any) is* E*-definite, and* $x^tEx \neq 0$ *for every vector* $x \in \mathbb{R}^n \setminus \{0\}$ *such that*

$$(X(\omega)^2 - 2\mu X(\omega) + I)x = 0, \qquad \lambda \pm i\mu \in \operatorname{Spec}(X(\omega)), \quad \lambda^2 + \mu^2 = 1.$$

Acknowledgments. We thank N. J. Higham for making available the paper [14] before its publication, and R. A. Horn and anonymous referees for several helpful suggestions.

REFERENCES

[1] L. AUTONNE, *Sur les Matrices Hyperhermitiennes et sur les Matrices Unitaires*, Annales de L'Université de Lyon, Nouvelle Série I, 38 (1915), pp. 1–77.
[2] A. BAKER, *Matrix Groups: An Introduction to Lie Group Theory*, Springer-Verlag, London, 2002.
[3] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, OT45, Birkhäuser-Verlag, Basel, Switzerland, 1990.
[4] J. BOGNÁR, *Indefinite Inner Product Spaces*, Springer-Verlag, New York, 1974.
[5] É. CARTAN, *The Theory of Spinors*, MIT Press, Cambridge, MA, 1967 (translation of 1938 French ed.).
[6] F. R. GANTMACHER, *Applications of the Theory of Matrices*, Interscience, New York, 1959 (translation from Russian).
[7] I. M. GELFAND AND V. B. LIDSKII, *On the structure of the regions of stability of linear canonical systems of differential equations with periodic coefficients*, Uspekhi Mat. Nauk, 10 (1955), pp. 3–40 (in Russian).
[8] W. GIVENS, *Factorization and signatures of Lorentz matrices*, Bull. Amer. Math. Soc., 46 (1940), pp. 81–85.
[9] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrices and Indefinite Scalar Products*, OT8, Birkhäuser-Verlag, Basel, Switzerland, 1983.
[10] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, Wiley Interscience, New York, 1986.
[11] I. GOHBERG AND B. REICHSTEIN, *On H-unitary and block-Toeplitz H-normal operators*, Linear and Multilinear Algebra, 30 (1991), pp. 17–48.
[12] I. GOHBERG AND B. REICHSTEIN, *Classification of block-Toeplitz H-normal operators*, Linear and Multilinear Algebra, 34 (1993), pp. 213–245.
[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[14] N. J. HIGHAM, *J-orthogonal matrices: Properties and generation*, SIAM Rev., 45 (2003), pp. 504–519.
[15] R. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[16] H. KIMURA, *Chain-scattering Approach to H^∞ Control*, Birkhäuser Boston, Boston, 1997.
[17] P. LANCASTER AND L. RODMAN, *Algebraic Riccati Equations*, Oxford University Press, Oxford, UK, 1995.
[18] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.
[19] A. I. MAL'TSEV, *Foundations of Linear Algebra*, W. H. Freeman, San Francisco-London, 1963 (translation from Russian).

[20] V. Mehrmann and H. Xu, *Structured Jordan canonical forms for structured matrices that are Hermitian, skew Hermitian or unitary with respect to indefinite inner products*, Electron. J. Linear Algebra, 5 (1999), pp. 67–103.
[21] L. Rodman, *Products of symmetric and skew-symmetric matrices*, Linear and Multilinear Algebra, 43 (1997), pp. 19–34.
[22] C. B. van Wyk, *The Lorentz operator revisited*, J. Math. Phys., 32 (1991), pp. 425–430.

SIAM J. Matrix Anal. Appl.
Vol. 25, No. 4, pp. 1163–1177

A RANK-k UPDATE PROCEDURE FOR REORTHOGONALIZING THE ORTHOGONAL FACTOR FROM MODIFIED GRAM–SCHMIDT*

LUC GIRAUD†, SERGE GRATTON†, AND JULIEN LANGOU‡

Abstract. The modified Gram–Schmidt algorithm is a well-known and widely used procedure to orthogonalize the column vectors of a given matrix. When applied to ill-conditioned matrices in floating point arithmetic, the orthogonality among the computed vectors may be lost. In this work, we propose an a posteriori reorthogonalization technique based on a rank-k update of the computed vectors. The level of orthogonality of the set of vectors built gets better when k increases and finally reaches the machine precision level for a large enough k. The rank of the update can be tuned in advance to monitor the orthogonality quality. We illustrate the efficiency of this approach in the framework of the seed-GMRES technique for the solution of an unsymmetric linear system with multiple right-hand sides. In particular, we report experiments on numerical simulations in electromagnetic applications where a rank-one update is sufficient to recover a set of vectors orthogonal to machine precision level.

Key words. Gram–Schmidt algorithm, reorthogonalization schemes, ill-conditioned matrix, seed-GMRES method

AMS subject classifications. 65F25, 65G50, 15A23

DOI. 10.1137/S0895479803424347

1. Introduction. Let A be an $m \times n$ real matrix, $m \geq n$ of full rank n. In exact arithmetic, the modified Gram–Schmidt algorithm (MGS) computes an $m \times n$ matrix Q with orthonormal columns and an $n \times n$ upper triangular matrix R such that $A = QR$. The framework of this paper is the study of the MGS algorithm in the presence of rounding errors. We call *computed quantities* quantities that are computed using a well-designed floating point arithmetic [1]. We denote by $\bar{Q}$ and $\bar{R}$ the computed quantities obtained by running MGS in the presence of rounding errors.

In [2], Björck and Paige show that $\bar{R}$ is as good as the triangular factor obtained using backward stable transformations such as Givens rotations or Householder reflections. This property of MGS explains why this algorithm can be safely used in applications where only the factor $\bar{R}$ is needed. This is namely the case in the solution of linear least squares problems where the R-factor of the QR-factorization of $[A, b]$ is needed [1, 2]. Another important feature of MGS is that the number of operations required to explicitly compute the Q-factor (known as the orthogonal basis problem) is approximatively half that of the methods using Givens rotations or Householder reflections [7, p. 232]. However, the computed factor $\bar{Q}$ has less satisfactory properties, since for an ill-conditioned matrix A, it may exhibit a very poor orthogonality as measured by the quantity $\|\bar{Q}^T\bar{Q} - I_n\|$, where $\|.\|$ denotes the spectral 2-norm and I_n denotes the identity matrix of order n [15]. This has stimulated significant work on modifications of MGS that enhance the orthogonality of $\bar{Q}$ at low computational cost. One of those strategies performs reorthogonalizations during the algorithm when a

*Received by the editors March 12, 2003; accepted for publication (in revised form) by A. J. Wathen January 13, 2004; published electronically August 4, 2004.
http://www.siam.org/journals/simax/25-4/42434.html
†CERFACS, 42 av. Gaspard Coriolis, 31057 Toulouse Cedex, France (giraud@cerfacs.fr, gratton@cerfacs.fr).
‡Department of Computer Science, University of Tennessee, Knoxville, TN 37996-3450 (langou@cs.utk.edu).

prescribed criterion is satisfied. This has given rise to the family of iterated modified Gram–Schmidt algorithms, which differ in the criterion they use to enforce the reorthogonalization (see, e.g., [3, 10, 16]). An alternative way to compensate for the lack of orthogonality in $\bar{Q}$ is derived in [2] for a wide class of problems, including the linear least squares problem and computation of the minimum 2-norm solution of an underdetermined linear system and the projection of a vector onto a subspace. A careful use of $\bar{Q}$ and $\bar{R}$, based on an equivalence of MGS on A and Householder QR on an augmented matrix obtained by putting a matrix of zeros on top of A, leads to a backward stable algorithm. Such a strategy implies—in general—that the use of $\bar{Q}$ is computationally more expensive than would be the use of a Q-factor with orthonormal columns.

The error analyses related to the loss of orthogonality, used to derive the successful methods mentioned above, are based on the study of the quantity $\|\bar{Q}^T\bar{Q} - I_n\|$. We propose here to adopt a different approach by inspecting not only the largest singular value, as actually done in the related literature, but each singular value of the matrices involved in MGS. We denote by σ_i, $i = 1, \ldots, n$, the singular values of A, $\sigma_1 \geq \cdots \geq \sigma_n > 0$, by $\kappa = \sigma_1/\sigma_n$ the spectral condition number of A. Also we define the reduced condition numbers via the following definition.

DEFINITION 1.1. *Let κ_i, the reduced condition number, be defined by $\kappa_i = \sigma_1/\sigma_{n-i+1}, i = 1, \ldots, n$.*

Finally let $\widetilde{Q}$ be the matrix obtained from $\bar{Q}$ by normalizing its columns. In this paper, we exhibit a series of low rank matrices F_k, $k = 0, \ldots, n-1$ that enables us to update the factor $\widetilde{Q}$ such that

- $\mathrm{rank}(F_k) \leq k$,
- the columns of $\widetilde{Q} + F_k$ are orthonormal up to machine precision times κ_k, if $k = n-1$, then the columns of $\widetilde{Q} + F_{n-1}$ are exactly orthonormal,
- $(\widetilde{Q} + F_k)\bar{R}$ represents A up to machine precision.

In the case $k = 0$, $F_0 = 0$ so $(\widetilde{Q} + F_0) = \widetilde{Q}$ and the results obtained are of the same essence as the ones by Björck [1]. Namely, MGS generates a Q-factor such that the columns of $\widetilde{Q}$ are orthonormal up to machine precision times $\kappa = \kappa_0$ and $\widetilde{Q}\bar{R}$ represents A up to machine precision. In the case $k = n-1$, $(\widetilde{Q} + F_{n-1})$ is indeed the same matrix as $\hat{Q}$, the matrix exhibited by Björck and Paige [2]. That is, $\hat{Q}$ has orthonormal columns and $\hat{Q}\bar{R}$ represents A up to machine precision. Our result can be seen as a theoretical bridge that links the result of Björck [1] to the result of Björck and Paige [2]. An algorithm to compute F_k, $k = 0, \ldots, n-1$, is also derived. In our experiments this algorithm behaves well in the presence of rounding errors. For example, when κ_k is close to one, the update of $\bar{Q}$ with F_k produces a Q-factor with columns orthonormal up to machine precision. The complexity of this algorithm increases with k. For small k, its complexity is competitive with other standard reorthogonalization techniques. We conclude our study with an application of this algorithm in the framework of the solution of unsymmetric linear systems with multiple right-hand sides where a seed variant of GMRES can be successfully used.

In the remainder of this paper, for any $m \times n$ matrix X, we denote by $\sigma_i(X)$, $i = 1, \ldots, n$ the singular values of X ordered such that $\sigma_1(X) \geq \cdots \geq \sigma_n(X)$. We note that the work of this paper can be extended to complex arithmetic as well.

2. Rank considerations related to the loss of orthogonality in MGS.

2.1. Introduction. A rigorous measure of the orthogonality of an $m \times n$ matrix $\bar{Q}$ can be defined to be the distance, in the spectral 2-norm, to the set $\mathcal{O}(m, n)$ of

$m \times n$ matrices with orthonormal columns

$$\min_{V \in \mathcal{O}(m,n)} \|\bar{Q} - V\|.$$

Fan and Hoffman in [4] for the case $m = n$ and Higham in [9] for the general case $n \leq m$ proved that the minimum is attained for V being the unitary polar factor of $\bar{Q}$. The easily computed quantity $\|I_n - \bar{Q}^T\bar{Q}\|$ is often preferred to measure the orthogonality, because, as shown in Lemma 2.1, it has the same order of magnitude as $\min_{V \in \mathcal{O}(m,n)} \|\bar{Q} - V\|$ when $\|\bar{Q}\|$ is close to one.

LEMMA 2.1 (see [9]). *Let $\bar{Q} \in \mathbb{R}^{m \times n}$, $n \leq m$,*

$$\frac{\|I_n - \bar{Q}^T\bar{Q}\|}{1 + \|\bar{Q}\|} \leq \min_{V \in \mathcal{O}(m,n)} \|\bar{Q} - V\| \leq \|I_n - \bar{Q}^T\bar{Q}\|.$$

Lemma 2.1 can be generalized into Lemma 2.2.

LEMMA 2.2 ([6, Lemma 1.2]). *Let $\bar{Q} \in \mathbb{R}^{m \times n}$, $n \leq m$,*

$$\frac{\sigma_i(\bar{Q}^T\bar{Q} - I_n)}{1 + \|\bar{Q}\|} \leq \sigma_i(\bar{Q} - U) \leq \sigma_i(\bar{Q}^T\bar{Q} - I_n),$$

where $i = 1, \ldots, n$ and U is the unitary polar factor associated with $\bar{Q}$.

An important consequence of Lemma 2.2 is that if $\bar{Q}$ does not have orthonormal columns, but if $\bar{Q}^T\bar{Q} - I_n$ has only k nonzero singular values, $\bar{Q}$ is at most a rank-k modification of a matrix with orthonormal columns (namely, U).

In section 2.3, we derive a result for MGS that is similar in essence to Lemma 2.2. However, for any $k \leq n$, the MGS context will enable us to find explicitly a rank-k matrix F_k such that $\bar{Q} + F_k$ has an improved orthogonality compared with $\bar{Q}$ and such that the product $(\bar{Q} + F_k)\bar{R}$ still accurately represents A.

2.2. Some useful background related to MGS in floating point arithmetic. A key result to understanding the loss of orthogonality in MGS in floating point arithmetic is that MGS on A can be interpreted as a Householder QR-factorization on $A_{aug} = \left[\begin{smallmatrix} O_n \\ A \end{smallmatrix}\right]$, where O_n is the square zero matrix of order n [2]. Since we elaborate our work on results and techniques presented in [2], we briefly outline them below.

The use of Wilkinson's analysis of Householder transformations [19, pp. 153–162] on A_{aug} enables Björck and Paige [2, equation (3.3)] to give an orthogonal transformation $\widetilde{P}$ such that

$$\begin{pmatrix} E_1 \\ A + E_2 \end{pmatrix} = \widetilde{P} \begin{pmatrix} \bar{R} \\ 0 \end{pmatrix} = \begin{pmatrix} \widetilde{P}_{11} \\ \widetilde{P}_{21} \end{pmatrix} \bar{R}, \qquad \|E_i\| \leq \bar{c}_i u \|A\|, \quad i = 1, 2, \tag{2.1}$$

where $\bar{c}_i$ are constant depending on m, n and the detail of the arithmetic, and u is the unit roundoff. Here $\widetilde{P}_{11}$ is strictly upper triangular; see [2, section 4 and (4.1)].

Let $\widetilde{Q} = [\widetilde{q}_1, \ldots, \widetilde{q}_n]$ be the matrix obtained from $\bar{Q} = [\bar{q}_1, \ldots, \bar{q}_n]$ by normalizing its columns ($\widetilde{q}_i = \bar{q}_i / \|\bar{q}_i\|$). The equality $\widetilde{P}_{21} = \widetilde{Q}(I_n - \widetilde{P}_{11})$ holds [2, equation (4.5)] and the residual error of the polar factor $\hat{Q}$ of $\widetilde{P}_{21}$ can be bounded, as follows,

$$\|A - \hat{Q}\bar{R}\| \leq \bar{c} u \|A\|, \tag{2.2}$$

where $\bar{c} = \bar{c}_1 + \bar{c}_2$, provided that $\bar{c}u\kappa < 1$ [2, equation (3.7)]. Finally, let $\bar{\sigma}_1 \geq \cdots \geq \bar{\sigma}_n$ be the singular values of $\bar{R}$. The singular values of $\bar{R}$ approximate those of A in the following sense: $|\bar{\sigma}_i - \sigma_i| \leq \bar{c}u\sigma_1$ [2, equation (3.8)]. This implies that under the assumption $\bar{c}u\kappa < 1$, $\bar{R}$ has full rank n.

2.3. Recapture of the orthogonality in MGS. As $\begin{pmatrix} \tilde{P}_{11} \\ \tilde{P}_{21} \end{pmatrix}$ has orthonormal columns and $n \leq m$, we consider its CS decomposition [7, p. 77] defined by

$$\begin{aligned} \tilde{P}_{11} &= UCW^T, \\ \tilde{P}_{21} &= VSW^T, \end{aligned} \tag{2.3}$$

where C is singular since $\tilde{P}_{11}$ is strictly upper triangular, the entries of S are in increasing order ($0 \leq s_1 \leq \cdots \leq s_n = 1$), the entries of C are in decreasing order ($1 \geq c_1 \geq \cdots \geq c_n = 0$), and $C^2 + S^2 = I_n$, and the three matrices U, V, W have orthonormal columns. C, S, U, and W are $n \times n$, V is $m \times n$. Similarly as in [2], we suppose that A is not too ill–conditioned, by assuming that $(\bar{c}_1 + \bar{c})u\kappa < 1$ or equivalently (since this implies both $\bar{c}u\kappa < 1$ and $\bar{c}_1 u\kappa < 1 - \bar{c}u\kappa$)

$$\bar{c}_1 u\eta\kappa < 1, \tag{2.4}$$

where $\eta = (1 - \bar{c}u\kappa)^{-1}$. This has the following consequence. Since the leading element of C is (using (2.1)) $c_1 = \|\tilde{P}_{11}\| = \|E_1\bar{R}^{-1}\| \leq \bar{c}_1 u\sigma_1/\bar{\sigma}_n$, and since from (2.2) $|\sigma_n - \bar{\sigma}_n| \leq \bar{c}u\sigma_1$, we see that $\bar{\sigma}_n \geq \sigma_n - \leq \bar{c}u\sigma_1 = \sigma_n\eta$, and it follows $c_1 \leq \bar{c}_1 u\eta\kappa < 1$ (see [2, equation (3.11)]), all the s_i are nonzero, and thus S is nonsingular.

Our goal is to improve the orthogonality of the Q-factor while maintaining the residual error, $\|A - Q\bar{R}\|/\|A\|$, at the level of the machine precision. Since $\hat{Q}$ has orthonormal columns and (2.2) holds, $\hat{Q}$ answers our question. Therefore, a straightforward but expensive way to achieve our goal would be to compute $\hat{Q}$ with $\hat{Q} = VW^T$ [7, p. 149]. Let us evaluate $F = \hat{Q} - \widetilde{Q}$ to find matrices that approximate the difference between $\hat{Q}$ and $\widetilde{Q}$ at low computational cost. Since $\hat{Q} = VW^T$, using $\widetilde{P}_{21} = \widetilde{Q}(I_n - \widetilde{P}_{11})$ (see section 2.3), the CS decomposition (2.3), and the fact that S is nonsingular, we get

$$\begin{aligned} F &= \widetilde{Q}\left((I_n - \widetilde{P}_{11})WS^{-1}(I_n - S)W^T - \widetilde{P}_{11}\right), \\ &= \widetilde{Q}\left(W(S^{-1} - I_n) - UCS^{-1}\right)W^T. \end{aligned} \tag{2.5}$$

We define the truncated matrices U_k, V_k, and W_k by retaining the first k columns in their counterparts U, V, and W. In Matlab-style notation, it reads $U_k = U(:, 1:k)$. We also denote by C_k (resp., S_k) the diagonal matrix of order k whose diagonal entries are the $c_i, i = 1, \ldots, k$ (resp., $s_i, i = 1, \ldots, k$).

We define the matrix F_k obtained by setting the c_l and the s_l, $l > k$, to zero and one, respectively, in (2.5); this gives

$$F_k = \widetilde{Q}(W_k(S_k^{-1} - I_k) - U_k C_k S_k^{-1})W_k^T, \tag{2.6}$$

so that $F_0 = 0$, $F_{n-1} = F_n = F$, since $s_n = 1$ and $c_n = 0$. The matrix $\widetilde{Q} + F$ has orthonormal columns and accurately represents A when multiplied on the right by $\bar{R}$. Theorem 2.3 shows how these properties are modified when the matrix $\widetilde{Q} + F_k$ is considered instead. The matrices Q_k are then a sequence of matrices going from the matrix of normalized vectors from MGS $Q_0 = \widetilde{Q}$ to the matrix of orthogonal vectors $Q_{n-1} = \hat{Q}$.

THEOREM 2.3. *Assume that* $\bar{c}_1 u\eta\kappa < 1$, *for* $k = 0, \ldots, n-1$; *the matrix* Q_k *defined by*

$$Q_k = \widetilde{Q} + F_k \tag{2.7}$$

enjoys the following properties:

(a)

$$rank(Q_k - \widetilde{Q}) \leq k,$$

(b)

$$\|A - Q_k\bar{R}\| \leq \left(\bar{c}_2 + 2\bar{c}_1 \frac{(1+\bar{c}_1 u\eta\kappa)}{(1-\bar{c}_1 u\eta\kappa)^2}\right) u\|A\|,$$

(c) *for* $k = 0, \ldots, n-2$,

$$\|I_n - Q_k^T Q_k\| \leq \left(2\bar{c}_1\eta \frac{(1+\bar{c}_1 u\eta\kappa)^2}{(1-\bar{c}_1 u\eta\kappa)^3}\right) u\kappa_{k+1}$$

for $k = n-1$, $Q_{n-1} = \hat{Q}$, *and so* $\|I_n - Q_{n-1}^T Q_{n-1}\| = 0$.

Proof. Part (a) is a consequence of the definition (2.6) of F_k. We then establish part (b) of this theorem. From (2.1), $\widetilde{P}_{11}\bar{R} = E_1$, and multiplying to the left by U_k^T implies that $U_k^T UCW^T\bar{R} = U_k^T E_1$. Using the definition of the truncated matrices C_k and W_k, one gets $C_k W_k^T\bar{R} = U_k^T E_1$, and, taking norms, $\|C_k W_k^T\bar{R}\| = \|U_k^T E_1\| \leq \|E_1\|$. From (2.1), $\|E_1\| \leq \bar{c}_1 u\|A\|$, we obtain a first intermediate result,

$$\|C_k W_k^T\bar{R}\| \leq \bar{c}_1 u\|A\|. \tag{2.8}$$

Let us bound the residual error $\|A - Q_k\bar{R}\|$. Using the triangular inequality yields

$$\|A - Q_k\bar{R}\| \leq \|A - \widetilde{Q}\bar{R}\| + \|F_k\bar{R}\|. \tag{2.9}$$

The first term of the right-hand side can be bounded using [6, Lemma A.2]. We study the second term of the right-hand side: $\|F_k\bar{R}\|$. By definition (2.6) of F_k,

$$F_k\bar{R} = \widetilde{Q}(W_k(S_k^{-1} - I_k) - U_k C_k S_k^{-1})(W_k^T\bar{R}).$$

Applying the result of [6, Lemma A.3] to $(S_k^{-1} - I_k)W_k^T\bar{R}$ and noticing that, from $c_i^2 + s_i^2 = 1$, we have

$$s_i^{-1} - 1 = (1-c_i^2)^{-1/2} - 1 = \frac{c_i^2}{\sqrt{1-c_i^2}(1+\sqrt{1-c_i^2})} \leq \frac{c_i^2}{2(1-c_i^2)}, \tag{2.10}$$

so $s_i^{-1} - 1 \leq c_i \frac{c_1}{2(1-c_1)}$, we get $\|(S_k^{-1} - I_k)W_k^T\bar{R}\| \leq \frac{\|C_k\|}{2(1-\|C_k\|)}\|C_k W_k^T\bar{R}\|$, from which follows that

$$\|F_k\bar{R}\| \leq \|\widetilde{Q}\|\left(\frac{\|C_k\|}{2(1-\|C_k\|)} + \|S_k^{-1}\|\right)\|C_k W_k^T\bar{R}\|, \tag{2.11}$$

where we have used the fact that the two matrices C_k and S_k^{-1} being diagonal, they commute. We recall after (2.4) that $\|C_k\| \leq \bar{c}_1 u\eta\kappa < 1$ and therefore $\|S_k^{-1}\| \leq (1-(\bar{c}_1 u\eta\kappa)^2)^{-1/2} \leq (1-(\bar{c}_1 u\eta\kappa))^{-1}$. Using [6, Lemma A.1],

$$\|F_k\bar{R}\| \leq \frac{(1+\bar{c}_1 u\eta\kappa)^2}{(1-\bar{c}_1 u\eta\kappa)^2}\bar{c}_1 u\|A\|.$$

With [6, Lemma A.2], we end the proof of part (b).

TABLE 2.1
Correspondence between the bounds in Theorem 2.3 and the results of Björck and Paige [2].

Theorem 2.3, Part (b), $k = 0$	[6, Lemma A.2] derived from Björck and Paige [2]
$\|A - \widetilde{Q}\bar{R}\| \leq \left(\bar{c}_2 + 2\bar{c}_1 \frac{(1+\bar{c}_1 u\eta\kappa)}{(1-\bar{c}_1 u\eta\kappa)^2}\right) u\|A\|$	$\|A - \widetilde{Q}\bar{R}\| \leq \left(\bar{c}_2 + \bar{c}_1 \frac{1+\bar{c}_1 u\eta\kappa}{1-\bar{c}_1 u\eta\kappa}\right) u\|A\|$
Theorem 2.3, Part (b), $k = n-1$	Björck and Paige [2, equation (3.7)]
$\|A - \hat{Q}\bar{R}\| \leq \left(\bar{c}_2 + 2\bar{c}_1 \frac{(1+\bar{c}_1 u\eta\kappa)}{(1-\bar{c}_1 u\eta\kappa)^2}\right) u\|A\|$	$\|A - \hat{Q}\bar{R}\| \leq (c_1 + c_2)u\|A\|$
Theorem 2.3, Part (c), $k = 0$	Björck and Paige [2, equation (5.3)]
$\|I_n - \widetilde{Q}^T\widetilde{Q}\| \leq \left(2\bar{c}_1\eta \frac{(1+\bar{c}_1 u\eta\kappa)^2}{(1-\bar{c}_1 u\eta\kappa)^3}\right) u\kappa$	$\|I_n - \widetilde{Q}^T\widetilde{Q}\| \leq \frac{2c_1}{1-(c+c_1)u\kappa} u\kappa$
Theorem 2.3, Part (c), $k = n-1$	Björck and Paige [2, equation (3.7)]
$\|I_n - \hat{Q}^T\hat{Q}\| = 0$	$\|I_n - \hat{Q}^T\hat{Q}\| = 0$

We now prove part (c) of the theorem. We define the matrices $U_{\bar{k}}$, $V_{\bar{k}}$, $W_{\bar{k}}$ so that $U = [U_k, U_{\bar{k}}]$, and similarly for V and W. In Matlab-style notation, $U_{\bar{k}} = U(:, k+1:n)$. We also define the matrices $C_{\bar{k}}$ (resp., $S_{\bar{k}}$), the diagonal matrix of order $n-k+1$ whose diagonal elements are the c_i, $i = k+1, \ldots, n$ (resp., s_i $i = k+1, \ldots, n$). One has

$$\hat{Q} - Q_k = F - F_k, \tag{2.12}$$

$$\hat{Q} - Q_k = \widetilde{Q}\left(W_{\bar{k}}(S_{\bar{k}}^{-1} - I_{n-k+1}) - U_{\bar{k}}C_{\bar{k}}S_{\bar{k}}^{-1}\right) W_{\bar{k}}^T, \tag{2.13}$$

$$\|\hat{Q} - Q_k\| \leq \|\widetilde{Q}\| \left(\|S_{\bar{k}}^{-1} - I_{n-k+1}\| + \|C_{\bar{k}}S_{\bar{k}}^{-1}\|\right). \tag{2.14}$$

Since both the s_i and the c_i belong to $[0, 1]$ and the c_i (resp., the s_i) are sorted in decreasing (resp., increasing) order, one obtains

$$\|\hat{Q} - Q_k\| \leq \|\widetilde{Q}\| \left((s_{k+1}^{-1} - 1) + s_{k+1}^{-1}c_{k+1}\right),$$

which yields, using (2.10),

$$\|\hat{Q} - Q_k\| \leq \|\widetilde{Q}\|\ c_{k+1}\frac{1 + c_{k+1}}{1 - c_{k+1}}.$$

From [6, Lemma A.1], and using the fact that $c_{k+1} \leq c_1 \leq \bar{c}_1 u\eta\kappa$, we get

$$\|\hat{Q} - Q_k\| \leq \frac{(1 + \bar{c}_1 u\eta\kappa)^2}{(1 - \bar{c}_1 u\eta\kappa)^2} c_{k+1}. \tag{2.15}$$

Since $\widetilde{P}_{11} = E_1\bar{R}^{-1}$, and c_{k+1} is the $(k+1)$th singular value of $\widetilde{P}_{11}$, one has $c_{k+1} \leq \|E_1\|\sigma_{k+1}(\bar{R}^{-1})$. From $\sigma_{k+1}(\bar{R}^{-1}) = 1/\bar{\sigma}_{n-k}$, $\|E_1\| \leq \bar{c}_1 u\|A\|$, $|\bar{\sigma}_{k+1} - \sigma_{k+1}| \leq \bar{c}u\sigma_1$, $\kappa_{k+1} = \frac{\|A\|}{\sigma_{n-k}}$, and $\eta = (1 - \bar{c}u\kappa)^{-1}$, we obtain $c_{k+1} \leq \bar{c}_1 u\frac{\|A\|}{\bar{\sigma}_{n-k}} \leq \bar{c}_1 u\eta\frac{\|A\|}{\sigma_{n-k}} = \bar{c}_1 u\eta\kappa_{k+1}$, and the conclusion for the case $k = 0, \ldots, n-2$ follows using Lemma 2.1. For the case $k = n-1$, this bound on c_n gives $c_n \leq \bar{c}_1 u\eta$, which is not satisfactory. Since $\widetilde{P}_{11}$ is strictly upper triangular, a better bound is $c_n = 0$ from which we recover with (2.15) that $\hat{Q} = Q_{n-1}$. ∎

Several remarks can be made. First, consistency, $\|A - Q_k\bar{R}\|/\|A\|$, is maintained close to machine precision independently of the rank-k of the update. In the introduction, we explain that in the case $k = 0$ and $k = n-1$, we recover the result of Björck [1] for $\widetilde{Q} = Q_0$ and Björck and Paige [2] for $\hat{Q} = Q_{n-1}$, respectively. A consequence of this unified framework is that the bounds given are larger than the original ones but remain very close. In Table 2.1, we summarize the relations to be compared. Note that the results of Björck [1] have been replaced by analogous results of Björck and Paige [2] in order to compare the same quantities.

TABLE 3.1
Algorithm 1: *MGS with an a posteriori reorthogonalization by a rank-k update.*

1. Run MGS on A to obtain $\bar{Q}$ and $\bar{R}$.
2. Compute $\bar{T}$, the strictly upper triangular matrix with entry (i,j), $\bar{q}_i^T\bar{q}_j$, $(i<j)$, then form $\bar{P}_{11} = (I_n + \bar{T})^{-1}\bar{T}$.
3. Compute the k largest singular values of $\bar{P}_{11}$, c_i, $i = 1,\ldots,n$, and the associated k right (resp., left) singular vectors U_k (resp., W_k) finally form $s_i = \sqrt{1-c_i^2}$, $i = 1,\ldots,k$. The matrix C_k (resp., S_k) is the $k \times k$ diagonal matrix with entry (i,i) equal to c_i (resp., s_i).
4. Form $Q_k = \bar{Q} + \bar{Q}(W_k(S_k^{-1} - I_k) - U_kC_kS_k^{-1})W_k^T$.

3. Numerical illustrations and examples of application.

3.1. Numerical illustrations of the bounds in Theorem 2.3. The aims of this section are twofold. First, we give an algorithm to compute the approximations $\bar{F}_k$ (resp., $\bar{Q}_k$) of the matrices F_k (resp., Q_k), then we verify numerically that Theorem 2.3 is satisfied with these computed quantities up to machine precision.

To ensure that the rank-k property of the $m \times n$ matrix F_k is inherited by the computed matrix $\bar{F}_k$, we define $\bar{F}_k$ as the product of the $m \times k$ computed quantities $\bar{Q}(W_k(S_k^{-1} - I_k) - U_kC_kS_k^{-1})$ times the $k \times n$ rectangular matrix W_k^T. Then by construction, the first statement (a) of Theorem 2.3 is satisfied and we can now focus on the last two statements and show that the bounds are sharp.

In the following, the notation F_k (resp., Q_k) stands for the the computed quantity $\bar{F}_k$ (resp., $\bar{Q}_k$). For the experiments, we proceed as follows. Starting from an initial matrix A, we run MGS to obtain $\bar{Q}$ and $\bar{R}$. Then for each k from $k = 0$ to $n-1$, we compute the associated matrix Q_k using formulas (2.6) and (2.7). In that respect, we need to compute $\widetilde{P}_{11}$. In [2, equation (4.1)], Björck and Paige show that $\widetilde{P}_{11}$ is strictly upper triangular with element (i,j) equal to $\tilde{q}_i^T(I_m - \tilde{q}_1\tilde{q}_1^T)\ldots(I_m - \tilde{q}_{j-1}\tilde{q}_{j-1}^T)\tilde{q}_j$ for $i<j$. We define $\widetilde{T}$ such that $\widetilde{T}$ is strictly upper triangular with element (i,j), $\tilde{q}_i^T\tilde{q}_j$, $(i<j)$. Since $\|\tilde{q}_i\| = 1$ for all i, one may notice that $(I_n + \widetilde{T})(I_n - \widetilde{P}_{11}) = I_n$, which can also be written

$$\widetilde{P}_{11} = (I_n + \widetilde{T})^{-1}\widetilde{T}. \tag{3.1}$$

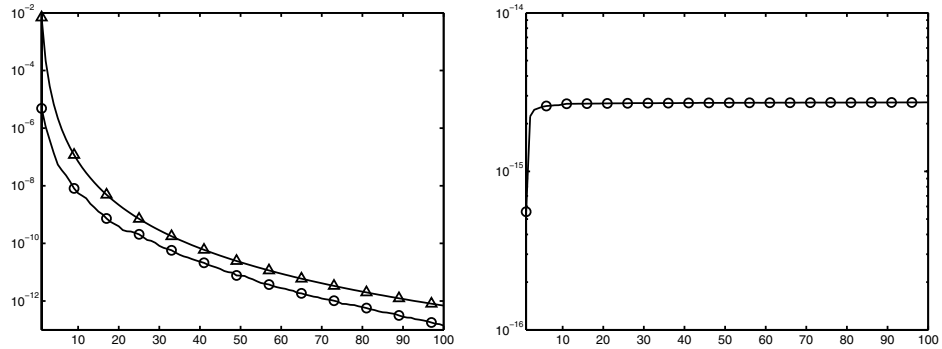
The matrix $\widetilde{P}_{11}$ is also closely related to the T-factor of the YTY-representation of the matrix $\widetilde{P}$ (see Schreiber and Van Loan [12]). Calling $\widetilde{Y} = \begin{pmatrix} -I_n \\ \tilde{Q} \end{pmatrix}$ and $\widetilde{T}_{YTY} = -I_n + \widetilde{P}_{11}$, we have

$$\widetilde{P} = I + \widetilde{Y}\widetilde{T}_{YTY}\widetilde{Y}^T,$$

and developing this expression leads directly to [2, equation (4.2)].

Note that in practice the mathematical quantities $\tilde{q}_i$ are replaced by the computed quantities $\bar{q}_i$. Equation (3.1) is preferred to the original equation of Björck and Paige [2, equation (4.1)] since it enables us to compute $\widetilde{P}_{11}$ with significantly fewer flops when m is large compared to n. We summarize the corresponding algorithm in Table 3.1.

In this section, the numerical experiments are run with Matlab 6, where the unit roundoff is $u \approx 1.1 \cdot 10^{-16}$. We consider two test matrices, which are the matrices



(a) $\triangle\ u\kappa_{k+1}$, $\circ\ \|I_n - Q_k^T Q_k\|$ (b) $\circ\ \|A - Q_k\bar{R}\|/\|A\|$

FIG. 3.1. *Illustrations of bounds* (b) *and* (c) *of Theorem* 2.3 *for matrix* $P(1500, 500, 1, 5)$.

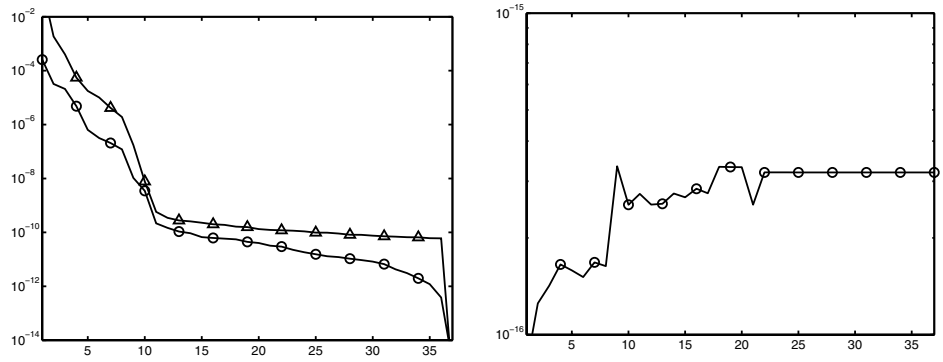
$P(1500, 500, 1, 5)$ from Paige and Saunders [13] and GRE_216B from Matrix Market.[1] The first is a 1500×500 matrix with condition number 10^{16} and the second is a 216×216 matrix with condition number $6 \cdot 10^{14}$. On those two matrices we investigate how sharp the bounds (b) and (c) in Theorem 2.3 are.

To quantify the orthogonality quality of the columns of different matrices, we define the level of orthogonality of Q as the quantity $\|I_n - Q^T Q\|$. In Figure 3.1(a), we plot the "recovered orthogonality" with $\circ$. For $k = 0$, we have $Q_0 = \bar{Q}$; therefore we simply plot the level of orthogonality obtained after the run of MGS on $P(1500, 500, 1, 5)$. For $k = 1$, we correct $\bar{Q}$ by the rank-one update F_1 to obtain Q_1 and then plot the level of orthogonality of Q_1. While k increases, we observe the benefit of adding F_k to $\bar{Q}$ on the orthogonality quality. We stop the plot at $k = 100$. At this step, the matrix Q_{100} has nearly reached its final level of orthogonality ($1.44 \cdot 10^{-14}$ for $k = n - 1$). With $\triangle$, we plot the corresponding $u\kappa_{k+1}$, $k = 0, \ldots, n-1$. The theorem predicts that for each k, $\|I_n - Q_k^T Q_k\|$ is bounded above by $u\kappa_{k+1}$ times a constant. In this experiment we observe that both curves fit well. This indicates that the constant can be taken close to one for these experiments and that the bound (c) of Theorem 2.3 is tight. In Figure 3.1(b), we illustrate that property (b) of Theorem 2.3 holds. In this case $\|A - Q_k\bar{R}\|$ is smaller than $u\|A\|$ times a constant where the constant is small.

Similar experiments are reported in Figure 3.2 for the matrix GRE_216B that also illustrates the tightness of the bounds.

Given the singular value distribution of A and the machine precision, Theorem 2.3 gives us a set of k for which all the associated matrices Q_k satisfy a prescribed level of orthogonality. Since the amount of work of Algorithm 1 (Table 3.1) increases with k, we can choose the lowest k of this set and update $\bar{Q}$ with the rank-k matrix F_k. Therefore an interesting feature of Algorithm 1 is that it is able to adapt its amount of work with respect to the level of orthogonality expected. For example, if the level of orthogonality required for the Q-factor of matrix GRE_216B is 10^{-9}, with both Theorem 2.3 and the knowledge of $u\kappa_{k+1}$, we can choose $k = 10$. Meanwhile, if the level of orthogonality required is 10^{-14}, we can estimate the value a priori $k = 37$. A posteriori we observe in Figure 3.2 and curve $\|I_n - Q_k^T Q_k\|$ that these two choices are correct.

[1] http://math.nist.gov/MatrixMarket/



(a) $\triangle\ u\kappa_{k+1}$, $\circ\ \|I_n - Q_k^T Q_k\|$ (b) $\circ\ \|A - Q_k\bar{R}\|/\|A\|$

FIG. 3.2. *Illustrations of bounds* (b) *and* (c) *of Theorem* 2.3 *for matrix GRE_216B.*

3.2. An application of choice: Seed-GMRES. A practical framework where our algorithm fits perfectly is the seed-GMRES method for solving a sequence of linear systems with the same coefficient matrix but for a sequence of different right-hand sides. Roughly speaking, one solves the linear system for one right-hand side at a time but uses the Krylov space associated with the current right-hand side to compute a good initial guess for the next ones.

Let us now briefly describe the seed-GMRES method and the various alternatives we consider to compare with our algorithm. Let Z be a square matrix of order m with full rank. We want to solve the linear systems $Zx^{(i)} = b^{(i)}$ for $i = 0, \ldots, p$ by using seed-GMRES with MGS (see, e.g., [14, 17]). For the sake of clarity, but without loss of generality, we describe the method assuming that the initial guesses for all the right-hand sides are zeros, and we illustrate it only when the first right-hand side has converged. For the next ones, the same algorithm applies but the initial guesses are no longer zero, making the notation more complicated for a purpose that is out of the scope of this paper.

We first run GMRES with MGS to solve the linear system $Zx^{(0)} = b^{(0)}$. This amounts to solving the linear least squares problem

$$\min_{y \in \mathbb{R}^{n-1}} \|b^{(0)} - ZV_{n-1}^{(0)} y\|,$$

where $V_{n-1}^{(0)}$ is a set of $n-1$ vectors built with an Arnoldi process on Z using the starting vector $b^{(0)}$ and orthogonalization scheme MGS. In most applications, the computational burden lies in the matrix-vector products and the scalar products required to solve this linear least squares problem. In seed-GMRES, the subsequent right-hand sides benefit from this work. An effective initial guess $x^{(i)} = V_{n-1}^{(0)} y^{(i)}$ for the system i is obtained by solving the same linear least squares problem but with another right-hand side, namely,

$$\min_{y \in \mathbb{R}^{n-1}} \|b^{(i)} - ZV_{n-1}^{(0)} y\|.$$

We first compare four approaches to solve this problem. In the first part, we present two standard algorithms and compare them in terms of floating point operations (flops) with an approach implementing Algorithm 1. In the second part, one

aspect of our problem is examined in more detail to show that—under reasonable assumptions—a rank-one update is enough to recover with Algorithm 1 a good level of orthogonality. In this particular case, a second algorithm is also derived based on an heuristic that enables us to substantially save computational work. Finally, we illustrate the effectiveness of our approach when embedded in large electromagnetism applications.

In what follows, the superscript $^{(0)}$ is omitted and the matrix A denotes the computed matrix $(b^{(0)}, ZV_{n-1}^{(0)})$, similar to the first section, and MGS is run on A of size $m \times n$ in a well-designed floating point arithmetic to obtain $\bar{Q}$ and $\bar{R}$. Indeed, the Arnoldi process gives $\bar{Q} = V_n^{(0)}$ but for the sake of generality this property is not taken into account.

3.2.1. The three approaches. Since we already have computed the QR-factorization of (b, ZV_{n-1}) via MGS, an efficient way to solve the linear least squares with $b^{(i)}$ is to compute the R-factor of the QR-factorization of $\left(b, ZV_{n-1}, b^{(i)}\right)$ via MGS. In practice it remains to compute the last column of this R-factor, that is, $c_{\mathrm{MGS}}^{(i)}$, such that

$$c_{\mathrm{MGS}}^{(i)} = \begin{pmatrix} \bar{q}_1^T b^{(i)} \\ \vdots \\ \bar{q}_n^T (I_m - \bar{q}_{n-1}\bar{q}_{n-1}^T) \dots (I_m - \bar{q}_1 \bar{q}_1^T) b^{(i)} \end{pmatrix}. \tag{3.2}$$

From a complexity point of view the MGS algorithm applied to A requires $2mn^2$ flops while the $(p-1)$ projections (3.2) for the remaining right-hand sides require $4mnp$ flops.

A second way is to reorthogonalize $\bar{Q}$, the Q-factor from MGS, before performing the set of projections. We reorthogonalize $\bar{Q}$ to obtain Q_k using formula (2.6), with Algorithm 1. The value of k is chosen large enough so that Q_k has columns orthonormal up to machine precision. Then we project the $(p-1)$ remaining right-hand sides with classical Gram–Schmidt type projections, which is

$$c_{\mathrm{CGS}}^{(i)} = \begin{pmatrix} \hat{q}_1^T b^{(i)} \\ \vdots \\ \hat{q}_n^T b^{(i)} \end{pmatrix}. \tag{3.3}$$

This latter approach still requires $2mn^2$ flops to get the QR-factorization of A but only $2mnp$ flops for the $(p-1)$ projections. However, we have to add the cost of the reorthogonalization that is mainly governed by the construction of T, which is mn^2 flops, plus the assembly of Q_k with (2.6), which is $4mnk$ flops.

A third approach consists in not using MGS as an orthogonalization scheme in GMRES but instead iterated modified Gram–Schmidt with a criterion denoted by MGS2(K) [10]. The extra costs compared with MGS come from the reorthogonalizations. We call ν the quantity so that the cost of MGS2(K) is $2mn^2\nu$; ν ranges from 1 (if no reorthogonalization is performed) to 2 (if one reorthogonalization per column is performed). The parameter K defines the criterion used to decide whether the reorthogonalization has to be performed. According to [3], we choose the value $K = \sqrt{2}$ and justify this choice later through numerical experiments. The aim here is to obtain directly an orthogonal basis to machine precision and then use (3.3) with the Q-factor obtained with MGS2(K).

TABLE 3.2
Flops required for the three orthogonalization schemes and associated projection considering m large over k, n, and p.

MGS and $(p-1)$ projections (3.2)	$2mn^2 + 4mnp$
Algorithm 1 and $(p-1)$ projections (3.3)	$2mn^2 + 2mnp + mn(n+4k)$
MGS2(K) and $(p-1)$ projections (3.3)	$2mn^2\nu + 2mnp$

We summarize the costs in flops of these three approaches in Table 3.2. From Table 3.2, for rather small p a good approach in term of flops seems to be MGS. However, our interest is in large p. For large p, Algorithm 1 is interesting over MGS2(K) when $\frac{3}{2} + \frac{2k}{n} \leq \nu$. We have seen that the parameter k is determined a priori by the level of orthogonality required by the user. In the rest of the paper, we consider k small compared to n, and the critical value is then $\nu = 1.50$. A larger value for ν would make our approach more efficient than MGS2(K)—and vice versa—since the construction of T which requires mn^2 is the main cost of Algorithm 1; therefore we compare $3mn^2$ (Algorithm 1) to $2mn^2\nu$ (MGS2(K)).

3.2.2. Special feature of $A = (b, ZV_{n-1})$. Greenbaum, Rozložník, and Strakoš [8] have shown that for GMRES with orthogonalization schemes MGS, the quantity $\sigma_n((b, ZV_{n-1}))$ is of the order of the residual of GMRES obtained at step $n-1$. When the residual is small, we expect $A = (b, ZV_{n-1})$ to be ill-conditioned and so an important loss of orthogonality is expected with MGS.

Since $\sigma_{n-1}((b, ZV_{n-1})) \geq \sigma_{n-1}(ZV_{n-1}) \geq \sigma_{n-1}(Z)\sigma_n(V_{n-1})$, if we assume Z and V_{n-1} well-conditioned, we get that κ_2 is close to one. We note that if the matrix (b, ZV_{n-1}) is numerically nonsingular, then as in [5], $\bar{Q}$ $(= V_n)$ is well-conditioned and we restrict our study only to well-conditioned matrix Z. From this analysis, the value $k = 1$ is enough for the reorthogonalization of $\bar{Q}$ with Algorithm 1 to obtain a Q-factor orthogonal up to machine precision. In the experimental part, we illustrate that $k = 1$ is indeed necessary and sufficient in the seed-GMRES context.

For small k compared to n, the cost of the a posteriori reorthogonalization procedure of MGS performed with Algorithm 1 is mainly governed by the computation of the $n(n+1)/2$ entries of the matrix $\bar{T}$ (section 3.2.1). We degrade Algorithm 1 to get a second algorithm, and this algorithm relies mainly on an heuristic that attempts to avoid the complete computation of $\bar{T}$. First, we consider that the rank of $\bar{P}_{11}$ is one—this is justified by the special feature of the problem: κ large and κ_2 close to one—and since P_{11} is strictly upper triangular and therefore nilpotent (i.e., $P_{11}^n = 0$), we have $P_{11}^2 = 0$, and so (3.1) reduces to $P_{11} = T$. Therefore in practice we just compute T and use it as P_{11}. But computing all the entries of a rank-one matrix may be considered a waste of time. In theory, it is enough to build a row i and a column j so that the entry (i, j) is nonzero. With rounding errors, the best choice is to build the row i and the column j such that the entry (i, j) is the largest in magnitude. In practice, if the entry (i, j) is not the largest but of the order of the largest entry of $\bar{T}$, the procedure is still reliable. A good candidate to be of the order of the largest entry of $\bar{T}$ is $|\bar{q}_1^T \bar{q}_n|$ since the orthogonality given by MGS of $\bar{q}_n$ over $\bar{q}_1$ assumes in theory the orthogonality of all the previous vectors; in practice, we expect the loss of orthogonality in V to be maximal between $\bar{q}_n$ and $\bar{q}_1$. This defines our heuristic:

$$|\bar{q}_1^T \bar{q}_n| \text{ is of the order of the largest entry in magnitude of } \bar{T}.$$

Thanks to this heuristic, only the first row and the last column of $\bar{T}$ are computed.

TABLE 3.3

Algorithm 2: MGS with an a posteriori reorthogonalization by a rank-one update using the heuristic.

1. Run MGS on $A = (b, ZV_{n-1})$ to obtain $\bar{Q}$ and $\bar{R}$.
2. Compute $u^T = (\bar{q}_n^T \bar{q}_1, \ldots, \bar{q}_n^T \bar{q}_{n-1}, 0)$, $c = u(1)$, and $w^T = (0, \bar{q}_1^T \bar{q}_2, \ldots, \bar{q}_1^T \bar{q}_n)$.
3. $c = u(1)$, $u = u/\|u\|$, $w = w/\|w\|$, $c = c/u(1)/w(n)$, $s = \sqrt{1 - c^2}$.
4. Compute $Q_1 = \bar{Q} + \bar{Q}(w(s^{-1} - 1) - ucs^{-1})w^T$.

Algorithm 2 in Table 3.3 uses the reorthogonalization based on this heuristic. The fourth approach to compute the orthogonalization and the projections in seed-GMRES is to use Algorithm 2 and then project the $(p-1)$ other right-hand sides with (3.3). The whole algorithm is very cheap and requires only $2mn^2 + 2mnp + 8mn$ flops, in which $8mn$ flops are necessary for the reorthogonalization. For comparison, $8mn$ corresponds to the extra cost of the reorthogonalization of about four columns. Finally, let us remark that if $p < n$, then it is worth using the factorized form of Q_1 instead of computing it explicitly as suggested by line 4 of Algorithm 2.

3.2.3. Numerical experiments in a large electromagnetism calculation. Our case study arises from large calculations in electromagnetism. The boundary element method is used to discretize the three-dimensional Maxwell's equations on the surface of an object. The formulation relies on the combined field integral equations, and the preconditioner used is a sparse approximate inverse [18], and this means that in practice the preconditioned matrix Z is well-conditioned. Moreover, one can notice that the matrix Z is not explicitly known and is accessed through matrix-vector product done via the fast multipole method. All the calculations are performed using double precision arithmetic. There are several linear systems $Zx^{(i)} = b^{(i)}$ to be solved; for this typical calculation we have $p = 180$, but this value might be much larger if several radar cross sections have to be computed, as is often the case in engineering applications. For each right-hand side, GMRES is stopped at iteration l if the approximate solution $x_l^{(i)}$ verifies $\|b^{(i)} - Ax_l^{(i)}\|/\|b^{(i)}\| \leq 10^{-14}$. We remark that the problem is defined in complex arithmetic; however, to be consistent with the whole paper the real notation is maintained.

Four geometries are considered, which represent standard test cases for electromagnetism calculations, namely, a cetaf, an Airbus airplane, a sphere, and an almond [18]. In Table 3.4, we give the characteristics of the matrices (b, ZV_{n-1}) obtained by a GMRES-MGS run on these matrices. The values obtained with GMRES-MGS2(K) are the same. For more information on the method and the test case, see [18].

In Table 3.4, # iter represents the number of GMRES iterations required to converge. The number of columns of the matrix $A = (b, ZV_{n-1})$ is $n = \#\text{ iter } + 1$, and the number of rows is m. As expected (see section 3.2.2), the condition number κ is such that $\kappa \cdot 10^{-14}$ is close to one, while κ_2 is of order $\mathcal{O}(1)$.

The fourth column of Table 3.4 corresponds to the average number of reorthogonalizations obtained with MGS2($\sqrt{2}$). In this cases, MGS2($\sqrt{2}$) systematically performs an extra reorthogonalization per matrix-vector product, which explains the constant value ($\nu = 2.00$).

In Table 3.5, we illustrate that all the residual errors $\|A - \bar{Q}\bar{R}\|$—where $\bar{Q}$ and $\bar{R}$ designated the QR-factor given by one the four algorithms—are of the order of the

TABLE 3.4
Characteristics of $A = (b, ZV_{n-1})$.

	m	# iter	κ	κ_2	ν
Cetaf	5391	31	$9.7 \cdot 10^{14}$	27	2.00
Airbus	23676	104	$3.6 \cdot 10^{14}$	14	2.00
Sphere	40368	59	$3.9 \cdot 10^{14}$	6.4	2.00
Almond	104973	71	$5.1 \cdot 10^{14}$	5.9	2.00

TABLE 3.5
Residual errors for the four test cases and the different algorithms.

	MGS	Algorithm 1	MGS2($\sqrt{2}$)	Algorithm 2
Cetaf	$2.8 \cdot 10^{-17}$	$2.8 \cdot 10^{-16}$	$1.8 \cdot 10^{-16}$	$2.9 \cdot 10^{-16}$
Airbus	$4.0 \cdot 10^{-17}$	$4.4 \cdot 10^{-16}$	$2.7 \cdot 10^{-16}$	$4.4 \cdot 10^{-16}$
Sphere	$5.8 \cdot 10^{-17}$	$2.7 \cdot 10^{-16}$	$1.6 \cdot 10^{-16}$	$2.7 \cdot 10^{-16}$
Almond	$3.9 \cdot 10^{-17}$	$3.9 \cdot 10^{-16}$	$3.9 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$

TABLE 3.6
$\|I_n - \bar{Q}^T\bar{Q}\|$ for the four test cases and the different algorithms.

	MGS	Algorithm 1	MGS2($\sqrt{2}$)	Algorithm 2
Cetaf	$1.6 \cdot 10^{-02}$	$1.9 \cdot 10^{-15}$	$2.8 \cdot 10^{-16}$	$2.4 \cdot 10^{-15}$
Airbus	$1.8 \cdot 10^{-02}$	$1.5 \cdot 10^{-15}$	$3.7 \cdot 10^{-16}$	$1.6 \cdot 10^{-15}$
Sphere	$3.9 \cdot 10^{-02}$	$5.4 \cdot 10^{-16}$	$3.0 \cdot 10^{-16}$	$7.8 \cdot 10^{-16}$
Almond	$4.1 \cdot 10^{-02}$	$6.8 \cdot 10^{-16}$	$2.8 \cdot 10^{-16}$	$7.9 \cdot 10^{-16}$

TABLE 3.7
$\|I_n - \bar{Q}^T\bar{Q}\|$.

	MGS2($\sqrt{2}$)	MGS2(2)	MGS2($\sqrt{5}$)
Cetaf	$2.8 \cdot 10^{-16}$ ($\nu = 2.00$)	$6.3 \cdot 10^{-16}$ ($\nu = 1.90$)	$1.2 \cdot 10^{-15}$ ($\nu = 1.87$)
Airbus	$3.7 \cdot 10^{-16}$ ($\nu = 2.00$)	$3.9 \cdot 10^{-03}$ ($\nu = 1.02$)	$8.8 \cdot 10^{-03}$ ($\nu = 1.01$)
Sphere	$3.0 \cdot 10^{-16}$ ($\nu = 2.00$)	$7.5 \cdot 10^{-15}$ ($\nu = 1.52$)	$4.9 \cdot 10^{-04}$ ($\nu = 1.07$)
Almond	$2.8 \cdot 10^{-16}$ ($\nu = 2.00$)	$1.7 \cdot 10^{-03}$ ($\nu = 1.06$)	$5.2 \cdot 10^{-03}$ ($\nu = 1.03$)

machine precision. In Table 3.6, the different levels of orthogonality characterized with $\|I_n - \bar{Q}^T\bar{Q}\|$ are given. As expected, MGS completely loses the orthogonality while the three other approaches give a set of vectors orthogonal up to machine precision. In the context of seed-GMRES, this enables us to use confidently (3.3) to project the $(p-1)$ remaining right-hand sides.

A conclusion drawn from Table 3.6 is that in the case of GMRES-MGS applied to a not too ill-conditioned matrix, the value $k = 1$ is satisfactory (Algorithm 1 with $k = 1$). Moreover, from Table 3.5 and Table 3.6, we observe that Algorithm 2 relying on the heuristic works fine in practice.

One might question the relevance of the choice $K = \sqrt{2}$ and its possible artificial high cost. In Table 3.7 we report on the sensitivity of the orthogonality quality with respect to the choice of the threshold. These experiments assess the choice of $K = \sqrt{2}$ for MGS2(K). This value gives a good orthogonality level for all the examples while the others tested ($K = 2$ and $K = \sqrt{5}$) fail. However, $K = \sqrt{2}$ implies in these cases $\nu = 2.00$, meaning that the criterion is unable to save any reorthogonalization. This result is not satisfactory and highlights a weakness of the MGS2(K) procedure.

Even if κ_2 is close to one, improving noticeably the condition number, κ cannot be obtained in the general case by removing only a column of (b, ZV_{n-1}); it is a global phenomenon that needs a global treatment (e.g., to add the singular vector associated to the smallest singular value to all the columns). In the same way, the loss of orthogonality is global and affects all the columns of $\bar{Q}$. An algorithm like MGS2(K) that acts locally on each column performs poorly in this case, whereas Algorithms 1 and 2 represent appealing strategies since the reorthogonalization—which has to be global—is expressed as a rank-one update.

Finally, there exist other examples where the value of $k > 1$ can be given a priori. Still, for the solution of linear systems with multiple right-hand sides, we mention, for instance, the Block(k)-seed-GMRES-MGS algorithm; that is, one run Block GMRES on k vectors, when the convergence is observed, a rank-k update is performed to recover an orthogonal set of vector, which we use to project the $p-k$ right-hand sides as in seed-GMRES.

4. Conclusion. In this paper we propose an a posteriori reorthogonalization technique based on a Rank-k update to reorthogonalize a set of vectors built by the modified Gram–Schmidt algorithm. We show that for large enough k, we can fully recover the orthogonality. We illustrate the effectiveness of our technique in the framework of the iterative solution of linear systems based on the GMRES algorithm. On a set of industrial test problems we demonstrate that our algorithm is efficient and outperforms classical approaches that also permit one to remedy the loss of orthogonality observed when GMRES has converged.

Acknowledgments. The authors would like to thank C. C. Paige for his careful reading of an earlier version of this paper and for his fruitful comments. After the submission of the present paper, and while he was refereeing the Ph.D. manuscript of the third author, C. C. Paige wrote a more direct proof of Theorem 2.3, which can be found in [11]. The work of J. Langou was supported by EADS, Corporate Research Centre, Toulouse.

REFERENCES

[1] A. Björck, *Solving linear least squares problems by Gram–Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–21.
[2] A. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190.
[3] J. W. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.
[4] K. Fan and A. J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math., 6 (1955), pp. 111–116.
[5] L. Giraud and J. Langou, *When modified Gram–Schmidt generates a well-conditioned set of vectors*, IMA J. Numer. Anal., 22 (2002), pp. 521–528.
[6] L. Giraud, S. Gratton, and J. Langou, *A Reorthogonalization Procedure for Modified Gram–Schmidt Algorithm Based on a Rank-k Update*, CERFACS Technical Report TR/PA/03/11, Toulouse, France, 2003.
[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.
[8] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behaviour of the modified Gram–Schmidt GMRES implementation*. BIT, 37 (1997), pp. 706–719.
[9] N. J. Higham, *Matrix nearness problems and applications*, in Applications of Matrix Theory, Oxford University Press, Oxford, UK, 1989.
[10] W. Hoffmann, *Iterative algorithms for Gram–Schmidt orthogonalization*, Computing, 41 (1989), pp. 335–348.

[11] J. Langou, *Iterative Methods for Solving Linear Systems with Multiple Right-Hand Sides*, Ph.D. dissertation, Institut National des Sciences Appliquées de Toulouse, 2003.
[12] R. Schreiber and C. F. Van Loan, *A storage-efficient WY representation for products of Householder transformations*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 53–57.
[13] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM TOMS, 8 (1982), pp. 43–71.
[14] B. N. Parlett, *A new look at the Lanczos algorithm for solving systems of linear equations*, Linear Algebra Appl., 29 (1980), pp. 323–346.
[15] J. R. Rice, *Experiments on Gram–Schmidt orthogonalization*, Math. Comp., 20 (1966), pp. 325–328.
[16] H. Rutishauser, *Description of Algol* 60, in Handbook for Automatic Computation, Vol. 1, Part A, Springer-Verlag, New York, 1967.
[17] Y. Saad, *On the Lanczos method for solving symmetric linear systems with several right-hand sides*, Math. Comp., 48 (1987), pp. 651–662.
[18] G. Sylvand, *La Méthode Multipôle Rapide en Électromagnétisme: Performances, Parallélisation et Applications*, Ph.D. dissertation, Ecole Nationale des Ponts et Chaussées, 2002.
[19] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 25, No. 4, pp. 1178–1192

COMPUTING THE POLAR DECOMPOSITION AND THE MATRIX SIGN DECOMPOSITION IN MATRIX GROUPS*

NICHOLAS J. HIGHAM†, D. STEVEN MACKEY†, NILOUFER MACKEY‡, AND FRANÇOISE TISSEUR†

Abstract. For any matrix automorphism group $\mathbb{G}$ associated with a bilinear or sesquilinear form, Mackey, Mackey, and Tisseur have recently shown that the matrix sign decomposition factors of $A \in \mathbb{G}$ also lie in $\mathbb{G}$; moreover, the polar factors of A lie in $\mathbb{G}$ if the matrix of the underlying form is unitary. Groups satisfying the latter condition include the complex orthogonal, real and complex symplectic, and pseudo-orthogonal groups. This work is concerned with exploiting the structure of $\mathbb{G}$ when computing the polar and matrix sign decompositions of matrices in $\mathbb{G}$. We give sufficient conditions for a matrix iteration to preserve the group structure and show that a family of globally convergent rational Padé-based iterations of Kenney and Laub satisfy these conditions. The well-known scaled Newton iteration for computing the unitary polar factor does not preserve group structure, but we show that the approach of the iterates to the group is precisely tethered to the approach to unitarity, and that this forces a different and exploitable structure in the iterates. A similar relation holds for the Newton iteration for the matrix sign function. We also prove that the number of iterations needed for convergence of the structure-preserving methods can be precisely predicted by running an associated scalar iteration. Numerical experiments are given to compare the cubically and quintically converging iterations with Newton's method and to test stopping criteria. The overall conclusion is that the structure-preserving iterations and the scaled Newton iteration are all of practical interest, and which iteration is to be preferred is problem-dependent.

Key words. automorphism group, bilinear form, sesquilinear form, adjoint, complex orthogonal matrix, symplectic matrix, perplectic matrix, pseudo-orthogonal matrix, polar decomposition, matrix sign decomposition, structure preservation, matrix iteration, Newton iteration, convergence tests

AMS subject classifications. 65F30, 15A18

DOI. 10.1137/S0895479803426644

1. Introduction.

The polar decomposition of $A \in \mathbb{C}^{n\times n}$ factors A as the product $A = UH$, where U is unitary and H is Hermitian positive semidefinite. The Hermitian factor H is always unique and can be expressed as $(A^*A)^{1/2}$, and the unitary factor is unique if A is nonsingular [13]. Here, the exponent $1/2$ denotes the principal square root: the one whose eigenvalues lie in the right half-plane. The polar decomposition is an important theoretical and computational tool, and much is known about its approximation properties, its sensitivity to perturbations, and its computation.

Closely related to the polar decomposition is the matrix sign decomposition, which is defined for $A \in \mathbb{C}^{n\times n}$ having no pure imaginary eigenvalues. The most concise definition of the decomposition is

$$A = SN \equiv A(A^2)^{-1/2} \cdot (A^2)^{1/2}.$$

*Received by the editors April 23, 2003; accepted for publication (in revised form) by D. Bini October 8, 2003; published electronically August 4, 2004. This work was supported by Engineering and Physical Sciences Research Council Visiting Fellowship GR/S15563/01.

http://www.siam.org/journals/simax/25-4/42664.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/; ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur/). The work of the first author was supported by Engineering and Physical Sciences Research Council grant GR/R22612. The work of the fourth author was supported by Engineering and Physical Sciences Research Council grant GR/R45079.

‡Department of Mathematics, Western Michigan University, Kalamazoo, MI 49008 (nil.mackey@wmich.edu, http://homepages.wmich.edu/~mackey/).

Here, $S = \mathrm{sign}(A)$ is the matrix sign function, introduced by Roberts [23]. Note that for scalar $z \in \mathbb{C}$ lying off the imaginary axis, $\mathrm{sign}(z) = 1$ or -1 according as z is in the right half-plane or left half-plane, respectively. An alternative definition is via the Jordan canonical form

$$A = ZJZ^{-1} = Z\mathrm{diag}(J_1, J_2)Z^{-1},$$

where the eigenvalues of J_1 are assumed to lie in the open left half-plane and those of J_2 in the open right half-plane. With this notation,

$$A = SN \equiv Z\mathrm{diag}(-I, I)Z^{-1} \cdot Z\mathrm{diag}(-J_1, J_2)Z^{-1},$$

from which it is clear that S is involutory and the eigenvalues of N lie in the right half-plane.

The polar and matrix sign decompositions are intimately connected [9]. For example, Roberts' integral formula [23],

$$\mathrm{sign}(A) = \frac{2}{\pi} A \int_0^\infty (t^2 I + A^2)^{-1} dt,$$

has an analogue in

$$U = \frac{2}{\pi} A \int_0^\infty (t^2 I + A^* A)^{-1} dt.$$

This example illustrates the rule of thumb that any property or iteration involving the matrix sign function can be converted into one for the polar decomposition by replacing A^2 by A^*A, and vice versa.

Practical interest in the polar decomposition stems mainly from the fact that the unitary polar factor of A is the nearest unitary matrix to A in any unitarily invariant norm [6]. The polar decomposition is therefore of interest whenever it is required to orthogonalize a matrix [8]. The matrix sign function was originally developed as a tool to solve algebraic Riccati equations [23] and it is also used more generally in determining invariant subspaces corresponding to eigenvalues lying in particular regions of the complex plane [1].

Almost all existing work on the polar decomposition and the matrix sign decomposition assumes no special properties of A. However, in an investigation of factorizations in structured classes of matrices, Mackey, Mackey, and Tisseur [22] consider, among other things, the structure of the polar and sign factors. The structures they work with include the automorphism group

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n\times n} : \langle Ax, Ay\rangle_{\mathrm{M}} = \langle x, y\rangle_{\mathrm{M}} \ \forall x, y \in \mathbb{K}^n \,\} \tag{1.1}$$

associated with a bilinear or sesquilinear form defined by any nonsingular matrix M:

$$(x, y) \mapsto \langle x, y\rangle_{\mathrm{M}} = \begin{cases} x^T My & \text{for real or complex bilinear forms,} \\ x^* My & \text{for sesquilinear forms.} \end{cases}$$

Here $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and the superscript $*$ denotes conjugate transpose. It is easy to see that $\mathbb{G}$ is indeed a group under matrix multiplication. Recall that the adjoint $A^\star$ of $A \in \mathbb{K}^{n\times n}$ with respect to $\langle\cdot,\cdot\rangle_{\mathrm{M}}$ is defined by

$$\langle Ax, y\rangle_{\mathrm{M}} = \langle x, A^\star y\rangle_{\mathrm{M}} \quad \forall x, y \in \mathbb{K}^{n\times n}.$$

TABLE 1.1
A sampling of automorphism groups $\mathbb{G} \in \mathfrak{U}$.

Space	M	$A^\star$	Automorphism group, $\mathbb{G}$
Groups corresponding to a bilinear form			
$\mathbb{R}^n$	I	$A^\star = A^T$	Real orthogonals
$\mathbb{C}^n$	I	$A^\star = A^T$	Complex orthogonals
$\mathbb{R}^n$	$\Sigma_{p,q}$	$A^\star = \Sigma_{p,q} A^T \Sigma_{p,q}$	Pseudo-orthogonals
$\mathbb{R}^n$	R	$A^\star = R A^T R$	Real perplectics
$\mathbb{R}^{2n}$	J	$A^\star = -J A^T J$	Real symplectics
$\mathbb{C}^{2n}$	J	$A^\star = -J A^T J$	Complex symplectics
Groups corresponding to a sesquilinear form			
$\mathbb{C}^n$	I	$A^\star = A^*$	Unitaries
$\mathbb{C}^n$	$\Sigma_{p,q}$	$A^\star = \Sigma_{p,q} A^* \Sigma_{p,q}$	Pseudo-unitaries
$\mathbb{C}^{2n}$	J	$A^\star = -J A^* J$	Conjugate symplectics

It can be shown that the adjoint is given explicitly by

$$A^\star = \begin{cases} M^{-1} A^T M & \text{for bilinear forms,} \\ M^{-1} A^* M & \text{for sesquilinear forms.} \end{cases}$$

The adjoint provides a useful alternative characterization of the automorphism group:

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n\times n} : A^\star = A^{-1} \,\}. \tag{1.2}$$

For further details of this background algebra see, for example, [14], [19], or [24].

Automorphism groups for which the matrix M defining the underlying form is unitary ($M^{-1} = M^*$) play an important role in this paper. We use $\mathfrak{U}$ to denote this set of groups. Table 1.1 lists some examples of automorphism groups in $\mathfrak{U}$; here, the matrix M is one of I,

$$R = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \qquad J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}, \qquad \Sigma_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

We need the following results, all from [22].

THEOREM 1.1. *Let $\mathbb{G} \in \mathfrak{U}$. Then any matrix in $\mathbb{G}$ has singular values that occur in reciprocal pairs σ and $1/\sigma$, with the same multiplicity.*

THEOREM 1.2. *Let $\mathbb{G} \in \mathfrak{U}$ and A be any matrix in $\mathbb{G}$. Then in the polar decomposition $A = UH$ the factors U and H also belong to $\mathbb{G}$.*

The following result places no restrictions on $\mathbb{G}$.

THEOREM 1.3. *Let $\mathbb{G}$ be any automorphism group and A be any matrix in $\mathbb{G}$ having a matrix sign decomposition $A = SN$. Then the factors S and N also belong to $\mathbb{G}$.*

We give proofs of Theorems 1.2 and 1.3 at the end of section 2 that provide alternatives to the proofs in [22].

For the real orthogonal and unitary groups, Theorems 1.2 and 1.3 are trivial. For the other groups the results are nontrivial, and indeed in three recent papers devoted to some of these groups the structured nature of the polar and sign factors of matrices in the groups is not noted [3], [4], [11].

This work is concerned with the exploitation of structure when computing the polar or sign factors of matrices from an automorphism group. In section 2 we identify

a general family of rational iterations that are structure-preserving, and we show that certain globally convergent Padé-based iterations of Kenney and Laub belong to this family. In sections 3–6 we concentrate on the polar decomposition, for $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$. In section 3 we identify the most efficient implementations of the cubically and quintically convergent iterations and compare them for efficiency with the scaled Newton iteration. In section 4 we show that although the Newton iteration does not preserve the group structure, under a suitable condition on the scaling parameter it has the remarkable property that its iterates X_k satisfy $X_k^\star = X_k^*$. This relation implies that the approach of the iterates to the group is precisely tethered to the approach to unitarity, and also that for certain automorphism groups the iterates have a different structure that can be exploited.

Numerical stability of the iterations is discussed in section 5. In section 6 we show that the number of iterations needed by one of our structure-preserving methods can be predicted by running the corresponding scalar iteration starting with the largest singular value of A. Corresponding results for the matrix sign decomposition are summarized in section 7. Numerical experiments are presented in section 8 that compare the Newton and quintic iterations and test different stopping criteria. Finally, conclusions are given in section 9.

2. Structure-preserving iterations. A great deal is known about matrix iterations of the form

$$X_{k+1} = f(X_k), \qquad X_0 = A,$$

for computing the unitary polar factor or the matrix sign function; see, for example, [8], [9], [15], [16]. Motivated by Theorems 1.2 and 1.3, we ask the question, "If A belongs to an automorphism group $\mathbb{G}$, when do all the iterates X_k also belong to $\mathbb{G}$?" When this property holds, we say the iteration is structure-preserving for $\mathbb{G}$. Sufficient conditions for such an iteration are given in the next theorem. For a polynomial p of degree m we introduce the notation $\mathrm{rev}p(x) = x^m p(1/x)$; thus $\mathrm{rev}p$ is p with its coefficients reversed. (Note that $\mathrm{rev}(\mathrm{rev}p)$ is not necessarily p, as the example $p(x) = x^2 + x$ shows.)

THEOREM 2.1. *Let p be any polynomial with real coefficients and let f be a matrix function having the form*

$$f(X) = Yp(Z)[\mathrm{rev}p(Z)]^{-1}. \tag{2.1}$$

Assume that the appropriate inverses exist, so that f is well defined.

(a) *If Y and Z are integer powers of $X \in \mathbb{G}$, then $f(X) \in \mathbb{G}$ for any automorphism group $\mathbb{G}$.*

(b) *If Y and Z are finite products of X, X^{-1} and X^*, in any combination, where $X \in \mathbb{G}$, then $f(X) \in \mathbb{G}$, for any automorphism group $\mathbb{G} \in \mathfrak{U}$.*

Proof. We note first the properties that $(ST)^\star = T^\star S^\star$ for all S and T and $(S^{-1})^\star = (S^\star)^{-1}$ for all nonsingular S, the latter equality implying that we can write $S^{-\star}$ without ambiguity. Observe also that since p has real coefficients, $p(T)^\star = p(T^\star)$ for all T.

For part (a), Y and Z are readily seen to belong to $\mathbb{G}$, since $\mathbb{G}$ is a group under multiplication. For part (b), $\mathbb{G} \in \mathfrak{U}$ implies $M^{-1} = M^*$, and so $(T^\star)^* = (T^*)^\star$ for all T. Consequently $X \in \mathbb{G}$ implies $X^* \in \mathbb{G}$; hence Y and Z belong to $\mathbb{G}$.

Marshalling these facts, and denoting by m the degree of p, we obtain

$$f(X)^\star f(X) = [\mathrm{rev}p(Z)]^{-\star} \cdot p(Z)^\star \cdot \underbrace{Y^\star \cdot Y}_{I} \cdot p(Z) \cdot [\mathrm{rev}p(Z)]^{-1}$$

$$
\begin{aligned}
&= \big([\mathrm{rev}p(Z)]^\star\big)^{-1} \cdot p(Z^\star) \cdot p(Z) \cdot [\mathrm{rev}p(Z)]^{-1} \\
&= \big([\mathrm{rev}p(Z^\star)]\big)^{-1} \cdot p(Z^{-1}) \cdot p(Z) \cdot [\mathrm{rev}p(Z)]^{-1} \\
&= \big(Z^{-m}p(Z)\big)^{-1} \cdot p(Z) \cdot p(Z^{-1}) \cdot [Z^m p(Z^{-1})]^{-1} \\
&= I. \qquad \square
\end{aligned}
$$

We mention that a converse of part (a) in Theorem 2.1 is proved in [12], from which it follows that any rational function f that maps $\mathbb{G}$ into itself for *all* $\mathbb{G}$ can be expressed in the form (2.1), with p a polynomial with real coefficients.

Theorem 2.1 says nothing about the convergence of the iteration $X_{k+1} = f(X_k)$, so further restrictions on f are needed to obtain a useful iteration. By using only elementary means, one can construct rational iteration functions of the form (2.1) with any specified odd order of convergence. The first two functions in this sequence are

$$x_{k+1} = f_{11}(x_k), \qquad f_{11}(x) = \frac{x(3+x^2)}{1+3x^2}, \tag{2.2}$$

$$x_{k+1} = f_{22}(x_k), \qquad f_{22}(x) = \frac{x(5+10x^2+x^4)}{1+10x^2+5x^4}, \tag{2.3}$$

which, for $x_0 \in \mathbb{C}$ not on the imaginary axis, converge to $\mathrm{sign}(x_0)$ at a cubic[1] and quintic rate, respectively. See [20] for details of this approach. It turns out that the functions f_{ii} thus constructed belong to the family of rational iterations

$$x_{k+1} = f_{\ell m}(x_k) = x_k \frac{P_{\ell m}(1-x_k^2)}{Q_{\ell m}(1-x_k^2)} \tag{2.4}$$

studied by Kenney and Laub [15], where $P_{\ell m}(t)/Q_{\ell m}(t)$ is the $[\ell/m]$ Padé approximant to $(1-t)^{-1/2}$, with the polynomials $P_{\ell m}$ and $Q_{\ell m}$ having degrees ℓ and m, respectively. These iterations are designed to compute $\mathrm{sign}(x_0)$, and the iterations with $\ell = m$ and $\ell = m-1$ are shown in [15] to be globally convergent, that is, they converge to $\mathrm{sign}(x_0)$ for any $x_0 \in \mathbb{C}$ not on the imaginary axis. For $\ell = m$ and $\ell = m-1$ it is also noted in [15] that $-xP_{\ell m}(1-x^2)$ and $Q_{\ell m}(1-x^2)$ are, respectively, the odd and even parts of $(1-x)^{\ell+m+1}$. For $\ell = m-1$, the iteration is easily verified not to be structure-preserving for $\mathbb{G}$. But from the odd-even property just mentioned, the iteration for $\ell = m$ can be seen to have the form (2.1) with $Y = Z = X$, and therefore by part (a) of Theorem 2.1 the iteration is structure-preserving for all automorphism groups $\mathbb{G}$.

THEOREM 2.2. *Let $A \in \mathbb{K}^{n\times n}$ and consider the iterations*

$$Y_{k+1} = Y_k\, P_{mm}(I - Y_k^2)\, Q_{mm}(I - Y_k^2)^{-1}, \qquad Y_0 = A, \tag{2.5}$$

and

$$Z_{k+1} = Z_k\, P_{mm}(I - Z_k^* Z_k)\, Q_{mm}(I - Z_k^* Z_k)^{-1}, \qquad Z_0 = A, \tag{2.6}$$

with $m \geq 1$. Assume that A has no eigenvalues on the imaginary axis for (2.5).

(a) *If $\mathbb{G}$ is any automorphism group and $A \in \mathbb{G}$, then $Y_k \in \mathbb{G}$ for all k, and Y_k converges to* $\mathrm{sign}(A)$.

[1] The iteration (2.2) is Halley's method for $x^2 - 1 = 0$ [7].

(b) *If $\mathbb{G}$ is any automorphism group in $\mathfrak{U}$ and $A \in \mathbb{G}$, then $Z_k \in \mathbb{G}$ for all k, and Z_k converges to the unitary polar factor of A.*

Moreover, both sequences have order of convergence $2m+1$.

Proof. The preservation of structure has already been shown. It remains to prove convergence. The existence of the inverse in (2.5) and the global convergence of (2.5) to $\operatorname{sign}(A)$ with order $2m+1$ are established in [15, Thm. 5.3]. That (2.6) converges globally to the unitary polar factor at the same rate can be shown by using the singular value decomposition of A to reduce (2.6) to n independent scalar iterations on the singular values, whose convergence to 1 follows from that of (2.5). ∎

A proof of Theorem 1.3 now follows immediately from part (a) of Theorem 2.2: since $\mathbb{G}$ is a closed set, $\lim Y_k = \operatorname{sign}(A)$ belongs to $\mathbb{G}$, and since $\mathbb{G}$ is a group under multiplication, the factor N in the matrix sign decomposition of A must also belong to $\mathbb{G}$. In an entirely analogous way, a proof of Theorem 1.2 follows from part (b) of Theorem 2.2.

In the next four sections we restrict our attention to the polar decomposition. In section 7 we explain to what extent our analysis for the polar decomposition can be adapted to the matrix sign decomposition.

3. Iterations for the polar decomposition. We begin by examining the first two iterations of the previous section and their computational cost.

The cubically convergent iteration (2.2) is, in matrix form for computing the polar decomposition,

$$X_{k+1} = X_k(3I + X_k^*X_k)(I + 3X_k^*X_k)^{-1}, \qquad X_0 = A. \tag{3.1}$$

We will measure the cost of iterations by counting the number of (general) matrix multiplications, *mult*, and the number of (general) matrix inversions, *inv*. When evaluating a term of the form AB^{-1} it is less expensive to factor B and then solve a multiple right-hand-side linear system than to explicitly invert B, so we will assume the former is done and record the cost as a corresponding multiple of *inv*. In our iterations, B is Hermitian positive definite and AB^{-1} is Hermitian; if we exploit this structure the cost of computing AB^{-1} is $(5/6)inv$.

One iteration of (3.1) costs $(3/2)\,mult + (5/6)\,inv$ per iteration. By rewriting the iteration the cost can be reduced: for

$$X_{k+1} = \frac{1}{3}X_k\big[I + 8\big(I + 3X_k^*X_k\big)^{-1}\big], \qquad X_0 = A, \tag{3.2}$$

the cost per iteration is $(3/2)\,mult + (1/2)inv$.

The quintically convergent iteration (2.3) becomes

$$X_{k+1} = X_k\big[5I + 10X_k^*X_k + (X_k^*X_k)^2\big]\big[I + 10X_k^*X_k + 5(X_k^*X_k)^2\big]^{-1}, \quad X_0 = A, \tag{3.3}$$

which costs $2\,mult + (5/6)\,inv$ per iteration. This iteration can be rewritten in various more efficient ways. We state the two of most interest in scalar form, for readability; for matrices, x^2 should be replaced by $X_k^*X_k$ and the divisions by matrix inversions. First, we have the continued fraction form

$$x_{k+1} = \frac{1}{5}x_k\left[1 + \cfrac{8}{x^2 + \cfrac{7}{5 + \cfrac{16}{7x^2+1}}}\right], \qquad x_0 = a, \tag{3.4}$$

which costs $(3/2)(mult + inv)$ per iteration. The alternative form

$$x_{k+1} = x_k \left[\frac{1}{5} + \frac{8}{5x_k^2 + 7 - \dfrac{16}{5x_k^2 + 3}} \right], \qquad x_0 = a, \tag{3.5}$$

requires just $(3/2)\,mult + inv$ per iteration, so is the least expensive of the three variants.

For any m, the iterations (2.6) can also be expressed in partial fraction form [17, (35)]. The cost of evaluation is $(3/2)\,mult + (m/2)\,inv$ per iteration, which in the cases $m = 1$ and $m = 2$ is the same as the cost of evaluating (3.2) and (3.5), respectively.

Also of interest is the well-known Newton iteration

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-*}), \qquad X_0 = A. \tag{3.6}$$

This iteration is not structure-preserving for automorphism groups $\mathbb{G}$, since $\mathbb{G}$ is not closed under addition, but, as we will see, the iteration is nevertheless of interest when $A \in \mathbb{G}$. In practice, the Newton iteration is usually implemented with scaling to accelerate the initial speed of convergence. The scaled iteration is

$$X_{k+1} = \frac{1}{2}\left[\gamma^{(k)} X_k + \frac{1}{\gamma^{(k)}} X_k^{-*} \right], \qquad X_0 = A, \tag{3.7}$$

where the scaling parameter $\gamma^{(k)} \in \mathbb{R}$ is intended to make X_{k+1} closer to U. Higham [8] identified the scaling

$$\gamma_{\mathrm{opt}}^{(k)} = \bigl(\sigma_{\min}(X_k)\sigma_{\max}(X_k)\bigr)^{-1/2},$$

where $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values, respectively, as optimal in the sense of minimizing a bound on $\|U - X_{k+1}\|_2 / \|U + X_{k+1}\|_2$, and this scaling leads to convergence in s iterations, where s is the number of distinct singular values of A [16, Lem. 2.2]. Among more economical choices analyzed in [16] is the Frobenius norm scaling

$$\gamma_F^{(k)} = \left(\frac{\|X_k^{-1}\|_F}{\|X_k\|_F} \right)^{1/2}, \tag{3.8}$$

which has the property that it minimizes $\|X_{k+1}\|_F$ over all $\gamma^{(k)}$ [5]. Both these scalings have the property that

$$X_0 \in \mathbb{G} \in \mathfrak{U} \quad \Rightarrow \quad \gamma^{(0)} = 1, \tag{3.9}$$

by virtue of the reciprocal pairing of the singular values when $\mathbb{G} \in \mathfrak{U}$ (see Theorem 1.1).

We do not investigate scaling for the structure-preserving iterations, because for $X_k \in \mathbb{G}$ and f in (2.1), $f(\gamma^{(k)} X_k) \notin \mathbb{G}$ in general, and so scaling destroys group structure.

We first ask which of the three iterations (3.2), (3.5), and (3.6) is the most computationally efficient, independent of structure considerations. In answering this question we need to take account of the fact that the iterations comprise two phases: the

TABLE 3.1
Cost estimates for (3.2), (3.5), *and* (3.6), *assuming* $\|X_0 - U\|_2 = 0.25$.

Iterations	Resulting error	Cost
2 quintic iterations	$0.25^{25} = 9 \times 10^{-16}$	$3\,mult + 2\,inv$
3 cubic iterations	$0.25^{27} = 6 \times 10^{-17}$	$(9/2)mult + (3/2)inv$
5 Newton iterations	$0.25^{32} = 5 \times 10^{-20}$	$5\,inv$

initial phase in which the error $\|X_k - U\|_2$ is reduced to safely less than 1, and then the phase in which asymptotic convergence sets in at a quadratic, cubic, or quintic rate. Working in IEEE double precision arithmetic, the furthest we iterate is until the error reaches the unit roundoff $u \approx 10^{-16}$, so the higher order iterations are not in their asymptotic phase for long. In the initial phase, our three iterations converge essentially linearly, with rate constants 1/2, 1/3, and 1/5, respectively (this can be seen by considering the scalar iterations with $0 < x_0 \ll 1$ and $x_0 \gg 1$). Hence for large $\|X_0 - U\|_2$ the quintic iteration requires the least work to reduce the error below 1, followed by the cubic and then the Newton iterations. Once the error is safely below 1, the three iterations cost roughly the same amount to reduce the error to the unit roundoff level; see Table 3.1. Our conclusion is that if $\|X_0 - U\|_2 \lesssim 1$ there is little to choose between the iterations in cost, but for $\|X_0 - U\|_2 \gg 1$ the quintic iteration has the advantage. In the scaled Newton iteration (3.7), with the Frobenius norm scaling (3.8), the first phase of convergence is shortened considerably. Practical experience, supported by theory [16], shows that about 9 or 10 iterations at most are required for any A in IEEE double precision arithmetic. Therefore scaled Newton is competitive in cost with the quintic iteration, albeit not structure-preserving.

4. Structure in the Newton iteration. We have seen that the cubic iteration (3.2) and the quintic iteration (3.5) are structure-preserving for automorphism groups $\mathbb{G} \in \mathfrak{U}$, while the Newton iteration (3.6) and its scaled form (3.7) are not. We now consider precisely how the Newton iteration affects structure, and to do so we first develop a measure of departure from $\mathbb{G}$-structure. Throughout the rest of this section we assume that $\mathbb{G} \in \mathfrak{U}$, that is, that M is unitary.

The characterization (1.2) says that A is in the automorphism group $\mathbb{G}$ if $A^\star = A^{-1}$. To obtain a measure of departure from $\mathbb{G}$-structure that is less dependent on the conditioning of A, we rewrite this relation as $A^\star A = I$. Consider $A + \Delta A$, where $A \in \mathbb{G}$ and $\|\Delta A\|_2 \le \epsilon \|A\|_2$. Using the fact that M is unitary, we have $\|A^\star\|_2 = \|A\|_2$ for all A, and hence

$$\begin{aligned}\|(A+\Delta A)^\star(A+\Delta A) - I\|_2 &= \|A^\star \Delta A + \Delta A^\star A + \Delta A^\star \Delta A\|_2 \\ &\le 2\|A\|_2\|\Delta A\|_2 + \|\Delta A\|_2^2 \\ &\le (2\epsilon + \epsilon^2)\|A\|_2^2.\end{aligned}$$

This inequality suggests that an appropriate relative measure of departure from $\mathbb{G}$-structure is

$$\mu_{\mathbb{G}}(A) = \frac{\|A^\star A - I\|_2}{\|A\|_2^2}. \tag{4.1}$$

In the particular case $\mathbb{G} = \mathbb{O}$, the unitary group, we have

$$\mu_{\mathbb{O}}(A) = \frac{\|A^* A - I\|_2}{\|A\|_2^2}, \tag{4.2}$$

which is a standard measure of departure from unitarity. Further justification for this measure is given by showing that if $\mu_{\mathbb{G}}(A)$ is small then A is close to a matrix in $\mathbb{G}$. For this, we use the *generalized* polar decomposition, which is closely related to the polar decompositions in indefinite scalar product spaces studied by Bolshakov et al. [2].

THEOREM 4.1 (generalized polar decomposition [22]). *Let $\mathbb{G}$ be an automorphism group corresponding to a bilinear or sesquilinear form for which $(A^\star)^\star = A$ for all $A \in \mathbb{K}^{n\times n}$. For any $A \in \mathbb{K}^{n\times n}$ such that $A^\star A$ has no eigenvalues on the nonpositive real axis, A has a unique decomposition $A = WS$, where $W \in \mathbb{G}$ (that is, $W^\star = W^{-1}$), $S^\star = S$, and $\operatorname{sign}(S) = I$.*

We note that the condition in this theorem on the adjoint being involutory holds precisely when $M^T = \pm M$ for bilinear forms and $M^* = \alpha M$ with $|\alpha| = 1$ for sesquilinear forms [22], and that these conditions hold for all the groups in Table 1.1.

LEMMA 4.2. *Let $A \in \mathbb{K}^{n\times n}$ have a generalized polar decomposition $A = WS$ with respect to an automorphism group $\mathbb{G} \in \mathfrak{U}$. If $\|W^{-1}(A - W)\|_2 < 1$, or equivalently $\|S - I\|_2 < 1$, then*

$$\frac{\|A^\star A - I\|_2}{\|A\|_2(\|A\|_2 + \|W\|_2)} \le \frac{\|A - W\|_2}{\|A\|_2} \le \frac{\|A^\star A - I\|_2}{\|A\|_2^2}\ \|A\|_2\|W\|_2.$$

The lower bound always holds.

Proof. Using $W^\star = W^{-1}$ and $S^\star = S$ we have

$$\begin{aligned}(A + W)^\star(A - W) &= A^\star A - A^\star W + W^\star A - W^\star W \\ &= A^\star A - S^\star W^\star W + W^\star W S - I = A^\star A - I.\end{aligned}$$

The lower bound follows immediately. For the upper bound, we need to bound $\|(A + W)^{-\star}\|_2$. Note that $W^{-1}(A - W) = W^{-1}(WS - W) = S - I$ and $A + W = 2W(I + (S - I)/2)$. Hence, using the fact that $\mathbb{G} \in \mathfrak{U}$,

$$\begin{aligned}\|(A + W)^{-\star}\|_2 = \|(A + W)^{-1}\|_2 &= \left\| \frac{1}{2}\big(I + (S - I)/2\big)^{-1} W^{-1} \right\|_2 \\ &\le \frac{1}{2}\|W^{-1}\|_2 \frac{1}{1 - \frac{1}{2}\|S - I\|_2} \\ &\le \|W^{-1}\|_2,\end{aligned}$$

which yields the result. ∎

Lemma 4.2 shows that there is a matrix $W \in \mathbb{G}$ within relative distance $\mu_{\mathbb{G}}(A)$ of A, modulo a factor $\|A\|_2\|W\|_2$, as we wanted to show.

We now present a numerical experiment in which we compute the orthogonal polar factor of a random symplectic matrix $A \in \mathbb{R}^{12\times 12}$ with $\|A\|_2 = 3.1 \times 10^2 = \|A^{-1}\|_2$. All our experiments were performed in MATLAB, for which $u \approx 1.1\times 10^{-16}$. Table 4.1 reports the behavior of the Newton iteration, both without scaling and with Frobenius norm scaling, the cubic iteration (3.2), and the quintic (3.5). We report iterations up to the last one for which there was a significant decrease in $\|X_k^* X_k - I\|_2$. First, note that the convergence is entirely consistent with our description earlier, with the quintic and scaled Newton iterations spending the least time in the first phase. Next, we see from the first line of the table that the matrix A is indeed symplectic to machine precision, but far from orthogonal. The Newton iterations destroy the symplectic structure on the first iteration, but gradually restore it, as they must,

TABLE 4.1
Results for a symplectic matrix $A \in \mathbb{R}^{12\times 12}$ with $\kappa_2(A) = 9.6 \times 10^4$. Here, $\mu_{\mathbb{G}}$ and $\mu_{\mathbb{O}}$ are defined in (4.1) *and* (4.2)*, and $E = \min_k \|U - X_k\|_2$.*

k	Newton		Newton (scaled)		Cubic, (3.2)		Quintic, (3.5)	
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$
0	1.0e+0	7.0e-18	1.0e+0	7.0e-18	1.0e+0	7.0e-18	1.0e+0	7.0e-18
1	1.0e+0	1.0e+0	1.0e+0	1.0e+0	1.0e+0	8.9e-17	1.0e+0	1.1e-15
2	1.0e+0	1.0e+0	8.6e-01	8.6e-01	1.0e+0	8.1e-16	9.9e-01	1.7e-14
3	1.0e+0	1.0e+0	2.0e-01	2.0e-01	9.9e-01	6.3e-15	8.5e-01	3.0e-13
4	1.0e+0	1.0e+0	3.2e-03	3.2e-03	9.4e-01	5.0e-14	7.0e-02	1.7e-12
5	9.9e-01	9.9e-01	9.0e-07	9.0e-07	5.7e-01	2.8e-13	7.6e-09	1.8e-12
6	9.6e-01	9.6e-01	6.0e-14	1.3e-13	3.6e-02	5.2e-13	4.8e-16	1.8e-12
7	8.5e-01	8.5e-01	4.3e-16	1.1e-13	3.2e-06	5.3e-13		
8	5.4e-01	5.4e-01			3.8e-16	5.3e-13		
9	1.4e-01	1.4e-01						
10	5.5e-03	5.5e-03						
11	7.7e-06	7.7e-06						
12	1.5e-11	1.5e-11						
13	4.4e-16	1.1e-13						
E	4.4e-13		4.4e-13		7.3e-13		1.9e-12	

since the limit U is symplectic. However, we see that for these two iterations the relation $\mu_{\mathbb{O}}(X_k) = \mu_{\mathbb{G}}(X_k)$, that is,

$$\|X_k^* X_k - I\|_2 = \|X_k^\star X_k - I\|_2,$$

holds from iteration 1 until close to convergence, at which point rounding errors vitiate the relation—thus the approach to symplecticity is precisely tethered to the approach to orthogonality in this example. In fact, this is always true, as is an even stronger condition: the Newton iterates satisfy $X_k^\star = X_k^*$ for $k \geq 1$, for both the unscaled and Frobenius norm scaled iterations! Hence although the Newton iteration destroys the group structure, from this structure it creates and preserves a *different* kind of structure.

THEOREM 4.3. *Let $\mathbb{G} \in \mathfrak{U}$, $A \in \mathbb{G}$, and X_k be defined by the Newton iteration* (3.6) *or by a scaled Newton iteration* (3.7) *for which* (3.9) *holds. Then, for $k \geq 1$, $X_k^\star = X_k^*$.*

Proof. We will use two properties that we recall from the proof of Theorem 2.1 and that hold for all $B \in \mathbb{K}^{n\times n}$: $(B^{-1})^\star = (B^\star)^{-1}$, and $\mathbb{G} \in \mathfrak{U}$ implies $(B^*)^\star = (B^\star)^*$. For the scaled iteration, (3.9) implies $\gamma^{(0)} = 1$, and hence for both the scaled and unscaled iterations

$$\begin{aligned} X_1^\star &= \frac{1}{2}(A + A^{-*})^\star = \frac{1}{2}\big(A^\star + (A^{-*})^\star\big) \\ &= \frac{1}{2}\big(A^{-1} + (A^{-\star})^*\big) = \frac{1}{2}(A^{-1} + A^*) = X_1^*. \end{aligned}$$

Now assume that $X_{k-1}^\star = X_{k-1}^*$. Then, writing $\gamma = \gamma^{(k-1)}$,

$$\begin{aligned} X_k^\star &= \frac{1}{2}\big(\gamma X_{k-1} + \gamma^{-1} X_{k-1}^{-*}\big)^\star \\ &= \frac{1}{2}\big((\gamma X_{k-1})^\star + (\gamma^{-1} X_{k-1}^{-*})^\star\big) \\ &= \frac{1}{2}\big(\gamma X_{k-1}^* + \gamma^{-1} X_{k-1}^{-1}\big) = X_k^*. \end{aligned}$$

The result follows by induction. ∎

COROLLARY 4.4. *Under the conditions of Theorem* 4.3*, for* $k \geq 1$*,*

(a) $MX_k = X_k M$ *and* $MX_k^* = X_k^* M$ *for real bilinear and complex sesquilinear forms,*

(b) $MX_k = \overline{X_k} M$ *and* $MX_k^* = X_k^T M$ *for complex bilinear forms.*

Proof. Theorem 4.3 gives $X_k^\star = X_k^*$ for $k \geq 1$.

(a) We therefore have $M^{-1} X_k^* M = X_k^*$, or $X_k^* M = M X_k^*$. Taking the conjugate transpose and using $M^* = M^{-1}$ gives $MX_k = X_k M$.

(b) Similarly, $M^{-1} X_k^T M = X_k^*$, or $X_k^T M = M X_k^*$. Taking the conjugate transpose and using $M^* = M^{-1}$ gives $MX_k = \overline{X_k} M$. ∎

While Theorem 4.3 establishes the tethering between our measures of departure from $\mathbb{G}$-structure and unitarity, Corollary 4.4 has some further implications. In the case where A is pseudo-orthogonal, $M = \Sigma_{p,q}$, and to commute with $\Sigma_{p,q}$ is to be block diagonal! So all iterates X_k, $k \geq 1$, and the unitary polar factor itself, are block diagonal. For symplectic A, all the Newton iterates have the block structure $\left[\begin{smallmatrix} E & F \\ -F & E \end{smallmatrix}\right]$, and for perplectic A all the Newton iterates are centrosymmetric, that is, $a_{i,j} = a_{n-i+1,n-j+1}$ for $1 \leq i, j \leq n$. Computational savings can readily be made in all these cases. For example, in the symplectic case we need compute only the first n rows of the iterates, since the last n rows can be obtained from them.

5. Numerical stability. All the iterations under consideration involve matrix inversion, either explicitly or via the solution of linear systems with multiple right-hand sides, and when the corresponding matrices are ill conditioned numerical instability is a concern. Many years of experience have shown that the Newton iteration (3.7) is less prone to instability than might be expected. Indeed, it performs better than the best available bounds suggest; for a recent rounding error analysis of the iteration see [18]. Table 4.1 provides some insight and is representative of the typical behavior of the four iterations it illustrates: the computed iterates converge to a matrix that is orthogonal to working precision, and the error $\|U - X_k\|_2$ is of order at most $\kappa_2(A)u$, as is the measure $\mu_{\mathbb{G}}(X_k)$ in (4.1) of departure from $\mathbb{G}$-structure.

Of the other iterations we have found empirically that (3.1) and (3.4) are numerically stable, but (3.3) is not; the latter iteration produces an error $\|U - X_k\|_2$ and loss of structure $\mu_{\mathbb{G}}(X_k)$ observed to be of order $\kappa_2(A)^2 u$ and fails to converge when $\kappa_2(A) \gtrsim u^{-1/2}$.

We have found the partial fraction form of the quintic, mentioned in section 3, to have the same numerical stability as (3.5).

6. Convergence tests. An important question is how to terminate these matrix iterations. Since the Padé-based iterations compute $X_k^* X_k$, a convergence test of the form $\|X_k^* X_k - I\| \leq \mathrm{tol}$ can be used at no extra cost. For small tol, this test directly controls the error, since from Lemma 4.2 with $\mathbb{G} = \mathbb{O}$, using in the upper bound a refinement specific to this case from [10, Prob. 19.14],

$$\frac{\|X_k^* X_k - I\|_2}{\sigma_{\max}(X_k) + 1} \leq \|U - X_k\|_2 \leq \frac{\|X_k^* X_k - I\|_2}{\sigma_{\min}(X_k) + 1}.$$

The Padé-based iterations also have special properties that can be exploited. For the iteration function $f_{\ell m}$ in (2.4), it can be shown that f_{mm} has the properties

$$f_{mm}(\sigma^{-1}) = f_{mm}(\sigma)^{-1}, \tag{6.1a}$$

$$1 < \sigma \;\Rightarrow\; 1 < f_{mm}(\sigma) < \sigma, \tag{6.1b}$$

$$1 \le \mu < \sigma \;\Rightarrow\; f_{mm}(\mu) < f_{mm}(\sigma), \tag{6.1c}$$

$$f_{mm}(1) = 1. \tag{6.1d}$$

Let $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$. Then, by Theorem 1.1, A has singular values that we may index

$$\sigma_1^{-1} \le \cdots \le \sigma_q^{-1} < \sigma_{q+1} = \cdots = \sigma_n = 1 < \sigma_q \le \cdots \le \sigma_1.$$

Using (6.1), we find that Z_k from (2.6) has singular values

$$\begin{aligned} f_{mm}^{(k)}(\sigma_1)^{-1} \le \cdots \le f_{mm}^{(k)}(\sigma_q)^{-1} < \sigma_{q+1} = \cdots \\ = \sigma_n = 1 < f_{mm}^{(k)}(\sigma_q) \le \cdots \le f_{mm}^{(k)}(\sigma_1). \end{aligned} \tag{6.2}$$

Applying this argument repeatedly, we deduce that Z_k from (2.6) satisfies

$$\|U - Z_k\|_2 = f_{mm}^{(k)}(\sigma_1) - 1. \tag{6.3}$$

The practical significance of this equality is that we can precisely predict the convergence of the matrix iteration simply by performing the iteration on σ_1, which is a scalar computation. If σ_1 is not known, or is too expensive to compute or estimate, then we can instead use

$$\|U - Z_k\|_2 \le f_{mm}^{(k)}(\|A\|_F) - 1.$$

The scalar computations can be done in advance of the matrix iteration, if required.

Another useful property of the iterates Z_k when $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$ can be derived from (6.1) and (6.2): the sequence $\|Z_k\|_F$ decreases monotonically to $\sqrt{n}$. This means that the iteration can be terminated when the computed iterates $\widehat{Z}_k$ satisfy

$$\frac{\|\widehat{Z}_{k+1}\|_F}{\|\widehat{Z}_k\|_F} \ge 1 - \delta, \tag{6.4}$$

for some tolerance δ depending on u, that is, when rounding errors start to dominate.

Similar techniques apply to the Newton iteration. Convergence prediction can be done for the unscaled Newton iteration (3.6) for any A, as observed by Kenney and Laub [16], though with a simplification when $A \in \mathbb{G}$ and $\mathbb{G} \in \mathfrak{U}$. The iteration $h(x) = (x + 1/x)/2$ shares the properties (6.1b)–(6.1d) of f_{mm} and satisfies $h(\sigma^{-1}) = h(\sigma)$ in place of (6.1a). Therefore[2]

$$\|U - X_k\|_2 = h^{(k)}(\sigma_1) - 1; \tag{6.5}$$

again, we can use $\|A\|_F$ in place of σ_1 and obtain an upper bound. Convergence prediction based on a scalar iteration is not possible for the Newton iteration with Frobenius norm scaling.

For any A, the Newton sequence norms $\|X_k\|_F$ decrease monotonically for $k \ge 1$, both for the unscaled iteration and for the iteration with Frobenius norm scaling. This follows from the properties of h in the unscaled case and for Frobenius norm scaling can readily be proved from its definition, as shown by Dubrulle [5]. Therefore the stopping criterion (6.4) is applicable, and indeed it is advocated by Dubrulle [5] for the Frobenius norm scaling.

[2]For general A, (6.5) holds for $k \ge 1$ with $h^{(k)}$ replaced by $h^{(k-1)}$ and with σ_1 now the largest singular value of X_1.

7. Matrix sign decomposition. Much, but not all, of the analysis of the previous four sections applies with minor modification to the matrix sign decomposition.

The rewritten forms of the cubic and quintic iterations remain valid, with $X_k^* X_k$ replaced by X_k^2. Their costs are slightly higher than in the polar decomposition case, since X_k^2 is not Hermitian. The scaled Newton iteration for the matrix sign function is

$$X_{k+1} = \frac{1}{2}\left[\gamma^{(k)} X_k + \frac{1}{\gamma^{(k)}} X_k^{-1}\right], \qquad X_0 = A. \tag{7.1}$$

Among many proposed scalings is the determinantal scaling, $\gamma_k = |\det(X_k)^{-1/n}|$. This scaling satisfies (3.9), which continues to be an important property.

Theorem 4.3, which shows that the (suitably scaled) Newton iterates satisfy $X_k^\star = X_k^*$, has the following analogue.

THEOREM 7.1. *Let $A \in \mathbb{G}$, where $\mathbb{G}$ is any automorphism group. Let X_k be defined by the Newton iteration* (7.1)*, either unscaled ($\gamma^{(k)} = 1$) or with a scaling for which $\gamma^{(0)} = 1$. Then $X_k^\star = X_k$ for $k \geq 1$.*

COROLLARY 7.2. *Under the conditions of Theorem* 7.1*, for $k \geq 1$,*

(a) *$MX_k = X_k^T M$ for bilinear forms,*
(b) *$MX_k = X_k^* M$ for sesquilinear forms.*

Theorem 7.1 implies that the Newton iterates for the matrix sign function satisfy the condition $X_k^2 - I = X_k^\star X_k - I$, and so the approach to the group structure is tethered to the approach to involutory structure.

The convergence tests discussed in section 6 are not applicable to the sign iteration. In particular, since A is generally nonnormal the errors are not determined solely by the eigenvalues of the iterates.

8. Numerical experiments. Returning to the polar decomposition, we now compare experimentally the quintic iteration (3.5) with the Newton iteration (3.7) with Frobenius norm scaling. We generated random complex orthogonal $A_1, A_{16} \in \mathbb{R}^{16\times 16}$, where A_k denotes a product of k random complex orthogonal $\mathbb{G}$-reflectors [21]. Results are shown in Tables 8.1 and 8.2. In these tables the term $\mathrm{err}_k = \|U - X_k\|_2$ is computed from (6.3) using a scalar recurrence. Also shown is the value $1 - \|X_{k+1}\|_F / \|X_k\|_F$ arising in the convergence test (6.4).

The results, and others from similar experiments, reveal a number of interesting features.

1. The monotonicity test (6.4), and, for the quintic iteration, convergence prediction based on (6.3), both provide reliable termination criteria. For the former, $\delta \approx \sqrt{u}$ seems an appropriate choice, and for the latter, $\mathrm{err}_k \approx u$.
2. The Newton iterations (scaled and unscaled) can produce a computed unitary polar factor with smaller errors and better structure preservation than the quintic iteration (by a factor of up to 10^4 in Table 8.2), though all these quantities are empirically bounded by about $\kappa_2(A)u$.
3. The quintic iteration's faster initial linear convergence and faster asymptotic convergence enable it to require fewer iterations than scaled Newton when $\|A - U\|_2 \lesssim 1$, but nevertheless the scaled Newton iteration usually requires the fewest flops.

TABLE 8.1

Results for a complex orthogonal matrix $A_1 \in \mathbb{R}^{16\times 16}$ with $\kappa_2(A) = 6.6$. Here, $\mu_{\mathbb{G}}$ and $\mu_{\mathbb{O}}$ are defined in (4.1) *and* (4.2), $\mathrm{err}_k = \|U - X_k\|_2$ *is computed from a scalar recurrence, and* $E = \min_k \|U - X_k\|_2$.

k	Newton (scaled)			Quintic, (3.5)			
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	err_k	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$
0	9.8e-01	6.7e-17		9.8e-01	6.7e-17	7.1e+0	
1	9.4e-01	9.4e-01	2.3e-01	7.0e-01	3.3e-15	8.2e-01	5.3e-01
2	6.5e-01	6.5e-01	3.4e-01	8.3e-03	7.5e-15	4.2e-03	4.7e-02
3	1.5e-01	1.5e-01	1.2e-01	1.5e-13	7.6e-15	7.7e-14	2.2e-06
4	4.8e-03	4.8e-03	1.2e-02	5.6e-16	7.5e-15	0.0e+0	0.0e+0
5	4.3e-06	4.3e-06	3.5e-04				
6	3.4e-12	3.4e-12	3.1e-07				
7	4.7e-16	1.1e-15	2.4e-13				
8	5.5e-16	1.2e-15	0.0e+0				
E	1.4e-15			5.6e-15			

TABLE 8.2

Results for a complex orthogonal matrix $A_{16} \in \mathbb{R}^{16\times 16}$ with $\kappa_2(A) = 6.5 \times 10^9$. Here, $\mu_{\mathbb{G}}$ and $\mu_{\mathbb{O}}$ are defined in (4.1) *and* (4.2), $\mathrm{err}_k = \|U - X_k\|_2$ *is computed from a scalar recurrence, and* $E = \min_k \|U - X_k\|_2$.

k	Newton (scaled)			Quintic, (3.5)			
	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$	$\mu_{\mathbb{O}}(X_k)$	$\mu_{\mathbb{G}}(X_k)$	err_k	$1 - \frac{\|X_{k+1}\|_F}{\|X_k\|_F}$
0	1.0e+0	2.9e-16		1.0e+0	2.9e-16	8.1e+04	
1	1.0e+0	1.0e+0	2.9e-1	1.0e+0	7.9e-15	1.6e+04	8.0e-1
2	1.0e+0	1.0e+0	1.0e+0	1.0e+0	1.9e-13	3.2e+03	8.0e-1
3	9.7e-01	9.7e-1	9.5e-1	1.0e+0	4.8e-12	6.4e+02	8.0e-1
4	5.3e-01	5.3e-1	6.2e-1	1.0e+0	1.2e-10	1.3e+02	8.0e-1
5	5.9e-02	5.9e-2	1.3e-1	1.0e+0	3.0e-09	2.5e+01	8.0e-1
6	4.8e-04	4.8e-4	8.7e-3	9.6e-01	7.3e-08	4.2e+0	7.5e-1
7	3.8e-08	3.8e-8	4.3e-5	4.4e-01	1.1e-06	3.4e-01	3.7e-1
8	6.1e-16	1.1e-8	2.9e-9	2.5e-04	1.9e-06	1.2e-04	1.1e-2
9	6.3e-16	1.1e-8	0.0e+0	1.5e-15	1.9e-06	2.2e-16	2.0e-9
10				7.3e-16	1.9e-06	0.0e+0	0.0e+0
E	3.6e-10			1.8e-6			

9. Conclusions. When a problem has structure it is important to exploit it to advantage. This work was motivated by the discovery of Mackey, Mackey, and Tisseur [22] that the polar and matrix sign factors of matrices from automorphism groups $\mathbb{G}$ also lie in the group: unconditionally for the sign decomposition, and provided the matrix of the underlying form is unitary for the polar decomposition. We have identified a family of globally convergent rational iterations that preserve group structure and shown how structure preservation leads to particularly convenient convergence tests in the case of the polar decomposition.

The most surprising results in this work concern Newton's method. Although Newton's method for the polar decomposition immediately destroys the underlying group structure, when $\mathbb{G} \in \mathfrak{U}$ it forces equality between the adjoint and the conjugate transpose of each iterate. This implies that the Newton iterates approach the group at the same rate that they approach unitarity. It also yields "commutativity" relations that for certain groups imply a different, exploitable structure. Similar properties hold for Newton's method for the matrix sign function, here with no restrictions on $\mathbb{G}$.

We have identified various pros and cons in the "structured iteration versus scaled Newton" comparison, including the slightly better empirically observed numerical stability of Newton, the convergence prediction possible with the structured iterations,

and the fact that, in practice, scaled Newton usually requires the fewest flops.

Our conclusion is that the Newton iteration (3.7) with Frobenius norm scaling (3.8) and the cubic (3.2) and quintic (3.5) structure-preserving iterations are all well-suited to computing the polar decomposition of a matrix from one of the automorphism groups under consideration. Likewise, for the matrix sign decomposition the scaled Newton iteration (7.1) and the obvious analogues of the cubic and quintic iterations are all suitable. Which of the iterations is to be preferred depends on the matrix A, the group $\mathbb{G}$, and the user's accuracy requirements.

REFERENCES

[1] Z. Bai and J. W. Demmel, *Using the matrix sign function to compute invariant subspaces*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 205–225.

[2] 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., 261 (1997), pp. 91–141.

[3] J. R. Cardoso, C. S. Kenney, and F. Silva Leite, *Computing the square root and logarithm of a real P-orthogonal matrix*, Appl. Numer. Math., 46 (2003), pp. 173–196.

[4] J. R. Cardoso and F. Silva Leite, *Extending Results from Orthogonal Matrices to the Class of P-Orthogonal Matrices*. manuscript, 2002.

[5] A. A. Dubrulle, *An optimum iteration for the matrix polar decomposition*, Electron. Trans. Numer. Anal., 8 (1999), pp. 21–25.

[6] K. Fan and A. J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[7] W. Gander, *On Halley's iteration method*, Amer. Math. Monthly, 92 (1985), pp. 131–134.

[8] N. J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[9] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 (1994), pp. 3–20.

[10] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[11] N. J. Higham, *J-orthogonal matrices: Properties and generation*, SIAM Rev., 45 (2003), pp. 504–519.

[12] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Functions Preserving Matrix Groups and Iterations for the Matrix Square Root*, Numerical Analysis Report 446, Manchester Centre for Computational Mathematics, Manchester, England, 2004.

[13] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[14] N. Jacobson, *Basic Algebra* I, W. H. Freeman and Company, San Francisco, 1974.

[15] C. S. Kenney and A. J. Laub, *Rational iterative methods for the matrix sign function*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 273–291.

[16] C. S. Kenney and A. J. Laub, *On scaling Newton's method for polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 688–706.

[17] C. S. Kenney and A. J. Laub, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.

[18] A. Kiełbasiński and K. Ziętak, *Numerical behaviour of Higham's scaled method for polar decomposition*, Numer. Algorithms, 32 (2003), pp. 105–140.

[19] S. Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1993.

[20] D. S. Mackey and N. Mackey, *On the Construction of Structure-Preserving Matrix Iterations*, in preparation.

[21] D. S. Mackey, N. Mackey, and F. Tisseur, *$\mathbb{G}$-Reflectors: Analogues of Householder Transformations in Scalar Product Spaces*, Numerical Analysis Report 420, Manchester Centre for Computational Mathematics, Manchester, England, Feb. 2003. Linear Algebra Appl., to appear.

[22] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured Factorizations in Scalar Product Spaces*, Numerical Analysis Report 432, Manchester Centre for Computational Mathematics, Manchester, England, 2004.

[23] 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.

[24] R. Shaw, *Linear Algebra and Group Representations. Vol.* I, Academic Press, London, 1982.