SIAM J. Matrix Anal. Appl.
Vol. 24, No. 1, pp. 1–16

MEANS AND AVERAGING IN THE GROUP OF ROTATIONS*

MAHER MOAKHER†

Abstract. In this paper we give precise definitions of different, properly invariant notions of mean or average rotation. Each mean is associated with a metric in $SO(3)$. The metric induced from the Frobenius inner product gives rise to a mean rotation that is given by the closest special orthogonal matrix to the usual arithmetic mean of the given rotation matrices. The mean rotation associated with the intrinsic metric on $SO(3)$ is the Riemannian center of mass of the given rotation matrices. We show that the Riemannian mean rotation shares many common features with the geometric mean of positive numbers and the geometric mean of positive Hermitian operators. We give some examples with closed-form solutions of both notions of mean.

Key words. special orthogonal group, rotation, geodesics, operator means, averaging

AMS subject classifications. 47A64, 65F30

PII. S0895479801383877

1. Introduction.

In many applications, such as the study of plate tectonics [22] or sequence-dependent continuum modeling of DNA [19], experimental data are given as a sequence of three-dimensional orientation data that usually contain a substantial amount of noise. A common problem is to remove or reduce the noise by processing the raw data, for example by the construction of a suitable filter, in order to obtain appropriately smooth data.

Three-dimensional orientation data are elements of the group of rotations that generally are given as a sequence of proper orthogonal matrices, or a sequence of Euler angles, or a sequence of unit quaternions, etc. As the group of rotations is not a Euclidean space, but rather a differentiable manifold, the notion of mean or average is not obvious so that appropriate filters are similarly not obvious. One might choose some local coordinate representation of the group—for instance, a set of Euler angles—then apply the usual averaging and smoothing techniques of Euclidean spaces. Although this approach is simple to implement, it is not properly invariant under the action of rigid transformations. In this article alternative approaches will be discussed.

There is extensive literature on the statistics of circular and spherical data; see [20, 27, 9, 8] and the references therein. In a more general context, Downs [4], Khatri and Mardia [18], and Jupp and Mardia [15] developed statistical methods for data in the Stiefel manifold, i.e., the Riemannian space $V_{n,p},\ 1 \leq p \leq n$, of $n \times p$ orthogonal matrices (the hypersphere S^n and the special orthogonal group $SO(n)$ are examples of such manifolds). The general approach of these statistical studies is to embed the given data into a Euclidean space of dimension larger than the dimension of the manifold (circle, sphere, hypersphere, etc.), then to pursue standard statistical approaches in this linear space, and finally to project the result onto the manifold. Prentice [22] used the parameterization of the group of rotations by four-dimensional axes (unsigned unit

*Received by the editors January 19, 2001; accepted for publication (in revised form) by U. Helmke January 8, 2002; published electronically May 15, 2002. This work was partially supported by the Swiss National Science Foundation.
http://www.siam.org/journals/simax/24-1/38387.html
†Faculté des Sciences de Base, Institut de Recherches Mathématiques Interdisciplinaires, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (Maher.Moakher@epfl.ch).

quaternions) and a slight modification of the algorithm of smoothing directional data on S^2 proposed in [16] to fit smooth spline paths to three-dimensional rotation data.

In this paper we are mainly concerned with a general mathematical theory of different possible notions of mean in the group of three-dimensional rotations rather than a statistical theory based on a specific notion of mean. In analogy with mean in Euclidean space, we define the mean rotation of a given sequence of rotations to be the minimizer of the sum of squared distances from the given rotations. The *projected arithmetic mean* is obtained when one uses the inherent Euclidean distance of the ambient space. We show that this is the orthogonal projection of the usual arithmetic mean in the space of 3×3 matrices onto the rotation group. It is the same as the directional mean of the statistics literature mentioned above. The *geometric mean* arises when one uses the Riemannian metric intrinsic to the group of rotations. We find close similarities between this mean and the geometric mean of positive numbers, as well as the geometric mean of positive Hermitian operators. We show that these two notions of mean are properly invariant under a change of frame and share many common properties with means of elements of Euclidean spaces.

To the best of our knowledge, the geometric mean rotation has not been discussed previously. In this paper, we show that the geometric mean and the Euclidean mean rotation, which we call the projected arithmetic mean, each arise from a least-squares error approach, but with different metrics. We also give some properties of the Euclidean mean rotation that have not been discussed in the literature, as well as its connection with the geometric mean.

The remainder of this paper is organized as follows. In section 2 we gather all the necessary background from Lie group theory, differential geometry, and optimization on manifolds that will be used in what follows. Further information on this condensed material can be found in [3, 1, 23, 21, 26, 13]. In section 3 we introduce two bi-invariant notions of mean rotation: the projected arithmetic mean and the geometric mean. We give the characterization and main features of these two notions of mean rotation. Examples of closed-form calculations of mean rotations are given in section 4. Finally, weighted means and power means of rotations are presented in section 5.

2. Geometry of the rotation group. Let $\mathcal{M}(3)$ be the set of 3×3 real matrices and $GL(3)$ be its subset containing only nonsingular matrices. The group of rotations in $\mathbb{R}^3$, denoted by $SO(3)$, is the Lie group of special orthogonal transformations in $\mathbb{R}^3$,

$$SO(3) = \left\{ \boldsymbol{R} \in GL(3) \mid \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I} \text{ and } \det \boldsymbol{R} = 1 \right\}, \tag{2.1}$$

where $\boldsymbol{I}$ is the identity transformation in $\mathbb{R}^3$ and the superscript T denotes the transpose. The corresponding Lie algebra, denoted by $\mathfrak{so}(3)$, is the space of skew-symmetric matrices

$$\mathfrak{so}(3) = \left\{ \boldsymbol{A} \in \mathfrak{gl}(3) \mid \boldsymbol{A}^T = -\boldsymbol{A} \right\}, \tag{2.2}$$

where $\mathfrak{gl}(3)$, the space of linear transformations in $\mathbb{R}^3$, is the Lie algebra corresponding to Lie group $GL(3)$.

2.1. Exponential and logarithm. The exponential of a matrix $\boldsymbol{X}$ in $GL(3)$ is denoted $\exp \boldsymbol{X}$ and is given by the limit of the convergent series $\exp \boldsymbol{X} = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{X}^k$. When a matrix $\boldsymbol{Y}$ in $GL(3)$ does not have eigenvalues in the (closed) negative real line, there exists a unique real logarithm, called the principal logarithm, denoted by $\operatorname{Log} \boldsymbol{Y}$,

whose spectrum lies in the infinite strip $\{z \in \mathbb{C} \,:\, -\pi < \operatorname{Im}(z) < \pi\}$ of the complex plane [3]. Furthermore, for any given matrix norm $\|\cdot\|$, if $\|\boldsymbol{I} - \boldsymbol{Y}\| < 1$, then the series $-\sum_{k=1}^{\infty} \frac{(\boldsymbol{I}-\boldsymbol{Y})^k}{k}$ converges, and hence one can write $\operatorname{Log} \boldsymbol{Y} = -\sum_{k=1}^{\infty} \frac{(\boldsymbol{I}-\boldsymbol{Y})^k}{k}$. However, as we will describe, the infinite series representations of the exponential of matrices in $\mathfrak{so}(3)$ and the logarithm of matrices in $SO(3)$ can be given as closed-form expressions.

The exponential of a skew-symmetric matrix $\boldsymbol{A}$, such that $a = \sqrt{\frac{1}{2}\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})}$ is in $[0, \pi)$, is the proper orthogonal matrix given by Rodrigues' formula

$$\exp \boldsymbol{A} = \begin{cases} \boldsymbol{I} & \text{if } a = 0, \\ \boldsymbol{I} + \dfrac{\sin a}{a}\boldsymbol{A} + \dfrac{1-\cos a}{a^2}\boldsymbol{A}^2 & \text{if } a \neq 0. \end{cases} \tag{2.3}$$

The principal logarithm for a matrix $\boldsymbol{R}$ in $SO(3)$ is the matrix in $\mathfrak{so}(3)$ given by

$$\operatorname{Log} \boldsymbol{R} = \begin{cases} \boldsymbol{0} & \text{if } \theta = 0, \\ \dfrac{\theta}{2\sin\theta}(\boldsymbol{R} - \boldsymbol{R}^T) & \text{if } \theta \neq 0, \end{cases} \tag{2.4}$$

where θ satisfies $\operatorname{tr} \boldsymbol{R} = 1 + 2\cos\theta$ and $|\theta| < \pi$. (This formula breaks down when $\theta = \pm\pi$.) An alternative expression for the logarithm of a matrix in $SO(3)$, where the parameter θ does not appear, is given in [14].

Solutions in $SO(3)$ of the matrix equation $\boldsymbol{Q}^k = \boldsymbol{R}$ with k a positive integer will be called kth roots of $\boldsymbol{R}$. These kth roots are given by

$$\exp\left(\frac{1}{k}\left(1 + \frac{2l\pi}{\theta}\right)\operatorname{Log} \boldsymbol{R}\right), \quad l = 0, \ldots, k-1,$$

where θ is the angle of rotation of $\boldsymbol{R}$. The kth root $\exp(\frac{1}{k}\operatorname{Log}\boldsymbol{R})$ is the one for which the eigenvalues have the largest positive real part and is the only one we denote by $\boldsymbol{R}^{1/k}$. In the case $k = 2$, it is the only square root with positive real part.

2.2. Metrics in $SO(3)$. A straightforward way to define a distance function in $SO(3)$ is to use the Euclidean distance of the ambient space $\mathcal{M}(3)$, i.e., if $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are two rotation matrices, then

$$\mathrm{d}_F(\boldsymbol{R}_1, \boldsymbol{R}_2) = \|\boldsymbol{R}_1 - \boldsymbol{R}_2\|_F, \tag{2.5}$$

where $\|\cdot\|_F$ is the Frobenius norm which is induced by the Euclidean inner product, known as the Frobenius inner product, defined by $\langle \boldsymbol{R}_1, \boldsymbol{R}_2 \rangle_F = \operatorname{tr}(\boldsymbol{R}_1^T\boldsymbol{R}_2)$. It is easy to see that this distance is bi-invariant in $SO(3)$, i.e., $\mathrm{d}_F(\boldsymbol{P}\boldsymbol{R}_1\boldsymbol{Q}, \boldsymbol{P}\boldsymbol{R}_2\boldsymbol{Q}) = \mathrm{d}_F(\boldsymbol{R}_1, \boldsymbol{R}_2)$ for all $\boldsymbol{P}, \boldsymbol{Q}$ in $SO(3)$.

Another way to define a distance function in $SO(3)$ is to use its Riemannian structure. The Riemannian distance between two rotations $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ is given by

$$\mathrm{d}_R(\boldsymbol{R}_1, \boldsymbol{R}_2) = \frac{1}{\sqrt{2}}\|\operatorname{Log}(\boldsymbol{R}_1^T\boldsymbol{R}_2)\|_F. \tag{2.6}$$

It is the length of the shortest geodesic curve that connects $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ given by

$$\boldsymbol{Q}(t) = \boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^t = \boldsymbol{R}_1 \exp(t \operatorname{Log}(\boldsymbol{R}_1^T\boldsymbol{R}_2)), \quad 0 \leq t \leq 1. \tag{2.7}$$

Note that the geodesic curve of minimal length may not be unique. If $\boldsymbol{R}_1^T\boldsymbol{R}_2$ is an involution, in other words if $(\boldsymbol{R}_1^T\boldsymbol{R}_2)^2 = \boldsymbol{I}$, i.e., a rotation through an angle π, then

$\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ can be connected by two curves of equal length. In such a case, the rotations $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are said to be antipodal points in $SO(3)$ and $\boldsymbol{R}_2$ is said to be the cut point of $\boldsymbol{R}_1$ and vice versa.

The Riemannian distance (2.6) is also bi-invariant in $SO(3)$. Indeed, using the fact $\operatorname{Log}(\boldsymbol{Q}^{-1}\boldsymbol{R}\boldsymbol{Q}) = \boldsymbol{Q}^{-1}(\operatorname{Log}\boldsymbol{R})\boldsymbol{Q}$ [3], we can show that $\mathrm{d}_R(\boldsymbol{P}\boldsymbol{R}_1\boldsymbol{Q}, \boldsymbol{P}\boldsymbol{R}_2\boldsymbol{Q}) = \mathrm{d}_R(\boldsymbol{R}_1, \boldsymbol{R}_2)$ for all $\boldsymbol{P}, \boldsymbol{Q}$ in $SO(3)$.

Remark 2.1. The Euclidean distance (2.5) represents the chordal distance between $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, i.e., the length of the Euclidean line segment in the space of $\mathcal{M}(3)$ (except for the end points $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, this line segment does not lie in $SO(3)$), whereas the Riemannian distance (2.6) represents the arc-length of the shortest geodesic curve (great-circle arc), which lies entirely in $SO(3)$, passing through $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$.

Remark 2.2. If θ denotes the angle of rotation of $\boldsymbol{R}_1^T\boldsymbol{R}_2$, then $\mathrm{d}_F(\boldsymbol{R}_1, \boldsymbol{R}_2) = 2\sqrt{2}|\sin\frac{\theta}{2}|$ and $\mathrm{d}_R(\boldsymbol{R}_1, \boldsymbol{R}_2) = |\theta|$. Therefore, when the rotations $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are sufficiently close, i.e., θ is small, we have $\mathrm{d}_F(\boldsymbol{R}_1, \boldsymbol{R}_2) \approx \sqrt{2}\,\mathrm{d}_R(\boldsymbol{R}_1, \boldsymbol{R}_2)$.

2.3. Covariant derivative and Hessian. We recall that the tangent space at a point $\boldsymbol{R}$ of $SO(3)$ is the space of all matrices $\boldsymbol{\Delta}$ such that $\boldsymbol{R}^T\boldsymbol{\Delta}$ is skew symmetric and that the normal space (associated with the Frobenius inner product) at $\boldsymbol{R}$ consists of all matrices $\boldsymbol{N}$ such that $\boldsymbol{R}^T\boldsymbol{N}$ is symmetric [5].

For a real-valued function $f(\boldsymbol{R})$ defined on $SO(3)$, the covariant derivative ∇f is the unique tangent vector at $\boldsymbol{R}$ such that

$$\operatorname{tr}(\boldsymbol{\Delta}^T\nabla f) = \left.\frac{d}{dt} f(\boldsymbol{Q}(t))\right|_{t=0}, \tag{2.8}$$

where $\boldsymbol{Q}(t)$ is a geodesic emanating from $\boldsymbol{R}$ in the direction of $\boldsymbol{\Delta}$, i.e., $\boldsymbol{Q}(t) = \boldsymbol{R}\exp(t\boldsymbol{A})$ and $\boldsymbol{R}^T\boldsymbol{\Delta} = \boldsymbol{A}$ is skew symmetric.

The Hessian of $f(\boldsymbol{R})$ is given by the quadratic form

$$\operatorname{Hess} f(\boldsymbol{\Delta}, \boldsymbol{\Delta}) = \left.\frac{d^2}{dt^2} f(\boldsymbol{Q}(t))\right|_{t=0}, \tag{2.9}$$

where $\boldsymbol{Q}(t)$ is a geodesic and $\boldsymbol{\Delta}$ is in the tangent space at $\boldsymbol{R}$ as above.

2.4. Geodesic convexity. We recall that a subset A of a Riemannian manifold M is said to be convex if the shortest geodesic curve between any two points x and y in A is unique in M and lies in A. A real-valued function defined on a convex subset A of M is said to be convex if its restriction to any geodesic path is convex, i.e., if $t \mapsto \hat{f}(t) \equiv f(\exp_x(tu))$ is convex over its domain for all $x \in M$ and $u \in T_x(M)$, where $\exp_x$ is the exponential map at x.

With these definitions, one can readily see that any geodesic ball $B_r(\boldsymbol{Q})$ in $SO(3)$ of radius r less than $\frac{\pi}{2}$ around $\boldsymbol{Q}$ is convex and that the real-valued function f defined on $B_r(\boldsymbol{Q})$ by $f(\boldsymbol{R}) = \|\operatorname{Log}(\boldsymbol{Q}^T\boldsymbol{R})\|_F$ is convex when r is less than $\frac{\pi}{2}$. Geodesic balls with radius greater than or equal to $\frac{\pi}{2}$ are not convex.

3. Mean rotation. For a given set of N points $\boldsymbol{x}_n$, $n = 1, \ldots, N$, in $\mathbb{R}^d$ the arithmetic mean $\bar{\boldsymbol{x}}$ is given by the barycenter $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n$ of the N points. The arithmetic mean also has a variational property; it minimizes the sum of the squared distances to the given points $\boldsymbol{x}_n$,

$$\bar{\boldsymbol{x}} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} \sum_{n=1}^N \mathrm{d}_e(\boldsymbol{x}, \boldsymbol{x}_n)^2, \tag{3.1}$$

where here $\mathrm{d}_e(\cdot,\cdot)$ represents the usual Euclidean distance in $\mathbb{R}^d$.

One can also use the arithmetic mean to average N positive real numbers $x_n > 0,\ n = 1,\dots,N$, and the mean is itself a positive number. In many applications, however, it is more appropriate to use the geometric mean to average positive numbers, which is possible because positive numbers form a multiplicative group. The geometric mean $\tilde{x} = x_1^{1/N}\cdots x_N^{1/N}$ also has a variational property; it minimizes the sum of the squared *hyperbolic distances* to the given data

$$\tilde{x} = \arg\min_{x>0} \sum_{n=1}^{N} \mathrm{d}_h(x_n, x)^2, \tag{3.2}$$

where $\mathrm{d}_h(x,y) = |\log x - \log y|$ is the hyperbolic distance[1] between x and y.

As we have seen, for the set of positive real numbers, different notions of mean can be associated with different metrics. In what follows, we will extend these notions of mean to the group of proper orthogonal matrices.

By analogy with $\mathbb{R}^d$, a plausible definition of the mean of N rotation matrices $\boldsymbol{R}_1,\dots,\boldsymbol{R}_N$ is that it is the minimizer in $SO(3)$ of the sum of the squared distances from that rotation matrix to the given rotation matrices $\boldsymbol{R}_1,\dots,\boldsymbol{R}_N$, i.e., $\mathfrak{M}(\boldsymbol{R}_1,\dots,\boldsymbol{R}_N) = \arg\min_{\boldsymbol{R}\in SO(3)} \sum_{n=1}^{N} \mathrm{d}(\boldsymbol{R}_n,\boldsymbol{R})^2$, where $\mathrm{d}(\cdot,\cdot)$ represents a distance in $SO(3)$. Now the two distance functions (2.5) and (2.6) define the two different means.

DEFINITION 3.1. *The mean rotation in the Euclidean sense, i.e., associated with the metric* (2.5)*, of N given rotation matrices $\boldsymbol{R}_1,\dots,\boldsymbol{R}_N$ is defined as*

$$\mathfrak{A}(\boldsymbol{R}_1,\dots,\boldsymbol{R}_N) := \arg\min_{\boldsymbol{R}\in SO(3)} \sum_{n=1}^{N} \|\boldsymbol{R}_n - \boldsymbol{R}\|_F^2. \tag{3.3}$$

DEFINITION 3.2. *The mean rotation in the Riemannian sense, i.e., associated with the metric* (2.6)*, of N given rotation matrices $\boldsymbol{R}_1,\dots,\boldsymbol{R}_N$ is defined as*

$$\mathfrak{G}(\boldsymbol{R}_1,\dots,\boldsymbol{R}_N) := \arg\min_{\boldsymbol{R}\in SO(3)} \sum_{n=1}^{N} \|\operatorname{Log}(\boldsymbol{R}_n^T\boldsymbol{R})\|_F^2. \tag{3.4}$$

The minimum here is understood to be the global minimum. We remark that in $\mathbb{R}^d$, or in the set of positive numbers, the objective functions to be minimized are convex over their domains, and therefore the means are well defined and unique. However, in $SO(3)$, as we shall see, the objective functions in (3.3) and (3.4) are not (geodesically) convex, and therefore the means may not be unique.

Before we proceed to study these two means, we note that both satisfy the following desirable properties that one would expect from a mean in $SO(3)$, and that are counterparts of properties of means of numbers, namely:

1. *Invariance under permutation.* For any permutation σ of the numbers 1 through N, we have $\mathfrak{M}(\boldsymbol{R}_{\sigma(1)},\dots,\boldsymbol{R}_{\sigma(N)}) = \mathfrak{M}(\boldsymbol{R}_1,\dots,\boldsymbol{R}_N)$.

2. *Bi-invariance.* If $\boldsymbol{R}$ is the mean rotation of $\{\boldsymbol{R}_n\},\ n = 1,\dots,N$, then $\boldsymbol{PRQ}$ is the mean rotation of $\{\boldsymbol{PR}_n\boldsymbol{Q}\},\ n = 1,\dots,N$, for every $\boldsymbol{P}$ and $\boldsymbol{Q}$ in $SO(3)$. This property follows immediately from the bi-invariance of the two metrics defined above.

[1] We borrow this terminology from the hyperbolic geometry of the Poincaré upper half-plane. In fact, the hyperbolic length of the geodesic segment joining the points $P(a,y_1)$ and $Q(a,y_2)$, $y_1,y_2>0$, is $|\log\frac{y_1}{y_2}|$ (see [26]).

3. *Invariance under transposition.* If $\boldsymbol{R}$ is the mean rotation of $\{\boldsymbol{R}_n\}$, $n = 1, \ldots, N$, then $\boldsymbol{R}^T$ is the mean rotation of $\{\boldsymbol{R}_n^T\}$, $n = 1, \ldots, N$.

We remark that the bi-invariance property is in some sense the counterpart of the homogeneity property of means of positive numbers (but here left and right multiplication are both needed because the rotation group is not commutative).

3.1. Characterization of the Euclidean mean. The following proposition gives a relation between the Euclidean mean and the usual arithmetic mean.

PROPOSITION 3.3. *The mean rotation* $\mathfrak{A}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)$ *of* $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N \in SO(3)$ *is the orthogonal projection of* $\overline{\boldsymbol{R}} = \sum_{n=1}^N \frac{\boldsymbol{R}_n}{N}$ *onto the special orthogonal group* $SO(3)$*. In other words, the mean rotation in the Euclidean sense is the projection of the arithmetic mean* $\overline{\boldsymbol{R}}$ *of* $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ *in the linear space* $\mathcal{M}(3)$ *onto* $SO(3)$*.*

Proof. As $\boldsymbol{R}_n$, $n = 1, \ldots, N$, and $\boldsymbol{R}$ are all orthogonal, it follows that

$$\mathfrak{A}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \sum_{n=1}^N \|\boldsymbol{R}_n - \boldsymbol{R}\|_F^2 = \operatorname*{arg\,max}_{\boldsymbol{R} \in SO(3)} \operatorname{tr}(\overline{\boldsymbol{R}}^T \boldsymbol{R}).$$

On the other hand, the orthogonal projection of $\overline{\boldsymbol{R}}$ onto $SO(3)$ is given by

$$\begin{aligned}
\Pi(\overline{\boldsymbol{R}}) &= \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \|\overline{\boldsymbol{R}} - \boldsymbol{R}\|_F = \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \|\overline{\boldsymbol{R}} - \boldsymbol{R}\|_F^2 \\
&= \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \left[\sum_{n=1}^N \sum_{m=1}^N \operatorname{tr}\left(\frac{\boldsymbol{R}_n}{N} \frac{\boldsymbol{R}_m^T}{N} \right) - 2 \operatorname{tr}\left(\sum_{n=1}^N \frac{\boldsymbol{R}_n^T}{N} \boldsymbol{R} \right) + \operatorname{tr} \boldsymbol{I} \right] \\
&= \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} -2 \operatorname{tr}\left(\sum_{n=1}^N \frac{\boldsymbol{R}_n^T}{N} \boldsymbol{R} \right) = \operatorname*{arg\,max}_{\boldsymbol{R} \in SO(3)} \operatorname{tr}\left(\overline{\boldsymbol{R}}^T \boldsymbol{R} \right). \qquad \square
\end{aligned}$$

Because of Proposition 3.3, the mean in the Euclidean sense will be termed the *projected arithmetic mean* to reflect the fact that it is the orthogonal projection of the usual arithmetic mean in $\mathcal{M}(3)$ onto $SO(3)$.

Remark 3.4. The projected arithmetic mean can now be seen to be related to the classical orthogonal Procrustes problem [10], which seeks the orthogonal matrix that most closely transforms a given matrix into a second one.

PROPOSITION 3.5. *If* $\det \overline{\boldsymbol{R}}$ *is positive, then the mean rotation in the Euclidean sense* $\mathfrak{A}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)$ *of* $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N \in SO(3)$ *is given by the unique polar factor in the polar decomposition* [10] *of* $\overline{\boldsymbol{R}}$*.*

Proof. Critical points of the objective function

$$F(\boldsymbol{R}) = \sum_{n=1}^N \|\boldsymbol{R} - \boldsymbol{R}_n\|_F^2 \tag{3.5}$$

defined on $SO(3)$ and corresponding to the minimization problem (3.3) are those elements of $SO(3)$ for which the covariant derivative of (3.5) vanishes. Using (2.8) we get $\nabla F = \sum_{n=1}^N \boldsymbol{R}(\boldsymbol{R}_n^T \boldsymbol{R} - \boldsymbol{R}^T \boldsymbol{R}_n)$. Therefore, critical points of (3.5) are the rotation matrices $\boldsymbol{R}$ such that $\sum_{n=1}^N \boldsymbol{R}(\boldsymbol{R}_n^T \boldsymbol{R} - \boldsymbol{R}^T \boldsymbol{R}_n) = \boldsymbol{0}$, or, equivalently, for which the matrix $\boldsymbol{S}$ defined by

$$\boldsymbol{S} = \boldsymbol{R}^T \sum_{n=1}^N \boldsymbol{R}_n = N \boldsymbol{R}^T \overline{\boldsymbol{R}} \tag{3.6}$$

is symmetric.

Since $\boldsymbol{R}$ is orthogonal, and both $\boldsymbol{S}$ and $\boldsymbol{M} = \overline{\boldsymbol{R}}^T\overline{\boldsymbol{R}}$ are symmetric, it follows that $\boldsymbol{S}^2 = N^2\boldsymbol{M}$. Therefore, there exists an orthogonal matrix $\boldsymbol{U}$ such that $\boldsymbol{S}^2 = N^2\boldsymbol{U}^T\boldsymbol{D}\boldsymbol{U}$, where $\boldsymbol{D} = \operatorname{diag}(\Lambda_1, \Lambda_2, \Lambda_3)$ with $\Lambda_1 \geq \Lambda_2 \geq \Lambda_3 \geq 0$ being the eigenvalues of $\boldsymbol{M}$. The eight possible square roots of $\boldsymbol{M}$ are $\boldsymbol{U}^T \operatorname{diag}(\pm\sqrt{\Lambda_1}, \pm\sqrt{\Lambda_2}, \pm\sqrt{\Lambda_3})\boldsymbol{U}$. To determine the square root $\boldsymbol{S} = \boldsymbol{U}^T \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)\boldsymbol{U}$ of $N^2\boldsymbol{M}$ that corresponds to the minimum of (3.5) we require that the Hessian of the objective function (3.5) at $\boldsymbol{R}$ given by (3.6) be positive for all tangent vectors $\boldsymbol{\Delta}$ at $\boldsymbol{R}$. From (2.9) we obtain $\operatorname{Hess} F(\boldsymbol{\Delta}, \boldsymbol{\Delta}) = 2N \operatorname{tr}\left(\overline{\boldsymbol{R}}^T\boldsymbol{R}\boldsymbol{\Delta}\boldsymbol{\Delta}^T\right)$, and therefore at $\boldsymbol{R}$ given by (3.6) we have

$$\operatorname{Hess} F(\boldsymbol{\Delta}, \boldsymbol{\Delta}) = 2[(\lambda_2 + \lambda_3)a^2 + (\lambda_1 + \lambda_3)b^2 + (\lambda_1 + \lambda_2)c^2],$$

where a, b, c are such that

$$\boldsymbol{\Delta} = \boldsymbol{U}^T\boldsymbol{R}\boldsymbol{B}\boldsymbol{U} \quad \text{and} \quad \boldsymbol{B} = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix}.$$

As we are looking for a proper rotation matrix, i.e., an orthogonal matrix with determinant one, it follows from (3.6) that $\det \boldsymbol{S} = N \det \overline{\boldsymbol{R}}$. We therefore conclude that $\operatorname{Hess} F(\boldsymbol{\Delta}, \boldsymbol{\Delta})$ is positive for all tangent vectors $\boldsymbol{\Delta}$ at $\boldsymbol{R}$ if and only if $\lambda_1 = N\sqrt{\Lambda_1}$, $\lambda_2 = N\sqrt{\Lambda_2}$, and $\lambda_3 = sN\sqrt{\Lambda_3}$, where $s = 1$ if $\det \overline{\boldsymbol{R}}$ is positive and $s = -1$ otherwise. In fact, (3.5) has four critical points belonging to $SO(3)$ which consist of a minimum $[(\lambda_1, \lambda_2, \lambda_3) = N(\sqrt{\Lambda_1}, \sqrt{\Lambda 2}, s\sqrt{\Lambda_3})]$, a maximum $[(\lambda_1, \lambda_2, \lambda_3) = N(-\sqrt{\Lambda_1}, -\sqrt{\Lambda_2}, -s\sqrt{\Lambda_3})]$, and two saddle points $[(\lambda_1, \lambda_2, \lambda_3) = N(-\sqrt{\Lambda_1}, s\sqrt{\Lambda_2}, -\sqrt{\Lambda_3})$ and $(\lambda_1, \lambda_2, \lambda_3) = N(s\sqrt{\Lambda_1}, -\sqrt{\Lambda_2}, -\sqrt{\Lambda_3})]$.

Hence, the projected arithmetic mean is given by

$$\boldsymbol{R} = \overline{\boldsymbol{R}}\boldsymbol{U} \operatorname{diag}\left(\frac{1}{\sqrt{\Lambda_1}}, \frac{1}{\sqrt{\Lambda_2}}, \frac{s}{\sqrt{\Lambda_3}}\right)\boldsymbol{U}^T, \tag{3.7}$$

which, when $\det \overline{\boldsymbol{R}} > 0$, coincides with the polar factor of the polar decomposition of $\overline{\boldsymbol{R}}$. Of course uniqueness fails when the smallest eigenvalue of $\boldsymbol{M}$ is not simple. ∎

Remark 3.6. The case where $\det \overline{\boldsymbol{R}} = 0$ is a degenerate case. However, if $\overline{\boldsymbol{R}}$ has rank 2, i.e., when $\Lambda_1 \geq \Lambda_2 > \Lambda_3 = 0$, one can still find a unique closest proper orthogonal matrix to $\overline{\boldsymbol{R}}$ (see [6] for details) and hence can define the mean rotation in the Euclidean sense.

3.2. Characterization of the Riemannian mean. First, we compute the derivative of the real-valued function $H(\boldsymbol{P}(t)) = \frac{1}{2}\|\operatorname{Log}(\boldsymbol{Q}^T\boldsymbol{P}(t))\|_F^2$ with respect to t, where $\boldsymbol{P}(t) = \boldsymbol{R}\exp(t\boldsymbol{A})$ is the geodesic emanating from $\boldsymbol{R}$ in the direction of $\boldsymbol{\Delta} = \dot{\boldsymbol{P}}(0) = \boldsymbol{R}\boldsymbol{A}$. As $\boldsymbol{\Delta}$ is in the tangent space at $\boldsymbol{R}$, we have $\boldsymbol{A} = \boldsymbol{R}^T\boldsymbol{\Delta} = -\boldsymbol{\Delta}^T\boldsymbol{R}$. Let $\theta(t) \in (-\pi, \pi)$ be the angle of rotation of $\boldsymbol{Q}^T\boldsymbol{P}(t)$, i.e., such that

$$\operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{P}(t)) = 1 + 2\cos\theta(t). \tag{3.8}$$

Differentiate (3.8) to get $\frac{d}{dt}H(\boldsymbol{P}(t))|_{t=0} = -\frac{\phi}{\sin\phi}\operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A})$, where $\phi = \theta(0)$ is the angle of rotation of $\boldsymbol{Q}^T\boldsymbol{R}$ and we have used the fact that $H(\boldsymbol{P}(t)) = \theta(t)^2$.

Recall that since $\boldsymbol{A}$ is skew symmetric, $\operatorname{tr}(\boldsymbol{S}\boldsymbol{A}) = 0$ for any symmetric matrix $\boldsymbol{S}$. It follows that $\operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A}) = \frac{1}{2}\operatorname{tr}([\boldsymbol{Q}^T\boldsymbol{R} - \boldsymbol{R}^T\boldsymbol{Q}]\boldsymbol{A})$. Hence

$$\operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A}) = \frac{1}{2}\operatorname{tr}[(\boldsymbol{Q}^T\boldsymbol{R} - \boldsymbol{R}^T\boldsymbol{Q})\boldsymbol{R}^T\boldsymbol{\Delta}] = \frac{1}{2}\operatorname{tr}[\boldsymbol{\Delta}^T\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{Q} - \boldsymbol{Q}^T\boldsymbol{R})].$$

Then, with the help of (2.4) we obtain $\frac{d}{dt}H(\boldsymbol{P}(t))|_{t=0} = \mathrm{tr}[\boldsymbol{\Delta}^T \boldsymbol{R}\,\mathrm{Log}(\boldsymbol{Q}^T\boldsymbol{R})]$. Therefore, the covariant derivative of H is given by

$$\nabla H = \boldsymbol{R}\,\mathrm{Log}(\boldsymbol{Q}^T\boldsymbol{R}). \tag{3.9}$$

The second derivative of (3.8) gives

$$\left.\frac{d^2}{dt^2}H(\boldsymbol{P}(t))\right|_{t=0} = \frac{\sin\phi - \phi\cos\phi}{4\sin^3\phi}[\mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A})]^2 - \frac{\phi}{2\sin\phi}\,\mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A}^2).$$

Let $\boldsymbol{U}$ be an orthogonal matrix and $\boldsymbol{B}$ the skew-symmetric matrix such that

$$\boldsymbol{Q}^T\boldsymbol{R} = \boldsymbol{U}^T\boldsymbol{V}\boldsymbol{U},\ \boldsymbol{B} = \boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^T = \begin{bmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{bmatrix},\ \text{where } \boldsymbol{V} = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then, as $\mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{V}\boldsymbol{B})$ and $\mathrm{tr}(\boldsymbol{Q}^T\boldsymbol{R}\boldsymbol{A}^2) = \mathrm{tr}(\boldsymbol{V}\boldsymbol{B}^2)$, it is easy to see that

$$\left.\frac{d^2}{dt^2}H(\boldsymbol{P}(t))\right|_{t=0} = \frac{\phi\sin\phi}{1-\cos\phi}(a^2+b^2) + 2c^2. \tag{3.10}$$

The right-hand side of (3.10) is always positive for arbitrary a, b, c in $\mathbb{R}$ and $\phi \in (-\pi, \pi)$. It follows that $\mathrm{Hess}\,H(\boldsymbol{\Delta}, \boldsymbol{\Delta})$ is positive for all tangent vectors $\boldsymbol{\Delta}$.

Now, let G denote the objective function of the minimization problem (3.4), i.e.,

$$G(\boldsymbol{R}) = \sum_{n=1}^{N} \|\,\mathrm{Log}(\boldsymbol{R}_n^T\boldsymbol{R})\|_F^2. \tag{3.11}$$

Using the above, the covariant derivative of G is found to be $\nabla G = \boldsymbol{R}\sum_{n=1}^{N}\mathrm{Log}(\boldsymbol{R}_n^T\boldsymbol{R})$. Therefore, a necessary condition for regular extrema of (3.11) is

$$\sum_{n=1}^{N}\mathrm{Log}(\boldsymbol{R}_n^T\boldsymbol{R}) = \boldsymbol{0}. \tag{3.12}$$

By (3.10) we conclude that the Hessian $\mathrm{Hess}\,G(\boldsymbol{\Delta}, \boldsymbol{\Delta})$ of the objective function (3.11) is positive for all tangent vectors $\boldsymbol{\Delta}$. Therefore, (3.12) characterizes local minima of (3.11) only. As a matter of fact, local maxima are not regular points, i.e., they are points where (3.11) is not differentiable.

It is worth noting that, as $\boldsymbol{R}_n^T = \boldsymbol{R}_n^{-1}$, the characterization for the Riemannian mean given in (3.12) is similar to the characterization

$$\sum_{n=1}^{N}\ln(x_n^{-1}x) = 0 \tag{3.13}$$

of the geometric mean (3.2) of positive numbers. However, while in the scalar case the characterization (3.13) has the geometric mean as unique solution, the characterization (3.12) has multiple solutions and hence is a necessary but not a sufficient condition for determining the Riemannian mean. The lack of uniqueness of solutions of (3.12) is akin to the fact that, due to the existence of a cut point for each element of $SO(3)$, the objective function (3.11) is not convex over its domain.

In general, closed-form solutions to (3.12) cannot be found. However, for some special cases solutions can be given explicitly. In the following subsections, we will present some of these special cases.

Remark 3.7. The Riemannian mean of $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ may also be called the *Riemannian barycenter* of $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$, which is a notion introduced by Grove, Karcher, and Ruh [11]. In [17] it was proven that for manifolds with negative sectional curvature, the Riemannian barycenter is unique.

3.2.1. Riemannian mean of two rotations. Intuitively, in the case $N = 2$, the mean rotation in the Riemannian sense should lie midway between $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ along the shortest geodesic curve connecting them, i.e., it should be the rotation $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$. Indeed, straightforward computation shows that $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$ does satisfy condition (3.12). Alternatively, (3.12) can be solved analytically as follows. First, we rewrite it as

$$\operatorname{Log}(\boldsymbol{R}_1^T\boldsymbol{R}) = -\operatorname{Log}(\boldsymbol{R}_2^T\boldsymbol{R}),$$

then we take the exponential of both sides to obtain $\boldsymbol{R}_1^T\boldsymbol{R} = \boldsymbol{R}^T\boldsymbol{R}_2$. After left multiplying both sides with $\boldsymbol{R}_1^T\boldsymbol{R}$ we get $(\boldsymbol{R}_1^T\boldsymbol{R})^2 = \boldsymbol{R}_1^T\boldsymbol{R}_2$. Such an equation has two solutions in $SO(3)$ that correspond to local minima of (3.11). However, the global minimum is the one that corresponds to taking the square root of the above equation that has eigenvalues with positive real part, i.e., $(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$. Therefore, for two nonantipodal rotation matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, the mean in the Riemannian sense is given explicitly by

$$\mathfrak{G}(\boldsymbol{R}_1, \boldsymbol{R}_2) = \boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2} = \boldsymbol{R}_2(\boldsymbol{R}_2^T\boldsymbol{R}_1)^{1/2}. \tag{3.14}$$

The second equality can be easily verified by premultiplying $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$ by $\boldsymbol{R}_2\boldsymbol{R}_2^T$, which is equal to $\boldsymbol{I}$. This makes it clear that $\mathfrak{G}$ is symmetric with respect to $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, i.e., $\mathfrak{G}(\boldsymbol{R}_1, \boldsymbol{R}_2) = \mathfrak{G}(\boldsymbol{R}_2, \boldsymbol{R}_1)$.

3.2.2. Riemannian mean of rotations in a one-parameter subgroup. In the case where all matrices $\boldsymbol{R}_n,\ n = 1, \ldots, N$, belong to a one-parameter subgroup of $SO(3)$, i.e., they represent rotations about a common axis, we expect that their mean is also in the same subgroup. Further, one can easily show that (3.12) reduces to saying that $\boldsymbol{R}$ is an Nth root of $\prod_{n=1}^N \boldsymbol{R}_n$. Therefore, the Riemannian mean is the Nth root that yields the minimum value of the objective function (3.11).

In this case, all rotations lie on a single geodesic curve. One can show that the geometric mean $\mathfrak{G}(\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_3)$ of three rotations $\boldsymbol{R}_1$, $\boldsymbol{R}_2$, and $\boldsymbol{R}_3$ such that $\mathrm{d}_R(\boldsymbol{R}_i, \boldsymbol{R}_j) < \pi,\ i, j = 1, 2, 3$, is the rotation that is located at $\frac{2}{3}$ of the length of the shortest geodesic segment connecting $\boldsymbol{R}_1$ and $\mathfrak{G}(\boldsymbol{R}_2, \boldsymbol{R}_3)$, i.e., the rotation $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2(\boldsymbol{R}_2^T\boldsymbol{R}_3)^{1/2})^{2/3}$. By induction, when $\mathrm{d}_R(\boldsymbol{R}_i, \boldsymbol{R}_j) < \pi,\ i, j = 1, \ldots, N$, we have

$$\mathfrak{G}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2(\boldsymbol{R}_2^T\boldsymbol{R}_3(\cdots\boldsymbol{R}_{N-1}(\boldsymbol{R}_{N-1}^T\boldsymbol{R}_N)^{\frac{1}{2}})^{\frac{2}{3}}\cdots)^{\frac{N-2}{N-1}})^{\frac{N-1}{N}}. \tag{3.15}$$

This explicit formula does not hold in the general case due to the inherent curvature of $SO(3)$; see the discussion at the end of Example 2 below.

When the rotations $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ belong to a geodesic segment of length less than π and centered at the identity, the above formula reduces to

$$\mathfrak{G}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \boldsymbol{R}_1^{1/N}\cdots\boldsymbol{R}_N^{1/N}. \tag{3.16}$$

Once again we see the close similarity between the geometric mean of positive numbers and the Riemannian mean of rotations. This is to be expected since both the set of positive numbers and $SO(3)$ are multiplicative groups, and we have used their intrinsic metrics to define the mean. For this reason, we will call the mean in the Riemannian sense the *geometric mean*.

3.3. Equivalence of both notions of mean of two rotations. In the following, we show that for two rotations the projected arithmetic mean and the geometric mean coincide. First, we prove the following lemma.

LEMMA 3.8. *Let $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ be two elements of $SO(3)$; then* $\det(\boldsymbol{R}_1 + \boldsymbol{R}_2) \geq 0$.

Proof. Consider the real-valued function defined on $[0, 1]$ by $f(t) = \det(\boldsymbol{R}_1 + t\boldsymbol{R}_2)$. We see that this function is continuous with $f(0) = 1$ and $f(1) = \det(\boldsymbol{R}_1 + \boldsymbol{R}_2)$. Assume that $f(1) < 0$, i.e., $\det(\boldsymbol{R}_1 + \boldsymbol{R}_2) < 0$; then there exists τ in $[0, 1]$ such that $f(\tau) = \det(\boldsymbol{R}_1 + \tau\boldsymbol{R}_2) = 0$. Since $\det \boldsymbol{R}_2 = 1$, it follows that $\det(\boldsymbol{R}_2^T\boldsymbol{R}_1 + \tau\boldsymbol{I}) = 0$. Hence, τ must be in the spectrum of $\boldsymbol{R}_2^T\boldsymbol{R}_1$, which is a proper orthogonal matrix. But this cannot happen, which contradicts the assumption that $\det(\boldsymbol{R}_1 + \boldsymbol{R}_2) < 0$. ∎

In general, the result of the above lemma does not hold for more than two rotations matrices. We will see examples of three rotation matrices for which the determinant of their sum can be negative.

PROPOSITION 3.9. *The polar factor of the polar decomposition of $\boldsymbol{R}_1 + \boldsymbol{R}_2$, where $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are two rotation matrices, is given by $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$.*

Proof. Let $\boldsymbol{Q}$ be the proper orthogonal matrix and $\boldsymbol{S}$ be the positive-definite matrix such that $\boldsymbol{QS}$ is the unique polar decomposition of $\boldsymbol{R}_1 + \boldsymbol{R}_2$. Then $\boldsymbol{S}^2 = (\boldsymbol{R}_1^T + \boldsymbol{R}_2^T)(\boldsymbol{R}_1 + \boldsymbol{R}_2) = 2\boldsymbol{I} + \boldsymbol{R}_1^T\boldsymbol{R}_2 + \boldsymbol{R}_2^T\boldsymbol{R}_1$. One can easily verify that $(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2} + (\boldsymbol{R}_1^T\boldsymbol{R}_2)^{-1/2}$ is the positive-definite square root of $2\boldsymbol{I} + \boldsymbol{R}_1^T\boldsymbol{R}_2 + \boldsymbol{R}_2^T\boldsymbol{R}_1$ and that the inverse of this square root is given by $\boldsymbol{S}^{-1} = (\boldsymbol{R}_1 + \boldsymbol{R}_2)^{-1}\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$. Hence, the polar factor is $\boldsymbol{Q} = (\boldsymbol{R}_1 + \boldsymbol{R}_2)\boldsymbol{S}^{-1} = \boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$. ∎

Since the polar decomposition is unique, the result of this proposition together with the previous lemma shows that both notions of mean agree for the case of two rotation matrices. For more than two rotations, however, both notions of mean coincide only in special cases that present certain symmetries. In Example 2 of section 4 below, we shall consider a two-parameter family of cases illustrating this coincidence.

4. Analytically solvable examples. In this section we present two cases in which we can solve for both the projected arithmetic mean and the geometric mean explicitly. These examples help us gain a deeper and concrete insight to both notions of mean. Furthermore, Example 2 confirms our intuitive idea that for "symmetric" cases, both notions of mean agree.

4.1. Example 1. We begin with a simple example where all rotation matrices for which we want to find the mean lie in a one-parameter subgroup of $SO(3)$. Using the bi-invariance property we can reduce the problem to that of finding the mean of

$$\boldsymbol{R}_n = \begin{bmatrix} \cos\theta_n & -\sin\theta_n & 0 \\ \sin\theta_n & \cos\theta_n & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad n = 1, \ldots, N. \tag{4.1}$$

Projected arithmetic mean. The arithmetic sum of these matrices has a positive determinant $r^2 = (\sum_{n=1}^N \cos\theta_n)^2 + (\sum_{n=1}^N \sin\theta_n)^2$. Hence, the projected arithmetic mean of the given matrices is given by the polar factor of the polar decomposition of

their sum. After performing such a decomposition we find that

$$\mathfrak{A}(\boldsymbol{R}_1,\dots,\boldsymbol{R}_N)=\begin{bmatrix}\cos\Theta_a & -\sin\Theta_a & 0\\ \sin\Theta_a & \cos\Theta_a & 0\\ 0 & 0 & 1\end{bmatrix},\quad \text{where}\quad \begin{cases}\cos\Theta_a=\dfrac{1}{r}\displaystyle\sum_{n=1}^{N}\cos\theta_n,\\[2ex] \sin\Theta_a=\dfrac{1}{r}\displaystyle\sum_{n=1}^{N}\sin\theta_n.\end{cases}$$

Such a mean is well defined as long as $r\neq 0$. This mean agrees with the notion of directional mean used in the statistics literature for circular and spherical data [20, 7, 9, 8]. The quantity $1-r/N$, which is called the circular variance, is a measure of dispersion of the circular data $\theta_1,\dots,\theta_N$. The direction defined by the angle Θ_a is called the mean direction of the directions defined by $\theta_1,\dots,\theta_N$.

Geometric mean. Solutions of (3.12) are given by

$$\begin{bmatrix}\cos\Theta_l & -\sin\Theta_l & 0\\ \sin\Theta_l & \cos\Theta_l & 0\\ 0 & 0 & 1\end{bmatrix},\quad \text{where}\quad \Theta_l=\frac{1}{N}\left(\sum_{n=1}^{N}\theta_n+2\pi l\right),\ l=0,\dots,N-1.$$

The geometric mean of these rotation matrices is therefore the solution that yields the minimum value of the objective function (3.11). Of course, as we have seen in section 3, the geometric mean is given explicitly by (3.15).

Note that even though elements of a one-parameter subgroup commute, the two rotations (3.15) and (3.16) are different. This is due to choosing the kth root of a rotation matrix to be the one with eigenvalues that have the largest positive real parts. To see this, consider the case $N=2$, $\theta_1=\frac{2\pi}{3}$, and $\theta_2=-\frac{2\pi}{3}$. Then $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}=\boldsymbol{P}$, where $\boldsymbol{P}$ is a rotation of an angle π about the z-axis while $\boldsymbol{R}_1^{1/2}\boldsymbol{R}_2^{1/2}=\boldsymbol{I}$.

If the rotation matrices $\boldsymbol{R}_n$ are such that $\alpha\leq\theta_n<\alpha+\pi,\ n=1,\dots,N$, for a certain number $\alpha\in\mathbb{R}$, then their geometric mean is a rotation about the z-axis of an angle $\Theta_g=\frac{1}{N}\sum_{n=1}^{N}\theta_n$.

The geometric mean rotation of the rotations given by (4.1) coincides with the concept of median direction of circular data [20, 7].

Remark 4.1. When $\theta_1=\theta$, $\theta_2=\theta+\pi$ and $N=2$ in (4.1), neither the projected arithmetic mean nor the geometric mean is well defined. On the one hand $\boldsymbol{R}_1+\boldsymbol{R}_2=\boldsymbol{0}$, so the projected arithmetic mean is not defined, while on the other hand the objective function (3.11) for the geometric mean has two local minima with the same value, namely, $\boldsymbol{R}_1(\boldsymbol{R}_1^T\boldsymbol{R}_2)^{1/2}$ and its cut value $\boldsymbol{R}_1\exp(\frac{\theta+\pi}{2\theta}\operatorname{Log}(\boldsymbol{R}_1^T\boldsymbol{R}_2))$, and therefore the global minimum is not unique.

Let $\tilde{F}$ and $\tilde{G}$ be the functions defined on $[-\pi,\pi]$ such that $\tilde{F}(\theta)=F(\boldsymbol{R})$ and $\tilde{G}(\theta)=G(\boldsymbol{R})$ for any rotation $\boldsymbol{R}$ about the z-axis through an angle θ, i.e., $\tilde{F}$ and $\tilde{G}$ are the restrictions of the objective functions (3.5) and (3.11) to the subgroup considered in this example. In Figure 4.1 we give the plots of $\tilde{F}$ and $\tilde{G}$ for the sets of data $N=4$ and $\theta_1=-\frac{\pi}{2}$, $\theta_2=0$, $\theta_3=\frac{\pi}{2}$, $\theta_4=\pi-\alpha$, where α takes several different values. It is clear that neither (3.5) nor (3.11) is convex. While the function (3.5) is smooth the function (3.11) has cusp points but only at local maxima. However, if the given rotations are located in a geodesic ball of radius less than $\pi/2$, i.e., in this example have angles θ_i such that $|\theta_i-\theta_j|<\pi, 1\leq i,j\leq N$, then the objective functions restricted to this geodesic ball are convex, and hence the means are well defined. Such a case is illustrated in Figure 4.2, which shows plots of $\tilde{F}$ and $\tilde{G}$ for the sets of data $N=3$ and $\theta_1=-\frac{\pi}{4}$, $\theta_2=\frac{\pi}{4}$, $\theta_3=\frac{3\pi}{4}-\alpha$, where α takes several different values.

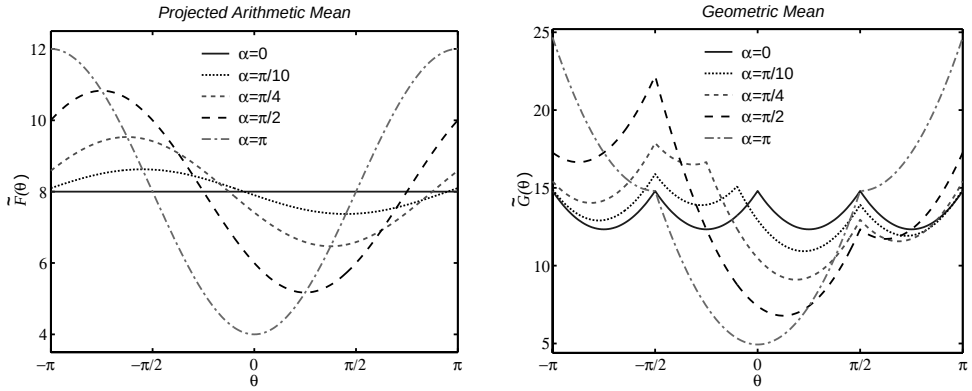


FIG. 4.1. *Plots of the objective functions $\tilde{F}(\theta)$ and $\tilde{G}(\theta)$ for different values of α. Note that when $\alpha = 0$, $\tilde{F}$ is constant and $\tilde{G}$ has four local minima with an equal value. Consequently, neither the projected arithmetic mean nor the geometric mean is well defined.*

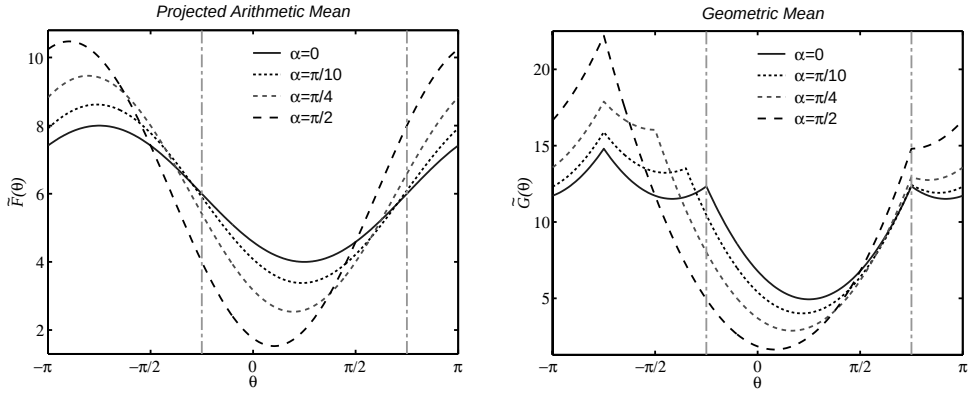


FIG. 4.2. *Plots of the objective functions $\tilde{F}(\theta)$ and $\tilde{G}(\theta)$ for different values of α. Restricted to $[-\pi/4, 3\pi/4]$, i.e., between the vertical dashed-dotted lines, the objective functions are indeed convex.*

4.2. Example 2. In the second example we consider N elements of $SO(3)$ that represent rotations through an angle θ about the axes defined by the unit vectors $\boldsymbol{u}_n = [\sin\alpha\cos\beta_n, \sin\alpha\sin\beta_n, \cos\alpha]^T$, where $\beta_n = \frac{2(n-1)\pi}{N}$, $n = 1, \ldots, N$, and $\alpha \in [0, \frac{\pi}{2})$.

Projected arithmetic mean. Straightforward computations show that the projected arithmetic mean is given by

$$\mathfrak{A}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \begin{bmatrix} \cos\Theta_a & -\sin\Theta_a & 0 \\ \sin\Theta_a & \cos\Theta_a & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{cases} \cos\Theta_a = \dfrac{2\cos\theta - \sin^2\alpha(\cos\theta - 1)}{2 + \sin^2\alpha(\cos\theta - 1)}, \\ \sin\Theta_a = \dfrac{2\cos\alpha\sin\theta}{2 + \sin^2\alpha(\cos\theta - 1)}. \end{cases}$$

By using half-angle tangent formulas in the above we obtain the following simple

relation between Θ_a and θ:

$$\tan\frac{\Theta_a}{2} = \cos\alpha\tan\frac{\theta}{2}. \tag{4.2}$$

Geometric mean. Since the rotation axes are symmetric about the z-axis, and the rotations share the same angle, we expect that their geometric mean is a rotation about the z-axis through a certain angle Θ_g. Furthermore, because of this symmetry we also expect that the mean in the Euclidean sense agrees with the one in the Riemannian sense.

From the Campbell–Baker–Hausdorff formula for elements of $SO(3)$ [23] we have

$$\operatorname{Log}(\boldsymbol{R}_n^T\boldsymbol{R}) = \phi\left(-a\operatorname{Log}\boldsymbol{R}_n + b\operatorname{Log}\boldsymbol{R} - c\left[\operatorname{Log}\boldsymbol{R}_n, \operatorname{Log}\boldsymbol{R}\right]\right),$$

where the coefficients a, b, c, and ϕ are given by

$$a\theta\sin\frac{\phi}{2} = \sin\frac{\theta}{2}\cos\frac{\Theta_g}{2}, \qquad b\Theta_g\sin\frac{\phi}{2} = \cos\frac{\theta}{2}\sin\frac{\Theta_g}{2},$$
$$c\theta\Theta_g\sin\frac{\phi}{2} = \sin\frac{\theta}{2}\sin\frac{\Theta_g}{2}, \qquad \cos\frac{\phi}{2} = \cos\frac{\theta}{2}\cos\frac{\Theta_g}{2} - \cos\alpha\sin\frac{\theta}{2}\sin\frac{\Theta_g}{2}.$$

Therefore, the characterization (3.12) of the geometric mean reduces to

$$a\sum_{n=1}^{N}\operatorname{Log}\boldsymbol{R}_n - bN\operatorname{Log}\boldsymbol{R} + c\sum_{n=1}^{N}\left[\operatorname{Log}\boldsymbol{R}_n, \operatorname{Log}\boldsymbol{R}\right] = \boldsymbol{0}.$$

This is a matrix equation in $\mathfrak{so}(3)$, which is equivalent to a system of three nonlinear equations. Because the axes of rotation of $\boldsymbol{R}_n$ are symmetric about the z-axis we have $\sum_{n=1}^{N}\cos\beta_n = \sum_{n=1}^{N}\sin\beta_n = 0$. It follows that $\sum_{n=1}^{N}\left[\operatorname{Log}\boldsymbol{R}_n, \operatorname{Log}\boldsymbol{R}\right] = \boldsymbol{0}$ and $\Theta_g\sum_{n=1}^{N}\operatorname{Log}\boldsymbol{R}_n = \theta\cos\alpha N\operatorname{Log}\boldsymbol{R}$. Therefore, this system reduces to the following single equation for the angle Θ_g:

$$\tan\frac{\Theta_g}{2} = \cos\alpha\tan\frac{\theta}{2}, \tag{4.3}$$

which when compared with (4.2) indeed shows that $\Theta_a = \Theta_g$ and therefore the projected arithmetic mean and the geometric mean coincide.

This example provides a family of mean problems parameterized by θ and α where the projected arithmetic and geometric mean coincide. We now further examine the problem of finding the mean of three rotations about the three coordinate axes through the same angle θ, which, by the bi-invariance property of both means, can be considered as a special case of this two-parameter family with $N = 3$ and $\cos\alpha = \frac{1}{\sqrt{3}}$. Therefore the mean of these three rotations is a rotation through an angle Φ about the axis generated by the vector $[1, 1, 1]^T$ with $\tan\frac{\theta}{2} = \sqrt{3}\tan\frac{\Phi}{2}$. The rotations $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ form a geodesic equilateral triangle in $SO(3)$. By symmetry arguments the geometric mean should be the intersection of the three geodesic medians, i.e., the geodesic segments joining the vertices of the geodesic triangle to the midpoints of the opposite sides. In flat geometry, this intersection is located at two-thirds from the vertices of the triangle. However, in the case of $SO(3)$, due to its intrinsic curvature, this is not true. The ratio γ of the length of the geodesic segment joining one rotation and the geometric mean to the length of the geodesic median joining this rotation and the midpoint of the geodesic curve joining the two other rotations is plotted as a function of the angles θ in Figure 4.3.

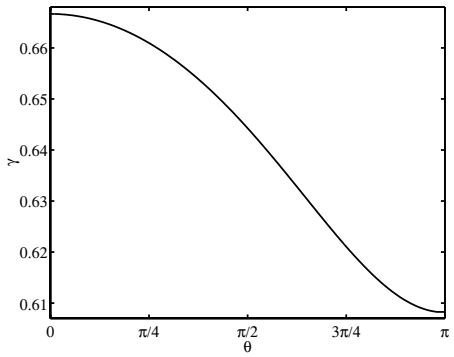


FIG. 4.3. *Plot of the ratio γ of the geodesic distance from one vertex to the barycenter over the geodesic distance from this vertex to the midpoint of the opposed edge in the geodesic equilateral triangle in $SO(3)$. The departure of γ from $2/3$, which is due to the curvature of $SO(3)$, increases with the length, θ, of the sides of the triangle.*

5. Weighted means and power means. Our motivation for this work was to construct a filter that smooths the rotation data giving the relative orientations of successive base pairs in a DNA fragment; see [19] for details. Such a filter can be a generalization of moving window filters, which are based on weighted averages, used in linear spaces to smooth noisy data. The construction of such filters and the direct analogy we have found between the arithmetic and geometric means in the group of positive numbers, and the projected arithmetic and geometric means in the group of rotations, have led us to the introduction of weighted means and power means of rotations that we discuss next.

DEFINITION 5.1. *The weighted projected arithmetic mean of N given rotations $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ with weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ is defined as*

$$\mathfrak{A}_w(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N; \boldsymbol{w}) := \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \sum_{n=1}^{N} w_n \|\boldsymbol{R} - \boldsymbol{R}_n\|_F^2. \tag{5.1}$$

This mean satisfies the bi-invariance property. Using similar arguments as for the projected arithmetic mean one can show that the weighted projected arithmetic mean is given by the polar factor of the polar decomposition of the matrix $\boldsymbol{A} = \sum_{n=1}^{N} w_n \boldsymbol{R}_n$ provided that $\det \boldsymbol{A}$ is positive.

DEFINITION 5.2. *The weighted geometric mean of N rotations $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ with weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ is defined as*

$$\mathfrak{G}_w(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N; \boldsymbol{w}) := \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \sum_{n=1}^{N} w_n \|\operatorname{Log}(\boldsymbol{R}^T \boldsymbol{R}_n)\|_F^2. \tag{5.2}$$

This mean also satisfies the bi-invariance property. Using arguments similar to those used for the geometric mean, we can show that the weighted geometric mean is characterized by $\sum_{n=1}^{N} w_n \operatorname{Log}(\boldsymbol{R}_n^T \boldsymbol{R}) = \boldsymbol{0}$.

DEFINITION 5.3. *For a real number s such that $0 < |s| \leq 1$, we define the weighted sth power mean rotation of N rotations $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_N$ with weights $\boldsymbol{w} =$*

$(w_1, \ldots, w_N)$ *as*

$$\mathfrak{M}_w^{[s]}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N; \boldsymbol{w}) := \operatorname*{arg\,min}_{\boldsymbol{R} \in SO(3)} \sum_{n=1}^{N} w_n \|\boldsymbol{R}^s - \boldsymbol{R}_n^s\|_F^2. \tag{5.3}$$

We note that $[\mathfrak{M}_w^{[s]}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N; \boldsymbol{w})]^s = \mathfrak{A}_w(\boldsymbol{R}_1^s, \ldots, \boldsymbol{R}_N^s; \boldsymbol{w})$. Of course for $s = 1$ this is the weighted projected arithmetic mean. Because elements of $SO(3)$ are orthogonal, and the trace operation is invariant under transposition, the weighted sth power mean is the same as the weighted $(-s)$th power mean. Therefore, it is immediate that the *weighted projected harmonic mean*, defined by

$$\mathfrak{H}_w(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N; \boldsymbol{w}) = \left[\mathfrak{M}_w(\boldsymbol{R}_1^{-1}, \ldots, \boldsymbol{R}_N^{-1}; \boldsymbol{w})\right]^{-1},$$

coincides with the weighted projected arithmetic mean.

This is a natural generalization of the sth power mean of positive numbers, and it is in line with the fact that for positive numbers $(x_1, \ldots, x_N)$ the sth power mean is given by the sth root of the arithmetic mean of $(x_1^s, \ldots, x_N^s)$ [12, 2]. One has to note, however, that for s such that $0 < |s| < 1$ this mean is not invariant under the action of elements of $SO(3)$. This is not a surprise, as the power mean of positive numbers also does not satisfy the homogeneity property.

For the set of positive numbers [12] and similarly for the set of Hermitian definite positive operators [25], there is a natural ordering of elements and the classical arithmetic-geometric-harmonic mean inequalities holds. Furthermore, it is well known [12, 25] that the sth power mean converges to the geometric mean as s goes to 0. However, for the group of rotations such a natural ordering does not exist. Nonetheless, one can show that if all rotations $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ belong to a geodesic ball of radius less than $\frac{\pi}{2}$ centered at the identity, then the projected power mean indeed converges to the geometric mean as s tends to 0.

Analysis of numerical algorithms for computing the geometric mean rotation and the use of the different notions of mean rotation for smoothing three-dimensional orientation data is forthcoming.

Acknowledgments. The author is grateful to Professor J. H. Maddocks for suggesting this problem and for his valuable comments on this paper. He also thanks the anonymous referee for his helpful comments.

REFERENCES

[1] M. Berger and B. Gostiaux, *Differential Geometry: Manifolds, Curves, and Surfaces*, Springer-Verlag, New York, 1988.
[2] P. S. Bullen, D. S. Mitrinović, and P. M. Vasić, *Means and Their Inequalities*, Math. Appl. (East European Ser.) 31, D. Reidel, Dordrecht, The Netherlands, 1988.
[3] M. L. Curtis, *Matrix Groups*, Springer-Verlag, New York, Heidelberg, 1979.
[4] T. D. Downs, *Orientation statistics*, Biometrika, 59 (1972), pp. 665–676.
[5] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.
[6] D. W. Eggert, A. Lorusso, and R. B. Fisher, *Estimating 3-D rigid body transformations: A comparison of four major algorithms*, Machine Vision Appl., 9 (1997), pp. 272–290.
[7] N. I. Fisher, *Spherical medians*, J. Roy. Statist. Soc. Ser. B, 47 (1985), pp. 342–348.
[8] N. I. Fisher, *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge, UK, 1993.
[9] N. I. Fisher, T. Lewis, and B. J. J. Embleton, *Statistical Analysis of Spherical Data*, Cambridge University Press, Cambridge, UK, 1987.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, London, 1989.
[11] K. Grove, H. Karcher, and E. A. Ruh, *Jacobi fields and Finsler metrics on compact Lie groups with an application to differentiable pinching problem*, Math. Ann., 211 (1974), pp. 7–21.
[12] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934.
[13] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[14] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numer., 9 (2000), pp. 215–365.
[15] P. E. Jupp and K. V. Mardia, *Maximum likelihood estimation for the matrix von Mises-Fisher and Bingham distributions*, Ann. Statist., 7 (1979), pp. 599–606.
[16] P. E. Jupp and J. T. Kent, *Fitting smooth paths to spherical data*, Appl. Statist., 36 (1987), pp. 34–46.
[17] H. Karcher, *Riemannian center of mass and mollifier smoothing*, Comm. Pure Appl. Math., 30 (1977), pp. 509–541.
[18] C. G. Khatri and K. V. Mardia, *The von Mises-Fisher matrix distribution in orientation statistics*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 95–106.
[19] R. S. Manning, J. H. Maddocks, and J. D. Kahn, *A continuum rod model of sequence-dependent DNA structure*, J. Chem. Phys., 105 (1996), pp. 5626–5646.
[20] K. V. Mardia, *Statistics of directional data*, Prob. Math. Statist. 13, Academic Press, London, New York, 1972.
[21] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1996.
[22] M. J. Prentice, *Fitting smooth paths to rotation data*, Appl. Statist., 36 (1987), pp. 325–331.
[23] J. M. Selig, *Geometrical Methods in Robotics*, Springer-Verlag, New York, 1996.
[24] S. T. Smith, *Optimization techniques on Riemannian manifolds*, in Hamiltonian and Gradient Flows, Algorithms and Control, A. Bloch, ed., AMS, Providence, RI, 1994, pp. 113–136.
[25] G. E. Trapp, *Hermitian semidefinite matrix means and related matrix inequalities—an introduction*, Linear and Multilinear Algebra, 16 (1984), pp. 113–123.
[26] C. Udrişte, *Convex Functions and Optimization Methods on Riemannian Manifolds*, Math. Appl. 297, Kluwer Academic, Dordrecht, The Netherlands, 1994.
[27] G. S. Watson, *Equatorial distributions on a sphere*, Biometrika, 52 (1965), pp. 193–201.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 17–24

ON THE ITERATIVE CRITERION FOR GENERALIZED DIAGONALLY DOMINANT MATRICES*

LEI LI†

Abstract. An iterative method for identifying generalized diagonally dominant matrices (GDDMs, or H-matrices) was given in [B. Li et al., *Linear Algebra Appl.*, 271 (1998), pp. 179–190], where the method is divergent when the matrix is not a GDDM. In this paper, we present an improved version. The new method is always convergent and needs fewer iterations than the earlier one. Some interesting features of the new method are presented. Spectral radii of nonnegative matrices with a constant diagonal entry also can be computed by our method.

Key words. H-matrix, diagonally dominant matrix, iteration, criterion

AMS subject classifications. 15A15, 15A09, 15A23

PII. S0895479898348829

1. Introduction.

Let $A = (a_{ij})$ be an $n \times n$ complex matrix, $N = \{1, 2, \ldots, n\}$, and $N_1(A) = \{i| \quad |a_{ii}| > \sum_{j \neq i} |a_{ij}| = S_i, i \in N\} \neq \Phi$. A matrix A is called a strictly diagonally dominant matrix if $N_1 = N$ and a *generalized diagonally dominant matrix* (GDDM) if there exists a positive diagonal matrix D such that AD is strictly diagonally dominant.

GDDM is a special class of matrices with wide applications in engineering and scientific computation [2]. Iterative algorithms for solving systems of linear equations with generalized diagonally dominant coefficient matrices have been studied; see, e.g., [3] for serial iterations and [4] for asynchronous parallel iterations with arbitrary splitting form. The extension to some nonlinear cases was studied in [5]. Another special class of matrices is the M-matrix. It has been proved that when the coefficient matrix of a linear system is an M-matrix, many iterative algorithms are convergent, e.g., multiple splitting methods [9] and some recent methods for solving elliptic PDEs with domain decomposition and Schwarz's relaxation [10]. A matrix $A = (a_{ij})$ is an M-matrix if $a_{ij} \leq 0$ for $i \neq j$ and $a_{ii} > 0$ and A is a GDDM. Then the identification of GDDMs plays an important role in analyzing the convergence of iterative algorithms.

The problem that interests us is how to identify the generalized diagonally dominant for a general matrix, particularly in a large scale. A matrix A is a GDDM if and only if $m(A)$ is an M-matrix, where $m(A)$ is the comparison matrix of A. More than forty equivalent conditions for the M-matrix have been given in [2]. All of these conditions are difficult for practical purposes. Several direct algorithms for identifying GDDMs were given in [3, 6, 7, 11]. However, those algorithms are successful only for some special cases. No direct algorithms have been explored for a general matrix. Recently, an iterative criterion for the identification of GDDMs was presented in [1, 12]. It has been proved that this iterative criterion is successful (convergent) only if the matrix is a GDDM. However, the method in [1, 12] is divergent when A is not a GDDM. The method has been applied in many engineering computations [13, 14].

*Received by the editors November 25, 1998; accepted for publication (in revised form) by P. Van Dooren January 29, 2002; published electronically May 15, 2002. This work was supported by Grant-in-Aid for Scientific Research contract 10680357 from the Japan Society for the Promotion of Science.

http://www.siam.org/journals/simax/24-1/34882.html

†Faculty of Engineering, Hosei University, Koganei, Tokyo 184-8584, Japan (lilei@k.hosei.ac.jp).

In this paper, we present an improved iterative method. The new method is always convergent in finite iterative steps for general matrices, and the number of iterations in the new method is less than that of the method in [1]. Also we present some interesting features of this method, which enable it to compute the spectral radii of nonnegative matrices with a constant diagonal entry.

2. An improved iterative criterion. An iterative criterion for general GDDMs is given as follows.

ALGORITHM 1.

For a given complex matrix $A = (a_{ij})_{n\times n}$, $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$:

1. Compute $S_i = \sum_{j\neq i} |a_{ij}|$, $i = 1, 2, \ldots, n$.
2. Set $t = 0$. For $i = 1, 2, \ldots, n$, if $|a_{ii}| > S_i$, then set $t = t + 1$.
3. If $t = 0$, then print "A is not a GDDM" : END.
4. If $t = n$, then print "A is a GDDM" : END.
5. For $i = 1, 2, \ldots, n$, compute

$$d_i = \frac{S_i + \varepsilon}{|a_{ii}| + \varepsilon}, \quad a_{ji} = a_{ji} \cdot d_i, \qquad j = 1, 2, \ldots, n,$$

where $\varepsilon > 0$ is a positive parameter.

6. Go to step 1.

Remark. Algorithm 1 generates two sequences of matrices, $\{A^{(i)}\}$ and $\{D^{(i)}\}$, where $A^{(i)}$ and $D^{(i)}$ denote the matrices A and D in the ith iteration, respectively. When $A^{(0)} = A$ is a GDDM, the sequence $\{A^{(i)}\}$ obtained in [1] satisfies the following monotonicity:

$$N_1(A^{(0)}) \subseteq N_1(A^{(1)}) \subseteq \cdots \subseteq N_1(A^{(k)}) \subseteq \cdots \to N,$$

where $N_1 = \{i||a_{ii}| > S_i, i = 1, 2, \ldots, n\}$. But it is not true for our algorithm.

In Algorithm 1, ε_0 is a small positive parameter to be determined by users. Usually, the number of iterations is a function of the parameter ε. Our numerical investigations illustrate that the number of iterations has its minimum at a small ε_0 and is a constant when $\varepsilon \leq \varepsilon_0$. However, it is difficult to find ε_0 in general. In fact, one only needs to find the smallest one (ε_{min}) that can be discriminated by computer.

Let δ denote the accuracy of the computer and let

$$\frac{S_i + \varepsilon_{min}}{|a_{ii}| + \varepsilon_{min}} \cdot |a_{ii}| - S_i > \delta, \qquad i \in N_1\,.$$

Then

$$\varepsilon_{min} > \frac{\delta|a_{ii}|}{|a_{ii}| - S_i - \delta}, \qquad i \in N_1\,.$$

So, we can take

$$\varepsilon_{min} = \min\left\{\frac{\delta(|a_{ii}| + 1)}{|a_{ii} - S_i|}, \quad i \in N_1\right\}.$$

Theoretically, one can choose ε based on the above formula in each iteration step. For simplicity, the above ε_{min} is used until the iteration stops. It was shown in [2] that one can choose $\varepsilon = 0$ when A is an irreducible matrix.

3. Theoretical analysis of algorithm. In this section we shall present some theoretical analysis.

THEOREM 1. *If Algorithm* 1 *stops in finite iterative steps, then its output is correct.*

Proof. There are only two possible outputs in Algorithm 1: "A is not a GDDM" and "A is a GDDM." We consider these two cases, respectively, as follows.

(i) If "A is not a GDDM" is the output in the kth iterative step, we need to prove that both $A^{(k)}$ and A are not GDDMs. By Algorithm 1, in this case, $t = 0$ and $N_1(A^{(k)}) = \Phi$, where

$$A^{(k)} = A^{(0)} \cdot D^{(0)} \cdot D^{(1)} \cdots D^{(k-1)} = A \cdot D$$

and $D = D^{(0)} \cdot D^{(1)} \cdots D^{(k-1)}$ is a positive diagonal matrix.

If $A^{(k)}$ is a GDDM, then there exists a positive diagonal matrix E such that $A^{(k)}E$ is a strictly diagonally dominant matrix. Let

$$e_i = \min\{e_1, e_2, \ldots, e_n\},$$

where e_j, $j = 1, 2, \ldots, n$, are the diagonal entries of E. We have

$$|a_{ii}^{(k)}|e_i > \sum_{j \neq i} |a_{ij}^{(k)}|e_j,$$

$$|a_{ii}^{(k)}| > \sum_{j \neq i} |a_{ij}^{(k)}|\frac{e_j}{e_i} \geq \sum_{j \neq i} |a_{ij}^{(k)}|,$$

and therefore $i \in N_1(A^{(k)})$. This contradicts $N_1(A^{(k)}) = \Phi$.

Similarly, if A is a GDDM, there exists a positive diagonal matrix F such that AF is a strictly diagonally dominant matrix. Since $AF = A^{(k)} \cdot D^{-1} \cdot F$ and $D^{-1}F$ is also a positive diagonal matrix, $A^{(k)}$ is a GDDM, which results in a contradiction. Thus, the output "A is not a GDDM" is correct.

(ii) If "A is a GDDM" is the output in the kth iterative step, then $t = n$, $N_1(A^{(k)}) = n$ and

$$A^{(k)} = A^{(0)} \cdot D^{(0)} \cdot D^{(1)} \cdots D^{(k-1)} = AD$$

is strictly diagonally dominant. Since $D = D^{(0)}D^{(1)} \cdots D^{(k-1)}$ is a positive diagonal matrix, A is a GDDM. The theorem is proved. □

THEOREM 2. *Let*

$$D^{(k)} = \operatorname{diag}\{d_1^{(k)}, \ldots, d_n^{(k)}\}$$

be the positive diagonal matrix in the k*th iterative step of Algorithm* 1*. Then*

$$\lim_{\varepsilon \to 0} \lim_{k \to \infty} D^{(k)} = uI, \tag{1}$$

where u *is a positive constant and* I *is the identity matrix.*

Proof. (i) Let $B^{(k)} = (b_{ij}^{(k)})$ be the Jacobi iterative matrix of the comparison matrix $m(A^{(k)})$, i.e.,

$$B^{(k)} := I - D(m(A^{(k)}))^{-1}m(A^{(k)}),$$

where $D(m(A^{(k)}))$ is a diagonal matrix in which the diagonal entries are the same as $m(A^{(k)})$. By the definition of $A^{(k)}$, we have

$$\begin{aligned} B^{(k)} &= I - D(m(A^{(k)}))^{-1} m(A^{(k)}) \\ &= I - [D(m(A^{(k-1)})) \cdot D^{(k-1)}]^{-1} \cdot [m(A^{(k-1)} \cdot D^{(k-1)})] \\ &= I - [D^{(k-1)}]^{-1} \cdot D(m(A^{(k-1)}))^{-1} \cdot m(A^{(k-1)}) \cdot D^{(k-1)} \\ &= [D^{(k-1)}]^{-1} [I - D(m(A^{(k-1)}))^{-1} \cdot m(A^{(k-1)})] \cdot D^{(k-1)} \\ &= [D^{(k-1)}]^{-1} B^{(k-1)} D^{(k-1)} \end{aligned}$$

and therefore

$$\lambda(B^{(k)}) = \lambda(B^{(k-1)}) = \cdots = \lambda(B^{(1)}) = \lambda(B^{(0)}).$$

(ii) We denote by $S_i(B^{(k)})$ the sum of entries in the ith row of $B^{(k)}$. Then we have

$$S_i(B^{(k)}) = d_i^{(k)}, \qquad i = 1, 2, \ldots, n, \tag{2}$$

by noting the fact that $B^{(k)}$ is nonnegative. Since $d_i^{(k)}$ is a continuous function of ε, we can consider only the case of $\varepsilon = 0$ to prove (1). It follows from Algorithm 1 that

$$\begin{aligned} d_i^{(k)} &= \frac{\sum_{j \neq i} |a_{ij}^{(k)}|}{|a_{ii}^{(k)}|} \\ &= \frac{\sum_{j \neq i} |a_{ij}^{(k-1)}| \cdot d_j^{(k-1)}}{|a_{ii}^{(k-1)}| \cdot d_i^{(k-1)}} \\ &\leq \frac{\sum_{j \neq i} |a_{ij}^{(k-1)}|}{|a_{ii}^{(k-1)}|} \cdot \frac{d_{max}^{(k-1)}}{d_i^{(k-1)}} \\ &= \frac{d_i^{(k-1)} \cdot d_{max}^{(k-1)}}{d_i^{(k-1)}} = d_{max}^{(k-1)} \end{aligned}$$

and

$$d_i^{(k)} = \frac{\sum_{j \neq i} |a_{ij}^{(k-1)}| d_j^{(k-1)}}{|a_{ii}^{(k-1)}| \cdot d_i^{(k-1)}} \geq \frac{\sum_{j \neq i} |a_{ij}^{(k-1)}| d_{min}^{(k-1)}}{|a_{ii}^{(k-1)}| \cdot d_i^{(k-1)}} = d_{min}^{(k-1)},$$

where $d_{max}^{(k-1)} := \max_j \{d_j^{(k-1)}\}$ and $d_{min}^{(k-1)} := \min_j \{d_j^{(k-1)}\}$. Let

$$L^{(k)} = d_{max}^{(k)} - d_{min}^{(k)}.$$

It follows that

$$L^{(k)} = d_{max}^{(k)} - d_{min}^{(k)} \leq d_{max}^{(k-1)} - d_{min}^{(k-1)} = L^{(k-1)}.$$

It can be seen that $L^{(k)} \geq 0$, $k = 1, 2, \ldots$, is decreasing monotonically. Then there exists L such that

$$L = \lim_{k \to \infty} L^{(k)}.$$

If $L > 0$, there exist a constant α and an integer $M > 0$ such that for $k \geq M$,

$$d_{max}^{(k)} - d_{min}^{(k)} > \alpha,$$

or

$$d_{max}^{(k)} > d_{min}^{(k)} + \alpha\,.$$

For convenience, we assume

$$d_{min}^{(k)} = d_1^{(k)} \leq d_2^{(k)} \leq \cdots \leq d_n^{(k)} = d_{max}^{(k)}\,.$$

Then

$$B^{(k+1)} = [D^{(k)}]^{-1} B^{(k)} \cdot D^{(k)}$$

and

$$b_{ij}^{(k+1)} = b_{ij}^{(k)} \cdot \frac{d_j^{(k)}}{d_i^{(k)}}\,.$$

In this case, if $i > j$, we have $b_{ij}^{(k+1)} \geq b_{ij}^{(k)}$, and if $i < j$, $b_{ij}^{(k+1)} \leq b_{ij}^{(k)}$; i.e., the upper triangle of $B^{(k+1)}$ is increasing monotonically, and the lower triangle is decreasing. Since

$$b_{1n}^{(k+1)} = b_{1n}^{(k)} \cdot \frac{d_n}{d_1} > b_{1n}^{(k)} \left(\frac{d_1 + \alpha}{d_1} \right) = b_{1n}^{(k)}(1 + \beta),$$

and $b_{ii}^{(k+1)} = b_{ii}^{(k)}$, where $\beta = \frac{\alpha}{d_1} > 0$, we have

$$d_{max}^{(k)} \geq S_1(B^{(k)}) \geq b_{1n}^{(k)} \to \infty \qquad \text{as } k \to \infty$$

and

$$d_{min}^{(k)} \leq S_n(B^{(k)}) \leq S_n(B^{(0)}),$$

which contradicts $|d_{max}^{(k)} - d_{min}^{(k)}|$ being convergent. The proof is complete. □

THEOREM 3. *Let u be the positive number defined in Theorem 2. Then*

$$u = \rho(B), \tag{3}$$

where B is the Jacobi iterative matrix of $m(A)$ and $\rho(B)$ is the spectral radius of B.

Proof. From the proof of Theorem 2, we have

$$\rho(B^{(0)}) = \rho(B^{(1)}) = \cdots = \rho(B^{(k)})\,.$$

Let

$$B^* = \lim_{k \to \infty} B^{(k)}\,.$$

By (1) and (2),

$$S_i(B^{(k)}) \to S_i(B^*) = u \qquad \text{as } k \to \infty$$

and

$$B^*e = ue,$$

where $e = (1, 1, \ldots, 1)^T$. Then u is an eigenvalue of B^* and e is the corresponding eigenvector. It follows that

$$\|B^*\|_\infty = u$$

by noting the fact that B^* is nonnegative. Equation (3) is obtained immediately. ∎

THEOREM 4. *For any given $n \times n$ matrix $A = (a_{ij})$, $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$, Algorithm 1 always stops in finite iterative steps.*

Proof. It is known that A is a GDDM if and only if $m(A)$ is an M-matrix, and $m(A)$ is an M-matrix if and only if the Jacobi iterative matrix $B = I - D(m(A))^{-1}m(A)$ of $m(A)$ is convergent, i.e., $\rho(B) < 1$.

From the proof of Theorem 2, we have

$$\rho(B) = \rho(B^{(k)}), \qquad k = 1, 2, \ldots,$$

and from the proof of Theorem 3,

$$\lim_{k\to\infty} d_i^{(k)} = u = \rho(B).$$

If $u < 1$, there exists an integer k such that $t = n$ (or $t \in N$) in the kth iteration; if $u \geq 1$, $m(A^{(k)})$ is not an M-matrix and there exists an integer k satisfying

$$d_i^{(k)} = \frac{S_i^{(k)}}{|a_{ii}^{(k)}|} \geq 1, \qquad i = 1, 2, \ldots, n\,,$$

i.e., $t = 0$ in the kth iteration. The proof is complete. ∎

4. Numerical examples.

Example 1. First we consider the matrix

$$A = \begin{pmatrix} 1 & -0.2 & -0.1 & -0.2 & -0.1 \\ -0.4 & 1 & -0.2 & -0.1 & -0.1 \\ -0.9 & -0.2 & 1 & -0.1 & -0.1 \\ -0.3 & -0.7 & -0.3 & 1 & -0.1 \\ -1 & -0.3 & -0.2 & -0.4 & 1 \end{pmatrix}$$

by using both the method in [1] and Algorithm 1. 13 iterations are needed when the method in [1] with $\epsilon = 0.001$ is used and only 3 iterations are needed for Algorithm 1 with $\epsilon = 0.1$.

Example 2. The second example is

$$A = \begin{pmatrix} 1 & -0.8 & -0.1 \\ -0.5 & 1 & c \\ -0.8 & -0.6 & 1 \end{pmatrix}$$

with a parameter c. It is easy to show that A is not a GDDM when $|c| > 0.3951$ and is a GDDM when $|c| \leq 0.3951$. Algorithm 1 is always successful for any c, while the method in [1] is divergent for $|c| > 0.3951$.

Example 3. Finally, we consider the Jacobi matrix $B = I - D(m(A))^{-1}m(A)$ where A is defined in Example 1. A simple calculation gives

$$B = \begin{pmatrix} 0 & 0.2 & 0.1 & 0.2 & 0.1 \\ 0.4 & 0 & 0.2 & 0.1 & 0.1 \\ 0.9 & 0.2 & 0 & 0.1 & 0.1 \\ 0.3 & 0.7 & 0.3 & 0 & 0.1 \\ 1 & 0.3 & 0.2 & 0.4 & 0 \end{pmatrix}.$$

We replace the convergence conditions $t = 0$ and $t = n$ in Algorithm 1 with

$$|d_i^{(k)} - d_j^{(k)}| < 0.0001 \qquad \text{for} \quad i \neq j\,.$$

Numerical calculation with 12 iterations gives

$$\rho(B) = d_i^{(k)} = 0.991878\cdots, \qquad i = 1, 2, \ldots, n\,,$$

which confirms our theoretical analysis in Theorem 3 and shows that the method can be used for computing the spectral radius of a nonnegative matrix with a constant diagonal entry.

5. Conclusions. We have presented a new iterative method for identifying the generalized diagonal dominant of a matrix and showed some interesting properties of the method. Comparing this method to that of [1], we find that the new one needs fewer iterations and is applicable to any matrices. The new method also can be used for computing the spectral radius of the Jacobi matrix of an M-matrix. The Gauss version of the new method remains for future research.

Acknowledgment. The author thanks Prof. Weiwei Sun, City University of Hong Kong, for his suggestions and comments on the paper.

REFERENCES

[1] B. Li, L. Li, M. Harada, H. Niki, and M.J. Tsatsomeros, *An iterative criterion for H-matrices*, Linear Algebra Appl., 271 (1998), pp. 179–190.
[2] A. Berman and R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.
[3] L. Li, *On some practical criteria for the convergence of splitting iterations*, Acta Math. Appl. Sinica, 17 (1994), pp. 429–436.
[4] L. Li, *Convergence of asynchronous iteration with arbitrary splitting form*, Linear Algebra Appl., 113 (1989), pp. 119–127.
[5] L. Li and T. Nakamura, *The convergence of asynchronous iterations for the fixed point of a splitting operator*, Parallel Algorithms Appl., 7 (1995), pp. 229–235.
[6] Y. Gao and X. Wang, *Criteria for generalized diagonally dominant matrices and M-matrices*, Linear Algebra Appl., 169 (1992), pp. 257–268.
[7] Y. Gao and X. Wang, *Criteria for generalized diagonally dominant matrices and M-matrices* II, Linear Algebra Appl., 248 (1996), pp. 339–353.
[8] L. Li, *On some practical criteria for positive definite matrices*, Trans. Japan SIAM, 7 (1997), pp. 91–96.
[9] A. Frommer and B. Pohl, *Comparison results for splittings based on overlapping blocks*, in Proceedings of the 5th SIAM Conference on Applied Linear Algebra, J. Lewis, ed., SIAM, Philadelphia, 1994, pp. 29–33.
[10] L. Kang, L. Sun, and Y. Chen, *Asynchronous Parallel Algorithms for Solving Mathematics and Physics Problems*, Science Press, Beijing, 1985.
[11] M. Harada, M. Usui, and H. Niki, *An extension of the criteria for generalized diagonally dominant matrices*, Int. J. Comput. Math., 60 (1996), pp. 115–119.

[12] L. Li, *On criteria for generalized diagonally dominant matrices*, in Proceedings of the 25th Japan Numerical Analysis Symposium, Japan Society for Industrial and Applied Mathematics, Tokyo, 1996, pp. 72–74.
[13] L. Li, *Gauss type criterion of GDDM and its applications*, Acta Math. Appl. Sinica, 21 (1998), pp. 155–158.
[14] L. Li, *Some applications of the criteria for generalized diagonally dominant matrices*, in Advance in Computational Engineering Science, S.N. Atluri and G. Yagawa, eds., Tech. Science Press, 1997, pp. 718–723.
[15] W. Li and W. Sun, *Comparison results for parallel multisplitting methods*, Linear Algebra Appl., 331 (2001), pp. 131–144.
[16] W. Li, W. Sun, and K.M. Liu, *Parallel multisplitting iterative methods for singular M-matrices*, Numer. Linear Algebra Appl., 8 (2001), pp. 181–190.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 25–39

SOME NEW RESULTS FOR THE SEMIDEFINITE LINEAR COMPLEMENTARITY PROBLEM*

M. SEETHARAMA GOWDA† AND Y. SONG†

Abstract. In this paper, we present some new results for the semidefinite linear complementarity problem (SDLCP). In the first part, we introduce the concepts of (i) nondegeneracy for a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ and (ii) the locally-star-like property of a solution point of an SDLCP(L, Q) for $Q \in \mathcal{S}^n$, and we relate them to the finiteness of the solution set of SDLCP(L, Q) as Q varies in $\mathcal{S}^n$. In the second part, we show that for positive stable matrices $A_1, \ldots, A_k$, the linear transformation $L := L_{A_1} \circ L_{A_2} \circ \cdots \circ L_{A_k}$ has the **Q**-property where $L_{A_i}(X) := A_i X + X A_i^T$. A similar result is proved for the transformation $S := S_{A_1} \circ S_{A_2} \circ \cdots \circ S_{A_k}$, where each A_i is Schur stable and $S_{A_i}(X) := X - A_i X A_i^T$. We relate these results to the simultaneous stability of a finite set of matrices.

Key words. nondegenerate, locally-star-like, semidefinite linear complementarity problem, **Q**-property, **P**-property

AMS subject classifications. 90C33, 93D05

PII. S0895479800377927

1. Introduction.

Let $\mathcal{S}^n$ be the vector space of all real symmetric $n \times n$ matrices and let $\mathcal{S}^n_+$ be the cone of symmetric positive semidefinite matrices in $\mathcal{S}^n$. Given a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ and a matrix $Q \in \mathcal{S}^n$, the *semidefinite linear complementarity problem*, SDLCP(L, Q), is

$$\text{Find} \quad X \in \mathcal{S}^n_+ \quad \text{such that} \quad Y := L(X) + Q \in \mathcal{S}^n_+ \quad \text{and} \quad \operatorname{trace}(XY) = 0 \quad (\Leftrightarrow XY = 0). \tag{1}$$

This problem, which is a generalization of the standard LCP [4], is equivalent to finding a pair

$$(X, Y) \in \mathcal{S} \text{ with } X, Y \in \mathcal{S}^n_+ \text{ and } \operatorname{trace}(XY) = 0,$$

where

$$\mathcal{S} = \{(X, Y) \in \mathcal{S}^n \times \mathcal{S}^n : Y - L(X) = Q\}$$

is an affine subspace of $\mathcal{S}^n \times \mathcal{S}^n$ of dimension $\frac{n(n+1)}{2}$. By considering a general affine subspace $\mathcal{F}$ (of dimension $\frac{n(n+1)}{2}$ in $\mathcal{S}^n \times \mathcal{S}^n$) instead of $\mathcal{S}$, Kojima, Shindoh, and Hara [14] introduced the geometric-SDLCP as a model unifying semidefinite linear programs and various problems arising from system and control theory and combinatorial optimization [21], [27], [3]. (In [13], Kojima, Shida, and Shindoh show that when $\mathcal{F}$ is monotone, the geometric-SDLCP is equivalent to a semidefinite linear program.)

It is easily seen (see Appendix A) that the geometric-SDLCP can be reformulated (at the expense of increase in dimension) as an "explicit" SDLCP (1). Thus, we

*Received by the editors September 8, 2000; accepted for publication (in revised form) by M. Overton July 2, 2001; published electronically May 15, 2002.
http://www.siam.org/journals/simax/24-1/37792.html

†Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 (gowda@math.umbc.edu, http://www.math.umbc.edu/~gowda, song@math.umbc.edu).

may regard the geometric-SDLCP as equivalent to our SDLCP. While the geometric-SDLCP may be computationally more attractive (particularly for semidefinite programs), our "explicit" formulation has certain advantages: It allows us to use cone LCP results (e.g., Karamardian's theorem [12]), ideas and results from variational inequality theory (e.g., the fixed point map; see section 2), and standard degree theoretic tools. In addition, our formulation allows us to study monotone and nonmonotone problems, whereas only the monotone problem has been studied in the geometric-SDLCP setting; see [19] and the references therein.

While the SDLCP (1) is a generalization of the standard LCP, the nonpolyhedrality of $\mathcal{S}^n_+$ does not allow us to routinely extend results of standard LCP to SDLCPs. However, because of extra structure available in $\mathcal{S}^n$, one can expect interesting and useful results for the SDLCP (that are not available for a general cone LCP).

Motivated by the study of nonmonotone matrices in the standard LCP theory, Gowda and Song [7] introduced and characterized, in the context of the SDLCP above, the $\mathbf{R_0}$-, **Q**-, **P**- and globally uniquely solvable (**GUS**) properties of a linear transformation on $\mathcal{S}^n$. In [7] and [6] these properties were specialized to transformations

$$L_A(X) := AX + XA^T \quad \text{and} \quad S_A(X) := X - AXA^T,$$

and complementarity forms of theorems of Lyapunov and Stein were obtained. In particular, it was shown in [7] and [6] that A is positive stable (which means that every eigenvalue of A has positive real part) if and only if L_A has the **P**-property and that A is Schur stable (that is, every eigenvalue of A has absolute value less than one) if and only if S_A has the **P**-property, where the **P**-property of a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ is defined by the condition

$$X \in \mathcal{S}^n,\ XL(X) = L(X)X \quad \text{negative semidefinite} \Rightarrow X = 0.$$

(As is well known [16], [23], these eigenvalue conditions are related to the (global) asymptotic stability of the continuous linear dynamical system $\frac{dx}{dt} = -Ax(t)$ and the discrete linear dynamical system $x(k+1) = Ax(k)$.)

In the standard LCP theory, a matrix M is said to have the nondegeneracy property if all principal minors of M are nonzero. This is equivalent to saying that for all $q \in R^n$, the solution set of standard LCP(M, q) is finite [4]. Motivated by this equivalence, we address the following question in the first part of the paper: When does a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ have the property that for all $Q \in \mathcal{S}^n$, SDLCP(L, Q) has a finite number of solutions? We provide an answer by introducing the concepts of nondegeneracy for a linear transformation and locally-star-like property of a solution of an SDLCP.

In the second part of the paper, motivated by a result regarding the simultaneous Lyapunov stability problem for a finite set of matrices, we prove the **Q**-property of the composite transformation $L := L_{A_1} \circ L_{A_2} \circ \cdots \circ L_{A_k}$, where each A_i is positive stable and $L_{A_i}(X) := A_iX + XA_i{}^T$. We prove a similar result for the composite transformation $S := S_{A_1} \circ S_{A_2} \circ \cdots \circ S_{A_k}$, where each A_i is Schur stable and $S_{A_i}(X) := X - A_iXA_i^T$. These results are proved using degree theoretic ideas.

2. Preliminaries. As noted earlier, $\mathcal{S}^n$ denotes the set of all real symmetric $n \times n$ matrices and $\mathcal{S}^n_+ \subset \mathcal{S}^n$ is the cone of (symmetric) positive semidefinite matrices. $\mathcal{S}^n$ is a Hilbert space under the inner product

$$\langle X, Y \rangle := \text{trace}(XY). \tag{2}$$

It is well known that $\mathcal{S}^n_+$ is a closed convex self-dual cone in $\mathcal{S}^n$. We use the symbol

$$X \succeq (\succ)\, 0$$

to say that X is symmetric and positive semidefinite (respectively, positive definite); the symbol $X \preceq 0$ means $-X \succeq 0$. For a vector x, we write $x \geq 0$ to mean that every component of x is nonnegative. Given a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ and a matrix $Q \in \mathcal{S}^n$, SOL(L, Q) denotes the solution set of SDLCP(L, Q). For $X, Y \in \mathcal{S}^n$, $[X, Y]$ denotes the line segment joining X and Y, i.e.,

$$[X, Y] = \{(1-t)X + tY : t \in [0, 1]\}.$$

For a real number α, we write $\alpha^+ := \max\{\alpha, 0\}$ and $\alpha^- := \alpha^+ - \alpha$; for a diagonal matrix $D = \text{diag}\,(d_1, d_2, \ldots, d_n)$, we write $D^+ := \text{diag}\,(d_1^+, d_2^+, \ldots, d_n^+)$. For $X \in \mathcal{S}^n$, writing $X = UDU^T$ with an orthogonal U and a diagonal D, we define $X^+ := UD^+U^T$ and $X^- := UD^-U^T$. (Note that $X = X^+ - X^-$.) For a real number $r > 0$, $\mathcal{B}(X, r)$ denotes a ball of radius r with the center X under the norm induced by the inner product in (2). Given $X, Y \in \mathcal{S}^n$ with $XY = YX$, it is well known that there exist an orthogonal matrix U and diagonal matrices D and E such that $X = UDU^T$ and $Y = UEU^T$ [10]. We use I to denote (depending on the context) either the identity matrix or the identity transformation.

A matrix $A \in R^{n\times n}$ is *positive stable* if every eigenvalue of A has positive real part. For such a matrix, we recall Lyapunov's result [16], [5]: *For any given matrix $G \succ 0\,(\succeq 0)$, there is a unique $X \succ 0\,(\succeq 0)$ such that*

$$AX + XA^T = G.$$

A matrix $A \in R^{n\times n}$ is *Schur stable* if every eigenvalue of A has absolute value less than one. For such a matrix, we recall Stein's result [23]: *For any given matrix $G \succ 0\,(\succeq 0)$, there is a unique $X \succ 0\,(\succeq 0)$ such that*

$$X - AXA^T = G.$$

We have the following from [7].

DEFINITION 1. *For a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$, we say that L has the*

(a) $\mathbf{Q}$*-property if for all $Q \in \mathcal{S}^n$, SDLCP(L, Q) has a solution;*
(b) $\mathbf{P}$*-property if $XL(X) = L(X)X \preceq 0 \Longrightarrow X = 0$;*
(c) $\mathbf{R_0}$*-property if SDLCP$(L, 0)$ has a unique solution (namely zero).*

We recall some results from [6] and [7].

PROPOSITION 2. *Let $L : \mathcal{S}^n \to \mathcal{S}^n$ be linear.*

(i) *If L has the $\mathbf{Q}$-property, then there exists $X \succ 0$ such that $L(X) \succ 0$.*
(ii) $\mathbf{P}$*-property implies* $\mathbf{Q}$*- and* $\mathbf{R_0}$*-properties.*
(iii) *If L has the $\mathbf{R_0}$-property, then for all $Q \in \mathcal{S}^n$, SDLCP(L, Q) has a bounded (compact) solution set (which may be empty).*
(iv) *A matrix $A \in R^{n\times n}$ is positive stable if and only if L_A, defined by*

$$L_A(X) := AX + XA^T,$$

has the $\mathbf{P}$-property.

(v) *A matrix $A \in R^{n\times n}$ is Schur stable if and only if S_A, defined by*

$$S_A(X) := X - AXA^T,$$

has the $\mathbf{P}$-property.

In the second part of the paper, we will use the equation-based reformulation of SDLCP(L, Q): The zero set of the fixed point map

$$F(X) := X - \Pi_{\mathcal{S}^n_+}(X - [L(X) + Q]),$$

where $\Pi_{\mathcal{S}^n_+}$ is the projection mapping from $\mathcal{S}^n$ onto $\mathcal{S}^n_+$, coincides with the solution set of SDLCP(L, Q) [8].

3. Nondegeneracy, locally-star-like property, and finiteness of SDLCP solution sets. In the standard LCP theory, the nondegeneracy of a matrix is defined as follows: A matrix $M \in R^{n\times n}$ is nondegenerate if every principal minor of M is nonzero. It is well known (see section 3.6 in [4]) that M is nondegenerate if and only if for all $q \in R^n$, the linear complementarity problem LCP(M, q) has a finite number of solutions, where LCP(M, q) is to find a vector $x \in R^n$ such that

$$x \geq 0, \quad y := Mx + q \geq 0, \quad \text{and} \quad x^T y = 0 \text{ (or equivalently, } x * y = 0)$$

with $x * y$ denoting the componentwise product of x and y. Here we make the observation (which is easy to verify) that M is nondegenerate if and only if

$$x * (Mx) = 0 \Longrightarrow x = 0.$$

This motivates us to introduce the concept of nondegeneracy for a linear transformation from $\mathcal{S}^n$ to $\mathcal{S}^n$ in the following way.

DEFINITION 3. *A linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ is said to be nondegenerate if*

$$XL(X) = 0 \Longrightarrow X = 0.$$

It is clear that if L has the **P**-property, then it is nondegenerate. Also, every nondegenerate transformation has the $\mathbf{R}_0$-property.

In the results below, we describe the nondegeneracy property for transformations L_A and S_A. But first we recall a result of Taussky and Wielandt [24]: For an $m \times m$ complex matrix C, the spectrum of the transformation $L_C : \mathcal{H}^m \to \mathcal{H}^m$ defined by $L_C(X) := CX + XC^*$ is $\sigma(C) + \sigma(C^*)$, where $\sigma(C)$ denotes the spectrum (i.e., the set of all eigenvalues) of C, etc. Here $\mathcal{H}^m$ denotes the space of all Hermitian $m \times m$ matrices.

THEOREM 4. *Let $A \in R^{n\times n}$. Then the following are equivalent:*

(i) $0 \notin \sigma(A) + \sigma(A)$.

(ii) *L_A is nondegenerate.*

(iii) *L_A is invertible as a transformation from $\mathcal{S}^n$ to itself.*

(iv) *L_A is invertible as a transformation from $\mathcal{H}^n$ to itself.*

Proof. (i) $\Longrightarrow$ (ii). Assume that (i) holds and that there is a nonzero X such that $XL_A(X) = 0$. Noting commutativity of X and $L_A(X)$, we may assume that X and $Y := L_A(X)$ are diagonal matrices. (This can be achieved by considering UXU^T, UYU^T, and UAU^T for an appropriate orthogonal matrix U [10].) We write

$$X = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 0 & 0 \\ 0 & E \end{bmatrix}, \tag{3}$$

where D and E are diagonal matrices with D invertible. Writing A in an appropriate block form as

$$A = \begin{bmatrix} C & * \\ N & * \end{bmatrix} \tag{4}$$

we get $DL_C(D) = 0$ and $DDN^T = 0$ from $XL_A(X) = 0$. Since D is invertible, we get $L_C(D) = 0$ and $N = 0$. From the block form of A it follows that the spectrum of C is a subset of the spectrum of A, and hence C inherits the property (i) from A. But then (because of the Taussky–Wielandt result above), L_C is invertible, and hence D must be zero, leading to a contradiction.

(ii) $\Longrightarrow$ (iii). The proof is obvious.

(iii) $\Longrightarrow$ (iv). Assume that (iii) holds. First we observe that A is invertible. (If $Au = 0$, then $A(uu^T) + (uu^T)A^T = 0$; since $uu^T \in \mathcal{S}^n$, from (iii), $uu^T = 0$, and so $u = 0$.) Suppose, if possible, that for some $Z \in \mathcal{H}^n$, $AZ + ZA^T = 0$. Writing $Z = X + iY$ with X and Y real, noting that A is real and using (iii), we see that $Z = iY$, where Y is skew-symmetric. From $AY + YA^T = 0$, we see that $S := AY \in \mathcal{S}^n$. But then $AS + SA^T = A(AY + YA^T) = 0$ implies that $S = 0$; since A is invertible, we see that $Y = 0$. The invertibility of $L_A : \mathcal{H}^n \to \mathcal{H}^n$ follows.

(iv) $\Longrightarrow$ (i). The proof follows from the above Taussky–Wielandt result. □

Similar to Theorem 4, we have the following for S_A.

THEOREM 5. *Let $A \in R^{n\times n}$. Then the following are equivalent:*

(i) $1 \notin \sigma(A)\,\sigma(A)$.
(ii) *S_A is nondegenerate.*
(iii) *S_A is invertible as a transformation from $\mathcal{S}^n$ to itself.*
(iv) *S_A is invertible as a transformation from $\mathcal{H}^n$ to itself.*

Proof. Here we provide (only) a sketch of the proof. Suppose any of the given conditions holds. Then -1 is not an eigenvalue of A and so $(I+A)$ is invertible. (This is clear when (i) holds. In the presence of other conditions, $Au = -u$ implies that $X - AXA^T = 0$ with $X = uu^T$ and so $u = 0$.) Let $B := (I + A)^{-1}(I - A)$. Then

$$\sigma(B) = \left\{ \frac{1-\zeta}{1+\zeta} \,:\, \zeta \in \sigma(A) \right\}. \tag{5}$$

Now it can be easily verified that

$$Y = BX + XB^T \quad \text{with} \quad X \in \mathcal{H}^n \quad \Longleftrightarrow \quad \frac{1}{2}(I + A)Y(I + A^T) = X - AXA^T. \tag{6}$$

It follows from (5) and (6) that (i), (iii), and (iv) are respectively equivalent to

(i′) $0 \notin \sigma(B) + \sigma(B)$;
(iii′) L_B is invertible as a transformation from $\mathcal{S}^n$ to itself;
(iv′) L_B is invertible as a transformation from $\mathcal{H}^n$ to itself.

Because of Theorem 4 (applied to B), we see that (i′), (iii′), and (iv′), and hence (i), (iii), and (iv), are equivalent. To complete the proof, we show that (i) implies (ii), the implication (ii) $\Longrightarrow$ (iii) being obvious. Assuming (i), we suppose that for some $X \in \mathcal{S}^n$, $XS_A(X) = 0$. Writing X and A as in (3) and (4) with D diagonal and invertible, we deduce that $D(D - CDC^T) = 0$ and $-DCDN^T = 0$. Since D is invertible, we get $D = CDC^T$ and (hence) the invertibility of C. We also see that $N = 0$. From the block form of A, we see that C inherits the property (i) from A. Thus $1 \notin \sigma(C)\sigma(C)$. Now let $\lambda \in \sigma(C)$ and $u \neq 0$ with $C^T u = \lambda u$. Then $Du = \lambda C(Du)$ implies that $\frac{1}{\lambda} \in \sigma(C)$, leading to a contradiction. Hence S_A is nondegenerate and the proof is complete. □

In view of the LCP result for nondegenerate matrices mentioned above, we may ask whether the SDLCP solution sets corresponding to a nondegenerate transformation are finite. The following example shows that this is false.

Example 1. In $R^{2\times 2}$, let $A = -\frac{1}{2}I$ and $Q = I$. Then the solution set of SDLCP(L_A, Q), consisting of all matrices of the form

$$X = \begin{bmatrix} \frac{1+\sqrt{1-4\lambda^2}}{2} & \lambda \\ \lambda & \frac{1-\sqrt{1-4\lambda^2}}{2} \end{bmatrix}$$

with λ real and $4\lambda^2 \leq 1$, is infinite. For any diagonal (or, more generally, symmetric) matrix $A \in R^{n\times n}$ with a repeated negative eigenvalue, we can modify X and Q appropriately so that the $SOL(L_A, Q)$ is infinite.

Now, to address the finiteness issue, we introduce the following.

DEFINITION 6. *For a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$ and $Q \in \mathcal{S}^n$, let X_0 be a solution of SDLCP(L, Q). We say that X_0 has the locally-star-like property if there exists a ball $\mathcal{B}(X_0, r)$ such that for all $X \in \mathcal{B}(X_0, r) \cap SOL(L, Q)$,*

$$[X_0, X] \subseteq SOL(L, Q),$$

or, equivalently,

$$(tX_0 + (1-t)X)(tY_0 + (1-t)Y) = 0 \quad \forall\, t \in [0, 1],$$

where $Y = L(X) + Q$ and $Y_0 = L(X_0) + Q$.

We note that if $SOL(L, Q)$ is convex, then every solution in $SOL(L, Q)$ has the locally-star-like property.

We now give a characterization of the finiteness of solution sets in SDLCPs; recall that a solution X_0 of SDLCP(L, Q) is locally unique if it is the only solution in a neighborhood of X_0.

THEOREM 7. *For a linear transformation $L : \mathcal{S}^n \to \mathcal{S}^n$, the following are equivalent:*

(a) *For all $Q \in \mathcal{S}^n$, SDLCP(L, Q) has a finite number of solutions.*
(b) *For all $Q \in \mathcal{S}^n$, each solution of SDLCP(L, Q) is locally unique.*
(c) *L is nondegenerate, and for all $Q \in \mathcal{S}^n$, every solution of SDLCP(L, Q) is locally-star-like.*

Proof. (a)$\Longrightarrow$ (b) is clear.

(b)$\Longrightarrow$ (a). Condition (b) implies that SDLCP$(L, 0)$ has the trivial solution. (This is because $SOL(L, 0)$ is a cone.) Hence L has the $\mathbf{R_0}$-property, which means that the $SOL(L, Q)$ is compact for all Q. This, with assumption (b), gives (a).

(b)$\Longrightarrow$(c). To show the nondegeneracy part, let $X \in \mathcal{S}^n$ be a nonzero matrix such that $XL(X) = 0$. Noting the commutativity, we write

$$X = UDU^T \quad \text{and} \quad L(X) = UEU^T$$

for some orthogonal matrix U and diagonal matrices D and E. From $DE = 0$, we get

$$X^+(L(X))^+ = X^-(L(X))^- = X^+(L(X))^- = X^-(L(X))^+ = 0.$$

Defining $Q := (L(X))^+ - L(X^+) = (L(X))^- - L(X^-)$, we see that SDLCP$(L, Q)$ has two distinct solutions X^+ and X^- with

$$(tX^+ + (1-t)X^-)(t(L(X))^+ + (1-t)(L(X))^-) = 0 \quad \forall\, t \in [0, 1],$$

i.e., $[X^-, X^+] \subseteq SOL(L, Q)$. This contradicts (b).

Now take any $Q \in \mathcal{S}^n$. For an $X_0 \in SOL(L, Q)$, the locally-star-like property is trivially satisfied since X_0 is locally unique.

(c)$\Longrightarrow$(b). Fix $Q \in \mathcal{S}^n$ and suppose that there is a sequence $\{X_k\} \subseteq SOL(L, Q)$ which converges to $X_0 \in SOL(L, Q)$ with $X_k \neq X_0$ for all k. By the locally-star-like condition, $[X_0, X_k] \subseteq SOL(L, Q)$ for all large k, resulting in $(X_k - X_0)(Y_k - Y_0) = 0$, where $Y_k = L(X_k) + Q$ for all $k = 0, 1, 2, \ldots$. But from the nondegeneracy property, this implies that $X_k = X_0$ for all large k, contradicting our assumption. This completes the proof. □

While the locally-star-like property of a solution point of an SDLCP comes up naturally in Theorem 7, it is not clear how to characterize (or verify) this property when a (nonlocally unique) solution of SDLCP is given. When $A \in R^{n\times n}$ is positive stable and positive semidefinite, it is known (see [7]) that for every Q, SDLCP(L_A, Q) has a unique solution and hence provides an instance of a situation where item (a) of Theorem 7 holds. It is not clear if item (a) holds for L_A when A is (merely) positive stable or, more generally, for an L that has the **P**-property. In the following example, we describe a matrix A such that A is neither positive stable nor positive semidefinite, yet $SOL(L_A, Q)$ is finite for every Q.

Example 2. Let

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

Then for all $Q \in \mathcal{S}^2$, SDLCP(L_A, Q) has a finite solution set; see Appendix B.

It is very likely, in light of Examples 1 and 2 above, that SOL(L_A, Q) is finite for all $Q \in \mathcal{S}^n$ when the following conditions hold:

(a) A is diagonal (or symmetric);
(b) $0 \notin \sigma(A) + \sigma(A)$; and
(c) every negative eigenvalue of A is simple.

4. The Q-property of a composite transformation and simultaneous stability of a commuting family. Given a set $\mathcal{A}$ of matrices, the simultaneous stability problem is as follows: Find a (symmetric) positive semidefinite X such that $AX + XA^T$ is positive definite for all $A \in \mathcal{A}$. As is well known, the above stability problem is related to the asymptotic stability of the linear time-varying system

$$\frac{dx}{dt} = -A(t)x,$$

where $x(t) \in R^n$ and $A(t) \in \mathcal{A}$ for all t [3].

In connection with this problem, Narendra and Balakrishnan [20] prove the following.

THEOREM 8. *Let $\{A_1, \ldots, A_k\}$ consist of (pairwise) commuting positive stable matrices. Then there exists $X \succ 0$ such that $L_{A_i}(X) := A_iX + XA_i^T \succ 0$ for all $i = 1, \ldots, k$.*

Their proof consists of proving the existence of the finite sequence $\{X_0, X_1, \ldots, X_k\}$ of (symmetric) positive definite matrices with $X_0 = I$ and $A_iX_i + X_iA_i^T = X_{i-1}$ (this is done by using the positive stable property of each A_i) and then showing (by commutativity of the A_i's) that $X := X_k$ satisfies the conclusion of the theorem. This proof reveals the existence of an $X \succ 0$ such that $L(X) \succ 0$, where $L : \mathcal{S}^n \to \mathcal{S}^n$ is defined by

$$L := L_{A_1} \circ \cdots \circ L_{A_k}.$$

Motivated by the equivalence of the **P**- and **Q**-properties of L_A to the positive stable property of A [7], we may ask whether the above L has the **P**- and **Q**-properties when each A_i is positive stable. We answer this by means of the following theorem and an example.

THEOREM 9. *Let $\{A_1, \ldots, A_k\}$ consist of positive stable matrices. Then $L := L_{A_1} \circ \cdots \circ L_{A_k}$ has the* **Q***-property. In particular, there exists an $X \succ 0$ such that $L(X) \succ 0$.*

Proof. We first claim that L has the $\mathbf{R}_0$-property. To see the claim, suppose $X \succeq 0$, $Y := L(X) \succeq 0$, and $XY = 0$. Without loss of generality, we can write

$$X = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 0 & 0 \\ 0 & E \end{bmatrix}, \tag{7}$$

where D and E are diagonal matrices with $D \succeq 0$ and $E \succeq 0$. Now writing $Z_1 := (L_{A_2} \circ \cdots \circ L_{A_k})(X)$, we see that $L_{A_1}(Z_1) = L(X) = Y \succeq 0$. Since A_1 is positive stable, by Lyapunov's theorem (see section 2), $Z_1 \succeq 0$. Repeating this argument, we get $Z_{k-1} := L_{A_k}(X) \succeq 0$. Now using the (block) form of X, we get

$$0 \preceq Z_{k-1} = L_{A_k}(X) = \begin{bmatrix} (Z_{k-1})_1 & (Z_{k-1})_2 \\ (Z_{k-1})_2{}^T & 0 \end{bmatrix}.$$

From this we get $(Z_{k-1})_2 = 0$. This implies that Z_{k-1} (in block form) looks like X with $(Z_{k-1})_1$ in place of D. By repeating this argument several times, we see that

$$Z_1 = \begin{bmatrix} (Z_1)_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

But then, because of the block forms of Z_1 and Y, $Z_1 L_{A_1}(Z_1) = Z_1 Y = 0$. Since A_1 is positive stable, L_{A_1} has the **P**-property and so $Z_1 = 0$. Since each transformation L_{A_i} is nonsingular (once again by the **P**-property), we see from $L_{A_1}(Z_1) = L(X)$ that $X = 0$. Thus we have shown that SDLCP$(L, 0)$ has only one solution, namely, the zero solution.

Now, fix any $Q \in \mathcal{S}^n$. We show that SDLCP(L, Q) has a solution by showing that the fixed point map

$$F(X) := X - \Pi_{\mathcal{S}^n_+}(X - [L(X) + Q]),$$

where $\Pi_{\mathcal{S}^n_+}$ denotes the projection mapping from $\mathcal{S}^n$ onto $\mathcal{S}^n_+$, has a zero in $\mathcal{S}^n$. That F has a zero is shown via degree theoretic arguments. Define a homotopy $H : \mathcal{S}^n \times [0,1] \to \mathcal{S}^n$ by

$$H(X,t) := X - \Pi_{\mathcal{S}^n_+}(X - [L_t(X) + tQ]),$$

where

$$L_t(X) := (L_{tA_1 + (1-t)\frac{1}{2}I} \circ L_{tA_2 + (1-t)\frac{1}{2}I} \circ \cdots \circ L_{tA_k + (1-t)\frac{1}{2}I})(X).$$

We see that

$$H(X,0) = I(X)$$

and

$$H(X,1) = X - \Pi_{\mathcal{S}^n_+}(X - [L(X) + Q]) = F(X).$$

Now for each index i and $t \in [0,1]$, $tA_i + (1-t)\frac{1}{2}I$ is positive stable, and from the first part of the proof, L_t has the $\mathbf{R_0}$-property for all $t \in [0,1]$. We now show that the zero sets of $H(\,\cdot\,,t)$ as t varies over $[0,1]$ are (uniformly) bounded. Suppose there exist sequences $\{X_k\} \subset \mathcal{S}^n$ and $\{t_k\} \subset [0,1]$ such that $H(X_k,t_k) = 0$ for all k and $\|X_k\| \to \infty$, where $\|X\| = \sqrt{\mathrm{trace}(X^2)}$. Then X_k solves SDLCP(L_{t_k}, t_kQ) and so

$$X_k \succeq 0, \quad Y_k := L_{t_k}(X_k) + t_kQ \succeq 0, \quad \text{and} \quad X_kY_k = 0. \tag{8}$$

Assuming $t_k \to t^*$ and $\frac{X_k}{\|X_k\|} \to X^*$, it follows from (8) that

$$X^* \succeq 0, \quad Y^* := L_{t^*}(X^*) \succeq 0, \quad \text{and} \quad X^*Y^* = 0.$$

Since X^* has norm one, it is a nonzero solution of SDLCP$(L_{t^*}, 0)$, contradicting the earlier observation. Hence we have the uniform boundedness of the zero sets of $H(\,\cdot\,,t)$ as t varies. Now let Ω be a bounded open set in $\mathcal{S}^n$ containing all of these zero sets (note that $0 \in \Omega$). Then, by the homotopy invariance of the degree [15, Thm. 2.1.2],

$$\deg(F, \Omega, 0) \;=\; \deg(I, \Omega, 0) \;=\; 1.$$

By Theorem 2.1.1 in [15], we conclude that SDLCP(L,Q) has a solution.

Now to see the second conclusion of the theorem, we consider a solution X_0 of SDLCP$(L,-I)$. Then $L(X_0) - I \succeq 0$ implies that $L(X_0) \succ 0$. Since $X_0 \succeq 0$, we may perturb it to get an $X \succ 0$ such that $L(X) \succ 0$. This completes the proof. □

We may ask if the transformation L in the above theorem has the $\mathbf{P}$-property. The following example shows that this is not the case even when the matrices commute.

Example 3. Let

$$A = \begin{bmatrix} -1 & -3 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -1 & -6 \\ 2 & 5 \end{bmatrix}.$$

It can be easily checked that A and B are commuting positive stable matrices. For

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \text{ we have } \quad XL_A(L_B(X)) = \begin{bmatrix} -8 & 0 \\ 0 & 0 \end{bmatrix} \preceq 0,$$

so $L := L_A \circ L_B$ does not have the $\mathbf{P}$-property.

For a matrix $A \in R^{n\times n}$, we recall that $S_A(X) := X - AXA^T$. Motivated by the previous theorem, we may ask if a similar result is valid for the composition of several transformations of the form S_A. The following theorem answers this question.

THEOREM 10. *Let $\{A_1, \ldots, A_k\}$ consist of Schur stable matrices. Then $S := S_{A_1} \circ \cdots \circ S_{A_k}$ has the $\mathbf{Q}$-property. In particular, there exists an $X \succ 0$ such that $S(X) \succ 0$. Moreover, if the matrices A_j $(j = 1,2,\ldots,k)$ commute pairwise, then there is an $X \succ 0$ such that*

$$S_{A_j}(X) \succ 0 \quad \forall j = 1,2,\ldots,k.$$

Proof. The proof is similar to that of the previous theorem. For completeness, we sketch a proof for two matrices A and B with $S(X) = S_A \circ S_B(X)$. We first claim that S has the $\mathbf{R}_0$-property. Let $0 \neq X \succeq 0$ be such that $Y := S(X) \succeq 0$ and $XY = 0$. If X is nonsingular, then $Y = 0$ and $X = 0$ (by the nonsingularity of S_A and S_B), leading to a contradiction. Without loss of generality, we may write X and Y as in (7), where D and E are diagonal with $D \succ 0$. From $0 \preceq Y = S_A(S_B(X))$ and Stein's

result mentioned before, we see that $S_B(X) \succeq 0$. Let B, in the block form, be given by the right-hand side of (4). Using the block form of X, we compute $S_B(X)$ and observe that the block in the lower right-hand corner, namely, $-NDN^T$, is positive semidefinite. But $-NDN^T$ is negative semidefinite, and hence $-NDN^T = 0$. Since D is positive definite, we must have $N = 0$. This leads to

$$Z := S_B(X) = \begin{bmatrix} D - CDC^T & 0 \\ 0 & 0 \end{bmatrix}.$$

Now $ZS_A(Z) = ZY = 0$. Since S_A has the **P**-property, we see that $Z = 0$, and hence $X = 0$ (because of the **P**-property of S_B). This is a contradiction, and so S has the $\mathbf{R}_0$-property. To see the **Q**-property of S, we fix a $Q \in \mathcal{S}^n$ and consider the homotopy

$$H(X,t) := X - \Pi_{\mathcal{S}^n_+}(X - [S_t(X) + tQ]),$$

where

$$S_t := S_{tA} \circ S_{tB}$$

and $t \in [0,1]$. We see that $H(X,0) = X$ and $H(X,1) = F(X)$, where

$$F(X) = X - \Pi_{\mathcal{S}^n_+}(X - [S(X) + Q]).$$

We proceed as in the previous theorem and show that F has a zero, say, X, in an appropriate bounded open set. This X will solve SDLCP(S,Q). By specializing $Q = -I$ (as in the proof of the previous theorem), we deduce the existence of $X \succ 0$ such that $S(X) \succ 0$. By Stein's theorem, $0 \prec S_A(S_B(X))$ implies that $S_B(X) \succ 0$. Finally, when A and B commute, S_A and S_B commute, and we see that $0 \prec S_B(S_A(X))$ implies $S_A(X) \succ 0$. This completes the proof. □

Remarks. The last conclusion in the previous theorem is well known in control theory; see [17]. The above two results motivate us to ask whether similar results exist in the standard LCP theory. To answer this, we first recall some definitions from the LCP theory [4]. We say that a matrix M is

(i) a **P**-matrix if all principal minors of M are positive, or, equivalently,

$$x * (Mx) \leq 0 \Longrightarrow x = 0,$$

where $x * (Mx)$ is the componentwise product of x and Mx;

(ii) a **Z**-matrix if all off-diagonal entries of M are nonpositive;

(iii) a **Q**-matrix if for all $q \in R^n$, LCP(M,q) has a solution.

The LCP analogue of Theorems 9 and 10 is the following.

PROPOSITION 11. *Let $\{M_1, \ldots, M_k\}$ be a set of $n \times n$ matrices such that each M_i is a* **P***-matrix with $M_i^{-1} \geq 0$; i.e., every entry in M_i^{-1} is nonnegative. Then $M := M_1 M_2 \cdots M_k$ is a* **Q***-matrix. In particular, this conclusion holds if each M_i is a* $\mathbf{P} \cap \mathbf{Z}$*-matrix.*

To compare this proposition with Theorem 9, we note that the **P**-matrix property (described with respect to the cone R^n_+ of nonnegative vectors in R^n) is analogous to the **P**-property of Definition 1. The condition $M_i^{-1} \geq 0$ (which is equivalent to $M_i^{-1}(R^n_+) \subseteq R^n_+$) is analogous to the condition $L_{A_i}^{-1}(\mathcal{S}^n_+) \subseteq \mathcal{S}^n_+$, which holds when A_i is positive stable.

Now, while a degree theoretic proof, similar to those of Theorems 9 and 10, can be given for this proposition, we present an elementary argument due to Parthasarathy based on the following well-known results from the LCP theory [4]:

(a) Every **P**-matrix is a **Q**-matrix. Also, the inverse of a **P**-matrix is a **P**-matrix.
(b) An (entrywise) nonnegative matrix M is a **Q**-matrix if and only if each diagonal entry of M is positive.
(c) If M is a $\mathbf{P} \cap \mathbf{Z}$-matrix, then M^{-1} is a nonnegative matrix with positive diagonal.
(d) The inverse of an invertible **Q**-matrix is a **Q**-matrix.

Now to justify the proposition, assume that each M_i is a **P**-matrix with $M_i^{-1} \geq 0$. Then M_i^{-1} is a **P**-matrix and hence a **Q**-matrix. Since $M_i^{-1} \geq 0$, the diagonal entries of M_i^{-1} are all positive. It follows that the inverse of $M := M_1 M_2 \cdots M_k$ is a nonnegative matrix with a positive diagonal. Hence M^{-1} is a **Q**-matrix. From this we conclude that M is a **Q**-matrix. The second part of the proposition follows from the first part and item (c) above.

At this stage one may wonder whether a product of $\mathbf{P} \cap \mathbf{Z}$-matrices is necessarily either a **P**-matrix or a **Z**-matrix. In the following example, we describe a $\mathbf{P} \cap \mathbf{Z}$-matrix whose third power is neither a **P**-matrix nor a **Z**-matrix.

Example 4. Let

$$A = \begin{bmatrix} 2 & -2 & -1 \\ 0 & 7 & -3 \\ -1 & 0 & 1 \end{bmatrix} \quad \text{so that} \quad A^3 = \begin{bmatrix} 7 & -136 & 52 \\ 30 & 337 & -174 \\ -8 & 20 & -1 \end{bmatrix}.$$

We see that A is a $\mathbf{P} \cap \mathbf{Z}$-matrix, while A^3 is neither a **P**-matrix nor a **Z**-matrix.

Remarks. In [22, p. 14], Parthasarathy presents two $\mathbf{P} \cap \mathbf{Z}$-matrices of size 4×4 whose product is neither a **P**-matrix nor a **Z**-matrix. He also notes [22, p. 13] that a product of two $\mathbf{P} \cap \mathbf{Z}$-matrices of size 3×3 must be a **P**-matrix.

Appendix A. Here we show that the geometric-SDLCP of Kojima, Shindoh, and Hara can be reformulated as SDLCP (1).

Let $\mathcal{F}$ be an affine subspace of $\mathcal{S}^n \times \mathcal{S}^n$ of dimension $\frac{n(n+1)}{2}$ and consider the geometric-SDLCP($\mathcal{F}$):

$$\text{Find } (X, Y) \in \mathcal{F} \cap (\mathcal{S}^n_+ \times \mathcal{S}^n_+) \text{ such that trace } (XY) = 0.$$

We may write, without loss of generality,

$$\mathcal{F} = \{(X, Y) \in \mathcal{S}^n \times \mathcal{S}^n : L_1(X) + L_2(Y) = B\},$$

where L_1 and L_2 are linear transformations from $\mathcal{S}^n$ to itself, and $B \in \mathcal{S}^n$.

We define $L : \mathcal{S}^{3n} \to \mathcal{S}^{3n}$ and $Q \in \mathcal{S}^{3n}$ by

$$L\left(\begin{bmatrix} X & * & * \\ * & Y & * \\ * & * & Z \end{bmatrix}\right) = \begin{bmatrix} Y & 0 & 0 \\ 0 & L_1(X) + L_2(Y) & 0 \\ 0 & 0 & -L_1(X) - L_2(Y) \end{bmatrix},$$

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -B & 0 \\ 0 & 0 & B \end{bmatrix}.$$

It is easily verified that if

$$W = \begin{bmatrix} X & * & * \\ * & Y & * \\ * & * & Z \end{bmatrix}$$

solves SDLCP(L, Q), then (X, Y) solves the geometric-SDLCP$(\mathcal{F})$. On the other hand, if (X, Y) solves the geometric-SDLCP$(\mathcal{F})$, then

$$W = \begin{bmatrix} X & 0 & 0 \\ 0 & Y & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

solves SDLCP(L, Q). Thus the solvability of geometric-SDLCP$(\mathcal{F})$ is equivalent to the solvability of SDLCP(L, Q).

Appendix B. Here we justify the assertion made in Example 2, namely, that for

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

SDLCP(L_A, Q) has a finite solution set for all $Q \in \mathcal{S}^2$.

Suppose, if possible, that there is a Q with SDLCP(L_A, Q) consisting of infinitely many solutions. Let $\{X_k\}$ be an infinite sequence of solutions for SDLCP(L_A, Q), where we write

$$X_k = \begin{bmatrix} x_k & y_k \\ y_k & z_k \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} p & q \\ q & r \end{bmatrix}.$$

From $X_k \succeq 0$, $L_A(X_k) + Q \succeq 0$, and $X_k[L_A(X_k) + Q] = 0$, we see the existence of infinitely many *positive* λ_k's satisfying

$$x_k(p - 2x_k) = -y_k(y_k + q) = z_k(4z_k + r) = \lambda_k \tag{9}$$

with

$${y_k}^2 = x_k z_k \tag{10}$$

for each k. From (9) we see that $p > 0$ and $q \neq 0$. Solving various equations in (9), we get

$$x_k = \frac{p \pm \sqrt{p^2 - 8\lambda_k}}{4}, \tag{11}$$

$$y_k = \frac{-q \pm \sqrt{q^2 - 4\lambda_k}}{2}, \quad \text{and} \tag{12}$$

$$z_k = \frac{-r \pm \sqrt{r^2 + 16\lambda_k}}{8}. \tag{13}$$

From (11) and (12), we see that $\{\lambda_k\}$ is bounded; without loss of generality, we may say that

$$\lambda_k \longrightarrow \lambda_* \in \left[0, \min\left\{\frac{p^2}{8}, \frac{q^2}{4}\right\}\right]. \tag{14}$$

Assuming that the signs ($+$ or $-$) in x_k, y_k, and z_k are fixed for all k, we let $x_k \longrightarrow x_*$, $y_k \longrightarrow y_*$, and $z_k \longrightarrow z_*$.

Case 1. $\lambda_* = \min\{\frac{p^2}{8}, \frac{q^2}{4}\}$. We consider the following subcases.

(1) $\frac{q^2}{4} < \frac{p^2}{8}$: From (10), we have

$$y_k{}^2 - y_*{}^2 = (x_k - x_*)z_k + x_*(z_k - z_*). \tag{15}$$

Dividing both sides of (15) by $(\lambda_k - \lambda_*)$ and taking the limit, we see that the left-hand side is infinite, whereas the right-hand side is finite. So this subcase is not possible.

(2) $\frac{q^2}{4} > \frac{p^2}{8}$: This is similar to item (1). The right-hand side is infinite yet the left-hand side is finite. Once again, this subcase is not possible.

(3) $\frac{q^2}{4} = \frac{p^2}{8}$: In this case, we see that

$$y_k = cx_k \quad \text{or} \quad x_k y_k = d\lambda_k,$$

where c and d are constants. From these relations and (10), we have (i) $x_k = \frac{1}{c^2} z_k$ or (ii) $y_k{}^3 = d\lambda_k z_k$. Since

$$\lim_{k\to\infty} \frac{x_k - x_*}{\lambda_k - \lambda_*} \quad \text{and} \quad \lim_{k\to\infty} \frac{y_k{}^3 - y_*{}^3}{\lambda_k - \lambda_*}$$

are infinite while

$$\lim_{k\to\infty} \frac{z_k - z_*}{\lambda_k - \lambda_*}$$

is finite, neither (i) nor (ii) can be true.

Therefore, Case 1 is not possible.

Case 2. $0 \le \lambda_* < \min\{\frac{p^2}{8}, \frac{q^2}{4}\}$. By suppressing k in (11)–(13) and putting $t = \sqrt{r^2 + 16\lambda}$, we may regard x, y, and z as functions of t with power series expansions valid in (α, β), where $\alpha := |r|$ and $\beta := \sqrt{\min\{(2p^2 + r^2), (4q^2 + r^2)\}}$:

$$x = \sum_{n=0}^{\infty} a_n (t^2)^n,$$

$$y = \sum_{n=0}^{\infty} b_n (t^2)^n, \tag{16}$$

and

$$z = \frac{1}{8}(-r \pm t).$$

Then (10) shows that

$$\left(\sum_{n=0}^{\infty} b_n (t^2)^n \right)^2 = \left(\sum_{n=0}^{\infty} a_n (t^2)^n \right) \frac{1}{8}(-r \pm t) \tag{17}$$

holds for all $t = t_k$. Since $t_k \longrightarrow t_* \in [\alpha, \beta) \subseteq (-\beta, \beta)$ and the power series in (17) are defined in $(-\beta, \beta)$, the above equality must hold for all $t \in (-\beta, \beta)$. But since the left-hand side has only even powers of t while the right-hand side has both even and odd powers of t, the series on the left must be identically zero. Thus y and hence y_k must be zero for all k. This implies that $\lambda_k = 0$, which is a contradiction. Therefore

Case 2 does not occur either. Thus we cannot have infinitely many solutions in the solution set of SDLCP(L_A, Q).

Acknowledgments. We wish to thank Mohamed Tawhid for discussions related to the local uniqueness issues in the first part of the paper. Our thanks are also due to T. Parthasarathy of the Indian Statistical Institute, New Delhi, for communicating the elementary proof of Proposition 11.

REFERENCES

[1] R. Bellman, *Introduction to Matrix Analysis*, SIAM, Philadelphia, 1997.
[2] A. Berman and R.J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[3] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
[4] R.W. Cottle, J.-S. Pang, and R.E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.
[5] F.R. Gantmacher, *Theory of Matrices*, Vol. II, Chelsea Publishing, New York, 1959.
[6] M.S. Gowda and T. Parthasarathy, *Complementarity forms of the theorems of Lyapunov and Stein, and related results*, Linear Algebra Appl., 320 (2000), pp. 131–144.
[7] M.S. Gowda and Y. Song, *On semidefinite linear complementarity problems*, Math. Program., 88 (2000), pp. 575–587.
[8] P.T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications*, Math. Program., 48 (1990), pp. 161–220.
[9] D. Hershkowitz, *Recent directions in matrix stability*, Linear Algebra Appl., 171 (1992), pp. 161–186.
[10] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[11] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[12] S. Karamardian, *An existence theorem for the complementarity problem*, J. Optim. Theory Appl., 19 (1976), pp. 227–232.
[13] M. Kojima, M. Shida, and S. Shindoh, *Reduction of monotone linear complementarity problems over cones to linear programs over cones*, Acta. Math. Vietnam, 22 (1997), pp. 147–157.
[14] M. Kojima, S. Shindoh, and S. Hara, *Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices*, SIAM J. Optim., 7 (1997), pp. 86–125.
[15] N.G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.
[16] A.M. Lyapunov, *Probléme général de la stabilité des mouvement*, Ann. Fac. Sci. Toulouse, 9 (1907), pp. 203–474; French translation of the original paper published in 1893 in Comm. Soc. Math. Kharkow; reprinted as Vol. 17 in Ann. of Math. Stud., Princeton University Press, Princeton, NJ, 1949.
[17] Y. Mori, T. Mori, and Y. Kuroe, *Classes of discrete linear systems having common quadratic Lyapunov functions*, Proc. Amer. Control Conf., 5 (1995), pp. 3364–3365.
[18] M. Mesbahi and G.P. Papavassilopoulos, *A cone programming approach to the bilinear inequality problem and its geometry*, Math. Program., 77 (1997), pp. 247–272.
[19] R.D.C. Monteiro and T. Tsuchiya, *Polynomiality of primal-dual algorithms for semidefinite linear complementarity problems based on Kojima-Shindoh-Hara family of directions*, Math. Program., 84 (1999), pp. 39–54.
[20] K.S. Narendra and J. Balakrishnan, *A common Lyapunov function for stable LTI systems with commuting $\mathcal{A}$-matrices*, Trans. Automat. Control, 39 (1994), pp. 2469–2471.
[21] M. Overton and H. Wolkowicz, eds., *Semidefinite programming*, Math. Program., 77 (1997), pp. 97–320.
[22] T. Parthasarathy, *On Global Univalence Theorems*, Lecture Notes in Math. 977, Springer-Verlag, New York, 1983.
[23] P. Stein, *Some general theorems on iterants*, J. Research Nat. Bur. Standards, 48 (1952), pp. 82–83.

[24] O. Taussky and H. Wielandt, *On the matrix function $AX + X'A'$*, Arch. Rational Mech. Anal., 9 (1962), pp. 93–96.
[25] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Program., 83 (1998), pp. 159–185.
[26] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[27] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite programming: Theory, Algorithms, and Applications*, Kluwer Academic, Boston, 2000.
[28] N. Yamashita and M. Fukushima, *A new merit function and a descent method for semidefinite complementarity problems*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, Kluwer Academic, Boston, 1999, pp. 405–420.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 40–58

BLOCK-ITERATIVE ALGORITHMS WITH DIAGONALLY SCALED OBLIQUE PROJECTIONS FOR THE LINEAR FEASIBILITY PROBLEM*

YAIR CENSOR† AND TOMMY ELFVING‡

Abstract. We formulate a block-iterative algorithmic scheme for the solution of systems of linear inequalities and/or equations and analyze its convergence. This study provides as special cases proofs of convergence of (i) the recently proposed component averaging (CAV) method of Censor, Gordon, and Gordon [*Parallel Comput.*, 27 (2001), pp. 777–808], (ii) the recently proposed block-iterative CAV (BICAV) method of the same authors [*IEEE Trans. Medical Imaging*, 20 (2001), pp. 1050–1060], and (iii) the simultaneous algebraic reconstruction technique (SART) of Andersen and Kak [*Ultrasonic Imaging*, 6 (1984), pp. 81–94] and generalizes them to linear inequalities. The first two algorithms are projection algorithms which use certain generalized oblique projections and diagonal weighting matrices which reflect the sparsity of the underlying matrix of the linear system. The previously reported experimental acceleration of the initial behavior of CAV and BICAV is thus complemented here by a mathematical study of the convergence of the algorithms.

Key words. block-iterative algorithms, component averaging (CAV), block-iterative CAV, simultaneous algebraic reconstruction technique, oblique projections, linear feasibility problem

AMS subject classifications. 90C25, 90C30

PII. S089547980138705X

1. Introduction.

Recently Censor, Gordon, and Gordon proposed and studied new iterative schemes for linear equations: In [7] the CAV (component averaging) method was presented as a simultaneous projection algorithm and in [8] BICAV was proposed as a block-iterative companion to CAV. In these methods the sparsity of the matrix is explicitly used when constructing the iteration formula. Using this new scaling we observed considerable improvement compared to traditionally scaled iteration methods. In [7] a proof of convergence was given for unity relaxation only, whereas no proofs at all were given for the block-iterative case [8].

The purpose of this paper is to describe a generalization to linear inequalities (with linear equations as a special case) of the Censor, Gordon, and Gordon schemes and study its convergence. It is shown that for the *consistent* case the block-iterative scheme (of which the fully simultaneous method is a special case) converges. For the inconsistent case we consider only linear equations and show that the simultaneous scheme converges to a weighted least squares solution. The treatment of the consistent case is based on our paper [6], in which an accelerated version of the fully simultaneous method with orthogonal projections for linear inequalities was proposed and studied.

*Received by the editors April 4, 2001; accepted for publication (in revised form) by M. Hanke October 16, 2001; published electronically May 15, 2002. This research was supported by research grants 293/97 and 592/00 from the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities.

http://www.siam.org/journals/simax/24-1/38705.html

†Department of Mathematics, University of Haifa, Mt. Carmel, Haifa 31905, Israel (yair@math.haifa.ac.il). The research of this author was supported by NIH grant HL-28438 at the Medical Image Processing Group (MIPG), Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA.

‡Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden (toelf@mai.liu.se). The research of this author was supported by the Swedish Natural Science Research Council under project M650-19981853/2000.

Recent relevant work of Byrne [5] and Jiang and Wang [19] is referred to at the end of Examples 7.1 and 7.2, respectively.

2. The CAV algorithm: Motivation and review. To motivate this work, let us consider linear equations and denote the hyperplanes

$$H_i := \left\{ x \in \mathbb{R}^n \mid \langle a^i, x \rangle = b_i \right\} \tag{2.1}$$

for $i = 1, 2, \ldots, m$, where $\langle \cdot, \cdot \rangle$ is the inner product and $a^i = (a^i_j)_{j=1}^n \in \mathbb{R}^n$, $a^i \neq 0$, and $b_i \in \mathbb{R}$ are given vectors and given real numbers, respectively. Then the *orthogonal (nearest Euclidean distance) projection* $P_i(z)$ of any $z \in \mathbb{R}^n$ onto H_i is

$$P_i(z) \;=\; z + \frac{b_i - \langle a^i, z \rangle}{\|a^i\|_2^2} a^i \,, \tag{2.2}$$

where $\| \cdot \|_2$ is the Euclidean norm.

In Cimmino's simultaneous projections method [11] (see also, e.g., Censor and Zenios [9, Algorithm 5.6.1] with relaxation parameters and with equal weights $w_i = 1/m$), the next iterate x^{k+1} is the average of the projections of x^k on the hyperplanes H_i, as follows.

ALGORITHM 2.1 (Cimmino).
Initialization: $x^0 \in \mathbb{R}^n$ *is arbitrary.*
Iterative Step: *Given* x^k*, compute*

$$x^{k+1} \;=\; x^k + \frac{\lambda_k}{m} \sum_{i=1}^{m} \left(P_i(x^k) - x^k \right), \tag{2.3}$$

where $\{\lambda_k\}_{k \geq 0}$ *are relaxation parameters.*

Expanding the iterative step (2.3) according to (2.2) produces, for every component $j = 1, 2, \ldots, n$,

$$x^{k+1}_j \;=\; x^k_j + \frac{\lambda_k}{m} \sum_{i=1}^{m} \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|_2^2} a^i_j \,. \tag{2.4}$$

When the $m \times n$ system matrix $A = (a^i_j)$ is sparse, only a relatively small number of the elements $\{a^1_j, a^2_j, \ldots, a^m_j\}$ in the jth column of A are nonzero, but in (2.4) the sum of their contributions is divided by the relatively large m. This observation led Censor, Gordon, and Gordon [7] to consider replacement of the factor $1/m$ in (2.4) by a factor that depends only on the *nonzero* elements in the set $\{a^1_j, a^2_j, \ldots, a^m_j\}$. For each $j = 1, 2, \ldots, n$, denote by s_j the number of nonzero elements of column j of the matrix A, and replace (2.4) by

$$x^{k+1}_j \;=\; x^k_j + \frac{\lambda_k}{s_j} \sum_{i=1}^{m} \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|_2^2} a^i_j \,. \tag{2.5}$$

Certainly, if A is sparse, then the s_j values will be much smaller than m. But this posed a theoretical difficulty. The iterative step (2.4) is a special case of

$$x^{k+1} \;=\; x^k + \lambda_k \sum_{i=1}^{m} w_i \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|_2^2} a^i \,, \tag{2.6}$$

where the fixed weights $\{w_i\}_{i=1}^m$ must be positive for all i and $\sum_{i=1}^m w_i = 1$. The attempt to use $1/s_j$ as weights in (2.5) does not fit into the scheme (2.6), unless one can prove convergence of the iterates of a fully simultaneous iterative scheme with component-dependent (i.e., j-dependent) weights of the form

$$x_j^{k+1} \;=\; x_j^k + \lambda_k \sum_{i=1}^m w_{ij} \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|_2^2} a_j^i \tag{2.7}$$

for all $j = 1, 2, \ldots, n$.

To derive a proof of convergence for (2.7), Censor, Gordon, and Gordon modified it further by replacing the orthogonal projections onto the hyperplanes H_i by certain *oblique projections* induced by appropriately defined weight matrices, as will be explained next. Consider a hyperplane $H := \{x \in \mathbb{R}^n \mid \langle a, x \rangle = b\}$, with $a = (a_j) \in \mathbb{R}^n$, $b \in \mathbb{R}$, and $a \neq 0$. Let G be an $n \times n$ symmetric positive definite matrix and let $\|x\|_G^2 := \langle x, Gx \rangle$ be the associated *ellipsoidal norm*; see, e.g., Bertsekas and Tsitsiklis [4, Proposition A.28]. Given a point $z \in \mathbb{R}^n$, the *oblique projection of z onto H with respect to G* is the unique point $P_H^G(z) \in H$ for which

$$P_H^G(z) \;=\; \arg\min \{\|x - z\|_G \mid x \in H\}\,. \tag{2.8}$$

Solving this minimization problem leads to

$$P_H^G(z) \;=\; z + \frac{b - \langle a, z \rangle}{\|a\|_{G^{-1}}^2} G^{-1} a\,, \tag{2.9}$$

where G^{-1} is the inverse of G. For $G = I$, the identity matrix, (2.9) yields the orthogonal projection of z onto H, as given by (2.2); see, e.g., Ben-Israel and Greville [3, section 2.6].

In order to consider oblique projections onto H with respect to a diagonal matrix $G = \mathrm{diag}(g_1, g_2, \ldots, g_n)$ for which some diagonal elements might be zero, the following definition is used.

DEFINITION 2.1 (see [7]). *Let $G = \mathrm{diag}(g_1, g_2, \ldots, g_n)$ with $g_j \geq 0$ for all $j = 1, 2, \ldots, n$, let $H = \{x \in \mathbb{R}^n \mid \langle a, x \rangle = b\}$ be a hyperplane with $a = (a_j) \in \mathbb{R}^n$ and $b \in \mathbb{R}$, and assume that $g_j = 0$ if and only if $a_j = 0$. The* generalized oblique projection *of a point $z \in \mathbb{R}^n$ onto H with respect to G is defined, for all $j = 1, 2, \ldots, n$, by*

$$(P_H^G(z))_j := \begin{cases} z_j + \dfrac{b - \langle a, z \rangle}{\sum_{\substack{l=1 \\ g_l \neq 0}}^n \frac{a_l^2}{g_l}} \cdot \dfrac{a_j}{g_j} & \text{if} \quad g_j \neq 0\,, \\ z_j & \text{if} \quad g_j = 0\,. \end{cases} \tag{2.10}$$

It is not difficult to verify that this $P_H^G(z)$ belongs to H, that it solves (2.8) if $\|x - z\|_G$ is replaced there by $\langle x - z, G(x - z) \rangle$, and that it is uniquely defined, although other solutions of (2.8) may exist due to the possibly zero-valued g_j's. This $P_H^G(z)$ reduces to (2.9) if $g_j \neq 0$ for all $j = 1, 2, \ldots, n$.

Consider next a set $\{G_i\}_{i=1}^m$ of real diagonal $n \times n$ matrices $G_i = \mathrm{diag}(g_{i1}, g_{i2}, \ldots, g_{in})$ with $g_{ij} \geq 0$ for all $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ and such that $\sum_{i=1}^m G_i = I$. Referring to the sparsity pattern of A, one needs the following definition.

DEFINITION 2.2 (see [7]). *A family $\{G_i\}_{i=1}^m$ of real diagonal $n \times n$ matrices with all diagonal elements $g_{ij} \geq 0$ and such that $\sum_{i=1}^m G_i = I$ is called* sparsity pattern

oriented (SPO) *with respect to an $m \times n$ matrix A if, for every $i = 1, 2, \ldots, m$, $g_{ij} = 0$ if and only if $a_j^i = 0$.*

The CAV algorithm of [7] combined three features: (i) Each orthogonal projection onto H_i in (2.3) was replaced by a generalized oblique projection with respect to G_i. (ii) The scalar weights $\{w_i\}$ in (2.6) were replaced by the diagonal weighting matrices $\{G_i\}$. (iii) The actual weights were set inversely proportional to the number of nonzero elements in each column, as motivated by the discussion preceding (2.5). The iterative step resulting from the first two features has the form

$$x^{k+1} \;=\; x^k + \lambda_k \sum_{i=1}^{m} G_i \left(P_{H_i}^{G_i}(x^k) - x^k \right), \tag{2.11}$$

or, equivalently, substituting from (2.10) for each $P_{H_i}^{G_i}$, one gets the following.

ALGORITHM 2.2 (diagonal weighting (DWE); see [7]).
Initialization: *$x^0 \in \mathbb{R}^n$ is arbitrary.*
Iterative Step: *Given x^k, compute x^{k+1} by using, for $j = 1, 2, \ldots, n$, the formula*

$$x_j^{k+1} \;=\; x_j^k + \lambda_k \sum_{\substack{i=1 \\ g_{ij} \neq 0}}^{m} \frac{b_i - \langle a^i, x^k \rangle}{\sum_{\substack{l=1 \\ g_{il} \neq 0}}^{n} \frac{(a_l^i)^2}{g_{il}}} \cdot a_j^i \,, \tag{2.12}$$

where $\{G_i\}_{i=1}^m$ is a given family of diagonal SPO (with respect to A) weighting matrices as in Definition 2.2, and $\{\lambda_k\}_{k \geq 0}$ are relaxation parameters.

Finally, the diagonal matrices $\{G_i\}_{i=1}^m$ are constructed in order to achieve the acceleration discussed above. Define

$$g_{ij} := \begin{cases} \dfrac{1}{s_j} & \text{if } a_j^i \neq 0 \,, \\ \\ 0 & \text{if } a_j^i = 0 \,. \end{cases} \tag{2.13}$$

With this particular SPO family of G_i's one obtains the CAV algorithm.

ALGORITHM 2.3 (component averaging (CAV); see [7]).
Initialization: *$x^0 \in \mathbb{R}^n$ is arbitrary.*
Iterative Step: *Given x^k, compute x^{k+1} by using, for $j = 1, 2, \ldots, n$, the formula*

$$x_j^{k+1} \;=\; x_j^k + \lambda_k \sum_{i=1}^{m} \frac{b_i - \langle a^i, x^k \rangle}{\sum_{l=1}^{n} s_l \left(a_l^i\right)^2} \cdot a_j^i \,, \tag{2.14}$$

where $\{\lambda_k\}_{k \geq 0}$ are relaxation parameters and $\{s_l\}_{l=1}^n$ are as defined above.

It was shown in [7] that Algorithm 2.2, with $\lambda_k = 1$ for all $k \geq 0$, generates sequences $\{x^k\}_{k \geq 0}$ which always converge regardless of the initial point x^0 and independently from the consistency or inconsistency of the underlying system $Ax = b$. Moreover, it always converges to a minimizer of a certain proximity function.

3. The block-iterative component averaging algorithm (BICAV). The basic idea of the block-iterative CAV (BICAV) algorithm is to break up the system $Ax = b$ into "blocks" of equations and treat each block according to the CAV methodology, passing cyclically over all the blocks. Throughout the following, T will be the number of blocks and, for $t = 1, 2, \ldots, T$, let the block of indices $B_t \subseteq \{1, 2, \ldots, m\}$ be an ordered subset of the form $B_t = \{i_1^t, i_2^t, \ldots, i_{m(t)}^t\}$, where $m(t)$ is the number of

elements in B_t. There is nothing preventing different blocks from containing common indices; we require, however, the following.

ASSUMPTION 3.1. *Every element of* $\{1, 2, \ldots, m\}$ *appears in at least one of the sets* B_t, $t = 1, 2, \ldots, T$.

For $t = 1, 2, \ldots, T$, let A_t denote the matrix formed by taking all the rows $\{a^i\}$ of A whose indices belong to the block of indices B_t, i.e.,

$$A_t := \begin{bmatrix} a^{i_1^t} \\ a^{i_2^t} \\ \vdots \\ a^{i_{m(t)}^t} \end{bmatrix}, \quad t = 1, 2, \ldots, T. \tag{3.1}$$

The iterative step of the BICAV algorithm, developed and experimentally tested by Censor, Gordon, and Gordon in [8], uses, for every block index $t = 1, 2, \ldots, T$, generalized oblique projections with respect to a family $\{G_i^t\}_{i \in B_t}$ of diagonal matrices which are SPO with respect to A_t. The same family is also used to perform the diagonal weighting. The resulting iterative step has the form

$$x^{k+1} \;=\; x^k + \lambda_k \sum_{i \in B_{t(k)}} G_i^{t(k)} \left(P_{H_i}^{G_i^{t(k)}}(x^k) - x^k \right), \tag{3.2}$$

where $\{t(k)\}_{k \geq 0}$ is a *control sequence* according to which the $t(k)$th block is chosen by the algorithm to be acted upon at the kth iteration, and thus, $1 \leq t(k) \leq T$ for all $k \geq 0$. The real numbers $\{\lambda_k\}_{k \geq 0}$ are user-chosen *relaxation parameters*. Substituting from (2.10) for each $P_{H_i}^{G_i^{t(k)}}$, one obtains the following.

ALGORITHM 3.1 (block-iterative diagonal weighting (BIDWE); see [8]).
Initialization: $x^0 \in \mathbb{R}^n$ *is arbitrary.*
Iterative Step: *Given* x^k, *compute* x^{k+1} *by using, for* $j = 1, 2, \ldots, n$, *the formula*

$$x_j^{k+1} \;=\; x_j^k + \lambda_k \sum_{\substack{i \in B_{t(k)} \\ g_{ij}^{t(k)} \neq 0}} \frac{b_i - \langle a^i, x^k \rangle}{\sum_{\substack{l=1 \\ g_{il}^{t(k)} \neq 0}}^{n} \frac{(a_l^i)^2}{g_{il}^{t(k)}}} \cdot a_j^i \,, \tag{3.3}$$

where, for each $t = 1, 2, \ldots, T$, $\{G_i^t\}_{i \in B_t}$ *is a given family of diagonal SPO (with respect to* A_t*) weighting matrices, as in Definition* 2.2, *the control sequence is cyclic, i.e.,* $t(k) = k \bmod T + 1$ *for all* $k \geq 0$, $\{\lambda_k\}_{k \geq 0}$ *are relaxation parameters, and* $G_i^t = \operatorname{diag}(g_{i1}^t, g_{i2}^t, \ldots, g_{in}^t)$.

Finally, in order to achieve the acceleration, the diagonal matrices $\{G_i^t\}_{i \in B_t}$ are constructed as in the original CAV algorithm [7], but with respect to each A_t. Let s_j^t be the number of nonzero elements $a_j^i \neq 0$ in the jth column of A_t and define

$$g_{ij}^t := \begin{cases} \dfrac{1}{s_j^t} & \text{if } a_j^i \neq 0, \\ \\ 0 & \text{if } a_j^i = 0. \end{cases} \tag{3.4}$$

It is easy to verify that, for each $t = 1, 2, \ldots, T$, $\sum_{i \in B_t} G_i^t = I$ holds for these matrices. With these particular SPO families of G_i^t's one obtains the block-iterative algorithm.

ALGORITHM 3.2 (block-iterative component averaging (BICAV); see [8]).
Initialization: $x^0 \in \mathbb{R}^n$ *is arbitrary.*
Iterative Step: *Given* x^k*, compute* x^{k+1} *by using, for* $j = 1, 2, \ldots, n$*, the formula*

$$x_j^{k+1} \;=\; x_j^k + \lambda_k \sum_{i \in B_{t(k)}} \frac{b_i - \langle a^i, x^k \rangle}{\sum_{l=1}^n s_l^{t(k)} (a_l^i)^2} \cdot a_j^i \,, \tag{3.5}$$

where $\{\lambda_k\}_{k \geq 0}$ *are relaxation parameters,* $\{s_l^t\}_{l=1}^n$ *are as defined above, and the control sequence is cyclic, i.e.,* $t(k) = k \bmod T + 1$ *for all* $k \geq 0$.

For the case $T = 1$ and $B_1 = \{1, 2, \ldots, m\}$, Algorithm 3.2 becomes fully simultaneous, i.e., it is the CAV algorithm of [7]. For $T = m$ and $B_t = \{t\},\ t = 1, 2, \ldots, m$, BICAV simply becomes the well-known ART (algebraic reconstruction technique) (see, e.g., Herman [17]), also known as Kaczmarz's algorithm [20] (see also, e.g., [9, Algorithm 5.4.3]).

4. The algorithmic schemes that cover the CAV and BICAV algorithms. We consider the system of linear inequalities

$$Ax \leq b, \tag{4.1}$$

where A is a real $m \times n$ matrix. We partition A into row blocks, precisely as done at the beginning of section 3. The right-hand-side vector b is partitioned similarly with b^t denoting those elements of b whose indices belong to the block of indices B_t,

$$b^t := \begin{bmatrix} b_{i_1^t} \\ b_{i_2^t} \\ \vdots \\ b_{i_{m(t)}^t} \end{bmatrix}, \quad t = 1, 2, \ldots, T. \tag{4.2}$$

The classical partitioning with fixed nonoverlapping blocks of equal sizes results by taking $m(t) = l,\ t = 1, 2, \ldots, T$, with $l \times T = m$. For each $i = 1, 2, \ldots, m$, the closed half-space

$$L_i := \{\, x \in \mathbb{R}^n \mid \langle a^i, x \rangle \leq b_i \,\} \tag{4.3}$$

has (2.1) as its bounding hyperplane. Define $L := \cap_{i=1}^m L_i$ and note that L is a closed convex set in $\mathbb{R}^n$. The task of finding a member of L, i.e., a solution of (4.1), is called the *linear feasibility problem*, which is a special case of the *convex feasibility problem*; see, e.g., Bauschke and Borwein [2] or [9, Chapter 5].

It is well known and easy to verify that the orthogonal projection $P_{L_i}(z)$ of a point $z \in \mathbb{R}^n$ onto L_i is

$$P_{L_i}(z) = z + c_i(z) a^i, \text{ where } \; c_i(z) = \min \left\{ 0, \frac{b_i - \langle a^i, z \rangle}{||a^i||_2^2} \right\}. \tag{4.4}$$

Note that if $z \notin L_i$, then $c_i(z) < 0$; otherwise $c_i(z) = 0$. Further define

$$I_t(z) := \{\, i \mid i_1^t \leq i \leq i_{m(t)}^t \text{ and } c_i(z) < 0 \,\} \tag{4.5}$$

as the set of indices of the half-spaces in the tth block which are violated by z. We also introduce diagonal matrices $\{D_t\}_{t=1}^T$, corresponding to the blocks $\{A_t\}_{i=1}^T$,

$$(D_t(z))_{jj} = \begin{cases} 1 & \text{if } j \in I_t(z), \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

Let $\{M_t\}_{t=1}^T$ be some given positive definite and symmetric matrices with nonnegative elements. Define

$$M_t(z) = D_t(z) M_t D_t(z), \quad t = 1, 2, \ldots, T. \tag{4.7}$$

If $\{x^k\}_{k\geq 0}$ is a sequence of vectors, then we use the following abbreviations: $c_i(x^k) \equiv c_i^k$, $I_t(x^k) \equiv I_t^k$, $D_t(x^k) \equiv D_t^k$, and $M_t(x^k) \equiv M_t^k$. We propose now the block-iterative algorithmic scheme which will work as an algorithmic structure that covers the CAV and BICAV algorithms and extends them from methods for solving linear equations to methods for solving the linear feasibility problem (i.e., both linear equations and linear inequalities). We use T to denote matrix transposition, but no ambiguity with the index T can arise.

ALGORITHM 4.1 (block-iterations for linear inequalities).
Initialization: $x^0 \in \mathbb{R}^n$ *is arbitrary.*
Iterative Step: *Given* x^k*, compute*

$$x^{k+1} \;=\; x^k + \lambda_k A_{t(k)}^T M_{t(k)}^k (b^{t(k)} - A_{t(k)} x^k), \tag{4.8}$$

where $\{\lambda_k\}_{k\geq 0}$ *are* relaxation parameters*, and* $\{t(k)\}_{k\geq 0}$ *is the* control sequence *governing which block is taken up at the* k*th iteration.*

For the choice $T = 1$ there is only one block, and we get the fully simultaneous version of Algorithm 4.1. In fact this method is then identical to Algorithm 2 of Censor and Elfving [6]. In addition to the cyclic control sequence, defined and used in Algorithms 3.1 and 3.2 above, we consider here two additional control sequences. These additional controls are problem-dependent. Denote by $d(x, L_i)$ the Euclidean distance between a point $x \in \mathbb{R}^n$ and the set L_i and define

$$\Phi(x) := \{\sup d(x, L_i) \,|\, 1 \leq i \leq m\}. \tag{4.9}$$

DEFINITION 4.1. (i) *We say that a sequence* $\{t(k)\}_{k\geq 0}$ *such that* $1 \leq t(k) \leq T$ *for all* $k \geq 0$ *is an* approximately remotest block control *sequence (with respect to the sequence* $\{x^k\}_{k\geq 0}$*, the family of sets* $\{L_i\}_{i=1}^m$*, and the blocks* $\{B_t\}_{t=1}^T$*) if, for every* $k \geq 0$*, there exists an* $i \in B_{t(k)}$ *such that*

$$\lim_{k\to\infty} d(x^k, L_i) = 0 \;\; \textit{implies that} \;\; \lim_{k\to\infty} \Phi(x^k) = 0. \tag{4.10}$$

(ii) *We say that a sequence* $\{t(k)\}_{k\geq 0}$ *such that* $1 \leq t(k) \leq T$ *for all* $k \geq 0$ *is a* remotest block control *sequence (with respect to the sequence* $\{x^k\}_{k\geq 0}$*, the family of sets* $\{L_i\}_{i=1}^m$*, and the blocks* $\{B_t\}_{t=1}^T$*) if, for every* $k \geq 0$*, there exists an* $i \in B_{t(k)}$ *such that*

$$\lim_{k\to\infty} d(x^k, L_i) = \Phi(x^k). \tag{4.11}$$

Every remotest block control is an approximately remotest block control. If all blocks consist of a single index, then these two definitions coincide with the definitions

of the *approximately remotest set control* and the *remotest set control*, respectively, of Gubin, Polyak, and Raik [16, section 1] (see also [9, section 5.1]). We will prove the next result in what follows.

THEOREM 4.1. *Assume that $L \neq \emptyset$ and that the relaxation parameters are restricted to*

$$0 < \epsilon \leq \lambda_k \leq (2-\epsilon)/\rho(A_{t(k)}^T M_{t(k)}^k A_{t(k)}) \text{ for all } k \geq 0, \tag{4.12}$$

where ϵ is an arbitrarily small but fixed constant and $\{M_t\}_{t=1}^T$ are given symmetric and positive definite matrices with nonnegative elements. If $\{t(k)\}_{k\geq 0}$ is a cyclic control or an approximately remotest block control, then any sequence $\{x^k\}_{k\geq 0}$, generated by Algorithm 4.1*, converges to a solution of the system* (4.1).

We also formulate the corresponding block-iterative algorithmic scheme for linear equalities

$$Ax = b. \tag{4.13}$$

ALGORITHM 4.2 (block-iterations for linear equalities).
Initialization: $x^0 \in \mathbb{R}^n$ *is arbitrary.*
Iterative Step: *Given x^k, compute*

$$x^{k+1} \;=\; x^k + \lambda_k A_{t(k)}^T M_{t(k)}(b^{t(k)} - A_{t(k)} x^k), \tag{4.14}$$

where $\{\lambda_k\}_{k\geq 0}$ are relaxation parameters*, and $\{t(k)\}_{k\geq 0}$ is the* control sequence *governing which block is taken up at the kth iteration.*

For this algorithm the following theorem will be proven in the next section.

THEOREM 4.2. *Assume that $H := \cap_{i=1}^m H_i \neq \emptyset$ and that the relaxation parameters are restricted to*

$$0 < \epsilon \leq \lambda_k \leq (2-\epsilon)/\rho(A_{t(k)}^T M_{t(k)} A_{t(k)}) \text{ for all } k \geq 0, \tag{4.15}$$

where ϵ is an arbitrarily small but fixed constant and $\{M_t\}_{t=1}^T$ are given symmetric and positive definite matrices with nonnegative elements. If $\{t(k)\}_{k\geq 0}$ is a cyclic control or an approximately remotest block control, then any sequence $\{x^k\}_{k\geq 0}$, generated by Algorithm 4.2*, converges to a solution of the system* (4.13)*. If, in addition, $x^0 \in R(A^T)$ (the range of A^T), then $\{x^k\}_{k\geq 0}$ converges to the solution of* (4.13)*, which has minimal Euclidean norm.*

5. Proofs of the convergence theorems. In proving Theorem 4.1 we use a convergence theory developed by Gubin, Polyak, and Raik [16]; see Bauschke and Borwein [2, Theorem 2.16 and Remark 2.17], which also contains a review and generalizations.

DEFINITION 5.1. *A sequence $\{x^k\}_{k\geq 0}$ is called* Fejér-monotone with respect to the set L *if, for every $x \in L$,*

$$||x^{k+1} - x||_2 \leq ||x^k - x||_2 \;\; \text{for all} \;\; k \geq 0. \tag{5.1}$$

It is easy to verify that every Fejér-monotone sequence is bounded. The convergence theory of Gubin, Polyak, and Raik applies to convex closed sets in general. For the sets L_i, defined here, their theorem is the following.

THEOREM 5.1. *Let $L = \cap_{i=1}^{m} L_i \neq \emptyset$. If, for a sequence $\{x^k\}_{k\geq 0}$, the following conditions hold, then $\lim_{k\to\infty} x^k = x^* \in L$:*

(i) *$\{x^k\}_{k\geq 0}$ is Fejér-monotone with respect to L, and*

(ii) $\lim_{k\to\infty} \Phi(x^k) = 0$.

Theorem 4.1 will be proved by establishing the conditions of Theorem 5.1. First we establish, in the next proposition, condition (i) of Theorem 5.1.

PROPOSITION 5.2. *Under the assumptions of Theorem 4.1, any sequence $\{x^k\}_{k\geq 0}$, generated by Algorithm 4.1, is Fejér-monotone with respect to L, provided that $x^k \notin L$ for all $k \geq 0$.*

Proof. We use the notation

$$r^{t(k),k} := b^{t(k)} - A_{t(k)}x^k \quad \text{and} \quad d^{t(k),k} = M^k_{t(k)} r^{t(k),k}. \tag{5.2}$$

Let $x \in L$ (i.e., $b - Ax \geq 0$), and define $e^k := x^k - x$. Then, by (4.8),

$$e^{k+1} = e^k + \lambda_k A^T_{t(k)} d^{t(k),k}. \tag{5.3}$$

It follows that

$$||e^{k+1}||_2^2 = ||e^k||_2^2 + \lambda_k^2 ||A^T_{t(k)} d^{t(k),k}||_2^2 + 2\lambda_k \langle A^T_{t(k)} d^{t(k),k}, e^k\rangle. \tag{5.4}$$

From $x \in L_{i_j^{t(k)}}$ we obtain (recall that $b_j^{t(k)}$ is the jth component of the block $b^{t(k)}$ of the vector b)

$$r_j^{t(k),k} = b_j^{t(k)} - \langle a^{i_j^{t(k)}}, x^k\rangle \geq -\langle a^{i_j^{t(k)}}, e^k\rangle, \quad j = 1, 2, \ldots, m(t(k)). \tag{5.5}$$

Hence we have for the last summand on the right-hand side of (5.4) that

$$\langle A^T_{t(k)} d^{t(k),k}, e^k\rangle = -\sum_{j=1}^{m(t(k))} d_j^{t(k),k} \langle -a^{i_j^{t(k)}}, e^k\rangle$$

$$\leq -\sum_{j=1}^{m(t(k))} d_j^{t(k),k} r_j^{t(k),k} = -\langle d^{t(k),k}, r^{t(k),k}\rangle, \tag{5.6}$$

provided that

$$d_j^{t(k),k} \leq 0 \quad \text{for } j = 1, 2, \ldots, m(t(k)) \text{ and for all } k \geq 0. \tag{5.7}$$

To see that (5.7) holds, observe that

$$d_j^{t(k),k} = \left(M^k_{t(k)} r^{t(k),k}\right)_j = \left(D^k_{t(k)} M_{t(k)} D^k_{t(k)} r^{t(k),k}\right)_j = \left(D^k_{t(k)}\right)_{jj} \sum_{s\in I^k_{t(k)}} m_s^{i_j^{t(k)}} r_s^{t(k),k}, \tag{5.8}$$

where $\{m_s^{i_j^{t(k)}}\}$ are the entries of the $i_j^{t(k)}$th row of $M_{t(k)}$, which are nonnegative by assumption, and observe that $r_s^{t(k),k} < 0$ whenever $s \in I^k_{t(k)}$.

Turning now to the second summand in the right-hand side of (5.4), we decompose the semidefinite matrix $M^k_{t(k)}$ as $M^k_{t(k)} = W^T W$ and use the well-known inequality

$$\langle Qy, y\rangle \leq \rho(Q)\langle y, y\rangle, \tag{5.9}$$

which holds for any symmetric and positive semidefinite matrix Q (where $\rho(Q)$ denotes the spectral radius of the matrix Q; see, e.g., Demmel [12, equation (5.2)]), to obtain

$$\begin{aligned} ||A_{t(k)}^T d^{t(k),k}||_2^2 &= \langle A_{t(k)}^T M_{t(k)}^k r^{t(k),k}, A_{t(k)}^T M_{t(k)}^k r^{t(k),k}\rangle \\ &= \langle M_{t(k)}^k A_{t(k)} A_{t(k)}^T M_{t(k)}^k r^{t(k),k}, r^{t(k),k}\rangle \\ &= \langle (W A_{t(k)} A_{t(k)}^T W^T) W r^{t(k),k}, W r^{t(k),k}\rangle \\ &\leq \rho(W A_{t(k)} A_{t(k)}^T W^T)\langle W r^{t(k),k}, W r^{t(k),k}\rangle \\ &= \rho(A_{t(k)}^T M_{t(k)}^k A_{t(k)})\langle d^{t(k),k}, r^{t(k),k}\rangle. \end{aligned} \tag{5.10}$$

Substituting (5.10) and (5.6) into (5.4), we get

$$||e^{k+1}||_2^2 \leq ||e^k||_2^2 + \lambda_k(\lambda_k \rho(A_{t(k)}^T M_{t(k)}^k A_{t(k)}) - 2)\langle d^{t(k),k}, r^{t(k),k}\rangle, \tag{5.11}$$

where $\langle d^{t(k),k}, r^{t(k),k}\rangle = \langle W r^{t(k),k}, W r^{t(k),k}\rangle \geq 0$. Now using (4.12), the desired conclusion $||e^{k+1}|| \leq ||e^k||$ follows. □

Note that if $I_{t(k)}^k = \emptyset$ (i.e., $A_{t(k)} x^k \leq b^{t(k)}$), then $D_{t(k)}^k = 0$, and hence $d^{t(k),k} = 0$ so that the second summand in the right-hand side of (5.11) disappears. The next proposition establishes condition (ii) of Theorem 5.1.

PROPOSITION 5.3. *Under the assumptions of Theorem* 4.1*, any sequence* $\{x^k\}_{k\geq 0}$*, generated by Algorithm* 4.1*, has the property*

$$\lim_{k\to\infty} \Phi(x^k) = 0. \tag{5.12}$$

Proof. Fejér-monotonicity, guaranteed by Proposition 5.2, implies that the sequence $\{||e^k||_2\}_{k\geq 0}$ is monotonically decreasing, and thus converging. It follows then from (5.11) that

$$\lim_{k\to\infty} \langle d^{t(k),k}, r^{t(k),k}\rangle = 0. \tag{5.13}$$

But

$$\langle d^{t(k),k}, r^{t(k),k}\rangle = \langle M_{t(k)}^k r^{t(k),k}, r^{t(k),k}\rangle = \langle M_{t(k)} D_{t(k)}^k r^{t(k),k}, D_{t(k)}^k r^{t(k),k}\rangle, \tag{5.14}$$

and thus

$$\lim_{k\to\infty} D_{t(k)}^k r^{t(k),k} = 0. \tag{5.15}$$

Using (4.4),

$$\left(D_{t(k)}^k r^{t(k),k}\right)_j = c_{i_j^{t(k)}}^k ||a^{i_j^{t(k)}}||_2^2, \quad j = 1, 2, \ldots, m(t(k)), \tag{5.16}$$

leads to

$$\begin{aligned} d(x^k, L_{i_j^{t(k)}}) &= ||P_{L_{i_j^{t(k)}}}(x^k) - x^k||_2 \\ &= ||c_{i_j^{t(k)}}^k a^{i_j^{t(k)}}||_2 = \left|\left(D_{t(k)}^k r^{t(k),k}\right)_j\right| / ||a^{i_j^{t(k)}}||_2 \end{aligned} \tag{5.17}$$

for all $j = 1, 2, \ldots, m(t(k))$. This shows, by (5.15), that

$$\lim_{k\to\infty} d(x^k, L_i) = 0 \text{ for all } i \in B_{t(k)}. \tag{5.18}$$

If $\{t(k)\}_{k\geq 0}$ is an approximately remotest block control, then the required result follows directly from (5.18) and Definition 4.1(i) and Assumption 3.1. For a cyclic control we argue as follows. From (4.8) and (4.7) we get

$$\begin{aligned} ||x^{k+1} - x^k||_2 &= \lambda_k ||A^T_{t(k)} D^k_{t(k)} M_{t(k)} D^k_{t(k)} r^{t(k),k}||_2 \\ &\leq \lambda_k ||A^T_{t(k)} D^k_{t(k)} M^{1/2}_{t(k)}||_2 \cdot ||M^{1/2}_{t(k)}||_2 ||D^k_{t(k)} r^{t(k),k}||_2. \end{aligned} \tag{5.19}$$

Therefore, using (4.12) and the fact that, for any matrix Q, it is true that $\rho(Q^TQ) = ||Q^T||_2^2$ (see, e.g., Demmel [12, Fact 9, p. 23]), we obtain

$$||x^{k+1} - x^k||_2 \leq \theta_1 \theta_2^{-1} ||D^k_{t(k)} r^{t(k),k}||_2, \tag{5.20}$$

where

$$\theta_1 := 2\max\{||M_i^{1/2}||_2 \mid 1 \leq i \leq T\} \text{ and } \theta_2 := \max\{||A_i^T D_i^k M_i^{1/2}||_2 \mid 1 \leq i \leq T\}. \tag{5.21}$$

The max in the expression of θ_2 exists and is independent of k because of the way these matrices were defined. If $\theta_2 = 0$, then, by (4.8), $x^{k+1} = x^k$. If, on the other hand, $\theta_2 \neq 0$, then θ_2 is bounded away from zero and, thus, (5.15) and (5.20) yield

$$\lim_{k\to\infty} ||x^{k+1} - x^k||_2 = 0. \tag{5.22}$$

Let $\epsilon > 0$ be such that for all $k \geq K$, we have $||x^{k+1} - x^k||_2 \leq \epsilon/T$. To reach the required conclusion (5.12) we look at $d(x^k, L_i) = ||P_{L_i}(x^k) - x^k||_2$ and observe that if $i \in B_{t(k)}$, then (5.18) shows that $||P_{L_i}(x^k) - x^k||_2 \leq \epsilon$ for all $k \geq K$. Otherwise, if $i \notin B_{t(k)}$, the cyclicality of $\{t(k)\}_{k\geq 0}$ guarantees that there exists a τ such that $1 \leq \tau < T$ and $i \in B_{t(k+\tau)}$. Then,

$$\begin{aligned} d(x^k, L_i) &= ||x^k - P_{L_i}(x^k)||_2 \leq ||x^k - P_{L_i}(x^{k+\tau})||_2 \\ &\leq ||x^k - x^{k+\tau}||_2 + ||x^{k+\tau} - P_{L_i}(x^{k+\tau})||_2 \\ &\leq ||x^k - x^{k+1}||_2 + \cdots + ||x^{k+\tau-1} - x^{k+\tau}||_2 + ||x^{k+\tau} - P_{L_i}(x^{k+\tau})||_2 \\ &\leq (T-1)(\epsilon/T) + \epsilon = \epsilon \end{aligned} \tag{5.23}$$

for all $k \geq K$. Therefore, $\Phi(x^k) \leq \epsilon$ for all $k \geq K$, and, using Assumption 3.1, the result follows. □

So, we see that the last two propositions, combined with Theorem 4.1, imply the truth of Theorem 4.1.

Proof of Theorem 4.2. Theorem 4.2 follows from Theorem 4.1. To simplify the discussion we deal only with the case that the weight matrices $\{M_t\}$ are positive diagonal matrices. This assumption actually holds in all three examples given in section 7. The general case can be proved along lines similar to the following argument. Any equation $\langle a^i, x\rangle = b_i$ can be written as a pair of inequalities $\langle a^i, x\rangle \leq b_i$ and $\langle -a^i, x\rangle \leq -b_i$. Now for a given linear system $Ax = b$, where $A \in \mathbf{R}^{m\times n}$, and given diagonal weight matrices $\{M_t\}$ we construct the inequalities $\tilde{A}x \leq \tilde{b}$ as follows:

$$\tilde{a}^{2i-1} = a^i, \quad \tilde{a}^{2i} = -a^i, \quad \tilde{b}_{2i-1} = b_i, \quad \tilde{b}_{2i} = -b_i, \qquad i = 1, 2, \ldots, m. \tag{5.24}$$

Denoting the (i,j)th element of a matrix A by $(A)_{i,j}$, we also set

$$\left(\tilde{M}_t\right)_{2i-1,2i-1} = \left(\tilde{M}_t\right)_{2i,2i} = (M_t)_{i,i} \quad \text{for all } i = 1, 2, \ldots, m. \tag{5.25}$$

Recall that $\tilde{M}_t^k = D_t(x^k)\tilde{M}_t D_t(x^k)$, where the matrix $D_t(z)$ is defined in (4.6). Then, for any x^k, one can verify that

$$\tilde{A}_{t(k)}^T \tilde{M}_{t(k)}^k (\tilde{b}^{t(k)} - \tilde{A}_{t(k)} x^k) = A_{t(k)}^T M_{t(k)} (b^{t(k)} - A_{t(k)} x^k) \tag{5.26}$$

so that the two iteration formulas (4.8) and (4.14) generate the same sequence of iterates, provided they are initialized with the same vector. It is also true, for any x^k, that

$$\rho(\tilde{A}_{t(k)}^T \tilde{M}_{t(k)}^k \tilde{A}_{t(k)}) = \rho(A_{t(k)}^T M_{t(k)} A_{t(k)}); \tag{5.27}$$

hence Theorem 4.2 follows. □

6. The inconsistent case. When there is just one block, i.e., $t = T = 1$, the resulting methods are fully simultaneous. We consider here the inconsistent case behavior only for linear equations. Let $M_1 = M$, $c = A^T M b$, and $\Gamma = A^T M A$. Then the iteration (4.14) can be written as

$$x^{k+1} = x^k + \lambda_k (c - \Gamma x^k). \tag{6.1}$$

This is the nonstationary *Richardson iteration method*; cf. Young [24, p. 361]. We observe that $c \in R(\Gamma)$ (the range of Γ) and, if we assume that $\hat{x}$ satisfies $c = \Gamma \hat{x}$, then $\hat{x} = \arg\min ||Ax - b||_M$ (with $||x||_M^2 = \langle x, Mx \rangle$). Let $u^k = \hat{x} - x^k$ and note that, with $v^k = c - \Gamma x^k$, it is true that $v^k = \Gamma u^k$. It follows that

$$u^k = \prod_{j=0}^{k-1} (I - \lambda_j \Gamma) u^0. \tag{6.2}$$

Assume first that Γ is a positive definite matrix. Then any sequence $\{x^k\}_{k \geq 0}$ generated by Algorithm 4.2, as given by (6.1), is convergent for any x^0 if and only if

$$\lim_{k \to \infty} \prod_{j=0}^{k-1} (I - \lambda_j \Gamma) = 0. \tag{6.3}$$

Since $|| \prod_{j=0}^{k-1} (I - \lambda_j \Gamma)||_2 \leq \prod_{j=0}^{k-1} \rho(I - \lambda_j \Gamma)$, it follows that any sequence $\{x^k\}_{k \geq 0}$, generated by Algorithm 4.2, as given by (6.1), converges to a weighted least squares solution if $0 < \epsilon \leq \lambda_k \leq (2 - \epsilon)/\rho(\Gamma)$. In case Γ is only positive semidefinite we have a similar result. All of these observations lead to the following theorem.

THEOREM 6.1. *Assume that M is a positive definite matrix. If $0 < \epsilon \leq \lambda_k \leq (2-\epsilon)/\rho(A^T M A)$ for all $k \geq 0$, where ϵ is an arbitrarily small but fixed constant, then any sequence $\{x^k\}_{k \geq 0}$, generated by Algorithm 4.2, as given by (6.1), converges to a weighted least squares solution $\hat{x} = \arg\min ||Ax - b||_M$. If, in addition, $x^0 \in R(A^T)$, then $\{x^k\}_{k \geq 0}$ converges to the unique solution of minimal Euclidean norm among all weighted least squares solutions.*

The proof of Theorem 6.1 can essentially be found in, e.g., Eggermont, Herman, and Lent [13, p. 44]; see also Elfving [14, p. 4].

We do not give a proof of convergence for the case of linear inequalities. We note, however, that a variant of Algorithm 4.1 for $T = 1$ (Cimmino's method; see Example 7.3 below) was shown to converge locally for the inconsistent case by Iusem and De Pierro [18] and to converge globally in that case by Combettes [10].

7. Applications. In this section we will consider only diagonal matrices $M_t = \operatorname{diag}\{\mu_j^t \mid j = 1, 2, \ldots, m(t)\}$ with positive diagonal elements. For such diagonal matrices let

(7.1)
$$W_t := A_t^T M_t A_t \text{ for all } t = 1, 2, \ldots, T \quad \text{and} \quad W_{t(k)}^k := A_{t(k)}^T M_{t(k)}^k A_{t(k)} \text{ for all } k \geq 0,$$

and note the expansions

$$W_{t(k)}^k = \sum_{j \in I_{t(k)}^k} \mu_j^{t(k)} a^{i_j^{t(k)}} \left(a^{i_j^{t(k)}}\right)^T, \quad W_{t(k)} = \sum_{j=1}^{m(t(k))} \mu_j^{t(k)} a^{i_j^{t(k)}} \left(a^{i_j^{t(k)}}\right)^T. \tag{7.2}$$

Hence the iterative step of Algorithm 4.1 takes the form

$$x^{k+1} = x^k + \lambda_k \sum_{j \in I_{t(k)}^k} \mu_j^{t(k)} \left(b_j^{t(k)} - \langle a^{i_j^{t(k)}}, x^k \rangle\right) a^{i_j^{t(k)}}, \tag{7.3}$$

and the iterative step of Algorithm 4.2 becomes

$$x^{k+1} = x^k + \lambda_k \sum_{j=1}^{m(t(k))} \mu_j^{t(k)} \left(b_j^{t(k)} - \langle a^{i_j^{t(k)}}, x^k \rangle\right) a^{i_j^{t(k)}}. \tag{7.4}$$

Also note that, by (7.2), for all $k \geq 0$,

$$\rho\big(W_{t(k)}^k\big) \leq \rho\big(W_{t(k)}\big). \tag{7.5}$$

In the following examples we show that several algorithms, including the BICAV and simultaneous algebraic reconstruction technique (SART) algorithms, are in fact special cases of the algorithmic schemes studied in the previous sections.

Example 7.1. The BICAV (Algorithm 3.2) and CAV (Algorithm 2.3) are both algorithms for equalities and of the form (7.4) with

$$\mu_j^{t(k)} = \frac{1}{||a^{i_j^{t(k)}}||^2_{S_{t(k)}}} = \frac{1}{\sum_{\nu=1}^n s_\nu^{t(k)} \left(a_\nu^{i_j^{t(k)}}\right)^2}, \quad j = 1, 2, \ldots, m(t(k)). \tag{7.6}$$

Here $\{t(k)\}_{k \geq 0}$ is the control sequence, $s_\nu^{t(k)}$ is the number of nonzero elements in the νth column of the block $A_{t(k)}$, and $S_{t(k)} := \operatorname{diag}\{s_\nu^{t(k)} \mid \nu = 1, 2, \ldots, n\}$. We first study the upper bound on the relaxation parameters for CAV, i.e., allowing one block only so that $t = T = 1$ and $m(1) = m$; cf. (3.1). The following result (Lemma 7.1) is due to Dr. Arnold Lent [22] (see the acknowledgments at the end of this paper).

LEMMA 7.1. *Let $t = T = 1$ and $m(1) = m$, let $M := \operatorname{diag}\{\mu_j \mid j = 1, 2, \ldots, m\}$ with $\mu_j = \mu_j^1$ obtained from (7.6) for $t = t(k) = 1$, and let $A_1 = A$, $s_\nu^1 = s_\nu$, $S_1 = S$, and $W := A^T M A$. Then $\rho(W) \leq 1$.*

Proof. Let a_j^i be the element in the ith row and jth column of A and write, by (7.6),

$$(\mu_i)^{-1} = \sum_{j=1}^n s_j \left(a_j^i\right)^2, \quad i = 1, 2, \ldots, m. \tag{7.7}$$

Let (λ, v) be an eigenpair (i.e., eigenvalue and eigenvector) of W so that $A^T MAv = \lambda v$ or $AA^T MAv = \lambda M^{-1} MAv$, or, with $w := MAv$, $AA^T w = \lambda M^{-1} w$. Hence $||A^T w||_2^2 = \lambda w^T M^{-1} w$ or, in component form, switching the order of summations and using (7.7),

(7.8)
$$||A^T w||_2^2 = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_j^i w_i \right)^2 = \lambda \sum_{i=1}^{m} w_i^2 \left(\sum_{j=1}^{n} s_j \left(a_j^i \right)^2 \right) = \lambda \sum_{j=1}^{n} s_j \left(\sum_{i=1}^{m} w_i^2 \left(a_j^i \right)^2 \right).$$

From Cauchy's inequality we have

$$\left(\sum_{i=1}^{m} a_j^i w_i \right)^2 \leq s_j \sum_{i=1}^{m} w_i^2 \left(a_j^i \right)^2, \tag{7.9}$$

and by summing both sides of (7.9) over j and comparing with (7.8), one finds that $\lambda \leq 1$. ∎

Remark 7.1. The critical estimate is (7.9). Let a, w, and e be three vectors of equal length. Denote by $z = a * w$ componentwise multiplication, i.e., $z_j = a_j w_j$ for all j. Further, let $e_j = 0$ if $z_j = 0$, and let $e_j = 1$ otherwise. Then

$$\langle a, w \rangle^2 = \langle e, z \rangle^2 \leq ||e||_2^2 \cdot ||z||_2^2 \leq s||z||_2^2, \tag{7.10}$$

where s is the number of nonzero elements in the vector a.

By applying Lemma 7.1 to each block A_t, $t = 1, 2, \ldots, T$, we obtain the following.

COROLLARY 7.1. *Let* $M_{t(k)} = \text{diag}\{\mu_j^{t(k)} \mid j = 1, 2, \ldots, m(t(k))\}$, $k \geq 0$, *with* $\mu_j^{t(k)}$ *obtained from* (7.6), *and let* $W_{t(k)} = A_{t(k)}^T M_{t(k)} A_{t(k)}$. *Then* $\rho(W_{t(k)}) \leq 1$ *for all* $k \geq 0$.

The next theorems establish the convergence of the BICAV algorithm in the consistent case for linear equations and linear inequalities, respectively, with relaxation parameters within the interval $[\epsilon, 2 - \epsilon]$.

THEOREM 7.1 (BICAV for linear equalities). *Let* $0 < \epsilon \leq \lambda_k \leq 2 - \epsilon$ *for all* $k \geq 0$, *where* ϵ *is an arbitrarily small but fixed constant. If the system* (4.13) *is consistent, then any sequence* $\{x^k\}_{k \geq 0}$, *generated by Algorithm* 3.2 *(BICAV), converges to a solution of the system* (4.13). *If, in addition,* $x^0 \in R(A^T)$, *then* $\{x^k\}_{k \geq 0}$ *converges to the solution of* (4.13), *which has minimal Euclidean norm.*

Proof. The proof follows from Theorem 4.2 and Corollary 7.1. ∎

THEOREM 7.2 (BICAV for linear inequalities). *Let* $0 < \epsilon \leq \lambda_k \leq 2 - \epsilon$ *for all* $k \geq 0$, *where* ϵ *is an arbitrarily small but fixed constant. If the system* (4.1) *is consistent, then any sequence* $\{x^k\}_{k \geq 0}$, *generated by Algorithm* 4.1, *with* $M_{t(k)} = \text{diag}\{\mu_j^{t(k)} \mid j = 1, 2, \ldots, m(t(k))\}$ *and* $\{\mu_j^{t(k)}\}$ *given by* (7.6), *converges to a solution of the system* (4.1).

Proof. The proof follows from Theorem 4.1, Corollary 7.1, and (7.5). ∎

The next theorem shows that any sequence $\{x^k\}_{k \geq 0}$, generated by the fully simultaneous Algorithm 2.3 (CAV), converges to a weighted least squares solution of the system of equations $Ax = b$, regardless of its consistency, for relaxation parameters in the interval $[\epsilon, 2 - \epsilon]$. Only the case of unity relaxation, i.e., $\lambda_k = 1$ for all $k \geq 0$, was proven by Censor, Gordon, and Gordon in [7], where CAV was first proposed and experimented with.

THEOREM 7.3 (CAV for linear equalities). *If $0 < \epsilon \leq \lambda_k \leq 2 - \epsilon$ for all $k \geq 0$, where ϵ is an arbitrarily small but fixed constant, then any sequence $\{x^k\}_{k\geq 0}$, generated by Algorithm 2.3 (CAV for linear equations), converges to a weighted least squares solution with weight matrix $M_1 = M_{CAV} = \text{diag}\{1/||a^i||_S^2 \mid i = 1, 2, \ldots, m\}$ and with $S = \text{diag}\{s_j \mid j = 1, 2, \ldots, n\}$, where s_j is the number of nonzero elements in the jth column of A. If, in addition, $x^0 \in R(A^T)$, then $\{x^k\}_{k\geq 0}$ converges to the unique solution of minimal Euclidean norm among all weighted least squares solutions.*

Proof. The proof follows from Theorem 6.1 and Lemma 7.1. □

Note that Theorems 7.1 and 7.2 assumed cyclic control of the blocks, as formulated in Algorithm 3.2; however, due to the analysis presented here, we may also allow approximately remotest block control of the blocks (by Theorems 4.2 and 6.1). Recently, and independently of our work, Byrne [5] derived convergence results analogous to Theorems 7.1 and 7.3, but only for the cyclic control and without explicit consideration of weighting. He also used Lent's result as expressed above in Lemma 7.1.

Example 7.2. The *simultaneous algebraic reconstruction technique (SART)* was proposed by Andersen and Kak [1] for solving the large and very sparse systems of linear equations arising from a fully discretized model of transmission computerized tomography problems; see also Kak and Slaney [21, section 7.4]. We show that a simplified version of SART falls within the convergence analysis presented here. First recall that the 1-*norm* of a vector $x \in \mathbb{R}^n$ is $||x||_1 = \sum_{j=1}^n |x_j|$ and that the induced matrix norm of an $m \times n$ matrix A is $||A||_1 = \max\{\sum_{i=1}^m |a_{ij}| \mid j = 1, 2, \ldots, n\}$. Let $a_c^{l,t}$ be the lth column of A_t. Then the iterative step of the original SART algorithm for linear equalities [1, equation (32)] (see also [23, equation (4)]) is

$$x_l^{k+1} = x_l^k + \frac{\lambda_k}{||a_c^{l,t(k)}||_1} \sum_{j=1}^{m(t(k))} \frac{b_j^{t(k)} - \langle a^{i_j^{t(k)}}, x^k \rangle}{||a^{i_j^{t(k)}}||_1} a_l^{i_j^{t(k)}}, \quad l = 1, 2, \ldots, n. \tag{7.11}$$

Note that in (7.11) it is tacitly assumed that all blocks A_t have nonzero columns. The formula (7.11) is slightly more general than the original algorithm in [1] since it allows (i) a relaxation parameter λ_k, (ii) a more flexible row-partitioning (originally the matrix was partitioned into nonoverlapping row blocks, where each block corresponds to all equations in one tomographic scan direction), (iii) arbitrary sign of the matrix elements (originally only nonnegative elements were considered), and (iv) apart from the cyclic control of blocks also the remotest block control.

We first note that (7.11) can be written in matrix-vector form, using our previous notation, as

$$x^{k+1} = x^k + \lambda_k D_{t(k)} A_{t(k)}^T M_{t(k)} (b^{t(k)} - A_{t(k)} x^k), \tag{7.12}$$

where

$$D_{t(k)} = \text{diag}\{1/||a_c^{l,t(k)}||_1 \mid l = 1, 2, \ldots, n\} \tag{7.13}$$

and

$$M_{t(k)} = \text{diag}\{1/||a^{i_j^t}||_1 \mid j = 1, 2, \ldots, m(t(k))\}. \tag{7.14}$$

We will not, however, analyze this iteration here. Instead, we consider a simplified version which fits into the class of methods (4.8) and (4.14), respectively. Let a_c^l be the lth column of A and put $D = \text{diag}\{1/||a_c^l||_1 \mid l = 1, 2, \ldots, n\}$.

Replacing $D_{t(k)}$ by D in (7.12) we get

$$x^{k+1} = x^k + \lambda_k D A_{t(k)}^T M_{t(k)} (b^{t(k)} - A_{t(k)} x^k). \tag{7.15}$$

We will call the method which uses the iterative step (7.15) *block simplified SART (BSSART)*. The method is a scaled version of (7.4). To see this we put

$$y^k = D^{-1/2} x^k \quad \text{and} \quad \bar{A}_{t(k)} = A_{t(k)} D^{1/2}, \tag{7.16}$$

which converts (7.15) into

$$y^{k+1} = y^k + \lambda_k \bar{A}_{t(k)}^T M_{t(k)} (b^{t(k)} - \bar{A}_{t(k)} y^k), \tag{7.17}$$

which is of the form (7.4) (or equivalently (4.14)). Next observe that with $W_{t(k)} = D^{1/2} A_{t(k)}^T M_{t(k)} A_{t(k)} D^{1/2}$, we have

$$\rho(W_{t(k)}) = \rho(A_{t(k)}^T M_{t(k)} A_{t(k)} D) \leq ||A_{t(k)}^T M_{t(k)}||_1 \cdot ||A_{t(k)} D||_1 = 1. \tag{7.18}$$

It follows from Theorem 4.2 that y^k converges to some y^*. Since, by (7.16), every row of A is postmultiplied by $D^{1/2}$, we also conclude that $AD^{1/2} y^* = b$. Then, using (7.16),

$$\lim_{k \to \infty} x^k = D^{1/2} y^* = x^*. \tag{7.19}$$

Hence $Ax^* = b$.

Now consider BSSART adapted to inequalities, i.e., the iterative step

$$x^{k+1} = x^k + \lambda_k D A_{t(k)}^T M_{t(k)}^k (b^{t(k)} - A_{t(k)} x^k). \tag{7.20}$$

It is clear, using (7.5) and Theorem 4.2, that the above analysis also holds for the iteration (7.20). Hence the following companion results to Theorems 7.1 and 7.2 hold.

THEOREM 7.4 (BSSART for linear equalities). *Let $0 < \epsilon \leq \lambda_k \leq 2 - \epsilon$ for all $k \geq 0$, where ϵ is an arbitrarily small but fixed constant. If the system* (4.13) *is consistent, then any sequence $\{x^k\}_{k \geq 0}$, generated by the iterative step* (7.15) *(BSSART), converges to a solution of the system* (4.13).

THEOREM 7.5 (BSSART for linear inequalities). *Let $0 < \epsilon \leq \lambda_k \leq 2 - \epsilon$ for all $k \geq 0$, where ϵ is an arbitrarily small but fixed constant. If the system* (4.1) *is consistent, then any sequence $\{x^k\}_{k \geq 0}$, generated by the iterative step* (7.20) *(BSSART for inequalities), converges to a solution of the system* (4.1).

When $T = 1$, SART (7.12) and BSSART (7.15) coincide and can be written

$$x^{k+1} = x^k + \lambda_k D A^T M (b - A x^k), \tag{7.21}$$

with $M = \text{diag}\{1/||a^j||_1 \mid j = 1, 2, \ldots, m\}$. Using the corresponding transformations as in (7.16),

$$y^k = D^{-1/2} x^k \quad \text{and} \quad \bar{A} = AD^{1/2}, \tag{7.22}$$

we find that

$$y^{k+1} = y^k + \lambda_k \bar{A}^T M (b - \bar{A} y^k). \tag{7.23}$$

It follows from Theorem 6.1, and by using (as above) the fact that $\rho(A^TMAD) \leq 1$, that

$$\lim_{k\to\infty} y^k = y^* \quad \text{such that } ||\bar{A}y^* - b||_M \quad \text{is minimal.} \tag{7.24}$$

But $\lim_{k\to\infty} x^k = D^{1/2}y^* = x^*$ so that x^* minimizes $||Ax - b||_M$. Also, by using

$$||y^*||_2 = ||D^{-1/2}D^{1/2}y^*||_2 = ||x^*||_{D^{-1}}, \tag{7.25}$$

it follows that x^* has minimal D^{-1}-norm. Hence the following result holds.

THEOREM 7.6. *If $0 < \epsilon \leq \lambda_k \leq 2 - \epsilon$ for all $k \geq 0$, where ϵ is an arbitrarily small but fixed constant, then any sequence $\{x^k\}_{k\geq 0}$, generated by Algorithm (7.21), converges to a weighted least squares solution with weight matrix $M = \text{diag}\{1/||a^i||_1 \mid i = 1, 2, \ldots, m\}$. If, in addition, $x^0 \in R(DA^T)$, then the limit point has minimal D^{-1}-norm.*

No proof of convergence was given in [1] or has, to the best of our knowledge, been published elsewhere since then. Recently, however, and independently of our work, Jiang and Wang [19] have also derived, under the additional assumption that the elements of the matrix A are nonnegative, Theorem 7.6.

Example 7.3. Block-Cimmino methods for linear equations and linear inequalities can also be viewed as special cases of Algorithms 4.1 and 4.2. To see this we define

$$\mu_j^{t(k)} = \frac{\theta_{i_j^{t(k)}}}{||a^{i_j^{t(k)}}||_2^2}, \quad j = 1, 2, \ldots, m(t(k)), \tag{7.26}$$

where $\theta_{i_j^{t(k)}} > 0$ and $\sum_{j=1}^{m(t(k))} \theta_{i_j^{t(k)}} = 1$. It follows, using (7.2), that $\rho(W_{t(k)}) = ||W_{t(k)}||_2 \leq \sum_{j=1}^{m(t(k))} \theta_{i_j^{t(k)}} = 1$ and that

$$\rho(W_{t(k)}^k) = ||W_{t(k)}^k||_2 \leq \sum_{j\in I_{t(k)}^k} \theta_{i_j^{t(k)}} \leq 1. \tag{7.27}$$

Therefore, also in this example, we may conclude convergence just as in Theorems 7.1, 7.2, and 7.3 with $M_1 = M_{CIM} = \text{diag}\{\theta_i/||a^i||_2^2 \mid i = 1, 2, \ldots, m\}$ in Theorem 7.3. The geometric interpretation of this scaling is as follows. By (2.2),

$$P_{H_i}(x) - x = (b_i - \langle a^i, x\rangle)\frac{a^i}{||a^i||_2^2}, \tag{7.28}$$

so that

$$\sum_{i=1}^m \theta_i ||P_{H_i}(x) - x||_2^2 = \sum_{i=1}^m \frac{\theta_i(b_i - \langle a^i, x\rangle)^2}{||a^i||_2^2} = ||b - Ax||_{M_{CIM}}^2. \tag{7.29}$$

Cimmino's original algorithm for linear equations [11] is purely simultaneous ($T = 1$), i.e., of the form (2.4). An interesting detail is that $\lambda_k = 2$ is used by Cimmino, and for this a special convergence analysis is furnished. We also remark that for *inequalities* the requirement on the relaxation parameters can be relaxed, using (7.27), to

$$0 < \epsilon \leq \lambda_k \leq \frac{2 - \epsilon}{\sum_{j\in I_{t(k)}^k} \theta_{i_j^{t(k)}}}. \tag{7.30}$$

In fact, the choice $\lambda_k = 2/\sum_{j\in I^k_{t(k)}} \theta_{i_j^{t(k)}}$ is also allowed but requires a special analysis, which appears, for the fully simultaneous case $T = 1$, assuming consistency, in Censor and Elfving [6]. See also Bauschke and Borwein [2, Remark 6.48] for a correction. A similar analysis can be done also for the block-iterative case. Iusem and De Pierro [18] have shown that this method (with $T = 1$) also converges (locally) for the inconsistent case and generalized it to closed convex sets in $\mathbb{R}^n$. A generalization to global convergence in infinite dimensional Hilbert spaces was done by Combettes [10].

We finally mention that if all block sizes are equal to 1 ($m(t) = 1$) and linear equations are considered, then we get the algebraic reconstruction technique (ART) of Gordon, Bender, and Herman [15], also known as Kaczmarz's method. For more on the history of this method and many of its variants, see, for example, Herman [17] and Censor and Zenios [9].

Acknowledgments. We thank Charles Byrne, Dan Gordon, Rachel Gordon, Arnold Lent, Robert Lewitt, and Samuel Matej for their enlightening comments and discussions on this work. Special thanks are due to Charles Byrne for sharing with us his paper [5] on this topic and to Ming Jiang and Ge Wang for making available to us their paper [19]. Lemma 7.1 is due to Dr. Arnold Lent from AT&T Labs, who made it available to us during a blackboard discussion with the first author and Professor Charles Byrne from the Department of Mathematical Sciences at the University of Massachusetts at Lowell. Part of this work was done during visits of Y. Censor to the Department of Mathematics of the University of Linköping, Linköping, Sweden. The support and hospitality of Professor Åke Björck, head of the Numerical Analysis Group there at the time, are gratefully acknowledged.

REFERENCES

[1] A.H. Andersen and A.C. Kak, *Simultaneous algebraic reconstruction technique (SART): A superior implementation of the ART algorithm*, Ultrasonic Imaging, 6 (1984), pp. 81–94.

[2] H.H. Bauschke and J.M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[3] A. Ben-Israel and T.N.E. Greville, *Generalized Inverses: Theory and Applications*, John Wiley & Sons, New York, 1974.

[4] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[5] C. Byrne, *Notes on Block-Iterative or Ordered Subset Methods for Image Reconstruction*, Technical Report, 2000.

[6] Y. Censor and T. Elfving, *New methods for linear inequalities*, Linear Algebra Appl., 42 (1982), pp. 199–211.

[7] Y. Censor, D. Gordon, and R. Gordon, *Component averaging: An efficient iterative parallel algorithm for large and sparse unstructured problems*, Parallel Comput., 27 (2001), pp. 777–808.

[8] Y. Censor, D. Gordon, and R. Gordon, *BICAV: An inherently parallel algorithm for sparse systems with pixel-dependent weighting*, IEEE Trans. Medical Imaging, 20 (2001), pp. 1050–1060.

[9] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, New York, 1997.

[10] P.L. Combettes, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., SP-42 (1994), pp. 2955–2966.

[11] G. Cimmino, *Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari*, Ric. Sci. Progr. Tecn. Econom. Naz., 1 (1938), pp. 326–333.

[12] J.W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[13] P.P.B. Eggermont, G.T. Herman, and A. Lent, *Iterative algorithms for large partitioned linear systems, with applications to image reconstruction*, Linear Algebra Appl., 40 (1981), pp. 37–67.

[14] T. Elfving, *Block-iterative methods for consistent and inconsistent linear equations*, Numer. Math., 35 (1980), pp. 1–12.
[15] R. Gordon, R. Bender, and G.T. Herman, *Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography*, J. Theoret. Biology, 29 (1970), pp. 471–481.
[16] L.G. Gubin, B.T. Polyak, and E.V. Raik, *The method of projections for finding the common point of convex sets*, Comput. Math. Math. Phys., 7 (1967), pp. 1–24.
[17] G.T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.
[18] A.N. Iusem and A.R. De Pierro, *Convergence results for an accelerated nonlinear Cimmino algorithm*, Numer. Math., 49 (1986), pp. 367–378.
[19] M. Jiang and G. Wang, *Convergence of the simultaneous algebraic reconstruction technique (SART)* (invited talk), in Proceedings of the 35th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2001.
[20] S. Kaczmarz, *Angenäherte Auflüsungen von Systemen linearer Gleichungen*, Bulletin de l'Académie Polonaise des Sciences et Lettres, A35 (1937), pp. 355–357.
[21] A.C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, Classics Appl. Math. 33, SIAM, Philadelphia, PA, 2001; also available online from http://www.slaney.org/pct/pct-toc.html.
[22] A. Lent, *Private communication.*
[23] K. Mueller, R. Yagel, and J.J. Wheller, *Anti-aliased three-dimensional cone-beam reconstruction of low-contrast objects with algebraic methods*, IEEE Trans. Medical Imaging, 18 (1999), pp. 519–537.
[24] D. Young, *Iterative Solution of Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 59–67

A FINER ASPECT OF EIGENVALUE DISTRIBUTION OF SELFADJOINT BAND TOEPLITZ MATRICES*

PETER ZIZLER†, ROB A. ZUIDWIJK‡, KEITH F. TAYLOR§, AND SHIGERU ARIMOTO¶

Abstract. The asymptotics of eigenvalues of Toeplitz operators has received a lot of attention in the mathematical literature and has been applied in several disciplines. This paper describes two such application disciplines and provides refinements of existing asymptotic results using new methods of proof. The following result is typical: Let $T(\varphi)$ be a selfadjoint band limited Toeplitz operator with a (real valued) symbol φ, which is a nonconstant trigonometric polynomial. Consider finite truncations $T_n(\varphi)$ of $T(\varphi)$, and a finite union of finite intervals of real numbers E. We prove a refinement of the Szegö asymptotic formula

$$\lim_{n\to\infty} \frac{N_n(E)}{n} = \frac{1}{2\pi} m(F).$$

Indeed, we show that

$$N_n(E) - \frac{1}{2\pi} m(F) n = O(1).$$

Here $m(F)$ denotes the measure of $F = \varphi^{-1}(E)$ on the unit circle, and $N_n(E)$ denotes the number of eigenvalues of $T_n(\varphi)$ inside E. We prove similar results for singular values of general Toeplitz operators involving a refinement of the Avram–Parter theorem.

Key words. Toeplitz matrix, eigenvalue distribution, Szegö formula, Avram–Parter theorem

AMS subject classifications. 15A18, 47A10, 47A58, 47B35

PII. S089547989834915X

1. Introduction.

The eigenvalue distribution of Toeplitz matrices and operators has been a fascinating and abundant source of topics of mathematical inquiries. The prominent monographs [9] and [10] respectively provide extensive analysis of Toeplitz matrices and operators. Among key historical papers are [11] (on operators), [19] (on matrices), and [20] (on block matrices). A comprehensive account on the theory involved is provided in [12].

From the interdisciplinary point of view, the above field also possesses a considerable potential, especially in terms of a wide range of applications and connections to disciplines outside mathematics. In the first part of this section, two application areas (see (I) and (II) below) which have motivated the authors to study the asymptotics of Toeplitz eigenvalues are addressed.

In the second part of the introduction, the mathematical contribution of this paper to the asymptotics of eigenvalues and singular values shall be outlined. We conclude the introduction with some clarification on notation used in the paper.

*Received by the editors December 10, 1998; accepted for publication (in revised form) by A. Bunse-Gerstner November 9, 2001; published electronically June 12, 2002.
http://www.siam.org/journals/simax/24-1/34915.html

†Department of Mathematics, Engineering and Physics, Mount Royal College, 4825 Richard Road S.W., Calgary, AB, T3E 6K6, Canada (zizler@mtroyal.ab.ca).

‡Rotterdam School of Management, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (rzuidwijk@fbk.fac.eur.nl).

§Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road, Saskatoon, SK, S7N 5E6, Canada (taylor@snoopy.usask.ca).

¶Department of Chemistry, University of Saskatchewan, 110 Science Place, Saskatoon, SK, S7N 5C9, Canada, and Institute for Fundamental Chemistry, 34-4 Takano-Nishihiraki-cho, Sakyo-ku, Kyoto 606-8103, Japan (arimoto@duke.usask.ca).

(I) Vast uncharted regions lie between mathematics and chemistry on the map of science. Communication across the border of these disciplines is still generally sporadic and uncoordinated, despite modern trends of cross-disciplinary investigations in each of these fields. In the present work, we have for the first time formed a linkage between

(i) the mathematical branch of Toeplitz matrices, and
(ii) the "repeat space theory" (RST) in theoretical chemistry, which originates in the study of the zero-point vibrational energies of hydrocarbons having repeating identical moieties [1].

Namely, in dealing with Toeplitz matrices in the proof of our main theorem, Theorem 2.3, we have recalled, sharpened, and applied a mathematical technique developed in the RST (to estimate quantum boundary effects in polymeric molecules). It is also remarkable that the sharpened technique in the proof of Theorem 2.3 can be applied to molecular problems by embedding the technique into the RST. In our opinion, researchers investigating in areas of (i) and (ii) can mutually benefit. The reader who is interested in cross-disciplinary mathematical investigations in chemistry is referred to [1, 2, 3, 4, 5, 6] and references therein, where he can find the genesis of the RST (in conjunction with experimental chemistry) and a variety of applications of the RST to quantum, thermodynamic, and structural chemistry.

Sequences of band circulant matrices are called "alpha sequences" and play a dominant role in the RST [1, 2, 3, 4, 5, 6]. The band circulant matrix associated with a band Toeplitz matrix has been used in the proof of the present paper based on the approach and technique originally developed in the RST, especially in [1] and [6]. Further, we remark that the study of asymptotic spectra of band Toeplitz matrices in [7] arises from the analysis of difference approximations of partial differential equations and that in [7] the asymptotic spectra of the band Toeplitz matrix and its associated circulant matrix were studied.

(II) The asymptotics of eigenvalues of Toeplitz operators is an important issue in the study of time-frequency localization of signals. Essentially time- and band-limited functions can be studied by means of Toeplitz matrix eigenvalue asymptotics; see [15, 17]. Quite recently, these results have been used in the analysis of seismic records [16].

It is hoped that the present work provides researchers of the asymptotic eigenvalue distribution of Toeplitz matrices with a fresh insight into the theme and that it contributes to dissolving the traditional boundary between the mathematical branch of Toeplitz matrices and other research areas such as quantum chemistry of molecules having repeating identical moieties, and time-frequency localization of (seismic) signals.

We shall now discuss the asymptotics of eigenvalues of Toeplitz matrices in further detail. Let φ be a real valued continuous function defined on the unit circle $\boldsymbol{T} = \{z \in \boldsymbol{C} : |z| = 1\}$. The Fourier coefficients of φ are given by $\varphi_k = (2\pi i)^{-1} \int_{\boldsymbol{T}} \varphi(z) z^{-k-1}\, dz$, $k \in \boldsymbol{Z}$. The corresponding Toeplitz operator $T(\varphi) = (\varphi_{i-j})_{i,j \in \boldsymbol{Z}^+}$ is selfadjoint and its finite truncations $T_n(\varphi) = (\varphi_{i-j})_{i,j=0}^{n-1}$ are Hermitian matrices. The spectrum of the operator $T(\varphi)$ coincides with the closed interval $\mathcal{I} = \{\varphi(z) : z \in \boldsymbol{T}\}$. In particular, the norm of $T(\varphi)$ is given by $\|T(\varphi)\| = \sup\{|\varphi(z)| : z \in \boldsymbol{T}\}$.

Moreover, the eigenvalues of the truncations $T_n(\varphi)$ are contained in the closed interval $\mathcal{I}$; see, for example, section 5.2b in [13] and Proposition 2.17 in [9]. However, much more can be said about the eigenvalue distribution of $T_n(\varphi)$. As a first step,

we mention that the asymptotic behavior of the eigenvalues is expressed by the well-known Szegö formula (cf. Theorem 5.2 in [13] and Theorem 5.10 in [9]): If f is a continuous function on the closed interval $\mathcal{I}$, and if $\{\lambda_{i,n}\}_{i=1}^n$ are the eigenvalues of $T_n(\varphi)$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(\lambda_{i,n}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))\, d\theta.$$

Moreover, if φ is smooth, e.g., $C^{1+\varepsilon}$ with $\varepsilon > 0$ and f is analytic in an open neighborhood of $\mathcal{I}$, then one has a second order formula

$$\frac{1}{n} \sum_{i=1}^{n} f(\lambda_{i,n}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))\, d\theta + \frac{E_f(\varphi)}{n} + o\left(\frac{1}{n}\right)$$

with some completely identified constant $E_f(\varphi)$ (see [20] or Theorem 5.6 in [9]).

We shall now make the further assumption that φ is actually a nonconstant trigonometric polynomial of degree $r \geq 1$, i.e., $\varphi(z) = \sum_{k=-r}^{r} \varphi_k z^k$. In this manner, $T(\varphi)$ becomes a selfadjoint band limited Toeplitz operator. Let E denote a finite union of compact intervals on the real line and χ_E be its characteristic function. Let m denote the Lebesgue measure on the unit circle. Since $m(\varphi^{-1}(\partial E)) = 0$, the Szegö formula can be extended to $f = \chi_E$ (see [22]) and we get

$$\lim_{n\to\infty} \frac{N_n(E)}{n} = \frac{1}{2\pi} m(F), \tag{1.1}$$

where $N_n(E)$ denotes the number of eigenvalues of $T_n(\varphi)$ in the set E and $F = \varphi^{-1}(E)$. The purpose of this paper is to sharpen the formula for the case of band limited Toeplitz operators. Indeed, (1.1) states that $N_n(E) - \frac{1}{2\pi} m(F)n = o(n)$. The main result of this paper refines this asymptotic result to $N_n(E) - \frac{1}{2\pi} m(F)n = O(1)$. In addition to such results for eigenvalues of selfadjoint Toeplitz operators, we prove similar results for singular values of general Toeplitz operators.

In the remaining part of this paper, $\operatorname{tr} A$ denotes the trace of the square matrix A. The space $BV(\mathcal{I})$ consists of functions of bounded variation on the closed interval $\mathcal{I} = [a, b]$. For such a function f, there exists a constant $V > 0$, such that for each partition $a = x_0 < x_1 < \cdots < x_m = b$, we get

$$\sum_{j=1}^{m} |f(x_j) - f(x_{j-1})| \leq V.$$

The minimum $V > 0$ which satisfies this condition is called the total variation of f on $\mathcal{I}$ and is denoted by $V_{\mathcal{I}}(f)$. If the natural domain of f contains $\mathcal{I}$ and $f|_{\mathcal{I}}$ is of bounded variation on $\mathcal{I}$, then $V_{\mathcal{I}}(f) = V_{\mathcal{I}}(f|_{\mathcal{I}})$. If $g : \mathbf{T} \to \mathbf{R}$, let $f(t) = g(e^{it})$, for $t \in \mathbf{R}$, and let $V_{\mathbf{T}}(g) = V_{[-\pi,\pi]}(f)$. Denote the eigenvalues of a Hermitian $n \times n$ matrix H by $\lambda_1(H) \leq \lambda_2(H) \leq \cdots \leq \lambda_n(H)$. The singular values of an arbitrary complex $m \times n$ matrix M are equal to the eigenvalues of the Hermitian matrix $(M^*M)^{1/2}$ and labeled so that $\sigma_1(M) \leq \sigma_2(M) \leq \cdots \leq \sigma_n(M)$. The spectral norm $\|M\|$ of M is equal to $\sigma_n(M)$.

2. Refined eigenvalue asymptotics. In this section, we prove a number of estimates which lead to the refined asymptotics result in Corollary 2.5. This corollary involves the characteristic function χ_E, while the preparatory results are stated for

general functions of bounded variation. First, we state Theorem 2.1 from [6]. For convenience of the reader and for reference later on, we include the proof.

THEOREM 2.1. *Consider the integers $1 \leq r < n$ and let $K = \{k_1, \dots, k_r\}$ be a subset of $\{1, 2, \dots, n\}$ consisting of r distinct elements. Define $L = \{1, 2, \dots, n\} \backslash K$. Let M and M' be $n \times n$ Hermitian matrices such that the ijth entries of M and M' coincide for all $(i, j) \in L \times L$, i.e., such that*

$$(M - M')_{ij} = 0$$

for all $(i, j) \in L \times L$. Consider a closed interval $\mathcal{I} = [a, b]$ which contains all the eigenvalues of both M and M'. Then we have

$$|\operatorname{tr} f(M) - \operatorname{tr} f(M')| \leq r V_{\mathcal{I}}(f)$$

for all $f \in BV(\mathcal{I})$.

Proof. Case 1: $r = 1$. We may and do assume that $K = \{n\}$, since this situation can be achieved by transforming $M - M'$ by means of a permutation similarity. Let M_0 denote the $(n-1) \times (n-1)$ matrix given by $(M_{ij})_{i,j=1}^{n-1}$. Observe that $M_0 = (M'_{ij})_{i,j=1}^{n-1}$. If we write $\lambda_0 = a$, $\lambda_j = \lambda_j(M_0)$ for $j = 1, \dots, n-1$, and $\lambda_n = b$, then by the Sturmian separation theorem [14], we get

$$\lambda_{j-1} \leq \lambda_j(M) \leq \lambda_j, \quad \lambda_{j-1} \leq \lambda_j(M') \leq \lambda_j, \quad j = 1, \dots, n.$$

Therefore, we arrive at

$$|\operatorname{tr} f(M) - \operatorname{tr} f(M')| = \left| \sum_{j=1}^{n} \{ f(\lambda_j(M)) - f(\lambda_j(M')) \} \right|$$

$$\leq \sum_{j=1}^{n} |f(\lambda_j(M)) - f(\lambda_j(M'))| \leq V_{\mathcal{I}}(f).$$

Case 2: $r > 1$. As in the first part of the proof, we may and do assume that K has a specific form, say $K = \{n - r + 1, \dots, n\}$. Define $n \times n$ Hermitian matrices $M^{(0)}, M^{(1)}, \dots, M^{(r)}$ such that $M^{(0)} = M$, $M^{(r)} = M'$ and such that the pairs $M^{(\nu-1)}, M^{(\nu)}$ for $\nu = 1, \dots, r$ each satisfy the conditions of Case 1. This can be achieved by setting $(0 \leq \nu \leq r)$

$$M_{ij}^{(\nu)} = \begin{cases} M_{ij}, & 1 \leq i, j \leq n - \nu, \\ M'_{ij}, & n - \nu < i \leq n \text{ or } n - \nu < j \leq n. \end{cases}$$

Let $[\widetilde{a}, \widetilde{b}] = \widetilde{\mathcal{I}} \supseteq \mathcal{I} = [a, b]$ be an interval which contains all eigenvalues of $M^{(\nu)}$ for $\nu = 1, \dots, r-1$, and let $\widetilde{f}$ be the extension of f to $\widetilde{\mathcal{I}}$ given by

$$\widetilde{f}(t) = \begin{cases} f(a), & \widetilde{a} \leq t \leq a, \\ f(t), & a \leq t \leq b, \\ f(b), & b \leq t \leq \widetilde{b}. \end{cases}$$

We have obtained

$$|\operatorname{tr} f(M) - \operatorname{tr} f(M')| \leq \sum_{\nu=1}^{r} \left| \sum_{j=1}^{n} \left\{ \widetilde{f}(\lambda_j(M^{(\nu-1)})) - \widetilde{f}(\lambda_j(M^{(\nu)})) \right\} \right|$$

$$\leq r V_{\widetilde{\mathcal{I}}}(\widetilde{f}) = r V_{\mathcal{I}}(f). \qquad \square$$

We now state and prove two new results for functions of bounded variation f and apply them to the characteristic function χ_E in the corollaries. If n is a positive integer, let P_n denote the cyclic shift $n \times n$ matrix with $(P_n)_{ij} = 1$ if $i - j \equiv 1 \bmod n$ and 0 otherwise. Let $A_n = A_n(\varphi) = \sum_{k=-r}^{r} \varphi_k P_n^k$.

THEOREM 2.2. *For any $f \in BV(\mathcal{I})$ and any positive integer n, we have*

(i) $|\mathrm{tr}(f(T_n)) - \mathrm{tr}(f(A_n))| \le rV_{\mathcal{I}}(f)$,

(ii) $|\mathrm{tr}(f(A_n)) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))\,d\theta| \le 2rV_{\mathcal{I}}(f)$,

(iii) $|\mathrm{tr}(f(T_n)) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))\,d\theta| \le 3rV_{\mathcal{I}}(f)$.

Proof. (i) If $n \le r$, then $\mathrm{tr}(f(T_n)) - \mathrm{tr}(f(A_n))$ is just the sum of the differences of the values of f at n pairs of points from $\mathcal{I}$. Thus,

$$|\mathrm{tr}(f(T_n)) - \mathrm{tr}(f(A_n))| \le nV_{\mathcal{I}}(f) \le rV_{\mathcal{I}}(f).$$

If $r < n$, then we make use of some auxiliary matrices. We have already introduced P_n in order to define the circulant matrix A_n. Further, for $|k| < n$, let $S_n(k)$ denote the $n \times n$ matrix with $(S_n(k))_{ij} = 1$ if $i - j = k$ and 0 otherwise. Let $S_n(k) = 0$ if $|k| \ge n$. Clearly $((P_n)^k)_{ij} = (S_n(k))_{ij}$ for $1 \le i, j \le n - |k|$. Since $T_n = T_n(\varphi) = \sum_{k=-r}^{r} \varphi_k S_n(k)$, we get

$$T_n - A_n = \sum_{k=-r}^{r} \varphi_k (S_n(k) - P_n^k).$$

Since $(T_n)_{ij} = (A_n)_{ij}$ for $1 \le i, j \le n - r$, we get, by Theorem 2.1,

$$|\mathrm{tr}(f(T_n)) - \mathrm{tr}(f(A_n))| \le rV_{\mathcal{I}}(f).$$

(ii) Let $h(\theta) = f(\varphi(e^{i\theta}))$ for $\theta \in \mathbf{R}$. Then

$$\mathrm{tr}(f(A_n)) = \sum_{j=1}^{n} h\left(\frac{2\pi j}{n}\right).$$

This implies

$$\begin{aligned}\left|\mathrm{tr}(f(A_n)) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))\,d\theta\right| &= \left|\sum_{j=1}^{n} h\left(\frac{2\pi j}{n}\right) - \frac{n}{2\pi}\int_{-\pi}^{\pi} h(\theta)\,d\theta\right| \\ &\le \frac{n}{2\pi}\sum_{j=1}^{n}\int_{\frac{2\pi(j-1)}{n}}^{\frac{2\pi j}{n}} \left|h\left(\frac{2\pi j}{n}\right) - h(\theta)\right| d\theta \\ &\le \frac{n}{2\pi}\sum_{j=1}^{n}\int_{\frac{2\pi(j-1)}{n}}^{\frac{2\pi j}{n}} V_{[\frac{2\pi(j-1)}{n}, \frac{2\pi j}{n}]}(h)\,d\theta = V_{[0,2\pi]}(h).\end{aligned}$$

Now, let $u(\theta) = \varphi(e^{i\theta})$. Since φ is a nonconstant trigonometric polynomial of degree r, u' has at least 2 and at most $2r$ distinct roots in $[0, 2\pi)$. Let $\theta_1 < \theta_2 < \cdots < \theta_l$ be the roots of u' in $[0, 2\pi)$. Then

$$V_{[0,2\pi]}(h) = V_{[\theta_1, \theta_1 + 2\pi]}(f \circ u)$$

$$= V_{[\theta_1,\theta_2]}(f \circ u) + V_{[\theta_2,\theta_3]}(f \circ u) + \cdots + V_{[\theta_l, \theta_1+2\pi]}(f \circ u) \le lV_{\mathcal{I}}(f).$$

It follows that

$$\left|\mathrm{tr}(f(A_n)) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))d\theta\right| \leq 2rV_{\mathcal{I}}(f).$$

Now (iii) follows immediately from (i) and (ii). □

THEOREM 2.3. *If $f \in BV(\mathcal{I})$ and if n is any positive integer, then*

$$\left|\sum_{i=1}^{n} f(\lambda_{i,n}) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta}))\,d\theta\right| \leq rV_{\mathcal{I}}(f) + V_{\boldsymbol{T}}(f\circ\varphi) \leq 3rV_{\mathcal{I}}(f).$$

Proof. If we apply Theorem 2.2 (i) to the setting of this theorem, we get

$$|\mathrm{tr}(f(T_n)) - \mathrm{tr}(f(A_n))| \leq rV_{\mathcal{I}}(f),$$

and the proof of Theorem 2.2 (ii) yields

$$\left|\mathrm{tr}(f(A_n)) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta})\,d\theta\right| \leq V_{\boldsymbol{T}}(f\circ\varphi) \leq 2rV_{\mathcal{I}}(f).$$

This, together with $\mathrm{tr}(f(T_n)) = \sum_{i=1}^{n} f(\lambda_{i,n})$, provides

$$\left|\sum_{i=1}^{n} f(\lambda_{i,n}) - \frac{n}{2\pi}\int_{-\pi}^{\pi} f(\varphi(e^{i\theta})\,d\theta\right| \leq rV_{\mathcal{I}}(f) + V_{\boldsymbol{T}}(f\circ\varphi) \leq 3rV_{\mathcal{I}}(f). \qquad \square$$

The following two corollaries are easy consequences of Theorem 2.3. We leave it to the reader to check the necessary minor details. Let E be a subset of $\mathbf{R}$ that is a finite union of compact intervals and $F = \varphi^{-1}(E)$ be the corresponding subset of $\mathbf{T}$. Note that if E is a union of N compact intervals and $\mathcal{I}$ is an interval in $\mathbf{R}$, then $V_{\mathcal{I}}(\chi_E) \leq 2N$.

COROLLARY 2.4. *Let T be a band limited selfadjoint Toeplitz operator with the symbol φ, a real-valued trigonometric polynomial of degree $r \geq 1$. Then*

$$\left|N_n(E) - \frac{1}{2\pi}m(F)n\right| \leq rV_{\mathcal{I}}(\chi_E) + V_{\boldsymbol{T}}(\chi_F) \leq 3rV_{\mathcal{I}}(\chi_E)$$

for every $n \geq 1$.

COROLLARY 2.5. *Let T be a band limited selfadjoint Toeplitz operator with the symbol φ, a real-valued trigonometric polynomial of degree $r \geq 1$. Then*

$$N_n(E) - \frac{n}{2\pi}m(F) \quad = \quad O(1).$$

3. Singular values. The results proved in the previous section for the eigenvalues of selfadjoint band Toeplitz matrices can easily be generalized to results concerning the singular values of arbitrary band Toeplitz matrices, although the constants in the new estimates are slightly worse. In this section, the main steps of the proofs are outlined. The analogue of Theorem 2.1 reads as follows.

THEOREM 3.1. *For $1 \leq r < n$, let $K = \{k_1, \ldots, k_r\}$ be a subset of $\{1, 2, \ldots, n\}$ having exactly r elements, and put $L = \{1, 2, \ldots, n\}\backslash K$. Let M and M' be two complex $n\times n$ matrices such that $(M - M')_{ij} = 0$ for all $(i, j) \in L\times L$. If $\mathcal{I} = [a, b]$ is a*

closed interval which contains the singular values of both M and M' and if $f \in BV(\mathcal{I})$, then

$$\sum_{j=1}^{n} |f(\sigma_j(M)) - f(\sigma_j(M'))| \le 2rV_{\mathcal{I}}(f).$$

Proof. We can proceed as in the proof of Theorem 2.1. The only difference is that we need to replace the Sturmian separation theorem by the following interlacing result (see, e.g., [8, pp. 81–82]). Let $A = (a_{ij})_{i,j=1}^{n}$ be a complex $n \times n$ matrix and $B = (a_{ij})_{i,j=1}^{n-1}$ be the $(n-1) \times (n-1)$ principal submatrix. Then

$$0 \le \sigma_1(A) \le \sigma_2(B),$$

$$\sigma_{j-1}(B) \le \sigma_j(A) \le \sigma_{j+1}(B), \quad j = 2, \dots, n-2,$$

and $\|B\| \le \|A\|$. If we abbreviate $\sigma_j = \sigma_j(M_0)$ for $j = 1, \dots, n-1$ (notation as in Theorem 2.1), then in the case of $r = 1$, we get

$$\sum_{j=1}^{n} |f(\sigma_j(M)) - f(\sigma_j(M'))|$$

$$\le V_{[a,\sigma_2]}(f) + \sum_{j=2}^{n-2} V_{[\sigma_{j-1},\sigma_{j+1}]}(f) + V_{[\sigma_{n-2},b]}(f) \le V_{[a,\sigma_{n-1}]}(f) + V_{[\sigma_1,b]}(f) \le 2V_{[a,b]}(f).$$

The case of $r > 1$ is dealt with in the same fashion as in Theorem 2.1. □

THEOREM 3.2. *Let ψ be a nonconstant trigonometric polynomial of degree $r \ge 1$, let $\mathcal{I} = [0, \|\psi\|_\infty]$, and let $g \in BV(\mathcal{I})$. Then for all $n \ge 1$,*

$$\left| \sum_{j=1}^{n} g(\sigma_j(T_n(\psi))) - \frac{n}{2\pi} \int_{-\pi}^{\pi} g(|\psi(e^{i\theta})|)\, d\theta \right| \le 2rV_{\mathcal{I}}(g) + V_{\boldsymbol{T}}(g \circ |\psi|) \le 6rV_{\mathcal{I}}(g).$$

Proof. The singular values of the circulant matrix A_n introduced in the proof of Theorem 2.2 (ii) are given by $|\psi(2\pi ij/n)|$ $(j = 1, \dots, n)$. Consequently, the reasoning of the proof of Theorem 2.2 (i), in conjunction with Theorem 3.1, gives

$$\left| \sum_{j=1}^{n} g(\sigma_j(T_n(\psi))) - \frac{n}{2\pi} \int_{-\pi}^{\pi} g(|\psi(e^{i\theta})|)\, d\theta \right| \le 2rV_{\mathcal{I}}(g) + V_{[0,2\pi]}(g \circ |\psi|).$$

Since $|\psi(e^{i\theta})|^2$ is a trigonometric polynomial of degree $2r$, we obtain as in the proof of Theorem 2.2 (ii) that $|\psi(e^{i\theta})|$ has at most $4r$ local extrema in $[0, 2\pi)$, whence $V_{[0,2\pi)}(g \circ |\psi|) \le 4rV_{\mathcal{I}}(g)$. This implies the assertion. □

While Theorem 2.3 is a refined version of Szegö's formula, Theorem 3.2 may be regarded as a refinement of the Avram–Parter theorem, which states that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(\sigma_i(T_n(\psi))) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(|\psi(e^{i\theta})|\, d\theta$$

if, for example, ψ is continuous on $\boldsymbol{T}$ and g is continuous on the range of $|\psi|$ (see [9] and [18] and the references therein). The counterpart of Corollaries 2.4 and 2.5 for singular values is as follows.

COROLLARY 3.3. *Let ψ be a trigonometric polynomial of degree $r \geq 1$, let $E \subset \boldsymbol{R}$ be a finite union of compact intervals, and let $F = \{t \in \boldsymbol{T} : \quad |\psi(t)| \in E\}$. If $N_n(E)$ denotes the number of singular values of $T_n(\psi)$ in E, then*

$$\left| N_n(E) - \frac{n}{2\pi} m(F) \right| \leq 2rV_{\mathcal{I}}(\chi_E) + V_{\boldsymbol{T}}(\chi_F) \leq 6rV_{\mathcal{I}}(\chi_E)$$

for every $n \geq 1$. In particular,

$$N_n(E) - \frac{n}{2\pi} m(F) \quad = \quad O(1).$$

Acknowledgment. The authors are indebted to professor A. Böttcher for pointing out the extension of the main results from eigenvalues to singular values for arbitrary band Toeplitz matrices, which has resulted in section 3. The authors would like to acknowledge professor A. Böttcher for his extensive comments, which have resulted in the description of refined singular value asymptotics of general Toeplitz matrices.

REFERENCES

[1] S. ARIMOTO, *Preliminary considerations towards a theoretical foundation of the zero point energy additivity rules*, Phys. Lett. A, 113 (1985), pp. 126–132.

[2] S. ARIMOTO AND K. FUKUI, *Fundamental mathematical chemistry, interdisciplinary research in fundamental mathematical chemistry and generalized repeat space,* IFC Bulletin (1998), pp. 7–13; PDF full text downloadable from http://duke.usask.ca/∼arimoto/.

[3] S. ARIMOTO, K. FUKUI, P. ZIZLER, K.F. TAYLOR, AND P.G. MEZEY, *Structural analysis of certain linear operators representing chemical network systems via the existence and uniqueness theorems of spectral resolution.* V, Int. J. Quantum Chem., 74 (1999), pp. 633–644.

[4] S. ARIMOTO, M. SPIVAKOVSKY, H. OHNO, P. ZIZLER, K.F. TAYLOR, T. YAMABE, AND P.G. MEZEY, *Structural analysis of certain linear operators representing chemical network systems via the existence and uniqueness theorems of spectral resolution.* VI, Int. J. Quantum Chem., 84 (2001), pp. 389–400.

[5] S. ARIMOTO, M. SPIVAKOVSKY, P. ZIZLER, R.A. ZUIDWIJK, K.F. TAYLOR, T. YAMABE, AND P.G. MEZEY, *Structural analysis of certain linear operators representing chemical network systems via the existence and uniqueness theorems of spectral resolution.* VII, Int. J. Quantum Chem., to appear.

[6] S. ARIMOTO AND M. SPIVAKOVSKY, *The asymptotic linearity theorem for the study of additivity problems of the zero-point vibrational energy of hydrocarbons and the total pi-electron energy of alternant hydrocarbons*, J. Math. Chem. 13 (1993), pp. 217–247.

[7] R.M. BEAM AND R.F. WARMING, *The asymptotic spectra of banded Toeplitz and quasi-Toeplitz matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 971–1006.

[8] R. BHATIA, *Matrix Analysis*, Springer-Verlag, New York, 1996.

[9] A. BÖTTCHER AND B. SILBERMANN, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, 1998.

[10] A. BÖTTCHER AND B. SILBERMANN, *Analysis of Toeplitz Operators*, Akademie-Verlag, Berlin, 1989 and Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[11] I. GOHBERG AND I.A. FELDMAN, *Convolution Equations and Projection Methods for Their Solution*, Transl. Math. Monogr. 41, AMS, Providence, RI, 1974.

[12] I. GOHBERG, S. GOLDBERG, AND M.A. KAASHOEK, *Classes of Linear Operators* II, Birkhäuser Verlag, Basel, 1993.

[13] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, 1958.

[14] R.A. HORN AND C.R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[15] H.J. LANDAU AND H. WIDOM, *Eigenvalue distribution of time and frequency limiting*, J. Math. Anal. Appl., 77 (1980), pp. 469–481.

[16] J.M. Lilly and J. Park, *Multiwavelet spectral and polarization analysis of seismic records*, Geophys. J. Int., 122 (1995), pp. 1001–1021.
[17] D. Slepian, *On bandwidth*, Proc. IEEE, 64 (1976), pp. 292–300.
[18] E.E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[19] H. Widom, *Toeplitz matrices*, in Studies in Real and Complex Analysis, I.I. Hirshman, Jr., ed., M.A.A. Studies in Mathematics 3, Mathematical Association of America, Washington, DC, 1965, pp. 179–209.
[20] H. Widom, *Asymptotic behaviour of block Toeplitz matrices and determinants* II, Advances in Math., 21 (1976), pp. 1–29.
[21] H. Widom, *On the singular values of Toeplitz matrices*, Z. Anal. Anwendungen, 8 (1989), pp. 221–229.
[22] H. Widom, *Eigenvalue distribution for nonselfadjoint Toeplitz matrices*, in Toeplitz Operators and Related Topics, Oper. Theory Adv. Appl. 71, Birkhäuser Verlag, Basel, 1994, pp. 1–8.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 68–77

COMPARISON OF CONVERGENCE OF GENERAL STATIONARY ITERATIVE METHODS FOR SINGULAR MATRICES*

IVO MAREK† AND DANIEL B. SZYLD‡

Abstract. New comparison theorems are presented comparing the asymptotic convergence factor of iterative methods for the solution of consistent (as well as inconsistent) singular systems of linear equations. The asymptotic convergence factor of the iteration matrix T is the quantity $\gamma(T) = \max\{|\lambda|, \lambda \in \sigma(T), \lambda \neq 1\}$, where $\sigma(T)$ is the spectrum of T. In the new theorems, no restrictions are imposed on the projections associated with the two iteration matrices being compared. The splittings of the well-known example of Kaufman [*SIAM J. Sci. Statist. Comput.*, 4 (1983), pp. 525–552] satisfy the hypotheses of the new theorems.

Key words. linear systems, iterative methods, comparison theorems, convergence factor, Markov processes, Markov chains, stochastic matrices

AMS subject classifications. 65F10, 15A48, 15A06, 15A51

PII. S0895479800375989

1. Introduction.

In this paper we study certain properties of iterative methods for the solution of $n \times n$ consistent (as well as inconsistent) singular linear systems of equations of the form

$$Ax = b. \tag{1.1}$$

One case important in applications is when

$$A = I - B, \quad B^T e = e, \quad e^T = [1, 1, \ldots, 1], \tag{1.2}$$

B is the stochastic matrix representing a Markov chain, and the solution of (1.1), for $b = 0$, is the stationary probability distribution of the Markov chain (normalized so that $x^T e = 1$); see, e.g., [3], [25]. In this case, $\rho(B) = 1$, where $\rho(B)$ denotes the spectral radius of B.

Iterative methods for the solution of (1.1) based on splittings of the form $A = M - N$, where M is nonsingular, have been successfully used for this problem; see, e.g., [1], [2], [8], [10], [14], [21]. These methods include point and block versions of the classical Jacobi, Gauss–Seidel, and SOR methods [3], [25], [29] and can be written as the following iteration, starting from an initial vector $x_{(0)}$:

$$x_{(k+1)} = Tx_{(k)} + c, \quad c = M^{-1}b. \tag{1.3}$$

The matrix $T = M^{-1}N$ is called the iteration matrix, and it is generally assumed to be nonnegative (denoted $T \geq O$), e.g., when the splittings are weak regular [3], i.e., $M^{-1} \geq O$ and $M^{-1}N \geq O$. A regular splitting is such that $M^{-1} \geq O$ and $N \geq O$

*Received by the editors August 1, 2000; accepted for publication (in revised form) by M. Eiermann January 17, 2002; published electronically June 12, 2002.
http://www.siam.org/journals/simax/24-1/37598.html

†Czech Institute of Technology, School of Civil Engineering, Thakurova 7, 16000 Praha 1, Czech Republic (marek@ms.mff.cuni.cz). This work was supported by the Grant Agency of the Czech Republic, grant 201/02/595, and by grant CEZ J04:210000010.

‡Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, PA 19122-6094 (szyld@math.temple.edu). This work was supported by National Science Foundation grant DMS-9973219.

[29]. A weak splitting is such that $M^{-1}N \geq O$ [13] (some authors call these splittings nonnegative splittings; see, e.g., [6], [31]). Since $A = M(I - T)$ it follows that A singular implies that 1 is an eigenvalue of T, and $\rho(T) = 1$ is implied in the case of stochastic matrices such as in the case of Markov chains. It also follows that the null space of A, $\mathcal{N}(A)$, coincides with $\mathcal{N}(I - T)$, the null space of $I - T$.

The rate of convergence of these iterative methods is governed by the quantity $\gamma(T) = \max\{|\lambda|, \lambda \in \sigma(T), \lambda \neq 1\}$, where $\sigma(T)$ is the spectrum of T. When $\gamma(T) = 1$ convergence is not guaranteed. When $\gamma(T) < 1$ and $ind(I - T) = 1$, there is convergence; see, e.g., [3] and section 2. We call the quantity $\gamma(T)$ the *asymptotic convergence factor* of the iterative method (1.3).

In the case of nonsingular A, the quantity governing the rate of convergence of the iterative methods is $\rho(T)$. The Perron–Frobenius theory provides the first comparison theorem for two iteration matrices; see, e.g., [3], [29].

THEOREM 1.1. *Let $0 \leq T_1 \leq T_2$; then $\rho(T_1) \leq \rho(T_2)$.*

There exists a rich literature comparing two splittings of the same matrix; see, e.g., [6], [7], [9], [12], [13], [18], [30], [31]. The following result goes back forty years to Varga [29].

THEOREM 1.2. *Let A be a nonsingular matrix with $A^{-1} \geq O$ and let $A = M_1 - N_1 = M_2 - N_2$ be two regular splittings. If*

$$N_1 \leq N_2, \tag{1.4}$$

then $\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2) < 1$.

The relation (1.4) means that $N_2 - N_1 \geq O$, i.e., that $(N_2 - N_1)x \geq 0$ whenever $x \geq 0$; in other words, if $\mathcal{K} = \mathbb{R}^n_+$, the nonnegative orthant, $(N_2 - N_1)\mathcal{K} \subset \mathcal{K}$. Woźnicki [30] was the first to prove that the hypothesis (1.4) can be replaced with

$$M_1^{-1} \geq M_2^{-1}; \tag{1.5}$$

see also [7], [31]. Condition (1.4) implies (1.5); see, e.g., [7], [15].

Comparison results such as Theorems 1.1 and 1.2 and their variants have been extended to nonnegative operators over Banach spaces, using partial orders defined by general cones $\mathcal{K}$ generating the appropriate Banach space; see, e.g., [6], [12], [22], [24], [27], [28]. See the appendix for the definition of a generating cone. The concept of nonnegativity carries over to any cone $\mathcal{K}$: $x \succeq O$ if $x \in \mathcal{K}$, and $T \succeq O$ if $T\mathcal{K} \subset \mathcal{K}$. The concepts of weak regular, regular splitting, etc., with respect to the cone $\mathcal{K}$ are based on this concept of $\mathcal{K}$-nonnegativity; see the mentioned references and [15].

When A is singular, several authors have provided examples where (1.4) holds, while $\gamma(M_1^{-1}N_1) \not\leq \gamma(M_2^{-1}N_2)$; see [4], [10]. The following example is due to Kaufman [10].

EXAMPLE 1.3. *Consider the matrix*

$$A = \begin{bmatrix} 1 & -1/2 & -1/2 & 0 \\ -1/2 & 1 & 0 & -1/2 \\ -1/2 & 0 & 1 & -1/2 \\ 0 & -1/2 & -1/2 & 1 \end{bmatrix}$$

and the two regular splittings $A = M_1 - N_1 = M_2 - N_2$ defined by

$$N_1 = \begin{bmatrix} 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad N_2 = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then $N_1 \leq N_2$, *but* $\gamma(M_1^{-1}N_1) = 1/9 > \gamma(M_2^{-1}N_2) = 0$.

In [15] we showed that conditions of the form (1.4) or (1.5) would imply the relation $\gamma(M_1^{-1}N_1) \leq \gamma(M_2^{-1}N_2)$ if these conditions are interpreted using a specific partial order, which is different than the usual partial order defined by the nonnegative orthant $\mathcal{K} = \mathbb{R}^n_+$. The new partial order is derived from the projection matrix associated with the iteration matrix, as described in the next section. In [15], our results required that both iteration matrices $T_1 = M_1^{-1}N_1$ and $T_2 = M_2^{-1}N_2$ be associated with the *same* projection (onto $\mathcal{N}(A)$). The splittings of Example 1.3 do not have this property; see Example 2.3 below.

In section 3 we present new comparison results without the requirement that the two projections be the same. In particular, unlike the results in [15], no restriction is imposed on $\dim\mathcal{N}(A)$. In other words, the new theorems can be applied to a much more general collection of splittings of A. In particular, our theorems apply to Example 1.3.

In these theorems we implicitly assume that $\gamma(T) \in \sigma(T)$. In section 4 we extend our theory to some splittings where this assumption is not needed.

2. The partial order. A matrix $T \in \mathbb{R}^{n\times n}$ is called *convergent* if $\lim_{k\to\infty} T^k$ exists. A splitting $A = M - N$ is called convergent if its iteration matrix $T = M^{-1}N$ is convergent. In this paper we consider the case where $\rho(T) = 1$. The following result indicates an equivalent definition of convergence; see [3, Lemma 7.6.9], [14] [15], [17]. For other equivalent conditions, see, e.g., [16], [19], [20], [26].

THEOREM 2.1. *Let* $T \in \mathbb{R}^{n\times n}$. T *is convergent if and only if*

$$T = P + Z, \text{ where } P^2 = P,\ PZ = ZP = O, \tag{2.1}$$

and $\rho(Z) < 1$. *Moreover,* P *is a projection onto* $\mathcal{N}(I - T)$.

It follows from Theorem 2.1 that $\lim_{k\to\infty} T^k = P$. In the case studied in this paper, i.e., when $A = M - N$ and $T = M^{-1}N$, the matrix P is a projection onto $\mathcal{N}(A)$. As is well known, an expression for this projection is $P = I - (I-T)^{\#}(I-T)$, where the notation $Q^{\#}$ stands for the (unique) group inverse of Q; see, e.g., [5], [16]. Thus, $I - P = (I-T)^{\#}(I-T)$.

REMARK 2.2. *If* $T \geq O$ *is irreducible, the Perron–Frobenius theorem implies that* $\dim\mathcal{N}(I-T) = \dim\mathcal{N}(A) = 1$. *In this case, any projection onto* $\mathcal{N}(A)$ *necessarily has the form*

$$P = \hat{x}\hat{z}^T, \quad \text{with} \quad \hat{z}^T\hat{x} = 1, \tag{2.2}$$

where $\hat{x} \in \mathcal{N}(A)$ *and* $\hat{z}$ *is some vector in* $\mathbb{R}^n$.

EXAMPLE 2.3. *Consider the matrix* $A = I - B$ *and the two splittings of Example* 1.3. *Let* $T_0 = B$, $T_i = M_i^{-1}N_i$, $i = 1, 2$. *We have* $\hat{x} = e \in \mathcal{N}(A)$, e *as in* (1.2). *Let* $T_i = P_i + Z_i$, *satisfying* (2.1), $i = 0, 1, 2$. *We obtain* $P_i = \hat{x}\hat{z}_i^T$, $i = 0, 1, 2$, *where* $\hat{z}_0^T = [1/4, 1/4, 1/4, 1/4]$, $\hat{z}_1^T = [0, 0, 1/2, 1/2]$, *and* $\hat{z}_2^T = [0, 1/4, 1/4, 1/2]$. *Note, however, that* $\rho(Z_0) = 1$ *and* T_0 *is not convergent.*

It follows from Example 2.3 that the iteration matrices obtained from different splittings of the same matrix A may have associated with them totally different projections P_i onto the same subspace $\mathcal{N}(A) = \mathcal{R}(P_i)$.

Given a convergent matrix $T_i = P_i + Z_i$ satisfying (2.1), the cone which we use for our comparison is a (pointed) cone generating the range of the projection $I - P_i = (I - T_i)^{\#}(I - T_i)$. In other words, we will use $\mathcal{K}_i$ such that for every element $u \in \mathcal{R}(I - P_i)$, there are v, $w \in \mathcal{K}_i$ (usually not unique) such that $u = v - w$, i.e.,

$\mathcal{K}_i - \mathcal{K}_i = \mathcal{R}(I - P_i)$. (We review the definition of a generating cone in the appendix.) Note that we always have $(I - P_i)\mathcal{K}_i = \mathcal{K}_i$, and furthermore $I - P_i$ is the identity operator on $\mathcal{K}_i$ and on $\mathcal{R}(I - P_i)$.

REMARK 2.4. *In the important and practical case of* $\dim\mathcal{N}(A) = 1$, *e.g., when* B *in* (1.2) *is irreducible, it was pointed out in* [15] *that we can compute* $\mathcal{R}(I-P_i)$ *even if we do not know* P_i. *This follows since from* (2.2), $P_i^T\hat{z} = \hat{z}$, *i.e., that* $(I - P_i)^T\hat{z} = 0$. *Then we can characterize* $\mathcal{R}(I - P_i)$ *as*

$$\mathcal{R}(I - P_i) = \{x \in \mathbb{R}^n \ : \ x^T\hat{z} = 0\}. \tag{2.3}$$

We can then choose

$$\mathcal{K}_i = \left\{ x \in \mathbb{R}^n \ : \ x = \sum_{k=1}^{n-1} \alpha_k v_k, \ \alpha_k \geq 0, k = 1, \ldots, n-1 \right\}, \tag{2.4}$$

where the $n - 1$ *vectors* $v_k \in \mathcal{R}(I - P_i)$ *(i.e.,* $v_k^T\hat{z} = 0$*) are linearly independent; cf.* (A.1).

Let $\mathcal{K}_i - \mathcal{K}_i = \mathcal{R}(I - P_i)$. By definition, the cone $\mathcal{K}_i$ generates a proper subspace, i.e., not the whole space. Therefore, to define a partial order on $\mathbb{R}^n$ using $\mathcal{K}_i$, these vectors and the matrices operating on them need to be restricted to the subspace $\mathcal{R}(I - P_i)$. Thus, we say that $x \overset{i}{\geq} y$, $x, \ y \in \mathbb{R}^n$, if $(I - P_i)(x - y) \in \mathcal{K}_i$. Similarly, a matrix $T \in \mathbb{R}^{n\times n}$ is said to be $\mathcal{K}_i$-nonnegative, denoted $T \overset{i}{\geq} O$ if $(I-P_i)Tx \in \mathcal{K}_i$ for all $x \in \mathcal{K}_i$. Similarly, a splitting $A = M - N$ is called $\mathcal{K}_i$-weak, $\mathcal{K}_i$-weak regular, or $\mathcal{K}_i$-regular if $M^{-1}N \overset{i}{\geq} O$, $M^{-1} \overset{i}{\geq} O$, and $M^{-1}N \overset{i}{\geq} O$, or $M^{-1} \overset{i}{\geq} O$ and $N \overset{i}{\geq} O$, respectively; see examples and further discussion in [15].

3. Comparison theorems. We begin with the observation that if one has two projections P_i and P_j onto the *same* subspace S, then

$$P_jP_i = P_i \ \text{ and consequently } \ (I - P_j)(I - P_i) = I - P_j \tag{3.1}$$

since for two projections P_i and P_j, there obviously holds $P_jP_i = P_i$ if and only if $\mathcal{R}(P_i) \subseteq \mathcal{R}(P_j)$.

In the particular case where S is one-dimensional and the two projections have the form (2.2), the identity (3.1) can be computed directly.

We are ready now to show an important tool for our comparisons.

LEMMA 3.1. *Let* A *be a singular matrix. Let* $A = M_1 - N_1 = M_2 - N_2$ *be two convergent splittings, and let* $T_i = M_i^{-1}N_i = P_i + Z_i$, $P_i^2 = P_i$, $P_iZ_i = Z_iP_i = O$, $\rho(Z_i) < 1$, $i = 1, 2$. *Then* $\sigma((I - P_i)Z_j) = \sigma(Z_j)$.

Proof. If $i = j$ there is nothing to prove. Thus, we assume $i \neq j$. Let $\lambda \in \sigma(Z_j)$ and x such that $Z_jx = \lambda x$. Since $Z_j(I - P_j) = Z_j$, we have, using (3.1), that

$$Z_j = Z_j(I - P_j)(I - P_i) = Z_j(I - P_i)$$

and therefore $Z_j(I - P_j)(I - P_i)x = \lambda x$. Consequently,

$$(I - P_i)Z_jx = (I - P_i)Z_j[(I - P_i)x] = \lambda(I - P_i)x,$$

and thus $\lambda \in \sigma((I - P_i)Z_j)$.

Conversely, let $\lambda \in \sigma((I - P_i)Z_j)$ and let v such that $(I - P_i)Z_jv = \lambda v$. Multiply the last equation by $(I - P_j)$ and, using (3.1), we have

$$(I - P_j)(I - P_i)Z_jv = (I - P_j)Z_jv = Z_jv = Z_j[(I - P_j)v] = \lambda(I - P_j)v$$

and thus $\lambda \in \sigma(Z_j)$. ∎

The following result was proved in [13], and the nonnegativity is with respect to any cone.

LEMMA 3.2. *Let $V \succeq O$, and let $x \succeq 0$, $x \neq 0$, be such that $Vx - \alpha x \succeq 0$. Then $\alpha \leq \rho(V)$.*

We can now proceed with the main result, which generalizes [15, Theorem 5.6] and is the general counterpart to Theorem 1.2 with the hypothesis (1.5).

THEOREM 3.3. *Let A be singular. Let $A = M_1 - N_1 = M_1(I - T_1) = M_2 - N_2 = M_2(I - T_2)$ be two (convergent) $\mathcal{K}_i$-regular splittings, where $\mathcal{K}_i$ is the cone generating $\mathcal{R}(I - P_i)$ for either $i = 1$ or $i = 2$, and $T_j = P_j + Z_j$, $P_j^2 = P_j$, $P_j Z_j = Z_j P_j = O$, $\rho(Z_j) < 1$, $j = 1, 2$. If*

$$M_1^{-1} \overset{i}{\succeq} M_2^{-1}, \tag{3.2}$$

then $\gamma(T_1) \leq \gamma(T_2)$.

Proof. We assume first that $i = 1$. If $\gamma(T_1) = 0$, there is nothing to prove, so we assume $\gamma(T_1) \neq 0$. Since $\mathcal{K}_1$ is the cone generating $\mathcal{R}(I - P_1)$, and by hypothesis $Z_1\mathcal{K}_1 = T_1\mathcal{K}_1 \subset \mathcal{K}_1$, there is a Perron eigenvector $x = (I - P_1)x \in \mathcal{K}_1$ for which $T_1 x = Z_1 x = \rho(Z_1)x = \gamma(T_1)x \succeq 0$. Here and in the rest of the proof we use the symbol $\succeq$ to indicate $\overset{1}{\succeq}$, since there is no possibility of confusion. Then

$$M_1 x = \frac{1}{\gamma(T_1)} N_1 x \succeq 0 \tag{3.3}$$

and

$$Ax = M_1(I - T_1)x = \frac{1 - \gamma(T_1)}{\gamma(T_1)} N_1 x \succeq 0.$$

Using (3.2), it follows that

$$(M_1^{-1} - M_2^{-1})Ax = (I - T_1)x - (I - T_2)x = T_2 x - \gamma(T_1)x \succeq 0. \tag{3.4}$$

Premultiply the last equation by $(I - P_1)$ which is not only $\mathcal{K}_1$-nonnegative but actually the identity on $\mathcal{K}_1$, and observe that because of (3.1), $(I - P_1)T_2 = (I - P_1)Z_2$. Thus, we have that

$$(I - P_1)Z_2 x \succeq \gamma(T_1)x,$$

which implies by Lemma 3.2 that $\rho((I - P_1)Z_2) \geq \gamma(T_1)$. Using Lemma 3.1, we can rewrite this as $\gamma(T_2) = \rho(Z_2) \geq \gamma(T_1)$, completing the proof for $i = 1$.

The proof for $i = 2$ is similar, using the eigenvector x of T_2, except that we need to require the additional hypothesis that x is in the interior of $\mathcal{K}_2$, so we can use [13, Lemma 3.3]. ∎

REMARK 3.4. *We point out that this theorem is valid with weaker hypotheses, using the same proof, namely, that the splittings be $\mathcal{K}_i$-weak splittings and convergent (or $\mathcal{K}_i$-weak regular splittings) and that if the Perron eigenvector x of Z_1 satisfies $N_1 x \overset{1}{\succeq} 0$. Alternatively the Perron eigenvector x of Z_2 (in the interior of $\mathcal{K}_2$) needs to satisfy $N_2 x \overset{2}{\succeq} 0$. We also remark that, as it can be seen from the hypotheses and the proof, no restriction on* $\dim \mathcal{N}(A)$ *is needed.*

The following result was shown in [15]; see also [7] or [31] for the nonsingular case.

LEMMA 3.5. *Let $A = M_1 - N_1 = M_2 - N_2$ be two $\mathcal{K}_i$-weak regular splittings, where $\mathcal{K}_i$ is a cone generating $\mathcal{R}(I - P_i)$, for either $i = 1$ or $i = 2$, and $T_j = P_j + Z_j$, $P_j^2 = P_j$, $P_j Z_j = Z_j P_j = O$, $\rho(Z_j) < 1$, $j = 1, 2$. If $N_2 \overset{i}{\geq} N_1$, then $M_1^{-1} \overset{i}{\geq} M_2^{-1}$.*

We can write the counterpart to Theorem 1.2. The proof follows directly from Lemma 3.5 and Theorem 3.3.

THEOREM 3.6. *Let A be singular. Let $A = M_1 - N_1 = M_1(I - T_1) = M_2 - N_2 = M_2(I - T_2)$ be two (convergent) $\mathcal{K}_i$-regular splittings, where $\mathcal{K}_i$ is the cone generating $\mathcal{R}(I - P_i)$, for either $i = 1$ or $i = 2$, and $T_j = P_j + Z_j$, $P_j^2 = P_j$, $P_j Z_j = Z_j P_j = O$, $\rho(Z_j) < 1$, $j = 1, 2$. If $N_2 \overset{i}{\geq} N_1$, then $\gamma(T_1) \leq \gamma(T_2)$.*

Again, this theorem is valid with weaker hypotheses; see Remark 3.4.

EXAMPLE 3.7. *Consider the matrix A and the splittings of Example* 1.3. *The projections P_1 and P_2 are shown in Example* 2.3. *A simple computation gives the matrix*

$$(I - P_1)(N_1 - N_2) = \begin{bmatrix} 0 & -1/2 & 0 & 1/4 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & -1/4 \\ 0 & 0 & 0 & 1/4 \end{bmatrix},$$

which is nonnegative with respect to the following cone generating $\mathcal{R}(I - P_1)$:

$$\mathcal{K}_1 = \left\{ \sum_{k=1}^{3} \alpha_k v_k, \alpha_k \geq 0, v_1^T = [-1, 0, 0, 0], v_2^T = [1, 1, -1, 1], v_3^T = [0, 1, 0, 0] \right\}.$$

Indeed, $(I - P_1)(N_1 - N_2)v_1 = 0$, $(I - P_1)(N_1 - N_2)v_2 = \frac{1}{2}v_1 + \frac{1}{4}v_2$, and $(I - P_1)(N_1 - N_2)v_3 = \frac{1}{2}v_1$. Furthermore, consider the matrix

$$(I - P_2)(N_1 - N_2) = \begin{bmatrix} 0 & -1/2 & 0 & 1/8 \\ 0 & 0 & 0 & 1/8 \\ 0 & 0 & 0 & -3/8 \\ 0 & 0 & 0 & 1/8 \end{bmatrix}.$$

This matrix is nonnegative with respect to the following cone generating $\mathcal{R}(I - P_2)$:

$$\mathcal{K}_2 = \left\{ \sum_{k=1}^{3} \alpha_k w_k, \alpha_k \geq 0, w_1^T = [-1, 0, 0, 0], w_2^T = [-2, 2, -2, 0], w_3^T = [1, 1, -3, 1] \right\}.$$

Indeed, $(I-P_2)(N_1-N_2)w_1 = 0$, $(I-P_2)(N_1-N_2)w_2 = w_1$, and $(I-P_2)(N_1-N_2)w_3 = \frac{1}{2}w_1 + \frac{1}{8}w_3$.

We have shown in [15] examples when two matrices cannot be compared in the usual partial order but are comparable with the appropriate choice of generating cone. Example 3.7 indicates that even in the case when two matrices are comparable in the usual partial order, the direction of the comparison can be reversed with the appropriate cone, and thus the comparison of the asymptotic convergence factors can be obtained.

We note that in the special case when $P_1 = P_2$, Theorems 3.3 and 3.6 reduce to the comparison theorems in [15], but these do not apply to Example 1.3.

We now present the counterpart to Theorem 1.1 in the singular case.

THEOREM 3.8. *Let A be singular. Let $A = M_1 - N_1 = M_1(I - T_1) = M_2 - N_2 = M_2(I - T_2)$ be two convergent $\mathcal{K}_i$-weak splittings, where $\mathcal{K}_i$ is the cone generating*

$\mathcal{R}(I-P_i)$ for either $i=1$ or $i=2$, and $T_j = P_j + Z_j$, $P_j^2 = P_j$, $P_jZ_j = Z_jP_j = O$, $\rho(Z_j) < 1$, $j = 1, 2$. If

$$T_2 \overset{i}{\geq} T_1, \tag{3.5}$$

then $\gamma(T_1) \leq \gamma(T_2)$.

Proof. We assume that $i = 1$. The proof for the case $i = 2$ is analogous. Premultiply (3.5) by $(I - P_1)$, the identity in $\mathcal{K}_1$, and using (3.1), we obtain

$$(I-P_1)Z_2 \overset{1}{\geq} Z_1 \overset{1}{\geq} O.$$

We now apply the Perron–Frobenius theorem in the subspace $\mathcal{R}(I - P_i)$ (see, e.g., [12], [22]) and obtain $\rho((I - P_1)Z_2) \geq \rho(Z_1)$. By Lemma 3.1 we then have

$$\gamma(T_2) = \rho(Z_2) = \rho((I-P_1)Z_2) \geq \rho(Z_1) = \gamma(T_1). \qquad \square$$

EXAMPLE 3.9. *Consider the matrix A and the splittings of Example* 1.3. *The projections P_1 and P_2 are shown in Example* 2.3. *One can directly compute the matrix*

$$(I-P_1)(T_1-T_2) = \begin{bmatrix} 0 & -1/4 & -1/12 & 1/3 \\ 0 & 0 & 1/6 & -1/6 \\ 0 & 0 & 1/18 & -1/18 \\ 0 & 0 & -1/18 & 1/18 \end{bmatrix},$$

which is nonnegative with respect to the following cone generating $\mathcal{R}(I - P_1)$:

$$\mathcal{K}_1 = \left\{ \sum_{k=1}^{3} \alpha_k v_k, \alpha_k \geq 0, v_1^T = [0,-1,0,0], v_2^T = [1,0,0,0], v_3^T = \left[-\frac{15}{4}, 0, 1, -1\right] \right\}.$$

Indeed, $(I-P_1)(T_1-T_2)v_1 = \frac{1}{4}v_2$, $(I-P_1)(T_1-T_2)v_2 = 0$, and $(I-P_1)(T_1-T_2)v_3 = \frac{1}{3}v_2 + \frac{1}{9}v_3$. Furthermore, the matrix

$$(I-P_2)(T_1-T_2) = \begin{bmatrix} 0 & -1/4 & -1/36 & 5/18 \\ 0 & 0 & -1/9 & 1/9 \\ 0 & 0 & 1/9 & -1/9 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

is nonnegative with respect to the following cone generating $\mathcal{R}(I - P_2)$:

$$\mathcal{K}_2 = \left\{ \sum_{k=1}^{3} \alpha_k w_k, \alpha_k \geq 0, w_1^T = [-1,0,0,0], w_2^T = [-1,3,-1,-1], w_3^T = [-1,1,-1,0] \right\}.$$

Indeed, $(I-P_2)(T_1-T_2)w_1 = 0$, $(I-P_2)(T_1-T_2)w_2 = w_1$, and $(I-P_2)(T_1-T_2)w_3 = \frac{1}{9}w_1 + \frac{1}{3}w_3$.

4. Majorizing splittings. We conclude with some observations which enlarge the class of splittings for which we can compare the asymptotic convergence factors. In Theorems 3.3, 3.6, and 3.8, we assume that the splittings are convergent $\mathcal{K}_i$-weak, and thus, we are implicitly assuming that the asymptotic convergence factor belongs to the spectrum, i.e., that $\gamma(T_i) \in \sigma(T_i)$, $T_i = M_i^{-1}N_i$, $A = M_i - N_i$. We can capture

some of the cases where the splittings are such that $\gamma(T_i) \notin \sigma(T_i)$ by the following construction.

DEFINITION 4.1. *Given a cone $\mathcal{K}$ and its induced partial order $\succeq$, one can define the absolute value of a matrix Z by $|Z| = Z^+ + Z^-$, where*

$$Z = Z^+ - Z^-, \quad \text{with } Z^+ \succeq O,\ Z^- \succeq O. \tag{4.1}$$

This definition of absolute value of an operator with respect to a partial order can be seen as a slight generalization of that defined in [23] in the case of a vector lattice space (Riesz space). Here, we do not need a vector lattice order but need only that the matrix Z be regular in the sense of [23, Definition 1.1]. The decomposition (4.1) is then possible (although not necessarily in a unique manner).

DEFINITION 4.2. *Let A be singular. Let $A = M_1 - N_1 = M_1(I - T_1) = M_2 - N_2 = M_2(I - T_2)$ be two splittings. Let $T_j = P_j + Z_j$, $P_j^2 = P_j$, $P_j Z_j = Z_j P_j = O$ for $j = 1,\ 2$. Let $A = M_2 - N_2$ be a $\mathcal{K}_2$-weak splitting, where $\mathcal{K}_2$ is a cone generating $\mathcal{R}(I - P_2)$. We say that the splitting $A = M_1 - N_1$ is majorized by the splitting $A = M_2 - N_2$ when $|Z_1| \overset{2}{\succeq} Z_2$. In a similar manner one defines a minorized splitting.*

REMARK 4.3. *Majorizing splittings were introduced in* [12, *section* 7] *for splittings of a nonsingular operator A and, in particular, were applied to SOR splittings. Many of the results from* [12] *using majorizing splittings can be easily extended to the singular case. Note that a basic hypothesis for deriving the results in* [12] *is the normality of the cones under consideration; see the appendix for definitions and comments.*

We are now ready to present a comparison result between a splitting for which the asymptotic convergence factor is not in the spectrum of the iteration matrix and another splitting for which it is.

THEOREM 4.4. *Let $A = M_1 - N_1 = M_1(I - T_1) = M_2 - N_2 = M_2(I - T_2)$ be two splittings. Let $T_j = P_j + Z_j$, $P_j^2 = P_j$, $P_j Z_j = Z_j P_j = O$ for $j = 1,\ 2$. Let $A = M_2 - N_2$ be a (convergent) $\mathcal{K}_2$-weak splitting, where $\mathcal{K}_2$ is a cone generating $\mathcal{R}(I - P_2)$, with $\rho(Z_2) < 1$. Assume that the splitting $A = M_2 - N_2$ majorizes the splitting $A = M_1 - N_1$. Then*

$$\gamma(T_1) = \rho(Z_1) \le \rho(|Z_1|) \le \rho(Z_2) = \gamma(T_2).$$

Proof. Relations

$$-|Z_1|y \ \overset{2}{\preceq}\ Z_1 y \ \overset{2}{\preceq}\ |Z_1|y,$$

valid for any $y \in \mathcal{K}_2$, imply that $\rho(Z_1) \le \rho(|Z_1|)$. Further, by hypothesis, we have

$$0 \ \overset{2}{\preceq}\ |Z_1|y \ \overset{2}{\preceq}\ Z_2 y \ \text{ for all } y \in \mathcal{K}_2,$$

and consequently $\rho(|Z_1|) \le \rho(Z_2)$. ∎

REMARK 4.5. *The fact that T_1 is convergent is a consequence of the hypothesis that Z_2 is convergent. Therefore, Z_1 need not be assumed to be convergent.*

5. Concluding remarks. We have demonstrated that the usual partial order ($\ge$) defined by the nonnegative orthant $\mathbb{R}^n_+$ is not the appropriate choice of order when comparing splittings of singular matrices.

We have provided two different partial orders with which the comparison of the splittings implies the comparison of the asymptotic convergence factors of the corresponding iteration matrices.

Example 1.3, due to Kaufman [10], was originally presented as a counterexample to possible theorems of the form of Theorem 1.2. It now becomes a good example to show that the alternative partial orders are the appropriate ones to use in the context of singular matrices.

Appendix.

DEFINITION A.1. *Let $\mathcal{E}$ be a real Banach space. A normal cone $\mathcal{K}$ is a subset of $\mathcal{E}$ with the following properties:*

(i) $\mathcal{K} + \mathcal{K} \subset \mathcal{K}$,

(ii) $\alpha\mathcal{K} \subset \mathcal{K}$ *for* $\alpha \geq 0$,

(iii) $\mathcal{K} \cap (-\mathcal{K}) = \{0\}$, *i.e., it is pointed,*

(iv) $\bar{\mathcal{K}} = \mathcal{K}$, *where* $\bar{\mathcal{K}}$ *denotes the norm-closure of* $\mathcal{K}$, *and*

(v) $\exists\ \sigma > 0$ *such that for* $x,\ y\ \in \mathcal{K}$ *one has* $\|x + y\| \geq \sigma\|x\|$.

We say that $\mathcal{K}$ is generating if $\mathcal{E} = \mathcal{K} - \mathcal{K}$. The typical example is $\mathcal{E} = \mathbb{R}^n$, and a generating cone is the standard cone

$$\begin{aligned} \mathcal{K} = \mathbb{R}^n_+ &= \{x \in \mathbb{R}^n \ :\ x \geq 0\} \\ &= \left\{ x \in \mathbb{R}^n \ :\ x = \sum_{k=1}^{n} \alpha_k e_k,\ \alpha_k \geq 0, k = 1, \ldots, n \right\}, \end{aligned} \tag{A.1}$$

where e_k is the standard kth canonical vector, i.e., the kth column of the identity.

We should remark that condition (v) is simply saying that the norm $\|.\|$ of the Banach space $\mathcal{E}$ is $\mathcal{K}$-semimonotone (and $\mathcal{K}$-monotone if it holds with $\sigma = 1$). The following result, which can be found, e.g., in [11], indicates when it holds.

PROPOSITION A.2. *Assume $\mathcal{E}$ is a Banach space over the field of reals with the norm $\|.\|_{\mathcal{E}}$. A cone $\mathcal{K} \subset \mathcal{E}$ satisfying* (i)–(iv) *is normal, i.e., it fulfills* (v), *if and only if the norm on $\mathcal{E}$ $\|.\|_*$ defined by*

$$\|x\|_* = \mathrm{Max}\left(\inf\{\|u\|_{\mathcal{E}} : u \in \mathcal{E}, (u - x) \in \mathcal{K}\}, \sup\{\|v\|_{\mathcal{E}} : v \in \mathcal{E}, (x - v) \in \mathcal{K}\}\right),\ x \in \mathcal{K},$$

is equivalent with $\|.\|_{\mathcal{E}}$.

As a consequence of Proposition A.2 we conclude that any closed cone in $\mathbb{R}^n$, i.e., any set satisfying (i)–(iv) of Definition A.1, is normal, since all the norms on a finite dimensional space are equivalent.

Acknowledgments. We thank Hans Schneider and Michael Eiermann, whose questions and observations led us to several improvements of the manuscript.

REFERENCES

[1] G. P. BARKER AND R. J. PLEMMONS, *Convergent iterations for computing stationary distributions of Markov chains,* SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 390–398.

[2] V. A. BARKER, *Numerical solution of sparse singular systems of equations arising from ergodic Markov chains,* Comm. Statist. Stochastic Models, 5 (1989), pp. 355–381.

[3] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, 3rd. ed., Academic Press, New York, 1979. Reprinted by SIAM, Philadelphia, 1994.

[4] J. J. BUONI, M. NEUMANN, AND R. S. VARGA, *Theorems of Stein–Rosenberg type.* III. *The singular case,* Linear Algebra Appl., 42 (1982), pp. 183–198.

[5] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, London, San Francisco, Melbourne, 1979. Reprinted by Dover, New York, 1991.

[6] J.-J. CLIMENT AND C. PEREA, *Some comparison theorems for weak nonnegative splittings of bounded operators,* Linear Algebra Appl., 275/276 (1998), pp. 77–106.

[7] G. Csordas and R. S. Varga, *Comparisons of regular splittings of matrices,* Numer. Math., 44 (1984), pp. 23–35.
[8] T. Dayar and W. J. Stewart, *Comparison of partitioning techniques for two-level iterative methods on large, sparse Markov chains,* SIAM J. Sci. Comput., 21 (2000), pp. 1691–1705.
[9] L. Elsner, *Comparisons of weak regular splittings and multisplitting methods,* Numer. Math., 56 (1989), pp. 283–289.
[10] L. Kaufman, *Matrix methods for queuing problems,* SIAM J. Sci. Statist. Comput., 4 (1983), pp. 525–552.
[11] M. A. Krasnoselskii, E. A. Lifshits, and A. V. Sobolev, *Positive Linear Systems,* Nauka, Moscow, 1985 (in Russian); Heldermann-Verlag, Berlin, 1989 (in English).
[12] I. Marek, *Frobenius theory of positive operators: Comparison theorems and applications,* SIAM J. Appl. Math., 19 (1970), pp. 607–628.
[13] I. Marek and D. B. Szyld, *Comparison theorems for weak splittings of bounded operators,* Numer. Math., 58 (1990), pp. 387–397.
[14] I. Marek and D. B. Szyld, *Iterative and semi-iterative methods for computing stationary probability vectors of Markov operators,* Math. Comput., 61 (1993), pp. 719–731.
[15] I. Marek and D. B. Szyld, *Comparison theorems for the convergence factor of iterative methods for singular matrices,* Linear Algebra Appl., 316 (2000), pp. 67–87.
[16] C. D. Meyer and R. J. Plemmons, *Convergent powers of a matrix with applications to iterative methods for singular systems,* SIAM J. Numer. Anal., 14 (1977), pp. 699–705.
[17] V. Migallón, J. Penadés, and D. B. Szyld, *Block two-stage methods for singular systems and Markov chains,* Numer. Linear Algebra Appl., 3 (1996), pp. 413–426.
[18] V. A. Miller and M. Neumann, *A note on comparison theorems for nonnegative matrices,* Numer. Math., 47 (1985), pp. 427–434.
[19] M. Neumann and R. J. Plemmons, *Convergent nonnegative matrices and iterative methods for consistent linear systems,* Numer. Math., 31 (1978), pp. 265–279.
[20] R. Oldenburger, *Infinite powers of matrices and characteristic roots,* Duke Math. J., 6 (1940), pp. 357–361.
[21] D. P. O'Leary, *Iterative methods for finding the stationary vector for Markov chains,* in Linear Algebra, Markov Chains and Queuing Models, C. D. Meyer and R. J. Plemmons, eds., IMA Vol. Math. Appl. 48, Springer, New York, Berlin, 1993, pp. 125–136.
[22] W. C. Rheinboldt and J. S. Vandergraft, *A simple approach to the Perron–Frobenius theory for positive operators on general partially-ordered finite-dimensional linear spaces,* Math. Comput., 27 (1973), pp. 139–145.
[23] H. H. Schaefer, *Banach Lattices and Positive Operators*, Springer, Berlin, Heidelberg, New York, 1974.
[24] H. Schneider, *Positive operators and an inertia theorem,* Numer. Math., 7 (1965), pp. 11–17.
[25] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1994.
[26] D. B. Szyld, *Equivalence of convergence conditions for iterative methods for singular equations,* Numer. Linear Algebra Appl., 1 (1994), pp. 151–154.
[27] J. S. Vandergraft, *Spectral properties of matrices which have invariant cones,* SIAM J. Appl. Math., 16 (1968), pp. 1208–1222.
[28] J. S. Vandergraft, *Applications of partial orderings to the study of positive definiteness, monotonicity, and convergence of iterative methods for linear systems,* SIAM J. Numer. Anal., 9 (1972), pp. 97–104.
[29] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962. 2nd ed., Springer, Berlin, 2000.
[30] Z. I. Woźnicki, *Two-Sweep Iterative Methods for Solving Large Linear Systems and Their Application to the Numerical Solution of Multi-Group Multi-Dimensional Neutron Diffusion Equations*, Ph.D. thesis, Institute of Nuclear Research, Otwock-Świerk, Poland, 1973.
[31] Z. I. Woźnicki, *Nonnegative splitting theory,* Japan J. Indust. Appl. Math., 11 (1994), pp. 289–342.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 78–90

MIRRORSYMMETRIC MATRICES, THEIR BASIC PROPERTIES, AND AN APPLICATION ON ODD/EVEN-MODE DECOMPOSITION OF SYMMETRIC MULTICONDUCTOR TRANSMISSION LINES*

GUO-LIN LI† AND ZHENG-HE FENG†

Abstract. Mirrorsymmetric matrices, which are the interaction matrices of mirrorsymmetric structures, are defined in this paper. The well-known centrosymmetric matrices, which can only reflect the mirror reflection relations of mirrorsymmetric structures with no component or one component on the mirror plane, are special cases of mirrorsymmetric matrices. However, almost all the properties of centrosymmetric matrices can be directly generalized to mirrorsymmetric matrices. It is proved that the eigenvectors of a mirrorsymmetric matrix are either mirrorsymmetric or skew-mirrorsymmetric corresponding to even-modes and odd-modes of the real physical systems. The application on odd/even-mode decomposition of symmetric multiconductor transmission lines is investigated in detail.

Key words. mirrorsymmetric matrices, centrosymmetric matrices, mode decomposition, multiconductor transmission lines

AMS subject classifications. 15A18, 15A57, 20B35, 78A50

PII. S0895479801393824

1. Introduction.

Real or complex matrices of rather high order are commonly encountered in real physical systems analysis. Usually, a physical system possesses certain geometrical symmetry. Mirror symmetry is the most common one. Interaction matrices of mirrorsymmetric structures are centrosymmetric while no component or only one component is on the mirror plane. Weeks exploits such symmetry in electrical packaging analysis in [1]. The properties of centrosymmetric matrices have already been thoroughly investigated; see, e.g., [2, 3, 4, 5] and the references therein. Centrosymmetric matrices cannot represent mirrorsymmetric structures with more than one component on the mirror plane. A new matrix type, which is much like the centrosymmetric matrix, is defined and called a mirrorsymmetric matrix because it is the interaction matrix of the mirrorsymmetric structure with any components on the mirror plane. Though centrosymmetric matrices are special cases of mirrorsymmetric matrices, almost all the properties of centrosymmetric matrices can be directly generalized to mirrorsymmetric matrices. Some basic properties of mirrorsymmetric matrices are discussed in section 2. In sections 3 and 4, multiconductor transmission line (MTL) equations with mirrorsymmetric per-unit-length (PUL) matrices are studied. Mirrorsymmetric MTL equations are divided into two subequations: odd-mode MTL equations and even-mode MTL equations. The order of equations (and matrices) is reduced from n to k and $k+p$ corresponding to odd-modes and even-modes, where p is the conductor number of the mirror plane.

*Received by the editors August 16, 2001; accepted for publication (in revised form) by A. H. Sayed January 8, 2002; published electronically June 12, 2002.
http://www.siam.org/journals/simax/24-1/39382.html
†Electronic Engineering Department, Tsinghua University, State Key Lab on Microwave and Digital Communications, Beijing, People's Republic of China, 100084 (guolinli98@mails.tsinghua.edu.cn, fzh-dee@mail.tsinghua.edu.cn).

2. Definition and basic properties of mirrorsymmetric matrices.

DEFINITION 1. *The (k,p)-mirror matrix $W_{(k,p)}$ is defined by*

$$W_{(k,p)} = \begin{bmatrix} & & J_k \\ & I_p & \\ J_k & & \end{bmatrix}, \tag{1}$$

where I_p, is the p-square identity matrix and J_k is the k-square backward identity matrix with ones along the secondary diagonal and zeros elsewhere.

The dimension of the (k,p)-mirror matrix is $n = 2k + p$, where $k \geq 1$, $p \geq 0$. The (k,p)-mirror matrix $W_{(k,p)}$ is orthogonal and symmetric, i.e., $W^{-1} = W^T = W$. When $p = 0$ or 1, mirror matrix $W_{(k,p)}$ is backward identity matrix J_n.

DEFINITION 2. *An n-dimensional vector a is called (k,p)-mirrorsymmetric if*

$$W_{(k,p)}a = a \tag{2a}$$

or (k,p)-skew-mirrorsymmetric if

$$W_{(k,p)}a = -a. \tag{2b}$$

From the definition, we know that (k,p)-mirrorsymmetric vector a may be written as

$$a = \begin{bmatrix} a_k \\ a_p \\ J_k a_k \end{bmatrix} \tag{3a}$$

and (k,p)-skew-mirrorsymmetric vector a may be written as

$$a = \begin{bmatrix} a_k \\ 0_p \\ -J_k a_k \end{bmatrix}, \tag{3b}$$

where a_k is the k-dimensional vector and a_p is the p-dimensional vector.

DEFINITION 3. *Let $\Omega_{(k,p)}$ be the set of $n \times n$ matrices such that $Q \in \Omega_{(k,p)}$ if and only if*

$$Q = W_{(k,p)} Q W_{(k,p)}, \tag{4}$$

where $n = 2k + p$, $k \geq 1$, $p \geq 0$.

DEFINITION 4. *$Q_{(k,p)} \in \Omega_{(k,p)}$ is called the (k,p)-mirrorsymmetric matrix.*

From the definitions, a (k,p)-mirrorsymmetric matrix $Q_{(k,p)}$ is of the form

$$Q_{(k,p)} = \begin{bmatrix} A_{k\times k} & B_{k\times p} & C_{k\times k}J_k \\ D_{p\times k} & E_{p\times p} & D_{p\times k}J_k \\ J_k C_{k\times k} & J_k B_{k\times p} & J_k A_{k\times k} J_k \end{bmatrix}, \tag{5}$$

where $A_{k\times k}$, $B_{k\times p}$, $C_{k\times k}$, $D_{p\times k}$, $E_{p\times p}$ are $k \times k$, $k \times p$, $k \times k$, $p \times k$, $p \times p$ matrices.

PROPOSITION 5. *Suppose that the dimension of centrosymmetric matrices is n. When n is odd $(n = 2k+1)$, centrosymmetric matrices are $(k,1)$-mirrorsymmetric matrices; when n is even $(n = 2k)$, centrosymmetric matrices are $(k,0)$-mirrorsymmetric matrices.*

That is to say, all centrosymmetric matrices are the special cases of mirrorsymmetric matrices. The proof is obvious, because when $p = 0$ or 1, mirror matrix $W_{(k,p)}$ is the backward identity matrix J_n. Then (4) becomes $Q = J_nQJ_n$, which is the definition of centrosymmetric matrices [4]. Mirrorsymmetric matrices are centrosymmetric only when $p = 0$ or 1. However, almost all properties of centrosymmetric matrices can be directly grafted onto mirrorsymmetric matrices.

LEMMA 6. *(k, p)-mirrorsymmetric matrix $Q_{(k,p)}$ and $\tilde{Q}$ are orthogonally similar, where*

$$\tilde{Q} = \begin{bmatrix} A_{k\times k} + C_{k\times k} & \sqrt{2}B_{k\times p} & 0_{k\times k} \\ \sqrt{2}D_{p\times k} & E_{p\times p} & 0_{p\times k} \\ 0_{k\times k} & 0_{k\times p} & J_k(A_{k\times k} - C_{k\times k})J_k \end{bmatrix}. \tag{6}$$

Proof. The matrix

$$K = \frac{1}{\sqrt{2}} \begin{bmatrix} I_k & & -J_k \\ & \sqrt{2}I_p & \\ J_k & & I_k \end{bmatrix} \tag{7}$$

is clearly orthogonal, and multiplication gives $K^TQ_{(k,p)}K = K^{-1}Q_{(k,p)}K = \tilde{Q}$. ∎

Especially, $W_{(k,p)} \in \Omega_{(k,p)}$ and

$$K^TW_{(k,p)}K = \begin{bmatrix} I_k & & \\ & I_p & \\ & & -I_k \end{bmatrix};$$

i.e., K is the eigenvector matrix of mirror matrix $W_{(k,p)}$ and the mirror matrix has $k+p$ repeated eigenvalues 1 and k repeated eigenvalues -1. Because the eigenvalues of the mirror matrix is repeated, the eigenvector matrix is not unique. K is the simplest one.

PROPOSITION 7. *The (k, p)-mirrorsymmetric matrix has $k + p$ mirrorsymmetric eigenvectors and k skew-mirrorsymmetric eigenvectors when it is diagonalizable.*

Proof. Using Lemma 6, $K^{-1}Q_{(k,p)}K = \tilde{Q}$. Suppose

$$G^{-1} \begin{bmatrix} A_{k\times k} + C_{k\times k} & \sqrt{2}B_{k\times p} \\ \sqrt{2}D_{p\times k} & E_{p\times p} \end{bmatrix} G = \mathrm{Diag}(\lambda_i^e),$$

$$H^{-1}(A_{k\times k} - C_{k\times k})H = \mathrm{Diag}(\lambda_i^o).$$

Then if $\tilde{S} = \begin{bmatrix} G & \\ & J_kHJ_k \end{bmatrix}$, it is clear that $\tilde{S}^{-1} = \begin{bmatrix} G^{-1} & \\ & J_kH^{-1}J_k \end{bmatrix}$ and

$$\tilde{S}^{-1}K^{-1}QK\tilde{S} = \tilde{S}^{-1}\tilde{Q}\tilde{S} = \begin{bmatrix} \mathrm{Diag}(\lambda_i^e) & \\ & J_k\mathrm{Diag}(\lambda_i^o)J_k \end{bmatrix}.$$

It follows that the columns of $K\tilde{S}$ are the eigenvectors of Q. Further suppose

$$G = \begin{bmatrix} G_{k\times k} & G_{k\times p} \\ G_{p\times k} & G_{p\times p} \end{bmatrix},$$

$$H = [H_{k\times k}].$$

Then the eigenvectors matrix of (k,p)-mirrorsymmetric matrix $Q_{(k,p)}$ is S:

$$S = K\tilde{S} = \frac{1}{\sqrt{2}} \begin{bmatrix} G_{k\times k} & G_{k\times p} & -H_{k\times k}J_k \\ \sqrt{2}G_{p\times k} & \sqrt{2}G_{p\times p} & 0_{p\times k} \\ J_kG_{k\times k} & J_kG_{k\times p} & J_kH_{k\times k}J_k \end{bmatrix}. \tag{8}$$

From Definition 2, we know that the first $k+p$ eigenvectors are mirrorsymmetric and the second k eigenvectors are skew-mirrorsymmetric.

LEMMA 8. *If matrices $P \in \Omega_{(k,p)}$ and $Q \in \Omega_{(k,p)}$, then*

(a) *$\alpha P + \beta Q \in \Omega_{(k,p)}$ for any complex α, β;*

(b) *$P^T \in \Omega_{(k,p)}$;*

(c) *if $\det(P) \neq 0$, then $P^{-1} \in \Omega_{(k,p)}$;*

(d) *$PQ \in \Omega_{(k,p)}$, especially when $P = W_{(k,p)}$ or $Q = W_{(k,p)}$;*

(e) *$W_{(k,p)}P = PW_{(k,p)}$.*

The proofs are elementary and are omitted.

3. Application on odd/even-mode decomposition of mirrorsymmetric MTLs. Multiconductor transmission lines (MTLs) is a system of $(n+1)$-conductor lines which are parallel to z-axis. The MTL equations for frequency domain analysis are

$$\frac{dv}{dz} = -Zi, \tag{9a}$$

$$\frac{di}{dz} = -Yv, \tag{9b}$$

where $v = (v_1, v_2, \ldots, v_n)^T$ are the line-voltages with respect to the reference conductor—the zeroth conductor (ground conductor)—and $i = (i_1, i_2, \ldots, i_n)^T$ are the line currents. Generally, the $n \times n$ complex PUL impedance matrix Z and admittance matrix Y are symmetric ($Z^T = Z, Y^T = Y$).

The general MTL with mirrorsymmetric structure is shown in Figure 1, where the right k lines are the mirror images of the left k lines and there are p lines on the mirror plane, i.e., $n = 2k + p$ ($k \geq 1, p \geq 0$). For the mirrorsymmetric structure, the PUL impedance matrix Z and admittance matrix Y are (k,p)-mirrorsymmetric matrices.

$$A = \begin{bmatrix} A_{ll} & A_{lc} & A_{lr} \\ A_{cl} & A_{cc} & A_{cl}J_k \\ J_kA_{lr}J_k & J_kA_{lc} & J_kA_{ll}J_k \end{bmatrix}, \tag{10}$$

where A denotes Z and Y. The subscripts l, r, c denote the *left*, *right*, and *central* parts of the mirrorsymmetric structure. Though we can transform mirrorsymmetric PUL matrices Z and Y by using K, the factor $\sqrt{2}$ in (6) is difficult to explain physically. In view of the two difference variables of voltages and currents, we define two transforming matrices $T_{V,I} = T(\kappa_{1,2})$:

$$v = T_V\tilde{v}, \tag{11a}$$

$$i = T_I\tilde{i}, \tag{11b}$$

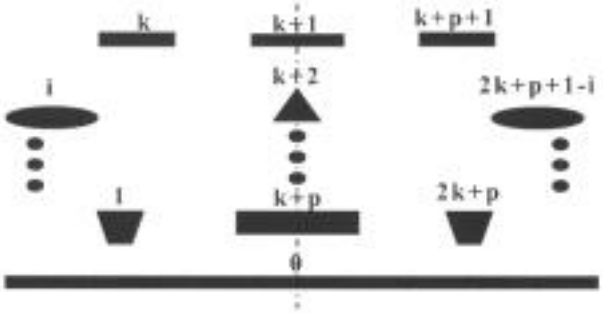


FIG. 1. *General MTL structure with mirror symmetry.*

where $T(\kappa)$ is the eigenvectors matrix of (k,p)-mirror matrix $W_{(k,p)}$, while $\kappa \neq 0$.

$$T(\kappa) = \begin{bmatrix} I_k & & -J_k \\ & \kappa I_p & \\ J_k & & I_k \end{bmatrix}. \tag{12}$$

Then the transforming voltages and currents are given by (13a), (13b), where $(T(\kappa))^{-1} = 0.5(T(2\kappa^{-1}))^T$:

$$\tilde{v} = T_V^{-1} v = \begin{bmatrix} \dfrac{v_l + J_k v_r}{2} \\ \kappa_1^{-1} v_c \\ \dfrac{v_r - J_k v_l}{2} \end{bmatrix} = \begin{bmatrix} v^e \\ -J_k v^o \end{bmatrix}, \tag{13a}$$

$$\tilde{i} = T_I^{-1} i = \begin{bmatrix} \dfrac{i_l + J_k i_r}{2} \\ \kappa_2^{-1} i_c \\ \dfrac{i_r - J_k i_l}{2} \end{bmatrix} = \begin{bmatrix} i^e \\ -J_k i^o \end{bmatrix}, \tag{13b}$$

where $v_l = (v_1, \ldots, v_k)^T$, $v_c = (v_{k+1}, \ldots, v_{k+p})^T$, and $v_r = (v_{k+p+1}, \ldots, v_n)^T$ are the line-voltages of the left, central, and right conductors shown in Figure 1. Three currents vectors are denoted as i_l, i_c, and i_r. Obviously, $\tilde{v}$ and $\tilde{i}$ include two parts: even-mode and odd-mode parts. For those mirrorsymmetric structures, $\kappa_1 = 1$ and $\kappa_2 = 2$ have definite physical sense—voltages on central conductors are unchanged ($\kappa_1^{-1} = 1$), but currents are divided into two equal parts ($\kappa_2^{-1} = 0.5$) when a magnetic wall is placed on the symmetric plane (even-mode parts). Substituting transformation

(13a), (13b) into (9a), (9b) yields

(14a)

$$\tilde{Z} = T_V^{-1} Z T_I = \mathrm{Diag}(Z^e, J_k Z^o J_k) = \begin{bmatrix} Z_{ll} + Z_{lr} J_k & 2Z_{lc} & \\ 2Z_{cl} & 2Z_{cc} & \\ & & J_k(Z_{ll} - Z_{lr} J_k) J_k \end{bmatrix},$$

(14b)

$$\tilde{Y} = T_I^{-1} Y T_V = \mathrm{Diag}(Y^e, J_k Y^o J_k) = \begin{bmatrix} Y_{ll} + Y_{lr} J_k & Y_{lc} & \\ Y_{cl} & 0.5 Y_{cc} & \\ & & J_k(Y_{ll} - Y_{lr} J_k) J_k \end{bmatrix}.$$

Here $\kappa_1 = 1$ and $\kappa_2 = 2$ are assumed. The odd/even-mode decomposition results (14a), (14b) can be testified by using the well-known electric/magnetic-wall analysis on mirrorsymmetric waveguide structures. In fact, if physical meaning is not required, $\kappa_{1,2}$ can be any nonzero factors. $\kappa_1 \kappa_2 = 2$ is needed if $\tilde{Z}^T = \tilde{Z}$ and $\tilde{Y}^T = \tilde{Y}$ are wanted. The odd/even-mode decomposition scheme proposed above reduces the order of MTL from n to k and $k + p$. Odd-mode and even-mode MTL equations can be solved independently.

PROPOSITION 9. *Mirrorsymmetric MTL systems can be divided into even-mode and odd-mode subsystems.*

PROPOSITION 10. *If the eigenvector matrices $S_{V,I}^{o,e}$ of odd/even-mode PUL matrices $Z^{o,e}$, $Y^{o,e}$ are found, then the eigenvector matrices $S_{V,I}$ of Z, Y can be gotten from $S_{V,I}^{o,e}$, i.e.,*

$$S_{V,I} = \begin{bmatrix} S^e_{V,I(k\times k)} & S^e_{V,I(k\times p)} & -S^o_{V,I(k\times k)} J_k \\ \kappa_{1,2} S^e_{V,I(p\times k)} & \kappa_{1,2} S^e_{V,I(p\times p)} & 0_{(p\times k)} \\ J_k S^e_{V,I(k\times k)} & J_k S^e_{V,I(k\times p)} & J_k S^o_{V,I(k\times k)} J_k \end{bmatrix}. \tag{15}$$

The proof of Proposition 10 can be directly obtained from relations $S_{V,I} = T_{V,I} \tilde{S}_{V,I}$, where

$$\tilde{S}_{V,I} = \mathrm{Diag}(S^e_{V,I}, J_k S^o_{V,I} J_k) = \begin{bmatrix} S^e_{V,I(k\times k)} & S^e_{V,I(k\times p)} & \\ S^e_{V,I(p\times k)} & S^e_{V,I(p\times p)} & \\ & & J_k S^o_{V,I(k\times k)} J_k \end{bmatrix}. \tag{16}$$

Here S_V is the eigenvector matrix of ZY, and S_I is eigenvector matrix of YZ. They simultaneously diagonalize PUL matrices Z and Y, i.e.,

$$D_z = S_V^{-1} Z S_I, \tag{17a}$$

$$D_y = S_I^{-1} Y S_V, \tag{17b}$$

where diagonal matrices $D_z = \mathrm{Diag}(z_1, z_2, \ldots, z_n)$ and $D_y = \mathrm{Diag}(y_1, y_2, \ldots, y_n)$ are decoupled circuit PUL parameters. And the eigenvalues of ZY or YZ are given by

$$D_\gamma^2 = D_z D_y, \tag{18}$$

where $D_\gamma = \mathrm{Diag}(\gamma_1, \gamma_2, \ldots, \gamma_3)$ are propagation constants of each mode. If PUL matrices are symmetric, the two eigenvector matrices must be biorthogonal when the eigenvalues ($\gamma_j^2 = z_j y_j, j = 1, 2, \ldots, n$) are distinct because of $[YZ]^T = Z^T Y^T = [ZY]$.

$$S_V^T S_I = D_d^2, \tag{19}$$

where $D_d = \mathrm{Diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix. There are several methods to normalize the eigenvector matrices S_V and S_I. Usually, they are normalized to satisfy the relation $S_V^T S_I = I$ [6, 7]. For example, they can be normalized as $S_V D_d^{-1}$ and $S_I D_d^{-1}$. Another normalization method is to make the diagonal elements of $S_{V,I}$ to be 1's, i.e., $S_{V,jj} = S_{I,jj} = 1$ ($j = 1, 2, \ldots, n$). Such normalization gives a simple calculation of decoupled circuit parameters:

$$z_j = \sum_{k=1}^{n} Z_{jk} S_{I,kj}, \tag{20a}$$

$$y_j = \sum_{k=1}^{n} Y_{jk} S_{V,kj}. \tag{20b}$$

The decoupled circuit parameters are given by the inner product of the row vectors of Z, Y and the corresponding column vectors of $S_{I,V}$. If this normalization method is adopted, the form of eigenvector matrices shown in (15) becomes

$$S_{V,I} = \begin{bmatrix} S^e_{V,I(k\times k)} & \kappa_{1,2}^{-1} S^e_{V,I(k\times p)} & -S^o_{V,I(k\times k)} J_k \\ \kappa_{1,2} S^e_{V,I(p\times k)} & S^e_{V,I(p\times p)} & 0_{(p\times k)} \\ J_k S^e_{V,I(k\times k)} & \kappa_{1,2}^{-1} J_k S^e_{V,I(k\times p)} & J_k S^o_{V,I(k\times k)} J_k \end{bmatrix}, \tag{21}$$

where the diagonal elements of $S^{e,o}_{V,I}$ have already been normalized to 1's.

4. Some symmetric MTL examples. Though the concept of odd/even-modes is well known in the electric/magnetic-wall analysis on mirrorsymmetric waveguide structures, and there are some isolated study cases of mirrorsymmetric MTL equations [6, 7], the general theory of decomposing the MTL equations with the structures was not proposed until the definition of mirrorsymmetric matrices. In section 3, we have given a general theory of odd/even-mode decomposition of mirrorsymmetric MTL equations. In this section, some examples are discussed, and it is proved that rotational symmetric MTL equations can also be solved from the view of mirror symmetry.

Taking the case of $p = 1$, $k = 1$, $n = 3$ as the first example, we can give the analytical modes-decomposition solutions under the light of the general theory proposed in section 3. Another analytical solution is given in [6], but the procedure is complex and not as clear as the following:

$$Z = \begin{bmatrix} Z_1 & Z_{m1} & Z_{m2} \\ Z_{m1} & Z_2 & Z_{m1} \\ Z_{m2} & Z_{m1} & Z_1 \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 & Y_{m1} & Y_{m2} \\ Y_{m1} & Y_2 & Y_{m1} \\ Y_{m2} & Y_{m1} & Y_1 \end{bmatrix}, \tag{22}$$

$$Z^e = \begin{bmatrix} Z_1 + Z_{m2} & 2Z_{m1} \\ 2Z_{m1} & 2Z_2 \end{bmatrix}, \quad Y^e = \begin{bmatrix} Y_1 + Y_{m2} & Y_{m1} \\ Y_{m1} & 0.5Y_2 \end{bmatrix}, \tag{23}$$

$$Z^o = [Z_1 - Z_{m2}], \quad Y^o = [Y_1 - Y_{m2}]. \tag{24}$$

The eigenvector matrices $S^e_{V,I}$ of even-modes can be gotten analytically, as shown in [8] because the order of even-mode MTL equations is 2.

$$S^e_V = \begin{bmatrix} 1 & -\beta^e \\ -\alpha^e & 1 \end{bmatrix}, \quad S^e_I = \begin{bmatrix} 1 & \alpha^e \\ \beta^e & 1 \end{bmatrix}. \tag{25}$$

Then the eigenvector matrices of the original MTL system can be represented by

$$S_V = \begin{bmatrix} 1 & -\beta^e & -1 \\ -\alpha^e & 1 & 0 \\ 1 & -\beta^e & 1 \end{bmatrix}, \quad S_I = \begin{bmatrix} 1 & 0.5\alpha^e & -1 \\ 2\beta^e & 1 & 0 \\ 1 & 0.5\alpha^e & 1 \end{bmatrix}. \tag{26}$$

From (20a), (20b) the decoupled circuit parameters are given by

$$z_1 = (Z_1 + Z_{m2}) + 2\beta^e Z_{m1}, \quad z_2 = Z_2 + \alpha^e Z_{m1}, \quad z_3 = Z_1 - Z_{m2}, \tag{27a}$$

$$y_1 = (Y_1 + Y_{m2}) - \alpha^e Y_{m1}, \quad y_2 = Y_2 - 2\beta^e Y_{m1}, \quad y_3 = Y_1 - Y_{m2}. \tag{27b}$$

The first two parameters correspond to even-modes and the last one corresponds to the odd-mode. Then the propagation constants of the MTL equations are given by

$$\gamma_i = \sqrt{z_i y_i} \qquad (i = 1, 2, 3). \tag{28}$$

The same procedure can give analytical solutions for the case of $p = 0$, $k = 2$, $n = 4$ and are omitted.

Next, we'll discuss a very special example—rotational symmetric MTL equation. The PUL matrices of MTL equations with rotational symmetric structure are symmetric circulant matrices [7].

$$Z = circ(z_0, z_{m1}, z_{m2}, \ldots, z_{m(n-1)}), \tag{29a}$$

$$Y = circ(y_0, y_{m1}, y_{m2}, \ldots, y_{m(n-1)}), \tag{29b}$$

where $z_{m(p)} = z_{m(n-p)}$, $y_{m(p)} = y_{m(n-p)}$, $p = 1, 2, \ldots, n-1$. From the view of rotational symmetry, both eigenvector matrices S_V and S_I are equal to the Fourier matrix [9]:

$$F = \frac{1}{\sqrt{n}}[\delta^{(p-1)(q-1)}] \qquad (p, q = 1, 2, \ldots, n), \tag{30}$$

where $\delta = e^{j\frac{2\pi}{n}} (j = \sqrt{-1})$. Because $S_V = S_I = F$, from (17a), (17b) we know that the eigenvalues of Z and Y are the decoupled circuit parameters. The eigenvalues of Z are given by

$$\lambda_p|_{n=2k} = z_0 + 2\sum_{q=1}^{k-1} z_{m(q)} \cos\frac{2pq\pi}{n} + (-1)^p z_{m(k)}, \tag{31a}$$

$$\lambda_p|_{n=2k+1} = z_0 + 2\sum_{q=1}^{k} z_{m(q)} \cos\frac{2pq\pi}{n} \qquad (p = 0, 1, \ldots, n-1). \tag{31b}$$

Because the eigenvalues of symmetric circulant matrices are repeated ($\lambda_{n-p} = \lambda_p$), the eigenvector matrix is not unique. The Fourier matrix is one that comes from the view of rotational symmetry.

PROPOSITION 11. *A symmetric circulant matrix is a symmetric centrosymmetric matrix.*

Proof. If circulant matrix Z is symmetric, then $z_{m(p)} = z_{m(n-p)}$ $(p = 1, 2, \ldots, n-1)$. Because $JZJ = circ(z_0, z_{m(n-1)}, z_{m(n-2)}, \ldots, z_{m(1)}) = circ(z_0, z_{m(1)}, z_{m(2)}, \ldots, z_{m(n-1)}) = Z$, Z is a centrosymmetric matrix. □

From Proposition 5, we know that Z is also mirrorsymmetric. Since the PUL matrices of rotational symmetric MTL equations are mirrorsymmetric, even if the rotational symmetric structure is not mirrorsymmetric, the modes can also be classified into k "odd-modes" and $k+p$ "even-modes" $(p = 0, 1)$. But the column vectors of the Fourier matrix used in [7, 9] are not mirrorsymmetric or skew-mirrorsymmetric. Here we give another set of orthogonal eigenvectors, which are mirrorsymmetric or skew-mirrorsymmetric:

$$s_j^e = \left(\cos\frac{(n-1)j\pi}{n}, \cos\frac{(n-3)j\pi}{n}, \ldots, \cos\frac{(n-3)j\pi}{n}, \cos\frac{(n-1)j\pi}{n}\right)^T, \tag{32a}$$

$$s_j^o = \left(\sin\frac{(n-1)j\pi}{n}, \sin\frac{(n-3)j\pi}{n}, \ldots, -\sin\frac{(n-3)j\pi}{n}, -\sin\frac{(n-1)j\pi}{n}\right)^T \tag{32b}$$

$$(j = 0, 1, 2, \ldots, k, \quad n = 2k, 2k+1).$$

Zero-vectors s_0^o and $s_k^e (n = 2k)$ in (32a), (32b) have no meaning and are not used. The first five normalized eigenvector matrices under a certain mode order are given below:

$$S_{2\times2} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad S_{3\times3} = \begin{bmatrix} 1 & -0.5 & -1 \\ 1 & 1 & 0 \\ 1 & -0.5 & 1 \end{bmatrix}, \quad S_{4\times4} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \end{bmatrix},$$

$$S_{5\times5} = \begin{bmatrix} 1 & \alpha-1 & -(\alpha+1)/2 & -\alpha & -1 \\ 1 & 1 & \alpha/2 & -1 & \alpha \\ 1 & -2\alpha & 1 & 0 & 0 \\ 1 & 1 & \alpha/2 & 1 & -\alpha \\ 1 & \alpha-1 & -(\alpha+1)/2 & \alpha & 1 \end{bmatrix}_{\alpha=\frac{\sqrt{5}-1}{2}},$$

$$S_{6\times6} = \begin{bmatrix} 1 & -0.5 & -1 & 1 & -0.5 & -1 \\ 1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -0.5 & 1 & -1 & -0.5 & -1 \\ 1 & -0.5 & 1 & 1 & 0.5 & 1 \\ 1 & 1 & 0 & 0 & 1 & -1 \\ 1 & -0.5 & -1 & -1 & 0.5 & 1 \end{bmatrix}.$$

From the view of rotational symmetry (circulant matrix), the eigenvector matrix is the Fourier matrix [7, 9], which is complex. From the view of the mirrorsymmetric matrix, the eigenvector matrix is real. Two views perfect the theory of rotational and mirror symmetries.

Two examples discussed above all involve the special cases of mirrorsymmetric matrices—centrosymmetric matrices. Thus the decomposition can be gotten directly by using the properties of centrosymmetric matrices [2, 3, 4, 5]. But if the number of central conductors is greater than one (as shown in Figure 1, if $p > 1$), the interaction matrices can no longer be represented by centrosymmetric matrices. Mirrorsymmetric matrices defined in section 2 have to be considered. Taking the case of $p = 2$, $k = 2$, $n = 6$ as the example, the odd/even-mode decomposition scheme is given by

$$Z = \begin{bmatrix} Z_1 & Z_{m1} & Z_{m3} & Z_{m4} & Z_{m7} & Z_{m8} \\ Z_{m1} & Z_2 & Z_{m5} & Z_{m6} & Z_{m9} & Z_{m7} \\ Z_{m3} & Z_{m5} & Z_3 & Z_{m2} & Z_{m5} & Z_{m3} \\ Z_{m4} & Z_{m6} & Z_{m2} & Z_4 & Z_{m6} & Z_{m4} \\ Z_{m7} & Z_{m9} & Z_{m5} & Z_{m6} & Z_2 & Z_{m1} \\ Z_{m8} & Z_{m7} & Z_{m3} & Z_{m4} & Z_{m1} & Z_1 \end{bmatrix}, \tag{33a}$$

$$Y = \begin{bmatrix} Y_1 & Y_{m1} & Y_{m3} & Y_{m4} & Y_{m7} & Y_{m8} \\ Y_{m1} & Y_2 & Y_{m5} & Y_{m6} & Y_{m9} & Y_{m7} \\ Y_{m3} & Y_{m5} & Y_3 & Y_{m2} & Y_{m5} & Y_{m3} \\ Y_{m4} & Y_{m6} & Y_{m2} & Y_4 & Y_{m6} & Y_{m4} \\ Y_{m7} & Y_{m9} & Y_{m5} & Y_{m6} & Y_2 & Y_{m1} \\ Y_{m8} & Y_{m7} & Y_{m3} & Y_{m4} & Y_{m1} & Y_1 \end{bmatrix}, \tag{33b}$$

$$Z^e_{4\times4} = \begin{bmatrix} Z_1 + Z_{m8} & Z_{m1} + Z_{m7} & 2Z_{m3} & 2Z_{m4} \\ Z_{m1} + Z_{m7} & Z_2 + Z_{m9} & 2Z_{m5} & 2Z_{m6} \\ 2Z_{m3} & 2Z_{m5} & 2Z_3 & 2Z_{m2} \\ 2Z_{m4} & 2Z_{m6} & 2Z_{m2} & 2Z_4 \end{bmatrix}, \tag{34a}$$

$$Y^e_{4\times4} = \begin{bmatrix} Y_1 + Y_{m8} & Y_{m1} + Y_{m7} & Y_{m3} & Y_{m4} \\ Y_{m1} + Y_{m7} & Y_2 + Y_{m9} & Y_{m5} & Y_{m6} \\ Y_{m3} & Y_{m5} & 0.5Y_3 & 0.5Y_{m2} \\ Y_{m4} & Y_{m6} & 0.5Y_{m2} & 0.5Y_4 \end{bmatrix}, \tag{34b}$$

(35)

$$Z^o_{2\times2} = \begin{bmatrix} Z_1 - Z_{m8} & Z_{m1} - Z_{m7} \\ Z_{m1} - Z_{m7} & Z_2 - Z_{m9} \end{bmatrix}, \quad Y^o_{2\times2} = \begin{bmatrix} Y_1 - Y_{m8} & Y_{m1} - Y_{m7} \\ Y_{m1} - Y_{m7} & Y_2 - Y_{m9} \end{bmatrix},$$

$$S_V = \begin{bmatrix} 1 & S^e_{V,12} & S^e_{V,13} & S^e_{V,14} & \beta^o & -1 \\ S^e_{V,21} & 1 & S^e_{V,23} & S^e_{V,24} & -1 & \alpha^o \\ S^e_{V,31} & S^e_{V,32} & 1 & S^e_{V,34} & 0 & 0 \\ S^e_{V,41} & S^e_{V,42} & S^e_{V,43} & 1 & 0 & 0 \\ S^e_{V,21} & 1 & S^e_{V,23} & S^e_{V,24} & 1 & -\alpha^o \\ 1 & S^e_{V,12} & S^e_{V,13} & S^e_{V,14} & -\beta^o & 1 \end{bmatrix}, \tag{36a}$$

$$S_I = \begin{bmatrix} 1 & S^e_{I,12} & 0.5S^e_{I,13} & 0.5S^e_{I,14} & -\alpha^o & -1 \\ S^e_{I,21} & 1 & 0.5S^e_{I,23} & 0.5S^e_{I,24} & -1 & -\beta^o \\ 2S^e_{I,31} & 2S^e_{I,32} & 1 & S^e_{I,34} & 0 & 0 \\ 2S^e_{I,41} & 2S^e_{I,42} & S^e_{I,43} & 1 & 0 & 0 \\ S^e_{I,21} & 1 & 0.5S^e_{I,23} & 0.5S^e_{I,24} & 1 & \beta^o \\ 1 & S^e_{I,12} & 0.5S^e_{I,13} & 0.5S^e_{I,14} & \alpha^o & 1 \end{bmatrix}. \tag{36b}$$

Here, PUL matrices Z, Y, $Z^{o,e}$, $Y^{o,e}$ are symmetric.

Rotational symmetric structure, discussed in the second example, needn't be mirrorsymmetric. However, since the PUL matrices of rotational symmetry, symmetric circulant matrices, are also mirrorsymmetric, we may as well suppose for convenience that rotational symmetric structures discussed here also have mirror symmetry. When n is even, the mirror plane may be chosen cutting no conductors, i.e., $p = 0$ (see Figure 2(a); zeroth conductor is not shown in the figure). This case has been discussed above. Another choice is that of the mirror plane cutting two conductors, i.e., $p = 2$ (see Figure 2(b), noticing the different line order of the two cases). Now the PUL matrices Z and Y are neither circulant matrices nor centrosymmetric matrices. However, they are mirrorsymmetric matrices, where $n = 2k + 2$ and

$$Z_{ll(kk)} = \begin{bmatrix} Z_0 & Z_{m1} & \bullet & Z_{m(k-1)} \\ Z_{m1} & Z_0 & \bullet & Z_{m(k-2)} \\ \bullet & \bullet & \bullet & \bullet \\ Z_{m(k-1)} & Z_{m(k-2)} & \bullet & Z_0 \end{bmatrix}, \tag{37a}$$

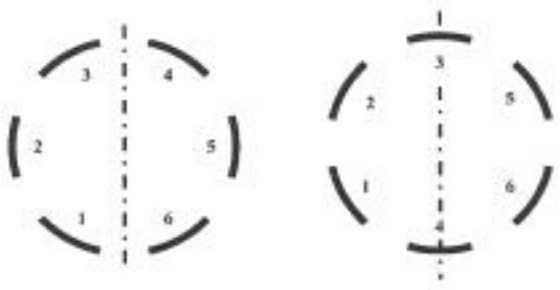
$$Z_{lr(kk)} = \begin{bmatrix} Z_{m(k+1)} & Z_{mk} & \bullet & Z_{m2} \\ Z_{mk} & Z_{m(k+1)} & \bullet & Z_{m3} \\ \bullet & \bullet & \bullet & \bullet \\ Z_{m2} & Z_{m3} & \bullet & Z_{m(k+1)} \end{bmatrix}, \tag{37b}$$

$$Z_{cl(2k)} = \begin{bmatrix} Z_{mk} & Z_{m(k-1)} & \bullet & Z_{m1} \\ Z_{m1} & Z_{m2} & \bullet & Z_{mk} \end{bmatrix}, \tag{37c}$$

$$Z_{cc(22)} = \begin{bmatrix} Z_0 & Z_{m(k+1)} \\ Z_{m(k+1)} & Z_0 \end{bmatrix}. \tag{37d}$$

There are a total of $k + 2$ distinct eigenvalues and k repeated eigenvalues that correspond to even-modes and odd-modes for this case. The corresponding orthogonal mirrorsymmetric and skew-mirrorsymmetric eigenvectors are given by

$$\begin{aligned} s^e_j = \Bigg(& \cos\frac{j\pi}{k+1}, \cos\frac{j2\pi}{k+1}, \ldots, \cos\frac{jk\pi}{k+1}, (-1)^j, 1, \\ & \cos\frac{jk\pi}{k+1}, \ldots, \cos\frac{j2\pi}{k+1}, \cos\frac{j\pi}{k+1} \Bigg)^T_{j=0,1,\ldots,k+1}, \end{aligned} \tag{38a}$$



(a) $n = 6$, $p = 0$, $k = 3$ (b) $n = 6$, $p = 2$, $k = 2$

FIG. 2. *Rotational symmetry with even n.*

$$S_j^o = \left(\sin\frac{j\pi}{k+1}, \sin\frac{j2\pi}{k+1}, \ldots, \sin\frac{jk\pi}{k+1}, 0, 0, \right.$$
$$\left. -\sin\frac{jk\pi}{k+1}, \ldots, -\sin\frac{j2\pi}{k+1}, -\sin\frac{j\pi}{k+1}\right)^T_{j=1,2,\ldots,k}. \tag{38b}$$

The first two normalized eigenvector matrices by a certain mode order are shown below ($p = 2, k = 1, 2$):

$$S_{4\times4} = \begin{bmatrix} 1 & -1 & 0 & -1 \\ 1 & 1 & -1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{bmatrix}, \quad S_{6\times6} = \begin{bmatrix} 1 & -1 & -0.5 & 0.5 & -1 & -1 \\ 1 & 1 & -0.5 & -0.5 & -1 & 1 \\ 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & -0.5 & -0.5 & 1 & -1 \\ 1 & -1 & -0.5 & 0.5 & 1 & 1 \end{bmatrix}.$$

Here $S_V = S_I = S$. We can verify $S_{6\times6}$ by using (33a)–(36b). When $n = 6$ (as shown in Figure 2(b)), there are 4 even-modes and 2 odd-modes. Compared with Figure 2(a), with the same structure but different mirror plane, there are 3 even-modes and 3 odd-modes. Although the structure and the modes are the same, the classification is different. But for general mirrorsymmetric structures as shown in Figure 1, there is no way to transform mirrorsymmetric matrices into centrosymmetric matrices when the number of central conductors is greater than 1 ($p > 1$). After all, centrosymmetric matrices are special cases of mirrorsymmetric matrices.

5. Conclusion. Mirrorsymmetric matrices, which are the interaction matrices of mirrorsymmetric structures, are defined in this paper. Some basic properties, especially eigenvectors of mirrorsymmetric matrices, are explored. It is proved that centrosymmetric matrices are special cases of mirrorsymmetric matrices, i.e., mirrorsymmetric matrices are centrosymmetric matrices only when $p = 0$ or 1, where p is the component number on the mirror plane. However, almost all properties of centrosymmetric matrices can be directly generalized to mirrorsymmetric matrices.

The application of mirrorsymmetric matrices on odd/even-mode decomposition of mirrorsymmetric MTL equations is investigated in detail. The order of MTL equations is reduced from n to k and $k+p$. Two transforming matrices $T_{V,I}$ are defined to

give a definite physical explanation on odd/even-mode decomposition. Some examples are discussed, especially rotational symmetric MTL equations, which can also be treated from the view of mirror symmetry. Because the interaction matrices of mirrorsymmetric structure (as shown in Figure 1) are (k, p)-mirrorsymmetric matrices, it is believed that mirrorsymmetric matrices will have wide applications in many scientific fields, since mirror symmetry is commonly encountered in real physical systems.

REFERENCES

[1] W. T. Weeks, *Exploiting symmetry in electrical packaging analysis*, IBM J. Res. Develop., 23 (1979), pp. 669–674.
[2] A. C. Aitken, *Determinants and Matrices*, Oliver and Boyd, Edinburgh, UK, 1958.
[3] A. L. Andrew, *Centrosymmetric matrices*, SIAM Rev., 40 (1998), pp. 697–698.
[4] A. Cantoni, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.
[5] J. R. Weaver, *Centrosymmetric (cross-symmetric) matrices, their properties, eigenvalues, and eigenvectors*, Amer. Math. Monthly, 92 (1985), pp. 711–717.
[6] V. K. Tripathi, *On the analysis of symmetrical three-line microstrip circuits*, IEEE Trans. Microwave Theory Tech., 25 (1977), pp. 726–729.
[7] C. R. Paul, *Decoupling the multiconductor transmission line equations*, IEEE Trans. Microwave Theory Tech., 44 (1996), pp. 1429–1440.
[8] G.-L. Li and Z.-H. Feng, *Line-modes decomposition of three conductor transmission line equations*, in Proceedings of the 2000 ASIA-PACIFIC Microwave Conference, Sydney, Australia, 2000, pp. 1031–1034.
[9] F.-Z. Zhang, *Matrix Theory—Basic Results and Techniques*, Springer, New York, 1999.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 91–105

E-OPTIMAL SPRING BALANCE WEIGHING DESIGNS FOR $n \equiv -1 \pmod 4$ OBJECTS*

MICHAEL G. NEUBAUER† AND WILLIAM WATKINS†

Abstract. Let $n \equiv -1 \pmod 4$ be a positive integer with $n \geq 7$ and let $M_{m,n}(0,1)$ be the set of all $m \times n$ (0,1)-matrices. Let $E(m,n)$ be the largest minimum eigenvalue for a matrix X^TX with $X \in M_{m,n}(0,1)$. Let $m = nt + r$, where $0 \leq r < n$. We show that for $r \neq n-4$,

$$E(nt+r,n) \leq \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor,$$

with equality for sufficiently large m. For $r = n-4$, we show that

$$E(nt+r,n) \leq \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n}$$

and for sufficiently large m,

$$\left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n} - \frac{4}{tn^3} < E(nt+r).$$

Similar inequalities are given for the case $n = 3$.

Key words. E-optimal weighing design, Hadamard design, spring balance scale

AMS subject classifications. Primary, 05B20, 62K05; Secondary, 15A18, 15A36, 15A42

PII. S0895479801393939

1. Introduction.

In this paper, we discuss E-optimality for spring balance weighing designs. Suppose we wish to estimate the weights of n objects. A *weighing* consists of placing a subset of the objects on a spring scale, which gives an estimate of the total weight of these objects. A weighing can be coded into a $\{0,1\}$-vector $(w_1, \ldots, w_n)$ by defining $w_i = 1$ if object i is placed on the scale, and $w_i = 0$ otherwise. In this way a series of m weighings can be represented by an $m \times n$ (0,1)-matrix X called a *design matrix*; each row of X corresponds to a weighing of the n objects.

Among all possible competing $m \times n$ design matrices X, optimality is often measured in terms of the eigenvalues of the Gram matrix (also known as the information matrix) X^TX. For example, X is *D-optimal* if $\det X^TX$ is maximal and *A-optimal* if $\operatorname{trace}(X^TX)^{-1}$ is minimal among all $m \times n$ design matrices. (See [Puk] for a full discussion of statistical designs and various types of optimality.)

E-optimality can be described in terms of the minimum eigenvalue, $\lambda_{\min}(X^TX)$ of the Gram matrix X^TX. Let $m \geq n$ be an integer and $M_{m,n}(0,1)$ be the set of all $m \times n$ (0,1)-matrices (design matrices). The first question about E-optimality is to determine how large this minimum eigenvalue can be—that is, to determine the value of

$$E(m,n) = \max\{\lambda_{\min}(X^TX) : X \in M_{m,n}(0,1)\}.$$

The following upper bounds for $\lambda_{\min}(X^TX)$ were established by Cheng [Che] using the statistical idea of an approximate design, which was developed by Kiefer

*Received by the editors August 20, 2001; accepted for publication (in revised form) by R. Brualdi January 1, 2002; published electronically June 12, 2002.
http://www.siam.org/journals/simax/24-1/39393.html
†Department of Mathematics, California State University Northridge, Northridge, CA 91330 (michael.neubauer@csun.edu, bill.watkins@csun.edu).

[Kie] in 1974:

$$\lambda_{\min}(X^TX) \le \frac{m(n+1)}{4n} \quad \text{if } n \text{ is odd,} \tag{1}$$

$$\lambda_{\min}(X^TX) \le \frac{mn}{4(n-1)} \quad \text{if } n \text{ is even.} \tag{2}$$

Actually Cheng showed that these upper bounds hold for a larger class of matrices and that similar upper bounds hold for a larger class of functions. To describe Cheng's results, let $\Omega = \{(x_1, \dots, x_n)\} : 0 \le x_i \le 1\}$ be the unit cube in $\mathbb{R}^n$. For a probability measure ξ (known as an *approximate design*) on Ω, define the $n \times n$ matrix $M(\xi)$ by

$$M(\xi) = \int_\Omega x^T x \, \xi(dx).$$

Cheng showed that for the class of functions $j_a(M) = n^{-1}(\sum \lambda^a)^{1/a}$ (where $a \le 1, a \ne 0$ and the sum is taken over the eigenvalues λ of M) the maximum value of $j_a(M(\xi))$ occurs if and only if

$$M(\xi) = \begin{cases} \frac{n+1}{4n}(I_n + J_n) & \text{if } n \text{ is odd,} \\ \\ \frac{1}{4(n-1)}(nI_n + (n-2)J_n) & \text{if } n \text{ is even.} \end{cases}$$

(Here, I_n is the $n \times n$ identity matrix and J_n is the $n \times n$ matrix all of whose entries are one.) One of the functions in this family is $j_{-\infty}(M) = \lambda_{\min}(M)$.

Now suppose $X \in M_{m,n}(0,1)$ is a design matrix and let ξ be the probability measure on Ω that assigns the value $1/m$ to each row of X. Then $M(\xi) = (1/m)X^TX$. Now since $\lambda_{\min}(I_n + J_n) = 1$ and $\lambda_{\min}(nI_n + (n-2)J_n) = n$, inequalities (2) and (1) follow from Cheng's result. However, the problem of maximizing $\lambda_{\min}(X^TX)$ over the smaller set $\{(1/m)X^TX : X \in M_{m,n}(0,1)\} \subset \Omega$ is quite different. Indeed, smaller upper bounds are available, and the actual value of $E(m,n)$ can be determined in many cases.

In this paper we are interested in the case where $n \equiv -1 \pmod 4$. We deal with the case $n = 3$ in section 5. Assume now that $n \ge 7$ and that $m = nt + r$, where t is a positive integer and $0 \le r < n$. Then inequality (1) becomes

$$\lambda_{\min}(X^TX) \le \left(\frac{n+1}{4}\right)\left(t + \frac{r}{n}\right). \tag{3}$$

For $X \in M_{m,n}(0,1)$ we obtain smaller upper bounds:

$$\lambda_{\min}(X^TX) \le \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor \tag{4}$$

if $r \ne n-4$ and

$$\lambda_{\min}(X^TX) \le \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n} \tag{5}$$

if $r = n-4$.

Our arguments depend on the congruence class of m modulo n. Thus we assume throughout that $m = nt + r$ with $0 \le r < n$ and that $n = 4p - 1$ for some positive integer $p \ge 2$.

THEOREM 1. *Let $n = 4p-1$ be a positive integer with $p \geq 2$, and let $m = nt+r$, with $0 \leq r < n$. If $r \neq n-4$, then*

$$E(nt+r,n) \leq \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor. \tag{6}$$

If $r = n-4 = 4p-5$, then

$$E(nt+r,n) \leq \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n}. \tag{7}$$

Furthermore, there exists a positive integer m_0 such that, if $m \geq m_0$ and $r \neq n-4$, then equality holds in inequality (6) *and, if $r = n-4$, then*

$$\left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n} - \frac{4}{tn^3} < E(nt+r,n). \tag{8}$$

The second problem about E-optimality is to find design matrices X for which the maximum value of $\lambda_{\min}(X^TX)$ is attained. If $X \in M_{m,n}(0,1)$ is a design matrix and $\lambda_{\min}(X^TX) = E(m,n)$, then X is said to be *E-optimal*. For $n = 7$, $m = 7t+r$, and $r \neq 3$, we construct an $m \times n$ E-optimal matrix in section 4. However, for $n > 7$, our methods are not constructive. Still, we prove that for sufficiently large m and $r \neq n-4$, design matrices X exist such that

$$\lambda_{\min}(X^TX) = \left(\frac{n+1}{4}\right)t + \left\lfloor \frac{r}{4} \right\rfloor,$$

so that equality holds in (6).

Before beginning the proofs, we establish some notation. For each $0 \leq r \leq n-1$, let $a = \left\lfloor \frac{r}{4} \right\rfloor$ so that $r = 4a + a_1$ for $a_1 = 0, 1, 2,$ or 3. For $r \neq n-4$, define

$$\alpha(r) = a$$

and define

$$\alpha(n-4) = a + \frac{1}{n} = p - 2 + \frac{1}{n}.$$

Let e be the n-tuple each of whose entries is one.

2. Upper bounds on $E(m,n)$. In this section we establish inequalities (6) and (7) of Theorem 1 by means of the following lemma.

LEMMA 2. *Let $n = 4p-1$ with $p \geq 2$, let X be a matrix in $M_{nt+r,n}(0,1)$, where $0 \leq r \leq 4p-2$, and let $R = X^TX - ptI_n$. Then there exists $u \in \langle e \rangle^{\perp}$ such that $u^TRu \leq \alpha(r)||u||^2$.*

Proof. Let $R = X^TX - ptI_n$, where $R = (r_{ij})$. For each pair $i \neq j$, the vector $u = \frac{1}{\sqrt{2}}(e_i - e_j)$ is a unit vector in $\langle e \rangle^{\perp}$ and $u^TRu = \frac{1}{2}(r_{ii} + r_{jj} - 2r_{ij})$. If $u^TRu \leq \alpha(r)$, for any pair $i \neq j$, then we are finished. So assume that $\frac{1}{2}(r_{ii} + r_{jj} - 2r_{ij}) > \alpha(r)$ for all $i \neq j$. Since $\alpha(r) \geq a$ and $r_{ii} + r_{jj} - 2r_{ij}$ is an integer, we have

$$r_{ii} + r_{jj} - 2r_{ij} \geq 2a + 1 \tag{9}$$

for all $i \neq j$.

Let $\rho = (r_{11}, \ldots, r_{nn})^T$ and $Q = (q_{ij}) = \rho e^T + e\rho^T - 2R$. Then

$$q_{ij} = r_{ii} + r_{jj} - 2r_{ij}. \tag{10}$$

Now let $u \in \langle e \rangle^\perp$. Since $u^T e = 0$, we have

$$u^T X^T X u = pt||u||^2 + u^T R u = pt||u||^2 - \frac{1}{2} u^T Q u.$$

Thus it is sufficient to show that $u^T Q u \geq -2\alpha(r) u^T u$, for some $u \in \langle e \rangle^\perp$.

Assume that r_{ii} is even for $i \leq k$ and odd for $i > k$. Partition Q as

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

where Q_{11} is $k \times k$ and Q_{22} is $l \times l$, with $k + l = n$. It follows from (10) that the off-diagonal entries in Q_{11} and Q_{22} are even, the entries in Q_{12} and Q_{21} are odd, and from inequality (9) that all off-diagonal entries in Q_{11} and Q_{22} are at least $2a + 2$. (The diagonal entries of Q are zero.) Now let

$$Q_0 = \begin{bmatrix} (2a+2)(J_k - I_k) & (2a+1)J_{k,l} \\ (2a+1)J_{l,k} & (2a+2)(J_l - I_l) \end{bmatrix}.$$

By the above remarks, the matrix $E = Q - Q_0$ has nonnegative entries.

We now obtain an upper bound on $e^T Q e$. Let k_s $(s = 0, 1, \ldots, n)$ denote the number of rows of the $(0,1)$-matrix X that have exactly s ones and $n - s$ zeros. Then $\operatorname{trace} X^T X = \sum s k_s$ and $e^T X^T X e = \sum s^2 k_s$, where the sums are taken over $s = 0, 1, \ldots, n$. Since $\operatorname{trace} I_n = e^T I_n e = n$, we have

$$\begin{aligned} \operatorname{trace} R &= -npt + \sum s k_s, \\ e^T R e &= -npt + \sum s^2 k_s. \end{aligned}$$

Thus

$$\begin{aligned} e^T Q e &= \sum_{i,j} (r_{ii} + r_{jj} - 2r_{ij}) \\ &= 2n \operatorname{trace} R - 2e^T R e \\ &= 2\left((n(n+1) - 2n^2)pt + \sum (ns - s^2) k_s \right) \\ &\leq 2\left(-n(n-1)pt + 2p(2p-1) \sum k_s \right) \qquad (11) \\ &= 2(-(4p-1)(4p-2)pt + 2p(2p-1)((4p-1)t + r)) \\ &= 4p(2p-1)r. \end{aligned}$$

The inequality comes from the fact that for $0 \leq s \leq n = 4p - 1$, the maximum value of $s(n - s)$ occurs only if $s = 2p$ or $2p - 1$.

One consequence of inequality (11) is that k and l are positive, for if one is zero, then each off-diagonal element of Q is at least $2a + 2$ and then $e^T Q e \geq (2a + 2)(4p-1)(4p-2)$. Thus from inequality (11), $4(a+1)(4p-1)(2p-1) \leq 4p(2p-1)r = 4p(2p-1)(4a+a_1)$. In this case, $p(4-a_1) \leq a+1$, which is impossible since $0 \leq a_1 \leq 3$, $a + 1 \leq p$, and $4a + a_1 = r \leq 4p - 2$.

Now let

$$u = (\overbrace{l, \dots, l}^{k}, \overbrace{-k, \dots, -k}^{l}) \in \mathbb{Z}^{4p-1}.$$

Then $u \in \langle e \rangle^{\perp}$ and

$$u^T E u \geq -kl(e^T E e) = -kl(e^T Q e - e^T Q_0 e) \geq -kl(4p(2p-1)r - e^T Q_0 e).$$

Thus

$$\begin{aligned} u^T Q u &= u^T Q_0 u + u^T E u \\ &\geq u^T Q_0 u - kl(4p(2p-1)r - e^T Q_0 e) \\ &= l^2 k(k-1)(2a+2) + k^2 l(l-1)(2a+2) - 2kl(kl)(2a+1) \\ &+ (kl)k(k-1)(2a+2) + (kl)l(l-1)(2a+2) + 2kl(kl)(2a+1) \\ &- kl(4p(2p-1)r \\ &= 2kl\left((a+1)(4p-1)(4p-3) - 2p(2p-1)r\right) \end{aligned}$$

and

$$u^T u = kl^2 + lk^2 = kl(4p-1).$$

We finish the proof by showing that $u^T Q u \geq -2\alpha(r) u^T u$.

First suppose $r = 4a + a_1$, where $a_1 = 0, 1$, or 2. Then $\alpha(r) = a$, and $a \leq p-1$. It follows that

$$\begin{aligned} u^T Q u + 2a u^T u &\geq 2kl\left((a+1)(4p-1)(4p-3) - 2p(2p-1)(4a+2) + a(4p-1)\right) \\ &= 2kl(8p^2 - 12p + 3 - 2a(2p-1)) \\ &\geq 2kl(8p^2 - 12p + 3 - 2(p-1)(2p-1)) \\ &= 2kl(4p^2 - 6p + 1) \\ &\geq 0, \end{aligned}$$

since $p \geq 2$.

Next suppose $r = 4a + 3$, where $a \leq p-3$. Then $\alpha(r) = a$ and

$$\begin{aligned} u^T Q u + 2a u^T u &\geq 2kl\left((a+1)(4p-1)(4p-3) - 2p(2p-1)(4a+3) + a(4p-1)\right) \\ &= 2kl(4p^2 - 10p + 3 - 2a(2p-1) \\ &\geq 2kl(4p^2 - 10p + 3 - 2(p-3)(2p-1)) \\ &= 2kl(4p-3) \\ &> 0. \end{aligned}$$

The remaining case is $r = 4p-5 = 4a+3$, where $a = p-2$ and $\alpha(r) = p-2+\frac{1}{4p-1}$. In this case $2\alpha(r) u^T u = 2kl(4p^2 - 9p + 3)$. Thus

$$\begin{aligned} u^T Q u + 2\alpha(r) u^T u &\geq 2kl\left((p-1)(4p-1)(4p-3) - 2p(2p-1)(4p-5) + 4p^2 - 9p + 3\right) \\ &= 0. \qquad \square \end{aligned}$$

Inequalities (6) and (7) follow readily from Lemma 2.

Proof of inequalities (6) *and* (7) *of Theorem* 1. Let $n = 4p-1$ with $p \geq 2$, let X be a matrix in $M_{nt+r,n}(0,1)$, and let $R = X^T X - ptI_n$. By Lemma 2, $\lambda_{\min}(R) \leq \alpha(r)$. Thus $\lambda_{\min}(X^T X) \leq pt + \alpha(r)$. Inequalities (6) and (7) hold since $p = \frac{n+1}{4}$ and $\alpha(r) = \left\lfloor \frac{r}{4} \right\rfloor$ for $r \neq n-4$ and $\alpha(r) = \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n}$ for $r \neq n-4$. $\square$

3. Existence of design matrices for $n \geq 7$. In this section we show that the upper bound on $E(m,n)$ in inequality (6) is also a lower bound and we establish the lower bound on $E(m,n)$ in inequality (8). Indeed since $p = \frac{n+1}{4}$ and $\alpha(r) = \left\lfloor \frac{r}{4} \right\rfloor$ for $r \neq n-4$ and $\alpha(r) = \left\lfloor \frac{r}{4} \right\rfloor + \frac{1}{n}$ for $r \neq n-4$, the second part of Theorem 1 follows from the following theorem.

THEOREM 3. *Let $n = 4p-1$ be a positive integer with $p \geq 2$. There exists a positive integer m_0 such that for each $m = nt + r \geq m_0$ there exists a matrix $X_m \in M_{m,n}(0,1)$ such that*

$$\lambda_{\min}(X_m^T X_m) = pt + \alpha(r) \tag{12}$$

if $r \neq n-4$ and

$$\lambda_{\min}(X_m^T X_m) \geq pt + \alpha(r) - \frac{4}{tn^3} \tag{13}$$

if $r = n-4$.

The Gram matrix, $X_m^T X_m$ of the matrix in Theorem 3 can be described as follows: For $0 \leq r < n$, let $a = \left\lfloor \frac{r}{4} \right\rfloor$, so that $r = 4a + a_1$, where $a_1 = 0, 1, 2,$ or 3. Define a diagonal matrix

$$D_r = \operatorname{diag}(\overbrace{a+1, \ldots, a+1}^{k}, a, \ldots, a), \tag{14}$$

where $k = \left\lfloor \frac{r}{4} \right\rfloor + p\left(r - 4\left\lfloor \frac{r}{4} \right\rfloor\right) = a + pa_1$. Notice that $4k = 4a + 4pa_1 = 4a + (n+1)a_1 = na_1 + r < 4n$ so that $k < n$. Also trace $D_r = na + k = pr$. We will show that for sufficiently large $m = nt + r$, there exists an $m \times n$ design matrix X_m such that $X_m^T X_m = tpI_n + bJ_n + D_r$ (for some b) and that $\lambda_{\min}(X_m^T X_m) = pt + \alpha(r)$ for $r \neq n-4$ and $\lambda_{\min}(X_m^T X_m) \geq pt + \alpha(r) - 4/tn^3$ if $r = n-4$.

The proof of Theorem 3 is not constructive. It involves the $\mathbb{Z}$-module generated by all (0,1)-matrices $v^T v$ where v is a (0,1)-vector in $\mathbb{Z}^n$ with exactly $2p$ or $2p-1$ ones. Let

$$\mathcal{M}(n) = \mathbb{Z}\text{-span } \{v^T v : v \in \{0,1\}^n, v^T e = 2p \text{ or } v^T e = 2p-1\}. \tag{15}$$

We begin with a lemma to show that $pI_n, J_n, D_r \in \mathcal{M}(n)$ for each $0 \leq r < n$.

LEMMA 4. *Let $n = 4p-1$ be a positive integer with $p \geq 2$ and $0 \leq r < n$. Then $pI_n, J_n, D_r \in \mathcal{M}(n)$.*

Proof. First we will show that $D_1 \in \mathcal{M}(n)$. It is easy to see that

$$D_1 = \operatorname{diag}(\overbrace{1, \ldots, 1}^{p}, 0, \ldots, 0).$$

Now define $(0,1)$-vectors $v_1, v_2, w_1, w_2, w_3, w_4$ in $\mathbb{Z}^n$:

$$\begin{aligned}
v_1 &= (1, 1, \overbrace{0, \ldots, 0}^{p-2}, 0, 0, 0, \overbrace{0, \ldots, 0}^{p-1}, \overbrace{1, \ldots, 1}^{2p-2}),\\
v_2 &= (0, 0, 0, \ldots, 0, 1, 1, 0, \ldots, 0, 1, \ldots, 1),\\
w_1 &= (0, 0, 0, \ldots, 0, 0, 1, 0, \ldots, 0, 1, \ldots, 1),\\
w_2 &= (0, 0, 0, \ldots, 0, 1, 0, 0, \ldots, 0, 1, \ldots, 1),\\
w_3 &= (0, 1, 0, \ldots, 0, 0, 0, 0, \ldots, 0, 1, \ldots, 1),\\
w_4 &= (1, 0, 0, \ldots, 0, 0, 0, 0, \ldots, 0, 1, \ldots, 1).
\end{aligned}$$

Then v_1, v_2 have $2p$ ones and $2p-1$ zeros, and w_1, w_2, w_3, w_4 have $2p-1$ ones and $2p$ zeros. Thus $v_i^T v_i, w_j^T w_j \in \mathcal{M}(n)$ for $i = 1, 2$ and $j = 1, 2, 3, 4$. A direct calculation gives

$$v_1^T v_1 - v_2^T v_2 + w_1^T w_1 + w_2^T w_2 - w_3^T w_3 - w_4^T w_4 = E_{1,2} + E_{2,1} - E_{p+1,p+2} - E_{p+2,p+1}.$$

(Here E_{ij} is the matrix whose only nonzero entry is a one in position (i, j).) Thus $E_{1,2} + E_{2,1} - E_{p+1,p+2} - E_{p+2,p+1}$, and similarly, $E_{ij} + E_{ji} - E_{p+i,p+j} - E_{p+j,p+i}$ are in $\mathcal{M}(n)$ for all distinct $1 \le i, j \le p$. Summing on all such i, j we get

$$M_1 = \begin{bmatrix} J_p - I_p & 0_p \\ 0_p & I_p - J_p \end{bmatrix} \oplus 0_{2p-1} \in \mathcal{M}(n).$$

Next let $A = [J_p, J_p - I_p, 0_{p,2p-1}]$. Each row of A has $2p-1$ ones and $2p$ zeros. Thus

$$A^T A = \begin{bmatrix} pJ_p & (p-1)J_p \\ (p-1)J_p & (p-2)J_p + I_p \end{bmatrix} \oplus 0_{2p-1} \in \mathcal{M}(n).$$

Finally, let $w = (\overbrace{1, \ldots, 1}^{2p}, \overbrace{0, \ldots, 0}^{2p-1})$. Then

$$w^T w = \begin{bmatrix} J_p & J_p \\ J_p & J_p \end{bmatrix} \oplus 0_{2p-1} \in \mathcal{M}(n).$$

Therefore,

$$D_1 = \begin{bmatrix} I_p & 0_p \\ 0_p & 0_p \end{bmatrix} \oplus 0_{2p-1} = A^T A - (p-1) w^T w - M_1 \in \mathcal{M}(n).$$

We now show that $D_r \in \mathcal{M}(n)$ for $r = 2, \ldots, n-1$. Since $\mathcal{M}(n)$ is invariant under permutation similarity and $D_1 \in \mathcal{M}(n)$, each diagonal matrix with p ones and $n-p$ zeros on the diagonal is also in $\mathcal{M}(n)$. But D_r is a diagonal matrix with trace pr and diagonal entries equal to a or $a+1$. Thus D_r is a sum of r diagonal matrices with p ones and $n-p$ zeros on the diagonal. It follows that $D_r \in \mathcal{M}(n)$.

It is clear that pI_n is a sum of n diagonal matrices each having p ones and $n-p$ zeros. Thus $pI_n \in \mathcal{M}(n)$.

Finally we show that $J_n \in \mathcal{M}(n)$. Let

$$u_1 = (1, \overbrace{1, \ldots, 1}^{2p-1}, 0, \ldots, 0) = \sum_{j=1}^{2p} e_j,$$
$$v_1 = (0, \overbrace{1, \ldots, 1}^{2p-1}, 0, \ldots, 0) = \sum_{j=2}^{2p} e_j.$$

Then

$$u_1^T u_1 - v_1^T v_1 = E_{1,1} + \sum_{j=2}^{2p} (E_{1,j} + E_{j,1}) \in \mathcal{M}(n).$$

Likewise for each $1 \le i \le n$ let

$$u_i = \sum_{j=i}^{i+2p-1} e_j,$$
$$v_i = \sum_{j=i+1}^{i+2p-1} e_j,$$

where the index j is taken modulo n. Then

$$u_i^T u_i - v_i^T v_i = E_{i,i} + \sum_{j=i+1}^{i+2p-1} (E_{i,j} + E_{j,i}) \in \mathcal{M}(n).$$

It is easy to see that $J_n = \sum_{i=1}^{n} (u_i^T u_i - v_i^T v_i)$ and thus $J_n \in \mathcal{M}(n)$. □

At this point, the arguments for the cases where n is a Hadamard number and the cases (if any) where n is not need to be separated. A *Hadamard number* is an integer $n = 4p - 1$ for which a there exists a $(4p-1, 2p, p)$-design, or equivalently an $n \times n$ design matrix H such that $H^T H = p(I_n + J_n)$. It is conjectured that every integer $n \equiv -1 \pmod 4$ is a Hadamard number, and $n = 427$ is the smallest such integer for which no Hadamard design is known.

In case n is a Hadamard number, we let H be a $(4p-1, 2p, p)$-design matrix so that $H^T H = p(I_n + J_n)$. Otherwise we need to define two larger design matrices of sizes $cn \times n$ and $2cn \times n$, where $c = \frac{1}{2p}\binom{4p-2}{2p-1}$ is a Catalan number and hence an integer.

Let C be the $nc \times n$ design matrix whose rows consist of all n-tuples with exactly $2p$ ones and $2p-1$ zeros. Let B be the $2nc \times n$ design matrix whose rows consist of all n-tuples with exactly $2p$ ones and $2p-1$ zeros and those with $2p-1$ ones and $2p$ zeros. It is easy to see that

$$C^T C = c(pI_n + pJ_n),$$
$$B^T B = c(2pI_n + (2p-1)J_n). \tag{16}$$

These matrices will be used in the next two lemmas.

LEMMA 5. *Let $n = 4p - 1$ be a positive integer with $p \ge 2$. If n is a Hadamard number and $0 \le r < n$, then there exists a positive integer τ_r and an $m \times n$ design-matrix X_r such that $m = 2\tau_r n + r$, each row of X_r has exactly $2p$ or $2p-1$ ones, and*

$$X_r^T X_r = \tau_r(2pI_n + (2p-1)J_n) + D_r, \tag{17}$$

where D_r is the $n \times n$ diagonal matrix defined in (14).

If n is not (necessarily) a Hadamard number and $0 \le R < nc$, then there exists positive integer τ_R and an $m \times n$ design matrix X_R such that $m = 2\tau_R cn + R$ and

$$X_R^T X_R = (c\tau_R + s)(2pI_n + (2p-1)J_n) + s_1(pI_n + pJ_n) + D_r, \tag{18}$$

where $R = (2s + s_1)n + r$ with $s \ge 0$, $s_1 = 0, 1$, $2s + s_1 < c$, and $0 \le r < n$.

Proof. Let $u_1, \ldots, u_N$ be the set of all (0,1)-vectors in $\mathbb{Z}^n$ with exactly $2p$ ones and $2p-1$ zeros or $2p-1$ ones and $2p$ zeros, that is, the $2cn$ rows of the matrix B. Then

$$\sum_{j=1}^{N} u_j^T u_j = B^T B = c(2pI_n + (2p-1)J_n). \tag{19}$$

To prove the first part of the lemma, assume n is a Hadamard number and that $0 \le r < n$. Since $D_r \in \mathcal{M}(n)$, there exist integers z_i (some of which may be negative) such that $D_r = \sum z_i u_i^T u_i$. Now let z be a nonnegative integer such that $z + z_i \ge 0$ for all i, and let X_r be the $m \times n$ design matrix whose rows consist of each of the u_i repeated $z + z_i$ times. Then

$$X_r^T X_r = \sum (z + z_i) u_i^T u_i = zc(2pI_n + (2p-1)J_n) + D_r.$$

Thus (17) holds for $\tau_r = zc$.

It remains to show that $m = 2\tau_r n + r$. Since each u_i has either $2p$ or $2p-1$ ones, we have

$$n \operatorname{trace} u_i^T u_i - e^T u_i^T u_i e = 2p(2p-1)$$

for all i. Thus,

$$\begin{aligned} 2p(2p-1)m &= 2p(2p-1)\sum (z + z_i) \\ &= n \operatorname{trace} X_r^T X_r - e^T X_r^T X_r e \\ &= (n-1)\operatorname{trace}(\tau_r(2pI_n) + D_r)) \\ &= (n-1)(2np\tau_r + pr) \\ &= 2p(2p-1)(2\tau_r n + r). \end{aligned}$$

(The third equation follows from the fact that $n \operatorname{trace} J_n - e^T J_n e = 0$.) Hence, $m = 2\tau_r n + r$. This completes the proof of the first part of the lemma.

Now we consider that case where n is not necessarily a Hadamard number. Let R be an integer satisfying $0 \le R < nc$, where $R = (2s + s_1)n + r$ with $s \ge 0$, $s_1 = 0, 1$, $2s + s_1 < c$, and $0 \le r < n$. Since $D_r, pI_n, J_n \in \mathcal{M}(n)$, there exist integers z_i such that

$$\sum z_i u_i^T u_i = s(2pI_n + (2p-1)J_n) + s_1(pI_n + pJ_n) + D_r.$$

As before, let z be a nonnegative integer such that $z + z_i \ge 0$ for all i and let X_R be the $m \times n$ design matrix whose rows are the u_i repeated $z + z_i$ times. Then

$$\begin{aligned} X_R^T X_R &= \sum (z + z_i) u_i^T u_i \\ &= (zc + s)(2pI_n + (2p-1)J_n) + s_1(pI_n + pJ_n) + D_r. \end{aligned}$$

Thus (18) holds for $\tau_R = z$.

It remains to show that $m = 2\tau_R cn + R$. Arguing as before,

$$\begin{aligned} 2p(2p-1)m &= 2p(2p-1)\sum (z + z_i) \\ &= n \operatorname{trace} X_R^T X_R - e^T X_R^T X_R e \\ &= (n-1)((\tau_R c + s)(2pn) + s_1 pn + pr) \\ &= 2p(2p-1)(2\tau_R cn + R). \qquad \square \end{aligned}$$

This completes the proof of the second part of the lemma.

In the next lemma we show that for sufficiently large $m = nt + r$, a matrix $X_m \in M_{m,n}(0,1)$ exists such that

$$X_m^T X_m = tpI_n + bJ_n + D_r,$$

where $t(p-\frac{1}{2}) \le b \le tp$.

LEMMA 6. *Let $n = 4p-1$ be a positive integer with $p \ge 2$ and let $0 \le r < n$. There exists a positive integer m_0 such that for each $m = nt + r \ge m_0$ there exists a matrix $X_m \in M_{m,n}(0,1)$ such that*

$$X_m^T X_m = tpI_n + bJ_n + D_r, \tag{20}$$

where $t(p-\frac{1}{2}) \le b \le tp$.

Proof. Let r be an integer satisfying $0 \le r < n$.

First we assume that n is a Hadamard number so that there exists an $n \times n$ Hadamard $(4p-1, 2p, p)$-design matrix H. By Lemma 5, there exists a positive integer τ_r and a $(2\tau_r n + r) \times n$ (0,1)-matrix X such that (17) holds.

Suppose $m \ge 2\tau_r n + r$ with $m = tn + r$. Then $t = 2\tau_r + k$ for some nonnegative integer k. Let H be an $n \times n$ Hadamard matrix. Then $H^T H = p(I_n + J_n)$. Let

$$X_m^T = [X^T, \overbrace{H^T, \ldots, H^T}^{k}]$$

be the $m \times n$ (0,1)-matrix obtained by adjoining k copies of H to X. Then

$$\begin{aligned} X_m^T X_m &= X^T X + kH^T H \\ &= \tau_r(2pI_n + (2p-1)J_n) + D_r + kp(I_n + J_n) \\ &= tpI_n + bJ_n + D_r, \end{aligned}$$

where $b = tp - \tau_r \le tp$. Since $t = 2\tau_r + k$, $\tau_r \le t/2$. Thus $t(p-\frac{1}{2}) \le b$. Now choose m_0 large enough that $m_0 \ge 2\tau_r n + r$ for $r = 0, \ldots, n-1$.

The proof without assuming that n is a Hadamard number is more complicated. With $m = nt + r$, we write $t = c\tau + 2s + s_1$, where c is the Catalan number from Lemma 5, $\tau = \lfloor \frac{t}{c} \rfloor$, $0 \le 2s + s_1 < c$, and $s_1 = 0$ or 1. Thus $m = nc\tau + R$, where $0 \le R = (2s + s_1)n + r < nc$. From Lemma 5, there exists a positive integer τ_R and a $(2nc\tau_R + R) \times n$ matrix X_R such that

$$X_R^T X_R = (c\tau_R + s)(2pI_n + (2p-1)J_n) + s_1(pI_n + pJ_n) + D_r. \tag{21}$$

Suppose $\tau = 2\tau_R + k$ for some nonnegative integer k. Let

$$X_m^T = [X_R^T, \overbrace{C^T, \ldots, C^T}^{k}].$$

Then from (16) and (21) we get

$$\begin{aligned} X_m^T X_m &= X_R^T X_R + kC^T C \\ &= (c\tau_R + s)(2pI_n + (2p-1)J_n) + s_1(pI_n + pJ_n) + kc(pI_n + pJ_n) + D_r \\ &= (\tau c + 2s + s_1)pI_n + bJ_n + D_r, \end{aligned}$$

where

$$\begin{aligned} b &= (2p-1)\tau_R c + p(2s + s_1) + kpc \\ &= pc(2\tau_R + k) + p(2s + s_1) - (\tau_R c + s) \\ &= tp - (\tau_R c + s) \\ &\le tp. \end{aligned}$$

Since $t = c(2\tau + k) + 2s + s_1$, we have $2(\tau_R c + s) \le t$. Thus, $t(p - \frac{1}{2}) \le b$. ∎

To prove Theorem 3, it remains only to show (12) and (13).

Proof of Theorem 3. Let $m \ge m_0$ with $m = nt + r$. By Lemma 6, there exists $X_m \in M_{m,n}(0,1)$ such that (20) holds. Let $a = \lfloor \frac{r}{4} \rfloor$ so that $r = 4a + a_1$ for $a_1 = 0, 1, 2,$ or 3. Let $k = a + pa_1 < n$. If $k = n - 1$, then $a_1 = 3, a = p - 2$, and $r = n - 4$; otherwise $k \le n - 2$. We distinguish two cases:

Case $r \ne n - 4$. In this case there are at least two diagonal entries of D_r equal to a. Thus it is clear that $\lambda_{\min}(X_m^T X_m) = pt + a$. (The vector $v = (0, \dots, 0, 1, -1)^T$ is an eigenvector of $X_m^T X_m$ corresponding to this eigenvalue.)

Case $r = n - 4$. In this case, $D_r = \mathrm{diag}(a + 1, \dots, a + 1, a)$ and $X_m^T X_m = ptI_n + bJ_n + D_r$. Since $b_0 := t(p - \frac{1}{2}) \le b$, we know that $\lambda_{\min}(X_m^T X_m) \ge (pt + a) + \lambda_{\min}(M)$, where $M = b_0 J_n + \mathrm{diag}(1, \dots, 1, 0)$. Let $e_1, \dots, e_n$ be the standard basis for $\mathbb{R}^n$. Then the $n - 2$ vectors $e_i - e_{i+1}$, $i = 1, \dots, n - 2$, are eigenvectors for M corresponding to the eigenvalue 1. Let $\mu_1 \le \mu_2$ be the other two eigenvalues of M. To finish the proof, we will show that $\lambda_{\min}(M) = \mu_1 > \frac{1}{n} - \frac{4}{tn^3}$.

We now find an explicit expression for μ_1. Clearly, $\mathrm{trace}(M) = nb_0 + 1$ and it is not hard to see that $\det(M) = b_0$. Thus

$$\begin{aligned} \mu_1 + \mu_2 &= nb_0 + 1, \\ \mu_1 \mu_2 &= b_0. \end{aligned}$$

Solving for μ_1, μ_2, we find that the smaller eigenvalue μ_1 is given by

$$\mu_1 = \frac{1}{2}\left(1 + b_0 n - \sqrt{(b_0 n + 1)^2 - 4b_0}\right).$$

It is easy to see that $\mu_1 < 1$, and thus $\mu_1 = \lambda_{\min}(M)$.

Finally, we show that

$$\frac{1}{n} - \frac{4}{tn^3} < \mu_1, \tag{22}$$

which is equivalent to each of the following inequalities:

$$\begin{aligned} 2(tn^2 - 4) &< tn^3\left(1 + b_0 n - \sqrt{(b_0 n + 1)^2 - 4b_0}\right), \\ tn^3\sqrt{(b_0 n + 1)^2 - 4b_0} &< tn^3(1 + b_0 n) - 2(tn^2 - 4), \\ t^2 n^6\left((b_0 n + 1)^2 - 4b_0\right) &< \left(tn^3(1 + b_0 n) - 2(tn^2 - 4)\right)^2, \end{aligned} \tag{23}$$

where $b_0 = t(p - \frac{1}{2}) = \frac{t(n-1)}{4}$. By a direct calculation, we see that 16 times the right side of inequality (23) is

$$\begin{aligned} &1024 - 512\,n^2\,t + 256\,n^3\,t \\ &+ 16\,n^6\,t^2 + 16\,n^6\,t^3 - 24\,n^7\,t^3 + 8\,n^8\,t^3 + n^8\,t^4 - 2\,n^9\,t^4 + n^{10}\,t^4, \end{aligned}$$

and 16 times the left side of inequality (23) is

$$16\,n^6\,t^2 + 16\,n^6\,t^3 - 24\,n^7\,t^3 + 8\,n^8\,t^3 + n^8\,t^4 - 2\,n^9\,t^4 + n^{10}\,t^4,$$

and their difference, $256(n^3 t - 2n^2 t + 4)$, is positive. Thus inequality (23), and hence inequality (22) and Theorem 3, is proved. ∎

4. Construction of E-optimal design matrices for $n = 7$. Although our methods do not show how to construct E-optimal design matrices for all Hadamard numbers, we can construct them for $n = 7$ and $r \neq 3$. (For $n = 7$, $p = 2$.) Let H_7 be any $(7, 4, 2)$ Hadamard design matrix. For example, we can take

$$H_7 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix},$$

so that $H_7^T H_7 = 2(I_7 + J_7)$. We now describe how to construct an E-optimal design matrix for $0 \leq r \leq 6$, except when $r = 3$. Let $m = 7t + r \geq 7$. Let v_1, v_2 denote the first two columns of H_7.

Case $r = 0, 1, 2$. In this case, Theorem 1 states that $E(7t + r, 7) = 2t$. First suppose $r = 0$ so that $m = 7t$, and let

$$X_{7t}^T = [\overbrace{H_7^T, \ldots, H_7^T}^{t}].$$

Then $X_m^T X_m = 2t(I_7 + J_7)$ so that $\lambda_{\min}(X_m^T X_m) = E(7t, 7) = 2t$.

Now let v_1, v_2 be any (0,1)-7-tuples and let $X_{7t+1}^T = [X_{7t}^T, v_1]$ and $X_{7t+2}^T = [X_{7t}^T, v_1, v_2]$, where X_{7t} is defined above. Clearly, $\lambda_{\min}(X_{7t+r}^T X_{7t+r}) \geq \lambda_{\min}(X_{7t}^T X_{7t}) = 2t$ for $r = 1, 2$, but Lemma 2 guarantees that $\lambda_{\min}(X_{7t+r}^T X_{7t+r}) \leq 2t$, and hence $\lambda_{\min}(X_{7t+r}^T X_{7t+r}) = 2t$.

Case $r = 3$. Let

$$X_{10} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Then

$$X_{10}^T X_{10} = \begin{bmatrix} 5 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 5 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 5 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 5 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 5 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 5 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 4 \end{bmatrix} = 2(I_7 + J_7) + D_6.$$

For $t \geq 2$, let $X_m^T = [\overbrace{H_7, \ldots, H_7}^{t-1}, X_{10}]$. Then $X_m^T X_m = 2(t-1)(I_7 + J_7) + X_{10}^T X_{10} = 2t(I_7 + J_7) + D_6$. The argument for the case $r = n - 4 = 3$ in the proof of Theorem 3 shows that $\lambda_{\min}(X_m^T X_m) \geq 2t + \frac{1}{7} - \frac{4}{343t}$.

Case $r = 4, 5, 6$. Let

$$X_{11} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Then

$$X_{11}^T X_{11} = \begin{bmatrix} 5 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 5 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 5 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 5 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 5 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 5 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 6 \end{bmatrix},$$

Then $\lambda_{\min}(X_{11}^T X_{11}) = pt + \left\lfloor \frac{r}{4} \right\rfloor = E(11, 7) = 3$. For $t \geq 2$, let

$$X_{7t+4}^T = [\overbrace{H_7^T, \ldots, H_7^T}^{t-1}, X_{11}].$$

Then $X_{7t+4}^T X_{7t+4} = 2(t-1)(I_7 + J_7) + X_{11}^T X_{11}$ and it is clear that $\lambda_{\min}(X_{7t+4}^T X_{7t+4}) = E(7t+4, 7) = 2t+1$.

For $r = 5, 6$, adjoin any (0,1)-7-tuples v_1 if $r = 5$ and v_1, v_2 if $r = 6$ to the design matrix described in the $r = 4$ case. The minimum eigenvalue of $X_{7t+r}^T X_{7t+r}$ does not increase. That is, $\lambda_{\min}(X_{7t+r}^T X_{7t+r}) = E(7t+r, 7) = 2t+1$.

5. E-optimality for $n = 3$. Let $n = 3$,

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

and A_i denote the ith row of A. Assume $X \in M_{m,3}(0, 1)$ has k_i rows equal to A_i.

Assume $m = 3t + r$. Then $k_1 + \cdots + k_7 = 3t + r$ and

$$X^T X = \begin{bmatrix} k_1 + k_5 + k_6 + k_7 & k_6 + k_7 & k_5 + k_7 \\ k_6 + k_7 & k_2 + k_4 + k_6 + k_7 & k_4 + k_7 \\ k_5 + k_7 & k_4 + k_7 & k_3 + k_4 + k_5 + k_7 \end{bmatrix}.$$

We prove the upper bound by contradiction and for that purpose we assume that $\lambda_{\min}(X^T X) > t$. In that case $u^t X^T X u > t$ for all unit vectors u. Using the three vectors $(1/\sqrt{2})(1, -1, 0)$, $(1/\sqrt{2})(1, 0, -1)$, and $(1/\sqrt{2})(0, 1, -1)$ in place of u and using the fact that the k_i are integers we get the following inequalities on the k_i:

$$\begin{aligned} (k_1 + k_4) + (k_2 + k_5) &\geq 2t + 1, \\ (k_1 + k_4) + (k_3 + k_6) &\geq 2t + 1, \\ (k_2 + k_5) + (k_3 + k_6) &\geq 2t + 1. \end{aligned}$$

Adding the three inequalities yields

$$2(k_1 + k_2 + k_3 + k_4 + k_5 + k_6) \geq 6t + 3$$

and hence

$$3t + r \geq k_1 + k_2 + k_3 + k_4 + k_5 + k_6 \geq 3t + 2,$$

a contradiction if $r = 0$ or $r = 1$. Thus for $r = 0$ or $r = 1$ we have $\lambda_{\min}(X^T X) \leq t$. For $r = 2$ we have $k_1+k_2+k_3+k_4+k_5+k_6 = 3t+2$, i.e., $k_7 = 0$. Furthermore, we may assume without loss of generality that $k_1 + k_4 = t+1 = k_2 + k_5$ and $k_3 + k_6 = t$. Now let $v = 1/\sqrt{6}(-1, -1, 2)$. Then $v^T X^T X v = t + 1/3$. Hence $\lambda_{\min}(X^T X) \leq t + 1/3$.

Next we construct a design matrix X with $\lambda_{\min}(X^T X) = t$ if $r = 0, 1$, and $\lambda_{\min}(X^T X) > t + \frac{1}{3} - \frac{2}{27t+3}$ if $r = 2$.

When $r = 0$ let $k_1 = k_2 = k_3 = 0$ and $k_4 = k_5 = k_6 = t$. Then it is easy to see that $X^T X = t(I_3 + J_3)$ and hence $\lambda_{\min}(X^T X) = t$.

When $r = 1$ we can produce many different matrices X with $\lambda_{\min}(X^T X) = t$ by adjoining any (0,1)-vector to the matrix given above for $r = 0$. In all cases, $\lambda_{\min}(X^T X) = t$. One such example is given by $k_1 = 1, k_2 = k_3 = 0$, and $k_4 = k_5 = k_6 = t$. It is easy to see that for this choice of $k_1, \ldots, k_6$ we have $X^T X = t(I_3 + J_3) + \mathrm{diag}(1, 0, 0)$ and hence $\lambda_{\min}(X^T X) = t$.

When $r = 2$ define $k_1 = k_2 = 1, k_3 = 0$, and $k_4 = k_5 = k_6 = t$. It is easy to see that for this choice of $k_1, \ldots k_6$ we have $X^T X = t(I + J) + Diagonal(1, 1, 0)$ and hence $\lambda_{\min}(X^T X) = 1/2(5t + 1 - \sqrt{9t^2 + 2t + 1}) > t + 1/3 - \frac{2}{27t+3}$.

We summarize the results for $n = 3$ in the following theorem.

THEOREM 7. *Let $n = 3$ and $m = 3t + r$, with $0 \leq r \leq 2$. If $r \neq 2$, then*

$$E(3t + r, 3) = t.$$

If $r = 2$, then

$$t + \frac{1}{3} - \frac{2}{27t + 3} \leq E(3t + 2, 3) \leq t + \frac{1}{3}.$$

This is not the best possible lower bound for $r = 2$. In fact, the matrix defined by $k_1 = 1, k_2 = k_3 = 0$, and $k_4 = k_5 = t, k_6 = t + 1$ has a larger minimal eigenvalue for the same t than the matrix defined above. However, this minimal eigenvalue is the root of an irreducible cubic polynomial which leads to cumbersome expressions that are not easily manipulated.

6. E-optimality versus D-optimality. A matrix $X \in M_{m,n}(0, 1)$ is *D-optimal* if $\det X^T X$ is maximal among all matrices in $M_{m,n}(0, 1)$. A referee suggested that we compare D-optimality with E-optimality. Actually, the results and techniques for these two kinds of optimality are quite different.

We consider the case $n = 4p - 1$ from the present paper and assume that n is a Hadamard number, that is, an integer for which there exists a $(4p - 1, 2p, p)$ design. As in section 3, let H be the corresponding design matrix so that $H^T H = p(I_n + J_n)$ and let X be the $tn \times n$ matrix defined by

$$X^T = [\overbrace{H^T, \ldots, H^T}^{t}].$$

Then $X^TX = pt(I_n + J_n)$. Thus $\lambda_{\min}(X^TX) = pt$ and so, by Theorem 1, X is E-optimal. It turns out that X is also D-optimal. See [NWZ, Thm. 3.1]. However, the complementary design matrix $Y = J_{tn,n} - X$ is E-optimal but not D-optimal. This follows from the general result [NWZ] that each row of an $m \times n$ D-optimal design matrix must contain $2p$ ones and $2p-1$ zeros if m is sufficiently large [NWZ]. Each row of X contains $2p$ ones, but each row of Y contains only $2p-1$ ones. Thus Y cannot be D-optimal if t is sufficiently large. (In fact, Y is not D-optimal for any t.)

Other comparisons between D- and E-optimality are difficult because neither theory has been completely worked out. Indeed, the only two integers $n \equiv -1 \pmod 4$ for which results on D-optimality are sufficiently developed are $n = 3, 7$. Thus we now compare D- and E-optimality for $n = 7$.

To produce a design matrix that is D-optimal but not E-optimal, consider the case $n = 7$ and $m = 7t + 4$. In [NW], a $(7t + 4) \times 7$ D-optimal matrix X is shown to exist (for sufficiently large t) such that $X^TX = 2t(I_7 + J_7) + R$, where $R = B^TB$ and

$$B = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

is part of a Hadamard matrix. It is easy to see that $\lambda_{\min}(X^TX) = 2t$. But by Theorem 1, $E(7t + 4, 7) = 2t + 1$. Thus X is D-optimal but not E-optimal. In fact, it is implicit in [NW] that $\lambda_{\min}(X^TX) = 2t$ for all $(7t + 4) \times 7$ D-optimal matrices X. Thus for sufficiently large t, *no* $(7t + 4) \times 7$ D-optimal matrix is E-optimal.

By contrast, if $m = 7t+2$ is sufficiently large, all $(7t+2)\times 7$ D-optimal matrices X have minimum eigenvalue equal to $2t$ [NW]. So for sufficiently large t, *every* $(7t+2)\times 7$ D-optimal is E-optimal.

REFERENCES

[Che] C.-S. Cheng, *An application of the Kiefer-Wolfowitz equivalence theorem to a problem in Hadamard transformation optics*, Ann. Statist., 15 (1987), pp. 1593–1603.

[Kie] J. Kiefer, *General equivalence theory for optimum designs (approximation theory)*, Ann. Statist., 2 (1974), pp. 849–879.

[JN] M. Jacroux and W. Notz, *On the optimality of spring balance weighing designs*, Ann. Statist., 11 (1983), pp. 970–978.

[NW] M.G. Neubauer and W. Watkins, *D-optimal designs for seven objects and a large number of weighings*, Linear and Multilinear Algebra, 50 (2002), pp. 61–74.

[NWZ] M.G. Neubauer, W. Watkins, and J. Zeitlin, *Notes on D-optimal designs*, Linear Algebra Appl., 280 (1998), pp. 109–127.

[Puk] F. Pukelsheim, *Optimal Design of Experiments*, Wiley, New York, 1993.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 106–120

THE FLOW OF A DAE NEAR A SINGULAR EQUILIBRIUM*

R. E. BEARDMORE† AND R. LAISTER‡

Abstract. We extend the *differential-algebraic equation (DAE) taxonomy* by assuming that the linearization of a DAE about a singular equilibrium has a particular index-2 Kronecker normal form. A Lyapunov–Schmidt procedure is used to reduce the DAE to a quasilinear normal form which is shown to posses quasi-invariant manifolds which intersect the singularity. In turn, this provides solutions of the DAE which pass through the singularity.

Key words. Lipschitz solutions, nonhyperbolic equilibrium, quasi-invariant manifolds

AMS subject classifications. 34A34, 34A09, 37G05

PII. S0895479800378660

1. Preliminaries.

We consider the differential-algebraic equation (DAE)

$$\dot{x} = f(x, y), \tag{1.1}$$

$$g(x, y) = 0, \tag{1.2}$$

where $x \in \mathbb{R}^n$ $(n \geq 2)$, $y \in \mathbb{R}^m$, and $f : \mathcal{U} \to \mathbb{R}^n$ and $g : \mathcal{U} \to \mathbb{R}^m$ are both C^ω (analytic) in an open neighborhood, $\mathcal{U}$, of $(0, 0)$ in $\mathbb{R}^{n+m}$. The motivation for this paper is to understand the orbit structure of (1.1)–(1.2) near $(0, 0)$, which is assumed to be a *singular equilibrium* in the sense that

A1. $f(0, 0) = 0,\ g(0, 0) = 0,$
A2. $\mathrm{N}(d_y g(0, 0)) = \langle k \rangle\,, k^T k = 1,$ where $\mathrm{N}(d_y g(0, 0)^T) = \langle u \rangle.$

We shall also make the following assumptions, which we introduce now in order to make the presentation as transparent as possible:

A3. $d_x g(0, 0) d_y f(0, 0) k \notin \mathrm{R}(d_y g(0, 0)),$
A4. $d(f \times g)(0, 0) \in GL(\mathbb{R}^{n+m})$, and
A5. $d^2_{yy} g(0, 0)[k, k] \notin \mathrm{R}(d_y g(0, 0)).$

There is one further condition to be imposed which will be introduced at the appropriate point in the paper. The regularity assumptions are imposed on f and g for brevity, and one could consider problems of finite smoothness in a similar manner.

First, let us define some terminology associated with (1.1)–(1.2). The *constraint manifold* for (1.1)–(1.2) is the set $\mathbf{C} = \{(x, y) \in \mathcal{U} : g(x, y) = 0\}$, and the *singularity* is $\mathbf{S} = \{(x, y) \in \mathbf{C} : \det(d_y g(x, y)) = 0\}$.

The main result of the paper is that one can use A1–A5 to reduce the DAE (1.1)–(1.2) to a *quasilinear normal form* of dimension n. This normal form is a differential equation which can be written as

$$\dot{\alpha} = L_0 \alpha + \mathcal{O}(2), \tag{1.3}$$

$$s(\alpha, \beta)\dot{\beta} = \beta + \mathcal{O}(2), \tag{1.4}$$

*Received by the editors September 20, 2000; accepted for publication (in revised form) by V. Mehrmann January 16, 2002; published electronically June 12, 2002.
http://www.siam.org/journals/simax/24-1/37866.html

†Department of Mathematics, Imperial College, South Kensington, University of London, London, England, SW7 2AZ (r.beardmore@ic.ac.uk).
‡School of Mathematical Sciences, University of the West of England, Frenchay Campus, Bristol, England, BS16 1QY (robert.laister@uwe.ac.uk).

where $(\alpha, \beta) \in \mathbb{R}^n, L_0 \in GL(\mathbb{R}^{n-1})$ is some mapping and $s(0,0) = 0$. We can then understand the nature of solutions of (1.3)–(1.4), and hence of the original DAE, by rescaling time and applying standard invariant manifold theory to the resulting ODE. The only proviso to be met in this process is that solutions of (1.3)–(1.4) will require a degree of differentiability that is not imposed by the formulation (1.1)–(1.2).

1.1. Background. A standard uniqueness theorem for differential equations implies that for any $(x_0, y_0) \in \mathbf{C}\backslash\mathbf{S}$ there exist $\alpha, \omega > 0$ and a unique C^ω solution of (1.1)–(1.2), $(-\alpha, \omega) \to \mathbb{R}^{n+m}$; $t \mapsto (x(t), y(t)) \in \mathbf{C}\backslash\mathbf{S}$, such that $(x(0), y(0)) = (x_0, y_0)$. The goal of this paper is therefore to try to understand the nature of solutions which encounter the singularity and to understand how uniqueness can break down.

The usual alternative for the global continuation of solutions of ODEs states that solutions either exist for all time or else become unbounded in finite time. There is a third alternative for solutions of DAEs: the solutions terminate at a singularity [8]. However, it is not true that all solutions which encounter the singularity must terminate there; some may be continued [11, 12]. Indeed, the *DAE taxonomy* described in these references gives conditions under which there are submanifolds of $\mathbf{S}$ where such a continuation is possible.

In [3], the authors discuss the possibility of using the *DAE taxonomy* to investigate a type of shock wave in a magneto-hydrodynamics equation which makes this paper also relevant to that study. In [7] März gives conditions to ensure that the semilinear DAE

$$\mathcal{A}\dot{z} + \mathcal{B}z = \varphi(z), \qquad \|\varphi(z)\| = O(\|z\|^2) \text{ as } z \to 0, \tag{1.5}$$

has a Lyapunov stable equilibrium. In particular, the author supposes that the *Kronecker index* of the matrix pencil $(\mathcal{A}, \mathcal{B})$ is two, and in due course we shall write (1.1)–(1.2) in this form.

1.2. Notation. The term *manifold* is taken as a synonym for *graph* and the tangent space of a manifold $\mathcal{M}$ at a point $z \in \mathcal{M}$ is written $T_z(\mathcal{M})$. If U is a linear space, then for each $u \in U$ we shall write the map $v \mapsto u^T v$ as u^T, and the span of u is written as $\langle u \rangle = \{\mu u : \mu \in \mathbb{R}\}$. Also, $\|u\|^2 := u^T u$ and a hash symbol (#) denotes set cardinality.

Let $(\mathcal{A}, \mathcal{B}) \in \mathcal{L}(\mathbb{R}^N) \times \mathcal{L}(\mathbb{R}^N)$ be a square matrix pencil. It is *regular* if there exists a $\lambda \in \mathbb{C}$ such that $\det(\lambda\mathcal{A} + \mathcal{B}) \neq 0$. The spectrum of $(\mathcal{A}, \mathcal{B})$ is $\sigma(\mathcal{A}, \mathcal{B}) := \{\lambda \in \mathbb{C} : \det(\lambda\mathcal{A} + \mathcal{B}) = 0\}$, and $(\mathcal{A}, \mathcal{B})$ is *hyperbolic* if $\sigma(\mathcal{A}, \mathcal{B})$ contains no purely imaginary elements. We write $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$, and $\mathbb{C}^-$ is defined similarly.

Let us stipulate the degree of smoothness of solutions of (1.1)–(1.2) as follows. If $I \subset \mathbb{R}$ is open, a solution of (1.1)–(1.2) is a map $t \mapsto (x(t), y(t)) \in C^1(I, \mathbb{R}^n) \times C^0(I, \mathbb{R}^m)$, such that (1.1)–(1.2) is satisfied for all $t \in I$. A set $K \subset \mathbf{C}$ is said to be *quasi-invariant* for (1.1)–(1.2) if for each $(x(0), y(0)) \in K$ there is *at least one* solution of (1.1)–(1.2), $(x, y) : I \to \mathbf{C}$, such that $(x(t), y(t)) \in K$ for all $t \in I$.

In order to discuss solutions of the quasilinear problem (1.3)–(1.4), we must impose some degree of differentiability. So, let $I \subset \mathbb{R}$ be a bounded interval and let us note at this stage that the setting for solutions of (1.3)–(1.4) will be the space of Lipschitz functions. Thus, let us denote the Sobolev space

$$W^{n,\infty}(I, \mathbb{R}) = \left\{\beta : I \to \mathbb{R} : \beta, \dot{\beta}, \ldots, \beta^{(n)} \in L^\infty(I)\right\},$$

endowed with the standard norm, $\|u\|_{W^{n,\infty}} = \sum_{j=0}^{n} \|u^{(j)}\|_{L^\infty}$, where a dot and superscript (j) represent the derivative in a weak sense. Due to the inequality

$$|u(x) - u(y)| \leq |x - y| \|u\|_{W^{1,\infty}} \qquad \forall x, y \in I,$$

we may consider elements of $W^{n,\infty}(I, \mathbb{R})$ as being those functions with a Lipschitz continuous nth derivative.

1.3. A Kronecker normal form. In [2] there is a Kronecker normal form (KNF) which will provide the basis for the construction of the quasilinear normal form (1.3)–(1.4). First, let us define the matrices

$$M := \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad L := \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathcal{L}(\mathbb{R}^{n+m}).$$

We then have the following result from [2] concerning the KNF of (M, L).

THEOREM 1.1. *Suppose that $n \geq 2$ and $\det L \neq 0$. If $\mathrm{N}(D) = \langle k \rangle$ for some nonzero $k \in \mathbb{R}^m$ such that $CBk \notin \mathrm{R}(D)$, then there are nonsingular transformations P and Q such that*

$$PMQ = \begin{bmatrix} I_u & 0 & 0 \\ 0 & 0 & 0 \\ 0 & C_0 & 0 \end{bmatrix} \quad \text{and} \quad PLQ = \begin{bmatrix} A_0^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & I_m \end{bmatrix},$$

where $C_0 : \mathbb{R} \to \mathbb{R}^m$ is a linear map such that $C_0(1) = k$. If we write $\mathrm{N}(D^T) = \langle u \rangle$ and $U = \langle C^T u \rangle^\perp$, then $A_0 \in GL(U)$ and $\sigma(M, L) = 1/\sigma(A_0)$, where both PMQ and PLQ are elements of $\mathcal{L}(U \oplus \mathbb{R} \oplus \mathbb{R}^m)$.

If one assumes A1–A6, it follows from Theorem 1.1 that the linear DAE obtained from linearizing (1.1)–(1.2) at the zero equilibrium has index 2.

1.4. An underlying vector field. By writing $z = (x, y) \in \mathbb{R}^{n+m}$ and setting

$$L = \begin{bmatrix} A & B \\ C & D \end{bmatrix} := \begin{bmatrix} d_x f(0,0) & d_y f(0,0) \\ d_x g(0,0) & d_y g(0,0) \end{bmatrix}, \tag{1.6}$$

we may write (1.1)–(1.2) as the semilinear problem

$$M\dot{z} - Lz = F(z), \tag{1.7}$$

where the C^ω mapping F is defined by $Lz + F(z) = (f \times g)(z)$ and $F(z)$ is $\mathcal{O}(2)$ at zero.

Now consider (1.2) along a solution of (1.1)–(1.2) which lies in $\mathbf{C}\backslash\mathbf{S}$. Differentiating this constraint with respect to time we find

$$\dot{y} = d_y g(x, y)^{-1} d_x g(x, y) f(x, y).$$

By defining the variable τ by

$$\frac{d\tau}{dt} = \frac{1}{\det\, d_y g(x(t), y(t))}, \qquad \tau(t_0) = \tau_0,$$

we can reduce (1.1)–(1.2) to a vector field in the new time-scale τ:

$$x' = f(x, y) \det(d_y g(x, y)), \tag{1.8}$$

$$y' = \mathrm{adj}(d_y g(x, y)) d_x g(x, y) f(x, y), \tag{1.9}$$

where a prime ($'$) denotes $\frac{d}{d\tau}$.

This procedure gives a smooth vector field for which $\mathbf{C}$ is an invariant manifold, and any invariant set of (1.8)–(1.9) in $\mathbf{C}$ is a quasi-invariant set for (1.1)–(1.2). Moreover, the orbits of (1.8)–(1.9) coincide geometrically with those of (1.1)–(1.2), and this allows us to infer the behavior of (1.1)–(1.2), even at the singularity. This approach is used in [11] as the basis for the *DAE taxonomy*.

This approach can be useful, as in the following result which shows that when orbits of (1.8)–(1.9) are transverse to $\mathbf{S}$ at some point, that singular point is an impasse point. First, let us define

$$\Delta(x, y) := \det(d_y g(x, y)).$$

PROPOSITION 1.2. *Suppose that $\tau \mapsto (x(\tau), y(\tau))$ is a solution of* (1.8)–(1.9) *with initial condition $(x(0), y(0)) = (x_0, y_0) \in \mathbf{S}$. If*

$$\frac{d}{d\tau}\Delta(x(\tau), y(\tau))\bigg|_{\tau=0} \neq 0,$$

then there is a $t_* \in \mathbb{R}$ such that (1.1)–(1.2) *has exactly two solutions, $(x(t), y(t))$, which are both defined on either $[t_*, t_* + T)$ or $(t_* - T, t_*]$ for some $T > 0$ and which satisfy $(x(t_*), y(t_*)) = (x_0, y_0)$. Moreover, $\|\dot{y}(t)\| \to \infty$ as $t \to t_*$.*

Proof. From Theorem 2.1 of [9], we have to show that there is some nonzero $k \in \mathbb{R}^m$ such that $\mathrm{N}(d_y g(x_0, y_0)) = \langle k \rangle$, $d_x g(x_0, y_0) f(x_0, y_0) \notin \mathrm{R}(d_y g(x_0, y_0))$ and $d^2_{yy} g(x_0, y_0)[k, k] \notin \mathrm{R}(d_y g(x_0, y_0))$.

Define $\delta(\tau) := \Delta(x(\tau), y(\tau))$, so that $\delta(0) = 0$. Differentiating we have

$$\begin{aligned}
\delta'(\tau) &= \frac{d}{d\tau}\Delta(x(\tau), y(\tau)) \\
&= d_x\Delta(x(\tau), y(\tau))x'(\tau) + d_y\Delta(x(\tau), y(\tau))y'(\tau) \\
&= -d_x\Delta \cdot \Delta \cdot f + d_y\Delta \cdot (\text{adj } d_y g) \cdot d_x g \cdot f.
\end{aligned}$$

Therefore $\delta'(0) = d_y\Delta\,(\text{adj } d_y g)\, d_x g\ f|_{(x_0, y_0)}$, which is nonzero by assumption. Since the dimension of $\mathrm{N}(d_y g(x_0, y_0))$ is greater than or equal to two if and only if the adjugate $\mathrm{adj}(d_y g(x_0, y_0))$ is the zero mapping, we have $\delta'(0) = 0$ if $\dim \mathrm{N}(d_y g(x_0, y_0)) \geq 2$. Therefore $\mathrm{N}(d_y g(x_0, y_0)) = \langle \kappa \rangle$ for a nonzero $\kappa \in \mathbb{R}^m$. Now apply Lemma 3 from [1] to deduce that $\mathrm{R}(\text{adj } d_y g(x_0, y_0)) = \langle \kappa \rangle$ and $\mathrm{N}(\text{adj } d_y g(x_0, y_0)) = \mathrm{R}(d_y g(x_0, y_0))$. Using Lemma 1 from [1] we have

$$d_y\Delta(x, y)[\cdot] = \det'(d_y g)\left[d^2_{yy} g(x, y)[\cdot]\right] = \mathrm{tr}\left((\text{adj } d_y g) d^2_{yy} g(x, y)[\cdot]\right) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}),$$

where $\det'$ is the derivative of the determinant. Hence

$$\delta'(0) = \mathrm{tr}\left((\text{adj } d_y g) d^2_{yy} g\left[(\text{adj } d_y g) d_x g\ f(x_0, y_0)\right]\right).$$

We now use the simple null-space of the derivative $d_y g$ to conclude that if

$$d_x g\ [f(x_0, y_0)] \in \mathrm{R}(d_y g(x_0, y_0)), \tag{1.10}$$

then $\delta'(0) = 0$, and (1.10) cannot be true. It follows that there is a *nonzero* l_0 such that $(\text{adj } d_y g) d_x g\ f(x_0, y_0) = l_0 \kappa$. Therefore $\delta'(0) = l_0\ \mathrm{tr}\left((\text{adj } d_y g) d^2_{yy} g(x_0, y_0)[\kappa]\right)$. Now define the linear mapping

$$T := (\text{adj } d_y g) d^2_{yy} g(x_0, y_0)[\kappa];$$

then $\mathrm{R}(T) \subset \langle \kappa \rangle$ and $Ty \equiv \kappa \ell^T y$ for some $\ell \in \mathbb{R}^m$. Hence $\sigma(T) = \{0, \ell^T \kappa\}$ so that $\ell^T \kappa = \mathrm{tr}(T)$. Using Lemma 3 from [1] again, if $d^2_{yy} g(x_0, y_0)[\kappa, \kappa] \notin \mathrm{R}(d_y g(x_0, y_0))$, then $T\kappa \neq 0$ from where $\delta'(0) \neq 0$, and the result follows. ∎

Generally, $y(t)$ has the form

$$y(t) = O(t - T_*)^{1/2}$$

as $t \to T_*$ at an impasse point. In the degenerate diffusion literature, solutions which have this form, where t represents a spatial variable, are said to be *sharp solutions* [10].

2. A quasilinear normal form. The principle tool in our approach to understanding the flow of (1.1)–(1.2) is given in this section and is based on the following idea. Rather than differentiating the constraint (1.2) to obtain a vector field, suppose instead that we eliminate (1.2) directly by applying the implicit function theorem. Clearly, one cannot solve the constraint for y as a function of x near $(0, 0)$, but since $dg(0, 0)$ has full rank then $\mathbf{C} = g^{-1}(0)$ is a manifold and the information contained in (1.1) will define trajectories on it. However, the way in which the implicit function theorem is used is crucial, and the location of the singularity must emerge from this process. If we choose the correct decomposition of the ambient space in order to apply this Lyapunov–Schmidt reduction, then we can limit the way in which the singularity appears in the reduced problem.

In fact, Theorem 1.1 gives a decomposition through which we can track the effect of the singularity on solutions, and this in turn will allow us to find solutions which are unaffected by the presence of the singularity.

First we prove a preliminary lemma.

LEMMA 2.1. *Suppose that* A1–A5 *hold; then* $\mathbf{C}$ *is a manifold of dimension* n*, and* $\mathbf{S}$ *is a codimension-*1 *submanifold of* $\mathbf{C}$*.*

Proof. Let $\mathrm{N}(D^T) = \langle u \rangle$ for some nonzero $u \in \mathbb{R}^m$ and note that $C^T u \neq 0$ by A4; recall from A2 that $\mathrm{N}(D) = \langle k \rangle$. Write $y = \alpha k + \kappa \in \langle k \rangle \oplus \langle k \rangle^\perp = \mathbb{R}^m$ and form the decomposition $\mathbb{R}^m = \langle u \rangle \oplus \langle u \rangle^\perp$. Let $P : \mathbb{R}^m \to \langle u \rangle$ and $I - P : \mathbb{R}^m \to \langle u \rangle^\perp$ be orthogonal projections, and write $x = \lambda C^T u + \xi \in \langle C^T u \rangle \oplus \langle C^T u \rangle^\perp$.

Then $g(x, y) = 0 \in \mathbb{R}^m$ if and only if $(I - P + P)g(x, y) = 0$, which suggests that we define the mapping $\Gamma : \mathbb{R} \times \langle k \rangle^\perp \times \mathbb{R} \times \langle C^T u \rangle^\perp \to \langle u \rangle^\perp \times \mathbb{R}$ by

$$\Gamma(\alpha, \kappa, \lambda, \xi) := \begin{bmatrix} (I - P) g(\lambda C^T u + \xi, \alpha k + \kappa) \\ u^T g(\lambda C^T u + \xi, \alpha k + \kappa) \end{bmatrix}.$$

Now

$$d_{\kappa,\lambda} \Gamma(0, 0, 0) = \left[\begin{array}{c|c} (I - P) D|_{\langle k \rangle^\perp} & (I - P) C C^T u \\ \hline u^T D|_{\langle k \rangle^\perp} & u^T C C^T u \end{array} \right] = \begin{bmatrix} (I - P) D|_{\langle k \rangle^\perp} & * \\ 0 & \| C^T u \|^2 \end{bmatrix},$$

where $(I - P) D|_{\langle k \rangle^\perp}$ is a bijection. Hence one can apply the implicit function theorem to solve $g(\lambda C^T u + \xi, \alpha k + \kappa) = 0$ for $\kappa = \kappa(\alpha, \xi)$ and $\lambda = \lambda(\alpha, \xi)$ in a neighborhood of the origin of $\mathbb{R}^{n+m}$.

To locate $\mathbf{S}$ we must solve $g(x, y) = 0,\ \det\left(d_y g(x, y)\right) = 0$, and these are satisfied in some neighborhood of the origin if and only if

$$\hat{g}(\alpha, \xi) := \det\left(d_y g(\lambda(\alpha, \xi) C^T u + \xi, \alpha k + \kappa(\alpha, \xi)) \right) = 0. \tag{2.1}$$

Now $\hat{g}$ is C^ω and $\hat{g}(0,0) = 0$, so that by Lemma 1 of [1], using the fact that $d_\alpha \lambda(0,0) = 0$ and $d_\alpha \kappa(0,0) = 0$, we have $d_\alpha \hat{g}(0,0) = \mathrm{tr}((\mathrm{adj}D) d^2_{yy} g(0,0)[k])$. Using Lemma 3 of [1] we have $\mathrm{R}(\mathrm{adj}D) = \langle k \rangle$, so that $\mathrm{tr}((\mathrm{adj}D) d^2_{yy} g(0,0)[k])$ coincides with the only nonzero element of $\sigma((\mathrm{adj}D) d^2_{yy} g(0,0)[k])$. But $d^2_{yy} g(0,0)[k,k] \notin \mathrm{R}(D) = \mathrm{N}(\mathrm{adj}D)$ and therefore $(\mathrm{adj}D) d^2_{yy} g(0,0)[k,k] = \eta k$ for some $\eta \neq 0$. Because $d_\alpha \hat{g}(0,0) = \eta$, we may locally solve $\hat{g} = 0$ for $\alpha = \alpha(\xi)$ by the implicit function theorem. ∎

2.1. The main result. From the following result we can deduce many properties concerning the flow of (1.1)–(1.2).

THEOREM 2.2. *Assume* A1–A5 *hold and recall* $U = \langle C^T u \rangle^\perp \subset \mathbb{R}^n$. *There is a* C^ω*-diffeomorphism* $\chi : B(0,0) \subset U \times \mathbb{R} \to \mathbf{C}$, *where* $B(0,0)$ *is a neighborhood of* $(0,0)$, *with the following properties. The map* $(x(\cdot), y(\cdot))$ *is a solution of* (1.1)–(1.2) *in* $\mathcal{U}$ *with* $k^T y(\cdot) \in W^{1,\infty}(I, \mathbb{R})$ *if and only if* $(x(t), y(t)) = \chi(\alpha(t), \beta(t))$, *where* (α, β) *satisfies*

$$\dot{\alpha} = L_0 \alpha + \rho_0(\alpha, \beta), \tag{2.2}$$

$$s(\alpha, \beta)\dot{\beta} = \beta + \rho_1(\alpha, \beta), \tag{2.3}$$

with $(\alpha, \beta) \in C^1(I, U) \times W^{1,\infty}(I, \mathbb{R})$ *and* (2.2)–(2.3) *satisfied for a.e.* $t \in I$.

The map $L_0 \in GL(U)$ *satisfies* $\sigma(L_0) = \sigma(M, L)$ *and* $\rho_0 \times \rho_1 : B(0,0) \to U \times \mathbb{R}$ *is* C^ω *and* $\mathcal{O}(2)$ *at zero. Moreover,* $s : B(0,0) \to \mathbb{R}$ *is* C^ω *and* $\chi(s^{-1}(0) \cap B(0,0)) = \mathbf{S}$, $s(0,0) = 0$, *and* $d_\beta s(0,0) \neq 0$. *Consequently,* $\Sigma := s^{-1}(0) \subset U \times \mathbb{R}$ *is an* $(n-1)$*-dimensional manifold.*

Proof. Using Theorem 1.1 we may write $\mathbb{R}^n = U \oplus \langle Bk \rangle$ and $\mathbb{R}^m = \langle k \rangle \oplus \langle k \rangle^\perp$. Now write $x = x_0 + x_1 Bk \in U \oplus \langle Bk \rangle$ and $y = y_1 k + y_0 \in \langle k \rangle \oplus \langle k \rangle^\perp$.

As in (1.7), we can write (1.1)–(1.2) as

$$\begin{aligned} \dot{x} &= Ax + By + \mathcal{F}(x, y), \\ 0 &= Cx + Dy + \mathcal{G}(x, y), \end{aligned} \tag{2.4}$$

where $\mathcal{F}$ and $\mathcal{G}$ are $\mathcal{O}(2)$ at $(0,0)$. Hence, the constraint (1.2) becomes

$$\begin{aligned} g(x, y) &= g(x_1 Bk + x_0, y_1 k + y_0) \\ &= x_1 CBk + Cx_0 + Dy_0 + \mathcal{G}(x_1 Bk + x_0, y_1 k + y_0) \\ &=: \Gamma(x_1, x_0, y_1, y_0) \\ &= 0. \end{aligned}$$

Now define the linear mapping $\Delta \in \mathcal{L}(\mathbb{R} \times \langle k \rangle^\perp, \mathbb{R}^m)$ by

$$\Delta[a, b] := d_{(x_1, y_0)} \Gamma(\mathbf{0})[a, b] = aCBk + D|_{\langle k \rangle^\perp} b$$

for $a \in \mathbb{R}$ and $b \in \langle k \rangle^\perp$. Since $\langle u \rangle = \mathrm{N}(D^T)$, then $\Delta[a, b] = 0$ implies $au^T CBk = 0$ so that $a = 0$. Since $Db = 0$ therefore follows and because b lies in a space complementary to $\langle k \rangle$, we find that $b = 0$ too. Since Δ is thus an injection of finite-dimensional spaces of the same dimension, it is a bijection. One can therefore solve $g(x, y) = 0$ locally and uniquely for C^ω functions X and Y such that $x_1 = X(x_0, y_1)$ and $y_0 = Y(x_0, y_1)$.

Now define the local diffeomorphism $\bar{\chi} \in C^\omega(U \times \mathbb{R}, \mathbf{C})$ by

$$\bar{\chi}(x_0, y_1) := (x_0 + X(x_0, y_1) Bk, y_1 k + Y(x_0, y_1)).$$

Denote, from (1.6),

$$L^{-1} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \in \mathcal{L}(\mathbb{R}^{n+m}); \tag{2.5}$$

define $(\mathcal{F}_0 \times \mathcal{G}_0) := L^{-1}(\mathcal{F} \times \mathcal{G})$; and note from [2] that $U = \mathrm{R}(A_1)$. Using Theorem 7 from [2] we find that

$$C_1 Bk = k, \mathrm{N}(A_1) = \langle Bk \rangle, \quad Bk \notin \mathrm{R}(A_1).$$

Recall also that the restricted map $A_0 := A_1|_{\mathrm{R}(A_1)} \in GL(U)$ satisfies $\sigma(M, L) = \sigma(A_0^{-1})$.

Multiplying (2.4) by L^{-1} we can write (1.1)–(1.2) as

$$A_1 \dot{x} = x + \mathcal{F}_0(x, y), \tag{2.6}$$

$$C_1 \dot{x} = y + \mathcal{G}_0(x, y). \tag{2.7}$$

By forming the decomposition $\mathcal{F}_0(x, y) = \mathcal{F}_r(x, y) + \mathcal{F}_b(x, y)Bk \in U \oplus \langle Bk \rangle$, where $\mathcal{F}_b(x, y) = u^T C \mathcal{F}_0(x, y)/u^T CBk$ and $\mathcal{F}_r = \mathcal{F}_0 - \mathcal{F}_b Bk$, we obtain

$$\begin{aligned} A_1 \dot{x} &= A_1(\dot{x}_1 Bk + \dot{x}_0) \\ &= A_1 \dot{x}_0 \\ &= x_1 Bk + x_0 + \mathcal{F}_b(x, y)Bk + \mathcal{F}_r(x, y). \end{aligned}$$

By projecting this onto U along $\langle Bk \rangle$, we then obtain

$$A_1 \dot{x}_0 = x_0 + \mathcal{F}_r(x_0 + X(x_0, y_1)Bk, Y(x_0, y_1) + y_1 k).$$

But A_0 is the restriction of A_1 to $\mathrm{R}(A_1)$, so that

$$\dot{x}_0 = A_0^{-1} x_0 + \rho(x_0, y_1), \tag{2.8}$$

where $\rho(x_0, y_1) = A_0^{-1} \mathcal{F}_r(x_0 + X(x_0, y_1)Bk, Y(x_0, y_1) + y_1 k)$ is a C^ω function and $\mathcal{O}(2)$ at the origin.

From (2.7) one may write $C_1 \dot{x} = \dot{x}_1 C_1 Bk + C_1 \dot{x}_0 = y_0 + y_1 k + \mathcal{G}_0(x, y)$. Taking the inner product of this with k yields

$$\dot{x}_1 + k^T C_1 \dot{x}_0 = y_1 + k^T \mathcal{G}_0(x, y),$$

recalling that $k^T k = 1$. This implies

$$\dot{x}_1 + k^T C_1 [A_0^{-1} x_0 + \rho(x_0, y_1)] = y_1 + \kappa(x_0, y_1),$$

where $\kappa(x_0, y_1) = k^T \mathcal{G}_0(x_0 + X(x_0, y_1)Bk, Y(x_0, y_1) + y_1 k)$.

Now we find another expression for $\dot{x}_1$, using the fact that $y_1(\cdot) = k^T y(\cdot) \in W^{1,\infty}$ by assumption gives

$$\dot{x}_1 = \frac{d}{dt} X(x_0, y_1) = d_{x_0} X[\dot{x}_0] + d_{y_1} X[\dot{y}_1] = d_{x_0} X[A_0^{-1} x_0 + \rho(x_0, y_1)] + d_{y_1} X[\dot{y}_1];$$

then

$$\begin{aligned} d_{x_0} X[A_0^{-1} x_0 + \rho(x_0, y_1)] + \dot{y}_1 d_{y_1} X[1] \\ + k^T C_1 [A_0^{-1} x_0 + \rho(x_0, y_1)] = y_1 + \kappa(x_0, y_1). \end{aligned} \tag{2.9}$$

The proof is essentially complete, but to simplify the notation a little, let us write

$$L_0 := A_0^{-1}, \quad p := x_0, \quad q := y_1, \quad \bar{s}(p,q) := d_q X(p,q)[1],$$

and $a := -k^T C_1 A_0^{-1}$. From (2.9) we find a function r, given by $r(p,q) = \kappa(p,q) - (k^T C_1 + d_{x_0} X(p,q))\rho(p,q) - d_{x_0} X(p,q) A_0^{-1} p$, such that

$$\dot{p} = L_0 p + \rho(p,q), \qquad \bar{s}(p,q)\dot{q} = a^T p + q + r(p,q). \tag{2.10}$$

We claim that

$$\bar{s}(0,0) = 0, \qquad d_q \bar{s}(0,0)[1] = -u^T d^2_{yy} g(0,0)[k,k]/u^T CBk, \tag{2.11}$$

and

$$d_p \bar{s}(0,0)[p] = -u^T d^2_{xy} g(0,0)[p,k]/u^T CBk \tag{2.12}$$

for all $p \in U$ and $q \in \mathbb{R}$. To prove this claim, we use the fact that

$$g(X(p,q)Bk + p, qk + Y(p,q)) \equiv 0;$$

differentiating and evaluating this expression at zero yield (2.11) and (2.12).

Now define new coordinates $(\alpha, \beta) := (p, a^T p + q)$, and let

$$\chi(\alpha,\beta) := \bar{\chi}(\alpha, \beta - a^T\alpha), \qquad s(\alpha,\beta) := \bar{s}(\alpha, \beta - a^T\alpha).$$

This provides the C^ω functions ρ_0 and ρ_1 such that (p,q) satisfies (2.10) if and only if (α,β) satisfies (2.2)–(2.3).

Since $\mathbf{S}$ and $\chi(\Sigma)$ have dimension equal to $n-1$, to prove $\chi(\Sigma \cap B(0,0)) = \mathbf{S}$ it suffices to prove that $\chi(\Sigma \cap B(0,0)) \subset \mathbf{S}$, and we know from Lemma 2.1 that $\mathbf{C}$ is an n-dimensional manifold containing $(0,0)$ and $\mathbf{S}$ is a codimension-1 submanifold of $\mathbf{C}$, also containing $(0,0)$. Thus, let $(x,y) = (x_0 + x_1 Bk, y_0 + y_1 k) \in \mathbf{C}\backslash\mathbf{S}$ satisfy $s(x_0, y_1) = 0$. One can solve $g(x,y) = 0$ uniquely for $y = y(x)$ near this point by the implicit function theorem. Hence, locally, $g(x,y_1) = g(x,y_2) = 0$ implies $y_1 = y_2 = y(x)$.

Define the smooth function $w : \mathbb{R}^2 \to \mathbb{R}$ by

$$w(\theta,\tau) := \theta - X(x_0,\tau),$$

and note that $w(x_1,y_1) = 0$. By definition, $d_\tau w(x_1,y_1) = -s(x_0,y_1) = 0, d_\theta w(x_1,y_1) = 1$, and when (x,y) is of sufficiently small norm we may assume without the loss of any generality that $d^2_{\tau\tau} w(x_1,y_1) \neq 0$ because $d_q s(0,0) \neq 0$. By the saddle-node bifurcation theorem there are two distinct solution branches of $w(\theta,\tau) = 0$ on which $\tau = \tau_\pm(\theta)$, say. Now suppose that a sequence $(x_1^m) \subset \mathbb{R}$ satisfies $x_1^m \to x_1$ as $m \to \infty$, so that the two sequences in $\mathbb{R}^{n+m}$ given by $((x_0 + \tau_\pm(x_1^m)Bk, y_0 + y_1 k))_m$ lie in $\mathbf{C}\backslash\mathbf{S}$ for m large enough. By uniqueness it follows that $\tau_+(x_1^m) \equiv \tau_-(x_1^m)$, a contradiction. Therefore, no such (x,y) exists and the result is proven. □

In light of Theorem 2.2, we define the following terminology. Suppose that $I \subset \mathbb{R}$ is a bounded, open interval. We call a map $(\alpha,\beta) \in C^1(I,U) \times W^{1,\infty}(I,\mathbb{R})$ a *sharp solution* of (2.2)–(2.3) if this differential equation holds for almost every $t \in I$, provided that $\dot{\beta} \notin C^0(I,\mathbb{R})$. A map $(\alpha,\beta) \in C^1(I, U \times \mathbb{R})$ is said to be a *smooth solution* of (2.2)–(2.3) if this differential equation is satisfied for all $t \in I$.

We assume throughout, without the loss of any generality, that $d_\beta s(\alpha,\beta) \neq 0$ for all $(\alpha,\beta) \in \Sigma$. Due to the fact that Σ is diffeomorphic to $\mathbf{S}$ and because existence and uniqueness of (2.2)–(2.3) may break down along Σ, we shall also describe Σ as *the singularity*.

For the moment let us record the fact, taken from the above proof, that

$$d_\alpha s(0,0)[p] = -u^T \left(d^2_{xy} g(0,0)[p,k] - a^T p \; d^2_{yy} g(0,0)[k,k] \right) / u^T CBk,$$

where k is defined in A2 and a is given in the proof. Let us note that the following assumption ensures that $d_\alpha s(0,0)$ is a nonzero map:

A6. $\exists p' \in U$ such that $d^2_{xy} g(0,0)[p',k] - a^T p' \; d^2_{yy} g(0,0)[k,k] \notin \mathrm{R}(d_y g(0,0))$.

Using Theorem 2.1 of [9] we can describe the impasse points of (2.2)–(2.3) as follows.

LEMMA 2.3. *Assuming* A1–A6*, if* $(\alpha,\beta) \in \Sigma$ *satisfies* $\beta + \rho_1(\alpha,\beta) \neq 0$*, then* (α,β) *is an impasse point for* (2.2)–(2.3).

Therefore, the set

$$P := \{(\alpha,\beta) \in \Sigma : \beta + \rho_1(\alpha,\beta) = 0\} \tag{2.13}$$

forms a subset of the singularity which does not necessarily contain impasse points, but the following lemma shows that P represents a nongeneric set of singular points.

LEMMA 2.4. *Assuming* A1–A6*, the set of pseudoequilibria of* (2.2)–(2.3)*,* $P \subset B(0,0)$*, is a codimension-*1 *submanifold of* Σ.

Proof. Use the implicit function theorem to solve the system $\beta + \rho_1(\alpha,\beta) = 0,\ s(\alpha,\beta) = 0$ near $(\alpha,\beta) = (0,0)$. □

Nevertheless, the following result shows that (2.2)–(2.3) is well behaved at P in the sense that there exists a smooth solution of this quasilinear ODE through every point in P.

THEOREM 2.5. *Suppose that* A1–A6 *hold and let* $r \in \mathbb{N}$*. There is a neighborhood* $B^{(r)}(0,0) \subset B(0,0)$ *and at least one* $(n-1)$*-dimensional, quasi-invariant* C^r *manifold* $W^R \subset B^{(r)}(0,0)$ *of* (2.2)–(2.3) *such that for each* $(\alpha_0,\beta_0) \in W^R$*, there exists an open interval* $I \ni 0$ *and a unique* C^r*-solution of* (2.2)–(2.3)*,* $(\alpha,\beta) : I \to W^R$ *such that* $(\alpha(0),\beta(0)) = (\alpha_0,\beta_0)$*. Moreover,* $W^R \cap \Sigma = P$.

Proof. Make the following change of time-scale: if $(\alpha(t),\beta(t))$ satisfies (2.2)–(2.3), define τ by

$$\frac{d\tau}{dt} = \frac{1}{s(\alpha(t),\beta(t))}, \qquad \tau(t_0) = \tau_0,$$

and write $\alpha(\tau) = \alpha(t(\tau)), \beta(\tau) = \beta(t(\tau))$. If a prime denotes $\frac{d}{d\tau}$, then

$$\alpha' = (L_0\alpha + \rho_0(\alpha,\beta)) s(\alpha,\beta), \tag{2.14}$$

$$\beta' = \beta + \rho_1(\alpha,\beta). \tag{2.15}$$

Linearizing (2.14)–(2.15) around the equilibrium point $(\alpha,\beta) = (0,0)$, we find *at least one* C^r, local center manifold $W^R := W^c_{loc}$. This is a quasi-invariant manifold for (2.2)–(2.3) on which $\beta = h(\alpha)$, where $h(0) = 0$ and $dh(0) = 0$. Now suppose that $s(\alpha_0,\beta_0) = 0$ and $(\alpha_0,\beta_0) \in W^R$, and let $(\alpha(\tau),\beta(\tau))$ be the solution of (2.14)–(2.15) in W^R with $(\alpha(0),\beta(0)) = (\alpha_0,\beta_0)$. Then

$$\beta + \rho_1(\alpha,\beta) = \beta' = dh(\alpha)\alpha' = dh(\alpha)(L_0\alpha + \rho_0(\alpha,\beta)) s(\alpha,\beta),$$

and setting $\tau = 0$ shows that $W^R \cap \Sigma \subseteq P$. However, the left-hand side of this inclusion is given by those α for which $s(\alpha, h(\alpha)) = 0$. This equation can be solved by the implicit function theorem, showing that $W^R \cap \Sigma$ is also an $(n-2)$-dimensional manifold. Since $W^R \cap \Sigma$ and P are manifolds of the same dimension and one is contained in the other, they coincide. The uniqueness of solutions of (2.2)–(2.3) in W^R follows from a standard ODE uniqueness theorem applied to $\dot{\alpha} = L_0\alpha + \rho_1(\alpha, h(\alpha))$, with $\beta(t) = h(\alpha(t))$. ∎

While the existence of W^R is assured from the center manifold theorem, it is not clear that there will be only one W^R with the properties outlined in Theorem 2.5. For this reason, we cannot claim that W^R is an invariant manifold, we can claim only quasi-invariance.

The following definition is given merely for completeness, and it provides the analogy of stable and unstable manifolds for (2.2)–(2.3).

DEFINITION 1 (local stable and unstable sets). *Let $B' \subset U \times \mathbb{R}$ be a neighborhood of $(0,0)$. The local stable set $W^s(0,0) \subset U \times \mathbb{R}$ is the set of $(\alpha, \beta) \in B'$ such that there exists a solution $(\alpha(t), \beta(t))$ of (2.2)–(2.3) with $(\alpha(0), \beta(0)) = (\alpha, \beta)$, $(\alpha(t), \beta(t)) \in B'$ for all $t \geq 0$ and $(\alpha(t), \beta(t)) \to 0$ as $t \to \infty$. The local unstable set $W^u(0,0)$ is defined analogously with $t \leq 0$ and the limit $t \to -\infty$ used above.*

PROPOSITION 2.6. *Suppose that* A1–A6 *hold and that (M, L) is a hyperbolic matrix pencil. Now define*

$$n_{\pm} := \#\left(\sigma(M, L) \cap \mathbb{C}^{\pm}\right),$$

both assumed to be nonzero, noting $n_- + n_+ = n-1$. Then there is an invariant subset of the stable set of (2.2)–(2.3), *$W^{Rs} \subset W^R$, which is an (n_-)-dimensional manifold, and an invariant subset of the unstable set of* (2.2)–(2.3), *$W^{Ru} \subset W^R$, which is an (n_+)-dimensional manifold.*

Proof. This uses the existence of the quasi-invariant manifold, W^R, of (2.2)–(2.3) on which $\beta = h(\alpha)$. The result follows since the ODE $\dot{\alpha} = L_0\alpha + \rho_1(\alpha, h(\alpha))$ has stable and unstable manifolds of the stated dimensions and using the fact that $\sigma(M, L) = \sigma(L_0)^{-1}$ from Theorem 2.2. ∎

Let us note that the fact that the stable and unstable sets $W^{s,u}(0,0)$ associated with (2.2)–(2.3) are not necessarily manifolds is simply due to the ellipticity of the zero equilibrium of (2.14)–(2.15).

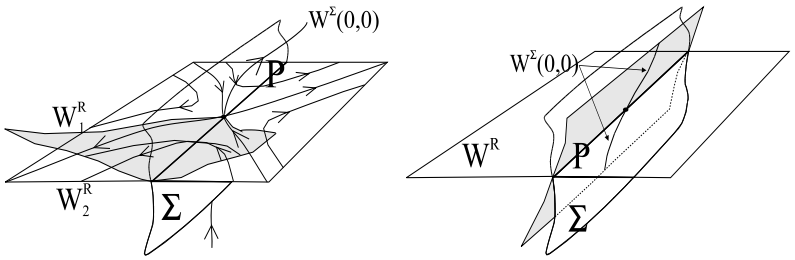
Now we use the remaining information in the normal form (2.2)–(2.3) to deduce that not only are there singularity-traversing solutions contained in W^R, there are other quasi-invariant manifolds which intersect the singularity Σ.

PROPOSITION 2.7. *Suppose that* A1–A6 *apply. Associated with each $(\alpha, \beta) \in P$ is a C^ω, one-dimensional, quasi-invariant manifold of* (2.2)–(2.3), *$W^\Sigma(\alpha, \beta)$, which is transverse to both W^R and Σ at (α, β). Moreover, if $(\alpha_0, \beta_0) \in W^\Sigma(0,0) \backslash (0,0)$, there exists a $T \in \mathbb{R}$ and a solution $(\alpha(t), \beta(t))$ of* (2.2)–(2.3) *on $[0, T]$ such that $(\alpha(0), \beta(0)) = (\alpha_0, \beta_0)$ and* sign $s(\alpha(T), \beta(T)) = -$sign $s(\alpha(0), \beta(0))$.

Proof. Suppose that $(\alpha, \beta) \in P$, so that (α, β) is an equilibrium of (2.14)–(2.15). Linearizing (2.14)–(2.15) around this equilibrium gives a smoothly parameterized mapping $T \in C^\omega\left(B(0,0), \mathcal{L}\left(U \times \mathbb{R}\right)\right)$ such that

$$T(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

where $B(0,0)$ is defined in Theorem 2.2. Since 1 is an algebraically simple eigenvalue of $T(0,0)$, by spectral perturbation results [5] there are C^ω functions $\lambda : B(0,0) \to \mathbb{R}$



(a) (b)

FIG. 2.1. *A typical flow near a singular equilibrium (solid dot).* (a) *Two instances of* W^R, $W^R_{1,2}$ *are shown, where* $W^\Sigma(0,0)$ *and* Σ *are shown transverse at* $(0,0)$. (b) *The relative positions of* Σ, W^R, *and* $W^\Sigma(0,0)$; *the shaded set is* $\bigcup_{(\alpha,\beta)} W^\Sigma(\alpha,\beta)$. *Elements of* $\Sigma\backslash P$ *are impasse points.*

and $e : B(0,0) \to U \times \mathbb{R}$ such that $\lambda(\alpha,\beta) \in \sigma(T(\alpha,\beta))$, with corresponding unit eigenvector $e(\alpha,\beta)$, such that $e(0,0) = (0,1)$ and $\lambda(0,0) = 1$. Hence we may assume with loss of generality that $\lambda(\alpha,\beta)$ is positive whenever $(\alpha,\beta) \in P \cap B(0,0)$. From this it follows that each $(\alpha,\beta) \in P$ has an associated local unstable manifold, $W^u(\alpha,\beta)$, which we write as $W^\Sigma(\alpha,\beta)$.

The representation of $W^\Sigma(0,0)$ is given by a graph of the form $\alpha = \ell(\beta)$, such that $\ell(0) = 0$ and $d\ell(0) = 0$. Therefore, the solutions of (2.2)–(2.3) on $W^\Sigma(0,0)$ are images of the solutions of the scalar ODE

$$\dot{\beta} = \frac{\beta + \rho_1(\ell(\beta),\beta)}{s(\ell(\beta),\beta)} = d_\beta s(0,0)^{-1} + O(\beta),$$

and the right-hand side of this is nonzero in a neighborhood of $\beta = 0$. Hence the solution passes through the regular point $\beta = 0$ in finite time.

Since $e(\cdot,\cdot)$ varies smoothly, it follows without the loss of any generality that each $W^\Sigma(\alpha,\beta)$ is transverse to Σ if $W^\Sigma(0,0)$ is transverse to Σ. Therefore, let us calculate $T_0(\Sigma)$, given that $T_0(W^\Sigma(0,0)) = U \times \{0\} \subset U \times \mathbb{R}$. Since we may solve $s(\alpha,\beta) = 0$ near $(0,0)$ for $\beta = \beta(\alpha)$ such that $s(\alpha,\beta(\alpha)) \equiv 0$, we find $d\beta(0) = -d_\beta s(0,0)^{-1} d_\alpha s(0,0) \neq 0$ and $T_0(\Sigma) = \{(\alpha, d\beta(0)\alpha) : \alpha \in U)\}$. It follows that $\dim(T_0(\Sigma) \oplus T_0(W^\Sigma(0,0))) = n$ and therefore the manifolds Σ and $W^\Sigma(\alpha,\beta)$ intersect transversally at (α,β). □

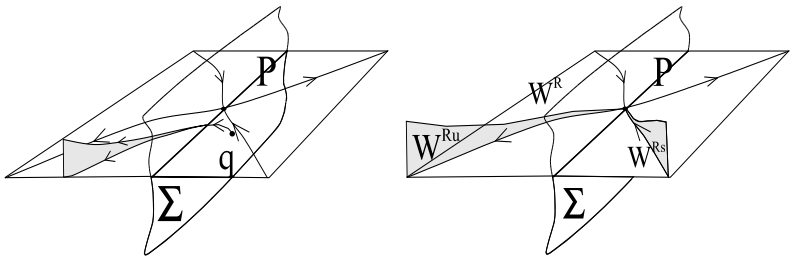
In [11], the authors use W^{sing} to denote a one-dimensional, quasi-invariant manifold containing the singular equilibrium. We use $W^\Sigma(\alpha,\beta)$ (and $W^\Sigma(0,0)$ is W^{sing}) to underline the fact that through every point on this set, there is a solution which can be extended to the singularity Σ. (See Figure 2.1(b).)

Let us note that it is possible for a subset of $W^\Sigma(0,0)$ to lie in either the stable or unstable set associated with (2.2)–(2.3). Indeed, we shall define

$$W^{\Sigma s}(0,0) := W^\Sigma(0,0) \cap W^s(0,0),$$

and similarly $W^{\Sigma u}(0,0) := W^\Sigma(0,0) \cap W^u(0,0)$. Theorem 2.8 below shows that $W^{\Sigma s}$ and $W^{\Sigma u}$ are not empty if A1–A6 apply, because of the existence of sharp solutions.

THEOREM 2.8. *Assuming* A1–A6, *through the singular equilibrium of* (2.2)–(2.3) *there pass two smooth solutions with* $(\alpha,\beta) \in C^\omega \times C^\omega$ *and two sharp solutions with* $(\alpha,\beta) \in C^1 \times W^{1,\infty}$. *Consequently,* $W^{\Sigma u}(0,0)$ *and* $W^{\Sigma s}(0,0)$ *are nonempty.*



(a) (b)

FIG. 2.2. *A typical flow near a singular equilibrium:* (a) *There are multiple instances of* W^R*, showing the possible lack of uniqueness of solutions along* $P = W^R \cap \Sigma$*. The shaded region shows the union of all possible forward orbits of* q *after encountering the singularity.* (b) *The local stable* (W^{Rs}) *and unstable* (W^{Ru}) *sets associated with* $(0,0)$.

Proof. One of the $C^\omega \times C^\omega$ solutions is the trivial equilibrium solution itself. The other smooth solution is obtained from the trajectory of (2.2)–(2.3) whose image forms $W^\Sigma(0,0)$.

Now concatenate a trajectory of (2.2)–(2.3) with initial condition on $W^\Sigma(0,0)$ to the equilibrium solution to form two sharp solutions. To see that this procedure forms a sharp solution, consider the solution (α, β) of (2.2)–(2.3) which is zero for $t \geq 0$ but lies on $W^\Sigma(0,0)$ for $t \in (-T, 0]$. Then, for $t \leq 0$, from the proof of Proposition 2.7 we have $\alpha = \ell(\beta)$, where $\ell(0) = 0$ and $d\ell(0) = 0$, but $\dot\beta = (\beta + \rho_1(\ell(\beta), \beta))/s(\ell(\beta), \beta)$. It follows that $\dot\beta(0_-) = 1/d_\beta s(0,0) \neq 0$ but $\dot\beta(0_+) = 0$, and $\dot\beta \in L^\infty(-T, T)$ for small enough $T > 0$. Another sharp solution is obtained by starting on the equilibrium solution before leaving the equilibrium along $W^\Sigma(0,0)$ in an analogous manner. ∎

An illustration of the invariant manifolds discussed in this section is given in Figures 2.1 and 2.2.

3. Discussion. Let us consider two examples which illustrate some of the fundamental ideas within the paper. The first example is somewhat artificial, but it clearly shows how smooth solutions can be concatenated to form less regular ones.

Example 1. Consider

$$\dot x = y, \qquad x^2 + y^2 = 1, \tag{3.1}$$

where $(x, y) = (\pm 1, 0)$ are both singular equilibrium points, so that there exists a trivial smooth solution passing through them. However, $(x(t), y(t)) = (\cos(t), \sin(t))$ is another smooth solution passing through these two points; this is precisely the behavior we observe at a singular equilibrium in higher dimensions, with $W^\Sigma(0,0)$ playing the role of the circle of this example. The concatenated function

$$(x(t), y(t)) = \begin{cases} (1,0), & t \leq 0, \\ (\cos(t), \sin(t)), & t \geq 0, \end{cases}$$

is a solution of (3.1) of class $C^1 \times W^{1,\infty}$. Indeed, because $\dot y(0_-) = 0$ and $\dot y(0_+) = -1$ we have $y \notin C^1(\mathbb{R})$, although $\dot y \in L^\infty(\mathbb{R})$.

This example demonstrates that the multiplicity of sharp solutions can be much greater than that of smooth solutions. This arises in this particular instance because of the existence of a connecting orbit between the two singular equilibria. So, in order

to form a continuum of sharp solutions, one can simply "wait" for some arbitrary time on arrival at the singular equilibrium before continuing around the circle to the other singular equilibrium.

Example 2. Consider a degenerate form of the *Fitzhugh–Nagumo* equation:

$$u_t = \frac{1}{2}(u^2)_{xx} + u(1-u) + v, \tag{3.2}$$

$$v_t = u - v, \qquad x \in \mathbb{R}, t > 0. \tag{3.3}$$

Scalar problems of this type can be found in [4, 10], where the authors are interested in the support of the waves, which may be finite, semi-infinite, or infinite. This equation is related to the much-studied Fitzhugh–Nagumo equation, except for the inclusion of the degenerate diffusive term. By seeking a traveling-wave solution of (3.2)–(3.3) which connects $(u, v) = (0, 0)$ to itself, we obtain the quasilinear problem

$$\left(cu - \frac{1}{2}(u^2)_z\right)_z = u(1-u) + v, \tag{3.4}$$

$$cv_z = u - v, \qquad u(\pm\infty), v(\pm\infty) = 0, \ z = x + ct, \tag{3.5}$$

where $c > 0$ is the wave speed. To study (3.4)–(3.5) as a DAE, we require $u \in C^0$ to also satisfy $u^2 \in C^1$ and $cu - \frac{1}{2}(u^2)_z \in C^1$, rather than simply allowing $u \in C^2$. This ensures that the resulting solutions are *weak* solutions if one considers (3.4)–(3.5) in a standard weak formulation [6]. A simple example of a function which satisfies such a regularity requirement is $u(t)$, where $u(t) = 0$ for $t < 0$ and $u(t) = t$ for $t \geq 0$ and $c = 1$.

We can manipulate this system to see explicitly how the results of the previous sections apply in this specific case. Thus, put $U = u - W$ and write (3.4)–(3.5) as a DAE,

$$w_z = cW, \tag{3.6}$$

$$cU_z = (W + U)(1 - W - U) + v, \tag{3.7}$$

$$cv_z = W + U - v, \tag{3.8}$$

$$0 = w - \frac{1}{2}(U + W)^2, \tag{3.9}$$

to which A1–A6 apply if $c \neq 0$. The constraint manifold for this problem is $\mathbf{C} = \{(w, U, v, W) \in \mathbb{R}^4 : (U + W)^2 = 2w\}$, and the singularity is $\mathbf{S} = \{(w, U, v, W) \in \mathbf{C} : U + W = 0\}$. Differentiating (3.9), we obtain a quasilinear ODE which is analogous to (2.2)–(2.3):

$$(U + W)W_z = cW - (U + W)[(U + W)(1 - U - W) - v]c^{-1}, \tag{3.10}$$

$$cv_z = W + U - v, \tag{3.11}$$

$$cU_z = (W + U)(1 - W - U) + v, \tag{3.12}$$

which, upon rescaling time, gives

$$W' = cW - (U + W)[(U + W)(1 - U - W) - v]c^{-1}, \tag{3.13}$$

$$cv' = (U + W)[W + U - v], \tag{3.14}$$

$$cU' = (U + W)((W + U)(1 - W - U) + v). \tag{3.15}$$

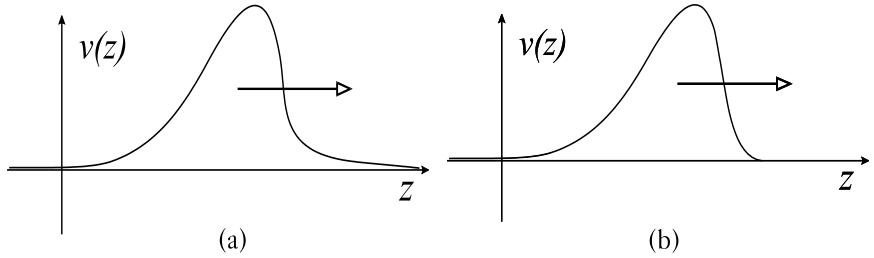


FIG. 3.1. (a) *A solution with a head and a tail when* W^{Rs} *intersects* W^{Ru}. (b) *A solution with a tail but no head when* $W^{\Sigma s}$ *intersects* W^{Ru}. *In this case, the solution is identically zero ahead of the wave.*

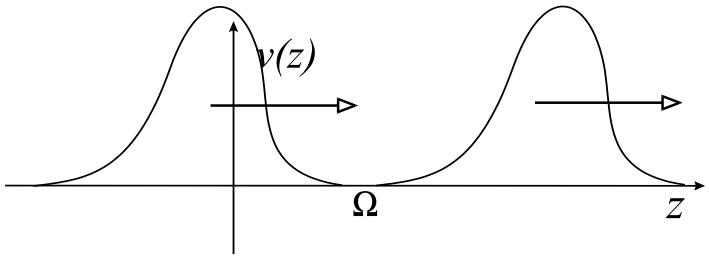


FIG. 3.2. *A solution when* $W^{\Sigma s}$ *intersects* $W^{\Sigma u}$, *and multiple waves arise. The solution is zero between each wave in the region* Ω.

It is straightforward to show that at the singular equilibrium point $(W, v, U) = (0, 0, 0)$, (3.10)–(3.12) has at least one quasi-invariant manifold, W^R, described by a graph of the form $W = h(U, v)$. For $c > 0$, there is another quasi-invariant manifold, $W^\Sigma(0, 0)$, on which $U = H_1(W)$ and $v = H_2(W)$. Now, restricting (3.4)–(3.5) induces a local dynamical system on W^R, given by the restricted flow of an ODE of the form

$$cv_z = U - v + \mathcal{O}(2), \tag{3.16}$$

$$cU_z = U + v + \mathcal{O}(2). \tag{3.17}$$

Since the equilibrium of this system, $(U, v) = (0, 0)$, has a stable and unstable manifold, it follows that (3.4)–(3.5) has at least two one-dimensional invariant manifolds, W^{Rs} and W^{Ru}, within its stable and unstable sets. The arrival time of solutions at the zero equilibrium along these manifolds must be infinite, by standard ODE uniqueness results, as applied to (3.16)–(3.17).

On $W^\Sigma(0, 0)$, using (3.10), we have an ODE

$$W_z = \frac{cW + O(W^2)}{W + H_1(W)} = c + O(W)$$

for small $|W|$. One can verify directly that $\mathbf{S}$ and $W^\Sigma(0, 0)$ intersect transversally in $\mathbf{C}$, so that there is at least one singularity-traversing smooth solution of (3.4)–(3.5). It follows that there are also sharp solutions which start and end at the equilibrium, existing on either side of the singularity. These are again formed by concatenating the trivial equilibrium solution to the trajectory which forms $W^\Sigma(0, 0)$.

While this does not provide any information as to whether a homoclinic orbit exists in (3.4)–(3.5), the intersections of the various manifolds involved will yield different types of traveling waves, as depicted in Figures 3.1 and 3.2.

Acknowledgments. Many thanks are due to the anonymous referees whose comments helped to improve an earlier version of the paper.

REFERENCES

[1] R. E. Beardmore, *Stability and bifurcation properties of index-1 DAEs*, Numer. Algorithms, 19 (1998), pp. 43–53.
[2] R. E. Beardmore, *The singularity-induced bifurcation and its Kronecker normal form*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 126–137.
[3] S. Campbell, R. Hollenbeck, and W. Marszalek, *Mixed symbolic-numeric computations with general DAE* I: *System properties*, Numer. Algorithms, 19 (1998), pp. 73–84.
[4] A. de Pablo and A. Sanchez, *Global travelling waves in reaction-convection diffusion equations*, J. Differential Equations, 165 (2000), pp. 377–413.
[5] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1980.
[6] M. Langlais and D. Phillips, *Stabilization of solutions of nonlinear and degenerate evolution equations*, Nonlinear Anal., 9 (1985), pp. 321–333.
[7] R. März, *Criteria for the trivial solution of differential algebraic equations with small nonlinearities to be asymptotically stable*, J. Math. Anal. Appl., 225 (1998), pp. 587–607.
[8] P. J. Rabier and W. C. Rheinboldt, *A geometric treatment of quasilinear differential-algebraic equations*, J. Differential Equations, 109 (1994), pp. 110–146.
[9] P. J. Rabier and W. C. Rheinboldt, *On impasse points of quasilinear differential-algebraic equations*, J. Math. Anal. Appl., 181 (1994), pp. 429–454.
[10] F. Sanchez-Garduno and P. K. Maini, *Travelling-wave phenomena in some degenerate reaction-diffusion equations*, J. Differential Equations, 117 (1995), pp. 281–319.
[11] V. Venkatasubramanian, H. Schattler, and J. Zaborszky, *A Stability Theory of Large Differential Algebraic Systems, a Taxonomy*, Tech. report SSM 9201, Part 1, Washington University, St. Louis, MO, 1992.
[12] H. von Sosen, *Part* I: *Folds and Bifurcations in the Solutions of Semi-Explicit DAEs*, Ph.D. dissertation, California Institute of Technology, Pasadena, CA, 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 121–125

A NOTE ON CONSTRAINT PRECONDITIONING FOR NONSYMMETRIC INDEFINITE MATRICES*

ZHI-HAO CAO†

Abstract. The constraint preconditioner for indefinite linear systems in [C. Keller, N. I. M. Gould, and A. J. Wathen, *SIAM J. Matrix Anal. Appl.*, 21 (2000), pp. 1300–1317] is extended to nonsymmetric matrices.

Key words. preconditioning, indefinite matrix, minimal polynomial

AMS subject classifications. 65F10, 65F15

PII. S0895479801391424

In [2] for the matrix of KKT form

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix},$$

the constraint preconditioner

$$\mathcal{G} = \begin{pmatrix} G & B^T \\ B & 0 \end{pmatrix}$$

is presented (see also Lukšan and Vlček [3]). Here $A, G \in \mathcal{R}^{n,n}, B \in \mathcal{R}^{m,n}, m \leq n$. $\mathcal{A}$ and $\mathcal{G}$ are assumed to be symmetric and nonsingular, thus matrix B must be of full rank. The eigensolution distribution of the preconditioned matrix $\mathcal{G}^{-1}\mathcal{A}$ is determined, and the convergence behavior of a Krylov subspace method such as GMRES is described.

In this note we extend the constraint preconditioner to obtain most of the results in [2] for the nonsymmetric case.

Let $\mathcal{F}$ be a matrix of KKT form

$$\mathcal{F} = \begin{pmatrix} F & B^T \\ B & 0 \end{pmatrix}, \tag{1}$$

and let $\mathcal{M}$ be a constraint preconditioner of $\mathcal{F}$ [2],

$$\mathcal{M} = \begin{pmatrix} M & B^T \\ B & 0 \end{pmatrix}. \tag{2}$$

Here we assume only that $\mathcal{F}$ and $\mathcal{M}$ are nonsingular; i.e., F and M may be nonsymmetric. An example is in the numerical solution of the Navier–Stokes equations of fluid dynamics, where $\mathcal{F} \neq \mathcal{F}^T$.

Following the approach in [1] let

$$B = U\Sigma V^T \tag{3}$$

*Received by the editors June 25, 2001; accepted for publication (in revised form) by M. Hanke February 1, 2002; published electronically July 1, 2002. This work was supported by NSFC project 10171021, the Foundation of National Key Laboratory of Computational Physics, and the Doctoral Point Foundation of China.

http://www.siam.org/journals/simax/24-1/39142.html

†Department of Mathematics and Laboratory of Mathematics for Nonlinear Sciences, Fudan University, Shanghai 200433, People's Republic of China (zhcao@cableplus.com.cn).

be the singular value decomposition of matrix B, where $\Sigma = [\Sigma_0, 0]$, with Σ_0 being an $m \times m$ diagonal matrix with positive diagonal entries, and U, V are orthonormal matrices of order m and n, respectively.

From (3) we have

$$BV \equiv B[V_1, V_2] = [U\Sigma_0, 0],$$

i.e., V_2 is an orthonormal basis for the null space of B.

Let $\mathcal{T} = \mathcal{M}^{-1}\mathcal{F}$; then

$$\begin{aligned}
\widehat{\mathcal{T}} &\equiv \begin{pmatrix} V^T & \\ & U^T \end{pmatrix} \mathcal{T} \begin{pmatrix} V & \\ & U \end{pmatrix} \\
&= \begin{pmatrix} V^T & \\ & U^T \end{pmatrix} \begin{pmatrix} M & B^T \\ B & 0 \end{pmatrix}^{-1} \begin{pmatrix} F & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} V & \\ & U \end{pmatrix} \\
&= \begin{pmatrix} V^T M V & \Sigma^T \\ \Sigma & 0 \end{pmatrix}^{-1} \begin{pmatrix} V^T F V & \Sigma^T \\ \Sigma & 0 \end{pmatrix} \\
&= \begin{pmatrix} V_1^T M V_1 & V_1^T M V_2 & \Sigma_0 \\ V_2^T M V_1 & V_2^T M V_2 & 0 \\ \Sigma_0 & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} V_1^T F V_1 & V_1^T F V_2 & \Sigma_0 \\ V_2^T F V_1 & V_2^T F V_2 & 0 \\ \Sigma_0 & 0 & 0 \end{pmatrix}.
\end{aligned}$$

Finally, $\widehat{\mathcal{T}}$ can be expressed as the following:

$$\widehat{\mathcal{T}} = \begin{pmatrix} \Sigma_0 & 0 & 0 \\ V_2^T M V_1 & V_2^T M V_2 & 0 \\ V_1^T M V_1 & V_1^T M V_2 & \Sigma_0 \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_0 & 0 & 0 \\ V_2^T F V_1 & V_2^T F V_2 & 0 \\ V_1^T F V_1 & V_1^T F V_2 & \Sigma_0 \end{pmatrix}. \tag{4}$$

Since $\mathcal{M}$ and $\mathcal{F}$ are assumed to be nonsingular, $V_2^T M V_2$ and $V_2^T F V_2$ are nonsingular. Let $S = (V_2^T M V_2)^{-1}(V_2^T F V_2)$; then $\widehat{\mathcal{T}}$ can be expressed as

$$\widehat{\mathcal{T}} = \begin{pmatrix} I & & \\ X & S & \\ Z & Y & I \end{pmatrix}. \tag{5}$$

Thus we have the following.

THEOREM 1 (extension of Theorem 2.1 in [2]). *The preconditioned matrix $\mathcal{T} \equiv \mathcal{M}^{-1}\mathcal{F}$ has an eigenvalue at 1 with multiplicity $2m$, and $n-m$ eigenvalues which are those of matrix $S \equiv (V_2^T M V_2)^{-1}(V_2^T F V_2)$.*

We now consider the degree of the minimal polynomial of $\mathcal{T}$ which determines the convergence behavior of a Krylov subspace method such as GMRES (cf. [2, 4]).

LEMMA 2. *Let $p_k(\lambda)$ be a monic polynomial of degree k; then*

$$(\widehat{\mathcal{T}} - I)p_k(\widehat{\mathcal{T}}) = \begin{pmatrix} 0 & 0 & 0 \\ p_k(S)X & (S-I)p_k(S) & 0 \\ W_k & Y p_k(S) & 0 \end{pmatrix} \tag{6}$$

for some W_k.

Proof. If $k = 1$, then $p_k(\lambda) = \lambda - c_1$. We have

$$(\widehat{\mathcal{T}} - I)p_1(\widehat{\mathcal{T}}) = \begin{pmatrix} 0 & 0 & 0 \\ X & S - I & 0 \\ Z & Y & 0 \end{pmatrix} \begin{pmatrix} I - c_1 I & & \\ X & S - c_1 I & \\ Z & Y & I - c_1 I \end{pmatrix}$$
$$= \begin{pmatrix} 0 & & \\ (S - c_1 I)X & (S - I)(S - c_1 I) & \\ W_1 & Y(S - c_1 I) & 0 \end{pmatrix},$$

where

$$W_1 = (1 - c_1)Z + YX.$$

Thus Lemma 2 holds for $k = 1$.

Assume that Lemma 2 holds for any monic polynomial with degree less than k; then any polynomial $p_k(\lambda)$ with degree k can be expressed as

$$p_k(\lambda) = p_{k-1}(\lambda)(\lambda - c_k)$$

with $p_{k-1}(\lambda)$ being a monic polynomial with degree $k - 1$. We have

$$(\widehat{\mathcal{T}} - I)p_k(\widehat{\mathcal{T}}) = (\widehat{\mathcal{T}} - I)p_{k-1}(\widehat{\mathcal{T}})(\widehat{\mathcal{T}} - c_k I)$$

$$= \begin{pmatrix} 0 & & \\ p_{k-1}(S)X & (S - I)p_{k-1}(S) & \\ W_{k-1} & Y p_{k-1}(S) & 0 \end{pmatrix} \begin{pmatrix} I - c_k I & & \\ X & S - c_k I & \\ Z & Y & I - c_k I \end{pmatrix}$$
$$= \begin{pmatrix} 0 & & \\ p_{k-1}(S)(S - c_k I)X & (S - I)p_{k-1}(S)(S - c_k I) & \\ W_k & Y p_{k-1}(S)(S - c_k I) & 0 \end{pmatrix}$$
$$= \begin{pmatrix} 0 & & \\ p_k(S)X & (S - I)p_k(S) & \\ W_k & Y p_k(S) & 0 \end{pmatrix},$$

where

$$W_k = (1 - c_k)W_{k-1} + Y p_{k-1}(S)X.$$

Lemma 2 follows by induction. ∎

THEOREM 3 (extension of Theorems 3.2, 3.5, and 3.7 in [2]). *If the degree of the minimal polynomial of* $S \equiv (V_2^T M V_2)^{-1}(V_2^T F V_2)$ *is* k $(0 \le k \le n - m)$, *then the degree of the minimal polynomial of the preconditioned matrix* $\mathcal{T} \equiv \mathcal{M}^{-1}\mathcal{F}$ *is at most* $2 + k$.

Proof. Let $p_k(\lambda)$ be the minimal polynomial of S; then $p_k(S) = 0$ and (6) implies

$$(\widehat{\mathcal{T}} - I)p_k(\widehat{\mathcal{T}}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ W_k & 0 & 0 \end{pmatrix};$$

thus

$$(\widehat{\mathcal{T}} - I)^2 p_k(\widehat{\mathcal{T}}) = 0.$$

Since $\mathcal{T}$ and $\widehat{\mathcal{T}}$ are similar, they have the same minimal polynomial. The theorem follows. □

Finally, we consider the eigenvector distribution of the preconditioned matrix $\mathcal{T} \equiv \mathcal{M}^{-1}\mathcal{F}$. Let

$$\begin{pmatrix} M & B^T \\ B & 0 \end{pmatrix}^{-1} \begin{pmatrix} F & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}; \tag{7}$$

then from (4) the eigenvalue problem is equivalent to the following generalized eigenvalue problem:

$$\begin{pmatrix} \Sigma_0 & & \\ V_2^T F V_1 & V_2^T F V_2 & \\ V_1^T F V_1 & V_1^T F V_2 & \Sigma_0 \end{pmatrix} \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{y} \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_0 & & \\ V_2^T M V_1 & V_2^T M V_2 & \\ V_1^T M V_1 & V_1^T M V_2 & \Sigma_0 \end{pmatrix} \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{y} \end{pmatrix}, \tag{8}$$

where $\widehat{x}_1 = V_1^T x$, $\widehat{x}_2 = V_2^T x$, and $\widehat{y} = U^T y$.

From (8) we can see that there are m linearly independent eigenvectors $[0^T, 0^T, \widehat{y}^{(j)^T}]^T$, $j = 1, \ldots, m$, corresponding to eigenvalue 1. From (8) we also know that there may be other eigenvectors corresponding to eigenvalue 1, the components $[\widehat{x}_1^T, \widehat{x}_2^T]^T$ of which should satisfy

$$\begin{aligned} V_2^T F V_1 \widehat{x}_1 + V_2^T F V_2 \widehat{x}_2 &= V_2^T M V_1 \widehat{x}_1 + V_2^T M V_2 \widehat{x}_2, \\ V_1^T F V_1 \widehat{x}_1 + V_1^T F V_2 \widehat{x}_2 &= V_1^T M V_1 \widehat{x}_1 + V_1^T M V_2 \widehat{x}_2, \end{aligned}$$

which can be rewritten as the following:

$$V^T (F - M) V \widehat{x} = 0,$$

where $\widehat{x} = [\widehat{x}_1^T, \widehat{x}_2^T]^T$, i.e.,

$$\widehat{x} \in \mathcal{N}((F - M)V).$$

Let $i = \dim \mathcal{N}((F - M)V) (0 \le i \le n)$. If $i > 0$, then let $\widehat{x}^{(l)} \equiv [\widehat{x}_1^{(l)^T}, \widehat{x}_2^{(l)^T}]^T \in \mathcal{N}((F - M)V), l = 1, \ldots, i$, be the linearly independent nullvectors. Then each of the vectors

$$[\widehat{x}^{(l)^T}, \widehat{z}^T]^T \in \mathcal{R}^{m+n}, \quad l = 1, \ldots, i,$$

where $\widehat{z} \in \mathcal{R}^m$ is arbitrary, is an eigenvector of $\widehat{\mathcal{T}}$ corresponding to eigenvalue 1. It is easy to see there are $m + i$ linearly independent eigenvectors of $\widehat{\mathcal{T}}$ corresponding to eigenvalue 1 which can be taken as

$$[0^T, 0^T, \widehat{y}^{(j)^T}]^T, \quad j = 1, \ldots, m; \quad [\widehat{x}_1^{(l)^T}, \widehat{x}_2^{(l)^T}, \widehat{z}^{(l)^T}]^T, \quad l = 1, \ldots, i. \tag{9}$$

Here $\widehat{y}^{(j)} \in \mathcal{R}^m, j = 1, \ldots, m$, are arbitrary m linearly independent vectors, and $\widehat{z}^{(l)} \in \mathcal{R}^m, l = 1, \ldots, i$, are arbitrary i vectors. Note $\dim \mathcal{N}((F - M)V) = \dim \mathcal{N}(F - M)$.

If eigenvalue $\lambda \neq 1$, then from (8) we have $\widehat{x}_1 = 0$ and

$$S\widehat{x}_2 = \lambda \widehat{x}_2, \quad (\lambda - 1)\Sigma_0 \widehat{y} = V_1^T (F - M) V_2 \widehat{x}_2. \tag{10}$$

From (10) we can see that if $\widehat{x}_2 = 0$, then $\widehat{y} = 0$; therefore, if $[\widehat{x}_1^T, \widehat{x}_2^T, \widehat{y}^T]^T$ is an eigenvector of $\widehat{\mathcal{T}}$ corresponding to an eigenvalue $\lambda \neq 1$, then $\widehat{x}_1 = 0$, $\widehat{x}_2$ is an eigenvector

of S, and $\widehat{y} = (\lambda-1)^{-1}\Sigma_0^{-1}V_1^T(F-M)V_2\widehat{x}_2$. Let j $(0 \le j \le n-m)$ be the number of linearly independent eigenvectors of S corresponding to eigenvalues not being 1; then the j linearly independent eigenvectors of $\widehat{\mathcal{T}}$ can be taken as

$$[0^T, \widehat{\chi}_2^{(l)^T}, \widehat{q}^{(l)^T}]^T, \quad l = 1, \ldots, j, \tag{11}$$

where $\widehat{\chi}_2^{(l)}, l = 1, \ldots, j$, are j linearly independent eigenvectors of S corresponding to eigenvalues $\lambda_l \neq 1, l = 1, \ldots, j$, and

$$\widehat{q}^{(l)} = (\lambda_l - 1)^{-1}\Sigma_0^{-1}V_1^T(F-M)V_2\widehat{\chi}_2^{(l)}, \quad l = 1, \ldots, j.$$

Now we have established the eigenvector distribution of the preconditioned matrix $\mathcal{T} \equiv \mathcal{M}^{-1}\mathcal{F}$.

THEOREM 4 (extension of Theorem 2.3 in [2]). *The preconditioned matrix $\mathcal{M}^{-1}\mathcal{F}$ has $n+m$ eigenvalues as defined by Theorem 1 and $m+i+j$ linearly independent eigenvectors. There are*

(i) *$m+i$ $(0 \le i \le n)$ eigenvectors corresponding to the eigenvalue $\lambda = 1$, where $i = \dim\mathcal{N}(F-M)$;*

(ii) *j $(0 \le j \le n-m)$ eigenvectors corresponding to eigenvalues not being 1, where j is the number of linearly independent eigenvectors of S corresponding to eigenvalues not being 1.*

Acknowledgments. I am grateful to the editor and to Professor Andy Wathen for their helpful comments and suggestions. In particular, I am indebted to Andy for his careful reading of the manuscript and pointing out some mistakes in my original argument of Theorem 4.

REFERENCES

[1] R. BANK, B.D. WELFERT, AND H. YSERENTANT, *A class of iterative methods for solving saddle point problems*, Numer. Math., 56 (1990), pp. 645–666.

[2] C. KELLER, N.I.M. GOULD, AND A.J. WATHEN, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.

[3] L. LUKŠAN AND J. VLČEK, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 219–247.

[4] M.F. MURPHY, G.H. GOLUB, AND A.J. WATHEN, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput., 21 (2000), pp. 1969–1972.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 126–131

INEQUALITIES ON SINGULAR VALUES OF BLOCK TRIANGULAR MATRICES*

CHI-KWONG LI† AND ROY MATHIAS†

Abstract. We prove inequalities on singular values for 2×2 block triangular matrices. Using the results, we answer the three questions of Ando on Bloomfield–Watson-type inequalities on eigenvalues and generalize the Kantorovich inequality.

Key words. eigenvalue, Kantorovich inequality, Bloomfield–Watson inequality, positive semidefinite matrix, triangular block matrix

AMS subject classifications. 15A18, 15A42

PII. S0895479801398517

1. Introduction.

Let X be a $p \times q$ matrix and let $k = \min\{p, q\}$.

Denote by $s(X) = (s_1(X), \ldots, s_k(X))$ the vector of decreasingly ordered singular values of X, i.e., $s_1(X) \geq \cdots \geq s_k(X)$ are the nonnegative square roots of the k largest eigenvalues of XX^*. For an $n \times n$ Hermitian matrix X let $\lambda(X) = (\lambda_1(X), \ldots, \lambda_n(X))$ denote the vector of decreasingly ordered eigenvalues. Given two real vectors $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$, we say that x is weakly majorized by y, denoted by $x \prec_w y$, if the sum of the m largest entries of x is not larger than that of y for $m = 1, \ldots, k$; for general background of the theory on majorization, see [7]. The algebra of $n \times n$ complex matrices will be denoted by M_n.

In this note, we prove inequalities on singular values for 2×2 block triangular matrices. Using the results, we answer Ando's questions on Bloomfield–Watson-type inequalities on eigenvalues and generalize the Kantorovich inequality and some results of Demmel.

2. Main theorem.

THEOREM 1. *Let* $A = \left(\begin{smallmatrix} R & 0 \\ S & T \end{smallmatrix}\right) \in M_n$ *be a block triangular matrix with singular values* $a_1 \geq \cdots \geq a_n$, *where* $R \in M_p$. *Let* $k = \min\{p, n-p\}$. *Then*

$$s(S) \prec_w (a_1 - a_n, \ldots, a_k - a_{n-k+1}). \tag{1}$$

If A *is invertible, then*

$$s(T^{-1}SR^{-1}) \prec_w \left(a_n^{-1} - a_1^{-1}, \ldots, a_{n-k+1}^{-1} - a_k^{-1}\right), \tag{2}$$

$$s(SR^{-1}) \prec_w \frac{1}{2}\left(\frac{a_1}{a_n} - \frac{a_n}{a_1}, \ldots, \frac{a_k}{a_{n-k+1}} - \frac{a_{n-k+1}}{a_k}\right), \tag{3}$$

and

$$s(T^{-1}S) \prec_w \frac{1}{2}\left(\frac{a_1}{a_n} - \frac{a_n}{a_1}, \ldots, \frac{a_k}{a_{n-k+1}} - \frac{a_{n-k+1}}{a_k}\right). \tag{4}$$

*Received by the editors September 24, 2001; accepted for publication by R. Bhatia November 26, 2001; published electronically July 1, 2002. This work was partially supported by NSF grants DMS-9704534 and DMS-0071994.

http://www.siam.org/journals/simax/24-1/39851.html

†Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187 (ckli@math.wm.edu, mathias@math.wm.edu).

Our proof of (1) relies on an elegant result of Thompson and Therianos [9]: Let

$$B = \begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix}$$

be an $n \times n$ Hermitian matrix with X being $q \times q$. Then for any indices $1 \le i_1 < \cdots < i_m \le q$ and $1 \le j_1 < \cdots < j_m \le n - q$

$$\sum_{l=1}^{m} \lambda_{i_l + j_l - l}(B) + \sum_{l=1}^{m} \lambda_{n-m+l}(B) \le \sum_{l=1}^{m} \lambda_{i_l}(X) + \sum_{l=1}^{m} \lambda_{j_l}(Z).$$

Proof. Let S have singular values $s_1 \ge \cdots \ge s_k$. Note that the matrix $\tilde{A} = \left(\begin{smallmatrix} 0 & A \\ A^* & 0 \end{smallmatrix}\right)$ has eigenvalues $a_1, \ldots, a_n, -a_n, \ldots, -a_1$ and is permutationally similar to

$$\begin{pmatrix} 0_p & S^* & R^* & 0 \\ S & 0_{n-p} & 0 & T \\ R & 0 & 0 & 0 \\ 0 & T^* & 0 & 0 \end{pmatrix} = \begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix},$$

where $Z = 0_n$ and $X \in M_n$ has the n eigenvalues $s_1, \ldots, s_k, 0, \ldots, 0, -s_k, \ldots, -s_1$. By the result of Thompson and Therianos, for any $m = 1, \ldots, k$, we have

$$\begin{aligned} -\sum_{j=1}^{m} s_j &= \sum_{j=1}^{m} \lambda_{n-m+j}(X) \\ &= \sum_{j=1}^{m} \lambda_{n-m+j}(X) + \sum_{j=1}^{m} \lambda_j(Z) \\ &\ge \sum_{j=1}^{m} \lambda_{n-m+j}(\tilde{A}) + \sum_{j=1}^{m} \lambda_{2n-j+1}(\tilde{A}) \\ &= \sum_{j=1}^{m} a_{n-m+j} - \sum_{j=1}^{m} a_j. \end{aligned}$$

Multiplying both sides by -1, we get (1).

If A is invertible, then

$$\hat{A} = (I_p \oplus -I_{n-p}) A^{-1} (I_p \oplus -I_{n-p}) = \begin{pmatrix} R^{-1} & 0 \\ T^{-1} S R^{-1} & T^{-1} \end{pmatrix}$$

has singular values $a_n^{-1}, \ldots, a_1^{-1}$. Applying the inequalities (1) to $\hat{A}$, we get (2).

Next, note that

$$A\hat{A} = \begin{pmatrix} I_p & 0 \\ 2SR^{-1} & I_{n-p} \end{pmatrix}.$$

Suppose U and V are unitary matrices such that $U^* S R^{-1} V$ has r_j as the (j, j) entry for $j = 1, \ldots, k$, and all other entries zero, where $s(SR^{-1}) = (r_1, \ldots, r_k)$. Then $A\hat{A}$ has the same singular values as the matrix

$$(V \oplus U)^* A\hat{A} (V \oplus U) = \begin{pmatrix} I_p & 0 \\ 2U^* S R^{-1} V & I_{n-p} \end{pmatrix},$$

which is permutationally similar to a direct sum of I_{n-2k} and 2×2 matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 2r_j & 1 \end{pmatrix}, \qquad j = 1, \ldots, k. \tag{5}$$

Matrices of the form (5) have singular values

$$r_j + \sqrt{r_j^2 + 1} \quad \text{and} \quad \left(r_j + \sqrt{r_j^2 + 1}\right)^{-1} = \sqrt{r_j^2 + 1} - r_j.$$

Thus,

$$s(A\hat{A}) = \left(r_1 + \sqrt{r_1^2 + 1}, \ldots, r_k + \sqrt{r_k^2 + 1}, \underbrace{1, \ldots, 1}_{n-2k}, \sqrt{r_k^2 + 1} - r_k, \ldots, \sqrt{r_1^2 + 1} - r_1 \right).$$

A well-known result of Alfred Horn (see [5], [6, Theorem 3.3.4], or [7, Chapter 9]) gives

$$\prod_{j=1}^{m} s_j(A\hat{A}) \le \prod_{j=1}^{m} s_j(A) s_j(\hat{A}) = \prod_{j=1}^{m} (a_j / a_{n-j+1}), \qquad m = 1, \ldots, k,$$

i.e.,

$$\left(\ln s_1(A\hat{A}), \ldots, \ln s_k(A\hat{A})\right) \prec_w \left(\ln(a_1/a_n), \ldots, \ln(a_k/a_{n-k+1})\right).$$

Consider the function $f(t) = e^t - e^{-t}$ for $t > 0$. Then $f(\ln(r_j + \sqrt{r_j^2 + 1})) = 2r_j$ for $j = 1, \ldots, k$. Since f is increasing and convex on $(0, \infty)$ it preserves weak majorization [7, Chapters 3, A.8, and C.1], and so we have

$$\begin{aligned} 2(r_1, \ldots, r_k) &= \left(f(\ln s_1(A\hat{A})), \ldots, f(\ln s_k(A\hat{A}))\right) \\ &\prec_w \left(f(\ln(a_1/a_n)), \ldots, f(\ln(a_k/a_{n-k+1}))\right) \\ &= \left(\frac{a_1}{a_n} - \frac{a_n}{a_1}, \ldots, \frac{a_k}{a_{n-k+1}} - \frac{a_{n-k+1}}{a_k}\right), \end{aligned}$$

which is (3). Applying a similar argument to $\hat{A}A$, we get (4). □

3. Questions of Ando. In [1], Ando raised several problems in connection with Bloomfield–Watson-type inequalities for eigenvalues that arise in statistics (see also [4, Problem 7.3]). The following theorem answers his questions in the affirmative and extends scalar inequalities of Demmel [3, equations (62), (63), (65), (66)] to majorizations.

THEOREM 2. *Let* $A = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$ *be positive definite such that* $A_{11} \in M_p$. *Suppose* $k = \min\{p, n-p\}$ *and* A *has eigenvalues* $\lambda_1 \ge \cdots \ge \lambda_n$. *Then*

$$s(A_{22}^{-1/2} A_{21}) \prec_w \left(\sqrt{\lambda_1} - \sqrt{\lambda_n}, \ldots, \sqrt{\lambda_k} - \sqrt{\lambda_{n-k+1}}\right) \tag{6}$$

and

$$s(A_{21} A_{11}^{-1}) \prec_w \frac{1}{2} \left(\sqrt{\frac{\lambda_1}{\lambda_n}} - \sqrt{\frac{\lambda_n}{\lambda_1}}, \ldots, \sqrt{\frac{\lambda_k}{\lambda_{n-k+1}}} - \sqrt{\frac{\lambda_{n-k+1}}{\lambda_k}} \right). \tag{7}$$

Proof. Let $a_j = \sqrt{\lambda_j}$ for $j = 1, \ldots, n$. Let

$$B = \begin{pmatrix} R & 0 \\ S & T \end{pmatrix}$$

with $T = A_{22}^{1/2}$, $S = A_{22}^{-1/2}A_{21}$, and $R = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{1/2}$. Then $A = B^*B$, and B has singular values $a_1, \ldots, a_n$. Applying Theorem 1 to the block triangular matrix B, we see that (6) is just (1).

Next, let

$$C = \begin{pmatrix} R & 0 \\ S & T \end{pmatrix}$$

with $R = A_{11}^{1/2}$, $S = A_{21}A_{11}^{-1/2}$, and $T = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{1/2}$. Then $A = CC^*$, and C has singular values $a_1, \ldots, a_n$. Applying Theorem 1 to the block triangular matrix C, we see that (7) is just (3). □

Suppose P is an $n \times k$ matrix such that $P^*P = I_k$. Then there exists a unitary U such that P is the first k columns of U. For any positive definite matrix A, we can apply Theorem 2 to the block matrix $U^*AU = \left(\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right)$. The results will take the more general form involving the eigenvalues and singular values of the matrices $P^*AP, P^*A^{-1}P, P^*A^2P$, etc. Many results in [1], [4] are stated in these forms, and they can be deduced from our results. We give a few examples in the following discussion. For easy reference and comparison, we state the next corollary in this manner.

COROLLARY 3. *Let A be a positive definite matrix with $\lambda_1 \geq \cdots \geq \lambda_n$. For any $n \times k$ matrix P such that $P^*P = I_k$, where $2k \leq n$, we have*

$$(8)\quad s(P^*AP - (P^*A^{-1}P)^{-1}) \prec_w \left(\left(\sqrt{\lambda_1} - \sqrt{\lambda_n}\right)^2, \ldots, \left(\sqrt{\lambda_k} - \sqrt{\lambda_{n-k+1}}\right)^2\right)$$

and

$$(9)\quad s\left((P^*AP)^{-1}(P^*A^2P)(P^*AP)^{-1}\right) \prec_w \left(\frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}, \ldots, \frac{(\lambda_k + \lambda_{n-k+1})^2}{4\lambda_k\lambda_{n-k+1}}\right).$$

Proof. Let $a_1 \geq \cdots \geq a_n > 0$ so that $a_j^2 = \lambda_j$ for $j = 1, \ldots, n$.

To prove (8), we may assume that P is the first k columns of a unitary matrix U, and

$$(10)\qquad U^*AU = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Then

$$\begin{aligned} P^*AP - (P^*A^{-1}P)^{-1} &= A_{11} - (A_{11} - A_{21}^*A_{22}^{-1}A_{21}) \\ &= A_{21}^*A_{22}^{-1}A_{21} = (A_{22}^{-1/2}A_{21})^*(A_{22}^{-1/2}A_{21}). \end{aligned}$$

By the majorization (6) and the fact that squaring preserves majorization (see [7, Chapters A.8 and C.1]), we have

$$\sum_{j=1}^{m} s_j(P^*AP - (P^*A^{-1}P)^{-1}) = \sum_{j=1}^{m} s_j^2(A_{22}^{-1/2}A_{21}) \leq \sum_{j=1}^{m}(a_j - a_{n-j+1})^2, \quad m = 1, \ldots, k.$$

Thus, (8) holds.

To prove (9), we may again assume that P is the first k columns of a unitary matrix U such that (10) holds. Then the left side of (9) is just

$$s(I_k + A_{11}^{-1}A_{21}A_{12}A_{11}^{-1}) = (1,\dots,1) + s(A_{21}A_{11}^{-2}A_{12}).$$

Using the square of (7), we have

$$(1,\dots,1)+s\left(A_{21}A_{11}^{-2}A_{12}\right) \prec_w (1,\dots,1)+\frac{1}{4}\left(\left(\frac{a_1}{a_n}-\frac{a_n}{a_1}\right)^2,\dots,\left(\frac{a_k}{a_{n-k+1}}-\frac{a_{n-k+1}}{a_k}\right)^2\right),$$

which is the right side of (9). □

We proved (8) and (9) by squaring (6) and (7). Ando proved (9) by another method in [1]. One may wonder whether it is possible to deduce (6) and (7) from (8) and (9) by taking square roots. It is not possible, since taking square roots does not preserve majorization.

Our bound (8) includes the inequality of Rao [8]:

$$\operatorname{tr}(P^*AP - (P^*A^{-1}P)^{-1}) \le \sum_{j=1}^{k}\left(\sqrt{\lambda_j} - \sqrt{\lambda_{n-j+1}}\right)^2.$$

The inequality (9) includes the Kantorovich inequality. To see this, given a unit vector x, take $k = 1$ and take $P = A^{-1/2}x/(x^*A^{-1}x)^{1/2}$. Then we have the Kantorovich inequality:

$$(x^*Ax)(x^*A^{-1}x) = s_1((P^*AP)^{-1}(P^*A^2P)(P^*AP)^{-1}) \le \frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n}.$$

In [1], Ando also asked whether the following is true for a positive definite matrix:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$, where $A_{11} \in M_{n-k}$ with $2k \le n$:

$$\sum_{j=1}^{m} s_j(A_{22}^{-1/2}A_{21}A_{11}^{-1/2}) \le \sum_{j=1}^{m}\frac{\lambda_j - \lambda_{n-j+1}}{\lambda_j + \lambda_{n-j+1}}, \qquad m = 1,\dots,k.$$

The result is indeed true for $m = 1$ [1], [3, Theorem 1], but not in general, as the following example shows.

EXAMPLE 4. *Let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad \textit{with} \quad A_{11} = A_{22} = \begin{pmatrix} 6 & 0 \\ 0 & 3 \end{pmatrix} \quad \textit{and} \quad A_{12} = A_{21} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then A has eigenvalues $8, 4, 4, 2$ and $A_{22}^{-1/2}A_{21}A_{11}^{-1/2}$ has singular values $1/3, 1/3$. However,

$$1/3 + 1/3 \not\le 3/5 = (8-2)/10 + (4-4)/8.$$

Note added in proof. In [2] (the final form of [1]), Professor Ando did not explicitly mentioned the problems stated in the earlier version [1]. He obtained some other results related to ours.

REFERENCES

[1] T. Ando, *Bloomfield-Watson type inequalities for eigenvalues*, notes of the lecture presented at the International Conference on Mathematical Analysis and Its Applications, National Sun Yat-Sen University, Kaoshiung, Taiwan, 2000.

[2] T. Ando, *Bloomfield-Watson type inequalities for eigenvalues*, Taiwanese J. of Math., 5 (2001), pp. 443–370.

[3] J. Demmel, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1983), pp. 599–610.

[4] S.W. Drury, S. Liu, C.-Y. Lu, S. Puntanen, and G.P.H. Styan, *Some Comments on Several Matrix Inequalities with Applications to Canonical Correlations: Historical Background and Recent Developments*, Report A332, Department of Mathematics, Statistics and Philosophy, University of Tampere, Tampere, Finland; Sankhya, to appear.

[5] A. Horn, *On the singular values of a product of completely continuous operators*, Proc. Nat. Acad. Sci. USA, 36 (1950), pp. 374–375.

[6] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[7] A.W. Marshall and I. Olkin, *Inequalities: The Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[8] R.C. Rao, *The inefficiency of least squares; extensions of the Kantorovich inequality*, Linear Algebra Appl., 70 (1985), pp. 249–255.

[9] R.C. Thompson and S. Therianos, *Inequalities connecting the eigenvalues of a Hermitian matrices with the eigenvalues of complementary principal submatrices*, Bull. Austral. Math. Soc., 6 (1972), pp. 117–132.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 132–149

ON POSITIVE SEMIDEFINITE MATRICES WITH KNOWN NULL SPACE*

PETER ARBENZ† AND ZLATKO DRMAČ‡

Abstract. We show how the zero structure of a basis of the null space of a positive semidefinite matrix can be exploited to determine a positive definite submatrix of maximal rank. We discuss consequences of this result for the solution of (constrained) linear systems and eigenvalue problems. The results are of particular interest if A and the null space basis are sparse. We furthermore execute a backward error analysis of the Cholesky factorization of positive semidefinite matrices and provide new elementwise bounds.

Key words. positive semidefinite matrices, Cholesky factorization, null space basis

AMS subject classifications. 65F05, 65F50

PII. S0895479800381331

1. Introduction.

The Cholesky factorization $A = R^T R$, R upper triangular, exists for any symmetric positive semidefinite matrix A. In fact, R is the upper triangular factor of the QR factorization of $A^{1/2}$ [14, section 10.3]. R can be computed with the well-known algorithm for positive definite matrices. However, for semidefinite matrices zero pivots may appear. When zero pivots appear, one can choose the corresponding row of R to be zero. We will do so in this paper in order to specify a unique factorization. To actually compute a numerically stable Cholesky factorization of a positive semidefinite matrix, one is advised to apply diagonal pivoting [14].

A semidefinite matrix A may be given implicitly, in factored form $A = F^T F$, where $F \in \mathbb{R}^{p\times n}$ is of full row rank $r = \operatorname{rank}(A)$. F, which does not need to be a Cholesky factor, exposes the singularity of A explicitly as $\mathcal{N}(A) = \mathcal{N}(F)$. In this case both the linear system and the eigenvalue problem can be solved efficiently and elegantly by working directly on the matrix F, never forming the matrix A explicitly. In fact, in some applications, not assembling the matrix A but its factor F instead is the most important step in the overall process of the numerical computation. One obvious reason is that the (spectral) condition number of F is the square root of the condition number of A. In finite element computation, F is the so-called natural factor of the stiffness matrix A [2]. In the framework of linear algebra, every symmetric positive semidefinite matrix is the Gram matrix of some set of vectors, the columns of F.

Another possibility to have the singularity of A explicit is to have available a basis of its null space $\mathcal{N}(A)$. This is the situation that we want to investigate in this paper. We will see that knowing a basis of $\mathcal{N}(A)$ allows us to determine a priori when the zero pivots will occur in the Cholesky factorization. It also permits us to give a positive definite submatrix of A right away. These results are of particular interest if A and the null space basis are sparse. This is the case in the application from

*Received by the editors November 17, 2000; accepted for publication (in revised form) by N. J. Higham January 14, 2002; published electronically July 1, 2002.
http://www.siam.org/journals/simax/24-1/38133.html

†Institute of Scientific Computing, Swiss Federal Institute of Technology (ETH), CH-8092 Zurich, Switzerland (arbenz@inf.ethz.ch).

‡Department of Mathematics, University of Zagreb, Bijenička 30, HR-10000 Zagreb, Croatia (drmac@math.hr). The work of this author was supported by Croatian Ministry of Science and Technology grant 037012.

electromagnetics that prompted this study [1]. There, a vector that is orthogonal to the null space corresponds to a discrete electric field that is divergence-free.

Our findings permit us to work with the positive definite part of A and to compute a rank-revealing Cholesky-like factorization $A = R^T R$, where the upper trapezoidal R has full row rank. What is straightforward in exact arithmetic amounts to simply *replacing by zero* potentially inaccurate small numbers. We analyze the error that is introduced by this procedure.

We complement this note with some implications of the above for solving eigenvalue problems and constrained systems of equations.

2. Cholesky factorization of a positive semidefinite matrix with known null space. In this section we consider joint structures of a semidefinite matrix A and its null space.

THEOREM 2.1. *Let $A = R^T R$ be the Cholesky factorization of the positive semidefinite matrix $A \in \mathbb{R}^{n\times n}$. Let $Y \in \mathbb{R}^{n\times m}$ with $\mathcal{R}(Y) = \mathcal{N}(A)$ and, for $i = 1, \ldots, m$, set $n_i := \max\{k \mid y_{ki} \neq 0\}$. If $n_1 < n_2 < \cdots < n_m$, then $r_{n_i n_i} = 0$, $i = 1, \ldots, m$. These are the only zero entries on the diagonal of R.*

Proof. Notice that the assumptions imply that $Y := [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ has full rank. By Sylvester's law of inertia R has precisely m zeros on its diagonal. Further,

$$(R\mathbf{y}_i)_{n_i} = r_{n_i n_i} y_{n_i i} = 0,$$

whence $r_{n_i n_i} = 0$ as $y_{n_i i} \neq 0$. ∎

If only $n_1 \leq n_2 \leq \cdots \leq n_m$, Y, flipped upside-down, can be transformed into column-echelon form in order to obtain strong inequalities.

The Cholesky factor R appearing in Theorem 2.1 is an $n \times n$ upper triangular matrix with m zero rows. These rows do not affect the product $R^T R$. Therefore, they can be removed from R to yield an $(n-m) \times n$ matrix $\widehat{R}$ with $\widehat{R}^T \widehat{R} = A$.

If the numbers n_i are known, it is convenient to permute the rows of Y and accordingly the rows and columns n_i of A to the end. Then Theorem 2.1 can be applied with $n_i = n - m + i$. The last m rows of R in Theorem 2.1 vanish. So, $\widehat{R}$ is upper trapezoidal.

After the above-mentioned permutation, the lowest $m \times m$ block of Y is nonsingular, in fact, upper triangular. This consideration leads to an alternative formulation of Theorem 2.1.

THEOREM 2.2. *Let $A = R^T R$ be the Cholesky factorization of the positive semidefinite matrix $A \in \mathbb{R}^{n\times n}$. Let $Y \in \mathbb{R}^{n\times m}$ with $\mathcal{R}(Y) = \mathcal{N}(A)$. If the last m rows of Y are linearly independent, then the leading principal $(n-m) \times (n-m)$ submatrix of A is positive definite and $R = \left[\begin{smallmatrix} \widehat{R} \\ O \end{smallmatrix}\right]$, where $\widehat{R}$ is $(n-m) \times n$ upper trapezoidal.*

Proof. Let

$$W := \begin{bmatrix} I_{n-m} & Y_1 \\ O & Y_2 \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \qquad Y_2 \in \mathbb{R}^{m\times m}. \tag{2.1}$$

Y_2 consists of the last m rows of Y. W is therefore invertible. Applying a congruence transformation with W on A gives

$$W^T A W = \begin{bmatrix} I_{n-m} & O \\ Y_1^T & Y_2^T \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I_{n-m} & Y_1 \\ O & Y_2 \end{bmatrix} = \begin{bmatrix} A_{11} & O \\ O & O \end{bmatrix}. \tag{2.2}$$

By Sylvester's law of inertia A_{11} must be positive definite.

Let $A_{11} = R_{11}^T R_{11}$ be the Cholesky factorization of A_{11}. Then the Cholesky factor of A is given by

$$\begin{bmatrix} R_{11} & O \\ O & O \end{bmatrix} W^{-1} = \begin{bmatrix} R_{11} & O \\ O & O \end{bmatrix} \begin{bmatrix} I_{n-m} & -Y_2^{-1}Y_1 \\ O & Y_2^{-1} \end{bmatrix} = \begin{bmatrix} R_{11} & -R_{11}Y_2^{-1}Y_1 \\ O & O \end{bmatrix}.$$

Thus we can write $A = \widehat{R}^T \widehat{R}$ with $\widehat{R} = [R_{11}, -R_{11}Y_1Y_2^{-1}]$. ∎

Theorem 2.2 is applicable as long as the last m rows of Y form an invertible matrix. If rows $i_1, \ldots, i_m$ of Y are linearly independent, we can permute Y such that these rows become the last ones. In particular, if we want A_{11} to be as sparse as possible, we may choose $i_1, i_2, \ldots$ to be the m most densely populated rows/columns of A with the following greedy algorithm: If we have determined $i_1, \ldots, i_k$, we choose i_{k+1} to be the index of the densest column of A such that rows $i_1, \ldots, i_{k+1}$ of Y are linearly independent. In this way we can hope for an A_{11} with sparse Cholesky factors. Notice that we have used the structure of Y to get sparse factors of A. We do not know how to exploit Y's structure to enhance the condition of A_{11}.

Remark 2.1. The equation

$$-\Delta u(\mathbf{x}) = 0 \text{ in } \Omega \subset \mathbb{R}^n, \qquad \partial_n u(\mathbf{x}) = 0 \text{ on } \partial\Omega, \tag{2.3}$$

in a simply connected domain Ω is satisfied by all constant functions u. The discretization of (2.3) with finite elements of Lagrange type [5] leads to a positive semidefinite matrix A with a one-dimensional null space spanned by the vector $\mathbf{e}$ with all entries equal to 1. Theorem 2.1 now implies that no matter how we permute A, in the Cholesky factorization the single zero on the diagonal of R will not appear before the very last elimination step.

Example 2.1. Let A and Y be given by

$$A := \begin{bmatrix} 1 & 0 & 1 & 1 & 3 \\ 0 & 9 & 3 & 9 & 9 \\ 1 & 3 & 3 & 6 & 8 \\ 1 & 9 & 6 & 14 & 16 \\ 3 & 9 & 8 & 16 & 22 \end{bmatrix}, \qquad Y := \begin{bmatrix} 2 & 3 \\ 0 & 1 \\ 0 & 6 \\ 1 & 0 \\ -1 & -3 \end{bmatrix}.$$

Then $AY = O$. As the last two rows of Y are linearly independent, Theorem 2.2 states that the principal 3×3 submatrix of A is positive definite and that its Cholesky factor is 3×5 upper triangular. In fact,

$$R = \begin{bmatrix} 1 & 0 & 1 & 1 & 3 \\ 0 & 3 & 1 & 3 & 3 \\ 0 & 0 & 1 & 2 & 2 \end{bmatrix}.$$

Let P be the permutation matrix, which exchanges the 2nd entry with the 4th and the 3rd entry with the 5th of a 5-vector. Then

$$A_1 := PAP^T = \begin{bmatrix} 1 & 1 & 3 & 0 & 1 \\ 1 & 14 & 16 & 9 & 6 \\ 3 & 16 & 22 & 9 & 8 \\ 0 & 9 & 9 & 9 & 3 \\ 1 & 6 & 8 & 3 & 3 \end{bmatrix}, \qquad Y_1 := PY = \begin{bmatrix} 2 & 3 \\ 1 & 0 \\ -1 & -3 \\ 0 & 1 \\ 0 & 6 \end{bmatrix}.$$

Now we have $n_1 = 3 < n_2 = 5$, and according to Theorem 2.1 the Cholesky factor R_1 of A_1 has zero diagonal elements at positions 3 and 5. Indeed,

$$R_1 = \frac{1}{\sqrt{13}} \begin{bmatrix} \sqrt{13} & \sqrt{13} & 3\sqrt{13} & 0 & \sqrt{13} \\ 0 & 13 & 13 & 9 & 5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3. Consistent semidefinite systems. In this section we discuss how to solve

$$A\mathbf{x} = R^T R\mathbf{x} = \mathbf{b} \in \mathcal{R}(A), \tag{3.1}$$

where A, R, and Y are as in Theorem 2.1. Without loss of generality, we can assume that $n_i = r + i$, $r := n - m$. We split matrices and vectors in (3.1),

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{bmatrix} R_{11}^T \\ R_{12}^T \end{bmatrix} [R_{11},\ R_{12}] \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \tag{3.2}$$

with $\mathbf{x}_1, \mathbf{b}_1 \in \mathbb{R}^r$ and $\mathbf{x}_2, \mathbf{b}_2 \in \mathbb{R}^m$. So, A_{11} is obtained from A by deleting rows and columns n_i, $i = 1, \ldots, m$. The factorization (3.2) yields

$$A_{11} = R_{11}^T R_{11}, \qquad A_{12} = R_{11}^T R_{12}. \tag{3.3}$$

Although A_{11} is invertible, its condition number can be arbitrarily high. To reduce fill-in during factorization [10] any symmetric permutations can be applied to A_{11} without affecting what follows. As R^T has full rank, $AY = O$ implies $RY = O$ or

$$R_{11}Y_1 + R_{12}Y_2 = O. \tag{3.4}$$

Since $n_i = r + i$, the $m \times m$ matrix Y_2 is upper triangular with nonzero diagonal elements. Because $\mathcal{R}(A) = \mathcal{N}(Y)^\perp$ the right side $\mathbf{b}$ of (3.1) has to satisfy

$$Y_1^T \mathbf{b}_1 + Y_2^T \mathbf{b}_2 = \mathbf{0}. \tag{3.5}$$

It is now easy to show that a particular solution of (3.1) is given by $\mathbf{x}$ with components

$$\mathbf{x}_1 = A_{11}^{-1}\mathbf{b}_1 = R_{11}^{-1}R_{11}^{-T}\mathbf{b}_1, \qquad \mathbf{x}_2 = \mathbf{0}.$$

In fact, employing (3.3)–(3.5) the second block row in (3.2) is

$$A_{12}^T\mathbf{x}_1 - \mathbf{b}_2 = R_{12}^T R_{11}\mathbf{x}_1 + Y_2^{-T}Y_1^T\mathbf{b}_1 = R_{12}^T R_{11}^{-T}(A_{11}\mathbf{x}_1 - \mathbf{b}_1) = \mathbf{0}.$$

The manifold $\mathcal{S}$ of the solutions of (3.1)–(3.2) is

$$\mathcal{S} = \left\{ \mathbf{x} = \begin{pmatrix} A_{11}^{-1}\mathbf{b}_1 \\ \mathbf{0} \end{pmatrix} + Y\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^m \right\}.$$

The vector $\mathbf{a}$ can be determined such that the solution $\mathbf{x}$ satisfies some constraints $C^T\mathbf{x} = \mathbf{0}$ with $C \in \mathbb{R}^{n\times m}$ provided that C^TY is invertible. In particular, if $C = Y$, then $\mathbf{x}$ is perpendicular to the null space of A.

Now let A be given implicitly as a Gram matrix $A = F^T F$ with $F \in \mathbb{R}^{p\times n}$, $p \geq n$, and let $Y \in \mathbb{R}^{n\times m}$ be as above. (This may require renumbering the columns of F.) As

$$FY = F_1 Y_1 + F_2 Y_2 = O,$$

and as Y_2 is nonsingular, the block F_2 depends linearly on F_1. Therefore, the QR factorization of F has the form

$$F = [F_1,\ F_2] = [Q_1,\ Q_2] \begin{bmatrix} R_{11} & R_{12} \\ O & O \end{bmatrix} = Q_1 [R_{11},\ R_{12}].$$

Since $A = F^T F = \widehat{R}\widehat{R}^T$, the factor $\widehat{R} = [R_{11},\ R_{12}]$ equals the upper trapezoidal Cholesky-like factor in (3.2).

4. Error analysis. In this section we give backward error analyses for the semidefinite Cholesky factorization and for the null space basis.

4.1. Semidefinite Cholesky factorization. The floating-point computation of the Cholesky factorization of a semidefinite matrix is classified by Higham [14, section 10.3.2] as unstable if no pivoting is used. However, with complete pivoting, backward stability in the normwise sense can be established [14, Theorem 10.14]. In this section we will establish elementwise backward error bounds for the factorization, where the leading positive definite submatrix is determined as explained in Theorem 2.2.

If we assume, as we do in this note, that a basis of the null space of the matrix under consideration is known a priori, then, of course, its rank is known. Let A be partitioned as in (3.2). We assume that $A_{11} \in \mathbb{R}^{r\times r}$ is positive definite numerically, i.e., that the Cholesky factorization does not break down in floating-point arithmetic with round-off unit $\mathbf{u}$. Due to a result by Demmel [6] (see also [14, Theorem 10.14]) this is the case if

$$\lambda_{\min}((A_{11})_s) \equiv \|(A_{11})_s^{-1}\|^{-1} > 2rf(r)\mathbf{u}, \qquad f(r) = \frac{r+1}{1-2(r+1)\mathbf{u}}, \tag{4.1}$$

where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue, $\|\cdot\|$ is the spectral norm, and

$$(A_{11})_s = \mathrm{diag}(A_{11})^{-1/2} A_{11} \mathrm{diag}(A_{11})^{-1/2}.$$

If (4.1) does not hold, A_{11} is not numerically definite. Note that $(A_{11})_s$ is symmetric positive definite with unit diagonal. The assumption on $\lambda_{\min}((A_{11})_s)$ can be relaxed if, for instance, we use double precision accumulation during the factorization. Then $f(r)$ can be replaced by a small integer for all r not larger than $1/\mathbf{u}$. We assume, however, that $2rf(r)\mathbf{u} < 1$.

The Cholesky decomposition of A is computed as indicated in (3.3). The Cholesky factor of A_{11} is computed first. Then the matrix R_{12} is obtained as the solution of the matrix equation $R_{11}^T X = A_{12}$.

Let $\widetilde{R}_{11}$ denote the computed floating-point Cholesky factor of A_{11}. Then the following two important facts are well known.

(1) There exists a symmetric δA_{11} such that $A_{11} + \delta A_{11} = \widetilde{R}_{11}^T \widetilde{R}_{11}$ and

$$\max_{1\leq i,j\leq r} \frac{|(\delta A_{11})_{ij}|}{\sqrt{(A_{11})_{ii}(A_{11})_{jj}}} \leq f(r)\mathbf{u}. \tag{4.2}$$

This is the backward error bound by Demmel [6], [14, Theorem 10.5].

(2) Let $(\delta A_{11})_s := \mathrm{diag}(A_{11})^{-1/2}\delta A_{11}\mathrm{diag}(A_{11})^{-1/2}$. Equations (4.1) and (4.2) imply that the Frobenius norm of $(\delta A_{11})_s$ satisfies $\|(\delta A_{11})_s\|_F \le rf(r)\mathbf{u} < \lambda_{\min}((A_{11})_s)$. Since assumption (4.1) implies $2\|(\delta A_{11})_s\|_F < \lambda_{\min}((A_{11})_s)$, one can show [8] that there exists an upper triangular matrix Γ such that

$$\widetilde{R}_{11} = (I+\Gamma)R_{11}, \qquad \|\Gamma\|_F \le \frac{\sqrt{2}\|(A_{11})_s^{-1}\|\|(\delta A_{11})_s\|_F}{1+\sqrt{1-2\|(A_{11})_s^{-1}\|\|(\delta A_{11})_s\|_F}} < \frac{1}{\sqrt{2}}.$$

Let $\widetilde{R}_{12}$ be the floating-point solution of the matrix equation $\widetilde{R}_{11}^T X = A_{12}$. Then $\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}]$ is the computed approximation of the exact Cholesky factor $R = [R_{11}, R_{12}]$. Let $\widetilde{A} = A + \delta A = \widetilde{R}^T\widetilde{R}$ be partitioned conforming with (3.2). Since $A+\delta A$ is positive semidefinite and of rank r by construction, the equation $\widetilde{A}_{22} = \widetilde{A}_{12}^T\widetilde{A}_{11}^{-1}\widetilde{A}_{12}$ holds.

If we compute $\widetilde{R}_{12}$ column by column, then, using Wilkinson's analysis of triangular linear systems [14, Theorem 8.5],

$$|\widetilde{R}_{11}^T\widetilde{R}_{12} - A_{12}| \le t(r)\mathbf{u}|\widetilde{R}_{11}|^T|\widetilde{R}_{12}|, \qquad t(r) = \frac{r}{1-r\mathbf{u}},$$

where the matrix absolute values and the inequality are to be understood entrywise. Thus, we can write $\widetilde{R}_{12}$ as

$$\widetilde{R}_{12} = \widetilde{R}_{11}^{-T}(A_{12}+\delta A_{12}), \qquad |\delta A_{12}| \le t(r)\mathbf{u}|\widetilde{R}_{11}|^T|\widetilde{R}_{12}|. \tag{4.3}$$

Also, if we define $\Psi = (I+\Gamma)^{-T} - I$, $\Omega = t(r)\mathbf{u}|\widetilde{R}_{11}^{-T}||\widetilde{R}_{11}^T|$, we have

$$\widetilde{R}_{12} = (I+\Psi)R_{12} + \widetilde{R}_{11}^{-T}\delta A_{12}, \qquad |\widetilde{R}_{11}^{-T}\delta A_{12}| \le \Omega|\widetilde{R}_{12}|. \tag{4.4}$$

Further, from the inequality $|\widetilde{R}_{12}| \le (I+|\Psi|)|R_{12}| + \Omega|\widetilde{R}_{12}|$ and using the M-matrix property of $I-\Omega$ we obtain

$$|\widetilde{R}_{12}| \le (I-\Omega)^{-1}(I+|\Psi|)|R_{12}|. \tag{4.5}$$

Hence, relations (4.2), (4.3), (4.5) imply that the backward error for all (i,j) in the $(1,2)$ block in (3.2) is bounded by

$$\begin{aligned}|(\delta A_{12})_{ij}| &\le t(r)\mathbf{u}\|\widetilde{R}_{11}\mathbf{e}_i\|\|\widetilde{R}_{12}\mathbf{e}_{j'}\| \le t(r)\mathbf{u}\sqrt{(\widetilde{A}_{11})_{ii}(\widetilde{A}_{22})_{j'j'}}, \qquad j' = j-r,\\ &\le t(r)\mathbf{u}(1+f(r)\mathbf{u})\frac{1+\||\Psi|\|}{1-\|\Omega\|}\sqrt{(A_{11})_{ii}(A_{22})_{j'j'}}.\end{aligned}$$

We first observe that $\||\Psi|\| \le \sqrt{r}\|\Gamma\|/(1-\|\Gamma\|)$ and that $\|\Omega\| \le rt(r)\mathbf{u}\sqrt{\|(\widetilde{A}_{11})_s^{-1}\|}$. Note that our assumptions imply that

$$\|\Omega\| \le \frac{1}{\sqrt{2}}\frac{\sqrt{r}}{r+1} < 1/2, \qquad \frac{\|\Gamma\|}{1-\|\Gamma\|} < 1+\sqrt{2}.$$

It remains to estimate the backward error in the $(2,2)$ block of the partition (3.2). Using relation (4.4), we compute $\delta A_{22} = \widetilde{R}_{12}^T\widetilde{R}_{12} - R_{12}^TR_{12}$ as follows:

$$\begin{aligned}\delta A_{22} = {} & R_{12}^T(\Psi^T+\Psi+\Psi^T\Psi)R_{12} + R_{12}^T(I+\Psi^T)\widetilde{R}_{11}^{-T}\delta A_{12}\\ & + \delta A_{12}^T\widetilde{R}_{11}^{-1}(I+\Psi)R_{12} + \delta A_{12}^T\widetilde{R}_{11}^{-1}\widetilde{R}_{11}^{-T}\delta A_{12}.\end{aligned}$$

Using the inequalities from relations (4.4), (4.5) we obtain, for all (i,j),

$$|(\delta A_{22})_{ij}| \le \sqrt{(A_{22})_{ii}(A_{22})_{jj}}\left(2\psi + 2\omega\frac{1+\psi'}{1-\omega} + \psi^2 + \omega^2\frac{(1+\psi)^2}{(1-\omega)^2}\right),$$

where $\omega = \|\Omega\|$, $\psi = \|\Psi\|$, $1+\psi' = (1+\psi)(1+\|\,|\Psi|\,\|)$.

We summarize the above analysis in the following.

THEOREM 4.1. *Let A be an $n\times n$ positive semidefinite matrix of rank r with block partition* (3.2)*, where the $r\times r$ matrix A_{11} is positive definite with the property* (4.1)*. Then the floating-point Cholesky factorization with round-off* $\mathbf{u}$ *will compute an upper trapezoidal matrix $\widetilde{R}$ of rank r such that $\widetilde{R}^T\widetilde{R} = A + \delta A$, where δA is a symmetric backward perturbation with the following bounds:*

$$\begin{aligned}
&|\delta a_{ij}| \le f(r)\mathbf{u}\sqrt{a_{ii}a_{jj}}, \quad 1 \le i,j \le r,\\
&|\delta a_{ij}| \le \left\{2t(r)(1+(1+\sqrt{2})\sqrt{r})(1+f(r)\mathbf{u})\right\}\mathbf{u}\sqrt{a_{ii}a_{jj}}, \quad 1 \le i \le r < j \le n,\\
&|\delta a_{ij}| \le \left\{2rt(r)\sqrt{\tilde\kappa} + \sqrt{8}rf(r)\kappa + \mathcal{O}(\mathbf{u})\right\}\mathbf{u}\sqrt{a_{ii}a_{jj}}, \quad r < i,j \le n.
\end{aligned}$$

In the last estimate, $\kappa = \|(A_{11})_s^{-1}\|$, $\tilde\kappa = \|(\widetilde{A}_{11})_s^{-1}\|$. Further, if $\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}]$ and if $R = [R_{11}, R_{12}]$ is the exact Cholesky factor of A, then

$$\begin{aligned}
&\widetilde{R}_{11} - R_{11} = \Gamma R_{11}, \qquad \|\Gamma\| \le \sqrt{2}rf(r)\kappa\mathbf{u},\\
&|\widetilde{R}_{12} - R_{12}| \le \Xi|R_{12}|, \qquad \|\Xi\| \le rt(r)\sqrt{\tilde\kappa}\mathbf{u} + \sqrt{2}rf(r)\kappa\mathbf{u} + \mathcal{O}(\mathbf{u}^2).
\end{aligned}$$

Here, the matrix Γ is upper triangular and Ξ is to the first order $|\Psi| + \Omega$.

Further, let the Cholesky factorization of A_{11} be computed with pivoting so that $(R_{11})_{ii} \ge \sum_{k=i}^{j}(R_{11})_{kj}^2$, $1 \le i \le j \le r$. Then the error $\delta R_{11} = \widetilde{R}_{11} - R_{11}$ is also rowwise small, that is,

$$\|\mathbf{e}_i^T\delta R_{11}\| \le \|\mathbf{e}_i^T\Gamma\|\sqrt{r-i+1}(R_{11})_{ii}, \quad i = 1,\dots,r. \tag{4.6}$$

Example 4.1. In this example we indicate that the error bound for the elements in the $(2,2)$ block A_{22} given in Theorem 4.1 correctly reflect reality. We again assume that $r = \mathrm{rank}(A)$ such that $A_{22} = A_{12}^TA_{11}^{-1}A_{12}$. In Theorem 4.1 the backward error in the $(2,2)$ block depends on the condition number of the scaled $(1,1)$ block. This is the consequence of enforcing the rank r by using the information on the null space or on any other criterion of threshold type; see, e.g., Higham [14]. In fact we think that this dependence is natural and that it cannot be removed.

Let

$$A = \begin{pmatrix} 1+\varepsilon & 1 & \varepsilon^2\\ 1 & 1 & 0\\ \varepsilon^2 & 0 & \varepsilon^3\end{pmatrix}, \qquad \mathrm{rank}(A) = 2, \quad \mathcal{N}(A) = \mathrm{span}\left(\begin{pmatrix}\varepsilon\\ -\varepsilon\\ -1\end{pmatrix}\right), \tag{4.7}$$

where $\varepsilon > 0$ is small, e.g., $\varepsilon = 10\cdot\mathbf{u}$. Here,

$$A_{11} = \begin{pmatrix}1+\varepsilon & 1\\ 1 & 1\end{pmatrix}, \qquad A_{11}^{-1} = \frac{1}{\varepsilon}\begin{pmatrix}1 & -1\\ -1 & 1+\varepsilon\end{pmatrix}, \qquad A_{22} = A_{12}^TA_{11}^{-1}A_{12} = \varepsilon^3.$$

Also, note that A_{11} is almost scaled, $\kappa_2(A_{11}) \approx 4/\varepsilon$, and $\|A_{11}^{-1}A_{12}\|_2 = \sqrt{2}\varepsilon$. Notice that A is prepivoted according to the classical pivoting strategy.

Now, suppose we introduce a perturbation δA_{11} that changes only the element $1+\varepsilon$ to $1+\varepsilon/2$. Let the block A_{12} remain unchanged, i.e., $\delta A_{12} = 0$. However, let us insist (as we do in enforcing the rank during the computation) that $A + \delta A$ is again of rank 2. Then A_{22} has to be changed to the value

$$A_{22} + \delta A_{22} = A_{12}^T (A_{11} + \delta A_{11})^{-1} A_{12} = 2\varepsilon^3.$$

We see that the relative change of A_{22} is of order 1, as indicated by the upper bound in Theorem 4.1, which is about $\kappa_2(A_{11}) \cdot \varepsilon = \mathcal{O}(1)$. Thus, this bound is realistic.

Example 4.2. This example was prompted by one of the anonymous referees who urged us to compare the statements of our Theorem 4.1 and of Theorem 10.14 in [14] that seem to give a sharper bound for the perturbation of A_{22}. Let

$$A = \left[\begin{array}{cc|c} 1+\epsilon & 1 & 1+\epsilon/2 \\ 1 & 1 & 1 \\ \hline 1+\epsilon/2 & 1 & 1+\epsilon/4 \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix}. \tag{4.8}$$

This matrix is positive semidefinite of rank 2. The block A_{11} is the same as in Example 4.1, but the block A_{22} is not tiny anymore. Again $\kappa_2(A_{11}) \approx 4/\varepsilon$, but now $\|A_{11}^{-1} A_{12}\|_2 = 1/\sqrt{2}$. Knowing that $\operatorname{rank}(A) = 2$, we compute R_{11} from $A_{11} = R_{11}^T R_{11}$ and then $R_{12} = R_{11}^{-T} A_{12}$. In exact arithmetic we then have $A_{22} = A_{12}^T A_{11}^{-1} A_{12}$.

Let us assume that the computed $\tilde{R}_{11}$ and $\tilde{R}_{12}$ are the exact factors of blocks that are perturbed entrywise (and with respect to the spectral norm) by $\mathcal{O}(\epsilon)$ quantities,

$$\tilde{R}_{11}^T \tilde{R}_{11} = \tilde{A}_{11} = \begin{bmatrix} 1+\epsilon & \sqrt{1+\epsilon-\epsilon^2} \\ \sqrt{1+\epsilon-\epsilon^2} & 1 \end{bmatrix}, \qquad \tilde{R}_{11}^T \tilde{R}_{12} = \tilde{A}_{12} = \begin{bmatrix} 1-\epsilon/2 \\ 1 \end{bmatrix}.$$

Then, in order that $[\tilde{R}_{11}, \tilde{R}_{12}]$ is the Cholesky factor of a matrix of rank 2, we must have

$$\tilde{A}_{22} = \tilde{A}_{12}^T \tilde{A}_{11}^{-1} \tilde{A}_{12} = 13/4 + \mathcal{O}(\epsilon).$$

With respect to the unperturbed $A_{22} = 1 + \epsilon/4$, this is an $\mathcal{O}(1)$ change. As in the previous example this corresponds to the upper bound in Theorem 4.1, which is $\kappa_2(A_{11}) \cdot \varepsilon = \mathcal{O}(1)$.

This result appears to contradict Theorem 10.14 in [14], which gives an $\mathcal{O}(\epsilon)$ bound for the modification of A_{22}. The problem can be traced back to Lemma 10.10, which is used in the proof of Theorem 10.14. In this lemma a Neumann series is used to expand the Schur complement $\tilde{A}_{22} - \tilde{A}_{12}^T \tilde{A}_{11}^{-1} \tilde{A}_{12}$ in terms of $E = \delta A$. For the matrix A of the present example, this Neumann series diverges. In fact, for the terms of the form $E_{11}(A_{11}^{-1} E_{11})^k$, $E_{11} = \delta A_{11}$, we have $\|E_{11}(A_{11}^{-1} E_{11})^k\| = \mathcal{O}(1/\epsilon)$. Therefore, Lemma 10.10 and consequently Theorem 10.14 in [14] cannot be applied to A in (4.8).

Notice that, formally, Theorem 4.1 cannot be applied directly, since the assumption that $\lambda_{\min}((A_{11})_s) > 2rf(r)\mathbf{u}$ is not satisfied. In fact, $\lambda_{\min}((A_{11})_s) \approx \epsilon/2$ is smaller than the perturbation, which is of size ϵ. However, the technique of the proof can be applied directly if we assume that the Cholesky factorization of $\tilde{A}_{11}$ exists (instead of giving a technical condition that ensures it). So, this excellent example fully complies with our theory. As our construction shows, no other theory can, in this situation, give an $O(\epsilon)$ estimate for the relative change in A_{22}.

Remark 4.1. Note that Theorem 4.1 states that in the positive definite case the Cholesky factorization with pivoting computes the triangular factor with small

column- and rowwise relative errors. This affects the accuracy of the linear equation solver (forward and backward substitutions following the Cholesky factorization) not only by ensuring favorable condition numbers but also by ensuring that the errors in the coefficients of the triangular systems are small.

4.2. Null space error. We now derive a backward error for the null space Y of A. We seek an $n \times (n-r)$ full rank matrix $\widetilde{Y} = Y + \delta Y$ such that δY is small and $\widetilde{A}\widetilde{Y} = 0$. As the null space and the range of A change simultaneously (being orthogonal complements of each other), the size of δY necessarily depends on a certain condition number of A; and the relevant condition number will depend on the form of the perturbation δA.

The equation that we investigate is $\widetilde{R}(Y + \delta Y) = 0$ or, equivalently, $\widetilde{R}\delta Y = -\delta R\, Y$. If $\widetilde{R}$ is sufficiently close to R (to guarantee invertibility of $\widetilde{R}R^+$), we can write

$$\delta Y = R^+(\widetilde{R}R^+)^{-1}\delta R\, Y = R^T(\widetilde{R}R^T)^{-1}\delta R\, Y. \tag{4.9}$$

Though simple, this equation is instructive. First of all, only the components of the columns of δR that lie in the null space $\mathcal{N}(A)$ affect the value of δY. Also, $Y + \delta Y$ keeps the full column rank of Y. Finally, $Y^T\delta Y = O$. Therefore, $\tan\angle(\mathcal{R}(Y), \mathcal{R}(\widetilde{Y})) = \|\delta Y\|/\sigma_{\min}(Y)$. It is easy to modify Y such that $\sigma_{\min}(Y) \geq 1$, e.g., if $Y_2 = I_m$. Thus, $\|\delta Y\|$ measures the angle between the true null space and the null space of the perturbed matrix $\widetilde{A}$. In what follows we try to bound $\|\delta Y\|$.

If we rewrite (4.9) as

$$\delta Y = R^+(I + \delta R\, R^+)^{-1}\delta R\, Y = (R')^+\left(I + \delta R'(R')^+\right)^{-1}\delta R'\, Y,$$

we get, after some manipulations, the following theorem.

THEOREM 4.2. *Let D be a nonsingular matrix and let $R = DR'$, $\delta R = D\delta R'$. If $\|\delta R'(R')^+\| < 1$, then, for $i = 1, \ldots, n-r$,*

$$\|\delta\mathbf{y}_i\| \leq \frac{\|(R')^+\|}{1 - \|\delta R'(R')^+\|}\|\delta R'\mathbf{y}_i\| \leq \frac{\|\delta R' P_{\mathcal{N}(A)}\|\|(R')^+\|}{1 - \|\delta R'(R')^+\|}\|\mathbf{y}_i\|. \tag{4.10}$$

Here, $\mathbf{y}_i = Y\mathbf{e}_i$, $\delta\mathbf{y}_i = Y\delta\mathbf{e}_i$, and $P_{\mathcal{N}(A)}$ denotes the orthogonal projection onto the null space of A.

We will discuss choices for D later. The theorem indicates that the crucial quantity for bounding $\|\delta Y\|$ is $\|\delta R'Y\|$. The following two examples detail this fact.

Example 4.3. Let β be big, of the order of $1/\mathbf{u}$, and let

$$A = R^TR = \begin{bmatrix} \sqrt{3} & 0 \\ 1 & 1 \\ \beta & 1 \end{bmatrix}\begin{bmatrix} \sqrt{3} & 1 & \beta \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & \sqrt{3} & \beta\sqrt{3} \\ \sqrt{3} & 2 & \beta+1 \\ \beta\sqrt{3} & \beta+1 & \beta^2+1 \end{bmatrix}.$$

The null space of A is spanned by $Y = [(1-\beta)/\sqrt{3}, -1, 1]^T$, which means that deleting any row and column of A leaves a nonsingular 2×2 matrix. Let's choose it be the last one, and let us follow the algorithm. For the sake of simplicity, let the only error be committed in the computation of the $(1,1)$ entry of $\widetilde{R}_{11}$, which is $\sqrt{3}(1+\varepsilon_1)$, $|\varepsilon_1| \leq \mathbf{u}$, instead of $\sqrt{3}$. Then we solve the lower triangular system for $\widetilde{R}_{12}$ and obtain

$$\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}] = \begin{bmatrix} \sqrt{3}\,(1+\varepsilon_1) & 1 & \beta(1+\varepsilon_2) \\ 0 & 1 & 1-\beta\varepsilon_2 \end{bmatrix}, \qquad |\varepsilon_2| \leq \mathbf{u} + \mathcal{O}(\mathbf{u}^2).$$

Thus,

$$\delta R = \begin{bmatrix} \sqrt{3}\,\varepsilon_1 & 0 & \beta\varepsilon_2 \\ 0 & 0 & -\beta\varepsilon_2 \end{bmatrix}, \qquad \delta R\, Y = \begin{pmatrix} (1-\beta)\varepsilon_1 + \beta\varepsilon_2 \\ -\beta\varepsilon_2 \end{pmatrix}.$$

If we take $\beta = 10^{15}$ and perform the computation in MATLAB, where $\mathbf{u} \approx 2.22 \cdot 10^{-16}$, then $\beta\varepsilon_2 = 0.25$. Thus, $\|\delta R\, Y\| = \mathcal{O}(1)$. However, $\sigma_{\min}(Y) = \|Y\| = \mathcal{O}(\beta)$ such that the angle between Y and δY is small.

Example 4.4. We alter the $(1,1)$ entry $\sqrt{3}$ of R of the previous example to get β,

$$A = R^T R = \begin{bmatrix} \beta & 0 \\ 1 & 1 \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \beta & 1 & \beta \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \beta^2 & \beta & \beta^2 \\ \beta & 2 & \beta+1 \\ \beta^2 & \beta+1 & \beta^2+1 \end{bmatrix}.$$

Now, $Y = [(1-\beta)/\beta, -1, 1]^T$. Again, we delete the last row and column of A and proceed as in Example 4.3. Let us again assume that the only error occurs in the $(1,1)$ entry of R_{11}, which becomes $\beta/(1+\varepsilon_1)$. Then

$$\widetilde{R} = \begin{bmatrix} \beta/(1+\varepsilon_1) & 1 & \beta(1+\varepsilon_1) \\ 0 & 1 & 1-\beta\varepsilon_1 \end{bmatrix}$$

and

$$\delta R = \begin{bmatrix} -\beta\varepsilon_1/(1+\varepsilon_1) & 0 & \beta\varepsilon_1 \\ 0 & 0 & -\beta\varepsilon_1 \end{bmatrix}, \qquad \delta R\, Y = \begin{pmatrix} \varepsilon_1(-1+2\beta+\beta\varepsilon_1)/(1+\varepsilon_1) \\ -\beta\varepsilon_1 \end{pmatrix}.$$

Again, $\|\delta R\, Y\| = \mathcal{O}(1)$. But now also $\|Y\| = \mathcal{O}(1)$. In fact, in computations with MATLAB, we observe an angle as large as $\mathcal{O}(10^{-2})$ between Y and δY.

Remark 4.2. Interestingly, if we set $\beta = 10^5$ in Example 4.3, the MATLAB function `chol()` computes the Cholesky factor

$$\widetilde{R} = \begin{bmatrix} \texttt{1.7321e+000} & \texttt{1.0000e+000} & \texttt{1.0000e+005} \\ \texttt{0} & \texttt{1.0000e+000} & \texttt{1.0000e+000} \\ \texttt{0} & \texttt{0} & \texttt{1.9531e-003} \end{bmatrix}.$$

It is clear that the computed and stored A is a perturbation of the true A. Therefore, numerically, it can be positive definite. It is therefore quite possible to know the rank $r < n$ of A exactly, to have a basis of the null space of A and a numerically stored positive definite floating-point A. Strictly speaking, this is a contradiction. Certainly, from an application or numerical point of view, it is advisable to be very careful when dealing with semidefiniteness.

In Examples 4.3 and 4.4 we excluded the largest diagonal entry of A. In fact, we can give an estimate that relates the error in R_{12} to the size of the deleted entries. Suppose we managed the deleted diagonal entries of A to be the $m = n - r$ smallest ones. Can we then guarantee that the relevant error in R will be small, and can we check the stability by a simple, inexpensive test?

According to Theorem 4.1, the matrix R_{11} is computed with rowwise small relative error, provided that the Cholesky factorization of A_{11} is computed with pivoting. If that is the case, then it remains to estimate the rowwise perturbations of R_{12}. If Ξ is as in Theorem 4.1, then the inequality

$$\|\mathbf{e}_i^T \delta R_{12}\| \le \|\mathbf{e}_i^T \Xi\| \sqrt{\operatorname{trace}(A_{22})} \le \|\mathbf{e}_i^T \Xi\| \left(\frac{\operatorname{trace}(A_{22})}{(A_{11})_{ii}} \right)^{1/2} \frac{\|R_{11}\mathbf{e}_i\|}{\|\mathbf{e}_i^T R\|} \|\mathbf{e}_i^T R\| \tag{4.11}$$

holds for all $i = 1, \ldots, r$ and

$$\frac{\|R_{11}\mathbf{e}_i\|}{\|\mathbf{e}_i^T R\|} = \frac{\|R_{11}\mathbf{e}_i\|}{|(R_{11})_{ii}|}\frac{|(R_{11})_{ii}|}{\|\mathbf{e}_i^T R\|} \leq \frac{\|R_{11}\mathbf{e}_i\|}{|(R_{11})_{ii}|} = \frac{1}{\sin\phi_i} \leq \sqrt{\|(A_{11})_s^{-1}\|} \tag{4.12}$$

with some $\phi_i \in (0, \pi/2]$. The angle ϕ_i has a nice interpretation. Let $A = F^T F$ be any factorization of A, with $F = [F_1, F_2]$, where F_1 has full column rank and $F_1^T F_1 = A_{11}$. Then ϕ_i is the angle between $F_1\mathbf{e}_i$ and the span of $\{F_1\mathbf{e}_1, \ldots, F_1\mathbf{e}_{i-1}\}$. (This is easily seen from the QR factorization of F_1.)

The following theorem states that well-conditioned $(A_{11})_s$ and a certain dominance of A_{11} over A_{22} ensure accurate rows of the computed matrix $\widetilde{R}$.

THEOREM 4.3. *With the notation of Theorem* 4.1, *let A (and accordingly Y) be arranged such that*

$$\max_i (A_{22})_{ii} \leq \min_i (A_{11})_{ii}. \tag{4.13}$$

If the Cholesky factorization of A_{11} is computed with (standard) pivoting, then

$$\|\mathbf{e}_i^T \delta R\| \leq \max\{\|\mathbf{e}_i^T \Gamma\|, \|\mathbf{e}_i^T \Xi\|\} \frac{\sqrt{n-i+1}}{\sin\phi_i} \|\mathbf{e}_i^T R_{11}\|, \quad i = 1, \ldots, r, \tag{4.14}$$

where $\sin\phi_i$ *is defined as in* (4.12).

Proof. This follows from relations (4.6), (4.11), (4.12) and the assumption (4.13). We only note that in (4.11) and (4.12) we can replace $\|\mathbf{e}_i^T R\|$ by $\|\mathbf{e}_i^T R_{11}\|$. ∎

Remark 4.3. If $A = SA_sS$ with $S^2 = \mathrm{diag}(A_{ii})$, then SY spans $\mathcal{N}(A_s)$, and any partition of A_s satisfies condition (4.13). If we apply the preceding analysis to A_s and SY, we get an estimate for δY in the elliptic norm generated by S.

Note that Theorem 4.2 is true for any diagonal D as long as $\|(R')^+\|$ is moderately big and $\|\delta R'\|$ is small. We have just seen that $\delta R'$ is nicely bounded if we choose $D = \mathrm{diag}(\|\mathbf{e}_i^T R_{11}\|)$. Moreover, $R' = D^{-1}R$ has an inverse nicely bounded independent of A_{11} because [14, section 10]

$$\|(R')^+\| \leq \|(D^{-1}R_{11})^{-1}\| \leq h(r).$$

Here the function $h(r)$ is in the worst case dominated by 2^r, and in practice one usually observes an $\mathcal{O}(r)$ behavior. In any case, $\|(D^{-1}R_{11})^{-1}\|$ is at most r times larger than $\|(A_{11})_s^{-1}\|^{1/2}$. More sophisticated pivoting can make sure that the behavior of $h(r)$ is not worse than Wilkinson's pivot growth factor. We skip the details for the sake of brevity.

To conclude, if the Cholesky factorization of A_{11} is computed with pivoting and relation (4.13) holds, then the backward error in Y can be estimated using (4.10) and (4.14), where $D = \mathrm{diag}(\|\mathbf{e}_i^T R_{11}\|)$.

4.3. Computation with implicit A. We consider now the backward stability of the computation with A given implicitly as $A = F^T F$, where $F \in \mathbb{R}^{p\times n}$ has rank r. Thus, the Cholesky factorization of A is accomplished by computing the QR factorization of F.

In the numerical analysis of the QR factorization we use the standard, well-known backward error analysis which can be found, e.g., in [14, section 18]. The simplest form of this analysis states that the backward error in the QR factorization is columnwise

small. For instance, if we compute the Householder (or Givens) QR factorization of F in floating-point arithmetic with round-off $\mathbf{u}$, then the backward error δF satisfies

$$\|\delta F \mathbf{e}_i\| \le \varepsilon_1 \|F\mathbf{e}_i\|, \quad \varepsilon_1 \le f_1(p,n)\mathbf{u}, \qquad 1 \le i \le n,$$

where $f_1(p,n)$ is a polynomial of moderated degree in the matrix dimensions.

Our algorithm follows the same ideas as in the direct computation of R from A. The knowledge of a null space basis admits that we can assume that F is in the form $F = [F_1, F_2]$, where the $p \times r$ matrix F_1 is of rank r; see section 3. We then apply r Householder reflections to F, which yields, in exact arithmetic, the matrix

$$Q^T F = R = \begin{pmatrix} R_{11} & R_{12} \\ O & R_{22} \end{pmatrix}, \qquad R_{22} = O,$$

where $R_{11} \in \mathbb{R}^{r\times r}$ is upper triangular and nonsingular. If $Q = [Q_1, Q_2]$ is partitioned conforming with F, then $F_1 = Q_1 R_{11}$ is the QR factorization of F_1.

In floating-point computation, R_{22} is unlikely to be zero. Our algorithm simply sets to zero whatever is computed as approximation of R_{22}. As we shall see, the backward error (in F) of this procedure depends on a certain condition number of the matrix F_1.

THEOREM 4.4. *Let $F \in \mathbb{R}^{p\times n}$ have rank r and be partitioned in the form $F = [F_1, F_2]$, where $F_1 \in \mathbb{R}^{p\times r}$ has the numerically well determined full rank r. More specifically, if $(F_1)_c$ is obtained from F_1 by scaling columns to have unit Euclidean norm, then we assume that $\sqrt{r}\varepsilon_1 \|(F_1)_c^+\| < 1/5$.*

Let the QR factorization of F be computed as described above, and let $\widetilde{R} = [\widetilde{R}_{11}, \widetilde{R}_{12}]$ be the computed upper trapezoidal factor.

Then there exist a backward perturbation ΔF and an orthogonal matrix $\widehat{Q}$ such that $F+\Delta F = \widehat{Q}\widetilde{R}$ is the QR factorization of $F+\Delta F$. The matrix $F+\Delta F$ has rank r. If $\Delta F = [\Delta F_1, \Delta F_2]$ and $\widehat{Q} = [\widehat{Q}_1, \widehat{Q}_2]$ are partitioned as F, and $\delta Q_1 := \widehat{Q}_1 - Q_1$, then

$$\begin{aligned}
&\|\Delta F \mathbf{e}_i\| \le \varepsilon_1 \|F\mathbf{e}_i\|, && 1 \le i \le r,\\
&\|\delta Q_1\|_F \le 11\eta + \mathcal{O}(\eta^2), && \eta = \|\Delta F_1 R_{11}^{-1}\|_F \le \sqrt{r}\varepsilon_1 \|(F_1)_c^+\|,\\
&\|\Delta F \mathbf{e}_i\| \le (\varepsilon_1 + \|\delta Q_1\|)\|F\mathbf{e}_i\|, && r+1 \le i \le n,\\
&\widetilde{R}_{11} - R_{11} = GR_{11}, && \|G\|_F \le \|\delta Q_1\|_F + \eta,
\end{aligned}$$

where $\varepsilon_1 \le f_1(p,r)\mathbf{u}$ bounds the round-off.

Proof. Let $\widetilde{F}^{(r)}$ be the matrix obtained after r steps of the Householder QR factorization. Then there exist an orthogonal matrix $\widehat{Q}$ and a backward perturbation δF such that

$$\begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ O & \widetilde{R}_{22} \end{bmatrix} \equiv \widetilde{F}^{(r)} = \widehat{Q}^T(F + \delta F), \qquad \|\delta F \mathbf{e}_i\| \le \varepsilon_1 \|F\mathbf{e}_i\|, \qquad 1 \le i \le n.$$

Our assumption on the numerical rank of F_1 implies that $F_1 + \delta F_1 = \widehat{Q}_1 \widetilde{R}_{11}$ is the QR factorization with nonsingular $\widetilde{R}_{11}$. Now, setting $\widetilde{R}_{22}$ to zero is, in the backward error sense, equivalent to the QR factorization of a rank r matrix,

$$\widehat{Q}\begin{bmatrix} \widetilde{R}_{11} & \widetilde{R}_{12} \\ O & O \end{bmatrix} = F + \Delta F, \qquad \Delta F = \delta F - \widehat{Q}\begin{bmatrix} O & O \\ O & \widetilde{R}_{22} \end{bmatrix}.$$

It remains to estimate $\widehat{Q}_2\widetilde{R}_{22} = \widehat{Q}_2\widehat{Q}_2^T(F_2+\delta F_2)$. First note that $F_2 = Q_1R_{12}$, where the ith column of R_{12} has the same norm as the corresponding column of F_2. Then

$$\widehat{Q}_2\widehat{Q}_2^T F_2 = \widehat{Q}_2\widehat{Q}_2^T Q_1R_{12} = \widehat{Q}_2\widehat{Q}_2^T(\widehat{Q}_1 - \delta Q_1)R_{12} = -\widehat{Q}_2\widehat{Q}_2^T\delta Q_1R_{12},$$

and we can write

$$\|\widehat{Q}_2\widehat{Q}_2^T F_2\mathbf{e}_i\| \le \|\delta Q_1\|\,\|F_2\mathbf{e}_i\|, \qquad 1 \le i \le n-r.$$

To estimate δQ_1, we first note that $F_1 = Q_1R_{11}$ and $F_1 + \Delta F_1 = \widehat{Q}_1\widetilde{R}_{11}$ imply that

$$\widehat{Q}_1 = (I + \Delta F_1F_1^+)Q_1(R_{11}\widetilde{R}_{11}^{-1}),$$

and that

$$R_{11}^{-T}\widetilde{R}_{11}^T\widetilde{R}_{11}R_{11}^{-1} = I + Q_1^T\Delta F_1R_{11}^{-1} + R_{11}^{-T}\Delta F_1^TQ_1 + R_{11}^{-T}\Delta F_1^T\Delta F_1R_{11}^{-1}.$$

Thus, $\widetilde{R}_{11}R_{11}^{-1}$ is the Cholesky factor of $I+E$, where

$$\|E\|_F \le 2\|\Delta F_1R_{11}^{-1}\|_F + \|\Delta F_1R_{11}^{-1}\|_F^2.$$

Now, by [8], $\|E\|_F < 1/2$ implies that $\widetilde{R}_{11}R_{11}^{-1} = I+\Gamma$, where Γ is upper triangular and

$$\|\Gamma\|_F \le \frac{\sqrt{2}\|E\|_F}{1+\sqrt{1-2\|E\|_F}} < \frac{1}{\sqrt{2}}.$$

Hence, $R_{11}\widetilde{R}_{11}^{-1} = I+\widehat{\Gamma}$, where $\|\widehat{\Gamma}\|_F \le \|\Gamma\|_F/(1-\|\Gamma\|_F) < (2+\sqrt{2})\|\Gamma\|_F$. Since $\widehat{Q}_1 = Q_1 + Q_1\widehat{\Gamma} + \Delta F_1R_{11}^{-1} + \Delta F_1R_{11}^{-1}\widehat{\Gamma}$, we obtain

$$\|\delta Q_1\|_F \le \|\widehat{\Gamma}\|_F + \|\Delta F_1R_{11}^{-1}\|_F + \|\widehat{\Gamma}\|\,\|\Delta F_1R_{11}^{-1}\|_F.$$

Finally, note that $\widetilde{R}_{11} - R_{11} = (\widehat{Q}_1^T\delta F_1R_{11}^{-1} - \delta Q_1^TQ_1)R_{11}$. ∎

We remark that

$$\widetilde{R}_{12} = R_{12} + \delta Q_1^TQ_1R_{12} + \widehat{Q}_1^T\Delta F_2,$$

which means that we can nicely bound $\delta R_{12} = \widetilde{R}_{12} - R_{12}$. We have, for instance,

$$\|\delta R_{12}\mathbf{e}_i\| \le (2\|\delta Q_1\| + \varepsilon_1)\|R_{12}\mathbf{e}_i\|, \qquad 1 \le i \le n-r.$$

If we use entrywise backward analysis of the QR factorization ($|\delta F| \le \varepsilon_2\mathbf{e}\mathbf{e}^T|F_2|$, $\mathbf{e} = (1,\dots,1)^T$), then we can also write

$$|\delta R_{12}| \le (|\delta Q_1^TQ_1| + \varepsilon_2|\widehat{Q}_1|^T\mathbf{e}\mathbf{e}^T|Q_1|)|R_{12}|,$$

where the matrix absolute values and inequalities are understood entrywise, and ε_2 is defined similarly as ε_1.

From the above analysis we see that the error in the computed matrix $\widetilde{R}$ is bounded in the same way as in Theorem 4.1. Also, the QR factorization can be computed with the standard column pivoting and R_{11} can have additional structure just as in the Cholesky factorization of A_{11}. Therefore, the analysis of the backward null space perturbation based on $\widetilde{R}^T$ holds in this case as well. However, the bounds of Theorem 4.4 are sharper than those of Theorem 4.1.

5. Constrained systems of equations. Again, let be $\mathcal{N}(A) = \mathcal{R}(Y)$ with $Y \in \mathbb{R}^{n\times m}$ having full rank. Let $C \in \mathbb{R}^{n\times m}$ be a matrix with full rank. Systems of equations of the form

$$\begin{bmatrix} A & C \\ C^T & O \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} \tag{5.1}$$

appear at many occasions, e.g., in mixed finite element methods [4] or constrained optimization [15]. They have a solution for every right side if $\mathbb{R}^n = \mathcal{R}(A) \oplus \mathcal{R}(C)$, which is the case if $H := Y^T C$ is nonsingular. In the computations of Stokes [4] or Maxwell equations [1] the second equation in (5.1) with $\mathbf{c} = \mathbf{0}$ imposes a divergence-free condition on the flow or electric field, respectively.

Duff et al. [9] discuss a multifrontal code for solving general sparse symmetric indefinite systems of the form (5.1). Here, we consider an algorithm that takes advantage of the knowledge of a basis of the null space of A. To that end we first construct a particular solution of the first block row. Premultiplying it by Y^T yields $\mathbf{y} = H^{-1}Y^T\mathbf{b}$. As $\mathbf{b} - C\mathbf{y} \in \mathcal{R}(A)$ we can proceed as in section 3 to obtain a vector $\tilde{\mathbf{x}}$ with $A\tilde{\mathbf{x}} = \mathbf{b} - C\mathbf{y}$. The solution $\mathbf{x}$ of (5.1) is obtained by setting $\mathbf{x} = \tilde{\mathbf{x}} + Y\mathbf{a}$ and determining $\mathbf{a}$ such that $C^T\mathbf{x} = \mathbf{c}$. Thus, $\mathbf{a} = H^{-T}(\mathbf{c} - C^T\tilde{\mathbf{x}})$.

This procedure can be described in an elegant way if a congruence transformation as in (6.2) is applied. Multiplying (5.1) by $W^T \oplus I_m$ (cf. (2.2)) yields

$$\begin{bmatrix} A_{11} & O & C_1 \\ O & O & H \\ C_1^T & H^T & O \end{bmatrix} \begin{pmatrix} \tilde{\mathbf{x}}_1 \\ \mathbf{a} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ Y^T\mathbf{b} \\ \mathbf{c} \end{pmatrix}, \qquad \begin{array}{l} \tilde{\mathbf{x}}_1 = \mathbf{x}_1 - Y_1 Y_2^{-1}\mathbf{x}_2, \\ \mathbf{a} = Y_2^{-1}\mathbf{x}_2, \\ \mathbf{b}_1 = I_{n,r}^T\mathbf{b}. \end{array} \tag{5.2}$$

Notice that $\tilde{\mathbf{x}}_1 \in \mathbb{R}^r$. From (5.2) we read that

$$\begin{array}{ll} \text{(i)} & \mathbf{y} = H^{-1}Y^T\mathbf{b}, \\ \text{(ii)} & \tilde{\mathbf{x}}_1 = A_{11}^{-1}(\mathbf{b}_1 - C_1\mathbf{y}), \\ \text{(iii)} & \mathbf{a} = H^{-T}(\mathbf{c} - C_1^T\tilde{\mathbf{x}}_1), \end{array} \qquad \begin{array}{ll} \text{(iv)} & \mathbf{x}_1 = \tilde{\mathbf{x}}_1 + Y_1\mathbf{a}, \\ \text{(v)} & \mathbf{x}_2 = Y_2\mathbf{a}. \end{array} \tag{5.3}$$

This geometric approach differs from the algebraic approach, also known as the null space algorithm [11], that is based on the factorization

$$\begin{bmatrix} A_{11} & A_{12} & C_1 \\ A_{12}^T & A_{22} & C_2 \\ C_1^T & C_2^T & O \end{bmatrix} = \begin{bmatrix} R_{11}^T & O & O \\ R_{12}^T & I_m & O \\ C_1^T R_{11}^{-1} & O & I_m \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} & R_{11}^{-T}C_1 \\ O & O & C_2 - R_{12}^T R_{11}^{-T} C_1 \\ O & C_2^T - C_1^T R_{11}^{-1} R_{12} & -C_1^T R_{11}^{-1} R_{11}^{-T} C_1 \end{bmatrix},$$

where the LU factorization of $C_2 - R_{12}^T R_{11}^{-T} C_1$ is employed to solve (5.1). In the geometric approach the LU factorization of H is used instead. Of course, there is a close connection between the two approaches: Using (3.4) we get $C_2^T - C_1^T R_{11}^{-1} R_{12} = H^T Y_2^{-1}$. Notice that the columns of C or Y can be scaled such that the condition numbers of H or $C_2 - R_{12}^T R_{11}^{-T} C_1$ are not too big. Notice also that Y can be chosen such that $Y_2 = I_m$, in which case $C_2^T - C_1^T R_{11}^{-1} R_{12} = H^T$. To enhance stability, instead of the LR factorization the QR factorization is often used in the algebraic approach; see [3] for references and an error analysis. A thorough perturbation analysis of (5.1)–(5.3) remains to be done in our future work.

Golub and Greif [12] use the algebraic approach to solve systems of the form (5.1) if the positive semidefinite A has a low-dimensional null space. As they do not have available a basis for the null space, they apply a trial-and-error strategy for finding a

permutation of A such that the leading $r \times r$ principal submatrix becomes nonsingular. They report that usually the first trial is successful. This is understandable because $n_i = r + i = n - m + i$ if the basis of the null space is dense, which is often the case.

If the null space of A is high-dimensional, then Golub and Greif use an augmented Lagrangian approach. They modify (5.1) such that the $(1,1)$ block becomes positive definite,

$$\begin{bmatrix} A + C\Delta C^T & C \\ C^T & O \end{bmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{b} + C\Delta\mathbf{c} \\ \mathbf{c} \end{pmatrix}.$$

Here, Δ is some symmetric positive definite matrix, e.g., a multiple of the identity. $A + C\Delta C^T$ is positive definite if $Y^T C$ is nonsingular. The determination of a good Δ is difficult. Golub and Greif thoroughly discuss how to choose Δ and how the "penalty term" $C\Delta C^T$ affects the condition of the problem. In contrast to this approach, where a term is added to A that is positive definite on the null space of A, $\mathcal{N}(A)$ can be avoided right away if a basis of it is known.

6. Eigenvalue problems. Let us consider the eigenvalue problem

$$A\mathbf{x} = \lambda M\mathbf{x}, \tag{6.1}$$

where A is symmetric positive semidefinite with $\mathcal{N}(A) = \mathcal{R}(Y)$ and M is symmetric positive definite. We assume that the last m rows of Y are linearly independent such that W in (2.1) is nonsingular. Then

$$W^T A W = \begin{bmatrix} A_{11} & O \\ O & O \end{bmatrix}, \qquad W^T M W = \begin{bmatrix} M_{11} & C_1 \\ C_1^T & H \end{bmatrix}, \tag{6.2}$$

where

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \qquad H = Y^T M Y = Y^T C.$$

Using the decomposition

$$W^T M W = \begin{bmatrix} M_{11} & C_1 \\ C_1^T & H \end{bmatrix} = P^T \begin{bmatrix} S & O \\ O & H \end{bmatrix} P, \qquad P = \begin{bmatrix} I & O \\ H^{-1}C_1^T & I \end{bmatrix} \tag{6.3}$$

with the Schur complement $S := M_{11} - C_1 H^{-1} C_1^T$ and noting that $P^{-T} W^T A W P^{-1} = W^T A W$, it is easy to see that the positive eigenvalues of (6.1) are the eigenvalues of

$$A_{11}\mathbf{y} = \lambda(M_{11} - C_1 H^{-1} C_1^T)\mathbf{y} = \lambda S\mathbf{y}. \tag{6.4}$$

Notice that S is dense, in general, whence, in sparse matrix computations, it should not be formed explicitly.

If $\mathbf{y}$ is an eigenvector of (6.4), then

$$\mathbf{x} = WP^{-1} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{y} - Y_1 H^{-1} C_1^T \mathbf{y} \\ -Y_2 H^{-1} C_1^T \mathbf{y} \end{pmatrix} = (I - Y H^{-1} C^T) \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \tag{6.5}$$

is an eigenvector of (6.1). By construction, $C^T\mathbf{x} = Y^T M\mathbf{x} = \mathbf{0}$, i.e., $\mathbf{x}$ is M-orthogonal to the null space of A.

We now consider the situation when A and M are given in factored form, $A = F^T F$ and $M = B^T B$, with $F = [F_1, F_2]$ and $B = [B_1, B_2]$ such that the rank of F_1

equals the rank of A. Let us find an implicit formulation of the reduced problem (6.4). With W from (2.1) we have $[F_1, F_2]W = [F_1, O]$. As before, $A_{11} = R_{11}^T R_{11}$, where R_{11} is computed by the QR factorization of F_1. It remains to compute a Cholesky factor of the Schur complement S, but directly from the matrix B. To that end we employ the QL factorization ("backward" QR factorization) of BW,

$$BW = QL = [Q_1, Q_2] \begin{bmatrix} L_{11} & O \\ L_{21} & L_{22} \end{bmatrix}, \qquad Q^T Q = I_n, \tag{6.6}$$

whence, with (6.3),

$$W^T M W = W^T B^T B W = \begin{bmatrix} L_{11}^T L_{11} + L_{21}^T L_{21} & L_{21}^T L_{22} \\ L_{22}^T L_{21} & L_{22}^T L_{22} \end{bmatrix} = \begin{bmatrix} M_{11} & C_1 \\ C_1^T & H \end{bmatrix}. \tag{6.7}$$

Straightforward calculation now reveals that

$$S = M_{11} - C_1 H^{-1} C_1^T = L_{11}^T L_{11}.$$

Thus, the eigenvalues of the matrix pencil (A_{11}, S) are the squares of the generalized singular values [13] of the matrix pair (R_{11}, L_{11}) or, equivalently, the squares of the singular values of $R_{11} L_{11}^{-1}$. An eigenvector $\mathbf{y}$ corresponds to a right singular vector $L_{11}\mathbf{y}$. The blocks L_{21} and L_{22} come into play when the eigenvectors of (6.1) are to be computed: using (6.7), equation (6.5) becomes

$$\mathbf{x} = \begin{pmatrix} \mathbf{y} \\ -L_{22}^{-1} L_{21} \mathbf{y} \end{pmatrix}.$$

It is known that the GSVD of (R_{11}, L_{11}) can be computed with high relative accuracy if the matrices $(R_{11})_c$ and $(L_{11})_c$ are well conditioned [7]. Here, $(R_{11})_c$ and $(L_{11})_c$ are obtained by R_{11} and L_{11}, respectively, by scaling their columns to make them of unit length. Obviously, $\kappa_2((R_{11})_c) = \kappa_2((F_1)_c)$, where $\kappa_2(\cdot)$ is the spectral condition number. It remains to determine $\kappa_2((L_{11})_c)$. From (6.6) we get

$$Q_1^T BW = Q_1^T [B_1, BY] = [L_{11}, O_{r,m}],$$

whence $Q_1^T B_1 = L_{11}$. Let the diagonal matrix D_1 be such that $(B_1)_c := B_1 D_1^{-1}$ has columns of unit length. Further, let $(B_1)_c = U_1 G_1$ be the QR factorization of $(B_1)_c$ and let $(L_{11})_s = L_{11} D_1^{-1} = Q_1^T U_1 G_1$. As Q_1 is orthogonal we have $\|(L_{11})_s\| \leq \|(B_1)_c\| = \sigma_{\max}((B_1)_c)$. Further,

$$\|(L_{11})_s^{-1}\| \leq \|G_1^{-1}\| \|(Q_1^T U_1)^{-1}\| = \frac{1}{\sigma_{\min}((B_1)_c) \cos\Phi},$$

where Φ is the largest principal angle [13] between $\mathcal{R}(B_1)$ and $\mathcal{R}(B_2)^\perp \bigcap \mathcal{R}(B)$. Therefore,

$$\kappa_2((L_{11})_s) \leq \frac{\sigma_{\max}((B_1)_c)}{\sigma_{\min}((B_1)_c) \cos\Phi} = \frac{\kappa_2((B_1)_c)}{\cos\Phi}.$$

Since $\kappa_2((L_{11})_c) \leq \sqrt{r} \min_{D=\text{diagonal}} \kappa_2(L_{11} D)$ [16], [14, Theorem 7.5], we have

$$\kappa_2((L_{11})_c) \leq \sqrt{r}\, \kappa_2((L_{11})_s) \leq \sqrt{r}\, \kappa_2((B_1)_c)/\cos\Phi. \tag{6.8}$$

So, we have identified condition numbers that do not depend on column scalings and that have a nice geometric interpretation. If the perturbations are columnwise small, then these condition numbers are the relevant ones.

7. Concluding remarks. In this paper we have investigated ways to exploit the knowledge of an explicit basis of the null space of a symmetric positive semidefinite matrix.

We have considered consistent systems of equations, constrained systems of equations, and generalized eigenvalue problems. First of all, the knowledge of a basis of the null space of a matrix A permits us to extract a priori a maximal positive semidefinite submatrix. The rest of the matrix is redundant information and is needed neither for the solution of systems of equations nor for the eigenvalue computation. The order of the problem is reduced by the dimension of the null space. In iterative solvers it is not necessary to complement preconditioners with projections onto the complement of the null space.

It is well known that a backward stable positive semidefinite Cholesky factorization exists if the principal $r \times r$ submatrix, $r = \mathrm{rank}(A)$, is well conditioned. This does not, however, mean that the computed Cholesky factor $\tilde{R}$ has a null space that is close to the known null space of R, $A = R^T R$. We observed that the backward error in the null space is small if the error in the Cholesky factor is (almost) orthogonal to the null space of A. We show that this is the case if the positive definite principal $r \times r$ submatrix after scaling is well conditioned and if its diagonal elements dominate those of the remaining diagonal block.

For systems of equations and eigenvalue problems, we considered the case when $A = F^T F$, where F is rectangular. This leads to interesting variants of the original algorithms and most of all leads to more accurate results.

What remains to be investigated is the relation between extraction of a positive definite matrix and fill-in during the Cholesky factorization. In future work we will use the new techniques in applications and, if possible, extend the theory to matrix classes more general than positive semidefinite ones.

REFERENCES

[1] P. Arbenz and R. Geus, *A comparison of solvers for large eigenvalue problems originating from Maxwell's equations*, Numer. Linear Algebra Appl., 6 (1999), pp. 3–16.

[2] J. H. Argyris and O. E. Brønlund, *The natural factor formulation of the stiffness for the matrix displacement method*, Comput. Methods Appl. Mech. Engrg., 5 (1975), pp. 97–119.

[3] M. Arioli, *The use of QR factorization in sparse quadratic programming and backward error issues*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 825–839.

[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978.

[6] J. Demmel, *On Floating Point Errors in Cholesky*, Tech. Report CS-89-87, Computer Science Department, University of Tennessee, Knoxville, TN, 1989; also available as LAPACK Working Note 14 from http://www.netlib.org/lapack/lawns/.

[7] Z. Drmač, *A tangent algorithm for computing the generalized singular value decomposition*, SIAM J. Numer. Anal., 35 (1998), pp. 1804–1832.

[8] Z. Drmač, M. Omladič, and K. Veselić, *On the perturbation of the Cholesky factorization*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1319–1332.

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

[10] A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[11] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, New York, 1981.

[12] G. H. Golub and C. Greif, *On solving block-structured indefinite linear systems*, SIAM J. Sci. Comput., to appear.

[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, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[15] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
[16] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 150–164

VARIANTS OF THE GREVILLE FORMULA WITH APPLICATIONS TO EXACT RECURSIVE LEAST SQUARES*

JIE ZHOU†, YUNMIN ZHU†, X. RONG LI‡, AND ZHISHENG YOU§

Abstract. In this paper, we present an order-recursive formula for the pseudoinverse of a matrix. It is a variant of the well-known Greville [*SIAM Rev.*, 2 (1960), pp. 578–619] formula. Three forms of the proposed formula are presented for three different matrix structures. Compared with the original Greville formula, the proposed formulas have certain merits. For example, they reduce the storage requirements at each recursion by almost half; they are more convenient for deriving recursive solutions for optimization problems involving pseudoinverses.

Regarding applications, using the new formulas, we derive recursive least squares (RLS) procedures which coincide exactly with the batch LS solutions to the problems of the unconstrained LS, weighted LS, and LS with linear equality constraints, respectively, including their simple and exact initializations. Compared with previous results, e.g., Albert and Sittler [*J. Soc. Indust. Appl. Math. Ser. A Control*, 3 (1965), pp. 384–417], our derivation of the explicit recursive formulas is much easier, and the recursions take a much simpler form. New findings include that the linear equality constrained LS and the unconstrained LS can have an identical recursion—their only difference is the initial conditions. In addition, some robustness issues, in particular, during the exact initialization of the RLS are studied.

Key words. Moore–Penrose generalized inverse, recursive least squares

AMS subject classifications. 65F20, 93E24

PII. S0895479801388194

1. Introduction.

Matrix pseudoinverses (Moore–Penrose generalized inverses) are often involved in the optimal solutions of various scientific and engineering problems. Their computation involves an increasing number of variables with a corresponding increase in the matrix order. To find a recursive version of such an optimal solution, a key technique is an order-recursive version of the pseudoinverse of a matrix.

For instance, consider the following minimization problem:

$$\min_{\theta} S_N = \sum_{i=1}^{N} |y_i - \mathbf{x}_i^* \theta|^2, \tag{1.1}$$

where $y_i \in \mathbb{C}^1$, $\mathbf{x}_i \in \mathbb{C}^r$, and the parameter to be estimated $\theta \in \mathbb{C}^r$. Here, $\mathbb{C}^1$ and $\mathbb{C}^r$ denote the spaces of the complex numbers and r-dimensional complex vectors, respectively. The superscript "$*$" stands for complex conjugate transpose. More generally, consider the problem of minimizing the objective function S_N in (1.1) subject to a linear equality constraint $A\theta = B$. The former is a special case of the latter with $A = 0$ and $B = 0$.

*Received by the editors April 20, 2001; accepted for publication (in revised form) by A. H. Sayed February 14, 2002; published electronically July 1, 2002. This work was supported in part by the National Key Project (grant 970211017) and the NNSF of China (grant 60074017).

http://www.siam.org/journals/simax/24-1/38819.html

†College of Mathematics, Sichuan University, Chengdu, Sichuan 610064, People's Republic of China (zhoujie68@263.net, ymzhu@scu.edu.cn).

‡Department of Electrical Engineering, University of New Orleans, New Orleans, LA 70148 (xli@uno.edu). The work of this author was supported in part by the ONR via grant N00014-00-1-0677 and the NSF via grant ECS-9734285.

§Department of Computer Science, Sichuan University, Chengdu, Sichuan 610064, People's Republic of China. The work of this author was supported in part by the NNSF of China (grant 69732010).

Such optimization problems can be found in various practical fields, including signal processing, control, and communications, to name a few.

Denote by θ_N the optimal solution in the above sense for θ using the data y_i and $\mathbf{x}_i$, $i = 1, 2, \ldots, N$. We call θ_N the least squares (LS) solution of θ. The unconstrained solution of θ has been known for two centuries, dating back to Gauss and Legendre.

Denote

$$Y_N = (y_1^* \quad y_2^* \quad \cdots \quad y_N^*)^* \in \mathbb{C}^{N\times 1}, \quad X_N = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_N)^* \in \mathbb{C}^{N\times r},$$

and rewrite S_N in (1.1) as

$$S_N = (Y_N - X_N\theta)^*(Y_N - X_N\theta).$$

Then it is well known that when $X_N^*X_N$ is nonsingular, (1.1) has a unique solution, given by

$$\theta_N = X_N^+Y_N = (X_N^*X_N)^{-1}X_N^*Y_N, \tag{1.2}$$

where A^+ denotes the pseudoinverse of A. When $X_N^*X_N$ is singular, the solution is not unique, the class of solutions is given by

$$\theta_N = X_N^+Y_N + (I - X_N^+X_N)\xi,$$

where ξ is any vector in $\mathbb{C}^r$, and the (unique) minimum-norm solution is

$$\theta_N = X_N^+Y_N. \tag{1.3}$$

We call both (1.2) and (1.3) *batch LS solutions*.

Half a century ago, an important progress on the studies of the LS method was made by Plackett [16] and others (such as Woodbury [24]), who demonstrated that when $X_{N_0}^*X_{N_0}$ is nonsingular, θ_N ($\forall N \geq N_0$) in (1.2) can be written recursively as

$$\theta_{N+1} = \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^*\theta_N), \tag{1.4}$$

$$K_{N+1} = P_N\mathbf{x}_{N+1}/(1 + \mathbf{x}_{N+1}^*P_N\mathbf{x}_{N+1}), \tag{1.5}$$

$$\begin{aligned} P_{N+1} &\overset{\text{def}}{=} (X_{N+1}^*X_{N+1})^{-1} \\ &= (P_N^{-1} + \mathbf{x}_{N+1}\mathbf{x}_{N+1}^*)^{-1} \\ &= (I - P_N\mathbf{x}_{N+1}\mathbf{x}_{N+1}^*/(1 + \mathbf{x}_{N+1}^*P_N\mathbf{x}_{N+1}))P_N \\ &= (I - K_{N+1}\mathbf{x}_{N+1}^*)P_N, \end{aligned} \tag{1.6}$$

where the third equality in (1.6) follows from the matrix inversion lemma (see [16] or [24]):

$$(A + BD^{-1}B^*)^{-1} = A^{-1} - A^{-1}B(D + B^*A^{-1}B)^{-1}B^*A^{-1}.$$

Here A and D are both Hermitian positive definite matrices.

This recursive least squares (RLS) solution greatly promotes the application of the LS method in many fields where real-time processing is required (cf. [5, 7, 9, 14]). Two significant advantages of the recursive solution (1.4)–(1.6) are (i) it is free of the matrix inverse operation and has a lower computational complexity; (ii) it is particularly suitable for real-time applications since the number of algebraic operations and

required memory locations at each iteration is fixed, rather than increases with N as the batch LS solution (1.2) or (1.3) does.

Although the RLS solution (1.4)–(1.6) has the above advantages, it could only be started when $X_{N_0}^* X_{N_0}$ is nonsingular. Note that $X_N^* X_N = \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^*$ cannot be nonsingular for any $N < r$.

To start the RLS from $N = 1$, Albert and Sittler in [1] discussed various properties of the limit of the following function, which in fact is the pseudoinverse of matrix H:

$$\lim_{\epsilon \to 0+} (H^* H + \epsilon I)^{-1} H^*.$$

Using these properties, they derived the unconstrained, linear equality constrained, and weighted RLS formulas exactly equal to the corresponding (minimum-norm) batch LS solution for $N = 1, 2, \ldots$. However, the derivation and the final recursive formulas presented are much more complicated than those of this paper.

In this paper, we present several modified order-recursive formulas of matrix pseudoinverses based on the Greville formula. Not only do the proposed formulas reduce the required memory locations of the Greville formula at each recursion by almost half, but they are also very useful to derive the recursive formulas for the optimal solutions involving matrix pseudoinverses. As applications, the proposed formulas are used in a straightforward way to derive the unconstrained, linear equality constrained, and weighted RLS procedures which coincide exactly with the corresponding unique batch LS solution (or the unique minimum-norm batch LS solution if more than one LS solution exists). In comparison with previous results of Albert and Sittler [1], not only is the derivation of the recursive formulas much easier, but the formulas themselves are also clearer and simpler. In particular, our results show that the linear equality constrained RLS can have the same recursion as that of the unconstrained RLS—they differ only in the initial values. This new finding has important practical implications. We expect to find more applications of the new order-recursive formulas in the future.

In the previous works on the exactly initialized RLS, the recursive QR decomposition method, which was described in detail in Haykin's book [9], can survive an exact start without resorting to special modifications of the algorithm and can also handle singular data well, provided the error sequence $\{y_i - \mathbf{x}_i^* \theta\}$ is of interest rather than the parameter vector θ itself, as is the case in certain signal processing applications such as echo cancellation or noise cancellation. Hubing and Alexander in [10] gave a very good statistical analysis of the parameter vector θ obtained using an exact initialization. It is worth noting that what they considered above are the initialization problems of the adaptive filtering; therefore, the first r data matrices have a special form, i.e., $\mathbf{x}_0^* = (0 \cdots 0), \mathbf{x}_1^* = (x^{(1)}\ 0 \cdots 0), \mathbf{x}_2^* = (x^{(2)}\ x^{(1)}\ 0 \cdots 0), \ldots, \mathbf{x}_r^* = (x^{(r)}\ x^{(r-1)} \cdots x^{(1)})$. This feature was also pointed out in Haykin's book (see [9, p. 519]). However, the initial data matrices considered in this paper are arbitrarily general. In addition, since there exist many results on the robustness issues of the RLS after the data matrix becomes full column rank (for example, see [4, 13, 15, 17, 22]), in this paper, we only derived the results on the error propagation and accumulation caused by the error of Q_N before X_N becomes full column rank. As for the robustness issues on P_N, it is still an open question.

The rest of this paper is organized as follows. In section 2, three forms of the proposed variant of the Greville's order-recursive formula, along with several corollaries, are presented. Then we apply the new formulas to derive the exact RLS, exact RLS with linear equality constraint, and exact weighted RLS in sections 3, 4, and 5,

respectively. In section 6, we discuss some robustness issues for the exactly initialized RLS. Finally, in section 7, we provide concluding remarks.

2. Order-recursive formulas for matrix pseudoinverses. Consider a matrix sequence $\{X_N\}_{N=1,2,\ldots}$, where $X_N \in \mathbb{C}^{N\times r}, X_{N+1} = (X_N^* \quad \mathbf{x}_{N+1})^*$, and $\mathbf{x}_{N+1}$ is an r-dimensional column vector, i.e., $\mathbf{x}_n^*$ is the nth row of X_N for any $n \le N$.

An order recursive formula was given by Greville in [8] as follows.

THEOREM 2.1 (Greville [8]). *For any* $N = 1, 2, \ldots,$

$$X_{N+1}^+ = (X_N^+ - K_{N+1}d_{N+1} \quad K_{N+1}), \tag{2.1}$$

where

$$\begin{aligned} d_{N+1} &= \mathbf{x}_{N+1}^* X_N^+, \\ c_{N+1} &= \mathbf{x}_{N+1}^* - d_{N+1}X_N, \\ K_{N+1} &= \begin{cases} c_{N+1}^+ & \text{if} \quad c_{N+1} \neq 0, \\ X_N^+ d_{N+1}^*/(1 + d_{N+1}d_{N+1}^*) & \text{if} \quad c_{N+1} = 0. \end{cases} \end{aligned} \tag{2.2}$$

Remark 2.1. The above formula is the complex conjugate transpose of the Greville formula in its original form. While the two versions are equivalent, this version fits our formulation of the problem better. Note also that both X_N^+ and X_N are used in the formula.

Using this recursive formula, we can compute the pseudoinverse X_N^+ of a high-dimensional matrix X_N from vector $(\mathbf{x}_1^*)^+$ recursively so as to eliminate the need to compute the pseudoinverse of a high-dimensional matrix. It is, however, not in a form handy for deriving the recursive versions of the optimal solutions involving the above matrix pseudoinverse. In light of this, we prove the following variant, which can be viewed as an improvement of the Greville formula.

THEOREM 2.2. *For any* $N = 0, 1, \ldots,$

$$\begin{aligned} X_1^+ &= K_1, \\ X_{N+1}^+ &= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} X_N^+ & 0 \\ 0 & 1 \end{pmatrix}, \quad N \ge 1, \end{aligned} \tag{2.3}$$

where K_{N+1} *is defined by the following:*

(i) *When* $\mathbf{x}_{N+1}^* Q_N = 0$,

$$K_{N+1} = P_N\mathbf{x}_{N+1}/(1 + \mathbf{x}_{N+1}^* P_N \mathbf{x}_{N+1}), \tag{2.4}$$

$$P_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*)P_N, \tag{2.5}$$

$$Q_{N+1} = Q_N; \tag{2.6}$$

(ii) *if* $\mathbf{x}_{N+1}^* Q_N \neq 0$,

$$K_{N+1} = Q_N\mathbf{x}_{N+1}/(\mathbf{x}_{N+1}^* Q_N \mathbf{x}_{N+1}), \tag{2.7}$$

$$P_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*)P_N(I - K_{N+1}\mathbf{x}_{N+1}^*)^* + K_{N+1}K_{N+1}^*, \tag{2.8}$$

$$Q_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*)Q_N; \tag{2.9}$$

and the initial values are

$$P_0 = 0, \quad Q_0 = I.$$

Proof. Denote $X_0 = 0$. For any $N = 0, 1, \ldots,$ let

$$P_N = X_N^+ (X_N^+)^*, \quad Q_N = I - X_N^+ X_N.$$

Clearly, P_N is Hermitian, Q_N is an orthogonal projection onto the orthogonal complement of the row space of X_N. We will show that P_N and Q_N satisfy (2.4)–(2.9).

Defining $K_1 = (\mathbf{x}_1^*)^+$, we have $X_1^+ = K_1$. Noticing that $0^+ = 0$ and $a^+ = a^*/(aa^*)$ for any nonzero row vector a, we can easily prove that K_1, P_1, Q_1 satisfy (2.4)–(2.9).

For $N \geq 1$, let

$$d_{N+1} = \mathbf{x}_{N+1}^* X_N^+, \quad c_{N+1} = \mathbf{x}_{N+1}^* - d_{N+1} X_N = \mathbf{x}_{N+1}^* Q_N.$$

Define K_{N+1} as in Theorem 2.1. Then,

$$\begin{aligned} X_{N+1}^+ &= (X_N^+ - K_{N+1} d_{N+1} \quad K_{N+1}) \\ &= (X_N^+ - K_{N+1} \mathbf{x}_{N+1}^* X_N^+ \quad K_{N+1}) \\ &= ((I - K_{N+1} \mathbf{x}_{N+1}^*) X_N^+ \quad K_{N+1}). \end{aligned}$$

From the above equation and the definitions of P_N, Q_N, we have

$$\begin{aligned} P_{N+1} &= X_{N+1}^+ (X_{N+1}^+)^* \\ &= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} X_N^+ & 0 \\ 0 & 1 \end{pmatrix} \\ &\quad \times \begin{pmatrix} (X_N^+)^* & 0 \\ 0 & 1 \end{pmatrix} (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1})^* \\ &= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} P_N & 0 \\ 0 & 1 \end{pmatrix} (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1})^* \\ &= (I - K_{N+1}\mathbf{x}_{N+1}^*) P_N (I - K_{N+1}\mathbf{x}_{N+1}^*)^* + K_{N+1} K_{N+1}^*, \end{aligned} \tag{2.10}$$

$$\begin{aligned} Q_{N+1} &= I - X_{N+1}^+ X_{N+1} \\ &= I - (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} X_N^+ & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_N \\ \mathbf{x}_{N+1}^* \end{pmatrix} \\ &= I - (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} X_N^+ X_N \\ \mathbf{x}_{N+1}^* \end{pmatrix} \\ &= I - (I - K_{N+1}\mathbf{x}_{N+1}^*)(I - Q_N) - K_{N+1}\mathbf{x}_{N+1}^* \\ &= (I - K_{N+1}\mathbf{x}_{N+1}^*) Q_N. \end{aligned} \tag{2.11}$$

When $c_{N+1} \neq 0$, by Theorem 2.1, we have

$$K_{N+1} = c_{N+1}^+ = Q_N \mathbf{x}_{N+1} / (\mathbf{x}_{N+1}^* Q_N \mathbf{x}_{N+1}).$$

If $c_{N+1} = 0$, by (2.11) and Theorem 2.1, we have

$$Q_{N+1} = Q_N - K_{N+1}\mathbf{x}_{N+1}^* Q_N = Q_N$$

and

$$K_{N+1} = P_N \mathbf{x}_{N+1} / (1 + \mathbf{x}_{N+1}^* P_N \mathbf{x}_{N+1}),$$

i.e.,

$$K_{N+1} + K_{N+1}\mathbf{x}_{N+1}^* P_N \mathbf{x}_{N+1} = P_N \mathbf{x}_{N+1}.$$

Then, (2.10) yields

$$\begin{aligned} P_{N+1} &= P_N - P_N\mathbf{x}_{N+1}K_{N+1}^* - K_{N+1}\mathbf{x}_{N+1}^* P_N \\ &\quad + (K_{N+1}\mathbf{x}_{N+1}^* P_N \mathbf{x}_{N+1} + K_{N+1})K_{N+1}^* \\ &= P_N - P_N\mathbf{x}_{N+1}K_{N+1}^* - K_{N+1}\mathbf{x}_{N+1}^* P_N + P_N\mathbf{x}_{N+1}K_{N+1}^* \\ &= (I - K_{N+1}\mathbf{x}_{N+1}^*)P_N. \end{aligned}$$

The theorem thus follows. □

In Theorem 2.2, $\mathbf{x}_{N+1}^* Q_N \neq 0$ implies $\mathbf{x}_{N+1}^*(\mathbf{x}_{N+1}^* Q_N)^+ = 1$, and (2.10)–(2.11) always hold. Hence, the two cases with $\mathbf{x}_{N+1}^* Q_N = 0$ and $\mathbf{x}_{N+1}^* Q_N \neq 0$ in Theorem 2.2 can be combined.

COROLLARY 2.3. *For any* $N = 0, 1, \ldots,$

$$\begin{aligned} X_1^+ &= K_1, \\ X_{N+1}^+ &= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} X_N^+ & 0 \\ 0 & 1 \end{pmatrix}, \quad N \geq 1, \end{aligned} \tag{2.12}$$

where

$$\begin{aligned} K_{N+1} &= (\mathbf{x}_{N+1}^* Q_N)^+ + (1 - \mathbf{x}_{N+1}^*(\mathbf{x}_{N+1}^* Q_N)^+)P_N\mathbf{x}_{N+1}/(1 + \mathbf{x}_{N+1}^* P_N \mathbf{x}_{N+1}), \\ P_{N+1} &= (I - K_{N+1}\mathbf{x}_{N+1}^*)P_N(I - K_{N+1}\mathbf{x}_{N+1}^*)^* + K_{N+1}K_{N+1}^*, \\ Q_{N+1} &= (I - K_{N+1}\mathbf{x}_{N+1}^*)Q_N, \end{aligned}$$

and the initial values are

$$P_0 = 0, \quad Q_0 = I.$$

Remark 2.2. $c_{N+1} = \mathbf{x}_{N+1}^* Q_N = 0$ if and only if $\mathbf{x}_{N+1}$ is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ because

$$Q_N\mathbf{x}_{N+1} = 0 \Longleftrightarrow \mathbf{x}_{N+1} \in \mathcal{N}(Q_N) = \mathcal{N}(I - X_N^+ X_N) = \mathcal{R}(X_N^+ X_N) = \mathcal{R}(X_N^*),$$

where $\mathcal{R}(A)$ and $\mathcal{N}(A)$ denote the range and null space of A, respectively. Hence, we have the following corollaries.

COROLLARY 2.4. *If* X_M *has full column rank (i.e.,* $X_M^* X_M$ *is nonsingular), then for any* $N \geq M$*, a recursion of* X_{N+1}^+ *is* (2.3), (2.4), *and* (2.5), *which includes* (1.5)–(1.6) *as a special case of Theorem* 2.2 *when* $\mathbf{x}_{N+1}^* Q_N = 0$.

COROLLARY 2.5. *If* $\mathbf{x}_1, \ldots, \mathbf{x}_M$ $(M = 1, \ldots, r)$ *are linearly independent, then for any* $N = 1, \ldots, r$*, a recursion of* X_N^+ *is* (2.3), (2.7), *and* (2.9).

Theorem 2.2 has certain advantages over Theorem 2.1. First, albeit a simple variant of (2.1), (2.3) is in a form more convenient to use, as demonstrated later in the derivations of recursive LS solutions. Further, (2.4)–(2.9) is much more efficient than (2.2) since they do not involve X_N directly. More specifically, since matrices P_N, K_N, and Q_N have fixed dimensions as N increases, Theorem 2.2 reduces the required memory locations of the Greville formula at each recursion by almost half when N is large.

Theorem 2.2 can be extended to the following more general version, to be used to derive the RLS with linear equality constraints.

THEOREM 2.6. *Let P be an orthogonal projection. For any $N = 0, 1, \ldots,$*

$$\begin{aligned} (X_1P)^+ &= K_1, \\ (X_{N+1}P)^+ &= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} (X_NP)^+ & 0 \\ 0 & 1 \end{pmatrix}, \quad N \geq 1, \end{aligned} \tag{2.13}$$

where K_{N+1} and the corresponding P_{N+1}, Q_{N+1} have the same recursion (2.4)–(2.9) *as given in Theorem* 2.2 *but with initial values*

$$P_0 = 0, \quad Q_0 = P.$$

Proof. Denote $X_0 = 0$. For any $N = 0, 1, \ldots,$ let

$$P_N = (X_NP)^+((X_NP)^+)^*, \quad \bar{Q}_N = I - (X_NP)^+(X_NP), \quad Q_N = P\bar{Q}_N.$$

Define $K_1 = (\mathbf{x}_{N+1}^*P)^+$. It is clear that $(X_1P)^+ = K_1$, and K_1, P_1, Q_1 satisfy (2.4)–(2.9). For $N \geq 1$, from the following property of pseudoinverse

$$(X_NP)^+ = (X_NP)^*((X_NP)(X_NP)^*)^+$$

and the definition of P, we have

$$P(X_NP)^+ = (X_NP)^+, \tag{2.14}$$

$$PP_N = P_NP = P_N, \tag{2.15}$$

$$P\bar{Q}_N = \bar{Q}_NP = Q_N. \tag{2.16}$$

To prove this theorem, we only need to use $\mathbf{x}_{N+1}^*P$ and X_NP to replace $\mathbf{x}_{N+1}^*$ and X_N in Theorem 2.2, respectively, and to define K_{N+1} according to (2.4) and (2.7). Note that $(\mathbf{x}_{N+1}^*P)\bar{Q}_N = \mathbf{x}_{N+1}^*Q_N$ from the definition of Q_N.

Using (2.14), we have

$$(I - K_{N+1}\mathbf{x}_{N+1}^*P)(X_NP)^+ = (I - K_{N+1}\mathbf{x}_{N+1}^*)(X_NP)^+,$$

and therefore (2.13) holds.

When $\mathbf{x}_{N+1}^*Q_N = 0$, (2.4)–(2.5) become

$$K_{N+1} = P_N(P\mathbf{x}_{N+1})/(1 + \mathbf{x}_{N+1}^*PP_NP\mathbf{x}_{N+1}), \tag{2.17}$$

$$P_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*P)P_N. \tag{2.18}$$

Equation (2.15) implies that (2.17) and (2.18) reduce to (2.4) and (2.5), respectively.

When $\mathbf{x}_{N+1}^*Q_N \neq 0$, (2.7)–(2.9) become

$$K_{N+1} = \bar{Q}_NP\mathbf{x}_{N+1}/(\mathbf{x}_{N+1}^*P\bar{Q}_NP\mathbf{x}_{N+1}), \tag{2.19}$$

$$P_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*P)P_N(I - K_{N+1}\mathbf{x}_{N+1}^*P)^* + K_{N+1}K_{N+1}^*, \tag{2.20}$$

$$\bar{Q}_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*P)\bar{Q}_N. \tag{2.21}$$

Because of (2.15) and (2.16), (2.19) and (2.20) reduce to (2.7) and (2.8), respectively. In addition,

$$Q_{N+1} = P\bar{Q}_{N+1} = P\bar{Q}_N - PK_{N+1}\mathbf{x}_{N+1}^*P\bar{Q}_N = (I - K_{N+1}\mathbf{x}_{N+1}^*)Q_N.$$

That is, (2.21) becomes (2.9). The theorem thus follows. ∎

Furthermore, to derive the solution of the weighted RLS problem, we now extend Theorem 2.2 to the pseudoinverse of $\Lambda_{N+1}X_{N+1}$, where the weight Λ_{N+1} is a diagonal matrix.

If

$$\Lambda_{N+1} = \begin{pmatrix} \Lambda_N & 0 \\ 0 & \lambda_{N+1} \end{pmatrix}, \tag{2.22}$$

where $\lambda_N > 0, N = 1, 2, \ldots, \Lambda_1 = 1$, then replacing $\mathbf{x}_n^*$ by $\lambda_n \mathbf{x}_n^* (n = 1, 2, \ldots)$ in Theorem 2.2, we can reduce the sought-after pseudoinverse to the pseudoinverse in Theorem 2.2 and derive the corresponding weighted RLS easily. More interestingly, consider the following forgetting-factor weighting matrix Λ_N:

$$\Lambda_{N+1} = \begin{pmatrix} \lambda_N \Lambda_N & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.23}$$

where $0 < \lambda_N \leq 1, N = 1, 2, \ldots, \Lambda_1 = 1$. Accordingly, we consider the pseudoinverse of matrix

$$\Lambda_{N+1}X_{N+1} = \begin{pmatrix} \lambda_N \Lambda_N & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_N \\ \mathbf{x}_{N+1}^* \end{pmatrix} = \begin{pmatrix} \lambda_N \Lambda_N X_N \\ \mathbf{x}_{N+1}^* \end{pmatrix}.$$

With this weight, the last (Nth) row vector of $\Lambda_N X_N$ for any N is always $\mathbf{x}_N^*$; for every $n < N$, the nth row vector of $\Lambda_N X_N$ is $(\prod_{i=n}^{N-1} \lambda_i)\mathbf{x}_n^*$. When $\lambda_n \equiv \lambda < 1$ for every $n < N$, it is the well-known exponentially decaying forgetting factor. Clearly, if $\lambda_n \equiv 1$ for every n, the weighted matrix becomes the original matrix without a weight.

THEOREM 2.7. *For any* $N = 0, 1, \ldots,$

$$\begin{aligned} &(\Lambda_1 X_1)^+ = K_1, \\ &(\Lambda_{N+1}X_{N+1})^+ = (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} \lambda_N^{-1}(\Lambda_N X_N)^+ & 0 \\ 0 & 1 \end{pmatrix}, \quad N \geq 1, \end{aligned} \tag{2.24}$$

where K_{N+1} *is defined by the following:*

(i) *When* $\mathbf{x}_{N+1}^* Q_N = 0$,

$$K_{N+1} = P_N \mathbf{x}_{N+1} / (\lambda_N^2 + \mathbf{x}_{N+1}^* P_N \mathbf{x}_{N+1}), \tag{2.25}$$

$$P_{N+1} = \lambda_N^{-2}(I - K_{N+1}\mathbf{x}_{N+1}^*)P_N, \tag{2.26}$$

$$Q_{N+1} = Q_N; \tag{2.27}$$

(ii) *if* $\mathbf{x}_{N+1}^* Q_N \neq 0$,

$$K_{N+1} = Q_N \mathbf{x}_{N+1} / (\mathbf{x}_{N+1}^* Q_N \mathbf{x}_{N+1}), \tag{2.28}$$

$$P_{N+1} = \lambda_N^{-2}(I - K_{N+1}\mathbf{x}_{N+1}^*)P_N(I - K_{N+1}\mathbf{x}_{N+1}^*)^* + K_{N+1}K_{N+1}^*, \tag{2.29}$$

$$Q_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*)Q_N; \tag{2.30}$$

and the initial values are

$$P_0 = 0, \quad Q_0 = I, \quad \lambda_0 > 0.$$

Proof. Denote $X_0 = 0$ and $\Lambda_0 = 1$. For any $N = 0, 1, \ldots,$ let

$$P_N = (\Lambda_N X_N)^+ ((\Lambda_N X_N)^+)^*, \quad Q_N = I - (\Lambda_N X_N)^+ (\Lambda_N X_N).$$

Note that Λ_N is a diagonal matrix, P_N is Hermitian, and Q_N is an orthogonal projection onto the orthogonal complement of the row space of $\Lambda_N X_N$.

Define $K_1 = (\mathbf{x}_1^*)^+$. It is clear that $(\Lambda_1 X_1)^+ = K_1$, and K_1, P_1, Q_1 satisfy (2.25)–(2.30). For $N \geq 1$, using $\Lambda_{N+1}X_{N+1}$ and $\lambda_N \Lambda_N X_N$ to replace X_{N+1} and X_N in Theorem 2.2, respectively, and defining K_{N+1} as in Theorem 2.1, we can prove that (2.24)–(2.30) hold by the same method as Theorem 2.2. □

Obviously, Theorem 2.2 is a special case of Theorem 2.7 with $\lambda_N \equiv 1$ for every $N > 0$.

As an application of Theorems 2.2, 2.6, and 2.7, we derive the RLS procedures that coincide exactly with the unique batch LS solutions (or the unique minimum-norm batch LS solutions if more than one LS solution exists) of the unconstrained problem, linear equality constrained problem, and weighted LS problem, respectively. It will be clear that the derivation is strikingly simple.

3. Exact RLS without constraint. From Theorem 2.2, we can derive directly the recursive form of the solution of the unconstrained LS problem (1.1).

THEOREM 3.1. *The batch LS solution given by* (1.2) *or* (1.3) *can always be written in the following recursive form:*

$$\theta_{N+1} = \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^* \theta_N), \quad N = 0, 1, \ldots, \tag{3.1}$$

where K_{N+1} *and the corresponding* P_{N+1}, Q_{N+1} *are given in Theorem* 2.2*, and the initial values are*

$$\theta_0 = 0, \quad P_0 = 0, \quad Q_0 = I.$$

Proof. When $N = 0$, (3.1) holds since $K_1 = X_1^+$. For $N \geq 1$, using (2.3), (1.2), and (1.3), we have

$$\begin{aligned}
\theta_{N+1} &= X_{N+1}^+ Y_{N+1} \\
&= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} X_N^+ & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Y_N \\ y_{N+1} \end{pmatrix} \\
&= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} \theta_N \\ y_{N+1} \end{pmatrix} \\
&= \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^* \theta_N).
\end{aligned}$$

The theorem follows. □

By Corollary 2.4, we have the following corollary.

COROLLARY 3.2. *When* X_N *has full column rank, a recursion of* θ_n $(n > N)$ *is* (3.1), (2.4), *and* (2.5), *which is the same as* (1.4)–(1.6).

Remark 3.1. Theorem 3.1 includes an exact and simplest possible initialization of the RLS algorithm.

4. Exact RLS with linear equality constraint. Consider the LS problem in (1.1) with the following linear equality constraint:

$$A\theta = B, \tag{4.1}$$

where $A \in \mathbb{C}^{M \times r}$ and $B \in \mathbb{C}^{M \times 1}$ $(M > 0)$. Denote the projector

$$P = I - A^+ A.$$

It is well known that if $(A^* \quad X_N^*)^*$ has full column rank (see [25]), then

$$\begin{aligned}\theta_N &= A^+B + (PX_N^*X_NP)^{-1}X_N^*(Y_N - X_NA^+B)\\ &= A^+B + (X_NP)^+(Y_N - X_NA^+B)\end{aligned} \tag{4.2}$$

is the unique solution to the LS problem (1.1) subject to (4.1). When $(A^* \quad X_N^*)^*$ does not have full column rank, the solution is not unique, and the class of the solutions is

$$\theta_N = A^+B + (X_NP)^+(Y_N - X_NA^+B) + P\xi,$$

where ξ is any vector satisfying $X_NP\xi = 0$ in $\mathbb{C}^r$. The minimum-norm solution is

$$\theta_N = A^+B + (X_NP)^+(Y_N - X_NA^+B). \tag{4.3}$$

We call both (4.2) and (4.3) *batch LS solutions* (with linear equality constraints).

Similar to Theorem 3.1, we have the following.

THEOREM 4.1. *The (minimum-norm) batch LS solution given by* (4.2) *or* (4.3) *can be written exactly in the following recursive form:*

$$\theta_{N+1} = \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^*\theta_N), \quad N = 0, 1, \ldots, \tag{4.4}$$

where K_{N+1} and the corresponding P_{N+1}, Q_{N+1} are defined in Theorem 2.6 *(but their recursive formulas are given in Theorem* 2.2*). The initial values are*

$$\theta_0 = A^+B, \quad P_0 = 0, \quad Q_0 = P.$$

Proof. When $N = 0$, (4.4) holds since $K_1 = (X_1P)^+$. For $N > 0$, noticing that P is an orthogonal projection, and using (2.13), (4.2), and (4.3), we have

$$\begin{aligned}\theta_{N+1} &= A^+B + (X_{N+1}P)^+(Y_{N+1} - X_{N+1}A^+B)\\ &= A^+B + (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1})\begin{pmatrix}(X_NP)^+ & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}Y_N - X_NA^+B\\ y_{N+1} - \mathbf{x}_{N+1}^*A^+B\end{pmatrix}\\ &= A^+B + (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1})\begin{pmatrix}(X_NP)^+(Y_N - X_NA^+B)\\ y_{N+1} - \mathbf{x}_{N+1}^*A^+B\end{pmatrix}\\ &= A^+B + (X_NP)^+(Y_N - X_NA^+B)\\ &\quad + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^*(A^+B + (X_NP)^+(Y_N - X_NA^+B)))\\ &= \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^*\theta_N).\end{aligned}$$

The theorem thus follows. ∎

Remark 4.1. Since the two pseudoinverses in Theorems 2.2 and 2.6 have the same recursion (but different initial values), and (4.4) and (3.1) are the same, Theorems 3.1 and 4.1 indicate that the solutions to the unconstrained LS problem and the linear equality constrained LS problem have an identical recursion; they differ only in the initial values.

COROLLARY 4.2. *When $(A^* \quad X_N^*)^*$ has full column rank, θ_n (for $n > N$) has a recursion identical to that of the unconstrained RLS* (1.4)–(1.6).

Proof. From (2.14) and the properties of pseudoinverses, we have

$$Q_NPX_N^* = P^2X_N^* - (X_NP)^+(X_NP)PX_N^* = PX_N^* - PX_N^* = 0.$$

Hence,

$$\mathcal{R}(PX_N^*) \subset \mathcal{N}(Q_N). \tag{4.5}$$

For any $\mathbf{x}_{N+1}$, $P\mathbf{x}_{N+1} \in \mathcal{R}(P(A^* \quad X_N^*)) = \mathcal{R}(PX_N^*)$ because $PA^* = 0$ and $(A^* \quad X_N^*)^*$ has full column rank. Then (4.5) implies

$$\mathbf{x}_{N+1}^* Q_N = \mathbf{x}_{N+1}^* P Q_N = (Q_N P \mathbf{x}_{N+1})^* = 0.$$

The corollary thus follows from Theorems 2.6 and 4.1. □

5. Exact weighted RLS. Consider the LS problem

$$\min_\theta S_N = (Y_N - X_N\theta)^* \Lambda_N^2 (Y_N - X_N\theta), \tag{5.1}$$

where Λ_N is defined by (2.23).

Remark 5.1. For convenience, we formulate the above weighted LS problem using as the weight Λ_N^2 rather than Λ_N, which is more common in the literature. If it is preferred to use the latter, then simply replace λ_N and Λ_N below by $\lambda_N^{1/2}$ and $\Lambda_N^{1/2}$, respectively.

It is well known that when $\Lambda_N X_N$ has full column rank,

$$\theta_N = (\Lambda_N X_N)^+ (\Lambda_N Y_N) \tag{5.2}$$

is the unique solution to the weighted LS problem (5.1); otherwise, the solution is not unique, the corresponding class of the solutions is

$$\theta_N = (\Lambda_N X_N)^+ (\Lambda_N Y_N) + (I - (\Lambda_N X_N)^+ (\Lambda_N X_N))\xi,$$

where ξ is any vector in $\mathbb{C}^r$, and the minimum-norm solution is

$$\theta_N = (\Lambda_N X_N)^+ (\Lambda_N Y_N). \tag{5.3}$$

We call both (5.2) and (5.3) *batch* (*weighted*) *LS solutions*.

THEOREM 5.1. *The* (*minimum-norm*) *batch LS solution given by* (5.2) *or* (5.3) *can be written exactly in the following recursive form:*

$$\theta_{N+1} = \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^* \theta_N), \quad N = 0, 1, \ldots, \tag{5.4}$$

where K_{N+1} *and the corresponding* P_{N+1}, Q_{N+1} *are given in Theorem* 2.7*, and the initial values are*

$$\theta_0 = 0, \quad P_0 = 0, \quad Q_0 = I.$$

Proof. When $N = 0$, (5.4) holds since $K_1 = X_1^+$. For $N > 0$, using (2.24) and (5.2), we have

$$\begin{aligned}
\theta_{N+1} &= (\Lambda_{N+1} X_{N+1})^+ (\Lambda_{N+1} Y_{N+1}) \\
&= \begin{pmatrix} \lambda_N \Lambda_N X_N \\ \mathbf{x}_{N+1}^* \end{pmatrix}^+ \begin{pmatrix} \lambda_N \Lambda_N Y_N \\ y_{N+1} \end{pmatrix} \\
&= (I - K_{N+1}\mathbf{x}_{N+1}^* \quad K_{N+1}) \begin{pmatrix} (\lambda_N \Lambda_N X_N)^+ & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_N \Lambda_N Y_N \\ y_{N+1} \end{pmatrix} \\
&= (I - K_{N+1}\mathbf{x}_{N+1}^*)(\lambda_N \Lambda_N X_N)^+ (\lambda_N \Lambda_N Y_N) + K_{N+1} y_{N+1} \\
&= (I - K_{N+1}\mathbf{x}_{N+1}^*)(\Lambda_N X_N)^+ (\Lambda_N Y_N) + K_{N+1} y_{N+1} \\
&= (I - K_{N+1}\mathbf{x}_{N+1}^*)\theta_N + K_{N+1} y_{N+1} \\
&= \theta_N + K_{N+1}(y_{N+1} - \mathbf{x}_{N+1}^* \theta_N).
\end{aligned}$$

The theorem thus follows. □

Similar to Corollary 3.2, we have the following.

COROLLARY 5.2. *When $\Lambda_N X_N$ has full column rank, a recursion of θ_n $(n > N)$ is* (5.4), (2.25), *and* (2.26).

Compared with previous results of Albert and Sittler [1], it is clear that not only are the derivations of the RLS formulas much easier, but the formulas themselves are also clearer and simpler. The simplicity of the recursive formulas and the almost parallel derivations enable us to identify the fact that the linear equality constrained RLS has the same recursion as the unconstrained RLS (they differ only in the initial values).

6. Robustness analysis of exactly initialized RLS.

6.1. On singularity of data matrix. In the conventional RLS (CRLS) algorithm, if the data matrix X_N has full column rank, from the following normal equation

$$(X_N^* X_N)\theta_N = X_N^* Y_N$$

we can obtain

$$\theta_N = (X_N^* X_N)^{-1} X_N^* Y_N.$$

Furthermore, we can derive the recursive formulas (1.4)–(1.6) from the recursive formula of $(X_N^* X_N)^{-1}$. However, in the numerical computations, even if X_N has full column rank, it is possible for $X_N^* X_N$ to be noninvertible. For example, we consider the following matrix:

$$X = \begin{pmatrix} 1 & 1 & 1 \\ \epsilon & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon \end{pmatrix}.$$

When ϵ is a constant close to the machine precision, $1 + \epsilon^2 \approx 1$. Thus,

$$X^* X \approx \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

becomes singular, and we cannot compute θ_N via matrix inverse as done above.

Since the derivation of the recursive formulas of θ_N in this paper is based on matrix pseudoinverses not on matrix inverses, we still can deal with the above RLS problem, as well as the recursive formulas for underdetermined systems. In [19], Stewart discussed the disturbance bound problem of matrix pseudoinverses. The necessary and sufficient condition for the continuity of the pseudoinverse of matrix A is $\operatorname{rank}(A) = \operatorname{rank}(A + E)$, where E is a disturbance for matrix A. That is to say, when a disturbance has not changed the $\operatorname{rank}(A)$, the algorithm may have robustness; otherwise, the algorithm may lose robustness.

6.2. Keep the orthogonal projection of Q_N. An issue in numerical computation of our exact RLS is to maintain the orthogonal projection of Q_N for any N before X_N becomes full column rank. For this purpose, we can modify the recursive formula of Q_N in (2.9) as

$$Q_{N+1} = (I - K_{N+1}\mathbf{x}_{N+1}^*)Q_N(I - K_{N+1}\mathbf{x}_{N+1}^*)^*.$$

6.3. Propagation of a single round-off error. Stewart in [20] studied perturbation theory of the pseudoinverse for the orthogonal projection onto the column space of a matrix, and for the linear LS problem. Van der Sluis in [18] also studied the stability of the LS solution of the linear equations. In essence, the unconstrained RLS, constrained RLS, and weighted RLS proposed in this paper are the extension of the CRLS. For the CRLS after X_N becomes full column rank, Ljung and Ljung [15], Slock [17], and Verhaegen [22] have done intensive research. They analyzed in detail the generation, propagation, and accumulation of the linear round-off error in the CRLS. Bottomley and Alexander [4] and Liavas and Regalia [13] discussed the nonlinear round-off error accumulation system of the CRLS algorithm. Since the exact RLS proposed here is also the CRLS after X_N becomes full column rank, we consider only the case before X_N becomes full column rank. For simplicity of analysis, suppose $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_r^*\}$ is linearly independent. Thus, the LS problem (1.1) has the recursive formulas (3.1), (2.7), and (2.9) no matter whether or not the exactly initialized RLS with constraints or weights is considered.

Straightforwardly using the basic results of error analysis for the LS problem by Stewart [20] and van der Sluis [18] and the round-off error made in a single recursion given by Verhaegen (see [22, Lemma 6]), as well as noting that the 2-norm of Q_N is 1, it is easy to obtain the following.

THEOREM 6.1. *Denoting the norms of absolute errors caused by round off during the construction of Q_N and θ_N by Δ_Q and Δ_θ, respectively, we have*

$$\begin{aligned} \Delta_Q &\le \epsilon_1, \\ \Delta_\theta &\le \epsilon_2(\|\theta_N\| + \|K_N\| \cdot \|y_N\|), \end{aligned}$$

where norms are 2-norms, and ϵ_i are constants close to the machine precision ϵ.

In the following, we consider the propagation of a single error at recursion instant N to subsequent recursions, assuming that no additional round-off errors are made.

Let us denote by $\tilde{x}$ the finite-precision version of x and denote by δx the round-off error in the quantity x. Then

$$\begin{aligned} \tilde{Q}_N &= Q_N + \delta Q_N, \\ \tilde{\theta}_N &= \theta_N + \delta\theta_N. \end{aligned} \tag{6.1}$$

Using the same argument given by Verhaegen, Liavas, and Regalia (e.g., see [22, Theorem 1]), it is easy to derive the following theorem.

THEOREM 6.2. *If the erroneous quantities at the recursive instant N are* (6.1)*, then these errors propagate to the next recursive instant $N+1$ as*

$$\begin{aligned} \delta Q_{N+1} &= (I - K_{N+1}\mathbf{x}_{N+1}^*)\delta Q_N (I - K_{N+1}\mathbf{x}_{N+1}^*)^* + O(\delta^2), \\ \delta\theta_{N+1} &= (I - K_{N+1}\mathbf{x}_{N+1}^*) \\ &\quad \times \left(\delta\theta_N + \frac{\delta Q_N \mathbf{x}_{N+1}}{\mathbf{x}_{N+1}^* Q_N \mathbf{x}_{N+1}}(y_{N+1} - \mathbf{x}_{N+1}^*\theta_N)\right) + O(\delta^2), \end{aligned}$$

where $O(\delta^2)$ indicates the order of magnitude of $\|\delta Q_N\|^2$.

It can be proved easily that

$$\|I - K_{N+1}\mathbf{x}_{N+1}^*\|_2 = \|K_{N+1}\mathbf{x}_{N+1}^*\|_2 = \frac{\|\mathbf{x}_{N+1}\|_2}{\|Q_N\mathbf{x}_{N+1}\|_2}.$$

Because of $\|Q_N\mathbf{x}_{N+1}\|_2 \le \|\mathbf{x}_{N+1}\|_2$, the round-off error at the instant N causes bigger round-off error at the instant $N+1$. The closer to $\mathbf{x}_{N+1}$ the projection of $\mathbf{x}_{N+1}$ onto

the orthogonal complement of $\mathcal{R}(X_N^*)$ is, the smaller the propagated round-off error at the instant N to the next recursive instant $N+1$ is.

6.4. Round-off error accumulation. By Theorem 6.2 and the recursive formulas (3.1), (2.7), and (2.9), the errors from time instant 1 to N can be given by

$$\delta Q_N = \phi(N,1)\delta Q_1 \phi(N,1)^* + O(\delta^2),$$
$$\delta\theta_N = \phi(N,1)\delta\theta_1 + \sum_{k=2}^{N-1} \phi(k,1)\mu_k + O(\delta^2),$$

where

$$\phi(k,k_0) = \prod_{i=k_0+1}^{k} (I - K_i \mathbf{x}_i^*)$$

and

$$\mu_k = \frac{\delta Q_{k-1}\mathbf{x}_k}{\mathbf{x}_k^* Q_{k-1}\mathbf{x}_k}(y_k - \mathbf{x}_k^*\theta_{k-1}).$$

For any N before X_N becomes full column rank, we have

$$\phi(N,1)(I - K_1\mathbf{x}_1^*) = \phi(N,1)Q_1 = Q_N;$$

thus

$$\|\phi(N,1)\|_2 \cdot \|Q_1\|_2 \geq \|Q_N\|_2,$$

i.e.,

$$\|\phi(N,1)\|_2 \geq 1.$$

As for the robustness issues on P_N, it is very complicated and still an open question.

7. Concluding remarks. A new order-recursive formula for the pseudoinverse of a matrix has been developed. It is an improved variant of the well-known Greville formula and reduces almost half of the required memory locations of the Greville formula at each recursion. Probably more importantly, it is in a more convenient form for deriving recursive solutions of optimization problems involving matrix pseudoinverses. Three forms of the proposed order-recursive formula have been given for three types of matrices, respectively. As applications of the proposed formulas, the unconstrained, linear equality constrained, and weighted RLS procedures that are completely equivalent to the corresponding (minimum-norm) batch LS solutions are derived in a straightforward way. It has also been shown that the linear equality constrained and unconstrained minimum-norm LS solutions have an identical recursion, with the only difference being in the initial values, a feature which has important applications. We expect that the proposed formulas will find more applications, particularly in the development of recursive algorithms. Since the robustness problems of the CRLS after X_N becomes full column rank have been studied extensively before, and once X_N becomes full column rank, the exact RLS proposed here is just the well-known CRLS, we derived the results on the error propagation and accumulation caused by the error of Q_N and θ_N before X_N becomes full column rank. As for the robustness issues on P_N, it is still an open question.

REFERENCES

[1] A. Albert and R. W. Sittler, *A method for computing least squares estimators that keep up with the data*, J. Soc. Indust. Appl. Math. Ser. A Control, 3 (1965), pp. 384–417.

[2] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.

[3] G. E. Bottomley and S. T. Alexander, *A theoretical basic for the divergence of conventional recursive least squares filters*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Glasgow, Scotland, 1989, pp. 908–911.

[4] G. E. Bottomley and S. T. Alexander, *A novel approach for stabilizing recursive least squares filters*, IEEE Trans. Signal Process., 39 (1991), pp. 1770–1779.

[5] H. F. Chen, *Recursive Estimation and Control for Stochastic Systems*, John Wiley and Sons, New York, 1985.

[6] O. L. Frost, *An algorithm for linearly constrained adaptive array processing*, Proc. IEEE, 60 (1972), pp. 926–935.

[7] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experimental Design and Data Analysis*, Academic Press, New York, 1977.

[8] T. N. E. Greville, *Some applications of the pseudoinverse of a matrix*, SIAM Rev., 2 (1960), pp. 15–22.

[9] S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1991.

[10] N. E. Hubing and S. T. Alexander, *Statistical analysis of initialization methods for RLS adaptive filters*, IEEE Trans. Signal Process., 39 (1991), pp. 1793–1804.

[11] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, SIAM, Philadelphia, PA, 1995.

[12] H. Lev-Ari, K.-F. Chiang, and T. Kailath, *Constrained-input/constrained-output stability for adaptive RLS lattice filters*, IEEE Trans. Circuit Systems, 38 (1991), pp. 1478–1483.

[13] A. P. Liavas and P. A. Regalia, *On the numerical stability and accuracy of the conventional recursive least squares algorithm*, IEEE Trans. Signal Process., 47 (1999), pp. 88–96.

[14] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[15] S. Ljung and L. Ljung, *Error propagation properties of recursive least-squares adaptation algorithms*, Automatica, 21 (1985), pp. 157–167.

[16] R. L. Plackett, *Some theorems in least squares*, Biometrika, 37 (1950), pp. 149–157.

[17] D. T. M. Slock, *Backward consistency concept and round-off error propagation dynamics in recursive least-squares algorithms*, Opt. Eng., 31 (1992), pp. 1153–1169.

[18] A. van der Sluis, *Stability of the solution of linear least squares problems*, Numer. Math., 23 (1975), pp. 241–254.

[19] G. W. Stewart, *On the continuity of the generalized inverse*, SIAM J. Appl. Math., 17 (1969), pp. 33–45.

[20] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.

[21] R. L. Streit, *Solution of systems of complex linear equations in the l_∞ norm with constraints on the unknowns*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 132–149.

[22] M. H. Verhaegen, *Round-off error propagation in four generally-applicable, recursive, least-squares estimation schemes*, Automatica, 25 (1989), pp. 437–444.

[23] M. H. Verhaegen and P. Van Dooren, *Numerical aspects of different Kalman filter implementations*, IEEE Trans. Automat. Control, 31 (1986), pp. 907–917.

[24] M. A. Woodbury, *Inverting Modified Matrices*, Memorandum Report 42, Statistical Research Group, Princeton University, Princeton, NJ, 1950.

[25] Y. M. Zhu and X. R. Li, *Recursive least squares with linear constraints*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 2414–2419.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 165–190

NUMERICAL COMPUTATION OF DEFLATING SUBSPACES OF SKEW-HAMILTONIAN/HAMILTONIAN PENCILS*

PETER BENNER†, RALPH BYERS‡, VOLKER MEHRMANN§, AND HONGGUO XU‡

Abstract. We discuss the numerical solution of structured generalized eigenvalue problems that arise from linear-quadratic optimal control problems, H_∞ optimization, multibody systems, and many other areas of applied mathematics, physics, and chemistry. The classical approach for these problems requires computing invariant and deflating subspaces of matrices and matrix pencils with Hamiltonian and/or skew-Hamiltonian structure. We extend the recently developed methods for Hamiltonian matrices to the general case of skew-Hamiltonian/Hamiltonian pencils. The algorithms circumvent problems with skew-Hamiltonian/Hamiltonian matrix pencils that lack structured Schur forms by embedding them into matrix pencils that always admit a structured Schur form. The rounding error analysis of the resulting algorithms is favorable. For the embedded matrix pencils, the algorithms use structure-preserving unitary matrix computations and are strongly backwards stable, i.e., they compute the exact structured Schur form of a nearby matrix pencil with the same structure.

Key words. eigenvalue problem, deflating subspace, Hamiltonian matrix, skew-Hamiltonian matrix, skew-Hamiltonian/Hamiltonian matrix pencil

AMS subject classifications. 49N10, 65F15, 93B40, 93B36

PII. S0895479800367439

1. Introduction and preliminaries.

In this paper we study eigenvalue and invariant subspace computations involving matrices and matrix pencils with the following algebraic structures.

DEFINITION 1.1. *Let* $\mathcal{J} := \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$, *where* I_n *is the* $n \times n$ *identity matrix.*

(a) *A matrix* $\mathcal{H} \in \mathbb{C}^{2n,2n}$ *is* Hamiltonian *if* $(\mathcal{H}\mathcal{J})^H = \mathcal{H}\mathcal{J}$. *The Lie algebra of Hamiltonian matrices in* $\mathbb{C}^{2n,2n}$ *is denoted by* $\mathbb{H}_{2n}$.

(b) *A matrix* $\mathcal{H} \in \mathbb{C}^{2n,2n}$ *is* skew-Hamiltonian *if* $(\mathcal{H}\mathcal{J})^H = -\mathcal{H}\mathcal{J}$. *The Jordan algebra of skew-Hamiltonian matrices in* $\mathbb{C}^{2n,2n}$ *is denoted by* $\mathbb{SH}_{2n}$.

(c) *If* $\mathcal{S} \in \mathbb{SH}_{2n}$ *and* $\mathcal{H} \in \mathbb{H}_{2n}$, *then* $\alpha\mathcal{S} - \beta\mathcal{H}$ *is a* skew-Hamiltonian/Hamiltonian *matrix pencil.*

(d) *A matrix* $\mathcal{Y} \in \mathbb{C}^{2n,2n}$ *is* symplectic *if* $\mathcal{Y}\mathcal{J}\mathcal{Y}^H = \mathcal{J}$. *The Lie group of symplectic matrices in* $\mathbb{C}^{2n,2n}$ *is denoted by* $\mathbb{S}_{2n}$.

(e) *A matrix* $\mathcal{U} \in \mathbb{C}^{2n,2n}$ *is* unitary symplectic *if* $\mathcal{U}\mathcal{J}\mathcal{U}^H = \mathcal{J}$ *and* $\mathcal{U}\mathcal{U}^H = I_{2n}$. *The compact Lie group of unitary symplectic matrices in* $\mathbb{C}^{2n,2n}$ *is denoted by* $\mathbb{US}_{2n}$.

(f) *A subspace* $\mathcal{L}$ *of* $\mathbb{C}^{2n}$ *is called* Lagrangian *if it has dimension* n *and* $x^H\mathcal{J}y = 0$ *for all* $x, y \in \mathcal{L}$.

*Received by the editors January 18, 2000; accepted for publication (in revised form) by D. Calvetti February 3, 2002; published electronically July 1, 2002. This work was partially supported by *Deutsche Forschungsgemeinschaft*, research grant Me 790/7-2, and Sonderforschungsbereich 393, "Numerische Simulation auf massiv parallelen Rechnern."

http://www.siam.org/journals/simax/24-1/36743.html

†Zentrum für Technomathematik, Fachbereich 3/Mathematik und Informatik, Universität Bremen, D-28334 Bremen, Germany (benner@math.uni-bremen.de).

‡Department of Mathematics, University of Kansas, Lawrence, KS 66045 (byers@math.ukans.edu, xu@math.ukans.edu). The second author was partially supported by National Science Foundation awards CCR-9732671, MRI-9977352, and by the NSF EPSCoR/K*STAR program through the Center for Advanced Scientific Computing. This work was completed while the fourth author was with the TU Chemnitz, Germany.

§Fachbereich 3 Mathematik, Technische Universität Berlin, Sekr. MA 4-5, Str. des 17. Juni 136, D-10623 Berlin, Germany (mehrmann@math.tu-berlin.de).

A matrix $\mathcal{S} \in \mathbb{C}^{2n,2n}$ is skew-Hamiltonian if and only if $i\mathcal{S}$ is Hamiltonian. Consequently, there is little difference between the structure of complex skew-Hamiltonian matrices and complex Hamiltonian matrices. However, *real* skew-Hamiltonian matrices are not *real* scalar multiples of Hamiltonian matrices, so there is a greater difference between the structure of real skew-Hamiltonian matrices and real Hamiltonian matrices.

The structures in Definition 1.1 arise typically in linear-quadratic optimal control [27, 33, 35] and H_∞ optimization [18, 39]. Moreover, instances of skew-Hamiltonian/Hamiltonian pencils appear in several other areas of applied mathematics, computational physics, and chemistry, e.g., gyroscopic systems [20], numerical simulation of elastic deformation [28, 34], and linear response theory [30]. Linear-quadratic optimal control and H_∞ optimization problems are related to skew-Hamiltonian/Hamiltonian pencils in [4, 5].

It is important to exploit and preserve algebraic structures (like symmetries in the matrix blocks or symmetries in the spectrum) as much as possible. Such algebraic structures typically arise from the physical properties of the problem. If rounding errors or other perturbations destroy the algebraic structures, then the results may be physically meaningless. Not coincidentally, numerical methods that preserve algebraic structures are typically more efficient as well as more accurate.

Despite the advantages associated with exploiting matrices with special structure, condensing data into a compact, structured matrix using finite precision arithmetic may be ill-advised. A discussion of avoiding normal-equations-like numerical instability when embedding linear-quadratic optimal control problems and H_∞ optimization problems into skew-Hamiltonian/Hamiltonian pencils appears in [4, 5].

Although the numerical computation of n-dimensional Lagrangian invariant subspaces of Hamiltonian matrices and the related problem of solving algebraic Riccati equations have been extensively studied (see [12, 22, 27, 35] and the references therein), finding completely satisfactory methods for general Hamiltonian matrices and matrix pencils remains an open problem. Completely satisfactory methods would be numerically backward stable, have complexity $\mathcal{O}(n^3)$, and preserve structure. There are several reasons for this difficulty, all of which are well demonstrated in the context of algorithms for Hamiltonian matrices. First of all, an algorithm based upon structure-preserving similarity transformations (including QR-like algorithms) would require a triangular-like Hamiltonian Schur form that displays the desired deflating subspaces. A Hamiltonian Schur form under unitary symplectic similarity transformations is presented in [31]. (See (1.1).) Unfortunately, not every Hamiltonian matrix has this kind of Hamiltonian Schur form. For example, the Hamiltonian matrix $\mathcal{J}$ in Definition 1.1 is invariant under arbitrary unitary similarity transformations but is not in the Hamiltonian Schur form described in [31]. (Similar difficulties arise in the skew-Hamiltonian/Hamiltonian pencil case for the Schur-like forms of skew-Hamiltonian/Hamiltonian matrix pencils in [25, 26] and for the other structures given in Definition 1.1 in [24].) A second problem comes from the fact that even when a Hamiltonian Schur form exists, there is no completely satisfactory structure-preserving numerical method to compute it. It has been argued in [2] that, except in special cases [13, 14], QR-like algorithms are impractically expensive because of the lack of a Hamiltonian Hessenberg–like form. For this reason other methods such as the multishift method of [1] and the structured implicit product methods of [6, 7, 38] do not follow the QR-algorithm paradigm. (The implicit product methods [6, 7] do come quite close to optimality. We extend the method of [6] to skew-Hamiltonian/Hamiltonian matrix pencils in section 4.) A third difficulty arises when the Hamiltonian matrix or

the skew-Hamiltonian/Hamiltonian matrix pencil has eigenvalues on the imaginary axis. In that case, the desired Lagrangian subspace is, in general, not unique [29]. Furthermore, if finite precision arithmetic or other errors perturb the matrix off the Lie algebra of Hamiltonian matrices, then it is typically the case that the perturbed matrix has no Lagrangian subspace or does not have the expected eigenvalue pairings; see, e.g., [7, 38].

We close the introduction by introducing some notation. To simplify notation, the term *eigenvalue* is used both for eigenvalues of matrices and, in the context of a matrix pencil $\alpha E-\beta A$, for pairs $(\alpha,\beta)\in\mathbb{C}\backslash(0,0)$ for which $\det(\alpha E-\beta A)=0$. These pairs are not unique. If $\beta\neq 0$, then we identify (α,β) with $(\alpha/\beta,1)$ and $\lambda=\alpha/\beta$. Pairs $(\alpha,0)$ with $\alpha\neq 0$ are called *infinite eigenvalues*.

By $\Lambda(E,A)$ we denote the set of eigenvalues of $\alpha E-\beta A$ including finite and infinite eigenvalues, both counted according to multiplicity. We will denote by $\Lambda_-(E,A)$, $\Lambda_0(E,A)$, and $\Lambda_+(E,A)$ the set of finite eigenvalues of $\alpha A-\beta E$ with negative, zero, and positive real parts, respectively. The set of infinite eigenvalues is denoted by $\Lambda_\infty(E,A)$. Multiple eigenvalues are repeated in $\Lambda_-(E,A)$, $\Lambda_0(E,A)$, $\Lambda_+(E,A)$, and $\Lambda_\infty(E,A)$ according to algebraic multiplicity. The set of all eigenvalues counted according to multiplicity is $\Lambda(E,A):=\Lambda_-(E,A)\cup\Lambda_0(E,A)\cup\Lambda_+(E,A)\cup\Lambda_\infty(E,A)$. Similarly, we denote by $\mathrm{Def}_-(E,A)$, $\mathrm{Def}_0(E,A)$, $\mathrm{Def}_+(E,A)$, and $\mathrm{Def}_\infty(E,A)$ the right deflating subspaces corresponding to $\Lambda_-(E,A)$, $\Lambda_0(E,A)$, $\Lambda_+(E,A)$, and $\Lambda_\infty(E,A)$, respectively.

Throughout this paper, the imaginary number $\sqrt{-1}$ is denoted by i. The inertia of a Hermitian matrix A consists of the triple $\mathrm{In}(A)=(\pi,\omega,\nu)$, where $\pi=\pi(A)$, $\omega=\omega(A)$, and $\nu=\nu(A)$ represent the number of eigenvalues with positive, zero, and negative real parts, respectively.

By abuse of notation, we identify a subspace and a matrix whose columns span this subspace by the same symbol.

We call a matrix *Hamiltonian block triangular* if it is Hamiltonian and has the form

$$\begin{bmatrix} F & G \\ 0 & -F^H \end{bmatrix}.$$

If, furthermore, F is triangular, then we call the matrix *Hamiltonian triangular*. The terms *skew-Hamiltonian block triangular* and *skew-Hamiltonian triangular* are defined analogously.

The Hamiltonian (skew-Hamiltonian) Schur form of a Hamiltonian (skew-Hamiltonian) matrix $\mathcal{H}$ is the factorization

$$\mathcal{H}=\mathcal{U}\mathcal{T}\mathcal{U}^H, \tag{1.1}$$

where $\mathcal{U}\in\mathbb{US}_{2n}$, and $\mathcal{T}$ is Hamiltonian (skew-Hamiltonian) triangular. As mentioned above, not all Hamiltonian matrices have a Hamiltonian Schur form. Real skew-Hamiltonian matrices always have one [38], but not all complex skew-Hamiltonian matrices do. For Hamiltonian matrices that have no purely imaginary eigenvalues the existence of a Hamiltonian Schur form was proved in [31]. Necessary and sufficient conditions for the existence of the Hamiltonian Schur form in the case of arbitrary spectra were suggested in [23], and a proof based on a structured Hamiltonian Jordan form was recently given in [24].

2. Schur-like forms of skew-Hamiltonian/Hamiltonian matrix pencils. In this section we derive the theoretical background for algorithms to compute eigenvalues and deflating subspaces of skew-Hamiltonian/Hamiltonian matrix pencils. A

primary theoretical and computational tool is the J-congruence. A *J-congruence transformation* of a $2n \times 2n$ pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ by a nonsingular matrix $\mathcal{Y} \in \mathbb{C}^{2n,2n}$ is the congruence transformation $\mathcal{J}\mathcal{Y}^H\mathcal{J}^T(\alpha\mathcal{S} - \beta\mathcal{H})\mathcal{Y}$, where $\mathcal{J}$ is as in Definition 1.1. The structure of skew-Hamiltonian/Hamiltonian matrix pencils is preserved by J-congruence transformations [25, 26]; i.e., if $\alpha\mathcal{S} - \beta\mathcal{H}$ is a skew-Hamiltonian/Hamiltonian pencil and $\mathcal{Y}$ is nonsingular, then $\mathcal{J}\mathcal{Y}^H\mathcal{J}^T(\alpha\mathcal{S} - \beta\mathcal{H})\mathcal{Y}$ is also skew-Hamiltonian/Hamiltonian.

The skew-Hamiltonian/Hamiltonian Schur form of a skew-Hamiltonian/Hamiltonian pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ is the factorization

$$\alpha\mathcal{S} - \beta\mathcal{H} = \mathcal{J}\mathcal{Q}\mathcal{J}^T \left(\alpha \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^H \end{bmatrix} - \beta \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^H \end{bmatrix} \right) \mathcal{Q}^H, \tag{2.1}$$

where $\mathcal{Q} \in \mathbb{C}^{2n,2n}$ is unitary, $S_{11} \in \mathbb{C}^{n,n}$ and $H_{11} \in \mathbb{C}^{n,n}$ are upper triangular, $S_{12} \in \mathbb{C}^{n,n}$ is skew-Hermitian, and $H_{12} \in \mathbb{C}^{n,n}$ is Hermitian. Note that the skew-Hamiltonian/Hamiltonian Schur form is a special case of the Schur form of a general matrix pencil and that it displays the eigenvalues and a nested system of deflating subspaces. This definition of a skew-Hamiltonian/Hamiltonian Schur form is essentially consistent with the definition of the Hamiltonian Schur form of a Hamiltonian matrix (1.1). If (2.1) holds with $\mathcal{S} = I$, then it is not difficult to show that $\mathcal{Q}$ is a unitary diagonal matrix multiple of a unitary symplectic matrix and that there is a unitary symplectic choice of $\mathcal{Q}$, $\mathcal{Q}^H = \mathcal{Q}^{-1} = J\mathcal{Q}^H J^T$, for which (2.1) holds with $S_{11} = I$ and $S_{12} = 0$.

Skew-Hamiltonian/Hamiltonian matrix pencils often have the characteristic that the skew-Hamiltonian matrix $\mathcal{S}$ is block diagonal [4, 5], i.e., $\mathcal{S} = \left[\begin{smallmatrix} E & 0 \\ 0 & E^H \end{smallmatrix}\right]$ for some matrix $E \in \mathbb{C}^{n,n}$. In this case (among others), the matrix $\mathcal{S}$ factors in the form

$$\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}, \tag{2.2}$$

where $\mathcal{Z} = \operatorname{diag}(I, E^H)$. Such a factorization may also be intrinsic to the problem formulation for nonblock diagonal skew-Hamiltonian matrices $\mathcal{S}$; see, e.g., [28].

Let $\langle x, y\rangle$ be the indefinite inner product on $\mathbb{C}^{2n} \times \mathbb{C}^{2n}$ defined by $\langle x, y\rangle = y^H\mathcal{J}x$. If $\mathcal{Z} \in \mathbb{C}^{2n,2n}$, then for all x, $y \in \mathbb{C}^{2n}$, $\langle(\mathcal{Z}x), y\rangle = \langle x, (\mathcal{J}^{-T}\mathcal{Z}^H\mathcal{J}^T)y\rangle$; i.e., the adjoint of $\mathcal{Z}$ with respect to $\langle\,.\,,\,.\,\rangle$ is $\mathcal{J}^{-T}\mathcal{Z}^H\mathcal{J}^T$. Because $\mathcal{J}^{-1} = \mathcal{J}^T = -\mathcal{J}$, the adjoint may also be expressed as $\mathcal{J}\mathcal{Z}^H\mathcal{J}^T$. From this point of view, (2.2) is a symmetric-like factorization of $\mathcal{S}$ into the product of adjoints $\mathcal{J}\mathcal{Z}\mathcal{J}^T$ and $\mathcal{Z}$. By analogy with the factorization of symmetric matrices, we will use the term *$\mathcal{J}$-semidefinite* to refer to skew-Hamiltonians matrices which have a factorization of the form (2.2). A *$\mathcal{J}$-definite* skew-Hamiltonian matrix is a skew-Hamiltonian matrix that is both $\mathcal{J}$-semidefinite and nonsingular.

The property of $\mathcal{J}$-semidefiniteness arises frequently in applications [3, 4, 5]. We show below that all real skew-Hamiltonian matrices are $\mathcal{J}$-semidefinite. We also show that if a skew-Hamiltonian/Hamiltonian matrix pencil has a skew-Hamiltonian/Hamiltonian Schur form, then the skew-Hamiltonian part is $\mathcal{J}$-semidefinite.

Although $\mathcal{J}$-semidefiniteness is a common property of skew-Hamiltonian matrices, it is not universal. The following lemma shows that neither $i\mathcal{J}$ nor any nonsingular, skew-Hamiltonian matrix of the form $i\mathcal{J}LL^T$ is $\mathcal{J}$-semidefinite.

LEMMA 2.1. *A nonsingular skew-Hamiltonian matrix $\mathcal{S}$ is $\mathcal{J}$-definite if and only if $i\mathcal{J}\mathcal{S}$ is Hermitian with n positive and n negative eigenvalues.*

Proof. If $\mathcal{S}$ is $\mathcal{J}$-definite, then $\mathcal{Z}$ in (2.2) is nonsingular and the Hermitian matrix $i\mathcal{J}\mathcal{S}$ is congruent to $-i\mathcal{J}^T = i\mathcal{J}$. It follows from Sylvester's law of inertia [16, p. 296],

[21, p. 188] that $i\mathcal{J}\mathcal{S}$ is a Hermitian matrix with n positive eigenvalues and n negative eigenvalues.

Conversely, suppose that $i\mathcal{J}\mathcal{S}$ is Hermitian with n positive and n negative eigenvalues. The matrix $i\mathcal{J}^T$ also has n positive and n negative eigenvalues, so, by an immediate consequence of Sylvester's law of inertia, there is a nonsingular matrix $\mathcal{Z} \in \mathbb{C}^{2n,2n}$ for which $i\mathcal{J}\mathcal{S} = \mathcal{Z}^H(i\mathcal{J}^T)\mathcal{Z}$. It follows that (2.2) holds with this matrix $\mathcal{Z}$. ∎

Lemma 2.1 suggests that $\mathcal{J}$-semidefiniteness might be a characteristic of the inertia of $i\mathcal{J}\mathcal{S}$. The next lemma shows that this is indeed the case.

LEMMA 2.2. *A matrix $\mathcal{S} \in \mathbb{SH}_{2n}$ is $\mathcal{J}$-semidefinite if and only if $i\mathcal{J}\mathcal{S}$ satisfies both $\pi(i\mathcal{J}\mathcal{S}) \le n$ and $\nu(i\mathcal{J}\mathcal{S}) \le n$.*

Proof. Suppose that $\mathcal{S} \in \mathbb{SH}_{2n}$ is $\mathcal{J}$-semidefinite. For some $\mathcal{Z}$ satisfying (2.2), define $\mathcal{S}(\epsilon)$ by $\mathcal{S}(\epsilon) = \mathcal{J}(Z+\epsilon I)^H\mathcal{J}^T(Z+\epsilon I)$. For ϵ small enough, $\mathcal{Z}+\epsilon I$ is nonsingular, and, by Lemma 2.1, $\pi(i\mathcal{J}\mathcal{S}(\epsilon)) = n$ and $\nu(i\mathcal{J}\mathcal{S}(\epsilon)) = n$. Because eigenvalues are continuous functions of matrix elements and $\mathcal{S} = \lim_{\epsilon\to 0}\mathcal{S}(\epsilon)$, it follows that $\pi(i\mathcal{J}\mathcal{S}) \le n$ and $\nu(i\mathcal{J}\mathcal{S}) \le n$.

For the converse, if $\pi(i\mathcal{J}\mathcal{S}) = p \le n$ and $\nu(i\mathcal{J}\mathcal{S}) = q \le n$, then there exists a nonsingular matrix $\mathcal{W}$ for which $i\mathcal{J}\mathcal{S} = \mathcal{W}^H\mathcal{L}\mathcal{W}$ with signature matrix

$$\mathcal{L} = \begin{array}{c} \\ p \\ n-p \\ q \\ n-q \end{array} \begin{array}{c} \begin{array}{cccc} p & n-p & q & n-q \end{array} \\ \begin{bmatrix} I_p & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -I_q & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

Because $p \le n$ and $q \le n$, $\mathcal{L}$ factors as $\mathcal{L} = \mathcal{L}\,\mathrm{diag}(I_n, -I_n)\mathcal{L}$, where I_n is the $n \times n$ identity matrix. The matrix $\mathrm{diag}(I_n, -I_n)$ is the diagonal matrix of eigenvalues of $i\mathcal{J}^T$, so $\mathcal{L} = \mathcal{L}(\mathcal{U}^H(i\mathcal{J}^T)\mathcal{U})\mathcal{L}$, where $\mathcal{U} = (1/\sqrt{2})\left[\begin{smallmatrix} I_n & I_n \\ iI_n & -iI_n \end{smallmatrix}\right]$ is the unitary matrix of eigenvectors of $i\mathcal{J}^T$. Hence, (2.2) holds with $\mathcal{Z} = \mathcal{U}\mathcal{L}\mathcal{W}$. ∎

The following immediate corollary also follows from [15].

COROLLARY 2.3. *Every real skew-Hamiltonian matrix $\mathcal{S}$ is $\mathcal{J}$-semidefinite.*

Proof. If $\mathcal{S}$ is real, then $\mathcal{J}\mathcal{S}$ is real and skew-symmetric. The eigenvalues of $\mathcal{J}\mathcal{S}$ appear in complex conjugate pairs with zero real part. Hence, the eigenvalues of $i\mathcal{J}\mathcal{S}$ lie on the real axis in $\pm$ pairs. In particular, $\pi(i\mathcal{J}\mathcal{S}) = \nu(i\mathcal{J}\mathcal{S})$. It follows from the trivial identity $\pi(i\mathcal{J}\mathcal{S}) + \omega(i\mathcal{J}\mathcal{S}) + \nu(i\mathcal{J}\mathcal{S}) = 2n$ that $\pi(i\mathcal{J}\mathcal{S}) \le n$ and $\nu(i\mathcal{J}\mathcal{S}) \le n$. ∎

The next lemma and its corollary show that $\mathcal{J}$-semidefiniteness of both $\mathcal{S}$ and $i\mathcal{H}$ are necessary conditions for a skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ to have a skew-Hamiltonian/Hamiltonian Schur.

LEMMA 2.4. *If $\mathcal{S} \in \mathbb{SH}_{2n}$ and there exists a nonsingular matrix $\mathcal{Y}$ such that*

$$\mathcal{J}\mathcal{Y}^H\mathcal{J}^T\mathcal{S}\mathcal{Y} = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^H \end{bmatrix}$$

with S_{11}, $S_{12} \in \mathbb{C}^{n,n}$, then $\mathcal{S}$ is $\mathcal{J}$-semidefinite.

Proof. Let $\mathcal{T}$ be the Hermitian matrix

$$\mathcal{T} = \mathcal{Y}^H(i\mathcal{J}\mathcal{S})\mathcal{Y} = \begin{bmatrix} 0 & iS_{11}^H \\ -iS_{11} & -iS_{12} \end{bmatrix},$$

and set $\mathcal{T}(\epsilon) = \mathcal{T} + \epsilon \left[\begin{smallmatrix} 0 & I_n \\ I_n & I_n \end{smallmatrix}\right]$. For ϵ sufficiently small, both $\epsilon I_n - iS_{12}$ and $\epsilon I_n - iS_{11}$ are nonsingular and $\mathcal{T}(\epsilon)$ is congruent to

$$\begin{bmatrix} -(\epsilon I_n - iS_{11})(\epsilon I_n - iS_{12})^{-1}(\epsilon I_n - iS_{11})^H & 0 \\ 0 & (\epsilon I_n - iS_{12}) \end{bmatrix}.$$

By Sylvester's law, the inertia of the negative of the $(1,1)$ block is equal to the inertia of the $(2,2)$ block. This implies $\pi(\mathcal{T}(\epsilon)) = \nu(\mathcal{T}(\epsilon)) = n$. Continuity of eigenvalues as $\epsilon \to 0$ implies $\pi(\mathcal{T}) \le n$ and $\nu(\mathcal{T}) \le n$. The assertion now follows from Lemma 2.2. ∎

COROLLARY 2.5. *If $\mathcal{H} \in \mathbb{H}_{2n}$ and there exists a nonsingular matrix $\mathcal{Y}$ such that*

$$\mathcal{J}\mathcal{Y}^H\mathcal{J}^T\mathcal{H}\mathcal{Y} = \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^H \end{bmatrix}$$

with H_{11}, $H_{12} \in \mathbb{C}^{n,n}$, then $i\mathcal{H}$ is $\mathcal{J}$-semidefinite.

Proof. Apply Lemma 2.4 to the skew-Hamiltonian matrix $i\mathcal{H}$. ∎

It follows from Lemma 2.4 and Corollary 2.5 that if $\alpha\mathcal{S} - \beta\mathcal{H}$ is a skew-Hamiltonian/Hamiltonian matrix pencil that has a skew-Hamiltonian/Hamiltonian Schur form, then $\mathcal{S}$ and $i\mathcal{H}$ are $\mathcal{J}$-semidefinite. As noted above, the factor $\mathcal{Z}$ in (2.2) is often given explicitly as part of the problem statement. It can also be obtained as in the proof of Lemma 2.2 or by a modification of Gaussian elimination [3]. The next theorem shows that if $\mathcal{S}$ is nonsingular, then the skew-Hamiltonian/Hamiltonian Schur form (if it exists) can be expressed in terms of block triangular factorizations of $\mathcal{Z}$ and $\mathcal{H}$ without explicitly using $\mathcal{S}$. This opens the possibility of designing numerical methods that work directly on $\mathcal{Z}$ and $\mathcal{H}$ and avoid the normal-equations-like numerical instability of forming $\mathcal{S}$ explicitly.

For regular skew-Hamiltonian/Hamiltonian matrix pencils, the following theorem gives necessary and sufficient conditions for the existence of a skew-Hamiltonian/Hamiltonian Schur form.

THEOREM 2.6 (see [25, 26]). *Let $\alpha\mathcal{S} - \beta\mathcal{H}$ be a regular skew-Hamiltonian/Hamiltonian matrix pencil, with ν pairwise distinct, finite, nonzero, purely imaginary eigenvalues $i\alpha_1, i\alpha_2, \ldots, i\alpha_\nu$ of algebraic multiplicity $p_1, p_2, \ldots, p_\nu$, and associated right deflating subspaces $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_\nu$. Let p_∞ be the algebraic multiplicity of the eigenvalue infinity and let $\mathcal{Q}_\infty$ be its associated deflating subspace. The following are equivalent.*

(i) *There exists a nonsingular matrix $\mathcal{Y}$ such that*

$$\mathcal{J}\mathcal{Y}^H\mathcal{J}^T(\alpha\mathcal{S} - \beta\mathcal{H})\mathcal{Y} = \alpha \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{11}^H \end{bmatrix} - \beta \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^H \end{bmatrix}, \tag{2.3}$$

where S_{11} and H_{11} are upper triangular while S_{12} is skew-Hermitian and H_{12} is Hermitian.

(ii) *There exists a unitary matrix $\mathcal{Q}$ such that $\mathcal{J}\mathcal{Q}^H\mathcal{J}^T(\alpha\mathcal{S} - \beta\mathcal{H})\mathcal{Q}$ is of the form on the right-hand side of* (2.3).

(iii) *For $k = 1, 2, \ldots, \nu$, $\mathcal{Q}_k^H\mathcal{J}\mathcal{S}\mathcal{Q}_k$ is congruent to a $p_k \times p_k$ copy of $\mathcal{J}$. (If $\nu = 0$, i.e., if $\alpha\mathcal{S} - \beta\mathcal{H}$ has no finite, nonzero, purely imaginary eigenvalue, then this statement holds vacuously.)*
Furthermore, if $p_\infty \neq 0$, then $\mathcal{Q}_\infty^H\mathcal{J}\mathcal{H}\mathcal{Q}_\infty$ is congruent to a $p_\infty \times p_\infty$ copy of $i\mathcal{J}$.

Similar results cover real Schur-like forms of real Hamiltonian matrices and skew-Hamiltonian/Hamiltonian matrix pencils [24, 25, 26].

Theorem 2.6 gives necessary and sufficient conditions for the existence of a structured triangular-like form for skew-Hamiltonian/Hamiltonian pencils. It also demonstrates that whenever a structured triangular-like form exists, then it also exists under unitary transformations. It is partly because of this fact that there exist structure-preserving, numerically stable numerical algorithms like those described here and in [4].

THEOREM 2.7. *Let $\alpha\mathcal{S} - \beta\mathcal{H}$ be a skew-Hamiltonian/Hamiltonian matrix pencil with nonsingular, $\mathcal{J}$-semidefinite skew-Hamiltonian part $\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$. If any of the equivalent conditions of Theorem* 2.6 *holds, then there exists a unitary matrix $\mathcal{Q}$ and a unitary symplectic matrix $\mathcal{U}$ such that*

$$\mathcal{U}^H\mathcal{Z}\mathcal{Q} = \begin{bmatrix} Z_{11} & Z_{12} \\ 0 & Z_{22} \end{bmatrix}, \tag{2.4}$$

$$\mathcal{J}\mathcal{Q}^H\mathcal{J}^T\mathcal{H}\mathcal{Q} = \begin{bmatrix} H_{11} & H_{12} \\ 0 & -H_{11}^H \end{bmatrix}, \tag{2.5}$$

where Z_{11}, Z_{22}^H, and H_{11} are $n \times n$ and upper triangular.

Proof. With $\mathcal{Q}$ as in Theorem 2.6(ii) we obtain (2.5) and $\mathcal{J}\mathcal{Q}^H\mathcal{J}^T\mathcal{S}\mathcal{Q} = \left[\begin{smallmatrix} S_{11} & S_{12} \\ 0 & S_{11}^H \end{smallmatrix}\right]$. Partition $\tilde{\mathcal{Z}} = \mathcal{Z}\mathcal{Q}$ as $\tilde{\mathcal{Z}} = [Z_1, Z_2]$, where $Z_1, Z_2 \in \mathbb{C}^{2n,n}$. Using $\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$, we obtain

$$\tilde{\mathcal{Z}}^H\mathcal{J}\tilde{\mathcal{Z}} = \begin{bmatrix} 0 & S_{11}^H \\ -S_{11} & -S_{12} \end{bmatrix}. \tag{2.6}$$

In particular, $Z_1^H\mathcal{J}Z_1 = 0$, i.e., the columns of Z_1 form a basis of a Lagrangian subspace, and therefore the columns of Z_1 form the first n columns of a symplectic matrix. (It is easy to verify from Definition 1.1 that using the nonnegative definite square root $[Z_1, -\mathcal{J}Z_1(Z_1^HZ_1)^{-1/2}]$ is symplectic.) It is shown in [11] that Z_1 has a unitary symplectic QR factorization

$$\mathcal{U}^HZ_1 = \begin{bmatrix} Z_{11} \\ 0 \end{bmatrix},$$

where $\mathcal{U} \in \mathbb{US}_{2n}$ is unitary symplectic and $Z_{11} \in \mathbb{C}^{n,n}$ is upper triangular. Setting

$$\mathcal{U}^H\mathcal{Z}\mathcal{Q} = \mathcal{U}^H\tilde{\mathcal{Z}} = \begin{bmatrix} Z_{11} & Z_{12} \\ 0 & Z_{22} \end{bmatrix}$$

we obtain from (2.6) that $Z_{22}^HZ_{11} = S_{11}$. Since S_{11} and Z_{11} are both upper triangular and Z_{11} is nonsingular, we conclude that Z_{22}^H is also upper triangular. ∎

Note that the invertibility of $\mathcal{Z}$ is only a sufficient condition for the existence of $\mathcal{U}$ as in (2.4) and (2.5). However, there is no particular pathology associated with $\mathcal{Z}$ being singular. The algorithms described below and in [4] do not require $\mathcal{Z}$ to be nonsingular.

If both $\mathcal{S}$ and $\mathcal{H}$ are nonsingular, then the following stronger form of Theorem 2.7 holds.

COROLLARY 2.8. *Let $\alpha\mathcal{S}-\beta\mathcal{H}$ be a skew-Hamiltonian/Hamiltonian matrix pencil with nonsingular $\mathcal{J}$-semidefinite skew-Hamiltonian part $\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$ and nonsingular $\mathcal{J}$-semidefinite Hamiltonian part $i\mathcal{H} = \mathcal{J}\mathcal{W}^H\mathcal{J}^T\mathcal{W}$. If any of the equivalent conditions of Theorem* 2.6 *holds, then there exist a unitary matrix $\mathcal{Q}$ and unitary*

symplectic matrices $\mathcal{U}$ *and* $\mathcal{V}$ *such that*

$$\mathcal{U}^H \mathcal{Z} \mathcal{Q} = \begin{bmatrix} Z_{11} & Z_{12} \\ 0 & Z_{22} \end{bmatrix}, \qquad \mathcal{V}^H \mathcal{W} \mathcal{Q} = \begin{bmatrix} W_{11} & W_{12} \\ 0 & W_{22} \end{bmatrix},$$

where Z_{11}, Z_{22}^H *and* W_{11}, W_{22}^H *are* $n \times n$ *and upper triangular.*

Proof. The proof is similar to that of Theorem 2.7. ∎

In the following we derive the theoretical background for algorithms to compute eigenvalues and deflating subspaces of skew-Hamiltonian/Hamiltonian matrix pencils.

We will obtain the structured Schur form of a complex skew-Hamiltonian/Hamiltonian matrix pencil from the structured Schur form of a real skew-Hamiltonian/skew-Hamiltonian matrix pencil of double dimension. The following theorem establishes that, in contrast to the complex skew-Hamiltonian/Hamiltonian case, every real, regular skew-Hamiltonian/skew-Hamiltonian pencil admits a structured real Schur form.

THEOREM 2.9. *If* $\alpha\mathcal{S} - \beta\mathcal{N}$ *is a real, regular skew-Hamiltonian/skew-Hamiltonian matrix pencil with* $\mathcal{S} = \mathcal{J}\mathcal{Z}^T\mathcal{J}^T\mathcal{Z}$, *then there exist a real orthogonal matrix* $\mathcal{Q} \in \mathbb{R}^{2n,2n}$ *and a real orthogonal symplectic matrix* $\mathcal{U} \in \mathbb{R}^{2n,2n}$ *such that*

$$\mathcal{U}^T \mathcal{Z} \mathcal{Q} = \begin{bmatrix} Z_{11} & Z_{12} \\ 0 & Z_{22} \end{bmatrix}, \tag{2.7}$$

$$\mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{N}\mathcal{Q} = \begin{bmatrix} N_{11} & N_{12} \\ 0 & N_{11}^T \end{bmatrix}, \tag{2.8}$$

where Z_{11} *and* Z_{22}^T *are upper triangular,* N_{11} *is quasi upper triangular, and* N_{12} *is skew-symmetric.*

Moreover,

$$\mathcal{J}\mathcal{Q}^T\mathcal{J}^T(\alpha\mathcal{S} - \beta\mathcal{N})\mathcal{Q} = \alpha \begin{bmatrix} Z_{22}^T Z_{11} & Z_{22}^T Z_{12} - Z_{12}^T Z_{22} \\ 0 & Z_{11}^T Z_{22} \end{bmatrix} - \beta \begin{bmatrix} N_{11} & N_{12} \\ 0 & N_{11}^T \end{bmatrix} \tag{2.9}$$

is a $\mathcal{J}$*-congruent skew-Hamiltonian/skew-Hamiltonian matrix pencil.*

Proof. A constructive proof for the existence of $\mathcal{Q}$ and $\mathcal{U}$ satisfying (2.7) and (2.8) is Algorithm 3 in [4]. To show (2.9), recall that $\mathcal{U}$ is orthogonal symplectic and therefore commutes with $\mathcal{J}$. Hence,

$$\begin{aligned} \mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{S}\mathcal{Q} &= \mathcal{J}\mathcal{Q}^T\mathcal{J}^T(\mathcal{J}\mathcal{Z}^T\mathcal{J}^T\mathcal{Z})\mathcal{Q} \\ &= \mathcal{J}\mathcal{Q}^T\mathcal{J}^T(\mathcal{J}\mathcal{Z}^T\mathcal{J}^T\mathcal{U})(\mathcal{U}^T\mathcal{Z}\mathcal{Q}) \\ &= \mathcal{J}(\mathcal{U}^T\mathcal{Z}\mathcal{Q})^T\mathcal{J}^T(\mathcal{U}^T\mathcal{Z}\mathcal{Q}). \end{aligned}$$

Equation (2.9) now follows from the block triangular form of (2.7). ∎

Note that this theorem does not easily extend to complex skew-Hamiltonian/skew-Hamiltonian matrix pencils.

A method for computing the structured Schur form (2.9) for real matrices was proposed in [32], but if $\mathcal{S}$ is given in factored form, then Algorithm 3 in [4] is more robust in finite precision arithmetic, because it avoids forming $\mathcal{S}$ explicitly.

Neither the method in [32] nor Algorithm 3 in [4] applies to complex skew-Hamiltonian/Hamiltonian matrix pencils because those algorithms depend on the fact that real diagonal skew-symmetric matrices are identically zero. This property is also crucial for the structured Schur form algorithms in [6, 38].

Algorithm 1 given below computes the eigenvalues of a complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ using an unusual embedding of $\mathbb{C}$ into

$\mathbb{R}^2$, which was recently proposed in [8]. Let $\alpha\mathcal{S} - \beta\mathcal{H}$ be a complex skew-Hamiltonian/Hamiltonian matrix pencil with $\mathcal{J}$-semidefinite skew-Hamiltonian part $\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$. Split the skew-Hamiltonian matrix $\mathcal{N} = i\mathcal{H} \in \mathbb{SH}_{2n}$ as $i\mathcal{H} = \mathcal{N} = \mathcal{N}_1 + i\mathcal{N}_2$, where $\mathcal{N}_1$ is real skew-Hamiltonian and $\mathcal{N}_2$ is real Hamiltonian, i.e.,

$$\mathcal{N}_1 = \begin{bmatrix} F_1 & G_1 \\ H_1 & F_1^T \end{bmatrix}, \quad G_1 = -G_1^T, \quad H_1 = -H_1^T,$$
$$\mathcal{N}_2 = \begin{bmatrix} F_2 & G_2 \\ H_2 & -F_2^T \end{bmatrix}, \quad G_2 = G_2^T, \quad H_2 = H_2^T,$$

and $F_j, G_j, H_j \in \mathbb{R}^{n\times n}$ for $j = 1, 2$. Setting

$$\mathcal{Y}_c = \frac{\sqrt{2}}{2}\begin{bmatrix} I_{2n} & iI_{2n} \\ I_{2n} & -iI_{2n} \end{bmatrix},$$
$$\mathcal{P} = \begin{bmatrix} I_n & 0 & 0 & 0 \\ 0 & 0 & I_n & 0 \\ 0 & I_n & 0 & 0 \\ 0 & 0 & 0 & I_n \end{bmatrix}, \tag{2.10}$$
$$\mathcal{X}_c = \mathcal{Y}_c\mathcal{P} \tag{2.11}$$

and using the embedding $\mathcal{B}_\mathcal{N} = \operatorname{diag}(\mathcal{N}, \bar{\mathcal{N}})$, we obtain that

$$\mathcal{B}_\mathcal{N}^c := \mathcal{X}_c^H\mathcal{B}_\mathcal{N}\mathcal{X}_c = \left[\begin{array}{cc|cc} F_1 & -F_2 & G_1 & -G_2 \\ F_2 & F_1 & G_2 & G_1 \\ \hline H_1 & -H_2 & F_1^T & F_2^T \\ H_2 & H_1 & -F_2^T & F_1^T \end{array}\right] \tag{2.12}$$

is a real skew-Hamiltonian matrix in $\mathbb{SH}_{4n}$. Similarly, set

$$\mathcal{B}_\mathcal{Z} := \begin{bmatrix} \mathcal{Z} & 0 \\ 0 & \bar{\mathcal{Z}} \end{bmatrix}, \tag{2.13}$$
$$\mathcal{B}_\mathcal{T} := \begin{bmatrix} \mathcal{J}\mathcal{Z}^H\mathcal{J}^T & 0 \\ 0 & \overline{\mathcal{J}\mathcal{Z}^H\mathcal{J}^T} \end{bmatrix}, \tag{2.14}$$
$$\mathcal{B}_\mathcal{S} := \begin{bmatrix} \mathcal{S} & 0 \\ 0 & \bar{\mathcal{S}} \end{bmatrix} = \mathcal{B}_\mathcal{T}\mathcal{B}_\mathcal{Z}. \tag{2.15}$$

Hence,

$$\alpha\mathcal{B}_\mathcal{S} - \beta\mathcal{B}_\mathcal{N} = \begin{bmatrix} \alpha\mathcal{S} - \beta\mathcal{N} & 0 \\ 0 & \alpha\bar{\mathcal{S}} - \beta\bar{\mathcal{N}} \end{bmatrix}.$$

One can easily verify that

$$\mathcal{B}_\mathcal{Z}^c := \mathcal{X}_c^H\mathcal{B}_\mathcal{Z}\mathcal{X}_c, \tag{2.16}$$
$$\mathcal{B}_\mathcal{T}^c := \mathcal{X}_c^H\mathcal{B}_\mathcal{T}\mathcal{X}_c = \mathcal{J}(\mathcal{B}_\mathcal{Z}^c)^T\mathcal{J}^T,$$
$$\mathcal{B}_\mathcal{S}^c := \mathcal{X}_c^H\mathcal{B}_\mathcal{S}\mathcal{X}_c = \mathcal{J}(\mathcal{B}_\mathcal{Z}^c)^T\mathcal{J}^T\mathcal{B}_\mathcal{Z}^c \tag{2.17}$$

are all real. Therefore,

$$\begin{aligned}\alpha\mathcal{B}_\mathcal{S}^c - \beta\mathcal{B}_\mathcal{N}^c &= \mathcal{X}_c^H(\alpha\mathcal{B}_\mathcal{S} - \beta\mathcal{B}_\mathcal{N})\mathcal{X}_c \\ &= \mathcal{X}_c^H\begin{bmatrix} \alpha\mathcal{S} - \beta\mathcal{N} & 0 \\ 0 & \alpha\bar{\mathcal{S}} - \beta\bar{\mathcal{N}} \end{bmatrix}\mathcal{X}_c \end{aligned} \tag{2.18}$$

is a real $4n \times 4n$ skew-Hamiltonian/skew-Hamiltonian matrix pencil. For this matrix pencil we can employ Algorithm 3 in [4] to compute the structured factorization (2.8); i.e., we can determine an orthogonal symplectic matrix $\mathcal{U}$ and an orthogonal matrix $\mathcal{Q}$ such that

$$\tilde{\mathcal{B}}^c_{\mathcal{Z}} := \mathcal{U}^T \mathcal{B}^c_{\mathcal{Z}} \mathcal{Q} = \begin{bmatrix} \mathcal{Z}_{11} & \mathcal{Z}_{12} \\ 0 & \mathcal{Z}_{22} \end{bmatrix}, \tag{2.19}$$

$$\tilde{\mathcal{B}}^c_{\mathcal{N}} := \mathcal{J}\mathcal{Q}^T \mathcal{J}^T \mathcal{B}^c_{\mathcal{N}} \mathcal{Q} = \begin{bmatrix} \mathcal{N}_{11} & \mathcal{N}_{12} \\ 0 & \mathcal{N}_{11}^T \end{bmatrix}. \tag{2.20}$$

Thus, if $\tilde{\mathcal{B}}^c_{\mathcal{S}} := \mathcal{J}(\tilde{\mathcal{B}}^c_{\mathcal{Z}})^T \mathcal{J}^T \tilde{\mathcal{B}}^c_{\mathcal{Z}}$, then

$$\alpha\tilde{\mathcal{B}}^c_{\mathcal{S}} - \beta\tilde{\mathcal{B}}^c_{\mathcal{N}} = \alpha(\mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{B}^c_{\mathcal{S}}\mathcal{Q}) - \beta(\mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{B}^c_{\mathcal{N}}\mathcal{Q})$$

is a $\mathcal{J}$-congruent skew-Hamiltonian/skew-Hamiltonian matrix pencil in Schur form. By (2.18) and the fact that the finite eigenvalues of $\alpha\mathcal{S} - \beta\mathcal{N}$ are symmetric with respect to the real axis, we observe that the spectrum of the extended matrix pencil $\alpha\mathcal{B}^c_{\mathcal{S}} - \beta\mathcal{B}^c_{\mathcal{N}}$ consists of two copies of the spectrum of $\alpha\mathcal{S} - \beta\mathcal{N}$. Consequently,

$$\Lambda(\mathcal{S}, \mathcal{H}) = \Lambda(\mathcal{S}, -i\mathcal{N}) = \Lambda(\mathcal{Z}_{22}^T \mathcal{Z}_{11}, -i\mathcal{N}_{11}).$$

In this way, Algorithm 1 below computes the eigenvalues of the complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H} = \alpha\mathcal{S} + i\beta\mathcal{N}$.

From this we can also derive the skew-Hamiltonian/Hamiltonian Schur form of $\alpha\mathcal{B}_{\mathcal{S}} - \beta\mathcal{B}_{\mathcal{H}}$, where

$$\mathcal{B}_{\mathcal{H}} = -i\mathcal{B}_{\mathcal{N}} = \begin{bmatrix} \mathcal{H} & 0 \\ 0 & -\bar{\mathcal{H}} \end{bmatrix} \tag{2.21}$$

and $\mathcal{B}_{\mathcal{S}}$ is as in (2.17). The spectrum of the extended matrix pencil $\alpha\mathcal{B}_{\mathcal{S}} - \beta\mathcal{B}_{\mathcal{H}}$ consists of two copies of the spectrum of $\alpha\mathcal{S} - \beta\mathcal{H}$ [6]. If

$$\mathcal{B}^c_{\mathcal{H}} = -i\mathcal{B}^c_{\mathcal{N}} = \mathcal{X}_c^H \mathcal{B}_{\mathcal{H}} \mathcal{X}_c, \tag{2.22}$$

then it follows from (2.19) and (2.20) that

$$\tilde{\mathcal{B}}^c_{\mathcal{Z}} := \mathcal{U}^T \mathcal{B}^c_{\mathcal{Z}} \mathcal{Q} = \begin{bmatrix} \mathcal{Z}_{11} & \mathcal{Z}_{12} \\ 0 & \mathcal{Z}_{22} \end{bmatrix}, \tag{2.23}$$

$$\tilde{\mathcal{B}}^c_{\mathcal{H}} := \mathcal{J}\mathcal{Q}^T \mathcal{J}^T \mathcal{B}^c_{\mathcal{H}} \mathcal{Q} = \begin{bmatrix} -i\mathcal{N}_{11} & -i\mathcal{N}_{12} \\ 0 & -(-i\mathcal{N}_{11})^H \end{bmatrix}, \tag{2.24}$$

and the matrix pencil $\alpha\tilde{\mathcal{B}}^c_{\mathcal{S}} - \beta\tilde{\mathcal{B}}^c_{\mathcal{H}} := \alpha\mathcal{J}(\tilde{\mathcal{B}}^c_{\mathcal{Z}})^H \mathcal{J}^T \tilde{\mathcal{B}}^c_{\mathcal{Z}} - \beta\tilde{\mathcal{B}}^c_{\mathcal{H}}$ is in skew-Hamiltonian/Hamiltonian Schur form. We have thus obtained the structured Schur form of the extended complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{B}^c_{\mathcal{S}} - \beta\mathcal{B}^c_{\mathcal{H}}$. Moreover,

$$\alpha\tilde{\mathcal{B}}^c_{\mathcal{S}} - \beta\tilde{\mathcal{B}}^c_{\mathcal{H}} = \mathcal{J}\mathcal{Q}^H\mathcal{J}^T(\alpha\mathcal{B}^c_{\mathcal{S}} - \beta\mathcal{B}^c_{\mathcal{H}})\mathcal{Q} = (\mathcal{X}_c\mathcal{J}\mathcal{Q}\mathcal{J}^T)^H (\alpha\mathcal{B}_{\mathcal{S}} - \beta\mathcal{B}_{\mathcal{H}})\, \mathcal{X}_c\mathcal{Q} \tag{2.25}$$

is in skew-Hamiltonian/Hamiltonian Schur form.

We have seen so far that we can compute structured Schur forms and thus are able to compute the eigenvalues of the structured matrix pencils under consideration using the embedding technique into a structured matrix pencil of double size.

3. Deflating subspaces of skew-Hamiltonian/Hamiltonian matrix pencils. For the solution of problems involving skew-Hamiltonian/Hamiltonian matrix pencils as described in the introduction it is usually necessary to compute n-dimensional deflating subspaces associated with eigenvalues in the closed left half plane. To get the desired subspaces we generalize the techniques developed in [6]. For this we need a structure-preserving method to reorder the eigenvalues along the diagonal of the structured Schur form so that all eigenvalues with negative real part appear in the $(1,1)$ block and eigenvalues with positive real part appear in the $(2,2)$ block. Such a reordering method is described in Appendix B of [4].

The following theorem uses this eigenvalue ordering to determine the desired deflating subspaces of the matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ from the structured Schur form (2.25).

THEOREM 3.1. *Let $\alpha\mathcal{S}-\beta\mathcal{H} \in \mathbb{C}^{2n,2n}$ be a skew-Hamiltonian/Hamiltonian matrix pencil with $\mathcal{J}$-semidefinite skew-Hamiltonian matrix $\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$. Consider the extended matrices*

$$\begin{aligned}
\mathcal{B}_\mathcal{Z} &= \operatorname{diag}(\mathcal{Z}, \bar{\mathcal{Z}}),\\
\mathcal{B}_\mathcal{T} &= \operatorname{diag}(\mathcal{J}\mathcal{Z}^H\mathcal{J}^T, \overline{\mathcal{J}\mathcal{Z}^H\mathcal{J}^T}),\\
\mathcal{B}_\mathcal{S} &= \mathcal{B}_\mathcal{T}\mathcal{B}_\mathcal{Z} = \operatorname{diag}(\mathcal{S}, \bar{\mathcal{S}}),\\
\mathcal{B}_\mathcal{H} &= \operatorname{diag}(\mathcal{H}, -\bar{\mathcal{H}}).
\end{aligned}$$

Let $\mathcal{U}, \mathcal{V}, \mathcal{W}$ be unitary matrices such that

$$\begin{aligned}
\mathcal{U}^H\mathcal{B}_\mathcal{Z}\mathcal{V} &= \begin{bmatrix} \mathcal{Z}_{11} & \mathcal{Z}_{12} \\ 0 & \mathcal{Z}_{22} \end{bmatrix} =: \mathcal{R}_\mathcal{Z},\\
\mathcal{W}^H\mathcal{B}_\mathcal{T}\mathcal{U} &= \begin{bmatrix} \mathcal{T}_{11} & \mathcal{T}_{12} \\ 0 & \mathcal{T}_{22} \end{bmatrix} =: \mathcal{R}_\mathcal{T},\\
\mathcal{W}^H\mathcal{B}_\mathcal{H}\mathcal{V} &= \begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ 0 & \mathcal{H}_{22} \end{bmatrix} =: \mathcal{R}_\mathcal{H},
\end{aligned} \tag{3.1}$$

where $\Lambda_-(\mathcal{B}_\mathcal{S}, \mathcal{B}_\mathcal{H}) \subset \Lambda(\mathcal{T}_{11}\mathcal{Z}_{11}, \mathcal{H}_{11})$ and $\Lambda(\mathcal{T}_{11}\mathcal{Z}_{11}, \mathcal{H}_{11}) \cap \Lambda_+(\mathcal{B}_\mathcal{S}, \mathcal{B}_\mathcal{H}) = \emptyset$. Here $\mathcal{Z}_{11}, \mathcal{T}_{11}, \mathcal{H}_{11} \in \mathbb{C}^{m,m}$. Suppose $\Lambda_-(\mathcal{S}, \mathcal{H})$ contains p eigenvalues. If $\left[\begin{smallmatrix} V_1 \\ V_2 \end{smallmatrix}\right] \in \mathbb{C}^{4n,m}$ are the first m columns of $\mathcal{V}$, $2p \le m \le 2n - 2p$, then there are subspaces $\mathbb{L}_1$ and $\mathbb{L}_2$ such that

$$\begin{aligned}
\operatorname{range} V_1 &= \operatorname{Def}_-(\mathcal{S}, \mathcal{H}) + \mathbb{L}_1, \qquad \mathbb{L}_1 \subseteq \operatorname{Def}_0(\mathcal{S}, \mathcal{H}) + \operatorname{Def}_\infty(\mathcal{S}, \mathcal{H}),\\
\operatorname{range} \overline{V_2} &= \operatorname{Def}_+(\mathcal{S}, \mathcal{H}) + \mathbb{L}_2, \qquad \mathbb{L}_2 \subseteq \operatorname{Def}_0(\mathcal{S}, \mathcal{H}) + \operatorname{Def}_\infty(\mathcal{S}, \mathcal{H}).
\end{aligned} \tag{3.2}$$

If $\Lambda(\mathcal{T}_{11}\mathcal{Z}_{11}, \mathcal{H}_{11}) = \Lambda_-(\mathcal{B}_\mathcal{S}, \mathcal{B}_\mathcal{H})$, and $\left[\begin{smallmatrix} U_1 \\ U_2 \end{smallmatrix}\right]$, $\left[\begin{smallmatrix} W_1 \\ W_2 \end{smallmatrix}\right]$ are the first m columns of $\mathcal{U}$, $\mathcal{W}$, respectively, then there exist unitary matrices Q_U, Q_V, Q_W such that

$$\begin{aligned}
U_1 &= [P_U^-, 0]Q_U, \qquad & U_2 &= [0, P_U^+]Q_U,\\
V_1 &= [P_V^-, 0]Q_V, \qquad & V_2 &= [0, P_V^+]Q_V,\\
W_1 &= [P_W^-, 0]Q_W, \qquad & W_2 &= [0, P_W^+]Q_W
\end{aligned}$$

and the columns of P_V^- and $\overline{P_V^+}$ form orthogonal bases of $\operatorname{Def}_-(\mathcal{S}, \mathcal{H})$ and $\operatorname{Def}_+(\mathcal{S}, \mathcal{H})$, respectively. Moreover, the matrices P_U^-, P_U^+, P_W^-, and P_W^+ have orthonormal columns and the following relations are satisfied:

$$\begin{aligned}
\mathcal{Z}P_V^- &= P_U^-\tilde{Z}_{11}, \qquad & \mathcal{J}\mathcal{Z}^H\mathcal{J}^TP_U^- &= P_W^-\tilde{T}_{11}, \qquad & \mathcal{H}P_V^- &= P_W^-\tilde{H}_{11},\\
\mathcal{Z}\overline{P_V^+} &= \overline{P_U^+}\,\overline{\tilde{Z}_{22}}, \qquad & \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\overline{P_U^+} &= \overline{P_W^+}\,\overline{\tilde{T}_{22}}, \qquad & \mathcal{H}\overline{P_V^+} &= -\overline{P_W^+}\,\overline{\tilde{H}_{22}}.
\end{aligned} \tag{3.3}$$

Here, $\tilde{Z}_{kk}$, $\tilde{T}_{kk}$, *and* $\tilde{H}_{kk}$, $k = 1, 2$, *satisfy* $\Lambda(\tilde{T}_{11}\tilde{Z}_{11}, \tilde{H}_{11}) = \Lambda(\tilde{T}_{22}\tilde{Z}_{22}, \tilde{H}_{22}) = \Lambda_-(\mathcal{S}, \mathcal{H})$.

Proof. The factorizations in (3.1) imply that $\mathcal{B}_\mathcal{S}\mathcal{V} = \mathcal{W}\mathcal{R}_\mathcal{T}\mathcal{R}_\mathcal{Z}$ and $\mathcal{B}_\mathcal{H}\mathcal{V} = \mathcal{W}\mathcal{R}_\mathcal{H}$. Comparing the first m columns and making use of the block forms, we have

$$\begin{aligned} \mathcal{S}V_1 &= W_1(\mathcal{T}_{11}\mathcal{Z}_{11}), \qquad & \mathcal{H}V_1 &= W_1\mathcal{H}_{11}, \\ \mathcal{S}\overline{V_2} &= \overline{W_2}\,(\overline{\mathcal{T}_{11}\mathcal{Z}_{11}}), \qquad & \mathcal{H}\overline{V_2} &= -\overline{W_2}\,\overline{\mathcal{H}_{11}}. \end{aligned} \tag{3.4}$$

Clearly, $\operatorname{range} V_1$ and $\operatorname{range}\overline{V_2}$ are both deflating subspaces of $\alpha\mathcal{S} - \beta\mathcal{H}$. Since

$$\Lambda_-(\mathcal{S}, \mathcal{H}) \subseteq \Lambda_-(\mathcal{B}_\mathcal{S}, \mathcal{B}_\mathcal{H}) \subseteq \Lambda(\mathcal{T}_{11}\mathcal{Z}_{11}, \mathcal{H}_{11})$$

and $\Lambda(\mathcal{T}_{11}\mathcal{Z}_{11}, \mathcal{H}_{11})$ contains no eigenvalue with positive real part, we get

$$\begin{aligned} \operatorname{range} V_1 &\subseteq \operatorname{Def}_-(\mathcal{S}, \mathcal{H}) + \mathbb{L}_1, \qquad & \mathbb{L}_1 &\subseteq \operatorname{Def}_0(\mathcal{S}, \mathcal{H}) + \operatorname{Def}_\infty(\mathcal{S}, \mathcal{H}), \\ \operatorname{range}\overline{V_2} &\subseteq \operatorname{Def}_+(\mathcal{S}, \mathcal{H}) + \mathbb{L}_2, \qquad & \mathbb{L}_2 &\subseteq \operatorname{Def}_0(\mathcal{S}, \mathcal{H}) + \operatorname{Def}_\infty(\mathcal{S}, \mathcal{H}). \end{aligned}$$

We still need to show that

$$\operatorname{Def}_-(\mathcal{S}, \mathcal{H}) \subseteq \operatorname{range} V_1, \qquad \operatorname{Def}_+(\mathcal{S}, \mathcal{H}) \subseteq \operatorname{range}\overline{V_2}. \tag{3.5}$$

Let $\tilde{V}_1$ and $\tilde{V}_2$ be full rank matrices whose columns form bases of $\operatorname{Def}_-(\mathcal{S}, \mathcal{H})$ and $\operatorname{Def}_+(\mathcal{S}, \mathcal{H})$, respectively. It is easy to show that the columns of $\left[\begin{smallmatrix} \tilde{V}_1 & 0 \\ 0 & \overline{\tilde{V}_2} \end{smallmatrix}\right]$ span $\operatorname{Def}_-(\mathcal{B}_\mathcal{S}, \mathcal{B}_\mathcal{H})$. This implies that

$$\operatorname{range}\begin{bmatrix} \tilde{V}_1 & 0 \\ 0 & \overline{\tilde{V}_2} \end{bmatrix} \subseteq \operatorname{range}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.$$

Therefore,

$$\operatorname{range}\begin{bmatrix} \tilde{V}_1 \\ 0 \end{bmatrix}, \quad \operatorname{range}\begin{bmatrix} 0 \\ \overline{\tilde{V}_2} \end{bmatrix} \subseteq \operatorname{range}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix},$$

and from this we obtain (3.5) and hence (3.2).

If $\Lambda(\mathcal{T}_{11}\mathcal{Z}_{11}, \mathcal{H}_{11}) = \Lambda_-(\mathcal{B}_\mathcal{S}, \mathcal{B}_\mathcal{H})$, where p is the number of eigenvalues in $\Lambda_-(\mathcal{S}, \mathcal{H})$, then from (3.2) we have $m = 2p$ and

$$\operatorname{range} V_1 = \operatorname{Def}_-(\mathcal{S}, \mathcal{H}), \qquad \operatorname{range}\overline{V_2} = \operatorname{Def}_+(\mathcal{S}, \mathcal{H}).$$

Hence, $\operatorname{rank} V_1 = \operatorname{rank} V_2 = p$ and furthermore $\mathcal{T}_{11}$, $\mathcal{Z}_{11}$, and $\mathcal{H}_{11}$ must be nonsingular. Using (3.4) we get

$$\begin{aligned} \mathcal{H}V_1 &= \mathcal{S}V_1((\mathcal{T}_{11}\mathcal{Z}_{11})^{-1}\mathcal{H}_{11}), \\ \mathcal{H}\overline{V_2} &= -\mathcal{S}\overline{V_2}\,(\overline{(\mathcal{T}_{11}\mathcal{Z}_{11})^{-1}\mathcal{H}_{11}}). \end{aligned}$$

Let $V_1 = [P_V^-, 0]Q_V$ be an RQ (triangular-orthogonal) decomposition [17] with P_V^- of full column rank. Since $\operatorname{rank} V_1 = p$ we have $\operatorname{rank} P_V^- = p$. Partition $V_2Q_V^H = [P_V, P_V^+]$ conforming to $V_1Q_V^H$. Since the columns of $\left[\begin{smallmatrix} V_1 \\ V_2 \end{smallmatrix}\right]$ are orthonormal, we obtain $(P_V^+)^H P_V^+ = I_p$ and hence $\operatorname{rank} P_V^+ = p$. Furthermore, since $\operatorname{rank} V_2 = p$, we have

$$\operatorname{range} P_V \subseteq \operatorname{range} P_V^+ = \operatorname{range} V_2,$$

and using orthonormality, we obtain $P_V = 0$. Therefore, the columns of P_V^- and $\overline{P_V^+}$ form orthogonal bases of $\mathrm{Def}_-(\mathcal{S},\mathcal{H})$ and $\mathrm{Def}_+(\mathcal{S},\mathcal{H})$, respectively.

From (3.1) we have

$$\mathcal{Z}V_1 = U_1\mathcal{Z}_{11}, \qquad \mathcal{J}\mathcal{Z}^H\mathcal{J}^TU_1 = W_1\mathcal{T}_{11}, \qquad \mathcal{H}V_1 = W_1\mathcal{H}_{11}, \tag{3.6}$$

and

$$\mathcal{Z}\overline{V_2} = \overline{U_2}\,\overline{\mathcal{Z}_{11}}, \qquad \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\overline{U_2} = \overline{W_2}\,\overline{\mathcal{T}_{11}}, \qquad \mathcal{H}\overline{V_2} = -\overline{W_2}\,\overline{\mathcal{H}_{11}}. \tag{3.7}$$

Let $U_1 = [P_U^-, 0]Q_U$ and $W_1 = [P_W^-, 0]Q_W$ be RQ (triangular-orthogonal) decompositions, with P_U^-, P_W^- of full column rank. Using $V_1 = [P_V^-, 0]Q_V$ and the fact that $\mathcal{Z}P_V^-$, $\mathcal{S}P_V^-$, and $\mathcal{H}P_V^-$ are of full rank (otherwise there would be a zero or infinite eigenvalue associated with the deflating subspace range P_V^-), from the first and third identity in (3.6) we obtain

$$\operatorname{rank} P_U^- = \operatorname{rank} P_W^- = \operatorname{rank} P_V^- = p.$$

Moreover, setting

$$\tilde{Z} = Q_U\mathcal{Z}_{11}Q_V^H, \qquad \tilde{T} = Q_W\mathcal{T}_{11}Q_U^H, \qquad \tilde{H} = Q_W\mathcal{H}_{11}Q_V^H,$$

we obtain

$$\tilde{Z} = \begin{bmatrix} \tilde{Z}_{11} & 0 \\ \tilde{Z}_{21} & \tilde{Z}_{22} \end{bmatrix}, \qquad \tilde{T} = \begin{bmatrix} \tilde{T}_{11} & 0 \\ \tilde{T}_{21} & \tilde{T}_{22} \end{bmatrix}, \qquad \tilde{H} = \begin{bmatrix} \tilde{H}_{11} & 0 \\ \tilde{H}_{21} & \tilde{H}_{22} \end{bmatrix},$$

where all diagonal blocks are $p \times p$.

Set $U_2Q_U^H =: [P_U, P_U^+]$, $W_2Q_W^H =: [P_W, P_W^+]$ and take $V_2Q_V^H =: [0, P_V^+]$. The block forms of $\tilde{Z}$, $\tilde{T}$, and $\tilde{H}$ together with the first identity of (3.7) imply that $\overline{P_U}\,\tilde{Z}_{11} = \overline{P_U^+}\,\tilde{Z}_{21}$. Since the columns of $\left[\begin{smallmatrix} U_1 \\ U_2 \end{smallmatrix}\right]$ are orthonormal, we have $(P_U^+)^HP_U^+ = I_p$ and $(P_U^+)^HP_U = 0$. Hence, $\tilde{Z}_{21} = 0$, and consequently $P_U = 0$. Similarly, from the third identity of (3.7) we get $P_W = 0$, $\tilde{H}_{21} = 0$, and from the second identity we obtain $\tilde{T}_{21} = 0$. Combining all these observations, we obtain

$$\begin{bmatrix} \mathcal{Z} & 0 \\ 0 & \bar{\mathcal{Z}} \end{bmatrix} \begin{bmatrix} P_V^- & 0 \\ 0 & P_V^+ \end{bmatrix} = \begin{bmatrix} P_U^- & 0 \\ 0 & P_U^+ \end{bmatrix} \begin{bmatrix} \tilde{Z}_{11} & 0 \\ 0 & \tilde{Z}_{22} \end{bmatrix},$$

$$\begin{bmatrix} \mathcal{J}\mathcal{Z}^H\mathcal{J}^T & 0 \\ 0 & \overline{\mathcal{J}\mathcal{Z}^H\mathcal{J}^T} \end{bmatrix} \begin{bmatrix} P_U^- & 0 \\ 0 & P_U^+ \end{bmatrix} = \begin{bmatrix} P_W^- & 0 \\ 0 & P_W^+ \end{bmatrix} \begin{bmatrix} \tilde{T}_{11} & 0 \\ 0 & \tilde{T}_{22} \end{bmatrix},$$

$$\begin{bmatrix} \mathcal{H} & 0 \\ 0 & -\bar{\mathcal{H}} \end{bmatrix} \begin{bmatrix} P_V^- & 0 \\ 0 & P_V^+ \end{bmatrix} = \begin{bmatrix} P_W^- & 0 \\ 0 & P_W^+ \end{bmatrix} \begin{bmatrix} \tilde{H}_{11} & 0 \\ 0 & \tilde{H}_{22} \end{bmatrix},$$

which gives (3.3). □

We remark that (3.1) can be constructed from (2.25) by reordering the eigenvalues properly.

Theorem 3.1 provides a way for obtaining the stable deflating subspace of a skew-Hamiltonian/Hamiltonian matrix pencil from the deflating subspaces of an embedded skew-Hamiltonian/Hamiltonian matrix pencil of double size. This will be used by the algorithms formulated in the next section.

4. Algorithms. The results of Theorem 3.1 together with the embedding technique lead to the following algorithm to compute the eigenvalues and the deflating subspaces $\mathrm{Def}_-(\mathcal{S},\mathcal{H})$ and $\mathrm{Def}_+(\mathcal{S},\mathcal{H})$ of a complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S}-\beta\mathcal{H}$. Since the algorithms are rather technical, we do not discuss details like eigenvalue reordering or explicit elimination orders in the construction of the structured Schur forms. Instead we refer the reader to the technical report [4] for these details.

In summary, Algorithm 1 proposed below transforms a $2n\times 2n$ complex skew-Hamiltonian/Hamiltonian matrix pencil with $\mathcal{J}$-semidefinite skew-Hamiltonian part into a $4n\times 4n$ complex skew-Hamiltonian/Hamiltonian matrix pencil in Schur form. The process passes through intermediate matrix pencils of the following types.

$2n\times 2n$ complex skew-Hamiltonian/Hamiltonian matrix pencil
$\alpha\mathcal{S}-\beta\mathcal{H}$ with $\mathcal{S}=\mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$.
$\Downarrow$
Equation (2.18)
$\Downarrow$
$4n\times 4n$ real skew-Hamiltonian/skew-Hamiltonian matrix pencil
$\alpha\mathcal{B}_{\mathcal{S}}^c-\beta\mathcal{B}_{\mathcal{N}}^c$ with $\mathcal{B}_{\mathcal{S}}^c=\mathcal{J}(\mathcal{B}_{\mathcal{Z}}^c)^T\mathcal{J}^T\mathcal{B}_{\mathcal{Z}}^c$
$\Downarrow$
Algorithm 3 in [4]
$\Downarrow$
$4n\times 4n$ real skew-Hamiltonian/skew-Hamiltonian matrix pencil in Schur form
$\alpha\tilde{\mathcal{B}}_{\mathcal{S}}^c-\beta\tilde{\mathcal{B}}_{\mathcal{N}}^c$ with $\tilde{\mathcal{B}}_{\mathcal{S}}^c=\mathcal{J}(\tilde{\mathcal{B}}_{\mathcal{Z}}^c)^T\mathcal{J}^T\tilde{\mathcal{B}}_{\mathcal{Z}}^c$
and $\tilde{\mathcal{B}}_{\mathcal{Z}}^c=\mathcal{U}^T\mathcal{B}_{\mathcal{Z}}^c\mathcal{Q}=\left[\begin{smallmatrix}\mathcal{Z}_{11} & \mathcal{Z}_{12}\\ 0 & \mathcal{Z}_{22}\end{smallmatrix}\right]$, $\tilde{\mathcal{B}}_{\mathcal{N}}^c=\mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{B}_{\mathcal{N}}^c\mathcal{Q}=\left[\begin{smallmatrix}\mathcal{N}_{11} & \mathcal{N}_{12}\\ 0 & \mathcal{N}_{11}^T\end{smallmatrix}\right]$ as in (2.19) and (2.20)
$\Downarrow$
Algorithm 4 in [4]
$\Downarrow$
$4n\times 4n$ complex skew-Hamiltonian/Hamiltonian matrix pencil in Schur form
with ordered eigenvalues.

The required deflating subspaces of the original skew-Hamiltonian/Hamiltonian matrix pencil are then obtained from the deflating subspaces of the final $4n\times 4n$ complex skew-Hamiltonian/Hamiltonian matrix pencil. (Unfortunately, if there are nonreal eigenvalues, then Algorithm 4 in [4] (the eigenvalue sorting algorithm) reintroduces complex entries into the $4n\times 4n$ extended real matrix pencil.)

ALGORITHM 1. *Given a complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S}-\beta\mathcal{H}$ with $\mathcal{J}$-semidefinite skew-Hamiltonian part $\mathcal{S}=\mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$, this algorithm computes the structured Schur form of the extended skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{B}_{\mathcal{S}}^c-\beta\mathcal{B}_{\mathcal{H}}^c$, the eigenvalues of $\alpha\mathcal{S}-\beta\mathcal{H}$, and orthonormal bases of the deflating subspace* $\mathrm{Def}_-(\mathcal{S},\mathcal{H})$ *and the companion subspace* $\operatorname{range}P_U^-$.

Input: Hamiltonian matrix $\mathcal{H}$ and the factor $\mathcal{Z}$ of $\mathcal{S}$.

Output: P_V^-, P_U^- as defined in Theorem 3.1.

Step 1:

Set $\mathcal{N}=i\mathcal{H}$ and form matrices $\mathcal{B}_{\mathcal{Z}}^c$, $\mathcal{B}_{\mathcal{N}}^c$ as in (2.16) and (2.12), respectively. Find the structured Schur form of the skew-Hamiltonian/skew-Hamiltonian matrix pencil $\alpha\mathcal{B}_{\mathcal{S}}^c-\beta\mathcal{B}_{\mathcal{N}}^c$ using Algorithm 3 in [4] to compute the factorization

$$\tilde{\mathcal{B}}_{\mathcal{Z}}^c=\mathcal{U}^T\mathcal{B}_{\mathcal{Z}}^c\mathcal{Q}=\begin{bmatrix}\mathcal{Z}_{11} & \mathcal{Z}_{12}\\ 0 & \mathcal{Z}_{22}\end{bmatrix},$$

$$\tilde{\mathcal{B}}_{\mathcal{N}}^c=\mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{B}_{\mathcal{N}}^c\mathcal{Q}=\begin{bmatrix}\mathcal{N}_{11} & \mathcal{N}_{12}\\ 0 & \mathcal{N}_{11}^T\end{bmatrix},$$

where $\mathcal{Q}$ is real orthogonal, $\mathcal{U}$ is real orthogonal symplectic, $\mathcal{Z}_{11}$, $\mathcal{Z}_{22}^T$ are upper triangular, and $\mathcal{N}_{11}$ is quasi upper triangular.

Step 2:

Reorder the eigenvalues using Algorithm 4 in [4] to determine a unitary matrix $\tilde{\mathcal{Q}}$ and a unitary symplectic matrix $\tilde{\mathcal{U}}$ such that

$$\tilde{\mathcal{U}}^H \check{\mathcal{B}}_{\mathcal{Z}}^c \tilde{\mathcal{Q}} = \begin{bmatrix} \tilde{\mathcal{Z}}_{11} & \tilde{\mathcal{Z}}_{12} \\ 0 & \tilde{\mathcal{Z}}_{22} \end{bmatrix} =: \check{\mathcal{B}}_{\mathcal{Z}}^c,$$
$$\mathcal{J}\tilde{\mathcal{Q}}^H \mathcal{J}^T (-i\check{\mathcal{B}}_{\mathcal{N}}^c)\tilde{\mathcal{Q}} = \begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ 0 & -\mathcal{H}_{11}^H \end{bmatrix} =: \check{\mathcal{B}}_{\mathcal{H}}^c,$$

with $\tilde{\mathcal{Z}}_{11}$, $\tilde{\mathcal{Z}}_{22}^H$, $\mathcal{H}_{11}$ upper triangular such that $\Lambda_-(\mathcal{J}(\check{\mathcal{B}}_{\mathcal{Z}}^c)^H \mathcal{J}^T \check{\mathcal{B}}_{\mathcal{Z}}^c, \check{\mathcal{B}}_{\mathcal{H}}^c)$ is contained in the spectrum of the $2p \times 2p$ leading principal subpencil of $\alpha\tilde{\mathcal{Z}}_{22}^H\tilde{\mathcal{Z}}_{11} - \beta\mathcal{H}_{11}$.

Step 3:

Set $V = [I_{2n}, 0]\mathcal{X}_c\mathcal{Q}\tilde{\mathcal{Q}}\left[\begin{smallmatrix} I_{2p} \\ 0 \end{smallmatrix}\right]$, $U = [I_{2n}, 0]\mathcal{X}_c\mathcal{U}\tilde{\mathcal{U}}\left[\begin{smallmatrix} I_{2p} \\ 0 \end{smallmatrix}\right]$ (where $\mathcal{X}_c$ is as in (2.11)) and compute P_V^-, P_U^-, orthogonal bases of range V and range U, respectively, using any numerically stable orthogonalization scheme.

End

Based on flop counts, we estimate the cost of this algorithm to be roughly 50% of the cost of the periodic QZ algorithm [10, 19] applied to the $2n \times 2n$ complex pencil $\alpha\mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z} - \beta\mathcal{H}$ (treating $\mathcal{J}\mathcal{Z}^H\mathcal{J}^T$ as one matrix).

If $\mathcal{S}$ is not factored, then the algorithm can be simplified by using the method of [32] to compute the real skew-Hamiltonian/Hamiltonian Schur form of $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ directly.

ALGORITHM 2. *Given a complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$, this algorithm computes the structured Schur form of the extended skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$, the eigenvalues of $\alpha\mathcal{S} - \beta\mathcal{H}$, and an orthogonal basis of the deflating subspace* $\mathrm{Def}_-(\mathcal{S}, \mathcal{H})$.

Input: A complex skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$.

Output: P_V^- as defined in Theorem 3.1.

Step 1:

Set $\mathcal{N} = i\mathcal{H}$ and form the matrices $\mathcal{B}_{\mathcal{S}}^c$, $\mathcal{B}_{\mathcal{N}}^c$ as in (2.17) and (2.12), respectively.

Find the structured Schur form of the skew-Hamiltonian/skew-Hamiltonian matrix pencil $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{N}}^c$ using Algorithm 5 in [4] to compute the factorization

$$\check{\mathcal{B}}_{\mathcal{S}}^c = \mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{B}_{\mathcal{S}}^c\mathcal{Q} = \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} \\ 0 & \mathcal{S}_{11}^T \end{bmatrix},$$
$$\check{\mathcal{B}}_{\mathcal{N}}^c = \mathcal{J}\mathcal{Q}^T\mathcal{J}^T\mathcal{B}_{\mathcal{N}}^c\mathcal{Q} = \begin{bmatrix} \mathcal{N}_{11} & \mathcal{N}_{12} \\ 0 & \mathcal{N}_{11}^T \end{bmatrix},$$

where $\mathcal{Q}$ is real orthogonal, $\mathcal{S}_{11}$ is upper triangular, and $\mathcal{N}_{11}$ is quasi upper triangular.

Step 2:

Reorder the eigenvalues using Algorithm 6 in [4] to determine a unitary matrix $\tilde{\mathcal{Q}}$ such that

$$\mathcal{J}\tilde{\mathcal{Q}}^H\mathcal{J}^T\check{\mathcal{B}}_{\mathcal{S}}^c\tilde{\mathcal{Q}} = \begin{bmatrix} \tilde{\mathcal{S}}_{11} & \tilde{\mathcal{S}}_{12} \\ 0 & \tilde{\mathcal{S}}_{11}^H \end{bmatrix},$$

$$\mathcal{J}\tilde{\mathcal{Q}}^H\mathcal{J}^T(-i\check{\mathcal{B}}^c_{\mathcal{N}})\tilde{\mathcal{Q}} = \begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ 0 & -\mathcal{H}_{11}^H \end{bmatrix},$$

with $\tilde{\mathcal{S}}_{11}$, $\mathcal{H}_{11}$ upper triangular and such that $\Lambda_-(\check{\mathcal{B}}^c_{\mathcal{S}}, -i\check{\mathcal{B}}^c_{\mathcal{N}})$ is contained in the spectrum of the $2p \times 2p$ leading principal subpencil of $\alpha\tilde{\mathcal{S}}_{11} - \beta\mathcal{H}_{11}$.

Step 3:

Set $V = [I_{2n}, 0]\mathcal{X}_c\mathcal{Q}\tilde{\mathcal{Q}}\left[\begin{smallmatrix} I_{2p} \\ 0 \end{smallmatrix}\right]$ (where $\mathcal{X}_c$ is as in (2.11)) and compute P_V^-, the orthogonal basis of range V, using any numerically stable orthogonalization scheme.

End

Algorithm 2 needs roughly 80% of the $1600n^3$ real flops required by the QZ algorithm applied to the $2n \times 2n$ complex pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ as suggested in [37]. If only the eigenvalues are computed, then Algorithm 2 without accumulation of V needs roughly 60% of the $960n^3$ real flops required by the QZ algorithm.

In this section we have presented numerical algorithms for the computation of (complex) structured triangular forms. Various details appear in [4]. In the next section we give an error analysis. The analysis is a generalization of the analysis for Hamiltonian matrices in [6, 7, 8].

5. Error and perturbation analysis. In this section we will give the perturbation analysis for eigenvalues and deflating subspaces of skew-Hamiltonian/Hamiltonian matrix pencils. Variables marked with a circumflex denote perturbed quantities.

We begin with the perturbation analysis for the eigenvalues of $\alpha\mathcal{S} - \beta\mathcal{H}$ and $\alpha\mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z} - \beta\mathcal{H}$. In principle, we could multiply out $\mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$ and apply the classical perturbation analysis of matrix pencils using the chordal metric [36], but this may give pessimistic bounds and would display neither the effects of perturbing each factor separately nor the effects of structured perturbations. Therefore, we make use of the perturbation analysis for formal products of matrices developed in [9].

If Algorithm 2 is applied to the skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$, then we compute the structured Schur form of the extended skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{B}^c_{\mathcal{S}} - \beta\mathcal{B}^c_{\mathcal{H}}$. The well-known backward error analysis of orthogonal matrix computations implies that rounding errors in Algorithm 2 are equivalent to perturbing $\alpha\mathcal{B}^c_{\mathcal{S}} - \beta\mathcal{B}^c_{\mathcal{H}}$ to a nearby matrix pencil $\alpha\hat{\mathcal{B}}^c_{\mathcal{S}} - \beta\hat{\mathcal{B}}^c_{\mathcal{H}}$, where

$$\hat{\mathcal{B}}^c_{\mathcal{S}} = \mathcal{B}^c_{\mathcal{S}} + \mathcal{E}_{\mathcal{S}}, \tag{5.1}$$

$$\hat{\mathcal{B}}^c_{\mathcal{H}} = \mathcal{B}^c_{\mathcal{H}} + \mathcal{E}_{\mathcal{H}}, \tag{5.2}$$

with $\mathcal{E}_{\mathcal{S}} \in \mathbb{SH}_{4n}$, $\mathcal{E}_{\mathcal{H}} \in \mathbb{H}_{4n}$ and

$$\|\mathcal{E}_{\mathcal{S}}\|_2 < c_{\mathcal{S}}\varepsilon\,\|\mathcal{B}^c_{\mathcal{S}}\|_2\,, \tag{5.3}$$

$$\|\mathcal{E}_{\mathcal{H}}\|_2 < c_{\mathcal{H}}\varepsilon\,\|\mathcal{B}^c_{\mathcal{H}}\|_2\,. \tag{5.4}$$

Here ε is the unit round of the floating point arithmetic and $c_{\mathcal{S}}$ and $c_{\mathcal{H}}$ are modest constants depending on the details of the implementation and arithmetic. Let x and y be unit norm vectors such that

$$\mathcal{H}x = \alpha_1 y, \qquad \mathcal{S}x = \beta_1 y, \tag{5.5}$$

and let $\lambda = \alpha_1/\beta_1$ be a simple eigenvalue of $\alpha\mathcal{S} - \beta\mathcal{H}$. If λ is finite and $\operatorname{Re}\lambda \neq 0$, then $-\bar{\lambda}$ is also a simple eigenvalue of $\alpha\mathcal{S} - \beta\mathcal{H}$. Let u, v be unit norm vectors such that

$$\mathcal{H}u = \alpha_2 v, \qquad \mathcal{S}u = \beta_2 v, \tag{5.6}$$

and $\alpha_2/\beta_2 = -\bar{\lambda}$. Then we have

$$-\bar{\mathcal{H}}\bar{u} = -\bar{\alpha}_2\bar{v}, \qquad \bar{\mathcal{S}}\bar{u} = \bar{\beta}_2\bar{v}. \tag{5.7}$$

Using the equivalence of the matrix pencils $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ and $\alpha\mathcal{B}_{\mathcal{S}} - \beta\mathcal{B}_{\mathcal{H}}$, and setting

$$\mathcal{U}_1 = \mathcal{X}_c^H \begin{bmatrix} y & 0 \\ 0 & \bar{v} \end{bmatrix}, \qquad \mathcal{U}_2 = \mathcal{X}_c^H \begin{bmatrix} x & 0 \\ 0 & \bar{u} \end{bmatrix}, \tag{5.8}$$

we obtain from (5.5) and (5.7) that

$$\mathcal{B}_{\mathcal{H}}^c\mathcal{U}_2 = \mathcal{U}_1 \begin{bmatrix} \alpha_1 & 0 \\ 0 & -\bar{\alpha}_2 \end{bmatrix}, \qquad \mathcal{B}_{\mathcal{S}}^c\mathcal{U}_2 = \mathcal{U}_1 \begin{bmatrix} \beta_1 & 0 \\ 0 & \bar{\beta}_2 \end{bmatrix},$$

which implies that λ is a double eigenvalue of $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ with a complete set of linearly independent eigenvectors. Similarly, $-\bar{\lambda}$ is a double eigenvalue of $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ with a complete set of linearly independent eigenvectors and

$$\mathcal{B}_{\mathcal{H}}^c\mathcal{V}_2 = \mathcal{V}_1 \begin{bmatrix} \alpha_2 & 0 \\ 0 & -\bar{\alpha}_1 \end{bmatrix}, \qquad \mathcal{B}_{\mathcal{S}}^c\mathcal{V}_2 = \mathcal{V}_1 \begin{bmatrix} \beta_2 & 0 \\ 0 & \bar{\beta}_1 \end{bmatrix},$$

where

$$\mathcal{V}_1 = \mathcal{X}_c^H \begin{bmatrix} v & 0 \\ 0 & \bar{y} \end{bmatrix}, \qquad \mathcal{V}_2 = \mathcal{X}_c^H \begin{bmatrix} u & 0 \\ 0 & \bar{x} \end{bmatrix}. \tag{5.9}$$

Note that the finite eigenvalues with nonzero real part appear in pairs as in (5.5) and (5.6), but infinite and purely imaginary eigenvalues may not appear in pairs. Consequently, in the following perturbation theorem, the bounds for purely imaginary and infinite eigenvalues are different from the bounds for finite eigenvalues with nonzero real part.

THEOREM 5.1. *Consider the skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ along with the corresponding extended matrix pencils $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c = \mathcal{X}_c^H(\alpha\mathcal{B}_{\mathcal{S}} - \beta\mathcal{B}_{\mathcal{H}})\mathcal{X}_c$, where $\mathcal{B}_{\mathcal{S}}$ is given by* (2.15), *$\mathcal{B}_{\mathcal{H}}$ by* (2.21), *$\mathcal{B}_{\mathcal{H}}^c$ by* (2.22), *$\mathcal{X}_c$ by* (2.11), *and $\mathcal{B}_{\mathcal{S}}^c$ by* (2.17). *Let $\alpha\hat{\mathcal{B}}_{\mathcal{S}}^c - \beta\hat{\mathcal{B}}_{\mathcal{H}}^c$ be a perturbed extended matrix pencil satisfying* (5.1)–(5.4) *with constants $c_{\mathcal{H}}$, $c_{\mathcal{S}}$ and let ε be equal to the unit round of the floating point arithmetic.*

If λ is a simple eigenvalue of $\alpha\mathcal{S} - \beta\mathcal{H}$ with vectors x and y as in (5.5) *and vectors u and v as in* (5.6), *then the corresponding double eigenvalue of $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ may split into two eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of the perturbed matrix pencil $\alpha\hat{\mathcal{B}}_{\mathcal{S}}^c - \beta\hat{\mathcal{B}}_{\mathcal{H}}^c$, each of which satisfies the following bounds.*

(i) *If λ is finite and $\operatorname{Re}\lambda \neq 0$, then*

$$\left|\frac{\hat{\lambda}_k - \lambda}{\lambda}\right| \leq \frac{\varepsilon}{|u^H\mathcal{J}y|}\left(\frac{c_{\mathcal{H}}}{|\alpha_1|}\|\mathcal{H}\|_2 + \frac{c_{\mathcal{S}}}{|\beta_1|}\|\mathcal{S}\|_2\right) + O(\varepsilon^2), \qquad k = 1, 2.$$

(ii) *If λ is finite and $\operatorname{Re}\lambda = 0$, then*

$$|\hat{\lambda}_k - \lambda| \leq \frac{\varepsilon}{|\beta_1||x^H\mathcal{J}y|}(c_{\mathcal{H}}\|\mathcal{H}\|_2 + c_{\mathcal{S}}|\lambda|\,\|\mathcal{S}\|_2) + O(\varepsilon^2), \qquad k = 1, 2.$$

(iii) *If $\lambda = \infty$, then*

$$\frac{1}{|\hat{\lambda}_k|} \leq \varepsilon\frac{c_{\mathcal{S}}\|\mathcal{S}\|_2}{|\alpha_1||x^H\mathcal{J}y|} + O(\varepsilon^2), \qquad k = 1, 2.$$

Proof. We first consider the case that λ is finite and $\operatorname{Re}\lambda \neq 0$. Let $\mathcal{U}_1$ and $\mathcal{U}_2$ be defined by (5.8) and $\mathcal{V}_1$ and $\mathcal{V}_2$ be defined by (5.9). Using the perturbation theory for formal products of matrices (see [9]), we obtain

$$\left|\frac{\hat{\lambda}_k - \lambda}{\lambda}\right| \leq \min\left(\left\|(\mathcal{V}_2^H \mathcal{J}\mathcal{U}_1 C_{\mathcal{S}})^{-1}\mathcal{V}_2^H \mathcal{J}\left(\frac{1}{\lambda}\mathcal{E}_{\mathcal{H}} - \mathcal{E}_{\mathcal{S}}\right)\mathcal{U}_2\right\|_2, \right.$$
$$\left. \left\|(\mathcal{V}_2^H \mathcal{J}\mathcal{U}_1)^{-1}\mathcal{V}_2^H \mathcal{J}\left(\frac{1}{\lambda}\mathcal{E}_{\mathcal{H}} - \mathcal{E}_{\mathcal{S}}\right)\mathcal{U}_2 C_{\mathcal{S}}^{-1}\right\|_2\right) + O(\varepsilon^2).$$

Here, $C_{\mathcal{S}} = \left[\begin{smallmatrix}\beta_1 & 0\\ 0 & \bar{\beta}_2\end{smallmatrix}\right]$ and $\mathcal{V}_2^H \mathcal{J}\mathcal{U}_1 = \left[\begin{smallmatrix}u & 0\\ 0 & \bar{x}\end{smallmatrix}\right]^H \mathcal{X}_c \mathcal{J}\mathcal{X}_c^H \left[\begin{smallmatrix}y & 0\\ 0 & \bar{v}\end{smallmatrix}\right] = \left[\begin{smallmatrix}u^H\mathcal{J}y & 0\\ 0 & x^T\mathcal{J}\bar{v}\end{smallmatrix}\right]$. The second equation in (5.6) implies $u^H\mathcal{J}\mathcal{S} = \bar{\beta}_2 v^H \mathcal{J}$. Combining this with the second equation of (5.5) we get $\bar{\beta}_2 v^H \mathcal{J}x = \beta_1 u^H \mathcal{J}y$. Hence,

$$\begin{aligned}\left|\frac{\hat{\lambda}_k - \lambda}{\lambda}\right| &\leq \left\|(\mathcal{V}_2^H \mathcal{J}\mathcal{U}_1 C_{\mathcal{S}})^{-1}\mathcal{V}_2^H \mathcal{J}\left(\frac{1}{\lambda}\mathcal{E}_{\mathcal{H}} - \mathcal{E}_{\mathcal{S}}\right)\mathcal{U}_2\right\|_2 + O(\varepsilon^2)\\ &\leq \left\|(\mathcal{V}_2^H \mathcal{J}\mathcal{U}_1 C_{\mathcal{S}})^{-1}\right\|_2 \left\|\frac{1}{\lambda}\mathcal{E}_{\mathcal{H}} - \mathcal{E}_{\mathcal{S}}\right\|_2 + O(\varepsilon^2)\\ &\leq \frac{1}{|u^H\mathcal{J}y|}\left(\frac{\|\mathcal{E}_{\mathcal{H}}\|_2}{|\beta_1\lambda|} + \frac{\|\mathcal{E}_{\mathcal{S}}\|_2}{|\beta_1|}\right) + O(\varepsilon^2)\\ &\leq \frac{\varepsilon}{|u^H\mathcal{J}y|}\left(\frac{c_{\mathcal{H}}}{|\alpha_1|}\|\mathcal{H}\|_2 + \frac{c_{\mathcal{S}}}{|\beta_1|}\|\mathcal{S}\|_2\right) + O(\varepsilon^2).\end{aligned}$$

If λ is purely imaginary or infinite, then the bounds are obtained by adapting the classical perturbation theory in [36] to a formal product of matrices (for details see [9]) and by replacing (5.7) with $-\bar{\mathcal{H}}\bar{x} = -\bar{\alpha}_1\bar{y}$ and $\bar{\mathcal{S}}\bar{x} = \bar{\beta}_1\bar{y}$ as well as replacing u, v, α_2, and β_2 by x, y, α_1, and β_1, respectively. ∎

The bound in part (i) appears to involve only u, y, α_1, and β_1 but not v, x, α_2, and β_2. However, note in the proof that $\bar{\beta}_2 v^H\mathcal{J}x = \beta_1 u^H\mathcal{J}y$, so the bound implicitly involves all the parameters. Note further that if $\mathcal{S}$ is nonsingular, then $v^H\mathcal{J}x$ and $u^H\mathcal{J}y$ are just the reciprocals of the condition number of λ as eigenvalue of $\mathcal{S}^{-1}\mathcal{H}$ and $\mathcal{H}\mathcal{S}^{-1}$, respectively; see [6].

If $\mathcal{S}$ is given in factored form, Algorithm 1 computes a unitary symplectic matrix $\mathcal{U}$ and a unitary matrix $\mathcal{Q}$ which reduce the perturbed matrices

$$\hat{\mathcal{B}}_{\mathcal{Z}}^c := \mathcal{B}_{\mathcal{Z}}^c + \mathcal{E}_{\mathcal{Z}}, \qquad \hat{\mathcal{B}}_{\mathcal{H}}^c := \mathcal{B}_{\mathcal{H}}^c + \mathcal{E}_{\mathcal{H}} \tag{5.10}$$

to block upper triangular form as in (2.23) and (2.24), where

$$\|\mathcal{E}_{\mathcal{Z}}\|_2 \leq c_{\mathcal{Z}}\varepsilon\|\mathcal{B}_{\mathcal{Z}}^c\|_2, \qquad \|\mathcal{E}_{\mathcal{H}}\|_2 \leq c_{\mathcal{H}}\varepsilon\|\mathcal{B}_{\mathcal{H}}^c\|_2, \tag{5.11}$$

and $c_{\mathcal{Z}}$ and $c_{\mathcal{H}}$ are constants. The eigenvalue perturbation bounds then are essentially the same as in Theorem 5.1.

THEOREM 5.2. *Consider the skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ with $\mathcal{J}$-semidefinite skew-Hamiltonian part $\mathcal{S} = \mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}$. Let $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c = \mathcal{X}_c^H(\alpha\mathcal{B}_{\mathcal{S}} - \beta\mathcal{B}_{\mathcal{H}})\mathcal{X}_c$ be the corresponding extended matrix pencils, where $\mathcal{B}_{\mathcal{S}}^c = \mathcal{J}(\mathcal{B}_{\mathcal{Z}}^c)^H\mathcal{J}^T\mathcal{B}_{\mathcal{Z}}^c$, $\mathcal{B}_{\mathcal{Z}}$ and $\mathcal{B}_{\mathcal{Z}}^c$ are given by* (2.13) *and* (2.16), *$\mathcal{B}_{\mathcal{H}}$ and $\mathcal{B}_{\mathcal{H}}^c$ by* (2.21) *and* (2.22), *and $\mathcal{X}_c$ by* (2.11). *Let $(\hat{\mathcal{B}}_{\mathcal{Z}}^c, \hat{\mathcal{B}}_{\mathcal{H}}^c)$ be the perturbed extended matrix pair in* (5.10), (5.11) *with constants $c_{\mathcal{H}}$, $c_{\mathcal{Z}}$.*

Let λ be a simple eigenvalue of $\alpha\mathcal{S}-\beta\mathcal{H}=\alpha\mathcal{J}\mathcal{Z}^H\mathcal{J}^T\mathcal{Z}-\beta\mathcal{H}$ with $\operatorname{Re}\lambda\neq 0$, and let x, y, z, u, v, w be unit norm vectors such that

$$\mathcal{J}\mathcal{Z}^H\mathcal{J}^T x=\alpha_1 y,\qquad \mathcal{H}z=\beta_1 y,\qquad \mathcal{Z}z=\gamma_1 x, \tag{5.12}$$

with $\lambda=\frac{\beta_1}{\alpha_1\gamma_1}$, and

$$\mathcal{J}\mathcal{Z}^H\mathcal{J}^T u=\alpha_2 v,\qquad \mathcal{H}w=\beta_2 v,\qquad \mathcal{Z}w=\gamma_2 u, \tag{5.13}$$

with $-\bar{\lambda}=\frac{\beta_2}{\alpha_2\gamma_2}$.

The corresponding double eigenvalue of $\alpha\mathcal{B}_{\mathcal{S}}^c-\beta\mathcal{B}_{\mathcal{H}}^c$ may split into two eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of the perturbed matrix pencil $\alpha\hat{\mathcal{B}}_{\mathcal{S}}^c-\beta\hat{\mathcal{B}}_{\mathcal{H}}^c$, each of which satisfies the following bounds.

(i) *If λ is finite and $\operatorname{Re}\lambda\neq 0$, then*

$$\left|\frac{\hat{\lambda}_k-\lambda}{\lambda}\right|\leq\varepsilon\left(\frac{c_{\mathcal{H}}}{|\beta_1 w^H\mathcal{J}y|}\left\|\mathcal{H}\right\|_2+2\frac{c_{\mathcal{Z}}}{\min\{|\gamma_1 u^H\mathcal{J}x|,|\alpha_1 w^H\mathcal{J}y|\}}\left\|\mathcal{Z}\right\|_2\right)+O(\varepsilon^2).$$

(ii) *If λ is purely imaginary, then*

$$|\hat{\lambda}_k-\lambda|\leq\varepsilon\left(\frac{c_{\mathcal{H}}}{|\alpha_1\gamma_1 y^H\mathcal{J}z|}\left\|\mathcal{H}\right\|_2+\frac{2|\lambda|c_{\mathcal{Z}}}{|\gamma_1 u^H\mathcal{J}x|}\left\|\mathcal{Z}\right\|_2\right)+O(\varepsilon^2).$$

(iii) *If $\lambda=\infty$, then $|\hat{\lambda}_k|^{-1}=O(\varepsilon^2)$.*

Proof. The perturbation analysis follows [9]. If λ is finite and $\operatorname{Re}\lambda\neq 0$, then

$$\left|\frac{\hat{\lambda}_k-\lambda}{\lambda}\right|\leq\left\|(\mathcal{V}_2^H\mathcal{J}\mathcal{U}_3)^{-1}(\tilde{C}_1\tilde{C}_3)^{-1}\left(\mathcal{V}_3^H\mathcal{E}_{\mathcal{Z}}^H\mathcal{J}\mathcal{U}_1C_3+\tilde{C}_3^H\mathcal{U}_1^H\mathcal{J}\mathcal{E}_{\mathcal{Z}}\mathcal{U}_3-\frac{1}{\lambda}\mathcal{V}_3^H\mathcal{J}\mathcal{E}_{\mathcal{H}}\mathcal{U}_3\right)\right\|_2+O(\varepsilon^2),$$

where $\mathcal{U}_1=\mathcal{X}_c^H\left[\begin{smallmatrix}x&0\\0&\bar{u}\end{smallmatrix}\right]\in\mathbb{C}^{4n,2}$, $\mathcal{U}_3=\mathcal{X}_c^H\left[\begin{smallmatrix}z&0\\0&\bar{w}\end{smallmatrix}\right]\in\mathbb{C}^{4n,2}$, $\mathcal{V}_2=\mathcal{X}_c^H\left[\begin{smallmatrix}v&0\\0&\bar{y}\end{smallmatrix}\right]\in\mathbb{C}^{4n,2}$, $\mathcal{V}_3=\mathcal{X}_c^H\left[\begin{smallmatrix}w&0\\0&\bar{z}\end{smallmatrix}\right]\in\mathbb{C}^{4n,2}$, and $\tilde{C}_1=\left[\begin{smallmatrix}\alpha_2&0\\0&\bar{\alpha}_1\end{smallmatrix}\right]\in\mathbb{C}^{2,2}$, $\tilde{C}_3=\left[\begin{smallmatrix}\gamma_2&0\\0&\bar{\gamma}_1\end{smallmatrix}\right]\in\mathbb{C}^{2,2}$, $C_3=\left[\begin{smallmatrix}\gamma_1&0\\0&\bar{\gamma}_2\end{smallmatrix}\right]\in\mathbb{C}^{2,2}$.

From $\mathcal{V}_2^H\mathcal{J}\mathcal{U}_3=\left[\begin{smallmatrix}v^H\mathcal{J}z&0\\0&y^T\mathcal{J}\bar{w}\end{smallmatrix}\right]$, it follows that

$$\left|\frac{\hat{\lambda}_k-\lambda}{\lambda}\right|\leq\frac{\max\{|\gamma_1|,|\gamma_2|\}\left\|\mathcal{E}_{\mathcal{Z}}\right\|_2+\frac{1}{|\lambda|}\left\|\mathcal{E}_{\mathcal{H}}\right\|_2}{\min\{|\bar{\alpha}_2\bar{\gamma}_2 v^H\mathcal{J}z|,|\alpha_1\gamma_1 w^H\mathcal{J}y|\}}+\frac{\left\|\mathcal{E}_{\mathcal{Z}}\right\|_2}{\min\{|\bar{\alpha}_2 v^H\mathcal{J}z|,|\alpha_1 w^H\mathcal{J}y|\}}+O(\varepsilon^2).$$

From (5.12) and (5.13), we also have

$$\bar{\alpha}_2 v^H\mathcal{J}z=\gamma_1 u^H\mathcal{J}x,\qquad \bar{\gamma}_2 u^H\mathcal{J}x=\alpha_1 w^H\mathcal{J}y,\qquad \bar{\beta}_2 v^H\mathcal{J}z=-\beta_1 w^H\mathcal{J}y. \tag{5.14}$$

It follows that

$$|\bar{\alpha}_2\bar{\gamma}_2 v^H\mathcal{J}z|=|\bar{\gamma}_2\gamma_1 u^H\mathcal{J}x|=|\gamma_1\alpha_1 w^H\mathcal{J}y|.$$

Hence,

$$\frac{\max\{|\gamma_1|,|\gamma_2|\}}{\min\{|\bar{\alpha}_2\bar{\gamma}_2 v^H\mathcal{J}z|,|\alpha_1\gamma_1 w^H\mathcal{J}y|\}}=\frac{1}{\min\{|\bar{\alpha}_2 v^H\mathcal{J}z|,|\alpha_1 w^H\mathcal{J}y|\}},$$

$$|\lambda|\min\{|\bar{\alpha}_2\bar{\gamma}_2 v^H\mathcal{J}z|,|\alpha_1\gamma_1 w^H\mathcal{J}y|\}=|\beta_1 w^H\mathcal{J}y|,$$

and

$$\left|\frac{\hat{\lambda}_k - \lambda}{\lambda}\right| \leq \varepsilon \left(\frac{c_{\mathcal{H}}}{|\beta_1 w^H \mathcal{J} y|} \|\mathcal{H}\|_2 + \frac{2c_{\mathcal{Z}}}{\min\{|\bar{\alpha}_2 v^H \mathcal{J} z|, |\alpha_1 w^H \mathcal{J} y|\}} \|\mathcal{Z}\|_2 \right) + O(\varepsilon^2).$$

Equation (5.14) implies that $\bar{\alpha}_2 v^H \mathcal{J} z = \gamma_1 u^H \mathcal{J} x$. The first part of the theorem follows.

If λ is purely imaginary, the proof is analogous.

If $\lambda = \infty$, then $\alpha_1 = 0$ or $\gamma_1 = 0$ and $\beta_1 \neq 0$. Using the first equation of (5.14), we have $\bar{\alpha}_1 y^H \mathcal{J} z = \gamma_1 x^H \mathcal{J} x$, where we have replaced u, v, and α_2 by x, y, and α_1, respectively ((5.12) and (5.13) are the same now). Since λ is simple, i.e., $y^H \mathcal{J} z \neq 0$ and $x^H \mathcal{J} x \neq 0$, we have $\alpha_1 = \gamma_1 = 0$ and hence

$$C_1 = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \bar{\alpha}_1 \end{bmatrix} = 0, \qquad C_3 = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \bar{\gamma}_1 \end{bmatrix} = 0, \qquad C_2 = \begin{bmatrix} \beta_1 & 0 \\ 0 & \bar{\beta}_1 \end{bmatrix} \neq 0.$$

Therefore,

$$E_\infty := C_1^H C_2^{-H} \mathcal{U}_3^H \mathcal{E}_{\mathcal{Z}}^H \mathcal{J} \mathcal{U}_1 - \mathcal{U}_1^H \mathcal{J} \mathcal{E}_{\mathcal{Z}} \mathcal{U}_3 C_2^{-1} C_1 - C_1^H C_2^{-H} \mathcal{U}_3^H \mathcal{J} \mathcal{E}_{\mathcal{H}} \mathcal{U}_3 C_2^{-1} C_1 = 0.$$

From [9, Theorem 23(b)], we get

$$\left|\frac{1}{\hat{\lambda}_k}\right| \leq \left\|(\mathcal{U}_1^H \mathcal{J} \mathcal{U}_1)^{-1} E_\infty\right\|_2 + O(\varepsilon^2) = O(\varepsilon^2). \qquad \square \tag{5.15}$$

If the matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ with $\mathcal{J}$-semidefinite skew-Hamiltonian part $\mathcal{S} = \mathcal{J}\mathcal{Z}^H \mathcal{J}^T \mathcal{Z}$ has semisimple, multiple, infinite eigenvalues, then the perturbation bound (5.15) weakens to $O(\varepsilon)$ [9].

To study the perturbations in the computed deflating subspaces we need to study the perturbations for the extended matrix pencil in more detail. As mentioned before, by applying Algorithm 2 to $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ we actually compute a unitary matrix $\hat{\mathcal{Q}}$ such that

$$\begin{aligned} \mathcal{J}\hat{\mathcal{Q}}^H \mathcal{J}^T (\alpha\hat{\mathcal{B}}_{\mathcal{S}}^c - \beta\hat{\mathcal{B}}_{\mathcal{H}}^c)\hat{\mathcal{Q}} &= \alpha\hat{\mathcal{R}}_{\mathcal{S}} - \beta\hat{\mathcal{R}}_{\mathcal{H}} \\ &=: \alpha \begin{bmatrix} \hat{\mathcal{S}}_{11} & \hat{\mathcal{S}}_{12} \\ 0 & \hat{\mathcal{S}}_{11}^H \end{bmatrix} - \beta \begin{bmatrix} \hat{\mathcal{H}}_{11} & \hat{\mathcal{H}}_{12} \\ 0 & -\hat{\mathcal{H}}_{11}^H \end{bmatrix}, \end{aligned} \tag{5.16}$$

where $\hat{\mathcal{B}}_{\mathcal{S}}^c$ and $\hat{\mathcal{B}}_{\mathcal{H}}^c$ are defined in (5.1) and (5.2), and $\Lambda(\hat{\mathcal{S}}_{11}, \hat{\mathcal{H}}_{11}) = \Lambda_-(\hat{\mathcal{B}}_{\mathcal{S}}^c, \hat{\mathcal{B}}_{\mathcal{H}}^c)$. If we assume that the matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ has no purely imaginary eigenvalues, then by Theorem 2.6 there exist unitary matrices $\mathcal{Q}_1$, $\mathcal{Q}_2$ such that

$$\mathcal{J}\mathcal{Q}_1^H \mathcal{J}^T (\alpha\mathcal{S} - \beta\mathcal{H})\mathcal{Q}_1 = \alpha \begin{bmatrix} S_{11}^- & S_{12}^- \\ 0 & (S_{11}^-)^H \end{bmatrix} - \beta \begin{bmatrix} H_{11}^- & H_{12}^- \\ 0 & -(H_{11}^-)^H \end{bmatrix}$$

with $\Lambda(S_{11}^-, H_{11}^-) = \Lambda_-(\mathcal{S}, \mathcal{H})$, and

$$\mathcal{J}\mathcal{Q}_2^H \mathcal{J}^T (\alpha\mathcal{S} - \beta\mathcal{H})\mathcal{Q}_2 = \alpha \begin{bmatrix} S_{11}^+ & S_{12}^+ \\ 0 & (S_{11}^+)^H \end{bmatrix} - \beta \begin{bmatrix} H_{11}^+ & H_{12}^+ \\ 0 & -(H_{11}^+)^H \end{bmatrix}$$

with $\Lambda(S_{11}^+, H_{11}^+) = \Lambda_+(\mathcal{S}, \mathcal{H})$, respectively. Set $\mathcal{Q} = \mathcal{X}_c^H \operatorname{diag}(\mathcal{Q}_1, \bar{\mathcal{Q}}_2)\mathcal{P}$ with $\mathcal{P}$ and $\mathcal{X}_c$ as in (2.10) and (2.11). Then $\mathcal{Q}$ is unitary and

$$
\begin{aligned}
&\mathcal{J}\mathcal{Q}^H\mathcal{J}^T(\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c)\mathcal{Q} \\
&= \alpha \left[\begin{array}{cc|cc} S_{11}^- & 0 & S_{12}^- & 0 \\ 0 & \overline{S_{11}^+} & 0 & \overline{S_{12}^+} \\ \hline 0 & 0 & (S_{11}^-)^H & 0 \\ 0 & 0 & 0 & (\overline{S_{11}^+})^H \end{array}\right] - \beta \left[\begin{array}{cc|cc} H_{11}^- & 0 & H_{12}^- & 0 \\ 0 & -\overline{H_{11}^+} & 0 & -\overline{H_{12}^+} \\ \hline 0 & 0 & -(H_{11}^-)^H & 0 \\ 0 & 0 & 0 & (\overline{H_{11}^+})^H \end{array}\right] \\
&=: \alpha \left[\begin{array}{cc} \mathcal{S}_{11} & \mathcal{S}_{12} \\ 0 & \mathcal{S}_{11}^H \end{array}\right] - \beta \left[\begin{array}{cc} \mathcal{H}_{11} & \mathcal{H}_{12} \\ 0 & -\mathcal{H}_{11}^H \end{array}\right] \\
&=: \alpha\mathcal{R}_{\mathcal{S}} - \beta\mathcal{R}_{\mathcal{H}}.
\end{aligned}
\tag{5.17}
$$

This is the structured Schur form of the extended skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$. Moreover, $\Lambda(\mathcal{S}_{11}, \mathcal{H}_{11}) = \Lambda_-(\mathcal{B}_{\mathcal{S}}^c, \mathcal{B}_{\mathcal{H}}^c)$.

In the following, we will use the linear space $\mathbb{C}^{n,n} \times \mathbb{C}^{n,n}$ endowed with the norm

$$\|(X, Y)\| = \max\{\|X\|_2, \|Y\|_2\}.$$

THEOREM 5.3. *Let $\alpha\mathcal{S} - \beta\mathcal{H}$ be a regular skew-Hamiltonian/Hamiltonian matrix pencil with neither infinite nor purely imaginary eigenvalues. Let $\mathcal{P}_V^-$ be the orthogonal basis of the deflating subspace of $\alpha\mathcal{S} - \beta\mathcal{H}$ corresponding to $\Lambda_-(\mathcal{S}, \mathcal{H})$, and let $\hat{\mathcal{P}}_V^-$ be the perturbation of $\mathcal{P}_V^-$ obtained by Algorithm 2 in finite precision arithmetic. Denote by $\Theta \in \mathbb{C}^{n,n}$ the diagonal matrix of canonical angles between $\mathcal{P}_V^-$ and $\hat{\mathcal{P}}_V^-$.*

Using the structured Schur form of the extended skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c$ (as in (2.17) and (2.22)) given by (5.17), define δ by

$$\delta = \min_{Y \in \mathbb{C}^{2n,2n} \setminus \{0\}} \frac{\left\|(\mathcal{H}_{11}^H Y + Y^H \mathcal{H}_{11}, \mathcal{S}_{11}^H Y - Y^H \mathcal{S}_{11})\right\|}{\|Y\|_2}. \tag{5.18}$$

If

$$8\,\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|\,(\delta + \|(\mathcal{S}_{12}, \mathcal{H}_{12})\|) < \delta^2, \tag{5.19}$$

then

$$\|\Theta\|_2 < c_b \frac{\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|}{\delta} < c_b \varepsilon \frac{\|(c_{\mathcal{S}}\mathcal{S}, c_{\mathcal{H}}\mathcal{H})\|}{\delta}, \tag{5.20}$$

where $c_{\mathcal{S}}$ and $c_{\mathcal{H}}$ are the modest constants in (5.3)–(5.4) and $c_b = 8(\sqrt{10}+4)/(\sqrt{10}+2) \approx 11.1$.

Proof. Let $\alpha\hat{\mathcal{R}}_{\mathcal{S}} - \beta\hat{\mathcal{R}}_{\mathcal{H}}$, $\hat{\mathcal{Q}}$ be the output of Step 2 in Algorithm 2 in finite precision arithmetic, where $\hat{\mathcal{B}}_{\mathcal{S}}^c$, $\hat{\mathcal{B}}_{\mathcal{H}}^c$ satisfy (5.1) and (5.2). Let $\tilde{\mathcal{Q}}$ be the unitary matrix computed by Algorithm 2 in exact arithmetic such that

$$
\begin{aligned}
\mathcal{J}\tilde{\mathcal{Q}}^H\mathcal{J}^T(\alpha\mathcal{B}_{\mathcal{S}}^c - \beta\mathcal{B}_{\mathcal{H}}^c)\tilde{\mathcal{Q}} &= \alpha\tilde{\mathcal{R}}_{\mathcal{S}} - \beta\tilde{\mathcal{R}}_{\mathcal{H}} \\
&= \alpha \left[\begin{array}{cc} \tilde{\mathcal{S}}_{11} & \tilde{\mathcal{S}}_{12} \\ 0 & \tilde{\mathcal{S}}_{11}^H \end{array}\right] - \beta \left[\begin{array}{cc} \tilde{\mathcal{H}}_{11} & \tilde{\mathcal{H}}_{12} \\ 0 & -\tilde{\mathcal{H}}_{11}^H \end{array}\right],
\end{aligned}
$$

with $\Lambda(\tilde{\mathcal{S}}_{11}, \tilde{\mathcal{H}}_{11}) = \Lambda_-(\mathcal{B}_{\mathcal{S}}^c, \mathcal{B}_{\mathcal{H}}^c)$. Since (5.17) is another structured Schur form with the same eigenvalue ordering, there exists a unitary diagonal matrix $\mathcal{G} = \operatorname{diag}(G_1, G_2)$

such that $\mathcal{Q} = \tilde{\mathcal{Q}}\mathcal{G}$. Therefore, we have

$$\left\|(\tilde{\mathcal{S}}_{12}, \tilde{\mathcal{H}}_{12})\right\| = \|(\mathcal{S}_{12}, \mathcal{H}_{12})\|,$$

and for δ given in (5.18) we also have

$$\delta = \min_{Y \in \mathbb{C}^{2n,2n} \setminus \{0\}} \frac{\left\|(\tilde{\mathcal{H}}_{11}^H Y + Y^H \tilde{\mathcal{H}}_{11}, \tilde{\mathcal{S}}_{11}^H Y - Y^H \tilde{\mathcal{S}}_{11})\right\|}{\|Y\|_2}.$$

Let

$$\tilde{\mathcal{E}}_{\mathcal{S}} := \mathcal{J}\tilde{\mathcal{Q}}^H \mathcal{J}^T \mathcal{E}_{\mathcal{S}} \tilde{\mathcal{Q}} =: \begin{bmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} \\ \mathcal{E}_{21} & \mathcal{E}_{11}^H \end{bmatrix}, \qquad \tilde{\mathcal{E}}_{\mathcal{H}} := \mathcal{J}\tilde{\mathcal{Q}}^H \mathcal{J}^T \mathcal{E}_{\mathcal{H}} \tilde{\mathcal{Q}} =: \begin{bmatrix} \mathcal{F}_{11} & \mathcal{F}_{12} \\ \mathcal{F}_{21} & -\mathcal{F}_{11}^H \end{bmatrix}$$

and set $\gamma = \|(\mathcal{E}_{21}, \mathcal{F}_{21})\|$, $\eta = \|(\tilde{\mathcal{S}}_{12} + \mathcal{E}_{12}, \tilde{\mathcal{H}}_{12} + \mathcal{F}_{12})\|$, and $\tilde{\delta} = \delta - 2\|(\mathcal{E}_{11}, \mathcal{F}_{11})\|$. Since we have $\|(\tilde{\mathcal{E}}_{\mathcal{S}}, \tilde{\mathcal{E}}_{\mathcal{H}})\| = \|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|$, condition (5.19) implies that

$$\tilde{\delta} \geq \delta - 2\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\| > \frac{3}{4}\delta,$$

and clearly

$$4\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\| \|(\mathcal{S}_{12}, \mathcal{H}_{12})\| < \delta^2 - 4\delta\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|.$$

Hence

$$\begin{aligned} \frac{\gamma\eta}{\tilde{\delta}^2} &\leq \frac{\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\| \{\|(\tilde{\mathcal{S}}_{12}, \tilde{\mathcal{H}}_{12})\| + \|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|\}}{(\delta - 2\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|)^2} \\ &< \frac{\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|^2 + (\delta^2 - 4\delta\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|)/4}{(\delta - 2\|(\mathcal{E}_{\mathcal{S}}, \mathcal{E}_{\mathcal{H}})\|)^2} = \frac{1}{4}. \end{aligned}$$

Following the perturbation analysis for a formal product of matrices in [9], it can be shown that there exists a unitary matrix

$$\mathcal{W} = \begin{bmatrix} (I + W^H W)^{-\frac{1}{2}} & -W^H (I + WW^H)^{-\frac{1}{2}} \\ W(I + W^H W)^{-\frac{1}{2}} & (I + WW^H)^{-\frac{1}{2}} \end{bmatrix}$$

with

$$\|W\|_2 < 2\frac{\gamma}{\tilde{\delta}} < \frac{8}{3}\frac{\gamma}{\delta} < \frac{1}{3} \tag{5.21}$$

such that

$$\mathcal{J}(\tilde{\mathcal{Q}}\mathcal{W})^H \mathcal{J}^T (\alpha\hat{\mathcal{B}}_{\mathcal{S}}^c - \beta\hat{\mathcal{B}}_{\mathcal{H}}^c)(\tilde{\mathcal{Q}}\mathcal{W})$$

is another structured Schur form of the perturbed matrix pencil. Since there are neither infinite nor purely imaginary eigenvalues, (5.16) implies that $\hat{\mathcal{Q}}^H \tilde{\mathcal{Q}}\mathcal{W}$ is unitary block diagonal.

Without loss of generality we may take $\hat{\mathcal{Q}} = \tilde{\mathcal{Q}}\mathcal{W}$. If $\mathcal{X}_c$ is as in (2.11) and $\mathcal{X}_c\tilde{\mathcal{Q}} = \left[\begin{smallmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{21} & \mathcal{Q}_{22} \end{smallmatrix}\right]$, then it follows from Theorem 3.1 that $\mathcal{P}_V^- = \operatorname{range} \mathcal{Q}_{11}$. Clearly $\hat{\mathcal{P}}_V^- = \operatorname{range}\{(\mathcal{Q}_{11} + \mathcal{Q}_{12}W)(I + W^H W)^{-\frac{1}{2}}\}$. The upper bound (5.20) can then be

derived from (5.21) by using the same argument as in the proof of Theorem 4.4 in [6]. ∎

If $\mathcal{S}$ is given in factored form, then we obtain a similar result. In this case, by using Algorithm 1 we compute a unitary matrix $\hat{\mathcal{Q}}$ and a unitary symplectic matrix $\hat{\mathcal{U}}$ such that

$$\hat{\mathcal{U}}^H \hat{\mathcal{B}}^c_{\mathcal{Z}} \hat{\mathcal{Q}} = \hat{\mathcal{R}}_{\mathcal{Z}} =: \begin{bmatrix} \hat{\mathcal{Z}}_{11} & \hat{\mathcal{Z}}_{12} \\ 0 & \hat{\mathcal{Z}}_{22} \end{bmatrix},$$

$$\mathcal{J}\hat{\mathcal{Q}}^H \mathcal{J}^T \hat{\mathcal{B}}^c_{\mathcal{H}} \hat{\mathcal{Q}} = \hat{\mathcal{R}}_{\mathcal{H}} =: \begin{bmatrix} \hat{\mathcal{H}}_{11} & \hat{\mathcal{H}}_{12} \\ 0 & -\hat{\mathcal{H}}_{11}^H \end{bmatrix}, \tag{5.22}$$

where $\hat{\mathcal{B}}^c_{\mathcal{Z}}$ and $\hat{\mathcal{B}}^c_{\mathcal{H}}$ are defined in (5.10) and (5.11), and $\Lambda(\hat{\mathcal{Z}}_{22}^H \hat{\mathcal{Z}}_{11}, \hat{\mathcal{H}}_{11}) = \Lambda_-(\hat{\mathcal{B}}^c_{\mathcal{S}}, \hat{\mathcal{B}}^c_{\mathcal{H}})$, where $\hat{\mathcal{B}}^c_{\mathcal{S}} = \mathcal{J}(\hat{\mathcal{B}}^c_{\mathcal{Z}})^H \mathcal{J}^T \hat{\mathcal{B}}^c_{\mathcal{Z}}$.

Analogous to Theorem 2.7, if $\alpha\mathcal{S} - \beta\mathcal{H}$ has no purely imaginary eigenvalues, then there exist unitary matrices $\mathcal{Q}_1$, $\mathcal{Q}_2$ and unitary symplectic matrices $\mathcal{U}_1, \mathcal{U}_2$ such that

$$\mathcal{U}_1^H \mathcal{Z} \mathcal{Q}_1 = \begin{bmatrix} Z_{11}^- & Z_{12}^- \\ 0 & Z_{22}^- \end{bmatrix}, \qquad \mathcal{J}\mathcal{Q}_1^H \mathcal{J}^T \mathcal{H} \mathcal{Q}_1 = \begin{bmatrix} H_{11}^- & H_{12}^- \\ 0 & -(H_{11}^-)^H \end{bmatrix},$$

with $\Lambda((Z_{22}^-)^H Z_{11}^-, H_{11}^-) = \Lambda_-(\mathcal{S}, \mathcal{H})$, and

$$\mathcal{U}_2^H \mathcal{Z} \mathcal{Q}_2 = \begin{bmatrix} Z_{11}^+ & Z_{12}^+ \\ 0 & Z_{22}^+ \end{bmatrix}, \qquad \mathcal{J}\mathcal{Q}_2^H \mathcal{J}^T \mathcal{H} \mathcal{Q}_2 = \begin{bmatrix} H_{11}^+ & H_{12}^+ \\ 0 & -(H_{11}^+)^H \end{bmatrix},$$

with $\Lambda((Z_{22}^+)^H Z_{11}^+, H_{11}^+) = \Lambda_+(\mathcal{S}, \mathcal{H})$, respectively. Set

$$\mathcal{Q} = \mathcal{X}_c^H \operatorname{diag}(\mathcal{Q}_1, \bar{\mathcal{Q}}_2)\mathcal{P}, \qquad \mathcal{U} = \mathcal{X}_c^H \operatorname{diag}(\mathcal{U}_1, \bar{\mathcal{U}}_2)\mathcal{P},$$

where $\mathcal{P}$ and $\mathcal{X}_c$ are as in (2.10) and (2.11). Then $\mathcal{Q}$ is unitary and $\mathcal{U} \in \mathbb{US}_{4n}$, and a simple calculation yields

$$\mathcal{U}^H \mathcal{B}^c_{\mathcal{Z}} \mathcal{Q} = \left[\begin{array}{cc|cc} Z_{11}^- & 0 & Z_{12}^- & 0 \\ 0 & \overline{Z_{11}^+} & 0 & \overline{Z_{12}^+} \\ \hline 0 & 0 & Z_{22}^- & 0 \\ 0 & 0 & 0 & \overline{Z_{22}^+} \end{array}\right] =: \begin{bmatrix} \mathcal{Z}_{11} & \mathcal{Z}_{12} \\ 0 & \mathcal{Z}_{22} \end{bmatrix} =: \mathcal{R}_{\mathcal{Z}}, \tag{5.23}$$

$$\mathcal{J}\mathcal{Q}^H \mathcal{J}^T \mathcal{B}^c_{\mathcal{H}} \mathcal{Q} = \left[\begin{array}{cc|cc} H_{11}^- & 0 & H_{12}^- & 0 \\ 0 & -\overline{H_{11}^+} & 0 & -\overline{H_{12}^+} \\ \hline 0 & 0 & -(H_{11}^-)^H & 0 \\ 0 & 0 & 0 & \overline{(H_{11}^+)^H} \end{array}\right] =: \begin{bmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} \\ 0 & -\mathcal{H}_{11}^H \end{bmatrix} =: \mathcal{R}_{\mathcal{H}}. \tag{5.24}$$

This leads to the structured Schur form of the extended skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{J}(\mathcal{B}^c_{\mathcal{Z}})^H \mathcal{J}^T \mathcal{B}^c_{\mathcal{Z}} - \beta\mathcal{B}^c_{\mathcal{H}}$ with $\Lambda(\mathcal{Z}_{22}^H \mathcal{Z}_{11}, \mathcal{H}_{11}) = \Lambda_-(\mathcal{B}^c_{\mathcal{S}}, \mathcal{B}^c_{\mathcal{H}})$.

THEOREM 5.4. *Consider the regular skew-Hamiltonian/Hamiltonian matrix pencil $\alpha\mathcal{S} - \beta\mathcal{H}$ with nonsingular, $\mathcal{J}$-definite skew-Hamiltonian part $\mathcal{S} = \mathcal{J}\mathcal{Z}^H \mathcal{J}^T \mathcal{Z}$. Suppose that $\alpha\mathcal{S} - \beta\mathcal{H}$ has no eigenvalue with zero real part. Let the extended skew-Hamiltonian and Hamiltonian matrix $\mathcal{B}^c_{\mathcal{Z}}$ and $\mathcal{B}^c_{\mathcal{H}}$ be as in* (2.16) *and* (2.22)*, respectively, with structured triangular form given by* (5.23) *and* (5.24)*. Define δ_p as*

$$\delta_p = \min_{(X,Y) \in \mathbb{C}^{2n,2n} \times \mathbb{C}^{2n,2n} \setminus \{(0,0)\}} \frac{\left\| (\mathcal{H}_{11}^H Y + Y^H \mathcal{H}_{11}, X\mathcal{Z}_{11} - \mathcal{Z}_{22} Y) \right\|}{\|(X,Y)\|_2}.$$

Define errors $\mathcal{E}_{\mathcal{Z}}$ *and* $\mathcal{E}_{\mathcal{H}}$ *by* (5.10) *and* (5.11). *Let* $\mathcal{P}_V^-$, $\mathcal{P}_U^-$, $\hat{\mathcal{P}}_V^-$, *and* $\hat{\mathcal{P}}_U^-$ *be the deflating subspaces computed by Algorithm* 1 *in exact and finite precision arithmetic, respectively. Denote by* $\Theta_V, \Theta_U \in \mathbb{C}^{n,n}$ *the diagonal matrices of canonical angles between* $\mathcal{P}_V^-$ *and* $\hat{\mathcal{P}}_V^-$, $\mathcal{P}_U^-$, *and* $\hat{\mathcal{P}}_U^-$, *respectively.*

If

$$8\,\|(\mathcal{E}_{\mathcal{Z}}, \mathcal{E}_{\mathcal{H}})\|\,(\delta_p + \|(\mathcal{Z}_{12}, \mathcal{H}_{12})\|) < \delta_p^2,$$

then

$$\|\Theta_V\|_2, \|\Theta_U\|_2 < c_b \frac{\|(\mathcal{E}_{\mathcal{Z}}, \mathcal{E}_{\mathcal{H}})\|}{\delta_p} < c_b \varepsilon \frac{\|(c_{\mathcal{Z}}\mathcal{Z}, c_{\mathcal{H}}\mathcal{H})\|}{\delta_p},$$

with c_b *as in Theorem* 5.3.

Proof. The proof is analogous to the proof of Theorem 5.3. ∎

It follows that the described numerical algorithms are numerically backwards stable. These algorithms can also be used to compute deflating subspaces which contain eigenvectors associated with infinite or purely imaginary eigenvalues. By Theorem 3.1 we get partial information also in these cases, but we face the difficulty that the desired deflating subspace may not be unique or may not exist. (See the recent analysis for Hamiltonian matrices [29].)

6. Conclusion. We have presented numerical procedures for the computation of structured Schur forms, eigenvalues, and deflating subspaces of matrix pencils with matrices having a Hamiltonian and/or skew-Hamiltonian structure. These methods generalize the recently developed methods for Hamiltonian matrices which use an extended, double dimension Hamiltonian matrix that always has a Hamiltonian Schur form.

The algorithms circumvent problems with skew-Hamiltonian/Hamiltonian matrix pencils that lack a structured Schur form by embedding them in extended matrix pencils that always admit a structured Schur form. For the extended matrix pencils, the algorithms use structure-preserving unitary matrix computations and are strongly backwards stable; i.e., they compute the exact structured Schur form of a nearby matrix pencil with the same structure. Such structured Schur forms can always be computed regardless of the regularity of the original matrix pencil.

It is still somewhat unsatisfactory that the algorithms do not efficiently exploit the microstructures of the extended matrix pencils, as, for example, in the matrix $\mathcal{B}_{\mathcal{N}}^e$ in (2.12). How best to use these microstructures is still an open question.

Practical implementation and numerical experiments are in progress and will be reported elsewhere. For detailed algorithms and implementation issues see [4].

Acknowledgment. We gratefully acknowledge Daniel Kressner for his assistance implementing and testing experimental versions of parts of the algorithms discussed here.

REFERENCES

[1] G. Ammar, P. Benner, and V. Mehrmann, *A multishift algorithm for the numerical solution of algebraic Riccati equations*, Electron. Trans. Numer. Anal., 1 (1993), pp. 33–48.

[2] G. Ammar and V. Mehrmann, *On Hamiltonian and symplectic Hessenberg forms*, Linear Algebra Appl., 149 (1991), pp. 55–72.

[3] P. Benner, R. Byers, H. Fassbender, V. Mehrmann, and D. Watkins, *Cholesky-like factorizations of skew-symmetric matrices*, Electron. Trans. Numer. Anal., 11 (2000), pp. 85–93.

[4] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Numerical Computation of Deflating Subspaces of Embedded Hamiltonian pencils*, Tech. Report SFB393/99-15, Fakultät für Mathematik, TU Chemnitz, Chemnitz, Germany, 1999; available online from http://www.tu-chemnitz.de/sfb393/sfb99pr.html.

[5] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Numerical methods for linear-quadratic and H_∞ control problems*, in Dynamical Systems, Control, Coding, Computer Vision: New Trends, Interfaces, and Interplay, G. Picci and D. Gilliam, eds., Progr. Systems Control Theory 25, Birkhäuser, Basel, 1999, pp. 203–222.

[6] P. Benner, V. Mehrmann, and H. Xu, *A new method for computing the stable invariant subspace of a real Hamiltonian matrix*, J. Comput. Appl. Math., 86 (1997), pp. 17–43.

[7] P. Benner, V. Mehrmann, and H. Xu, *A numerically stable, structure preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils*, Numer. Math., 78 (1998), pp. 329–358.

[8] P. Benner, V. Mehrmann, and H. Xu, *A note on the numerical solution of complex Hamiltonian and skew-Hamiltonian eigenvalue problems*, Electron. Trans. Numer. Anal., 8 (1999), pp. 115–126.

[9] P. Benner, V. Mehrmann, and H. Xu, *Perturbation Analysis for the Eigenvalue Problem of a Formal Product of Matrices*, Berichte aus der Technomathematik, Report 00–01, FB3 – Mathematik und Informatik, Universität Bremen, Bremen, Germany, 2000; available online from http://www.math.uni-bremen.de/zetem/berichte.html.

[10] A. Bojanczyk, G. Golub, and P. Van Dooren, *The periodic Schur decomposition; algorithms and applications*, Proc. SPIE, 1770 (1992), pp. 31–42.

[11] A. Bunse-Gerstner, *Matrix factorization for symplectic QR-like methods*, Linear Algebra Appl., 83 (1986), pp. 49–77.

[12] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *Numerical methods for algebraic Riccati equations*, in Proceedings of the Workshop on the Riccati Equation in Control, Systems, and Signals, S. Bittanti, ed., Como, Italy, 1989, pp. 107–116.

[13] R. Byers, *Hamiltonian and Symplectic Algorithms for the Algebraic Riccati Equation*, Ph.D. thesis, Cornell University, Ithaca, NY, 1983.

[14] R. Byers, *A Hamiltonian QR algorithm*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.

[15] H. Fassbender, D. Mackey, N. Mackey, and H. Xu, *Hamiltonian square roots of skew-Hamiltonian matrices*, Linear Algebra Appl., 287 (1998), pp. 125–159.

[16] F. Gantmacher, *Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.

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

[18] M. Green and D. Limebeer, *Linear Robust Control*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[19] J. Hench and A. Laub, *Numerical solution of the discrete-time periodic Riccati equation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1197–1210.

[20] P. Lancaster, *Strongly stable gyroscopic systems*, Electron. J. Linear Algebra, 5 (1999), pp. 53–66.

[21] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.

[22] A. Laub, *Invariant subspace methods for the numerical solution of Riccati equations*, in The Riccati Equation, S. Bittanti, A. Laub, and J. Willems, eds., Springer-Verlag, Berlin, 1991, pp. 163–196.

[23] W.-W. Lin and T.-C. Ho, *On Schur Type Decompositions for Hamiltonian and Symplectic Pencils*, Tech. Report, Institute of Applied Mathematics, National Tsing Hua University, Taiwan, 1990.

[24] W.-W. Lin, V. Mehrmann, and H. Xu, *Canonical forms for Hamiltonian and symplectic matrices and pencils*, Linear Algebra Appl., 301–303 (1999), pp. 469–533.

[25] C. Mehl, *Compatible Lie and Jordan Algebras and Applications to Structured Matrices and Pencils*, dissertation, Fakultät für Mathematik, TU Chemnitz, Chemnitz, Germany, 1998.

[26] C. Mehl, *Condensed forms for skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 454–476.

[27] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Heidelberg, 1991.

[28] V. Mehrmann and D. Watkins, *Structure-preserving methods for computing eigenpairs of large sparse skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Sci. Comput., 22 (2001), pp. 1905–1925; also available online from http://www.tu-chemnitz.de/sfb393/sfb00pr.html.

[29] V. Mehrmann and H. Xu, *Lagrangian Invariant Subspaces of Hamiltonian Matrices*, Tech. Report SFB393/98-25, Fakultät für Mathematik, TU Chemnitz, Chemnitz, Germany, 1998;

available online from http://www.tu-chemnitz.de/sfb393/sfb98pr.html.
[30] J. Olson, H. Jensen, and P. Jørgensen, *Solution of large matrix equations which occur in response theory*, J. Comput. Phys., 74 (1988), pp. 265–282.
[31] C. Paige and C. Van Loan, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 14 (1981), pp. 11–32.
[32] R. Patel, *On computing the eigenvalues of a symplectic pencil*, Linear Algebra Appl., 188/189 (1993), pp. 591–611.
[33] P. Petkov, N. Christov, and M. Konstantinov, *Computational Methods for Linear Control Systems*, Prentice-Hall, Hertfordshire, UK, 1991.
[34] A.-M. Sändig and W. Wendland, *Asymptotic expansions of elastic fields in domains with boundary and structural singularities*, in Boundary Element Topics (Stuttgart, 1995), Springer, Berlin, 1997, pp. 419–444.
[35] V. Sima, *Algorithms for Linear-Quadratic Optimization*, Pure Appl. Math. 200, Marcel Dekker, New York, 1996.
[36] G. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[37] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[38] C. Van Loan, *A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix*, Linear Algebra Appl., 61 (1984), pp. 233–251.
[39] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River, NJ, 1995.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 191–218

ALGEBRAIC MULTILEVEL METHODS AND SPARSE APPROXIMATE INVERSES*

MATTHIAS BOLLHÖFER† AND VOLKER MEHRMANN†

Abstract. In this paper we introduce a new approach to algebraic multilevel methods and their use as preconditioners in iterative methods for the solution of symmetric positive definite linear systems. The multilevel process and, in particular, the coarsening process are based on the construction of sparse approximate inverses and their augmentation with corrections of smaller size. We present comparisons of the effectiveness of the resulting multilevel technique and numerical results.

Key words. sparse approximate inverse, large sparse matrices, algebraic multilevel method

AMS subject classifications. 65F05, 65F10, 65F50, 65Y05

PII. S0895479899364441

1. Introduction.

For the solution of large sparse linear systems of the form

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

sparse approximate inverses, i.e., sparse matrices that are good approximations of the inverse of a sparse matrix [22, 21, 10, 15, 5], have become as popular as preconditioners for Krylov-subspace [12, 30, 14] techniques. There are several techniques to construct such sparse approximate inverses. One may, for example, minimize the norm of $\|AB - I\|$ subject to some prescribed pattern [21, 10, 15]. Another technique is to construct upper triangular matrices $Z, W^\top$ such that for a diagonal matrix D, $W^\top AZ$ is a good approximation to D [5]. Moreover, success has been made over the years in using approximate inverses in combination with multilevel methods [11, 25, 24, 32, 33]. Especially in [33] it has been shown that by adjusting the quality of the approximate inverse, the smoothing property can be improved significantly.

We assume in the following that A is symmetric positive definite and that the approximate inverse B is factored as $B = LL^\top$. We set $M = L^\top AL$ and assume for simplicity that $\|M\|_2 \leqslant 1$. This can always be achieved by an appropriate scaling. We will concentrate on sparse approximate inverses for which M is still sparse. This is, for example, the case if the approximate inverse is diagonal or block diagonal. Even factored sparse approximate inverses from [21, 22] can be used as long as the pattern of L is moderate—for example, if the pattern of L is the same as the pattern of A (or the same pattern as the lower triangular part of A). There also exist sparse approximate inverse approaches that cannot be applied here, because they are only sparse with respect to certain basis transformations like wavelet-based sparse approximate inverses [9]. For large classes of matrices, sparse approximate inverses have proved very effective as preconditioners. But there are problems where the sparse

*Received by the editors November 19, 1999; accepted for publication (in revised form) by R. Freund October 30, 2001; published electronically July 1, 2002.
http://www.siam.org/journals/simax/24-1/36444.html

†Institut für Mathematik, MA 4-5, TU Berlin, D–10623 Berlin, Germany (bolle@math.tu-berlin.de, http://www.math.tu-berlin.de/∼bolle, mehrmann@math.tu-berlin.de, http://www.math.tu-berlin.de/∼mehrmann/). The research of the first author was supported by the DFG under grant BO 1680/1-1 and by the University of Minnesota. Part of this research was performed while the first author was visiting the University of Minnesota at Minneapolis. The research of the second author was supported by SFB 393 "Numerische Simulation auf massiv parallelen Rechnern."

approximate inverse needs a large number of nonzero entries to become a suitable approximation to the inverse of A. When using sparse approximate inverses based on norm-minimizing techniques, one often observes that many eigenvalues [13] of the residual matrix $E = I - M$ are quite small, while a small number of eigenvalues stay big. And allowing more fill-in in the sparse approximate inverse B does not cure this. For an example, see [6].

The observation that many eigenvalues are small but some stay large means that B approximates A^{-1} well on a subspace of large size, while there is almost no approximation on the complementary subspace. In the context of multigrid methods for the numerical solution of partial differential equations, this effect is typically called the smoothing property [16]. Algebraically this means that the residual $E = I - M$ can be written as

$$E = E_p + F, \tag{1.2}$$

where $E_p \in \mathbb{R}^{n,n}$ has rank $p < n$ and $\|F\| \leq \eta \ll 1$, i.e., the residual can be approximated well by a matrix of lower rank p. Typically one cannot expect that the size p of E_p is independent of the dimension n of A. More realistic is the assumption that $p \approx cn$, where, for example, $c = \frac{1}{2}$ or $c = \frac{1}{4}$.

If one is solving a symmetric positive definite linear system $Ax = b$ and one has already determined some sparse approximate inverse B, it is therefore desirable (and our primary goal) to improve the preconditioner $LL^\top$. Our goal is to construct an updated preconditioner of the form

$$L(I + PZ^{-1}P^\top)L^\top \tag{1.3}$$

with sparse matrices P, Z, where Z is another symmetric positive definite matrix of smaller size. Since A and the augmented preconditioner are positive definite, this means that we are interested in the small eigenvalues (since $\|M\|_2 \leqslant 1$) of the preconditioned system

$$AL(I + PZ^{-1}P^\top)L^\top. \tag{1.4}$$

In other words we have to achieve that

$$\|I - M^{1/2}(I + PZ^{-1}P^\top)M^{1/2}\|_2 = \|E - M^{1/2}PZ^{-1}P^\top M^{1/2}\|_2 \tag{1.5}$$

is small, while at the same time P and Z are sparse.

Since the matrix E is symmetric positive semidefinite by assumption, it is well known [13] that the best approximation of E by a matrix of rank p is given by the matrix

$$\hat{E}_p = U_p \Sigma_p U_p^\top = \begin{bmatrix} u_1, \ldots, u_p \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \end{bmatrix} \begin{bmatrix} u_1 & \ldots & u_p \end{bmatrix}^\top, \tag{1.6}$$

where $\sigma_1 \geqslant \cdots \sigma_n \geqslant 0$ are the eigenvalues of E and u_i, $i = 1, \ldots, n$, are the eigenvectors.

But in general this best approximation will be a full matrix, since U_p is full even if E is sparse, and hence we cannot directly use $\hat{E}_p$ in the construction of sparse preconditioners.

Since we have assumed that the given approximate inverse B has the property that $E = I - L^\top AL$ is approximated well by $\hat{E}_p$ in the sense of (1.2), we have that the entries of $\hat{E}_p$ differ only slightly from the entries of E. So we may expect that taking an appropriate selection of columns of E as V will be a good choice for U and the approximation of E by a lower rank matrix. This expectation is justified by the following lemma.

LEMMA 1.1. *Let $E \in \mathbb{R}^{n,n}$ be symmetric positive semidefinite and let*

$$E = U\Sigma U^\top = [U_1, U_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} [U_1, U_2]^\top$$

be the spectral decomposition of E, where U is orthogonal, $U_1 \in \mathbb{R}^{n,p}$, and the diagonal entries of Σ are ordered in decreasing order. If E satisfies (1.2) (i.e., $E = E_p + F$ for a rank-p matrix E_p and $\|F\|_2 \leqslant \eta$), then there exists a permutation matrix $\Pi = [\Pi_1, \Pi_2]$, partitioned analogously, such that

$$\inf_{X \in \mathbb{R}^{p,p}} \|U_1 X - M^{1/2}\ (E\Pi_1)\|_2 \leqslant \eta. \tag{1.7}$$

Proof. Applying the QR decomposition with column pivoting [13] to $M^{1/2}\hat{E}_p = U_1(I - \Sigma_1)^{1/2}\Sigma_1 U_1^\top$, we obtain $Q, R^\top \in \mathbb{R}^{n,p}$, where Q is orthogonal, $R = [R_1, R_2]$, with $R_1 \in \mathbb{R}^{p,p}$, is upper triangular, and $\Pi = [\Pi_1, \Pi_2]$ is a permutation matrix with Π_1 having p columns such that

$$M^{1/2}\hat{E}_p\Pi = QR.$$

It immediately follows that $M^{1/2}\hat{E}_p\Pi_1 = QR_1$ and thus there exists a nonsingular $p \times p$ matrix X such that $M^{1/2}\hat{E}_p\Pi_1 = QR_1 = U_1 X$, and we have

$$\begin{aligned} \|M^{1/2}E\Pi_1 - U_1 X\|_2 &= \|M^{1/2}(E - \hat{E}_p)\Pi_1\|_2 \leqslant \|(E - \hat{E}_p)\Pi_1\|_2 \\ &= \min_{\substack{E_p \\ \operatorname{rank} E_p = p}} \|E - E_p\|_2 \leqslant \|F\|_2 = \eta. \qquad \square \end{aligned}$$

Lemma 1.1 gives us subspaces that consist of suitably chosen columns of E, which are close to the subspace U_1 of E associated with the large eigenvalues of E in the sense of (1.7).

Using such subspaces in the construction of appropriate sparse representations of the updates $PZ^{-1}P^\top$ as in (1.3) is the topic of this paper, which is organized as follows.

We first discuss the theoretical background for this problem, i.e., to construct optimal preconditioners of this form, and show that they are closely related to algebraic multilevel methods. We derive two types (multiplicative and additive) of algebraic multilevel preconditioners in section 2.

The approximation properties of the multiplicative correction term $I + PZ^{-1}P^\top$ in (1.3) for the two multilevel schemes are studied in detail in section 3.

In view of Lemma 1.1, we may in principle use a QR-like decomposition of $M^{1/2}E$ to construct the desired updated preconditioners. The key in this construction is the appropriate pivoting strategy in the QR decomposition with column pivoting. We will present two heuristic pivoting strategies and interpret them as the coarsening process of the multilevel scheme in section 4.

Finally in section 5 we present numerical examples that demonstrate the properties of this new approach and also indicate the effectiveness of the heuristics that have been used.

In what follows, for symmetric matrices A, B we will use the notation $A \succeq B$ if $A - B$ has nonnegative eigenvalues. We also identify a matrix with the space spanned by its columns.

2. Multilevel preconditioners. In this section we present two multilevel preconditioners for symmetric positive definite systems. Algebraic multilevel preconditioners have become popular in recent years. Several algebraic multigrid (AMG) approaches focus on incomplete LU or Schur-complement approaches [2, 3, 34, 4, 27, 28], while others are based on the analogy to geometric multigrid methods [8, 29, 20, 18, 26, 19]. Here we will concentrate on the second class of approaches.

Let $A \in \mathbb{R}^{n,n}$ be symmetric positive definite and let $L \in \mathbb{R}^{n,n}$ be a given sparse matrix such that $LL^\top$ is a symmetric positive definite matrix in factored form that approximates A^{-1}.

Suppose that the approximation of A^{-1} by $LL^\top$ is not satisfactory, e.g., the condition number of $L^\top AL$ is not small enough to get good convergence in the conjugate gradient method, and we wish to improve the preconditioning properties. To do this we like to determine a matrix of the form

$$M^{(1)} = LL^\top + PZ^{-1}P^\top, \tag{2.1}$$

with $P \in \mathbb{R}^{n,p}$, $Z \in \mathbb{R}^{p,p}$ nonsingular, P, Z sparse, and, furthermore, $p \leqslant cn$ with $0 < c < 1$, so that $M^{(1)}$ is a better approximation to A^{-1} than $LL^\top$.

The particular form (2.1) is chosen close to the form of an algebraic two-level method, where multiplication with $P, P^\top$ corresponds to the mapping between fine and coarse grids and Z represents the coarse grid system. Note further that using the representation $LL^\top + PZ^{-1}P^\top$ as a preconditioner for A, only a system with Z has to be solved. As shown in Lemma 1.1, skillfully chosen columns/rows of the residual matrix $E = I - L^\top AL$ can be used to approximate the invariant subspace of E associated with its large eigenvalues. As we will see, precisely this invariant subspace has to be approximated by P. In the sense of the underlying undirected graph of E, we refer to the nodes associated with the columns/rows of E that will be used to approximate the invariant subspace of E associated with the largest eigenvalues as *coarse grid nodes,* while the remaining nodes are called *fine grid nodes.* The process of detecting a suitable set of coarse grid nodes will be called the *coarsening process.* Once we have selected certain nodes as coarse grid nodes, they are in a natural way embedded in the initial graph. In addition the graph of $W = P^\top AP$ is a natural graph associated with the coarse grid nodes. We will call it *coarse grid* in analogy to the notation arising in discretized partial differential equations.

Recalling the well-known techniques of constructing preconditioners for the conjugate gradient method applied to symmetric positive definite systems (e.g., [13, 17, 30]), we should choose P and Z such that

$$\mu A^{-1} \preceq M^{(1)} \preceq \mu\kappa^{(1)} A^{-1}, \tag{2.2}$$

with $\kappa^{(1)}$ as small as possible and $\mu > 0$. Clearly $\kappa^{(1)} \geqslant 1$ is the condition number of $M^{(1)}A$, i.e., the ratio of the largest by the smallest eigenvalue of $M^{(1)}A$, and thus $\kappa^{(1)} = 1$ would be optimal. The importance of the condition number is justified from the well-known results on the performance of the conjugate gradient method with preconditioner $M^{(1)}$; see, e.g., [13]. We discuss the construction of P, Z with minimal $\kappa^{(1)}$ below.

For discretized elliptic partial differential equations, often—but not always—one can construct optimal preconditioners using multigrid methods [16]. In order to obtain

a similar preconditioner augmented with a suitably chosen coarse grid correction, consider the use of $LL^\top$ in a linear iteration scheme [35] for the solution of $Ax = b$ with initial guess $x^{(0)} \in \mathbb{R}^n$. Such an iteration is given by

$$x^{(k+1)} = x^{(k)} + LL^\top(b - Ax^{(k)}), \qquad k = 0, 1, 2, \ldots.$$

The error propagation matrix $I - LL^\top A$ satisfies $x - x^{(k+1)} = (I - LL^\top A)(x - x^{(k)})$. In multilevel techniques [16] one uses such an iteration for pre- and postsmoothing and, in addition, one has to add a coarse grid correction. In terms of the error propagation matrix this means that instead of $I - LL^\top A$ we have $(I - LL^\top A)(I - PZ^{-1}P^\top A)(I - LL^\top A)^\top$ as error propagation matrix. A simple calculation shows that this product can be rewritten as $I - M^{(2)}A$ with

$$M^{(2)} = 2LL^\top - LL^\top ALL^\top + (I - LL^\top A)PZ^{-1}P^\top(I - ALL^\top). \tag{2.3}$$

Again we are interested in choosing P, Z such that

$$\mu A^{-1} \preceq M^{(2)} \preceq \mu\kappa^{(2)}A^{-1}, \tag{2.4}$$

with $\kappa^{(2)}$ as small as possible.

In the following we discuss the approximation properties of $M^{(1)}, M^{(2)}$. The first step will be the construction of optimal P, Z for given A, L based on the spectral decomposition

$$E \equiv I - L^\top AL = \Psi\Lambda\Psi^\top, \tag{2.5}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1 \geqslant \cdots \geqslant \lambda_n$, and $\Psi = [\psi_1, \ldots, \psi_n]$ is orthogonal. We use the notation $\Psi_p = [\psi_1, \ldots, \psi_p]$, $\Lambda_p = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$.

LEMMA 2.1. *Let $A, L \in \mathbb{R}^{n,n}$ with A symmetric positive definite, L nonsingular, and $E = I - L^\top AL$ positive semidefinite, and let $p < n$.*

1. *The minimal $\kappa^{(1)}$ in (2.2) is obtained with $P \in \mathbb{R}^{n,p}$, $Z \in \mathbb{R}^{p,p}$ defined via*

$$P = L[v_1, \ldots, v_p] \in \mathbb{R}^{n,p}, \quad Z = P^\top AP\left(I - P^\top AP\right)^{-1} \in \mathbb{R}^{p,p}. \tag{2.6}$$

In this case we have $\mu = 1 - \lambda_{p+1}$, $\kappa^{(1)} = (1 - \lambda_n)/(1 - \lambda_{p+1})$.

2. *For P from (2.6) and*

$$\hat{Z} = P^\top AP \tag{2.7}$$

we have

$$\gamma M^{(1)} \preceq LL^\top + P\hat{Z}^{-1}P^\top \preceq \Gamma M^{(1)}, \tag{2.8}$$

where $\gamma = 2 - \lambda_1 \geqslant 1$, $\Gamma = 2 - \lambda_p \leqslant 2$.

3. *The matrices P from (2.6) and $\hat{Z}$ from (2.7) yield the minimal $\kappa^{(2)}$ in (2.4) with $\mu = 1 - \lambda_{p+1}^2$, $\kappa^{(2)} = (1 - \lambda_n^2)/(1 - \lambda_{p+1}^2)$.*

Proof. 1. For P, Z as in (2.6) we have

$$Z = (I - E)E^{-1} = (I - \Lambda_p)\Lambda_p^{-1},$$

and condition (2.2) is equivalent to

$$\mu(I - E)^{-1} \preceq I + \Psi_p\Lambda_p(I - \Lambda_p)^{-1}\Psi_p^\top \preceq \mu\kappa^{(1)}(I - E)^{-1}. \tag{2.9}$$

Multiplying with $V^\top$ from the left and V from the right we obtain an inequality for diagonal matrices as

$$\mu \begin{bmatrix} \frac{1}{1-\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{1-\lambda_n} \end{bmatrix} \preceq \begin{bmatrix} \frac{1}{1-\lambda_1} & & & \\ & \ddots & & \\ & & \frac{1}{1-\lambda_p} & \\ & & & I \end{bmatrix} \preceq \mu\kappa^{(1)} \begin{bmatrix} \frac{1}{1-\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{1-\lambda_n} \end{bmatrix},$$

and for $\mu = 1 - \lambda_{p+1}$, $\kappa^{(1)} = (1-\lambda_n)/(1-\lambda_{p+1})$ these inequalities are satisfied. The optimality of $\kappa^{(1)}$ in (2.9) follows directly from the Courant–Fischer min–max characterization [13], which implies that $\mu \leqslant 1 - \lambda_{p+1}$ and $\mu\kappa^{(1)} \geqslant 1 - \lambda_n$. Thus the choice of $\kappa^{(1)}$ is optimal and with P, Z we obtain the optimal $\kappa^{(1)}$.

2. For $\hat{Z}$ as in (2.7), we note that we have $\lambda_i \in [0, 1)$, and therefore inequalities (2.8) immediately follow.

3. For $M^{(2)}$ we proceed analogously. The desired inequality has the form

$$\mu(I-E)^{-1} \preceq I + E + E\Psi_p(I-\Lambda_p)^{-1}\Psi_p^\top E \preceq \mu\kappa^{(2)}(I-E)^{-1}. \tag{2.10}$$

Multiplying with Ψ from the right and its transpose from the left, we obtain that

$$\Psi(I+E+E\Psi_p(I-\Lambda_p)^{-1}\Psi_p^\top E)\Psi^\top = \operatorname{diag}\left(\frac{1}{1-\lambda_1}, \ldots, \frac{1}{1-\lambda_p}, 1+\lambda_{p+1}, \ldots, 1+\lambda_n\right)$$

and the optimal choices are clearly $\mu = 1 - \lambda_{p+1}^2$ and $\mu\kappa^{(2)} = 1 - \lambda_n^2$. ∎

A similar result for $M^{(1)}$ was obtained in [26]. Note that the optimal choice $M^{(1)}$ can be viewed as approximation to A^{-1} of *first order,* since $\kappa^{(1)} \approx 1/(1-\lambda_{p+1}^1)$, while $M^{(2)}$ is an approximation of *second order,* since $\kappa^{(2)} \approx 1/(1-\lambda_{p+1}^2)$.

Lemma 2.1 shows how the optimal choices for P, Z may be computed. However, in practice we usually cannot determine these optimal choices, since the spectral decomposition is not available; even if it were available, it would be very expensive to apply, since the matrix P would be a full matrix. Instead we would like to determine P, Z (or $P, \hat{Z}$), which are inexpensive to apply and still produce good approximation properties in $M^{(1)}$ ($M^{(2)}$). By the results of Lemma 2.1 it seems natural to set $Z = P^\top AP$ or to choose Z such that

$$\gamma Z \preceq P^\top AP \preceq \Gamma Z.$$

An inequality of this form is also useful if we intend to recursively repeat the technique in a multilevel way. To do this we replace in

$$LL^\top + P(P^\top AP)^{-1}P^\top \tag{2.11}$$

the term $(P^\top AP)^{-1}$ by an additive approximation $L_1L_1^\top + P_1(P_1^\top P^\top APP_1)^{-1}P_1^\top$. For the construction of $M^{(2)}$ the procedure is analogous. Recursively applied, this idea leads to the following algebraic multilevel scheme.

Let $A \in \mathbb{R}^{n,n}$ be symmetric positive definite and let $n = n_l > n_{l-1} > \cdots > n_0 > 0$ be integers. For chosen full rank matrices $P_k \in \mathbb{R}^{n_k, n_{k-1}}$, $k = l, l-1, \ldots, 1$, define A_k via

$$A_k = \begin{cases} A, & k = l, \\ P_{k+1}^\top A_{k+1} P_{k+1}, & k = l-1, l-2, \ldots, 1. \end{cases}$$

Choose a nonsingular matrix $L_k \in \mathbb{R}^{n_k,n_k}$ such that $L_k L_k^\top \approx A_k^{-1}$, $k = 0,\dots,l$; then *multilevel sparse approximate inverse preconditioners* $M_l^{(1)}, M_l^{(2)}$ are recursively defined via

$$M_k^{(1)} = \begin{cases} A_0^{-1}, & k = 0, \\ L_k L_k^\top + P_k M_{k-1}^{(1)} P_k^\top, & k = 1,\dots,l, \end{cases} \tag{2.12}$$

and

$$M_k^{(2)} = \begin{cases} A_0^{-1}, & k = 0, \\ L_k(2I - L_k^\top A_k L_k)L_k^\top \\ \quad + (I - L_k L_k^\top A_k) P_k M_{k-1}^{(2)} P_k^\top (I - A_k L_k L_k^\top), & k = 1,2,\dots,l, \end{cases} \tag{2.13}$$

respectively.

For $l = 1$ we obviously obtain the operators $M^{(1)}$ and $M^{(2)}$ in (2.1) and (2.3), respectively.

If we exactly decompose the matrix on the coarsest level (i.e., $A_0^{-1} = L_0 L_0^\top$), for example, by the Cholesky decomposition and set $\Pi_k = P_l P_{l-1} \cdots P_{k+1}$, then we can rewrite $M_l^{(1)}$ as

$$M_l^{(1)} = \sum_{k=0}^{l} \Pi_k L_k L_k^\top \Pi_k^\top. \tag{2.14}$$

For $M_l^{(2)}$ one obtains that

$$I - M_l^{(2)} A = (I - \Pi_l L_l L_l^\top \Pi_l^\top A) \cdots (I - \Pi_0 L_0 L_0^\top \Pi_0^\top A) \cdots (I - \Pi_l L_l L_l^\top \Pi_l^\top A). \tag{2.15}$$

We see from (2.14), (2.15) that $M_l^{(1)}$ can be viewed as an *additive multilevel method,* since all the projections Π_k are formally performed simultaneously, while $M_l^{(2)}$ can be viewed as a *multiplicative multilevel method,* since the projections Π_k are performed successively. In what follows we also refer to $M_l^{(1)}$ as an additive algebraic multilevel preconditioner and to $M_l^{(2)}$ as a multiplicative algebraic multilevel preconditioner.

The operator $M_l^{(2)}$ is immediately derived from V-cycle multigrid methods in the numerical solution of partial differential equations. A special case for the operator $M_l^{(1)}$ is that $L_k L_k^\top = \frac{1}{\alpha_k} I$ is a multiple of the identity. In this case for $E = I - \alpha_k A_k$, the choice of some columns of E can be expressed by applying a permutation $\Phi_k \in \mathbb{R}^{n_k,n_{k-1}}$ to E, i.e., $P_k = (I - \alpha_k A_k)\Phi_k$. In this case $M_l^{(1)}$ reduces to

$$M_l^{(1)} = \frac{1}{\alpha_l}(I + \alpha_l P_l M_{l-1} P_l^\top) = \frac{1}{\alpha_l}\left(I + \frac{\alpha_l}{\alpha_{l-1}} P_l \left(I + \alpha_{l-1} P_{l-1} M_{l-2} P_{l-1}^\top\right) P_l^\top\right) = \cdots,$$

where the dots indicate that M_{l-2} has to be successively substituted in a similar way. For operators of this form in [19] optimal choices for α_k have been discussed according to a wisely a priori chosen permutation matrix Φ_k. Such operators have also been studied in detail in [1, 26].

3. Approximation properties. In this section we discuss the approximation properties of $M^{(1)}, M^{(2)}$ from (2.1), (2.3) for the case $l = 1$ and later for arbitrary $l \geqslant 1$.

For given Z, P we compare the approximation properties of $M^{(1)}, M^{(2)}$ in (2.2), (2.4) with the optimal choices in Lemma 2.1. For this we use the following theorem.

THEOREM 3.1 (see [17]). *Consider a symmetric positive definite matrix $M \in \mathbb{R}^{n,n}$ and matrices $P_k \in \mathbb{R}^{n,n_k}$ with $\operatorname{rank} P_k = n_k$ for $k = 1, \ldots, l$ and $\operatorname{rank}[P_1, \ldots, P_l] = n$. Consider, furthermore, positive definite matrices $B_k \in \mathbb{R}^{n_k,n_k}$ and*

$$M_S^{-1} := \sum_{k=1}^{l} P_k B_k^{-1} P_k^\top. \tag{3.1}$$

If $K > 0$ is a constant such that for every $x \in \mathbb{R}^n$ there exists a decomposition $x = \sum_{k=1}^{l} P_k x_k$ satisfying

$$\sum_{k=1}^{l} x_k^\top B_k x_k \preceq K x^\top M x, \tag{3.2}$$

then $M_S \preceq KM$.

Applying this theorem we can prove the following result.

THEOREM 3.2. *Let $A \in \mathbb{R}^{n,n}$ be symmetric positive definite and let $L \in \mathbb{R}^{n,n}$ be nonsingular such that $M = L^\top A L \preceq I$. Set $E = I - M$ and $P = LV$, where $V \in \mathbb{R}^{n,p}$ has $\operatorname{rank} V = p$ and let $W \in \mathbb{R}^{n,n-p}$ be such that $\operatorname{rank} W = n - p$ and $W^\top M V = 0$. Finally let $Z \in \mathbb{R}^{p,p}$ be symmetric positive definite such that*

$$\gamma P^\top A P \preceq Z \preceq \Gamma P^\top A P \tag{3.3}$$

with positive constants γ, Γ.

1. *If*

$$W^\top W \preceq \Delta\, W^\top M W \tag{3.4}$$

for some positive constant Δ, then for the matrix $M^{(1)}$ in (2.1) we have

$$\frac{\gamma}{\gamma + 1} A \preceq \left(M^{(1)}\right)^{-1} \preceq \max\{\Gamma, \Delta\} A. \tag{3.5}$$

2. *If in (3.3) $\gamma \geqslant 1$ and*

$$\begin{bmatrix} 0 & 0 \\ 0 & W^\top M W \end{bmatrix} \preceq \Delta\, [V, W]^\top (M - EME)\, [V, W] \tag{3.6}$$

for some positive constant Δ, then for the matrix $M^{(2)}$ in (2.3) we have

$$A \preceq \left(M^{(2)}\right)^{-1} \preceq \max\{\Gamma, \Delta\} A. \tag{3.7}$$

Proof. 1. We apply Theorem 3.1 to the matrices M, $B_1 = I$, $B_2 = Z$, $P_1 = I, P_2 = L^{-1}P = V$. Set $\Pi = P_2(P_2^\top M P_2)^{-1} P_2^\top M$ and $\Omega = I - \Pi$. Since $\Pi^\top M(I - \Pi) = 0$, we have $\Omega = W(W^\top M W)^{-1} W^\top M$. It follows that every $x \in \mathbb{R}^n$ can be written as

$$x = \underbrace{(I - \Pi)x}_{P_1 x_1} + \underbrace{\Pi x}_{P_2 x_2} = P_1 x_1 + P_2 x_2,$$

where $x_2 = (P_2^\top P_2)^{-1} P_2^\top x$ and $x_1 = \Omega x$. By Theorem 3.1 it suffices to find a constant $K > 0$ such that

$$x_1^\top x_1 + x_2^\top Z x_2 \leqslant K x^\top M x.$$

From (3.3) it follows that

$$\Omega^\top \Omega \preceq \Delta\ \Omega^\top M \Omega.$$

Substituting the representations of x_1, x_2 we obtain

$$\begin{aligned} x_1^\top x_1 + x_2^\top Z x_2 &= x^\top \Omega^\top \Omega x + x_2^\top Z x_2 \\ &\leqslant \max\{\Gamma, \Delta\}(x^\top \Omega^\top M \Omega x + x_2^\top (P_2^\top M P_2) x_2) \\ &= \max\{\Gamma, \Delta\}(x^\top \Omega^\top M \Omega x + x^\top \Pi^\top M \Pi x) \\ &= \max\{\Gamma, \Delta\} x^\top (\Omega + \Pi)^\top M (\Omega + \Pi) x \\ &= \max\{\Gamma, \Delta\} x^\top M x. \end{aligned}$$

Thus we have $K = \max\{\Gamma, \Delta\}$ in Theorem 3.1.

For the other inequality, we obtain from

$$M + M^{1/2} P_2 Z^{-1} P_2 M^{1/2} \preceq M + \frac{1}{\gamma} M^{1/2} P_2 (P_2^\top M P_2)^{-1} P_2^\top M^{1/2} \preceq M + \frac{1}{\gamma} I$$

that

$$I + P_2 Z^{-1} P_2 \preceq I + \frac{1}{\gamma} M^{-1} \preceq \left(1 + \frac{1}{\gamma}\right) M^{-1}.$$

Hence we get

$$M^{(1)} = L L^\top + P Z^{-1} P^\top \preceq \left(1 + \frac{1}{\gamma}\right) A^{-1}.$$

2. To derive the inequalities for $M^{(2)}$ we multiply $M^{(2)}$ by $M^{1/2} L^{-1}$ from the left and its transpose from the right. We obtain

$$\begin{aligned} M^{1/2} L^{-1} M^{(2)} L^{-\top} M^{1/2} &= 2M - M^2 + E M^{1/2} V Z^{-1} (M^{1/2} V)^\top E \\ &= I - E \left(I - (M^{1/2} V) Z^{-1} (M^{1/2} V)^\top\right) E. \end{aligned}$$

Setting $\hat{V} = M^{1/2} V$, $T = I - \hat{V}(\hat{V}^\top \hat{V})^{-1} \hat{V}^\top$, and $\tilde{T} = I - \hat{V} Z^{-1} \hat{V}^\top$, it follows that $P^\top A P = \hat{V}^\top \hat{V}$ and

$$\begin{aligned} M^{1/2} L^{-1} M^{(2)} L^{-\top} M^{1/2} &= I - E \tilde{T} E \\ &\preceq I - E \left(\left(1 - \frac{1}{\gamma}\right) I + \frac{1}{\gamma} T\right) E \\ &\preceq I - \left(1 - \frac{1}{\gamma}\right) E^2. \end{aligned}$$

If $\gamma \geq 1$, then the last term is bounded by I; otherwise the bound will be $\frac{1}{\gamma}$, and hence it follows that

$$(M^{(2)})^{-1} \succeq \min\{\gamma, 1\} A.$$

For the other direction we can adapt the proof of Theorem 3.1 in [29]. We have to estimate $E\tilde{T}E$ by a multiple of the identity from above. Note that since $W^\top M^{1/2}\hat{V} = W^\top MV = 0$, inequality (3.6) is equivalent to

$$M^{1/2}TM^{1/2} \preceq \Delta\ (M - EME)$$

or

$$E^2 \preceq I - \frac{1}{\Delta}T.$$

Observe that $E\tilde{T}E \preceq \beta I$ if and only if $\tilde{T}^{1/2}E^2\tilde{T}^{1/2} \preceq \beta I$, and since $\gamma \geqslant 1$, we therefore have that $\tilde{T}^{1/2}$ exists and it follows that

$$\tilde{T} = T + \hat{V}\left((\hat{V}^\top\hat{V})^{-1} - Z^{-1}\right)\hat{V}^\top \preceq T + \left(1 - \frac{1}{\Gamma}\right)\hat{V}(\hat{V}^\top\hat{V})^{-1}\hat{V}^\top.$$

Since $\tilde{T}T = T = T\tilde{T}$ we obtain

$$\begin{aligned}
\tilde{T}^{1/2}E^2\tilde{T}^{1/2} &\preceq \tilde{T} - \frac{1}{\Delta}\tilde{T}^{1/2}T\tilde{T}^{1/2} \\
&= \tilde{T} - \frac{1}{\Delta}T \\
&\preceq \left(1 - \frac{1}{\Delta}\right)T + \left(1 - \frac{1}{\Gamma}\right)\hat{V}(\hat{V}^\top\hat{V})^{-1}\hat{V}^\top \\
&\preceq \max\left\{1 - \frac{1}{\Delta}, 1 - \frac{1}{\Gamma}\right\}\left(T + \hat{V}(\hat{V}^\top\hat{V})^{-1}\hat{V}^\top\right) \\
&= \max\left\{1 - \frac{1}{\Delta}, 1 - \frac{1}{\Gamma}\right\}I.
\end{aligned}$$

From this we finally obtain that

$$\begin{aligned}
(M^{(2)})^{-1} &= L^{-\top}M^{1/2}(I - E\tilde{T}E)^{-1}M^{1/2}L^{-1} \\
&\preceq \max\{\Delta, \Gamma\}L^{-\top}ML^{-1} = \max\{\Delta, \Gamma\}A. \qquad \square
\end{aligned}$$

For the operator $M^{(1)}$ the condition number of $M^{(1)}A$ may also be estimated in terms of the angle between the invariant subspaces associated with the p smallest eigenvalues of M and V. We refer to [26] for this approach. Note that in (3.4), (3.6) we always have $\Delta \geqslant 1$, since $M \preceq I$. Thus if we set $Z = P^\top AP$ in Theorem 3.2, then $\gamma = \Gamma = 1$ and the bounds for $M^{(1)}$ are determined by Δ only. Via (3.4) we see that the inequality for M is needed only on the subspace W which is the M-orthogonal complement of $\operatorname{span} V$. Especially for the choice P in Lemma 2.1 it is easy to verify that $\Delta = 1/(1 - \lambda_{p+1})$. Thus we obtain a condition number $\kappa^{(1)} = 2/(1 - \lambda_{p+1})$ in Theorem 3.2, which is only slightly worse than the optimal condition number obtained via Lemma 2.1, which would give $\kappa^{(1)} = (1-\lambda_n)(2-\lambda_p)/(1-\lambda_{p+1})(2-\lambda_1)$. In a similar way we can compare the bound for $M^{(2)}$ obtained by Theorem 3.2 with the result of Lemma 2.1. In this case we obtain $\Delta = 1/(1 - \lambda_{p+1}^2)$ and thus $\kappa^{(2)} = 1/(1 - \lambda_{p+1}^2)$. Again this is almost the bound of Lemma 2.1, which would give $\kappa^{(2)} = (1 - \lambda_n^2)/(1 - \lambda_{p+1}^2)$. In this respect, the bounds in Theorem 3.2 are (almost) as sharp as the optimal bounds in Lemma 2.1. In contrast to Lemma 2.1, Theorem 3.2 can be applied to any prescribed choice of P that has full rank!

Our next theorem extends Theorem 3.2 to the case $l \geqslant 1$.

THEOREM 3.3. *Let $A \in \mathbb{R}^{n,n}$ be symmetric positive definite and consider the algebraic multilevel operators $M_l^{(1)}, M_l^{(2)}$ in (2.12) and (2.13), respectively. Suppose that the matrices L_k are chosen such that $M_k = L_k^\top A L_k \preceq I$ for all $k = 1, \ldots, l$. Set $E_k = I - M_k$, $P_k = L_k V_k$ and let $W_k \in \mathbb{R}^{n_k, n_k - n_{k-1}}$ be such that* $\operatorname{rank} W_k = n_k - n_{k-1}$ *and $W_k^\top M_k V_k = 0$ for all $k = 1, \ldots, l$.*

1. *If Δ is a constant such that*

$$W_k^\top W_k \preceq \Delta\ W_k^\top M_k W_k \tag{3.8}$$

for all $k = 1, \ldots, l$, then we have

$$\frac{1}{l+1} A \preceq \left(M_l^{(1)}\right)^{-1} \preceq \Delta A. \tag{3.9}$$

2. *If Δ is a constant such that*

$$\begin{bmatrix} 0 & 0 \\ 0 & W_k^\top M_k W_k \end{bmatrix} \preceq \Delta\ \left[V_k, W_k\right]^\top \left(M_k - E_k M_k E_k\right) \left[V_k, W_k\right] \tag{3.10}$$

for all $k = 1, \ldots, l$, then we have

$$A \preceq \left(M_l^{(2)}\right)^{-1} \preceq \Delta\ A. \tag{3.11}$$

Proof. We proceed by induction on l. For $l = 1$ the assertion follows by Theorem 3.2 applied to $Z = P^\top A P$. If we apply Theorem 3.2 to $A_{l-1}, M_{l-1}^{(1)}$, i.e., let Δ be a constant such that

$$\frac{1}{l} A_{l-1} \preceq (M_{l-1}^{(1)})^{-1} \preceq \Delta\ A_{l-1},$$

then, with $Z = (M_{l-1}^{(1)})^{-1}$, we obtain $\gamma = \frac{1}{l}, \Gamma = \Delta$. But $\frac{\gamma}{1+\gamma} = \frac{1}{l+1}$ and hence (3.9) follows.

Inequality (3.11) follows analogously. ∎

By Theorem 3.3 we lose only a factor $\frac{1}{l+1}$ in the condition number by using $l+1$ levels compared with the case $l = 1$ (exact two-level method). If the reduction in size of A_k in every step is sufficient, i.e., if the size of A_{k-1}, for example, is half the size of A_k or less, then we need at most $l \leqslant \log_2(n)$ levels. In this case the factor $1/(l+1) \approx 1/\log_2(n)$ is (almost) negligible.

For the multilevel method we still need a method for the construction of a well-suited matrix P_k in each step. This will be the topic of the next section.

4. The coarsening process. So far we have not discussed the construction of the coarse grid projection matrix P for given L, A. As before we set $L^\top A L = M$, $E = I - M$ and assume that $E \succeq 0$.

4.1. Construction of P via the QR decomposition. We have already seen in Lemma 2.1 that in terms of conditioning, an invariant subspace V of E associated with the large eigenvalues of E yields the optimal choice for $P = LV$. But in practice we do not have this invariant subspace available, nor is this a favorable choice, because in this case P would typically be full and a further coarsening of $P^\top A P$ will be almost impossible, since this matrix is no longer sparse. So we need a different choice for $P = LV$.

By Lemma 1.1 we may use a suitably chosen set of columns of E as V to approximate the space spanned by the eigenvectors associated with the large eigenvalues. But Lemma 1.1 does not give bounds on the preconditioning property of the resulting preconditioner.

On the other hand the approximation results from section 3 and especially (3.6) show that choosing a suitable space V will give the desired approximation properties. To find this suitable space V, we need to establish the connection between the approximation results and Lemmas 1.1 and 2.1. According to the proof of Lemma 1.1 we need a QR-like decomposition $M^{1/2}E = QR$ (or more precisely of $M^{1/2}\hat{E}_p = QR$) if we want to approximate the eigenvectors associated with the large eigenvalues. Equivalently we can compute $E = QR$, where $Q^\top MQ = I$. So if V, satisfying (3.6), arises from a QR decomposition of E with $Q^\top MQ = I$, then Lemma 1.1 is applicable. In other words this choice of V should ensure that E is well approximated by a rank-p matrix up to a small error. Lemma 4.1 gives precisely this connection.

LEMMA 4.1. *Let $M \in \mathbb{R}^{n,n}$ be symmetric positive definite and let $E = I - M$. Suppose that we have a decomposition*

$$E \underbrace{[\Pi_1, \Pi_2]}_{\Pi} = \underbrace{[V, W]}_{Q} \underbrace{\begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}}_{R}, \tag{4.1}$$

where Π is a permutation matrix, $Q = [V, W]$ is nonsingular, and $V^\top MW = 0$. Then there exist matrices R, F such that

$$E = M^{1/2}(E\Pi_1)R + F. \tag{4.2}$$

If there exists a constant Δ that satisfies (3.6), then $\|F\|_2^2 \leqslant 1 - \frac{1}{\Delta}$.

Proof. Since $[V, W]$ is nonsingular and $W^\top MV = 0$, we have

$$I = M^{1/2}V(V^\top MV)^{-1}V^\top M^{1/2} + M^{1/2}W(W^\top MW)^{-1}W^\top M^{1/2}.$$

With $R = R_{11}^{-1}(V^\top MV)^{-1}V^\top M^{1/2}E$ we have

$$\begin{aligned} F &\equiv E - M^{1/2}(E\Pi_1)R \\ &= E - M^{1/2}VR_{11}R \\ &= E - M^{1/2}V(V^\top MV)^{-1}V^\top M^{1/2}E \\ &= M^{1/2}W(W^\top MW)^{-1}W^\top M^{1/2}E, \end{aligned}$$

and it follows that

$$\begin{aligned} \|F\|_2^2 &= \|(W^\top MW)^{-1/2}W^\top M^{1/2}E\|_2^2 \\ &= \sup_{x \neq 0} \frac{x^\top W^\top EMEWx}{x^\top W^\top MWx} \\ &= 1 - \sup_{x \neq 0} \frac{x^\top W^\top (M - EME)Wx}{x^\top W^\top MWx} \\ &\leqslant 1 - \frac{1}{\Delta}. \end{aligned}$$

∎

Lemma 4.1 shows that if V satisfies (3.6) with a small Δ, then by Lemma 1.1 the spaces spanned by the columns of $M^{1/2}V$ and those of $M^{1/2}E\Pi_1$ are good approximations to the invariant subspace of E associated with the p largest eigenvalues.

As a consequence of Lemma 4.1 we may use a QR decomposition with column pivoting of E,

$$E[\Pi_1, \Pi_2] = [V, W]R,\ V^\top MW = 0, \tag{4.3}$$

to obtain a projection matrix $P = LE\Pi_1 = LVR_{11}^{-1}$ such that the remaining error matrix F has small norm. Clearly there is no restriction in replacing V by $E\Pi_1$, since by $V = E\Pi_1 R_{11}^{-1}$ both sets of columns span the same space. But the preconditioners $M^{(1)}, M^{(2)}$ do not change when replacing V by VR_{11}. In contrast to V, $E\Pi_1$ is typically sparse. Moreover, we can determine $P = LE\Pi_1$ as a coarse grid projection matrix from the QR decomposition (4.3) for which the bounds of Lemma 3.2 hold. Here the columns of V, W are not required to be orthogonal in the standard inner product, as one typically requires in a QR decomposition (see, e.g., [13, 31]), but they are orthogonal with respect to the inner product defined by M. We will not discuss in detail how to compute an approximate QR decomposition. One possibility is to adapt a QR-like decomposition as in [31], but other constructions are possible as well. See [6] for a detailed description of this quite technical construction.

4.2. Selection of coarse grid nodes. The next issue that has to be discussed is the pivoting strategy in the QR decomposition. Clearly the best we can do is to locally maximize Δ in the inequalities (3.4), (3.6) to obtain a feasible coarse grid matrix $P = LE\Pi_1$ for the preconditioners $M^{(1)}$ in (2.1) and $M^{(2)}$ in (2.2). Since we only have the freedom to choose the permutation Π_1 in each step, we could choose p columns of E to locally optimize (3.4), (3.6). It is clear that for a fixed number of columns p there exist $\binom{n}{p}$ permutations which have to be checked, and for any of these choices one has to compute a QR decomposition of an $n \times p$ matrix $E\Pi_1$ to get the corresponding Δ. Already for small p the costs are prohibitively expensive, e.g., for $p = 2$, $n(n-1)/2$ possibilities have to be checked. So in practice not more than $p = 1$ can be used in one step. Using the M-orthogonality of V, i.e., that $V^\top MV = I$, we set

$$T = I - VV^\top M. \tag{4.4}$$

Then it is easy to see that the M-orthogonal complement W of V is given by

$$W = TE\Pi_2. \tag{4.5}$$

Using T from (4.4), identity (3.4) can be written as

$$\frac{1}{\Delta} = \min_{y \neq 0} \frac{y^\top W^\top MWy}{y^\top W^\top Wy} \tag{4.6}$$

or, equivalently, as

$$\frac{1}{\Delta} = \min_{Tx \neq 0} \frac{x^\top T^\top MTx}{x^\top T^\top Tx}. \tag{4.7}$$

Likewise we can reformulate (3.6) as

$$\frac{1}{\Delta} = \min_{Tx \neq 0} \frac{x^\top (M - EME)x}{x^\top T^\top MTx}. \tag{4.8}$$

The minimal quotient (4.7) is obtained if Tx is the eigenvector associated with the smallest eigenvalue of M.

After a certain pivot index has been chosen in step p, we can compute the best pivot index from the remaining matrix using (4.7), (4.8) and get the next pivot column.

Expressions (4.7), (4.8) require the solution of an eigenvalue problem in every step. Since even for small matrices it is almost impossible to solve all the eigenvalue problems completely for any possible choice in step $p+1$, the eigenvector of M associated with the smallest eigenvalue can serve as a test vector. Initially the minimum is achieved for the eigenvector associated with the smallest eigenvalue λ. Suppose that x with $x^\top x = 1$ is a normalized eigenvector of M associated to the smallest eigenvalue, say λ. Then we have

$$\begin{aligned}\hat{\lambda} &:= \frac{x^\top T^\top M T x}{x^\top T^\top T x} \\ &= \frac{x^\top (M - MVV^\top M)x}{x^\top (I - 2VV^\top M + MVV^\top VV^\top M)x} \\ &= \lambda \frac{1 - \lambda \|V^\top x\|_2^2}{1 - 2\lambda \|V^\top x\|_2^2 + \lambda^2 (x^\top V) V^\top V (V^\top x)}. \end{aligned} \tag{4.9}$$

If $V^\top V$ is not too big, then, once a projection operator T is applied, the change in $\hat{\lambda}$ is essentially determined by the norm of $V^\top x$. Examining (4.9) we see that if $\|V^\top x\|_2$ is large, then $\hat{\lambda}$ will still be close to λ, while if $\|V^\top x\|_2$ is small, then λ and $\hat{\lambda}$ will be even much closer.

We can do similar calculations for (4.8) and obtain

$$\hat{\lambda} = \frac{x^\top (M - EME)x}{x^\top T^\top M T x} = \frac{1 - (1-\lambda)^2}{1 - \lambda \|V^\top x\|_2^2}.$$

Here the changes are precisely driven by the angle $\|V^\top x\|_2$ independently of $V^\top V$.

This analysis justifies replacing both (4.7), (4.8) by $\|V^\top x\|_2$. In [6] approximations to x were computed using a simple heuristic approach, but clearly there exist many other strategies. Let us postpone the concrete choice of a test vector that approximates the eigenvector x for a moment and instead discuss pivoting strategies based on a given angle $\|V^\top x\|_2$. A first strategy would be that, after p coarse grid nodes have been chosen, we choose the next coarse grid node such that $\|V^\top x\|_2$ is maximized for all possible T of the form

$$T = I - VV^\top M, \qquad V = [V_p, v_{p+1}].$$

Here V_p corresponds to the already chosen first p coarse grid nodes in the QR decomposition (4.1), while v_{p+1} represents column $p+1$ and we want that $[V_p, v_{p+1}]^\top M [V_p, v_{p+1}] = I$.

A second and better approach is the following block strategy. Since V spans the same space as suitably chosen columns of E, we have that two columns i, j of V or E are M-orthogonal if their distance is larger than 3 in the graph of M. This can be seen from the fact that E, M have the same graph and $E^\top M E$ may have nonzero elements only for pairs (i, j) that have a distance less than or equal to 3. For this reason, for $k = 1, \ldots, n$ we introduce the sets

$$\mathcal{N}^t(k) = \{l : \ e_k^\top |E|^t e_l \neq 0\}, \tag{4.10}$$

which contain the nodes of distance t from k in the undirected graph associated with E. Since any two possible choices for v_{p+1} commute if their distance in the undirected

graph of M is larger than 3, we can choose as many new nodes in step $p+1$ as there are nodes with distance 4 or more between each other. Hence, after p coarse grid nodes have been chosen, we may choose the next coarse grid node such that $V^\top x$ is maximized for all T of the form

$$T = I - VV^\top M, \qquad V = [V_p, v_{p+1}^{(1)}].$$

Then we can continue this procedure for every node of distance larger than 3 from node $p+1$ and obtain

$$T = I - [V_p, v_{p+1}^{(1)}, v_{p+1}^{(2)}][V_p, v_{p+1}^{(1)}, v_{p+1}^{(2)}]^\top M.$$

We can repeat this strategy until there exists no new nodes outside $\mathcal{N}^3(k)$ for any selected coarse grid node k. Since all these new nodes are independent of each other, eigenvalue problems (4.7), (4.8) need not be updated during this step, and likewise $V^\top x$ is maximized independently.

Numerical experiments with these two strategies have shown that in practice the second strategy is preferable, since it does not run into a local but nonglobal optimum as often as the first strategy.

We will also introduce a locking mode. Suppose that one pass of the block strategy has determined a certain set of coarse grid nodes, while the remaining nodes so far are not considered, since they are within a distance of 3 to one of the members of the set. Let us omit indices for a moment and set $T = I - VV^\top M$. Suppose that in step $p+1$, the index l is chosen as coarse grid node in the second strategy. For all neighboring nodes k we can compute the arithmetic mean of $(v_k^\top x)^2$. Then we lock all those nodes m for which the value $(v_k^\top x)^2$ is smaller than the arithmetic mean, i.e., we do not consider m as coarse grid node anymore. In our experience this strategy is safe when applied a posteriori after a set of coarse grid nodes has been determined such that all remaining nodes are within a distance of 3 to at least one coarse grid node or more. We also apply this strategy during the detection of the coarse grid nodes to all nodes within distance 3 of the recently detected coarse grid node. But in contrast to the strategy that locks nodes a posteriori we need to be much more careful when locking nodes during the construction of coarse grid nodes. In other words we add some constraint before we lock nodes in order to make sure that we do not lock nodes that might become potential coarse grid nodes later on. For this reason we lock only those nodes which are within a distance of 3 to the coarse grid node that is currently determined and require that for any of these nodes there exists a neighbor node belonging to the coarse grid. This is much more restrictive but accelerates the process, since during the construction the number of nodes that need to be updated or that are considered as coarse grid nodes decreases significantly.

The basic form of the coarsening process then appears as follows.

Set $x^\top E = \alpha$, $\nu_i = (Ee_i)^\top MEe_i$, and $p = 0$.
while nodes available
  Choose node $p+1$ subject to maximize α_{p+1}^2/ν_{p+1} among all available nodes.
  Exclude nodes within distance 3 or less.
  Perform one step of the QR decomposition (4.1).
  Replace α by $T\alpha$ and ν_i by $(TEe_i)^\top MTEe_i$.
  $p = p+1$.
Lock nodes.

To perform this procedure, we have to sort the list of angles $(\|v_j^\top x\|_2^2)_j$. This could, for example, be done initially, and then the list can be updated whenever angles

change. Our experiments have shown that after a step of the QR decomposition (full or approximate) was performed, the angles often drastically changed. Although this is a local effect for the case of an approximate QR decomposition, updating a sorted list of angles was very costly. So instead of the first step in the described procedure we take the maximum only among the nodes of $\mathcal{N}^t(i)$, where i is the coarse grid node that has just been chosen in the previous step. Since the nodes of $\mathcal{N}^t(i)$ are locked from the previous step, they cannot serve as coarse grid nodes. But what one could do is to use one node $j \in \mathcal{N}^t(i)$, which maximizes $\|v_j^\top x\|_2$. Instead of taking this node j as coarse grid node, we simulate only one step of the approximate QR decomposition and take a related unlocked node from $\mathcal{N}^t(j)$ as the next coarse grid node that maximizes $(\|v_j^\top x\|_2^2)_j$. The step of maximizing $\|V^\top x\|_2$ is carried out only if $\mathcal{N}^t(j)$ consists of nothing but locked nodes or coarse grid nodes. In general there are typically much more than only one node that might serve as the next coarse grid node. Therefore the set of candidates is stored in a list, and candidates from this list can serve as coarse grid nodes in a later steps (following the first-in first-out principle). This simplifies the detection of coarse grid nodes massively, and steps that require a simulation of an additional QR step become relatively rare.

At the end of the procedure we will end up in a situation in which every node either is locked or belongs to the coarse grid. Then we keep those nodes j locked for which $\|v_j^\top x\|_2$ was below the arithmetic mean taken over $\mathcal{N}^t(j)$. After that, new unlocked nodes appear and the process to detect coarse grid nodes is repeated. We also unlock nodes j if there is either no coarse grid node in $\mathcal{N}^1(j)$ or no unlocked node with a larger angle in $\mathcal{N}^1(j)$.

In every step of the procedure that determines the new coarse grid we need a step of the QR decomposition. To do this exactly would again be too expensive. In the next subsection we therefore discuss an approximate QR decomposition.

4.3. A simple approximate QR decomposition. To derive an approximate QR decomposition we have to discuss which problems occur. One problem is that a full QR decomposition will typically end up in a full matrix Q even if the original matrix is sparse. But there is a simple way to work around this large memory requirement. If a partitioned matrix $A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ is factored as

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ \mathrm{O} & R_{22} \end{bmatrix},$$

then Q_1 can be obtained from A_1 by solving a linear system with R_{11}. As long as only the last column, say column k, of Q_1 is required, then we can compute $Q_1 e_k := A_1 r$ via the solution of the linear system $R_{11} r = e_k$ and $e_k^\top R_{12}$ from $e_k^\top Q_1^\top M A_2$; see [31] for an application of this approach. Clearly $Q_1 e_k$ will still be full and the costs are increasing as k increases, since one has to solve a linear system with a $k \times k$ matrix R_{11}, but solving a linear system with R_{11} corresponds to a reorthogonalization of $A_1 e_k$ against the leading $k-1$ columns of Q_1 in the modified Gram–Schmidt process. So a natural simplification is to restrict the reorthogonalization procedure to a neighborhood of k in the sense of the graph of A. Here the matrix for which a QR decomposition is performed is the residual matrix E and the inner product is given by the preconditioned matrix M. So a natural way to define a neighborhood of k is given by sparsity pattern of $E^\top M E$, i.e., we consider the nodes of $\mathcal{N}^3(k)$ and reduce R_{11} to the diagonal block associated with $\mathcal{N}^3(k)$.

Now suppose that we have generated a test vector (see subsection 4.4). Even if we can detect a reasonable set of coarse grid nodes from the approximate QR

decomposition, we lose our test vector x_0 from the initial grid. Also we need a new test vector when the coarsening process is repeated on the next coarser grid. Of course one can use those components of x or x_0 that are associated with the coarse grid nodes. More sensible is to modify the coarsening process such that recycling of components of the test vector is supported. In principle we should have $Mx \approx 0$, i.e., $Ex \approx x$. Suppose that $\mathcal{C}$ is the set of coarse grid nodes. We should try to modify the selection of coarse grid nodes subject to $Ex \approx \sum_{k\in\mathcal{C}} Ee_k x_k$. In this case we can recycle the test vector and use $(x_k)_{k\in\mathcal{C}}$ as a test vector for the repeated coarsening process applied to the second grid. Since $E \preceq I$ we can add a postprocessing step to the algorithm in which the condition of maximizing the angle $\|V^\top x\|_2$ by taking all nodes j such that $\|v_j^\top x\|_2 \geqslant c \max_l \|v_l^\top x\|_2$ is supplemented with additional nodes k that maximize $\|\sum_{k\in\mathcal{C}} Ee_k x_k + Ee_j x_j\|_2$. In practice we use $c = \frac{3}{4}$. To complete this postprocessing step the final set $\mathcal{C}$ is supplemented with additional nodes j subject to minimize $\|Ex - \rho(\sum_{k\in\mathcal{C}} Ee_k x_k + Ee_j x_j)\|_1$. Here ρ is chosen to minimize $\|Ex - \rho \sum_{k\in\mathcal{C}} Ee_k x_k\|_1$, since we typically do not obtain $\rho = 1$.

4.4. Construction of a test vector. We cannot afford to compute the exact smallest eigenvector, since this would typically be more expensive than solving the linear system. We need to find a test vector that can be easily generated. Throughout the computations we use $x_0 = (1, \ldots, 1)^\top$ for the initial matrix A and start with $x = L^{-1}x_0$ for the preconditioned system M. This test vector is known to satisfy $Ax_0 \approx 0$ in many applications which arise from partial differential equations, but other choices for x_0 may also be used. To use x as a test vector, more work is necessary. Small components of x may be important, but they do not contribute to the measure $\|V^\top x\|_2$. This is even more serious if x is only an approximate eigenvector and if $V^\top MV \neq I$, which is the case for an approximate QR factorization.

To make sure that the information on x is not overlaid by the approximation errors, we split the approximate test vector x as

$$x = x^{(1)} + x^{(2)},$$

where $\|x^{(1)}\| \gg \|x^{(2)}\|$, and then instead of one test vector x, we use the pair of normalized vectors

$$[x^{(1)}/\|x^{(1)}\|,\ x^{(2)}/\|x^{(2)}\|]$$

together as test vectors. This means that for a potential coarse grid node k, the measure $|v_k^\top x|^2$, which reflects the angle, is replaced by

$$\left\| v_k^\top \left[x^{(1)}/\|x^{(1)}\|,\ x^{(2)}/\|x^{(2)}\| \right] \right\|_2^2,$$

which is the angle between v_k and the space spanned by $x^{(1)}$, $x^{(2)}$.

The same strategy is recursively applied to $x^{(2)}$. For the small contribution $x^{(2)}$ it is no longer clear whether $\|Mx^{(2)}\| \ll \|M\| \cdot \|x^{(2)}\|$. For this reason we check for each component of $x^{(2)}$ if its sign should be changed. In principle we could simply take the large components of x as $x^{(1)}$ and the small components as $x^{(2)}$. But one has to examine the situation in more detail. There are simple cases where small components x_j of x most likely do not contribute to Mx. This is the case if $\|Mx\| \approx \|M(x - e_j x_j)\|$. To detect these cases we compare $\|Me_j x_j\|_\infty$ with all $\|Me_k x_k\|_\infty$, $k \in \mathcal{N}^1(j)$. If

$$\|Me_j x_j\|_\infty \leqslant c \max_{k\in\mathcal{N}^1(j)} \|Me_k x_k\|_\infty, \qquad c \ll 1, \tag{4.11}$$

then x_j is considered to be a component of $x^{(2)}$ but not a component of $x^{(1)}$. In practice we used $c = 1/4$. Condition (4.11) can be viewed as small local contribution with respect to j's neighbors $\mathcal{N}^1(j)$.

Another case in which we should take x_j as part of $x^{(2)}$ is when (4.11) is not fulfilled but

$$\|Me_jx_j\|_\infty \leqslant (1+c) \sum_{k\in\mathcal{N}^1(j)} \|Me_kx_k\|_\infty/|\mathcal{N}^1(j)|.$$

(Here $|\mathcal{N}^1(j)|$ denotes the cardinality of $\mathcal{N}^1(j)$.) This means that with respect to the average over the neighbors of j, $\|Me_jx_j\|_\infty$ is relatively large. If in this case

$$\|Me_jx_j\|_\infty \leqslant c \max_{k=1,\dots,n} \|Me_kx_k\|_\infty,$$

then $\|Me_jx_j\|_\infty$ can be viewed as a globally small contribution, but not necessarily as noise, since the neighbors k of j do not have significantly larger $\|Me_kx_k\|_\infty$ in the average.

This strategy is repeated with x replaced by $x^{(2)}$.

In the strategies that we have presented so far, splitting and modifying the test vector x are based on examining contributions of x that may be small but become big once the parts are rescaled. The final modification of x is based on contributions that do no not immediately show up because they (almost) cancel each other. In other words, we might find proper subsets $J \subset \{1,\dots,n\}$ such that $(m_{ij})_{i,j\in J}(x_j)_{j\in J} \approx 0$. To detect these sets we check for any $i = 1,\dots,n$ row i of Mx. We try to detect a subset $J_0 \subset \mathcal{N}^1(i)$ such that

$$\sum_{j\in J_0} m_{ij}x_j \approx 0.$$

J_0 is constructed starting with $J_0 = \{i\}$ and adding additional nodes step by step. Additional nodes j are added if $m_{ij}x_j$ has a sign other than $m_{ii}x_i$. This is done until $|\sum_{j\in J_0} m_{ij}x_j|$ has reached its minimal value or at most a tolerance (we used $0.05\,|m_{ii}|\,\|\,(x_j)_{j\in\mathcal{N}^1(i)}\,\|_\infty$). Additional nodes j with the same sign as $m_{ii}x_i$ are added if $|\sum_{j\in J_0} m_{ij}x_j|$ can be reduced further. After J_0 has been detected, we repeat this strategy for all remaining $i \in J_0$. If new i are found with an analogous property, then J_0 is enlarged to obtain a new set J_1. It is clear that nodes which were excluded when J_0 was constructed will not be added in a later step. This limits the nodes i which might be considered in the next step. Finally this strategy yields one or more sets J.

4.5. Final comments. Note that in order to approximately satisfy $M \preceq I$, we used four steps of the Lanczos method to compute an approximation to the largest eigenvalue of M.

Since the use of approximate inverses introduces entries that are small in absolute value compared with the other entries in the row, we used diagonal compensation for M for any entry $|m_{ij}|$ that was less than $10^{-4}\cdot\max_k |m_{ik}|$. For E we also used diagonal compensation but with $5\cdot 10^{-2}$ instead of 10^{-4}. We used different tolerances because M with diagonal compensation should well approximate the original M, while E is only used for the coarse grid projection.

As iterative solver, cg with initial solution x_0 was used. As stopping criterion we used $\|Ax_k - b\|_2 \leqslant \sqrt{\text{eps}}\|Ax_0 - b\|_2$, where $\text{eps} = 2.2203\cdot 10^{-16}$ denotes the machine precision.

We have described several heuristic ideas to generate the updating procedure for a given preconditioner. We have seen that this updating can be viewed as an AMG process. In the next section we give several numerical examples and compare them with other multigrid techniques.

5. Numerical results. In this section we illustrate the effectiveness of the new procedures and, in particular, our chosen heuristic approximations. Our computations were done in MATLAB 5.3 [23] on a LINUX PC with a Pentium III/400 processor.

In all our examples we start with a given sparse approximate inverse for the initial matrix. There are several choices that we discuss. These are (depending on the example) the classical Jacobi preconditioner, i.e., the diagonal of the matrix, a factored approximate inverse using the graph of the initial matrix (again from [22, 21]), and finally a factored block Jacobi preconditioners. For this latter type of preconditioner a diagonal block is factored using the eigenvalue decomposition of the block.

We updated the preconditioner recursively and at each level we stopped the coarsening process if there were no more nodes available (because of the locking strategy). In the multigrid process we always used diagonal preconditioning on the coarser levels. We terminated the coarsening process when at some level the reduction of the system size was no longer significant, i.e., more than 75% of the previous system. In this case the coarse grid system was solved via the Cholesky factorization.

The algebraic multilevel method based on the approximate QR decomposition will be denoted by AMG–QR. We will denote the geometric multigrid by GMG, and the algebraic multigrid from [29] will be denoted by AMG–RS.

Example 1. Our first example is the matrix LANPRO/NOS2 from the Harwell–Boeing collection. Table 1 shows the results for the QR-based AMG compared with AMG from [29]. The original system has size $n = 957$ and an average of 4.3 nonzero entries per row. The condition number of the initial system is $5.1 \cdot 10^9$. The matrix has large positive off-diagonal entries.

Table 2 gives the results for the number of iteration steps. From Table 2 we can see that the coarse grids generated by the QR-based AMG perform very well, while in contrast to this AMG–RS constructs an unsatisfactory coarse grid hierarchy.

For a tridiagonal preconditioner obtained from a factored sparse approximate inverses in [21, 22], the results for the coarsening process as well as for the iterative process are essentially identical for all three methods.

For the factored sparse approximate inverse from [21, 22] with the same sparsity pattern as the initial matrix, the results for the coarsening process can be found in Table 3.

Here the use of a sparse approximate inverse does not improve the coarsening

TABLE 1
NOS2, *diagonal preconditioner, coarsening.*

			Levels: size and nonzeros (average per row)					
AMG		Flops	2	3	4	5	6	7
RS	Size	$3.9 \cdot 10^5$	477	237	117	19		
	Nonzeros		4.3	4.3	4.3	2.9		
QR	Size	$1.4 \cdot 10^6$	477	238	114	55	22	11
	Nonzeros		4.3	4.3	4.2	4.2	3.2	3.4

TABLE 2
NOS2, *diagonal preconditioning, iteration.*

Type of precond.	No prec.	Diag.	AMG–RS $M_l^{(1)}$	AMG–RS $M_l^{(2)}$	AMG–QR $M_l^{(1)}$	AMG–QR $M_l^{(2)}$
cg steps	59765	6920	4632	2103	92	39
Flops	$1.1{\cdot}10^9$	$1.5{\cdot}10^8$	$1.7{\cdot}10^8$	$1.7{\cdot}10^8$	$4.0{\cdot}10^6$	$3.5{\cdot}10^6$

TABLE 3
NOS2, *pattern of A for preconditioning, coarsening.*

			Levels: size and nonzeros (average per row)			
AMG		Flops	2	3	4	5
RS	Size	$5.7 \cdot 10^5$	426	212	79	13
	Nonzeros		5.5	4.5	2.9	2.8
QR	Size	$2.0 \cdot 10^6$	449	152	19	
	Nonzeros		8.3	5.3	2.8	

TABLE 4
NOS2, *pattern of A for preconditioning, iteration.*

Type of precond.	No prec.	Pattern of A	AMG–RS $M_l^{(1)}$	AMG–RS $M_l^{(2)}$	AMG–QR $M_l^{(1)}$	AMG–QR $M_l^{(2)}$
cg steps	59765	3360	4518	2087	1340	714
Flops	$1.1{\cdot}10^9$	$9.4{\cdot}10^7$	$1.9{\cdot}10^8$	$1.9{\cdot}10^8$	$7.3{\cdot}10^7$	$7.7{\cdot}10^7$

process. The results are significantly worse than for the case where diagonal preconditioning is used. However, the QR-based AMG still performs much better than AMG–RS. This is no surprise, since this example has large positive off-diagonal entries, which is known to cause problems for AMG–RS. The numerical results for the iterative solution are given in Table 4.

Finally we will consider a block diagonal preconditioner. The matrix NOS2 is block tridiagonal with blocks of size 3×3. So natural block diagonal preconditioners should have block size $3, 6, 9, \ldots$. We will use a block Jacobi preconditioner of block size 6. Table 5 shows the results for the generation of the coarse grid hierarchy and Table 6 the numerical results.

Again the numerical results are not as good for the diagonal case, but still one can observe a smaller coarse grid hierarchy and a significantly smaller number of iteration steps for the QR-based AMG.

The last two preconditioners, i.e., the block diagonal preconditioner and the factored sparse approximate inverse preconditioner with the same sparsity pattern as A, illustrate that even the QR-based AMG does not always construct a satisfactory grid, but it is still better than that of AMG–RS.

Example 2. Consider the problem

$$\begin{aligned} -\operatorname{div}(a \operatorname{grad} u) &= f \text{ in } [0,1]^2, \\ u &= g \text{ on } \partial[0,1]^2, \end{aligned}$$

TABLE 5
NOS2, *block diagonal* (6×6) *preconditioning, coarsening.*

			Levels: size and nonzeros (average per row)			
AMG		Flops	2	3	4	5
RS	Size	$5.2 \cdot 10^5$	476	158	39	19
	Nonzeros		4.6	3.0	2.9	2.9
QR	Size	$1.7 \cdot 10^6$	323	99	36	
	Nonzeros		7.0	5.6	4.5	

TABLE 6
NOS2, *block diagonal* (6×6) *preconditioning, iteration.*

Type of precond.	No prec.	Block diag.	AMG–RS $M_l^{(1)}$	AMG–RS $M_l^{(2)}$	AMG–QR $M_l^{(1)}$	AMG–QR $M_l^{(2)}$
cg steps	59765	5037	3517	1675	657	299
Flops	$1.1{\cdot}10^9$	$1.5{\cdot}10^8$	$1.5{\cdot}10^8$	$1.6{\cdot}10^8$	$3.3{\cdot}10^7$	$2.9{\cdot}10^7$

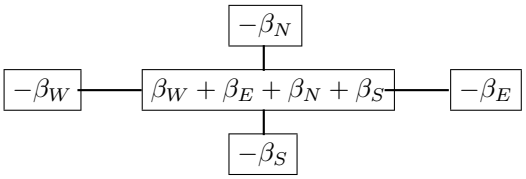


FIG. 1. *Dirichlet, 5-point difference star.*

where $a : [0,1]^2 \to \mathbb{R}$ has different weights in parts of the domain. In detail we consider in each quarter the weights

100	1
1	100

.

The discretization is done using a uniform grid and a 5-point star difference discretization. With local weights $\beta_N, \beta_W, \beta_E, \beta_S$, then the discretization is described by Figure 1.

In every subdomain the value of β is identical to the weights, and for nodes on the interface between the subdomain the arithmetic mean is used.

In this case we will also compare the results with those of geometric multigrid, for which the compact 7-point stencil

$$1/2 \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

is used. Since for this problem the vector $x = (1, \ldots, 1)^\top$ represents the constant function, it makes sense to modify the QR-based AMG slightly. In general we have

TABLE 7
Dirichlet, diagonal preconditioning, coarsening.

		Levels: size and nonzeros (average per row)							
AMG	Level	2	3	4	5	6	7	8	9
RS	Size	32509	8756	2547	822	280	102	37	10
	Nonz.	8.9	9.7	11.2	13.9	15.9	17.2	15.5	8.2
QR	Size	32509	7313	1534	463	84	12		
	Nonz.	8.9	9.7	10.6	15.6	12.5	5.5		
GMG	Size	16129	3969	961	225	49	9	1	
	Nonz.	5.0	4.9	4.9	4.7	4.4	3.7	1.0	

TABLE 8
Dirichlet, diagonal preconditioning, iteration.

Type of precond.	No prec.	Diag.	AMG–RS		AMG–QR		GMG	
			$M_l^{(1)}$	$M_l^{(2)}$	$M_l^{(1)}$	$M_l^{(2)}$	$M_l^{(1)}$	$M_l^{(2)}$
cg steps	5129	862	84	26	61	24	42	16
Flops	$6.7{\cdot}10^9$	$1.3{\cdot}10^9$	$2.4{\cdot}10^8$	$1.8{\cdot}10^8$	$1.7{\cdot}10^8$	$1.6{\cdot}10^8$	$1.3{\cdot}10^8$	$1.0{\cdot}10^8$

TABLE 9
Dirichlet, diagonal preconditioning, scalability.

Size	961	3969	16129	65025
AMG	Flops for the coarse grid generation			
RS	$6.0{\cdot}10^5$	$2.6{\cdot}10^6$	$1.0{\cdot}10^7$	$4.2{\cdot}10^7$
QR	$1.6{\cdot}10^6$	$7.1{\cdot}10^6$	$3.0{\cdot}10^7$	$1.3{\cdot}10^8$
	Flops for the iteration (using prec. $M^{(2)}$)			
RS	$1.9{\cdot}10^6$	$8.8{\cdot}10^6$	$3.9{\cdot}10^7$	$1.8{\cdot}10^8$
QR	$1.1{\cdot}10^6$	$5.9{\cdot}10^6$	$3.1{\cdot}10^7$	$1.6{\cdot}10^8$

adapted the AMG such that the coarse grid projection matrix $E(:, C)$ with the set of coarse grid nodes C roughly satisfies $Ev \approx E(:, C)x(C)$. In this specific problem we may satisfy this constraint exactly by replacing $E(:, C)$ with $DE(:, C)$, where D is a diagonal scaling such that $Ev = DE(:, C)x(C)$.

We use $n = 65025$ and the initial system has on average 5 entries per row. Table 7 shows the results of the coarsening process, i.e., the size of the coarser systems and also the average amount of nonzero elements per row. Table 8 gives the number of iteration steps and flops using multigrid (geometric/algebraic) as preconditioner for cg.

In order to see how the new method scales we compare the flops for the generation of the coarsening process for $n = 961, 3969, 16129, 65025$; see Table 9.

The results so far demonstrate that AMG–QR performs well, even better than

TABLE 10
Dirichlet, sparsity of A for preconditioning, coarsening.

		Levels: size nonzeros (average per row)					
AMG	Level	2	3	4	5	6	7
RS	Size	14386	7249	2241	466	98	10
	Nonzeros	13.4	20.8	22.0	15.1	9.4	2.6
QR	Size	9010	1737	338	72	13	
	Nonzeros	13.6	12.7	11.2	10.7	6.4	

TABLE 11
Dirichlet, sparsity of A for preconditioning, iteration.

Type of precond.	No prec.	Pattern of A	AMG–RS $M_l^{(1)}$	AMG–RS $M_l^{(2)}$	AMG–QR $M_l^{(1)}$	AMG–QR $M_l^{(2)}$
cg steps	5129	436	210	66	64	25
Flops	$6.7{\cdot}10^9$	$9.1{\cdot}10^8$	$6.6{\cdot}10^8$	$4.8{\cdot}10^8$	$2.0{\cdot}10^8$	$1.6{\cdot}10^8$

TABLE 12
Dirichlet, sparsity of A for preconditioning, scalability.

Size	961	3969	16129	65025
AMG	Flops for the coarse grid generation			
RS	$1.2{\cdot}10^6$	$4.6{\cdot}10^6$	$1.8{\cdot}10^7$	$7.2{\cdot}10^7$
QR	$3.0{\cdot}10^6$	$1.4{\cdot}10^7$	$5.9{\cdot}10^7$	$2.4{\cdot}10^8$
	Flops for the iteration (using prec. $M^{(2)}$)			
RS	$1.8{\cdot}10^6$	$9.4{\cdot}10^6$	$6.3{\cdot}10^7$	$4.8{\cdot}10^8$
QR	$1.1{\cdot}10^6$	$6.0{\cdot}10^6$	$2.9{\cdot}10^7$	$1.6{\cdot}10^8$

classical AMG–RS. It is better with respect to the coarsening process as well as with respect to the iterative process. One problem that can be seen from Table 9, however, is that the QR-based AMG is more expensive (by a factor of 3) than AMG–RS. This is no surprise, since its construction involves an approximate QR factorization. Despite this construction it also scales linearly.

For a sparse approximate inverse as in [22, 21] with the same sparsity pattern as the initial matrix, the preconditioned system is still an M-matrix, which has been observed to be helpful for the application of the classical AMG. Table 10 shows that both methods use a much coarser grid than in the case of diagonal preconditioning. But still AMG–QR needs fewer and smaller levels. The iterative process is also faster for the QR-based AMG, as shown in Table 11.

Scalability is shown in Table 12. It is interesting that for this approximate inverse the QR-based AMG performs better (fewer flops) than in the diagonal case, while the classical AMG becomes slower. The overhead in the construction is now more than compensated for by the accelerated iterative part. Again, both methods scale linearly

TABLE 13
Dirichlet, block diagonal (4×4) preconditioning, coarsening.

		Levels: size and nonzeros (average per row)							
AMG	Level	2	3	4	5	6	7	8	9
RS	Size	32592	16251	5773	2185	763	274	106	33
	Nonz.	13.8	17.7	20.7	27.7	25.8	24.7	24.7	16.8
QR	Size	8142	1824	382	75	17			
	Nonz.	9.1	9.7	9.4	8.2	5.6			

TABLE 14
Dirichlet, block diagonal (4×4) preconditioning, iteration.

Type of precond.	No prec.	Pattern of A	AMG–RS $M_l^{(1)}$	AMG–RS $M_l^{(2)}$	AMG–QR $M_l^{(1)}$	AMG–QR $M_l^{(2)}$
cg steps	5129	640	76	23	83	32
Flops	$6.7{\cdot}10^9$	$1.5{\cdot}10^9$	$3.1{\cdot}10^8$	$2.5{\cdot}10^8$	$2.6{\cdot}10^8$	$2.0{\cdot}10^8$

TABLE 15
Dirichlet, block diagonal (4×4) preconditioning, scalability.

Size	961	3969	16129	65025
AMG	Flops for the coarse grid generation			
RS	$1.1{\cdot}10^6$	$5.0{\cdot}10^6$	$2.1{\cdot}10^7$	$8.4{\cdot}10^7$
QR	$1.6{\cdot}10^6$	$6.7{\cdot}10^6$	$2.8{\cdot}10^7$	$1.1{\cdot}10^8$
	Flops for the iteration (using prec. $M^{(2)}$)			
RS	$2.7{\cdot}10^6$	$1.3{\cdot}10^7$	$5.6{\cdot}10^7$	$2.5{\cdot}10^8$
QR	$1.7{\cdot}10^6$	$8.0{\cdot}10^6$	$4.2{\cdot}10^7$	$2.0{\cdot}10^8$

with respect to the coarsening process, but AMG–QR is much faster and scales much better in the iterative part.

Finally, we use a block diagonal preconditioner with small blocks. For a block diagonal matrix where each diagonal block has size 4×4, we see in Table 13 that the coarse grid generation for the QR-based AMG is much superior to AMG–RS. Here it is important to note that due to the use of block diagonal approximate inverses, the preconditioned system has many positive off-diagonal entries, which causes problem for the classical AMG. But the QR-based AMGs can exploit the benefits of the sparse approximate inverses to construct only a few small coarser grids. The number of iteration steps between both AMG methods here is not very different, as shown in Table 14. For scalability see Table 15. The construction of much smaller grids for AMG–QR is reflected by a much faster coarse grid generation and a significant acceleration when applying the preconditioner in the iteration process. For the coarse grid generation this can be seen from the surprisingly small difference between the number of flops needed by both AMGs. For the iterative part one can observe that

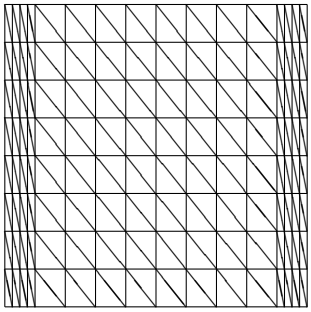
AMG–QR needs fewer flops, although it requires more iteration steps.

The numerical results for this problem show that the QR-based AMG better adapts to the given initial sparse approximate inverse. This should be the case because they have been constructed to do so. The drawback is that this approach consumes more time for its construction because it uses an approximate QR factorization. However, AMG–QR scales as good as AMG–RS.

Example 3. Finally, consider the problem

$$\begin{aligned} -\varepsilon^2 u_{xx} - u_{yy} &= f \text{ in } [0,1]^2, \\ u &= g \text{ on } \partial[0,1]^2, \end{aligned}$$

where ε strongly varies from 10^0 to 10^{-4}. For this problem we use the variational formulation and piecewise quadratic finite elements; cf., e.g., [7]. The discretization is done using a uniform triangulation with two additional boundary layers of size $\frac{\varepsilon}{4} \times 1$ near the left and also near the right boundary, as shown below:



Within these boundary layers the triangles are condensed by an additional factor of $\varepsilon/4$ in the x-direction. We examine the aspect of scalability (with respect to the system size) and robustness (with respect to ε).

Table 16 shows the number of cg iteration steps for both AMGs for the case of a diagonal approximation using $M^{(2)}$ as preconditioner. The same comparison is made in Table 17 for the case of the sparse approximate inverse with the same pattern as A.

Next we examine the computational amount of work in flops.

As the number of iteration steps in Tables 16 and 17 have indicated, the scalability of AMG–RS performs poorer with increasing system size than AMG–QR, which

TABLE 16
Anisotropic Dirichlet, diagonal precond., cg steps using $M^{(2)}$.

		ε versus scalability			
AMG	ε	961	3969	16129	65025
RS	10^0	31	47	89	170
	10^{-2}	42	84	174	268
	10^{-4}	38	65	135	264
QR	10^0	23	33	56	109
	10^{-2}	23	33	60	98
	10^{-4}	23	31	49	85

TABLE 17
Anisotropic Dirichlet, pattern of A for precond., cg steps using $M^{(2)}$.

		ε versus scalability			
AMG	ε	961	3969	16129	65025
RS	10^0	24	56	103	232
	10^{-2}	47	90	196	318
	10^{-4}	48	91	177	336
QR	10^0	19	24	47	61
	10^{-2}	22	38	49	84
	10^{-4}	16	31	43	60

TABLE 18
Anisotropic Dirichlet, diagonal precond., flops (coarsening + cg).

		ε versus scalability		
AMG	ε	3969	16129	65025
RS	10^0	$3.5\cdot10^6+2.5\cdot10^7$	$1.4\cdot10^7+2.0\cdot10^8$	$5.9\cdot10^7+1.5\cdot10^9$
	10^{-2}	$2.9\cdot10^6+4.1\cdot10^7$	$1.2\cdot10^7+3.5\cdot10^8$	$5.0\cdot10^7+2.2\cdot10^9$
	10^{-4}	$2.9\cdot10^6+3.2\cdot10^7$	$1.2\cdot10^7+2.7\cdot10^8$	$5.0\cdot10^7+2.2\cdot10^9$
QR	10^0	$1.4\cdot10^7+1.5\cdot10^7$	$7.1\cdot10^7+1.1\cdot10^8$	$4.3\cdot10^8+8.4\cdot10^8$
	10^{-2}	$9.8\cdot10^6+1.4\cdot10^7$	$5.1\cdot10^7+1.0\cdot10^8$	$3.3\cdot10^8+7.0\cdot10^8$
	10^{-4}	$9.4\cdot10^6+1.3\cdot10^7$	$4.9\cdot10^7+8.5\cdot10^7$	$3.3\cdot10^8+6.2\cdot10^8$

TABLE 19
Anisotropic Dirichlet, pattern of A for precond., flops (coarsening + cg).

		ε versus scalability		
AMG	ε	3969	16129	65025
RS	10^0	$6.1\cdot10^6+3.0\cdot10^7$	$2.5\cdot10^7+2.2\cdot10^8$	$1.0\cdot10^8+2.0\cdot10^9$
	10^{-2}	$5.4\cdot10^6+4.3\cdot10^7$	$2.2\cdot10^7+3.8\cdot10^8$	$9.0\cdot10^7+2.5\cdot10^9$
	10^{-4}	$5.5\cdot10^6+4.4\cdot10^7$	$2.2\cdot10^7+3.5\cdot10^8$	$9.0\cdot10^7+2.6\cdot10^9$
QR	10^0	$2.8\cdot10^7+9.0\cdot10^6$	$1.2\cdot10^8+7.2\cdot10^7$	$5.2\cdot10^8+3.8\cdot10^8$
	10^{-2}	$2.5\cdot10^7+2.0\cdot10^7$	$1.3\cdot10^8+1.1\cdot10^8$	$6.7\cdot10^8+7.3\cdot10^8$
	10^{-4}	$2.2\cdot10^7+1.5\cdot10^7$	$1.1\cdot10^8+8.7\cdot10^7$	$6.5\cdot10^8+5.0\cdot10^8$

roughly needs only half as many flops (see Tables 18 and 19). One additional observation can be made. AMG–QR is designed as a supplement for a given sparse approximate inverse. This does not mean that it will always be able to compensate a poor smoothing property of the initial sparse approximate inverse. This can be seen when looking at the scalability of the coarse grid generation. Although AMG–QR needs more flops for the coarse grid generation when a sparse approximate inverse with same pattern as A is used than with the diagonal approximate inverse, it scales better than in the diagonal case. Apparently the sparse approximate inverse with same pattern as A compensates the anisotropy much better than the diagonal approximate inverse, and this property is detected by AMG–QR. Although this is not

part of this kind of AMG, we expect an improvement if the initial sparse approximate inverse is more adapted to the anisotropic behavior than those simple two sparse approximate inverses that were chosen in these examples.

6. Conclusions. We have derived new approaches for the construction of algebraic multilevel methods that automatically detect the coarse grid by suitably chosen columns of the residual matrix. We have presented the mathematical theory to develop optimal preconditioners. The key feature of the new approach is the choice of an effective pivoting strategy to detect the correct set of columns. The numerical examples indicate that obtaining a good choice is a challenging problem. Simple techniques, such as locking of some nodes or taking several nodes in one step, seem to be useful. Clearly none of these strategies is successful if the sparse approximate preconditioner does not have a smoothing property, i.e., if most of the eigenvalues of the preconditioned system are clustered at the large end of the spectrum. A more detailed analysis of methods for constructing good pivoting strategies needs further research.

REFERENCES

[1] O. Axelsson, M. Neytcheva, and B. Polman, *An application of the bordering method to solve nearly singular systems,* Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet., 1 (1996), pp. 3–25.

[2] O. Axelsson and P. Vassilevski, *Algebraic multilevel preconditioning methods* I, Numer. Math., 56 (1989), pp. 157–177.

[3] O. Axelsson and P. S. Vassilevski, *Algebraic multilevel preconditioning methods* II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.

[4] R. E. Bank and C. Wagner, *Multilevel ILU decomposition,* Numer. Math., 82 (1999), pp. 543–576.

[5] M. Benzi, C. D. Meyer, and M. Tůma, *A sparse approximate inverse preconditioner for the conjugate gradient method,*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.

[6] M. Bollhöfer and V. Mehrmann, *A New Approach to Algebraic Multilevel Methods Based on Sparse Approximate Inverses,* Preprint SFB393/99–22, Department of Mathematics, TU Chemnitz, Germany, 1999.

[7] D. Braess, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics,* Cambridge University Press, Cambridge, UK, 2001.

[8] A. Brandt, *Algebraic multigrid theory: The symmetric case,* Appl. Math. Comput., 19 (1986), pp. 23–65.

[9] T. Chan, W.-P. Tang, and W. L. Wan, *Fast wavelet based sparse approximate inverse preconditioner,* BIT, 37 (1997), pp. 644–660.

[10] E. Chow and Y. Saad, *Approximate inverse preconditioners via sparse-sparse iterations,* SIAM J. Sci. Comput., 19 (1998), pp. 995–1023.

[11] W. Dahmen and L. Elsner, *Hierarchical iteration,* in Robust Multi-Grid Methods, Notes Numer. Fluid Mech. 23, W. Hackbusch, ed., Vieweg, Braunschweig, 1988.

[12] R. Freund, G. Golub, and N. Nachtigal, *Iterative solution of linear systems,* Acta Numer., 1992, pp. 1–44.

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

[14] A. Greenbaum, *Iterative Methods for Solving Linear Systems,* Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.

[15] M. J. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses,* SIAM J. Sci. Comput., 18 (1997), pp. 838–853.

[16] W. Hackbusch, *Multigrid Methods and Applications,* Springer-Verlag, New York, 1985.

[17] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations,* Springer-Verlag, New York, 1994.

[18] V. E. Henson and P. S. Vassilevski, *Element-Free AMG-e: General Algorithms for Computing Interpolation Weights,* Technical report UCRL–VG–138290, Lawrence Livermore National Laboratory, Livermore, CA, 2000.

[19] T. Huckle and J. Staudacher, *Matrix multilevel methods and preconditioning,* BIT, to appear.
[20] J. E. Jones and P. S. Vassilevski, *AMGe based on element agglomeration,* SIAM J. Sci. Comput, 23 (2001), pp. 109–133.
[21] I. E. Kaporin, *New convergence results and preconditioning strategies for the conjugate gradient method,* Numer. Linear Algebra Appl., 1 (1994), pp. 179–210.
[22] L. Yu. Kolotilina and A. Yu. Yeremin, *Factorized sparse approximate inverse preconditionings.* I. *Theory,* SIAM J. Matrix Anal. Appl., 14 (1993), pp. 45–58.
[23] The MathWorks Inc., *MATLAB—The Language of Technical Computing,* 1996.
[24] Y. Notay, *Optimal V-cycle algebraic multilevel preconditioner,* Numer. Linear Algebra Appl., 5 (1998), pp. 441–459.
[25] Y. Notay, *Using approximate inverses in algebraic multigrid methods,* Numer. Math., 80 (1998), pp. 397–417.
[26] A. Padiy, O. Axelsson, and B. Polman, *Generalized augmented matrix preconditioning approach and its application to iterative solution of ill-conditioned algebraic systems,* SIAM J. Matrix Anal. Appl., 22 (2000), pp. 793–818.
[27] A. Reusken, *Approximate cyclic reduction preconditioning,* in Multigrid Methods 5, Proceedings of the Fifth European Multigrid Conference, W. Hackbusch and G. Wittum, eds., Springer-Verlag, New York, 1998, pp. 243–259.
[28] A. Reusken, *On the approximate cyclic reduction preconditioner,* SIAM J. Sci. Comput., 21 (2000), pp. 565–590.
[29] J. Ruge and K. Stüben, *Algebraic multigrid,* in Multigrid Methods, S. McCormick, ed., SIAM, Philadelphia, 1987, pp. 73–130.
[30] Y. Saad, *Iterative Methods for Sparse Linear Systems,* PWS Publishing, Boston, 1996.
[31] G. Stewart, *Four algorithms for the efficient computation of truncated pivoted QR approximations to a sparse matrix,* Numer. Math., 83 (1999), pp. 313–323.
[32] W.-P. Tang, *Toward an effective sparse approximate inverse preconditioner,* SIAM J. Matrix Anal. Appl., 20 (1999), pp. 970–986.
[33] W.-P. Tang and W. L. Wan, *Sparse approximate inverse smoother for multigrid,* SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1236–1252.
[34] A. van der Ploeg, E. Botta, and F. Wubs, *Nested grids ILU–decomposition (NGILU),* J. Comput. Appl. Math., 66 (1996), pp. 515–526.
[35] R. S. Varga, *Matrix Iterative Analysis,* Prentice-Hall, Englewood Cliffs, NJ, 1962.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 219–237

ON THE RELATIONS BETWEEN ILUs AND FACTORED APPROXIMATE INVERSES*

MATTHIAS BOLLHÖFER† AND YOUSEF SAAD‡

Abstract. This paper discusses some relationships between ILU factorization techniques and factored sparse approximate inverse techniques. While ILU factorizations compute approximate LU factors of the coefficient matrix A, approximate inverse techniques aim at building triangular matrices Z and W such that $W^\top AZ$ is approximately diagonal. The paper shows that certain forms of approximate inverse techniques amount to approximately inverting the triangular factors obtained from some variants of ILU factorization of the original matrix. A few useful applications of these relationships will be discussed.

Key words. sparse matrices, ILU, variants of ILU, sparse approximate inverse

AMS subject classifications. 65F05, 65F10, 65F50

PII. S0895479800372110

1. Introduction.

Preconditioned Krylov-subspace iterations are among the most efficient techniques for solving linear systems of the form

$$Ax = b, \tag{1}$$

where $A \in \mathbb{R}^{n,n}$ is nonsingular and $b \in \mathbb{R}^n$ is a given right-hand side; see, e.g., [22, 12, 1, 14]. Among the most popular preconditioners are those based on approximate factorizations obtained from direct solution methods, such as the LU factorization [11, pp. 92ff]. Alternative techniques appeared in recent years which compute approximate solutions of (1) via an approximate inverse of A, instead of a factorization. One of the main motivations for using preconditioners of this type is parallelism. Another important reason is that ILU preconditioners, which have been developed for M-matrices [19], often fail for indefinite matrices.

A few of the approximate inverse techniques are based on minimizing $\|I - AM\|$ in some appropriate norm [17, 15, 13, 9]. Others compute the approximate inverse in factored form by seeking two sparse unit upper triangular matrices W and Z and a diagonal D, such that $W^\top AZ \approx D$; see, e.g., [3, 4, 2, 16, 22]. As it turns out, the latter class of preconditioners show an algebraic behavior that is similar to that of the well-known ILU decompositions. For example, they are stable for M- and H-matrices, in perfect analogy with known results on ILU decompositions in [19, 18].

It is worth mentioning that there has been some work on methods for inverting triangular matrices which are computed from a standard LU factorization, based on the same motivations; see [24]. However, our paper does not consider these methods.

*Received by the editors May 10, 2000; accepted for publication (in revised form) by E. Ng October 3, 2001; published electronically July 1, 2002.
http://www.siam.org/journals/simax/24-1/37211.html

†Institute of Mathematics, MA 4–5, Berlin University of Technology, D–10623 Berlin, Germany (bolle@math.tu-berlin.de, http://www.math.tu-berlin.de/∼bolle/). The research of this author was supported by the University of Minnesota and by grants of the DFG BO 1680/1-1. This research was performed while the author was visiting the University of Minnesota at Minneapolis.

‡Department of Computer Science and Engineering, University of Minnesota, 4–192 EE/CSci Building, 200 Union St., SE, Minneapolis, MN 55455–0154 (saad@cs.umn.edu, http://www.cs.umn.edu/∼saad/). The research of this author was supported by the U.S. Army Research Office under contract DAAD19-00-1-0485 and by the Minnesota Supercomputing Institute.

We also point out that all the results in this paper are valid in the presence of exact arithmetic.

The purpose of this paper is to take an in-depth look at the relationships between factored approximate inverse preconditioners and ILU decomposition methods. In particular, it will be shown that AINV methods generate factors which can be viewed as approximations of the inverses of the triangular factors obtained by certain variants of ILU. Using a slight modification of the strategies to drop entries we will also show that matrices resulting from these methods can be viewed as the exact inverses of triangular factors obtained via an ILU decomposition. Specifically, what is required is to suitably modify or construct modified approximate Schur complements such that the inverse factors are those (or at least close to those) obtained by factored approximate inverse techniques.

2. ILU factorizations. ILU factorizations construct approximate L, D, U factors of A such that

$$A \approx LDU,$$

where $L, U^\top$ are lower triangular matrices with unit diagonal. A partial LU factorization, when it exists, can be recursively expressed by considering the first step:

$$\begin{bmatrix} a_{11} & f \\ e & C \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ g & I \end{bmatrix} \begin{bmatrix} \delta & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} 1 & h \\ 0 & I \end{bmatrix}, \tag{2}$$

with $\delta = a_{11}$. The terms δ, g, h, and S satisfy $g\delta = e \in \mathbb{R}^{n-1,1}$, $\delta h = f \in \mathbb{R}^{1,n-1}$, and

$$S = C - g\ \delta\ h \in \mathbb{R}^{n-1,n-1}\ . \tag{3}$$

The matrix S denotes the so-called Schur complement. An exact LU decomposition is obtained by applying (2) recursively on the resulting Schur complement. The process is completed by substituting the factorization $S = L_S D_S U_S$, when it exists, into (2) to obtain

$$\begin{bmatrix} a_{11} & f \\ e & C \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ g & L_S \end{bmatrix} \begin{bmatrix} \delta & 0 \\ 0 & D_S \end{bmatrix} \begin{bmatrix} 1 & h \\ 0 & U_S \end{bmatrix}, \tag{4}$$

which is the final LU factorization.

In incomplete factorizations, entries are dropped during this procedure in the L, U factors and in the Schur complement. A common strategy is to drop entries in the first column of L according to a certain "dropping rule" and apply a similar dropping rule to the first row of U. As a result of this procedure, the row $h = \delta^{-1} f$ and column $g = e\delta^{-1}$ are replaced by sparsified approximations

$$\tilde{h} \approx h, \qquad \tilde{g} \approx g,$$

leading to the approximate Schur complement

$$\tilde{S} = C - \tilde{g}\ \delta\ \tilde{h}, \tag{5}$$

which is a sparsified version of (3). However, there are several other ways of defining approximate Schur complements from approximations to g and h. For example, we can multiply both sides of (2) to the left by the inverse of the approximate L factor obtained by replacing g by $\tilde{g}$. Equating the resulting $(2,2)$ blocks leads to

$$\tilde{S} = C - \tilde{g}\ f. \tag{6}$$

From an algorithmic point of view, the process amounts to multiplying the current matrix, i.e., the matrix on the left-hand side of (2), to the left by

$$\begin{bmatrix} 1 & 0 \\ -\tilde{g} & I \end{bmatrix}.$$

In other words, the next Schur complement is obtained by performing the usual row operations in Gaussian elimination using a sparsified version of L, obtained by dropping some elements.

Similarly, a column-based version of this process consists of multiplying both sides of (2) to the right by the inverse of the approximate U factor obtained by replacing h by $\tilde{h}$. This leads to the approximation

$$\tilde{S} = C - e\ \tilde{h}. \tag{7}$$

A fourth option we mention consists of a combination of these two operations. First, operate with the approximation to the inverse of L to the left of the matrix A, and then operate with the approximation to the inverse of U to the right of the resulting matrix. The $(2,2)$ block of the resulting matrix is the Schur complement

$$\tilde{S} = C - \tilde{g}\ f - (e - \tilde{g}\ \delta)\,\tilde{h}. \tag{8}$$

Other ways of defining an approximate Schur complement can be derived from other equivalent expressions of the Schur complement. In the case of an exact factorization (no dropping) the update formulas (5), (6), (7), and (8) will all lead to the same S. In practice, (5) is the most common scheme for defining ILU factorizations; see, e.g., [19] or [21]. Typically, (5) produces the smallest amount of fill-in compared with the other formulas. The update (8) has also been used in a number of papers [23, 2, 6, 8]. In the symmetric positive definite case, it is guaranteed to produce a stable ILU factorization; see [23].

2.1. Update variants. In order to simplify the description of the algorithms to be considered we make the following observation which allows us to express all four types of updates just described in a concise manner. Consider, for example, the update (5). The update for entry (i,j) of C is performed only when $\tilde{g}_i$ and $\tilde{h}_j$ are both nonzero, i.e., when their original terms in g and h have both not been dropped. We now notice that if we call S the current Schur complement matrix, i.e., the matrix on the left-hand side of (2), then (5) is equivalent to performing the following update for each pair (i,j) such that $s_{ik} \cdot s_{kj} \neq 0$:

$$s_{ij} = s_{ij} - \frac{s_{ik} s_{kj}}{d_{kk}}. \tag{9}$$

This update is restricted to the cases when g_i and h_j have both not been dropped. Thus, (5) can be expressed as "Perform (9) when $\tilde{g}_i \neq 0$ and $\tilde{h}_j \neq 0$." Interestingly, each of (5), (6), (7), and (8) can be expressed in this manner.

- Update (5): Perform (9) when $\tilde{g}_i \neq 0$ and $\tilde{h}_j \neq 0$.
- Update (6): Perform (9) when $\tilde{g}_i \neq 0$.
- Update (7): Perform (9) when $\tilde{h}_j \neq 0$.
- Update (8): Perform (9) when $\tilde{g}_i \neq 0$ or $\tilde{h}_j \neq 0$.

A little explanation is required for the last case. If $\tilde{g}_i \neq 0$ and $\tilde{h}_j = 0$, then the formula will coincide with (6), which is the same as (9) for this particular situation.

The opposite case, when $\tilde{g}_i = 0$ and $\tilde{h}_j \neq 0$, is similar and leads to formula (7). When both $\tilde{g}_i$ and $\tilde{h}_j$ are nonzero, then the entry in position (i, j) of the matrix $C - \tilde{g}f - e\tilde{h}$ can be viewed as the term s_{ij} which has undergone two updates of which only one is required. Therefore, we need to correct this update by adding $\tilde{g}\delta\tilde{h}$, leading to (8).

Throughout the paper we will use the above formalism, i.e., all updates (5)–(8) will be expressed in the form "*if* version$(\tilde{g}_i, \tilde{h}_j)$*, then perform update* (9)," in which version$(\tilde{g}_i, \tilde{h}_j)$ is a boolean function which takes the following values for the four different cases under consideration:

- Update (5): version$(\tilde{g}_i, \tilde{h}_j) = \{\ \tilde{g}_i \neq 0$ and $\tilde{h}_j \neq 0\ \}$.
- Update (6): version$(\tilde{g}_i, \tilde{h}_j) = \{\ \tilde{g}_i \neq 0\ \}$.
- Update (7): version$(\tilde{g}_i, \tilde{h}_j) = \{\ \tilde{h}_j \neq 0\ \}$.
- Update (8): version$(\tilde{g}_i, \tilde{h}_j) = \{\ \tilde{g}_i \neq 0$ or $\tilde{h}_j \neq 0\ \}$.

2.2. Block versions. It is useful to generalize the above arguments to the case when the (1,1) term a_{11} in (2) is replaced by a block B of size $k \times k$ of the matrix A, with $k > 1$. The partial LU factorization, when it exists, is now expressed by

$$\begin{bmatrix} B & F \\ E & C \end{bmatrix} = \begin{bmatrix} L_B & 0 \\ G & I \end{bmatrix} \begin{bmatrix} D_B & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} U_B & H \\ 0 & I \end{bmatrix}, \tag{10}$$

where $L_B, U_B^\top \in \mathbb{R}^{k,k}$ are lower triangular matrices with unit diagonal and $D_B \in \mathbb{R}^{k,k}$ is diagonal. Here, L_B, D_B, U_B refer to an already computed LU decomposition of B. The matrices D_B, G, H, and S satisfy $GD_BU_B = E \in \mathbb{R}^{n-k,k}$, $L_BD_BH = F \in \mathbb{R}^{k,n-k}$, and the Schur complement now becomes

$$S = C - GD_BH \in \mathbb{R}^{n-k,n-k}. \tag{11}$$

The same four versions of the approximate Schur complement as those defined by (5)–(8) can be defined similarly. We list them all below for future reference:

$$\tilde{S} = C - \tilde{G}D_B\tilde{H}, \tag{12}$$

$$\tilde{S} = C - \tilde{G}L_B^{-1}F, \tag{13}$$

$$\tilde{S} = C - EU_B^{-1}\tilde{H}, \tag{14}$$

$$\tilde{S} = C - \tilde{G}L_B^{-1}F - \left(E - \tilde{G}L_B^{-1}B\right)U_B^{-1}\tilde{H}. \tag{15}$$

At this point we make an important observation regarding the approximate Schur complement. For convenience we call the pth Schur complement the Schur complement obtained by eliminating unknowns $i = 1, \ldots, p$. The zeroth Schur complement is, by definition, the original matrix and the $(i+1)$st Schur complement can be obtained by applying (10), (11) to the ith Schur complement. When dropping is applied, the pth Schur complement, a matrix of size $n - p$, will vary depending on which of the four formulas (12)–(15) is used. Instead of this p-step procedure, we could alternatively obtain an approximate Schur complement directly by using one step of the above process with $k = p$, taking the same equations from (12)–(15). The important property which we point out is that these two methods would lead to the same approximate Schur complement.

PROPERTY 1. *The pth (approximate) Schur complement S obtained from applying p consecutive steps of one of the four formulas* (12)–(15) *with $k = 1$ is identical with the (approximate) pth Schur complement obtained from* 1 *step of the same formula among* (12)–(15)*, with $k = p$.*

2.3. Dropping strategies. There are two broad classes of dropping strategies. In the first category there are strategies which drop elements based only on the pattern of the matrix. This includes the level-of-fill strategy [19]. A second category of methods drops elements dynamically, based on their magnitude [20, 21]. Other strategies combine graph-based methods with threshold dropping.

It is important to point out here that the results we show concern not only the "static" dropping strategies but also some dynamic dropping, e.g., with respect to a prescribed drop tolerance τ, similar to the threshold-based ILUT preconditioning [21]. To be more specific, throughout the paper we assume that any dropping rule we use for sparsifying a vector has information about its numerical values and its associated coordinates. For example, a dropping rule applied to the entries $g_{i,k}$ of G uses only information on $g_{i,k}$ and the related coordinates (i, k).

Possible dropping rules of this type could be

- drop $g_{i,k}$ if $|g_{i,k}| \leqslant \tau$,
- drop $g_{i,k}$ if (i, k) is outside a specific pattern,
- drop $g_{i,k}$ if $|g_{i,k}| \leqslant \tau \|e_i^\top A\|$,

where τ is a fixed drop tolerance and e_i is the ith column of the identity matrix. For more complex dynamic dropping, different versions of Gaussian elimination may produce different ILU factors even if the corresponding exact Gaussian elimination versions would produce the same factors. This is because the dropping strategies may yield different patterns. In general, threshold-based methods are harder to analyze than pattern-based algorithms.

2.4. K, I, J implementations. A sample routine for performing an ILU decomposition is given by Algorithm 2. Algorithm 2 is based on the so-called K, I, J version (or "rank-one" update version) of Gaussian elimination. We make use of our earlier observation on a unified way to handle the approximate Schur complements (12), (15), (13), and (14) in Algorithm 2. The different updates $s_{ij} = s_{ij} - \frac{s_{ik}s_{kj}}{d_{kk}}$ of the approximate Schur complement can be expressed in terms of a logical value version$(\tilde{g}_i, \tilde{h}_j)$ which were defined earlier. The notation changes in the algorithm and the variables $\tilde{g}_i$ and $\tilde{h}_j$ are now called p_i and q_j.

ALGORITHM 2 (ILU).

Input: $A = (a_{ij}) \in \mathbb{R}^{n,n}$. *Output: ILU factorization* $A \approx LDU$.

0. $p = q = 0 \in \mathbb{R}^n$, $L = U = I$, $S = A$.
1. ***for*** $k = 1, \ldots, n$
2. $\quad d_{kk} = s_{kk}$
3. $\quad$ ***for*** $i = k+1, \ldots, n$ *and when* $s_{ik} \neq 0$ *or* $s_{ki} \neq 0$
4. $\quad\quad p_i = s_{ik}/d_{kk}, \quad q_i = s_{ki}/d_{kk}$
5. $\quad\quad$ *Apply a dropping rule to* p_i *and* q_i
6. $\quad\quad l_{ik} = p_i, \quad u_{ki} = q_i$
7. $\quad\quad$ ***for*** $j = k+1, \ldots, n$ *and when* $s_{ik} \neq 0$ *and* $s_{kj} \neq 0$
8. $\quad\quad\quad$ ***if*** version(p_i, q_j) *then:* $\quad s_{ij} = s_{ij} - \frac{s_{ik}s_{kj}}{d_{kk}}$
9. $\quad\quad$ ***end***
10. $\quad$ ***end***
11. ***end***

A significant drawback of Algorithm 2 lies in its practical implementation. Each step of the procedure alters rows $k+1$ to n of the matrix S, which is typically held in a single data structure. This leads to the use of expensive linked lists, or elbow room. In spite of these drawbacks the algorithm is attractive for several reasons, and it has been used by a few authors to develop incomplete factorizations [10, 25]. One of its

advantages is the ease with which powerful pivoting and reordering strategies can be implemented. The next section describes a different implementation which consists of swapping the k and i loops in Algorithm 2.

2.5. I, K, J variants of ILU. A more common alternative to implement ILU factorizations is based on the I, K, J version of Gaussian elimination. This is sketched in Algorithm 3.

ALGORITHM 3 (ILU).

Input: $A = (a_{ij}) \in \mathbb{R}^{n,n}$. *Output: ILU factorization* $A \approx LDU$.

0. $L = D = U = I$.
1. ***for*** $i = 1, \ldots, n$
2. $w = e_i^\top A$
3. ***for*** $k = 1, \ldots, i-1$ *and when* $w_k \neq 0$
4. $w_k = w_k / d_{kk}$
5. *Apply a dropping rule to* w_k.
6. ***for*** $j = k+1, \ldots, n$ *and if* ($w_k \neq 0$ *and* $u_{kj} \neq 0$)
7. $w_j := w_j - w_k u_{kj}$
8. ***end***
9. ***end***
10. $d_{ii} = w_i$
11. ***for*** *all* $j < i$: $l_{ij} = w_j$
12. ***for*** *all* $j > i$: $u_{ij} = w_j / w_i$. *Apply a dropping rule to* u_{ij}
13. ***end***

When the same static dropping strategy is used, e.g., one that is based on level-of-fill, it is known that Algorithm 3 and Algorithm 2, with S defined by (12), will deliver the same factors. However, this relation is still true for dynamic dropping strategies if the dropping rule is applied in the same way. Recall that Algorithms 3 and 2 perform the same sequence of operations in a different order. If an element is dropped in one, it will also dropped in the other if the same criterion is applied. For this to be true, one should be careful that the same rule is applied for partial results in the factorizations.

In practice, incomplete factorization algorithms are typically organized such that the L, D, U factors are stored in one single data structure. The attraction of the implementation in Algorithm 3 is clear: the rows of L and U are determined one at a time and are easily added to the existing data structure.

3. Relations between AINV and ILUs. There are two broad classes of approximate inverse methods. The first includes methods which compute directly an approximate inverse M to A; see, e.g., [9, 13]. The second includes those methods which obtain this approximate inverse in the form of a product of two triangular factors. A method in this category, called AINV, was proposed in [3, 4]. It is briefly outlined next.

3.1. Factored approximate inverse. The method in [3, 4] computes a decomposition of the form $W^\top AZ = D$, where W, Z are unit upper triangular matrices and D is a diagonal. In the exact factorization case, the matrices W and Z are the inverses of the factors $L^\top$ and U, respectively, of the standard LDU decomposition $A = LDU$, when this decomposition exists. The matrices W and Z can be directly computed by a biorthogonalization procedure. Indeed, since

$$W^\top A = DU$$

is upper triangular, we immediately get $e_i^\top W^\top A e_j = 0$ for any $j < i$, which means that column i of W is orthogonal to the first $i-1$ columns of A. A procedure can be devised to make the ith column of W orthogonal to the columns $1, \ldots, i-1$ of A via linear combinations with the first $i-1$ columns of W. Alternatively, columns $i+1, \ldots, n$ of W can be made orthogonal to the first i columns of A. This makes it possible to successively orthogonalize all columns of W against each of the columns of A. During this procedure one can drop small entries, or entries outside a certain sparsity pattern. The resulting incomplete biorthogonalization process, which is sketched next, produces an approximate factored inverse.

ALGORITHM 4 (factored approximate inverse (right-looking AINV)).

0. *Input:* $A = (a_{ij}) \in \mathbb{R}^{n,n}$. *Output:* Z, D, W *such that* $A^{-1} \approx Z D^{-1} W^\top$.
1. *Let* $p = q = (0, \ldots, 0) \in \mathbb{R}^n$, $Z = [z_1, \ldots, z_n] = I_n$, $W = [w_1, \ldots, w_n] = I_n$.
2. ***for*** $k = 1, \ldots, n$
3a. $p_k = w_k^\top A e_k, \quad q_k = e_k^\top A z_k$
4. ***for*** $i = k+1, \ldots, n$
5a. $p_i = \left(w_i^\top A e_k\right)/p_k, \quad q_i = \left(e_k^\top A z_i\right)/q_k$
6. *Apply a dropping rule to* p_i, q_i
7. $w_i = w_i - w_k p_i, \quad z_i = z_i - z_k q_i$
8. *Apply a dropping rule to* $w_{j,i}$ *and* $z_{j,i}$, *for* $j = 1, \ldots, i$.
9. ***end***
10. ***end***
11. *Choose diagonal entries of* D *as the components of* p *or* q.

Lines 3 and 5 are labeled with an "a" because they represent only one of two available options. An alternative way of computing W and Z is based on the fact that $W^\top A Z$ should become approximately diagonal. Instead of orthogonalizing W (respectively, Z) with respect to the columns of A, we can apply a biconjugation process that enforces the biorthogonality of the columns of W and Z. For this we must enforce $e_k^\top W^\top A Z e_j = 0$ for all $k \neq j$, $1 \leqslant k, j \leqslant i$. This will result in simple changes to Algorithm 11. Specifically, the second option which we label with a "b" consists of changing lines 3a and 5a into the following lines:

3b. $p_k = w_k^\top A z_k, \quad q_k = w_k^\top A z_k,$
5b. $p_i = \left(w_i^\top A z_k\right)/p_k, \quad q_i = \left(w_k^\top A z_i\right)/q_k.$

Clearly, if no entry is dropped and if there exists an LDU decomposition of A, then $W = L^{-\top}$, $Z = U^{-1}$. In this case it can be immediately seen by induction that after step i of the algorithm, columns $i+1, \ldots, n$ of W are orthogonal to column $1, \ldots, i$ of A, and likewise columns $i+1, \ldots, n$ of Z are orthogonal to rows $1, \ldots, i$ of A. Remarkably, the computations of Z and W can be performed independently of each other for option a.

It is important to note that in the original version of AINV [3, 4], no dropping is applied to p_i or q_i. One is only applied to w_i and z_i by discarding entries in W and Z that are less than a certain drop tolerance. Moreover, it has been pointed out in [3] that dropping entries of p and q produces poor results. The problem with dropping elements in p, q is that small entries $|p_j/p_i|$ may multiply large entries of $Z_{:,i}$, resulting in discarded entries in the approximate inverse that might not be small at all.

We still consider this variant because it shows very strong direct connections with various implementations of ILU. More general results that concern practical variants will be shown in section 3.4. In [3, 4] p and q were defined using option a, while option b was used in [16, 2] for symmetric positive definite matrices.

Note that the strict biorthogonality property of the exact factors no longer holds

if dropping is introduced. Interestingly, however, stability can still be proved for H-matrices in the case of ILU factorizations as well as for AINV; see, for example, [3].

3.2. ILU with progressive factor inversion. In order to establish a bridge between the AINV and the ILU approaches, we introduce an intermediate algorithm that can be viewed as an ILU process with a simultaneous inversion of the factors which it produces. Specifically, if at step $k-1$ we have a matrix U of the form

$$U = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ \mathrm{O} & 1 & \mathrm{O} \\ \mathrm{O} & \mathrm{O} & I \end{bmatrix},$$

the kth step will compute the entries in position (2,3) of the above matrix and add them to the current U to get U_{new}. Consider the row vector $q^\top = e_k^\top U - e_k^\top$. Note that the "diagonal" element q_k of q is zero. Then

$$U_{new} = U + e_k q^\top.$$

Because of the structure of U and q it is easy to see that $q^\top U = q^\top$, and so

$$U_{new} = (I + e_k q^\top)U.$$

Hence we have the relation

$$U_{new}^{-1} = U^{-1}(I + e_k q^\top)^{-1} = U^{-1}(I - e_k q^\top) = U^{-1} - U^{-1} e_k q^\top. \tag{16}$$

If we were to compute the inverse of U progressively, the columns $z_j, j = 1, \ldots, n$, of this corresponding progressive approximate inverse could therefore be updated by the following formula at the kth step:

$$z_i := z_i - z_k q_i, \qquad i = k+1, \ldots, n.$$

Analogous arguments hold for the L factor. This provides a formula for progressively computing $L^{-\top}$ and U^{-1} throughout the ILU factorization algorithm. We call the inverse factors W and Z as in Algorithm 4.

ALGORITHM 5 (ILU with progressive inversion of L and U).

Input: $A = (a_{ij}) \in \mathbb{R}^{n,n}$. *Output: ILU factorization* $A \approx LDU$.

0. $p = q = 0 \in \mathbb{R}^n$, $L = U = I$, $W = [w_1, \ldots, w_n] = Z = [z_1, \ldots, z_n] = I$, $S = A$.
1. ***for*** $k = 1, \ldots, n$
2. $\quad d_{kk} = s_{kk}$
3. $\quad$ ***for*** $i = k+1, \ldots, n$ *and when* $s_{ik} \neq 0$ *or* $s_{ki} \neq 0$
4. $\quad\quad p_i = s_{ik}/d_{kk}, \quad q_i = s_{ki}/d_{kk}$
5. $\quad\quad$ *Apply a dropping rule to* p_i *and* q_i
6. $\quad\quad l_{ik} = p_i, \quad u_{ki} = q_i$
7. $\quad\quad w_i = w_i - w_k p_i, \quad z_i = z_i - z_k q_i$
8. $\quad\quad$ *for all* $l \leqslant i$: $\quad$ *apply a dropping rule to* w_{li} *and to* z_{li}
9. $\quad\quad$ ***for*** $j = k+1, \ldots, n$ *and when* $s_{ik} \neq 0$ *and* $s_{kj} \neq 0$
10. $\quad\quad\quad$ ***if*** version(p_i, q_j) *then:* $\quad s_{ij} = s_{ij} - \frac{s_{ik} s_{kj}}{d_{kk}}$
11. $\quad\quad$ ***end***
12. $\quad$ ***end***
13. ***end***

3.3. The case of no dropping in W and Z. The notation we used in Algorithm 5 already suggests that Z, W coincide with those of Algorithm 4. This is confirmed by the next theorem which establishes a few relations between various versions of AINV and ILU.

THEOREM 6. *Assume that in Algorithms* 4 *and* 5 *the same dropping rules are applied to* p *and* q *and that no dropping rule is applied to* W, Z*. Then certain relationships between the factors* L*,* D*, and* U *on the one hand and the matrices* W, Z *and the vectors* p*,* q *on the other can be stated:*

1. *If Algorithm* 5 *with choice* (13) *and Algorithm* 4 *with option* a *are used, then*

$$L^{-\top} = W_{Alg.5/4}, \qquad \operatorname{diag}(D_{Alg.5}) = p_{Alg.4}.$$

2. *If Algorithm* 5 *with choice* (15) *and Algorithm* 4 *with option* b *are used, then*

$$L^{-\top} = W_{Alg.5/4}, \qquad U^{-1} = Z_{Alg.5/4}, \qquad \operatorname{diag}(D_{Alg.5}) = \begin{cases} p_{Alg.4}, \\ q_{Alg.4}. \end{cases}$$

3. *If Algorithm* 5 *with choice* (14) *and Algorithm* 4 *with option* a *are used, then*

$$U^{-1} = Z_{Alg.5/4}, \qquad \operatorname{diag}(D_{Alg.5}) = q_{Alg.4}.$$

Proof. We will prove only the first result for W and p, since the proof for the other cases is analogous. We will show by induction on k that W is identical in both methods after any step k, that the first k diagonal entries of D coincide with $p_1, \ldots, p_k$, and that

$$s_{ij} = w_i^\top A e_j \text{ for all } i, j > k. \tag{17}$$

Initially, for $k = 0$ there is nothing to show since obviously $W = I$ in both algorithms and $S = A$. Now suppose that W is identical before we enter step k of each algorithm. Suppose that the first $k-1$ diagonal entries of D coincide with the first $k-1$ components of p and that

$$s_{ij} = w_i^\top A e_j \text{ for all } i, j \geqslant k.$$

We immediately obtain $p_k = w_k^\top A e_k = s_{kk} = d_{kk}$ and

$$p_i = w_i^\top A e_k / d_{kk} = s_{ik}/d_{kk} \text{ for all } i = k+1, \ldots, n.$$

From this it follows that $p_i^{(new)}$ from both algorithms is identical and satisfies $p_i^{(new)} = l_{ik}$ for any $i > k$. Since we choose the same dropping rule for both algorithms, this equality still holds after sparsifying these entries.

Obviously the update procedure

$$w_i^{(new)} = w_i^{(old)} - w_k^{(old)} p_i^{(new)} \text{ for all } i > k$$

from Algorithm 4 is the nontrivial update part on

$$W^{(new)} = W^{(old)} \left(I - e_k p^\top\right)$$

in Algorithm 5. Now for the new entries s_{ij}, $i,j > k$, we have the update procedure

$$\begin{aligned} s_{ij}^{(new)} &:= s_{ij}^{(old)} - p_i^{(new)} s_{kj}^{(old)} \\ &= e_i^\top \begin{bmatrix} I_{k-1} & & & & \\ & 1 & & & \\ & -p_{k+1}^{(new)} & 1 & & \\ & \vdots & & \ddots & \\ & -p_n^{(new)} & & & 1 \end{bmatrix} S^{(old)} e_j \\ &= e_i^\top \begin{bmatrix} I_{k-1} & & & & \\ & 1 & & & \\ & -p_{k+1}^{(new)} & 1 & & \\ & \vdots & & \ddots & \\ & -p_n^{(new)} & & & 1 \end{bmatrix} (W^{(old)})^\top A e_j = e_i^\top (W^{(new)})^\top A e_j. \end{aligned}$$

This completes the proof. ∎

3.4. Dropping elements in *W* and *Z*. As mentioned earlier, Algorithm 4 is more general than the original AINV algorithm [3, 4], which does not allow dropping entries in the update factors from p, q but only in the updated matrices Z, W. The previous theorem does not address this case since its assumptions do not allow dropping in W and Z. The key to getting a connection between AINV and ILU-type factorizations lies in (17). If a Schur complement is constructed so that this relation holds between AINV and an ILU factorization, it is easy to see that both algorithms will result in comparable W's and Z's. The results to be proved next concern such update versions, i.e., they are valid for the Schur complements defined by any one of the following three expressions:

$$S = (W^\top A)_\square\,, \quad S = (AZ)_\square\,, \quad \text{or} \quad S = (W^\top A Z)_\square\,, \tag{18}$$

where the square subscripts indicate that an appropriate submatrix is extracted. In these situations, we may expect for example $W^\top$ to be close to L^{-1} in some sense, i.e., that $W^\top$ can be viewed as an approximation to the inverse of L.

We will need two simple lemmas before establishing the general result. We begin with some required additional notation. The matrix W at the kth step of Algorithm 5 is denoted by $W^{(k)}$, starting with $W^{(0)} = I$. It is obtained from $W^{(k-1)}$ by the relation

$$W^{(k)} = W^{(k-1)} \left[I - e_k (p^{(k)})^\top \right] - G_k, \tag{19}$$

where G_k is the matrix of elements that have been dropped in the process and $p^{(k)}$ is the vector denoted by p in the algorithm, as step k. The vector $p^{(k)}$ has zero elements in positions 1 through k, i.e., $e_j^\top p^{(k)} = 0$ for all $j \leqslant k$.

LEMMA 7. *Denote by Q_k the matrix*

$$Q_k = I - e_k (p^{(k)})^\top, \tag{20}$$

and let G_k be the matrix of elements dropped in the matrix $W^{(k)}$ at step k. Then

$$G_k\, Q_{k-l} = G_k, \quad 0 \leq l \leq k-1. \tag{21}$$

Proof. Note that $(G_k)_{ij} = 0$ for $j \leqslant k$ or $i > k$. Therefore, we can write

$$G_k = \sum_{i \leqslant k,\ j>k} g_{ij} e_i e_j^\top,$$

and so

$$G_k Q_{k-l} = \sum_{i \leqslant k,\ j>k} g_{ij} e_i e_j^\top \left(I - e_{k-l} (p^{(k-l)})^\top \right) = \sum_{i \leqslant k,\ j>k} g_{ij} e_i e_j^\top = G_k. \qquad \square$$

This relation is key to establishing the next lemma.

LEMMA 8. *Let $W^{(k)}, G_k$ be defined by* (19) *and* (20) *and let*

$$L_k^{-\top} = Q_1 \times Q_2 \cdots Q_k.$$

Then

$$I - W^{(k)} L_k^\top = \sum_{i=1}^{k} G_i. \tag{22}$$

Proof. Exploiting the result of Lemma 7 we can write

$$\begin{aligned} W^{(k)} &= W^{(k-1)} Q_k - G_k = (W^{(k-1)} - G_k) Q_k \\ &= [(W^{(k-2)} - G_{k-1}) Q_{k-1} - G_k] Q_k = [W^{(k-2)} - G_{k-1} - G_k] Q_{k-1} Q_k \\ &= [W^{(k-3)} - G_{k-2} - G_{k-1} - G_k] Q_{k-2} Q_{k-1} Q_k \\ &= \cdots \\ &= [W^{(0)} - G_1 - \cdots - G_k] Q_1 Q_2 \cdots Q_k. \end{aligned}$$

This essentially gives the result by recalling that $W^{(0)} \equiv I$. $\square$

We now need to link the AINV algorithm (Algorithm 4) with Algorithm 5. To interpret AINV as a form of ILU, the definition of the approximate Schur complement must be adapted. Standard computations of the Schur complement in Algorithm 2 correspond to the definition in (15), (12). We now consider a hypothetical version of Algorithm 5, in which the Schur complement is defined via one of the options in (18).

An important observation is that we will obtain the same W matrices in Algorithms 4 and 5 if the same dropping rule is used for p in both algorithms and if the Schur complement is defined from (18) in Algorithm 5.

Lemma 8 indicates that $W^{(k)}$ is an approximate inverse of $L_k^\top$ if the sum of the matrices G_i remains small, a statement which can be made more precise if a drop tolerance strategy is invoked. Putting these observations together leads to the following result.

THEOREM 9. *Assume that in Algorithm* 5 *w_{ij} is dropped if $|w_{ji}| \leqslant \varepsilon$, $i \leqslant k, j > k$. Then, the L-factor and the matrix W produced by Algorithm* 5 *are such that*[1]

$$|(I - W L^\top)_{ij}| \leqslant (j-i)\varepsilon, \quad 1 \leqslant i \leqslant j \leqslant n. \tag{23}$$

If in addition the Schur complement in Algorithm 5 *is defined through* (18)*, and if the related version of Algorithm* 4 *uses the same dropping rules for W as Algorithm* 5*,*

[1]The factor $(j-i)$ is a rough overestimate which, as the proof indicates, can be reduced significantly. It is bounded by the number of times that dropping has occurred in position (i,j) during the algorithm. A good reordering strategy and a graph-theoretical approach may also lead to a lower factor.

whereas no dropping is applied to p, q, then the matrices W produced by both algorithms are identical.

Proof. The first part of the theorem follows by applying the previous lemma with $k \equiv n$ and, noting that in position (i, j) of W, dropping occurs at most $(j - i)$ times since at step k dropping takes place only in the rectangle of pairs (i, j) such that $i < k < j$.

The second part of the theorem was stated above without proof. A rigorous proof would be by induction. In short, both sequences satisfy the same recurrence relation,

$$W^{(k)} = W^{(k-1)}(I - e_k(p^{(k)})^\top) - G_k,$$

because $p^{(k)}$ and G_k are the same in both algorithms due to the common dropping rules. This leads to the same sequence of W's for both algorithms. ∎

Though all the analysis has been made for the lower triangular factor L and the associated W, it is clear that analogous relationships can be established between U^{-1} and Z (apply Theorem 9 to $A^\top$). We mention that an algorithm of this type was recently presented in [5] for the symmetric positive definite case where the Cholesky factor was obtained as a by-product of the AINV factor.

We now consider the more general situation when no dropping is applied to p and q in Algorithm 4 while Algorithm 5 does perform dropping. In this case the W matrices obtained by both algorithms are no longer (easily) comparable. This is because the vectors p, q in the recurrence (19) are no longer the same. We could modify Algorithm 5 so that dropping is also not done in p and q but only in L after p, q have been used to update W and Z. This amounts to simply moving line 7 of the algorithm to behind line 4. Specifically, only lines 5–7 change in the algorithm and they become

5a. $w_i = w_i - w_k p_i, \quad z_i = z_i - z_k q_i.$
6a. Apply a dropping rule to p_i and q_i.
7a. $l_{ik} = p_i, \quad u_{ki} = q_i.$

We will refer to this algorithm as the a-version of Algorithm 5. If the goal is to mimic the behavior of the actual AINV (no dropping in p, q), then clearly this version is more suitable and practical.

There are now two sequences of L matrices produced by this version of the algorithm. One is the sequence L_k seen before which uses the vectors $p^{(k)}$ before dropping. The second is a sequence $\tilde{L}_k$ which corresponds to the actual L-factors produced by the factorization and which uses the vectors p, q after dropping is applied. Therefore, we define the elementary factors corresponding to this second sequence:

$$\tilde{Q}_k = I - e_k(\tilde{p}^{(k)})^\top = I - e_k(p^{(k)} - f_k)^\top \tag{24}$$

in which f_k is the column vector of elements that have been dropped in $p^{(k)}$, and

$$\tilde{L}_k^{-\top} = \tilde{Q}_1 \times \tilde{Q}_2 \cdots \tilde{Q}_k,$$

which is the transpose of the inverse L-factor produced at the end of step k of algorithm 5. A standard result of LU factorizations is that $\tilde{L}_k$ is simply the matrix with column vectors $\tilde{p}^{(i)}$, $i = 1, \ldots, k$, to which we add the identity. Similarly for L_k. Therefore, it is clear that

$$L_k^\top - \tilde{L}_k^\top = \sum_{i=1}^{k} e_i f_i^\top. \tag{25}$$

We define

$$F_k = \sum_{i=1}^{k} e_i f_i^\top. \tag{26}$$

Putting (25) into (22) gives the following generalization of Theorem 9.

THEOREM 10. *Assume that the* a*-version of Algorithm* 5 *is used, and let* $W^{(k)}$, G_k, *and* F_k *be defined by* (19) *and* (26) *and* $\tilde{L}_k$ *be the L-factor obtained at step* k *of the same algorithm. Then the following equality holds:*

$$I - W^{(k)}\tilde{L}_k^\top = \sum_{l=1}^{k} G_l + W^{(k)}F_k. \tag{27}$$

Furthermore, assume that at step k *of Algorithm* 5 *an entry* l_{ik} *is discarded if*

$$|l_{ik}| \ \times \max_{j=k,\ldots,n} |w_{jk}| \leqslant \varepsilon,$$

whereas no dropping is applied for p, q *in Algorithm* 4. *In both algorithms it is assumed that* w_{ij} *is dropped if*

$$|w_{ij}| \leqslant \varepsilon, \quad i \leqslant k, j > k. \tag{28}$$

Then for Algorithm 5 *the following holds for any* $j > i$*:*

$$|\ (I - W\tilde{L}^\top)_{ij}\ | \leqslant 2(j-i)\varepsilon. \tag{29}$$

If, in addition, the Schur complement in Algorithm 5 *is defined through* (18), *then the matrices* W *produced by Algorithm* 5 *and the related version of Algorithm* 4 *are identical.*

Proof. Relation (27) follows immediately from (25) and (22). Denote $W^{(n)}$ by W, and similarly F_n by F. In the remainder of the proof, we write W as

$$W = \sum_{k=1}^{n} w_k e_k^\top$$

from which we infer that

$$WF = \sum_{k=1}^{n} w_k e_k^\top \sum_{k=1}^{n} e_k f_k^\top = \sum_{l=1}^{n} w_l f_l^\top.$$

We now consider the entry (i,j) on both sides of (27):

$$|e_i^\top (I - W\tilde{L}^\top)e_j| \leqslant \left| \sum_{l=1}^{n} e_i^\top G_l e_j \right| + \left| e_i^\top WF e_j \right|. \tag{30}$$

From Theorem 9, we already have a bound for the first term on the right-hand side:

$$\left| \sum_{l=1}^{n} e_i^\top G_l e_j \right| \leqslant (j-i)\varepsilon, \quad 1 \leqslant i \leqslant j \leqslant n. \tag{31}$$

For the second term, we write

$$|e_i^\top WFe_j| = \left| \sum_{k=1}^{n} e_i^\top w_k f_k^\top e_j \right| \leqslant \sum_{k=1}^{n} |e_i^\top w_k| \, |f_k^\top e_j|.$$

Notice that $e_i^\top w_k = 0$ for $k < i$ and similarly $f_k^\top e_j = 0$ for $k \geqslant j$, so the above inequality becomes

$$|e_i^\top WFe_j| \leqslant \sum_{k<j,\ k\geqslant i} |e_i^\top w_k| \, |f_k^\top e_j| \leqslant \sum_{k<j,\ k\geqslant i} \max_i |w_{ki}| \, |f_k^\top e_j|.$$

According to the dropping strategy each term in the sum does not exceed ε. Therefore,

$$|e_i^\top WFe_j| \leqslant \sum_{k<j,\ k\geqslant i} \varepsilon = (j-i)\varepsilon. \tag{32}$$

Substituting (31) and (32) into (30) yields the desired result (29). ∎

As in Theorem 9 all the analysis can be carried over to establish analogous relationships between U^{-1} and Z.

3.5. Left-looking AINV. An equivalent alternative to Algorithm 4, at least without dropping, was suggested in [4] and was referred to as the "left-looking" version of AINV. The method consists essentially of computing the approximate inverses W and Z column-wise instead of using rank-1 updates as in Algorithm 4.

ALGORITHM 11 (factored approximate inverse (left-looking AINV)).

Input: $A = (a_{ij}) \in \mathbb{R}^{n,n}$. *Output:* Z, D, W *such that* $A^{-1} \approx ZD^{-1}W^\top$.

0. $p = q = 0 \in \mathbb{R}^n$, $p_1 = q_1 = a_{11}$; $W = Z = D = I_n$
1. ***for*** $i = 2, \ldots, n$
2. ***for*** $j = 1, \ldots, i-1$
3a. $P_j = \left(w_i^\top A e_j\right)/p_j \qquad Q_j = \left(e_j^\top A z_i\right)/q_j$
4. *apply a dropping rule to* P_j *and* Q_j.
5. $w_i = w_i - w_j \, P_j, \qquad z_i = z_i - z_j \, Q_j$
6. *for all* $l \leqslant i$: *apply a dropping rule to* w_{li}, z_{li}.
7. ***end***
8a. $p_i = w_i^\top A e_j, \quad q_i = e_j^\top A z_i$
9. ***end***
10. *Choose diagonal entries of* D *as the components of* p *or* q.

This algorithm is almost identical to Algorithm 4 except that the updates in Z, W are now performed in sequence, column by column, while in Algorithm 4 the updates are performed simultaneously for all columns. This difference corresponds to the difference between two equivalent formulations of the modified Gram–Schmidt orthogonalization procedure, one which completes the computation of the k-column of the orthogonal matrix Q at step k and the other which updates columns $k+1$ to n of Q at each step k.

Similarly to Algorithm 4, Algorithm 11 also has a b option which consists of the following changes to lines 3a and 8a:

3b. $P_j = \left(w_i^\top A z_j\right)/p_j, \qquad Q_j = \left(w_j^\top A z_i\right)/q_j,$
8b. $p_i = w_i^\top A z_j, \quad q_i = w_j^\top A z_i.$

The simple relation between Algorithms 4 and 11 is stated in the following proposition, which is straightforward to verify.

PROPOSITION 12. *Assume that the same dropping rule is applied to p, q and P, Q and that the same dropping rule is also applied to W and Z in Algorithm* 4 *and Algorithm* 11. *Then both algorithms will compute the same W, Z. They also compute the same D if the same choice is made for the D entries in their lines* 11 *and* 10, *respectively.*

In fact, the equality between both algorithms also includes the case when each column is sparsified only once. For Algorithm 11 this would be a more natural dropping rule, i.e., entries of z_{li}, w_{li} would be discarded only if $j = i - 1$. For step k of Algorithm 4 the associated dropping rule would sparsify only column $k+1$ of W and Z which might lead to large amounts of fill-in for W and Z.

3.6. Bordering methods. An analysis similar to the one developed in the previous sections was discussed in the earlier report [8] which established links between ILU and approximate inverse methods based on "bordering." An approximate inverse method of this type was discussed in [22]. The main idea is to partition the (k, k) principal submatrix of A as

$$A_k = \begin{pmatrix} B & f \\ e^\top & c \end{pmatrix},$$

where $B \equiv A_{k-1}$ is of dimension $k-1$. Assume that we already know the factorized approximate inverse of B in the form $W^\top BZ = D$, where W, Z are unit upper triangular and D is diagonal. Then the factored inverse of A_k can be obtained by writing

$$W_{new}^\top A_k Z_{new} \equiv \begin{pmatrix} W & g \\ 0 & 1 \end{pmatrix}^\top \begin{pmatrix} B & f \\ e^\top & c \end{pmatrix} \begin{pmatrix} Z & h \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & s \end{pmatrix}.$$

The above relation immediately shows that h, s, g must satisfy the equations

$$B^\top g = -e, \tag{33}$$

$$Bh = -f, \tag{34}$$

$$s = c + g^\top f + e^\top h + g^\top Bh. \tag{35}$$

To develop approximate inverse methods, we can simply use vectors g, h provided from approximately solving systems (33) and (34) and then computing s. In fact, we could simply utilize the relation $W^\top BZ \approx D$ to approximate h, g and discard some entries according to a dropping rule. This means that we compute g and h from

$$g := -WD^{-1}Z^\top e, \qquad h := -ZD^{-1}W^\top f$$

and apply a dropping rule to g and h. The algorithm now becomes clear. Start with the (1,1) matrix which has a trivial factorized approximate inverse and then build, recursively, the approximate inverses of the (k, k) principal submatrix of A_k from that of the $(k-1, k-1)$ principal submatrix, for $k = 2, \ldots, n$. It is also possible to develop additional variants to the algorithm depending on how the diagonal elements of D are selected. Denote the columns of the final W and Z matrices by w_j and z_j, $j = 1, \ldots, n$. Then, (35) corresponds to the choice $s = w_k^\top A_k z_k$. Two other choices are obtained by taking $s = c + g^\top f \equiv w_k A_k e_k$ which corresponds to (13) or $s = c + e^\top h \equiv e_k^\top A_k z_k$ which is analogous with (14).

In [8] a result similar to Theorem 6 was mentioned, though this applies again to nonpractical versions. However, it was also shown that there is a practical relation

between bordered factored inverse methods and Algorithm 11. Specifically, a first result is that both algorithms compute the same W and D when (1) the choice $s = w_k^\top A_k e_k$ is used in the bordered approximate inverse algorithm; (2) the same dropping rule is used for W in both algorithms; and (3) no dropping is applied to Z in the bordering method. A second result is that both algorithms compute the same D and Z under analogous conditions.

4. Consequences. The comparison results established in the previous sections can provide theoretical insight into known algorithms by exploiting the body of existing literature on ILU and AINV. On the practical side, they can also help develop improved variants of both ILU and AINV. In fact, new algorithms have already been developed by exploiting these relationships in both directions. In the following we briefly discuss a few of these known results and point to other potential applications yet to be explored.

4.1. AINV with pivoting. In [7] we applied what is known from ILU algorithms to devise pivoting techniques for approximate inverse methods. This technique can be easily inferred from the following relation which holds at step k:

$$W^\top A Z \approx \begin{pmatrix} D & 0 \\ 0 & S \end{pmatrix},$$

where $W^\top$ and Z are the inverses of the matrices L and U, respectively, of the LU factorization. As was seen in earlier sections, these are also close to the W and Z matrices obtained at step k of the AINV procedure. If we apply permutations $\Pi^\top$ and Σ to S, on the left and right, respectively, it is easy to determine how this permutation must be also applied to W and Z for consistency:

$$\begin{pmatrix} I & 0 \\ 0 & \Pi^\top \end{pmatrix} W^\top A Z \begin{pmatrix} I & 0 \\ 0 & \Sigma \end{pmatrix} \approx \begin{pmatrix} D & 0 \\ 0 & \Pi^\top S \Sigma \end{pmatrix}.$$

This means that the corresponding rows of W and Σ need to be permuted according to the permutation applied to S. As for which permutation to apply, we can use the parallel with ILU, since S is more or less the same matrix that is obtained from the ILU factorization. For example, we can simply do a column permutation as is done in ILUTP [21]. The strategy suggested in [7] is to use row and column pivoting successively a few times (in the same step) until the pivot satisfies a certain stability condition both for the kth row and the kth column of S. For details see [7]. Numerical experiments do confirm that this procedure is much more robust than AINV with no pivoting.

4.2. An ILU based on monitoring the growth factors. Proceeding in the reverse direction, the relationships established in this paper have also allowed us to design more robust ILU techniques. Here, we cite two independent works [6, 5]. The paper [6] introduces dropping strategies in ILU that are more rigorous than simple threshold techniques by exploiting the parallel between ILU and AINV [6].

The fundamental relation which was exploited in [6] is (28). As shown by Theorem 10 this relation ensures that the W matrix is close to the inverse of the factor L. Therefore the L-factor will clearly be stable, in the sense that its inverse will have a moderate norm. Similarly for the U factor.

In [5], an incomplete Cholesky factorization was extracted as a by-product of the AINV process for the symmetric positive definite case [2]. This can be seen as another

way of exploiting the relationships between ILU and AINV. Numerical observation has shown that AINV preconditioning often outperforms the standard incomplete Cholesky factorization for the conjugate gradient. In [5], it was shown that only the by-product incomplete Cholesky decomposition was able to obtain results comparable with those of AINV. However, the crucial dropping strategy (28) is not employed. We believe that such a dropping strategy may substantially enhance the quality of the factor produced by the method.

4.3. Theory: Results for SPD matrices and for *H*-matrices. From a theoretical point of view, some results on approximate inverse methods can be derived by exploiting the relationship with ILU, for which much is known. This line of argument was indeed exploited, for example, in [2] by transferring the related incomplete Cholesky decomposition [23]. An immediate corollary for the symmetric positive definite case is the following.

COROLLARY 13. *Let A be symmetric positive definite. Suppose that Algorithms* 4 *and* 5 *apply the same dropping rule to p and q and that no dropping is applied to W and Z.*

If option b *is used in Algorithm* 4 *and if S in Algorithm* 5 *is defined via* (15)*, then both algorithms do not break down. In addition, both methods compute the same W and Z and $W = Z$. The diagonal entries of S in Algorithm* 5 *are positive and coincide with the entries of $p = q$ in Algorithm* 4.

Proof. This follows immediately from Theorem 6 and Property 1. ∎

It is well known that the ILU decomposition of an H-matrix exists for any of the dropping strategies discussed in section 2.3; see, e.g., [19, 18]. It immediately follows that W and Z of Algorithm 4 exist for this case. Likewise for M-matrices we know that the computed L and U are again M-matrices. Consequently W and Z have to be nonnegative in this case. However, this argument applies only to the theoretic way of dropping in p and q. A proof for the natural way of dropping is given in [3].

4.4. Further applications. The few applications just described indicate that much can be gained by exploiting good qualities of a technique from one class to improve the corresponding algorithm from the other class. Another possible application which does not seem to have been explored is to exploit level-of-fill strategies used in ILU techniques, for developing pattern-based dropping strategies for AINV methods. Finding good patterns for dropping in AINV methods remains poorly understood. For matrices with good diagonal dominance properties, level-of-fill techniques work quite well, and when combined with blocking they are often the preferred techniques for solving certain types of problems in fluid dynamics, for example.

In ILU(p) a level-of-fill *lev* is attributed to each element during factorization. Each element that is updated by formula such as (9) will have its `lev` value updated by the formula

$$\texttt{lev}(s_{ij}) = \min\{\texttt{lev}(s_{ij}) \ , \ \texttt{lev}(s_{ik}) + \texttt{lev}(s_{kj}) + 1\}.$$

Initially, any nonzero element is assigned a `lev` value of 0, and any zero element is (implicitly) assigned an infinite `lev` value. It is typical to process the ILU factorization in two phases, a symbolic one and a numeric one. The pattern of ILU(p) is determined in the symbolic factorization. This pattern can now be used for obtaining a pattern for AINV. Consider, in line 5 of Algorithm 4, the update to w_j the jth column of W. This update is $w_j = w_j - w_k p_j$, or, component-wise $w_{ij} = w_{ij} - w_{ik} p_j$. Now recall

that p_j is nothing but s_{jk}, so

$$w_{ij} = w_{ij} - w_{ik}s_{jk}.$$

Using the same model for decrease of the elements in the factorization, we can easily see that a good way to define the level-of-fill of w_{ij} is

$$\texttt{lev}(w_{ij}) = \min\{\texttt{lev}(w_{ij}) \ , \ \texttt{lev}(w_{ik}) + \texttt{lev}(s_{jk}) + 1\}.$$

Notice that computing the `lev` values for the L factors is inexpensive.

A hint at another potential class of applications is provided by the recent paper [6]. There, some information about W, Z is exploited to gain insight on suitable dropping strategies when building L and U. ILU and AINV can be viewed as some kind of optimization methods, which produce factors that approximate either A directly (for ILU) or its inverse (for AINV). A rule of thumb seems to be that ILU works better than AINV methods when it produces factors that are stable. In other words, *accuracy+stability* $\rightarrow$ *fast convergence*. If we find factors L from ILU, such that $L^{-1} \approx W$ and W is well behaved, then clearly both criteria of accuracy and stability are satisfied. This suggests that strategies which combine both criteria should be developed. In [6] a dropping strategy was found which ensured that $L^{-1} \approx W$ – using the result of Theorem 10. Other strategies may exist.

5. Conclusions. We have shown a number of interrelations between factored approximate inverse and related incomplete factorizations of ILU type. We also established relations between different approaches to compute factored approximate inverses. It was shown that approximate inverse techniques are intimately related to ILU factorizations. Indeed, they can be viewed as a process for obtaining the inverses of the L and U factors directly from the elementary subfactors that arise in Gaussian elimination. What is interesting is that with an appropriate set of assumptions on the patterns used for dropping, many other relationships can be established. This equivalence permits one to establish some results on existence and, more generally, to better understand the algorithms. For example, it is now clear that ILU and AINV factorizations are two extremes where elementary factors are all inverted (in AINV) or kept as are they are (in ILUs). It is also clear, however, that there is a multitude of variation in between these two extremes and it is quite conceivable that better methods would be adaptive algorithms that lie in between—where adaptivity here is understood in relation to stability.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, New York, 1994.
[2] M. Benzi, J. K. Cullum, and M. Tůma, *Robust approximate inverse preconditioning for the conjugate gradient method*, SIAM J. Sci. Comput., 22 (2000), pp. 1318–1332.
[3] M. Benzi, C. D. Meyer, and M. Tůma, *A sparse approximate inverse preconditioner for the conjugate gradient method*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.
[4] M. Benzi and M. Tůma, *A sparse approximate inverse preconditioner for nonsymmetric linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 968–994.
[5] M. Benzi and M. Tůma, *A robust incomplete factorization preconditioner for positive definite matrices*, Numer. Linear Algebra Appl., to appear.
[6] M. Bollhöfer, *A robust ILU with pivoting based on monitoring the growth of the inverse factors*, Linear Algebra Appl., 338 (2001), pp. 201–218.
[7] M. Bollhöfer and Y. Saad, *A factored approximate inverse preconditioner with pivoting*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 692–705.

[8] M. Bollhöfer and Y. Saad, *ILUs and Factorized Approximate Inverses Are Strongly Related. Part* I: *Overview of Results*, Technical Report umsi–2000-39, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2000.

[9] E. Chow and Y. Saad, *Approximate inverse preconditioners via sparse-sparse iterations*, SIAM J. Sci. Comput., 19 (1998), pp. 995–1023.

[10] E. F. D'Azevedo, F. A. Forsyth, and W. P. Tang, *Towards a cost-effective high order ILU preconditioner*, BIT, 31 (1992), pp. 442–463.

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

[12] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.

[13] M. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses*, SIAM J. Sci. Comput., 18 (1997), pp. 838–853.

[14] W. Hackbusch, *Iterative Solution of Large Linear Systems of Equations*, Springer-Verlag, New York, 1994.

[15] I. E. Kaporin, *New convergence results and preconditioning strategies for the conjugate gradient method*, Numer. Linear Algebra Appl., 1 (1994), pp. 179–210.

[16] S. Kharchenko, L. Kolotilina, A. Nikishin, and A. Yeremin, *A reliable AINV–type preconditioning method for constructing sparse approximate inverse preconditioners in factored form*, Numer. Linear Algebra Appl., 8 (2001), pp. 165–179.

[17] L. Y. Kolotilina and A. Y. Yeremin, *Factorized sparse approximate inverse preconditionings I. Theory*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 45–58.

[18] T. Manteuffel, *An incomplete factorization technique for positive definite linear systems*, Math. Comp., 34 (1980), pp. 473–490.

[19] J. Meijerink and H. A. Van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M–matrix*, Math. Comp., 31 (1977), pp. 148–162.

[20] N. Munksgaard, *Solving sparse symmetric sets of linear equations by preconditioned conjugate gradient method*, ACM Trans. Math. Software, 6 (1980), pp. 206–219.

[21] Y. Saad, *ILUT: A dual threshold incomplete ILU factorization*, Numer. Linear Algebra Appl., 1 (1994), pp. 387–402.

[22] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.

[23] M. Tismenetsky, *A new preconditioning technique for solving large sparse linear systems*, Linear Algebra Appl., 154–156 (1991), pp. 331–353.

[24] A. C. Van Duin, *Scalable Parallel Preconditioning with the Sparse Approximate Inverse of Triangular Matrices*, Preprint, Department of Computer Science, Rijksuniversiteit Leiden, Leiden, The Netherlands, 1997.

[25] Z. Zlatev, *Use of iterative refinement in the solution of sparse linear systems*, SIAM J. Numer. Anal., 19 (1982), pp. 381–399.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 238–259

A CASE FOR A BIORTHOGONAL JACOBI–DAVIDSON METHOD: RESTARTING AND CORRECTION EQUATION*

ANDREAS STATHOPOULOS†

Abstract. We propose a biorthogonal Jacobi–Davidson method (biJD), which can be viewed as an explicitly biorthogonalized, restarted Lanczos method, that uses the approximate solution of a correction equation to expand its basis. Through an elegant formulation, the algorithm allows for all the functionalities and features of the Jacobi–Davidson method (JD), but it also includes some of the advantages of nonsymmetric Lanczos.

The main motivation for this work stems from a correction equation and a restarting scheme that are possible with biJD but not with JD. Specifically, a correction equation using the left approximate eigenvectors available in biJD yields cubic asymptotic convergence, as opposed to quadratic with the JD correction equation. In addition, a successful restarting scheme for symmetric JD depends on the Lanczos three-term recurrence and thus can only apply to biJD. Finally, methods that require a multiplication with the adjoint of the matrix need to be reconsidered on today's computers with memory hierarchies, as this multiplication can be performed with minimal additional cost.

We describe the algorithm, its features, and the possible functionalities. In addition, we develop an appropriate correction equation framework and analyze the effects of the new restarting scheme. Our numerical experiments confirm that biJD is a highly competitive method for a difficult problem.

Key words. Jacobi–Davidson, BCG, eigenvalues, Lanczos, three-term recurrence, preconditioning

AMS subject classification. 65F15

PII. S0895479800373371

1. Introduction.

The solution of the large eigenvalue problem $A\tilde{x} = \tilde{\lambda}\tilde{x}$ for a few eigenvalues closest to a given value σ and their corresponding eigenvectors (together eigenpairs) is recognized as a harder problem than the solution of a linear system of equations with A. Because the eigenvalues are not known a priori, the system to be solved is nonlinear [46, 45]. Even if the eigenvalues were known, the resulting linear system would be indefinite for any eigenvalue that lies inside the spectrum. This usually implies slow convergence of the linear solver, and moreover it is hard to obtain good preconditioners [35, 1]. Nonnormality and ill-conditioning exacerbate these problems further.

Preconditioning is also not straightforward to apply on eigenvalue iterative solvers. Early attempts included variants of the Davidson method [11] and shift-and-invert methods [31], but Jacobi–Davidson-type methods (JD) have provided an appropriate preconditioning framework for eigensolvers [38].

Another important problem with eigensolvers is their high storage requirements. For linear systems, storage is less of an issue, because three-term recurrence methods, such as CG and BCG, are as effective as full orthogonalization Arnoldi-type methods [23, 14]. In contrast, the three-term recurrence Lanczos method for eigenproblems needs to store the basis vectors to recover eigenvector approximations [22]. Moreover, most eigenvalue methods that use preconditioning do not build a Krylov space, and thus full orthogonalization methods, like JD, are necessary [11, 38, 34].

*Received by the editors October 18, 2000; accepted for publication (in revised form) by R. Freund February 11, 2002; published electronically July 1, 2002. This work was supported by the College of William and Mary.

http://www.siam.org/journals/simax/24-1/37337.html

†Department of Computer Science, College of William and Mary, Williamsburg, VA 23187-8795 (andreas@cs.wm.edu).

For all the above reasons, the more complicated and expensive-per-step methods of Arnoldi and JD are usually preferred over the computational simplicity of the Lanczos method. The JD method can be viewed as an inner-outer method. At each outer step, JD incorporates into the basis an approximate solution of the correction equation

$$(I - xx^*)(A - \sigma I)(I - xx^*)\ \delta = (I - xx^*)(\mu I - A)x, \tag{1.1}$$

where (μ, x), with $\|x\| = 1$, is an approximation of the required eigenpair. Typically, a few steps of a Krylov iterative solver are applied to (1.1), and available preconditioners for A can be used through a technique described in [38, 15, 37].

Despite its improved convergence, for many hard problems the JD method may still require a large number of steps, with overwhelming storage requirements. This problem is controlled by restarting the method when the basis size reaches a user-specified upper limit. Because the JD basis is not required to be Krylov, a variety of restarting techniques can be used, including implicit restarting with various shift strategies [40] and thick restarting which, at restart, retains either Ritz vectors [43] or Schur vectors [15]. A host of other improvements on targeting eigenvalues, harmonic Ritz approximations, preconditioning techniques, and extensions to the generalized eigenvalue problem have helped make JD a robust and widely used method [15, 37].

Yet, the nonsymmetric Lanczos method has two very appealing characteristics that could offer significant advantages if exploited in a JD framework. The Lanczos method maintains both a left and a right biorthogonal basis, generated by A^* and A, respectively. As a result, approximations for the left eigenvectors are also obtained, and biorthogonality is implicitly maintained through a three-term recurrence.

The availability of approximations to left eigenvectors suggests a natural alternative correction equation, based on an approximate spectral projector:

$$(I - xy^*)(A - \sigma I)(I - xy^*)\ \delta = (I - xy^*)(\lambda I - A)x, \tag{1.2}$$

where (λ, x) is an approximation of the right eigenpair of A, and y is an approximation of the corresponding left eigenvector, such that $\|x\| = 1$ and $y^*x = 1$. An interesting observation is that we can solve both (1.2) and its conjugate transpose by a single BCG iteration, improving simultaneously both left and right eigenpairs. Performing inverse iteration with these two conjugate matrices is known to converge cubicly [45], and Sleijpen et al. have proved the same convergence rate for JD with the above correction equation [37]. However, for the latter to hold, y must converge to the left eigenvector, which does not hold in general if only a right space is considered.

The left and right biorthogonal bases also suggest an effective restarting scheme. Restarting has significant performance shortcomings, since important components of the invariant subspace may be discarded. Restarting techniques attempt to identify and retain these important components to help future convergence. One class of techniques achieves this by retaining certain Ritz vectors that tend to improve convergence toward the desired eigenpair in a deflation-like way [43, 9, 15]. Another class traces the problem to the orthogonality lost when restarting and tries to reinstate it by keeping those vector directions against which the basis tends to lose orthogonality [12, 13]. Because the Lanczos method maintains biorthogonality implicitly through the three-term recurrence, it is natural to ask whether these recurrence vectors can be used in restarting. For the symmetric case, this idea has been explored in restarting the JD method with impressive results [27, 42, 41]. For the nonsymmetric case,

however, a short-term recurrence is not possible for the JD and Arnoldi methods. A Lanczos-based method seems to be the only alternative.

In this paper, we propose a biorthogonal Jacobi–Davidson method (biJD), which combines the Lanczos two-sided iteration with the solution of the correction equation (1.2) for both left and right Ritz pairs. The goal is twofold: provide a faster converging algorithm, and exploit an effective restarting scheme. First, we describe the biJD algorithm and review its various advantageous features. We also show how the multiplication of the adjoint with a vector can be performed with minimal computational cost, which is of more general interest. In the second part, we examine how the correction equation (1.2) can be set up, how fast it converges, and how to use preconditioning with it. In the third part, we adapt the recurrence-based restarting idea to biJD and show why it is expected to be beneficial. We conclude with numerical examples and a summarizing discussion.

2. The biorthogonal Jacobi–Davidson method (biJD). Throughout this paper, we assume that the matrix A is nonsymmetric, diagonalizable of order N, with eigenpairs $(\tilde{\lambda}_i, \tilde{x}_i)$, of which the one closest to a complex value σ is sought. The Davidson method first appeared as a diagonally preconditioned version of the Lanczos method for the symmetric eigenproblem. Extensions to more general preconditioners and to the nonsymmetric case have since been given [25, 10]. Morgan and Scott [26] proposed to solve approximately with some preconditioner the generalized Davidson correction equation: $(A - \sigma I)\ \delta = r = (A - \mu I)x$. In [38], Sleijpen and van der Vorst showed that for stability and robustness, as well as efficiency, the operator in the correction equation should have a range orthogonal to x, yielding (1.1). Several extensions have been proposed for the JD method, including general projections for (1.1), restarting schemes, and the use of harmonic Ritz vectors for interior eigenpairs [15, 37, 42]. The convergence analysis of JD using (1.2) is described in [37], but the formulation of a two-sided biJD method appears to be new. Algorithm 2.1 outlines the basic steps of biJD and introduces most of the notation in this paper.

We focus on finding only one eigenpair closest to σ, but the extension to finding more eigenpairs is straightforward. When an eigenpair converges, it can stay in the basis and never be targeted again or it can be locked out of the basis [33]. If a preconditioner $M \approx A$ is known, we can apply it to the BCG iteration, but the details are discussed later. Also, only one of the projections needs to be applied at every BCG iteration, and when λ is far from σ, we solve (1.2) instead of the equations in step 3 of biJD (see [15]). Because the three-term and the coupled, two-term recurrence versions of BCG span the same space, the more stable two-term recurrence is used in our implementation. When the basis reaches a maximum size of m, thick restarting is performed by keeping the k Ritz vectors closest to the shift σ. The above algorithm uses complex arithmetic, but as with JD, a real-only version is also possible [15]. Finally, a block biJD based on the Davidson–Liu block variant can be developed if more than one Ritz pairs are targeted at each step.

Computationally, biJD is more expensive per step than JD. First, it requires a matrix-vector multiplication with A^*, which may not be available in some applications. Second, in step 3, each iteration of BCG performs two matrix-vector multiplications, one with A and one with A^*. Similarly at step 6, the update of the auxiliary matrices K and L requires two such multiplications, twice the number required by JD. However, even JD would require two multiplications per inner iteration if a short recurrence method such as BCGSTAB or QMR were used. Finally, biJD requires twice as much storage as JD because of the left space W and its image $L = A^*W$. If needed, we can

ALGORITHM 2.1. BiJD.
Input: σ: *a complex value,* m: *maximum size allowed for the bases.*
k: *the number of vectors to be retained at restart.*
Initial right and left spaces: $V = [v_1, \ldots, v_{k_0}]$ *and* $W = [w_1, \ldots, w_{k_0}]$, *with* $W^*V = I$, *and* $\|v_i\| = 1,\ i = 1, \ldots, k_0$.
Output: Finds an approximate eigenpair (λ, x) *of* A, *with* λ *closest to* σ.

Compute initial $K = AV$, $L = A^*W$, *and* $H = W^*AV$. *Set* $j = k_0$.
Repeat
1. *Compute right* (g_i) *and left* (f_i) *eigenvectors of* H, *with* $\|g_i\| = 1, f_i^* g_i = 1,\ i = 1, \ldots, j$
2. *Target the Ritz triplet* (f, g, λ) *with Ritz value* λ *closest to* σ:
$x = Vg$ *with residual* $r_r = Kg - \lambda x$ *(right Ritz pair)*
$y = Wf$ *with residual* $r_l = Lf - \bar{\lambda} y$ *(left Ritz pair)*
If $(\|r_r\| < tolerance)$ **return**
3. *Run* p *steps of BCG simultaneously on the two correction equations:*
$(I - xy^*)(A - \lambda I)(I - xy^*)\ \delta_r = r_r$
$(I - yx^*)(A^* - \bar{\lambda} I)(I - yx^*)\ \delta_l = r_l$
4. *Set* $V = [V, \delta_r]$, *and* $W = [W, \delta_l]$
5. *Biorthogonalize the new basis vectors such that* $W^*V = I$, *and* $\|v_{j+1}\| = 1$
6. *Set* $j = j + 1$, *and compute* $K_j = AV_j$ *and* $L_j = A^*W_j$
7. *Compute the last column and row of the matrix* $H = W^*AV$

Until $(j == m)$
If $(\|r_r\| \geq tolerance)$ **then**
8. *Compute* $k < m$ *current Ritz vectors and restart:*
9. *Set* $V = [x_1, \ldots, x_k]$, $W = [y_1, \ldots, y_k]$, $K = AV$, $L = A^*W$ *and* $H = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$
10. *Set* $j = k$, *and goto step* 1.

endif

save the space of the arrays K and L by computing the residuals at step 2 by explicit matrix-vector multiplications.

2.1. Computational efficiency. From the above discussion biJD seems to incur about twice as many floating point operations per outer step than JD when the latter is using GMRES rather than BCGSTAB for the correction equation. However, in today's multiple memory hierarchy computers the role of memory accesses is more relevant than the flops. In this context, biJD can be performed with only a slight computational overhead over JD.

The fact that the matrix-vector multiplications with A and with A^* can be performed with only one access to the matrix A has not received any attention in the literature. Let us assume for simplicity a compressed sparse row (CSR) storage of the matrix A [35, 32]. For BCG-like methods two approaches are traditionally discussed [6]. The first performs the multiplication with A using the code in Figure 2.1(a) and then with A^* using the code in Figure 2.1(b). However, the CSR data structure increases memory traffic for the latter operation. The second approach explicitly transposes A into a CSR stored matrix A^*. Then, it applies matrix-vector multiplications for both matrices using the code in Figure 2.1(a). Besides the extra storage, two matrices are still read from memory at each operation.

(a)

```
%% A x, A in CSR
for i = 1,n
  t = 0
  for k=ia(i), ia(i+1)-1
    t = t + a(k)*x(ja(k))
  end
  y(i) = t
end
```

(b)

```
%% A^T u, A in CSR
for i=1,n
  w(i) = 0
end
for i = 1,n
  for k=ia(i), ia(i+1)-1
    w(ja(k)) = w(ja(k)) + u(i)*a(k)
  end
end
```

FIG. 2.1. *Traditional methods for matrix-vector multiplication when A is stored in CSR format.* (a) $y = Ax$, (b) $w = A^*u$.

(a)

```
%% Both Ax and A^Tu, A in CSR
for i=1,n
  w(i) = 0.0
end
for i = 1,n
  t = 0
  for k=ia(i), ia(i+1)-1
    ix = ja(k)
    w(ix) = w(ix) + u(i)*a(k)
    t = t + a(k)*x(ix)
  end
  y(i) = t
end
```

(b)

```
%% Ax and A^Tu, using temp vector
for i = 1,n
  Tmp(2*i-1) = x(i)
  Tmp(2*i)   = 0
end
for i = 1,n
  t = 0
  for k=ia(i), ia(i+1)-1
    ix = 2*ja(k)
    Tmp(ix) = Tmp(ix) + u(i)*a(k)
    t = t + a(k)*Tmp(ix-1)
  end
  y(i) = t
end
for i = 1,n
   w(i) = Tmp(2*i)
end
```

FIG. 2.2. *Proposed methods that perform the two matrix-vector multiplications simultaneously.* (a) $y = Ax$ *and* $w = A^*u$, (b) $y = Ax$ *and* $w = A^*u$ *but using a temporary vector.*

A first improvement would be to perform both Ax and A^*u while the same row of A has been brought in from memory. The code in Figure 2.2(a) shows how this is performed by simply merging the codes of Figure 2.1. Each sparse row of A is brought in once (both `a` and `ja`), and it is used to accumulate an inner product and to update various elements of $w = A^*u$. Depending on the cache size and the read/write channels available on the computer, this code can significantly reduce execution time.

Despite the locality of the array `a`, the vector `x` and the result vector `w` are accessed in a nonlocal but identical pattern. The idea of the code in Figure 2.2(b) is to create a temporary vector `Tmp` with the elements of `x` and `w` interleaved in it. With this scheme, the two nonlocal accesses to `x` and `w` become one nonlocal access to `Tmp` with the second access being in the adjacent memory location. If the number of nonzero elements in the matrix is large, and if the machine can perform both a read and a write on `Tmp` efficiently, this modification can provide further reduction in execution time. Note that the overhead from the initialization of the arrays can be hidden if these initializations are embedded in the calling BCG function. Finally, in contrast to BCGSTAB-like methods, the two matrix-vector multiplications can be

TABLE 2.1

Time in seconds to execute the two matrix-vector multiplications $y = Ax$ and $w = A^*u$, where A is a matrix of size N, and with nz nonzero elements per row randomly placed. The method numbers refer to the algorithms in Figures 2.1–2.2. *The machine is a SUN Ultra* 2300.

	Matrix size N / nonzero elements per row nz				
Method	50000/30	50000/10	200000/15	200000/5	400000/5
2.1(a-b)	0.28	0.10	0.63	0.25	0.52
2.1(a) with A^*	0.27	0.10	0.63	0.25	0.52
2.2(a)	0.39	0.06	0.40	0.16	0.34
2.2(b)	0.19	0.07	0.46	0.22	0.44

TABLE 2.2

Time in seconds to execute the two matrix-vector multiplications $y = Ax$ and $w = A^*u$, where A is a matrix of size N, and with nz nonzero elements per row randomly placed. The method numbers refer to the algorithms in Figures 2.1–2.2. *The machine is a* 1 *GHz Pentium* III.

	Matrix size N / nonzero elements per row nz				
Method	50000/150	50000/10	200000/30	400000/40	800000/20
2.1(a-b)	0.69	0.058	0.59	1.59	1.63
2.1(a) with A^*	0.71	0.057	0.61	1.61	1.64
2.2(a)	0.53	0.045	0.47	1.27	1.28
2.2(b)	0.56	0.055	0.52	1.34	1.44

performed in parallel. Tables 2.1 and 2.2 present some timing results using a g77 compiler on a SUN Ultra 2300 with 1 MB cache, and on a 1 GHz Pentium III with 256 KB cache, respectively. On both machines we see that the proposed algorithms consistently improve execution time by about 25%. We expect bigger improvements with advanced optimizing compilers.

Another computational requirement that seems to limit biJD applicability is the storage of K and L. In practice this turns out to be a minor problem for several reasons. First, the limiting factor is the expensive orthogonalization procedure, and for this reason the basis size is not allowed to grow very large. Second, with ever decreasing memory prices, storage for this limited basis size is not an issue, unlike the Lanczos process where hundreds or even thousands of vectors might be needed. Third, the availability of good preconditioners and in particular the advanced restarting techniques that we propose allow the basis size to shrink even further without significant convergence deterioration.

Finally, for computational efficiency on cache-based and parallel computers, we use an iterative Gram–Schmidt biorthogonalization. When there is no preconditioner and the number of BCG steps equals 1, the method reduces to a stable implementation of restarted nonsymmetric Lanczos [36].

2.2. Features of the biJD method. Besides the attractive properties of biJD for restarting and the correction equation, there is a host of features that enhance the overall performance and robustness of the algorithm.

2.2.1. Benefits from the left/right bases. The intrinsic advantage of biJD is its ability to obtain the left eigenvectors almost for free, and with accuracy similar to that of obtaining the right ones. Left eigenvectors can be extremely useful, even if they are not specifically needed by the application. First, they can be used in the spectral projector to deflate converged eigenpairs. Second, left and right Ritz

pairs provide an estimate to the condition number of the required eigenvalue, which is a measure of how reliably this eigenpair has been computed. Third, even before convergence is achieved, detecting an ill-conditioned eigenvalue might help speed up the correction equation by approximately removing this ill-conditioning through a similarity transformation.

Another significant advantage of biJD is that the Ritz values are the generalized Rayleigh quotients (GRQs): $\lambda = y^*Ax/(y^*x)$. If the eigenvalue $\tilde{\lambda}$ is not too ill-conditioned, the GRQ is known to be more accurate than the Rayleigh quotient (RQ): $\mu = x^*Ax/(x^*x)$. In fact, this is true even when the left eigenvectors are known to a lesser accuracy than the right ones. As explained in [45] (see also [4]), let x, y be approximations to the right and left eigenvectors, with $x = \tilde{x} + \epsilon_x$ and $y = \tilde{y} + \epsilon_y$. If we assume for simplicity that $\epsilon_x \perp \tilde{y}$ and $\epsilon_y \perp \tilde{x}$, then

$$\frac{y^*Ax}{y^*x} = \tilde{\lambda} + \frac{\epsilon_y^*(A-\tilde{\lambda})\epsilon_x}{\tilde{y}^*\tilde{x} + \epsilon_y^*\epsilon_x} \qquad \Rightarrow$$

$$|\lambda - \tilde{\lambda}| \leq \left(\|A\| + |\tilde{\lambda}|\right) \frac{\|\epsilon_y\|\|\epsilon_x\|}{|\tilde{y}^*\tilde{x}| - \|\epsilon_x\|\|\epsilon_y\|}, \tag{2.1}$$

which implies that the error in the Ritz value is $\mathcal{O}(\|\epsilon_y\|\|\epsilon_x\|)$, provided that the eigenvalue is not too ill-conditioned. Interestingly, if we substitute $x = \tilde{x} + \epsilon_x$ with $\epsilon_x \perp x$ in the RQ, the term $\tilde{x}^*A\epsilon_x$ is not zero unless the matrix is normal. Thus, assuming $\|\tilde{x}\| = 1$, the RQ is given by

$$\frac{x^*Ax}{x^*x} = \tilde{\lambda} + \frac{\epsilon_x^*(A-\tilde{\lambda})\epsilon_x + \tilde{x}^*A\epsilon_x}{\|\tilde{x}\|^2 + \|\epsilon_x\|^2} \qquad \Rightarrow$$

$$|\mu - \tilde{\lambda}| \leq \left(\|A\| + |\tilde{\lambda}|\right) \frac{\|\epsilon_x\|^2}{1 + \|\epsilon_x\|^2} + \frac{\|A\|\|\epsilon_x\|}{1 + \|\epsilon_x\|^2}. \tag{2.2}$$

Thus, the error in the RQ is $\mathcal{O}(\|\epsilon_x\|)$ in general and $\mathcal{O}(\|\epsilon_x\|^2)$ in the normal case.

Note that both GRQs and RQs can be close to the required eigenvalue even though the vectors in those quotients are linear combinations of unrelated eigenvectors. In such cases, however, the GRQ and RQ differ substantially. Because the RQ can be computed inexpensively in biJD, contrasting it to the GRQ provides an excellent means of assessing eigenvalue convergence.

For nonsymmetric matrices, neither the Galerkin nor the Petrov–Galerkin projection methods provide any useful optimality for the Ritz pairs [33]. It has been observed that sometimes approximations are extracted faster and more accurately from the Lanczos process than from an orthogonal projection method (like Arnoldi), but also the contrary is often true. The biJD method inherits these characteristics, which for some problems may prove advantageous over JD. However, differences solely caused by the Petrov–Galerkin are expected to be minor because of the use of preconditioning and restarting.

2.2.2. Flexibility of the biJD algorithm. The biJD algorithm uses explicit biorthogonalization, thus avoiding problems from the loss of orthogonality of nonsymmetric Lanczos. More interestingly, it also retains the flexibility of JD. For example, it can accommodate a variety of restarting techniques because it does not have to maintain a tridiagonal projection matrix. In our description of the algorithm, we have used thick restarting where the k Ritz pairs closest to σ are retained. The left and right spaces facilitate an elegant extension of JD thick restarting to biJD, since

left and right Ritz vectors are biorthogonal by construction, and $H = (y_i^* A x_j)_{i,j}$ is the diagonal of the corresponding Ritz values. Implicit restarting with user-defined shifts can also be applied in a way similar to the implicitly restarted nonsymmetric Lanczos method [36], but the benefits in the absence of a Krylov space are not clear.

The biJD method can also restart with any arbitrary vectors $Vc \in V$ and $Wz \in W$. Biorthogonality can be maintained inexpensively in the coefficient space by biorthogonalizing the vectors c and z instead, and H can be updated by inner products of the coefficient restarting vectors (see [42]). This flexibility is used in the restarting scheme proposed in a later section, and it is also useful with harmonic Ritz vectors.

When looking for interior eigenpairs, harmonic Ritz vectors often provide better approximations and may result in a more effective correction equation [7, 28, 15, 41]. The main idea is to perform a Petrov–Galerkin on the matrix $(A - \sigma I)^{-1}$, for which the required eigenpairs lie on the extreme of its spectrum. The inversion of the matrix is avoided if the space $(A - \sigma)V$ is used instead in the projection. To compute the harmonic pairs for biJD, we proceed similarly to JD, with the exception that we modify both left and right projection spaces:

$$W_h = (A - \sigma I)^* W = A^* W - \bar{\sigma} W = L - \bar{\sigma} W,$$
$$V_h = (A - \sigma I) V = AV - \sigma V = K - \sigma V.$$

The W_h and V_h can be computed without matrix-vector multiplications. Moreover,

$$W_h^* V_h = W^* (A - \sigma I)^2 V,$$

and we can formulate the Petrov–Galerkin projection with W_h and V_h solving for g_h:

$$W_h^* (A - \sigma I)^{-1} V_h g_h = \frac{1}{\nu} W_h^* V_h g_h \qquad \Leftrightarrow$$
$$W^* (A - \sigma I) V g_h = \frac{1}{\nu} W^* (A - \sigma I)^2 V g_h. \tag{2.3}$$

Equation (2.3) is similar to the one for the JD iteration, and it involves computations with only V, W, K, and L. As with JD, to obtain the harmonic Ritz vectors we apply implicitly one step of inverse iteration to $V_h g_h = (A - \sigma I) V g_h$, yielding vector $V g_h$.

Finally, biJD can incorporate into its bases any arbitrary vector in $\mathcal{C}^N$ that carries useful information. Both a left and a right vector would be needed, so the user must guarantee that they are not orthogonal. The vectors are appended in the bases, biorthogonalized, and the algorithm resumes. Besides allowing for external information to be used, this feature allows for the flexible preconditioning required in the biJD/JD methods, but more importantly it provides a straightforward way of dealing with breakdown.

2.2.3. Resolving breakdown. Breakdown can occur in biJD whenever the two vectors added in the left and right spaces are orthogonal. For the nonsymmetric Lanczos method, (near) breakdown is usually remedied through look-ahead schemes [30, 17, 18, 5], although an incurable breakdown is also possible [29]. A simple alternative is to restart the Lanczos method with a slightly modified residual vector [35], or to perform an implicit restarting with "nonexact" shifts (i.e., non-Ritz values) [36]. Note that if we use exact shifts or, equivalently, if we thick restart with the current Ritz vectors, the breakdown will recur immediately after restarting [36]. These techniques for avoiding (near) breakdown situations are also readily applicable to biJD, as no special structure is required in the projection matrix.

In addition, the ability to include arbitrary vectors at step 4 of the biJD algorithm offers a much simpler solution to the problem without the need to restart the iteration. If a (near) breakdown is detected at step 5 of the algorithm, we can insert a small random perturbation of the vectors δ_r or δ_l. It is more reasonable to change only one of the vectors, e.g., the left one if we are interested in the right eigenpair. This is usually enough to overcome the breakdown but still retains the basic direction of δ_l.

Breakdown is also possible during the BCG iteration, but it is handled easily by early termination of BCG, which does not need to run to convergence. Specifically, there are two possible BCG breakdowns; first when the left and right BCG residuals are orthogonal, and second when the LU decomposition cannot be carried out. If the first breakdown occurs in the first step of BCG, the original biJD residuals are orthogonal and the situation is treated as a biJD breakdown. If the LU breakdown occurs in the first step of BCG, we simply add the residuals in the bases and resume biJD. If any of the two breakdowns occurs during the ith BCG iteration, we terminate BCG and return to biJD the approximate solutions from iteration $i-1$. These are not orthogonal, because the ith iteration is the first time that breakdown occurs, and thus the biJD algorithm can resume.

3. The biJD correction equation. There is a multitude of choices for projectors in the correction equation of JD. A general framework that describes the use and convergence properties of arbitrary projectors for JD has been given in [37, 15], and some recent developments on preconditioning can be found in [39, 19, 20].

Despite the variety of possible correction equations, the choices for biJD are limited because the operators of the left and right correction equations have to be adjoint to each other for BCG to apply. We show next that from the two natural choices, the orthogonal projector $(I - xx^*)$ and the spectral one $(I - xy^*)$, only the spectral projector solves a meaningful correction equation for the left eigenvector and thus has better convergence properties. In addition, the biorthogonal bases maintained by biJD provide an elegant framework for using the spectral projector.

3.1. Forming the appropriate equation. Let x be an approximation to an eigenvector of A, say $\tilde{x}$ with eigenvalue $\tilde{\lambda}$. We are interested in solving for the correction δ that satisfies $\tilde{x} = x + \delta$. Because of the scale invariance of $\tilde{x}$, we can look for δ in a space orthogonal to x (original JD) or orthogonal to some other vector $p_2^*\delta = 0$ [37]. Assuming that $p_2^*x \neq 0$, we consider the projector $(I - p_1p_2^*)$, with $p_2^*p_1 = 1$, that can represent both operators in (1.1) and (1.2):

$$B = (I - p_1p_2^*)(A - \tilde{\lambda}I)(I - p_1p_2^*). \tag{3.1}$$

Starting from the eigenvalue equation for the required eigenpair and following the same algebraic manipulations as in [38], we obtain the correction equation

$$B\delta = -(A - \tilde{\lambda}I)x - p_1p_2^*(A - \tilde{\lambda}I)\delta. \tag{3.2}$$

Because $B\delta$ is orthogonal to p_2, the same applies for the right-hand side, which yields the condition $\tilde{\lambda} = \rho + \epsilon$, with

$$\rho = \frac{p_2^*Ax}{p_2^*x} \quad \text{and} \quad \epsilon = \frac{p_2^*(A - \tilde{\lambda}I)\delta}{p_2^*x}. \tag{3.3}$$

Substituting ρ and ϵ into (3.2) we obtain

$$B\delta = (\rho x - Ax) + \epsilon(x - p_1(p_2^*x)). \tag{3.4}$$

To be able to form and solve this correction equation, the right-hand side should not include any unknowns. Because ϵ is not known, this term has to vanish, which is possible in general only if p_1 and x are colinear. Let $p_1 = x/\|x\|$.

In biJD, the left equation solved by BCG should involve the adjoint operator $B^* = (I - p_2 x^*)(A - \tilde{\lambda})^*(I - p_2 x^*)$. We want to identify a p_2 so that we can form an appropriate right-hand side for the correction equation for an approximation y to the left eigenvector $\tilde{y}$ of A. According to the above analysis, p_2 must be colinear with y. Thus, the projector must be of the form $(I - xy^*)$. If the orthogonal projector $I - xx^*$ is chosen instead, the right correction equation has a proper right-hand side, but the same does not hold for the left one.

The above does not incapacitate a biJD method that uses BCG on this inappropriate left correction equation. It implies only that the convergence of the left eigenvector will not be as fast, which may not be relevant if we are interested only in the right eigenvector. However, as we show next, the asymptotic convergence of biJD with (1.1) is inferior to the use of (1.2).

3.2. Asymptotic convergence. When we apply two coupled inverse iterations for finding both left and right eigenpairs using the generalized Rayleigh quotient

$$(A - \lambda_s I)x_{s+1} = x_s, \qquad (A - \lambda_s I)^* y_{s+1} = y_s,$$
$$\text{with } \lambda_s = y_s^* A x_s / y_s^* x_s,$$

convergence is known to be ultimately cubic [45]. A large condition number of the sought eigenvalue, $\kappa(\tilde{\lambda}) = \|\tilde{y}\|\|\tilde{x}\|/\tilde{y}^*\tilde{x}$, would only delay the cubic convergence phase.

The situation is very similar in the coupled solution of the left and right correction equations with BCG. In [37, Theorem 3.4, Remark 3.5], Sleijpen et al. prove that if y converges to the left eigenvector, a stationary iteration that corrects an approximation x to the right eigenvector with the solution of (1.2) has locally cubic convergence to the right eigenpair. The JD method accelerates the stationary method by performing Galerkin over the basis of all x iterates, providing also global convergence.

Exactly the same result holds for the biJD method, because it only applies a different acceleration method to the correction equation (1.2). The difference is that biJD treats left and right eigenvectors symmetrically, with y converging to the left eigenvector with speed similar to that of x, thus guaranteeing the local cubic convergence. The same is not true in general for the JD method, since the right space may never contain sufficient components of the left eigenvector.

As with classical JD, biJD converges quadratically [37, 46] if (1.1) is solved accurately for the right eigenpair, regardless of any left equation used. Even if an appropriate equation were solved for the left pair, since (1.1) does not use any y information, convergence to the right pair would still be quadratic. Therefore, we expect faster biJD convergence with the correction equation (1.2), even when the equations are not solved accurately.

Note that the conditioning of the operator $(I - xy^*)(A - \lambda I)(I - xy^*)$ depends on $\|y\|$ or, equivalently, on the angle between x and y. At the limit, the operator is equivalent to a deflated matrix, and thus $\|A - \tilde{\lambda}\ \tilde{x}\tilde{y}^*/\tilde{x}^*\tilde{y}\| \leq \|A\| + |\tilde{\lambda}|\ \kappa(\tilde{\lambda})$. On the other hand, the operator $(I - xx^*)(A - \lambda I)(I - xx^*)$ does not increase the norm of the matrix. This suggests that, for stability reasons, the biJD method could switch to the orthogonally projected equation if an ill-conditioned Ritz pair is detected. However, the correction equations are never solved accurately, and moreover in all our experiments we have observed that the ill-conditioning stems only from an increase in the largest singular value, while the rest are not affected. Also, at the limit, the

spectral projector preserves both left and right eigenvectors, while the orthogonal projector preserves only the left ones.

3.3. Using preconditioning in biJD. Performing preconditioning for the correction equations of JD and especially of biJD is involved, because we must approximate the inverse of a projected matrix. In the common case of $M \approx A - \sigma I$, the appropriate application of this preconditioner would be to invert the operator $(I - xy^*)M(I - xy^*)$, which is not practical and often not even feasible.

For JD with correction equation (1.1), Sleijpen and van der Vorst [38] described a way to apply such a preconditioner implicitly by solving systems with M^{-1} and by applying a few additional orthogonalizations. In [37] they extended this scheme to arbitrary projections for the correction equation. Considering the correction equation (1.2) for the right eigenpair and a preconditioner $M \approx A - \sigma I$, Theorem 7.3 in [37] states that the appropriate preconditioned correction equation can be written as

$$\left(I - \frac{M^{-1}xy^*}{y^*M^{-1}x}\right) M^{-1}(A - \sigma I) \left(I - \frac{M^{-1}xy^*}{y^*M^{-1}x}\right) \delta_r = -\left(I - \frac{M^{-1}xy^*}{y^*M^{-1}x}\right) M^{-1}r_r. \tag{3.5}$$

Let us consider the correction equation for the left eigenpair, with operator the adjoint of (1.2), and $M^* \approx (A - \lambda I)^*$. For BCG to solve both systems simultaneously, the appropriate correction equation for the left eigenpair must be the adjoint of (3.5). Therefore, we need to apply right instead of left preconditioning. If we let $\delta_l = M^{-*}t$, then a version of the above theorem for right preconditioning [16] states

$$\left(I - \frac{yx^*M^{-*}}{x^*M^{-*}y}\right) (A - \sigma I)^* M^{-*} \left(I - \frac{yx^*M^{-*}}{x^*M^{-*}y}\right) t = -\left(I - \frac{yx^*M^{-*}}{x^*M^{-*}y}\right) r_l. \tag{3.6}$$

Because $x^*M^{-*}t = x^*M^{-*}M^*\delta_l = x^*\delta_l$, the orthogonality condition $\delta_l \perp x$ is equivalent to the orthogonality condition $t \perp M^{-1}x$ in (3.6). Note also that $P_M = (I - yx^*M^{-*}/x^*M^{-*}y)$ is a projector with $P_M y = 0$ and $P_M t = t$. Thus, (3.6) is a correction equation for the left eigenpair.

4. Efficient restarting for biJD. The idea of thick restarting is based on the observation that as Krylov methods approximate extreme eigenvectors, these vectors become gradually deflated from the iteration, and the method converges faster. The goal of thick restarting is to retain those Ritz vectors that the method tends to approximate better, so that they can be improved and thus cause the superlinear convergence (for linear systems, see [9, 24, 2]). The Ritz vectors with Ritz values closest to the required eigenvalue are thus a natural choice. The dynamic thick restarting scheme retains also Ritz vectors with Ritz values in the extreme part of the spectrum, since Krylov methods approximate these vectors better [43]. In the symmetric case, the results of the dynamic scheme have been impressive [43, 41]. In the nonsymmetric case, it still performs well, but the improvements are not as dramatic and are more matrix dependent. However, thick and especially dynamic thick restarting increase the iteration costs because they retain a large number of vectors at restart.

A different class of restarting strategies is based on the observation that all Krylov methods enforce some kind of orthogonality in order to guarantee new directions in the basis [13, 12]. With restarting, some directions are discarded, and the loss of full orthogonality causes the convergence to deteriorate. This behavior is common not only in explicitly restarted methods such as Arnoldi and GMRES, but also in methods based on short recurrences, such as CG. The much researched loss of orthogonality in CG is known to cause slower convergence. Restarting strategies in this class attempt

to identify and retain those directions that the algorithm tends to repeat. A typical and effective example is the truncation strategy of de Sturler [12].

Interestingly, the two restarting classes often overlap. In the symmetric case, selective orthogonalization against converged eigenvectors can be viewed as both a deflation- and an orthogonality-based method. In the nonsymmetric case, eigenvector deflation can also be viewed as a special case of orthogonality conditions, but the problem is more complicated and other directions become important.

The above suggests that maintaining orthogonality against all visited directions is a critical issue in restarted iterative methods. The three-term recurrence of the symmetric Lanczos method achieves full orthogonality implicitly, so it is natural to seek ways to use this recurrence to restart efficiently the symmetric JD method.

4.1. Restarting idea for symmetric JD. Even though explicit full orthogonalization is avoided in the Lanczos algorithm through the three-term recurrence, the basis vectors still need to be stored for computing the Ritz vector. However, if the exact eigenvalue is known, the eigenvector can be obtained by the CG method storing only three vectors [44, 21]. If the eigenvalue is not known but converges rapidly, methods based on CG can still be used [44].

A more useful variant of this idea was proposed in [27] and extended and analyzed in [42]. It is based on the observation that, in the absence of preconditioning, the space built by CG for solving the correction equation differs from the Krylov space of the Lanczos method (JD with no correction step) only in the starting vector. In addition, if the Ritz value at step k were known, the two methods would yield exactly the same vector at the kth step. Note that CG minimizes the A-norm of the error on a three-vector space, which is close to the space spanned by $\{x^{(k-1)}, x^{(k)}, r\}$, where $x^{(k-1)}, x^{(k)}$ are successive Ritz vectors from JD iterations $k-1$ and k, respectively, and r is the residual of $x^{(k)}$.

We have argued that if the JD method is restarted at the kth iteration, it is beneficial to keep the Ritz vector from the previous iteration ($x^{(k-1)}$) along with the current one. In fact, if these three-vector spaces from CG and JD were identical, there would be no information loss by this restarted JD variant. In general, the two spaces are not the same but close if the Ritz value does not vary significantly between steps.

This technique works extremely well for extreme eigenpairs and still performs well for interior eigenpairs because it retains some orthogonality memory. Combining this scheme with thick restarting provides in addition a deflation-like character, and it is the only technique that has managed to improve on the dynamic thick restarting.

4.2. Extending to nonsymmetric matrices. Extending the above restarting technique to the nonsymmetric JD is not possible because orthogonality must be maintained explicitly. However, if we trade the more stable orthogonality for biorthogonality, the nonsymmetric Lanczos method fits the description. This is the second motivation—besides the faster outer convergence—for proposing the biJD algorithm.

The changes in Algorithm 2.1 are analogous to the symmetric case. We denote new steps by decimal numbers to show between which biJD steps they are inserted.

ALGORITHM 4.1. ADDITIONS TO BIJD FOR NEW RESTARTING.
1.1 $x_{prev} = x,\ y_{prev} = y$
9.1 *Biorthogonalize* (x_{prev}, y_{prev}) *against* (V, W)
9.2 *set* $V = [V, x_{prev}]$, $W = [W, y_{prev}]$, *and* $H_{k+1,k+1} = y_{prev}^* A x_{prev}$.

Note that the restarting applies symmetrically to both left and right spaces. For clarity, the algorithm above presents the restarting scheme in terms of the long vectors x_{prev} and y_{prev}. In practice, all of the above operations can be performed without extra matrix-vector multiplications or long vector biorthogonalizations. Because $x = Vg$ and $x_{prev} = V\,[c_1, \ldots, c_{m-1} 0]^T$ for some coefficient vector $c \in \mathcal{C}^{m-1}$, biorthogonalizations can be performed in the coefficient space. In addition, the updates of matrices $K \leftarrow Kg_i$ and $L \leftarrow Lf_i$ for $i = 1, \ldots, k+1$ during restarting can be computed using the coefficient vector c. Finally, in the above algorithm, the new restarting scheme is coupled with thick restarting by adding in the restarted basis both the previous Ritz vector and the k current ones. The rationale is analogous to the symmetric case.

4.2.1. Extending the theory. In this section we extend to the nonsymmetric case the theory developed in [42]. The goal is to explain why restarting based on the three-term recurrence yields future Ritz vectors that are close to the Ritz vectors we would have obtained without restarting.

To facilitate presentation clarity, we use a single subscript that denotes the iteration number for any variable, e.g., x_i is the Ritz vector at the ith iteration. We assume that the matrix A is diagonalizable, with no multiple eigenvalues.

LEMMA 4.1. *Let x, y be vectors of $\mathcal{C}^N$ such that $y^*x = 1$ and $\lambda = y^*Ax$. Let $\pi = (I - xy^*)$ denote the oblique projector onto $y^\perp$, and let the residual of x be denoted by $r = (A - \lambda I)x = \pi r$. Then, for every $k > 1$,*

$$span(\{x, Ax, \ldots, A^k x\}) = span(\{x, r, (\pi A\pi)r, \ldots, (\pi A\pi)^{k-1} r\}).$$

Proof. Denote by $\mathcal{K}_k$ and $\mathcal{L}_k$ the spaces of the left- and right-hand sides, respectively. Obviously, for $k = 1$, $\mathcal{K}_1 = \mathcal{L}_1$. We assume that $\mathcal{K}_i = \mathcal{L}_i$ for all $i < k$. Let $q \in \mathcal{L}_k$. There is $u \in \mathcal{L}_{k-1} = \mathcal{K}_{k-1}$ and $\alpha \in \mathcal{C}$ such that

$$q = u + \alpha(\pi A\pi)^{k-1} r = u + \pi A\pi\ z,$$

where $z = \alpha(\pi A\pi)^{k-2} r \in \mathcal{L}_{k-1} = \mathcal{K}_{k-1}$. Since $\pi = I - xy^*$ and $Az \in \mathcal{K}_k$, we have

$$\begin{aligned} q &= u + (I - xy^*)A(z - (y^*z)x) \\ &= u + Az - (y^*z)Ax - (y^*Az)x + (y^*Ax)(y^*z)x \in \mathcal{K}_k. \end{aligned}$$

Thus, $\mathcal{L}_k \subseteq \mathcal{K}_k$. If $\mathcal{L}_k$ is of full dimension, its dimension is $k+1$, the same as $\mathcal{K}_k$, and thus the two spaces must be equal. If $\mathcal{L}_k$ is not of full dimension, then it forms a smaller invariant subspace of dimension $i < k+1$, which is also included in $\mathcal{K}_k$. Then, from the inductive hypothesis, $\mathcal{L}_k = \mathcal{L}_{i-1} = \mathcal{K}_{i-1} = \mathcal{K}_k$. ∎

The lemma says that the right (left) Krylov space built by Lanczos in k steps is the same as the right (left) Krylov space that BCG builds in k steps when solving the correction equation (1.2), appended with the initial vector x (y). The use of the spectral projector is important for the left equation. If $(I - xx^*)$ were used instead, it would introduce a multiple of x term, which does not belong in the Krylov space.

THEOREM 4.2. *Let $x_0, y_0 \in \mathcal{C}^N$, with $\|x_0\| = 1, y_0^* x_0 = 1, \lambda_0 = y_0^* A x_0$, and $\sigma \in \mathcal{C}$.*

Let (x_k, y_k, λ_k) be the right and left Ritz vectors and their GRQ after k steps of the biJD method with no correction equation (Lanczos), with (x_0, y_0) as right and left starting vectors.

Let $z_k = x_0 + \delta_r$ and $w_k = y_0 + \delta_l$ be the approximate right and left eigenvectors, where δ_r and δ_l are the right and left corrections obtained by applying k steps of the BCG method to (1.2) *with shift σ. Then*

$$z_k = x_k \quad \text{and} \quad w_k = y_k \Leftrightarrow \sigma = \lambda_k.$$

Proof. Because there is no correction equation being solved, biJD builds right and left Krylov spaces:

$$span(\{x_0, Ax_0, \ldots, A^k x_0\}) \quad \text{and} \quad span(\{y_0, A^* y_0, \ldots, A^{*k} y_0\}).$$

Let $\pi = I - xy^*$, and denote $r_r = \alpha(A - \lambda_0)x_0$ and $r_l = \alpha(A - \lambda_0)^* y_0$, where $\alpha \in \mathcal{C}$ is chosen such that $r_l^* r_r = 1$. Note that $\pi r_r = r_r$ and $\pi^* r_l = r_l$. The BCG method on (1.2), starting with zero right and left initial guesses, builds the spaces

$$span(\{r_r, (\pi A\pi)r_r, \ldots, (\pi A\pi)^{k-1} r_r\}) \text{ and } span(\{r_l, (\pi A\pi)^* r_l, \ldots, (\pi A\pi)^{*(k-1)} r_l\}).$$

By construction, the Lanczos biorthogonal bases for the biJD spaces have (x_0, r_r) and (y_0, r_l) as their first two vectors. Therefore, if we consider bases $\{x_0, X\}$ and $\{y_0, Y\}$ for these two subspaces, with $Y^*X = I$, then X and Y are also bases of the spaces generated by BCG. In the following, we focus only on the right Ritz pair, because the arguments for the left one are identical.

With the above bases, and normalizing the Ritz vector x_k so that its coefficient of x_0 is one, the Petrov–Galerkin projection at the kth step of biJD solves the following problem (note the matrix is tridiagonal):

$$\begin{bmatrix} \lambda_0 & y_0^* AX \\ Y^* Ax_0 & Y^* AX \end{bmatrix} \begin{pmatrix} 1 \\ c_k \end{pmatrix} = \lambda_k \begin{pmatrix} 1 \\ c_k \end{pmatrix} \tag{4.1}$$

or, equivalently, the following system, which has $k+1$ Ritz pairs as solutions. We fix the equations for a specific (λ_k, c_k), $c_k \in \mathcal{C}^k$, so that $x_k = x_0 + Xc_k$:

$$\lambda_0 + y_0^* AXc_k = \lambda_k, \tag{4.2}$$

$$Y^* Ax_0 + Y^* AXc_k = \lambda_k c_k. \tag{4.3}$$

Consider the bases X, Y for the Petrov–Galerkin condition of the BCG method. BCG computes a correction to x_0 and sets $z_k = x_0 + Xc'_k$. Because $\pi X = X$ and $Y^*\pi^* = Y^*$, the projected problem solved is

$$Y^*(A - \sigma I)Xc'_k = Y^*(\lambda_0 I - A)x_0 = -Y^* Ax_0. \tag{4.4}$$

From (4.3) and (4.4) we obtain

$$(Y^* AX - \lambda_k)c_k = (Y^* AX - \sigma)c'_k. \tag{4.5}$$

If $c_k = c'_k$ (and thus $x_k = z_k$), then obviously $\sigma = \lambda_k$. Conversely, if $\sigma = \lambda_k$, we have

$$(Y^* AX - \lambda_k)(c_k - c'_k) = 0, \tag{4.6}$$

which implies that $c_k = c'_k$, and thus $x_k = z_k$. Note that $(\lambda_k, c_k - c'_k)$ could not be an eigenpair of $Y^* AX$, because the $k \times k$ matrix in (4.1) is tridiagonal, irreducible (since k steps of biJD/BCG can be carried out), and already has λ_k as an eigenvalue.

The proof for the left eigenvectors is identical to the above. ∎

The following diagram depicts the iterates of Lanczos (biJD with no correction equation) and of BCG with the Ritz value λ_k as shift. If started with the same vector x_0, they end in the same Ritz vector x_k at the kth step. Intermediate vectors differ in general. In this case, we can recover the information lost in restarting Lanczos.

$$\begin{array}{ccccccccccc} & \nearrow & x_1 & x_2 & x_3 & \ldots & x_{k-2} & x_{k-1} & \searrow & & \\ x_0 & & & & & & & & & & x_k \\ & \searrow & z_1 & z_2 & z_3 & \ldots & z_{k-2} & z_{k-1} & \nearrow & & \end{array}$$

However, λ_k is usually computed during the Lanczos (biJD) procedure. What is pertinent to our restarting scheme is whether the three-term BCG recurrence still produces an accurate approximation to the Ritz vector, if an inexact eigenvalue shift is used in (1.2). The following lemma quantifies the distance of the vectors x_k and z_k when $\sigma \neq \lambda_k$.

LEMMA 4.3. *Let the assumptions and notations of Theorem* 4.2 *hold, and denote by* s *the smallest singular value of the matrix* $Y^*AX - \lambda_k I$. *Then*

$$\|z_k - x_k\| \leq |\sigma - \lambda_k| \frac{\|X\| \|c'_k\|}{s}.$$

Proof. Let $S = Y^*AX$ in (4.5), and since λ_k is not an eigenvalue of S,

$$c_k = (S - \lambda_k I)^{-1}(S - \sigma I)c'_k.$$

From the definitions of x_k, z_k and from the above equation we have

$$\begin{aligned}\|z_k - x_k\| &= \|X(c'_k - c_k)\| = \|X(I - (S - \lambda_k I)^{-1}(S - \sigma I))c'_k\| \\ &= \|(\sigma - \lambda_k)X(S - \lambda_k I)^{-1}c'_k\| \leq |\sigma - \lambda_k| \frac{\|X\| \|c'_k\|}{s}. \qquad \square\end{aligned}$$

COROLLARY 4.4. *If in addition to the assumptions of Lemma* 4.3, $\|z_k\|$ *is larger than the correction term, i.e.,* $\|z_k\| > \|Xc'_k\|$, *then we have a relative bound involving the condition number of the basis* X*:*

$$\frac{\|z_k - x_k\|}{\|z_k\|} \leq |\sigma - \lambda_k| \frac{\kappa(X)}{s}.$$

Proof. We have $\|z_k\| \geq \|Xc'_k\| = \sqrt{c'^*_k X^* X c_k} \geq \sigma_{min}(X)\|c'_k\|$. Dividing both sides of the bound in Lemma 4.3 yields the result. $\square$

These bounds imply that when the Ritz value is almost constant, which usually occurs near convergence or when convergence is slow, BCG computes a close approximation to the Ritz vector of biJD. In the context of restarting, assume that we need to compute the (λ_{k+1}, x_{k+1}) Ritz pair and that biJD (no correction equation) is restarted after $k-1$ steps, retaining only the Ritz pair (λ_{k-1}, x_{k-1}). After restarting, biJD generates the Ritz pair (λ_k, x_k), but after a second iteration the new Ritz pair differs from (λ_{k+1}, x_{k+1}). Consider a hypothetical BCG recurrence that uses the unknown λ_{k+1} to produce the wanted Ritz pair in $k+1$ steps. If we apply Corollary 4.4 on the vectors z_i of BCG, but consider instead x_{k-1} and x_k as the end points, we get two inequalities:

$$\begin{aligned}\|z_{k-1} - x_{k-1}\|/\|z_{k-1}\| &\leq \mathcal{O}(|\lambda_{k+1} - \lambda_{k-1}|), \\ \|z_k - x_k\|/\|z_k\| &\leq \mathcal{O}(|\lambda_{k+1} - \lambda_k|).\end{aligned}$$

When the Ritz value is almost constant between steps, the Ritz vectors x_{k-1} and x_k approximate the BCG iterates for the still uncomputed $k+1$ step. Because x_{k+1} is a linear combination of the unknown z_k, z_{k-1}, a good restarting basis for biJD is one consisting of both Ritz vectors $\{x_{k-1}, x_k\}$.

However, proximity may not be as good as in the symmetric case [42]. As expected, the bounds include both the condition number of the basis matrix X and the smallest singular value of S, which incorporates information on the conditioning of the eigenvectors of S and the distance of other eigenvalues from λ_k. In case of highly

ill-conditioned bases or eigenvalues, the effects of the restarting scheme seem arbitrary, although in such cases the problem should be traced rather in the near ill-posedness of the eigenproblem. Finally, eigenvalue convergence in the nonsymmetric case is not monotonic and sometimes is even irregular, which complicates the runtime interpretation of the bounds to decide whether the restarting scheme should be applied. Yet, if we know when to apply it, the new restarting scheme works very well on a variety of matrices, as shown in the experiments in the following section.

5. Numerical experiments. We have implemented the above algorithms in Matlab and conducted an extensive set of tests on nonsymmetric matrices from the collection in [3] and from the Matrix Market [8]. In our experiments, we look for the right eigenpair that is of interest in the application domain of the matrix. We iterate until the residual norm reduces by 10^{-8}, and we plot residual convergence versus the number of outer iterations. Experiments are run on a SUN Ultra 2300 and on a Pentium III. In the figure notation, `JD` is the JD method, `biJD(I-xy')` (or simply `biJD`) and `biJD(I-xx')` is biJD with correction equation (1.2) and (1.1), respectively. `biJD+1` denotes biJD whose basis is augmented by the previous Ritz vector at restart.

5.1. biJD vs. JD without restarting. In the first set of experiments, biJD and JD each apply 10 steps of BCG or GMRES, respectively, to its correction equation. There is no restarting, and no preconditioner or harmonic eigenpairs are used for the correction equation.

The experiments suggest three general observations that agree with the theory discussed in this paper. First, although 10 steps on the correction equation are not enough for biJD and JD to demonstrate cubic or quadratic convergence, respectively, the convergence of biJD is usually faster asymptotically. The *semiquadratic* convergence of nonsymmetric Lanczos also contributes to this [4]. Second, the superiority of the Petrov–Galerkin method over the Galerkin process is problem dependent. Third, the projection $I - xx^*$ in the biJD correction equation does not usually help convergence.

In Figure 5.1, the left graph shows the convergence for the pde225 matrix. We look for the eigenvalue with the largest real part. In this case, biJD has better global convergence than JD, suggesting that the Petrov–Galerkin may be finding the correct components early in the iteration. Note also that there is practically no difference between the two ways of projecting the correction equation. The right graph in Figure 5.1 shows the convergence for the Tolosa 340 matrix. The goal is to compute an eigenvalue with largest imaginary part. The observations are the same as with matrix pde225, except that the gap between JD and biJD is even larger.

The results in Figure 5.2 show that, as with Lanczos versus Arnoldi, there are also cases where JD performs better than biJD. The left graph involves matrix west0479 from Harwell-Boeing, and we seek the interior eigenpair closest to (-17.825 - 4.6376 i). The `biJD(I-xx')` method is not shown as it converged to the wrong eigenvalue. Note that although the JD curve is below the biJD one, the asymptotic convergence of biJD looks more concave (superlinear). The right graph involves the matrix bwm200, and we seek the eigenvalue with the largest real part. In this case, the global convergence of JD is particularly fast. However, the credit should not go to the correction equation, because the same equation seems to hurt the convergence of `biJD(I-xx')`.

5.2. biJD vs. JD with restarting. In the second set of experiments, we examine the effects of restarting on JD and biJD. We allow 20 vectors for the JD basis and 20 vectors for each of the left and right bases of biJD. We thick restart JD and

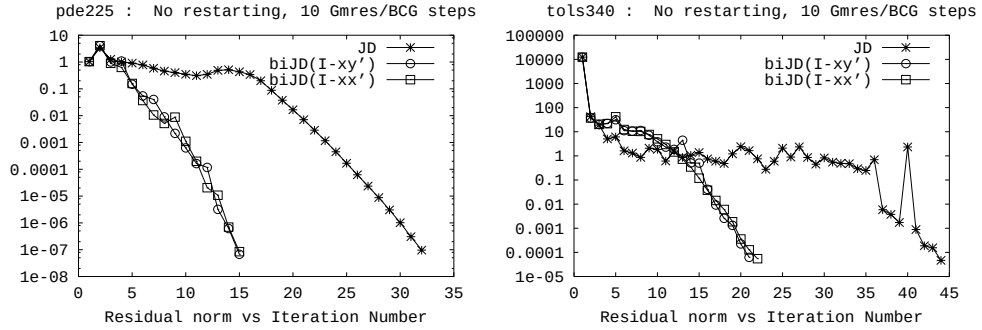


FIG. 5.1. *Convergence history for the residual norm of JD and variants of biJD in terms of outer iterations. There is no restarting, and in each outer iteration* 10 *GMRES or BCG steps are applied to the correction equation. Left graph: matrix pde*225*. Right graph: matrix tols*340*.*

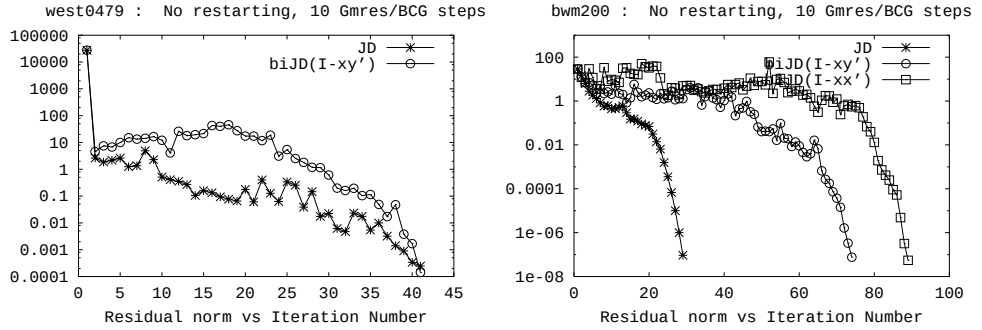


FIG. 5.2. *Convergence history for the residual norm of JD and variants of biJD in terms of outer iterations. There is no restarting, and in each outer iteration* 10 *GMRES or BCG steps are applied to the correction equation. Left graph: matrix west*0479*. Right graph: matrix bwm*200*.*

biJD with 5 Ritz vectors, while `biJD+1` thick restarts with four Ritz vectors and the Ritz vector from the previous step. In certain cases, we switch to the `biJD+1` scheme only after relatively good eigenvalue approximations have been obtained.

Our observations confirm that both JD and biJD outperform each other depending on the problem. However, while JD can use only thick restarting variants, biJD can use the combined restarting scheme, which can result in a substantial reduction of the number of iterations.

In Figure 5.3, the left graph involves the Tolosa matrix, but in this case, JD is faster than biJD. The `biJD+1` matches the performance of JD, assuming a fast, superlinear convergence, which for smaller thresholds would supersede JD. In the right graph, we look for the rightmost eigenvalue of the matrix rdb450. In this case, JD does not perform as well as biJD. The reason for the minor differences between biJD and `biJD+1` is that the algorithm converges before a second restart takes place. Finally, we note that the convergence of the `biJD(I-xx')` with thick restarting is not as good for this problem either.

In Figure 5.4, we examine restarted methods for the bwm200 matrix, both with the correction equation (left graph) and without it (right graph). Once again, the situation is reversed between the two methods. When solving the correction equation, JD is far better than biJD (see also the nonrestarted JD in Figure 5.2). `biJD+1` im-

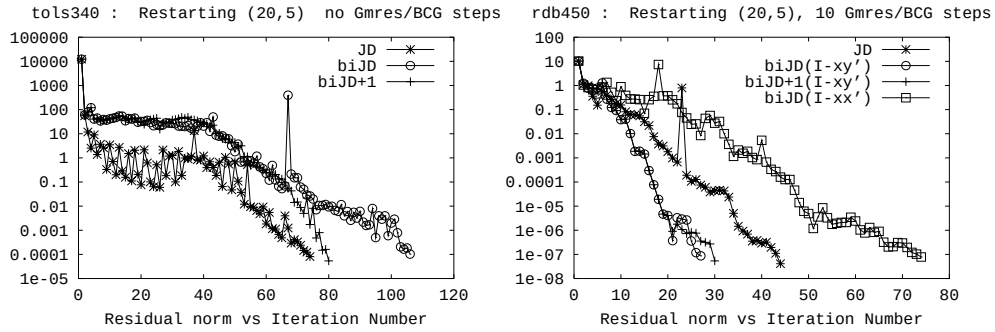


FIG. 5.3. *Convergence of the residual norm of the restarted JD and biJD. Maximum basis size is* 20, *and thick restarting is* 5. *For* `biJD+1` *thick restarting is* 4 *plus the previous Ritz vector. Left graph: matrix tol*340 *(no correction equation). Right graph: matrix rdb*450 *(with correction equation).*

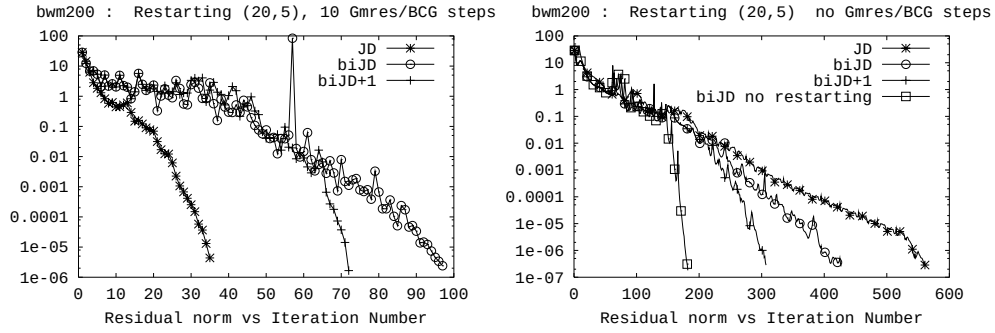


FIG. 5.4. *Convergence of the residual norm of the restarted JD and biJD. Maximum basis size is* 20, *and thick restarting is* 5. *For* `biJD+1` *thick restarting is* 4 *plus the previous Ritz vector. Matrix: bwm*200. *Left graph: with correction equation. Right graph: no correction equation.*

proves convergence, but it is still far from JD. On the other hand, when no correction equation is solved (right graph), JD converges the slowest, while using `biJD+1` comes surprisingly close to the nonrestarted method.

In Figure 5.5 we examine the interior problem from the matrix west0479. Note from Figure 5.2 that a subspace of 40 is enough to converge rapidly to the solution. By limiting the bases to 20 vectors, the iteration count increases dramatically, even with 10 steps on the correction equation. In this case, JD does not converge for at least 1300 steps, while biJD converges in 320 steps. The `biJD+1` scheme converges in more steps, if applied from the beginning, but it improves slightly on the biJD convergence, if applied after the Ritz value has relatively stabilized (residual norm less than 0.1). When no correction equation is solved (right graph), JD outperforms biJD. `biJD+1` applied during all restarts is substantially worse, possibly because in early iterations the restarting was locking onto a wrong eigenpair, discarding useful information. However, when applied dynamically only after the residual norm is less than 0.1, `biJD+1` can improve significantly the performance of the method.

5.3. Solving the correction equation more accurately. In this experiment we explore the effects on the `biJD+1` restarting scheme of applying more BCG steps on the correction equation. Because biJD is an inner-outer method, it is expected that

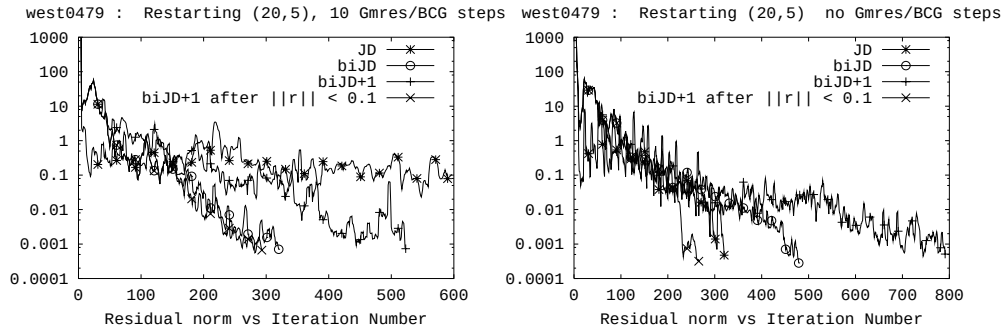


FIG. 5.5. *Convergence of the residual norm of the restarted JD and biJD. Maximum basis size is* 20*, and thick restarting is* 5*. A variant of* `biJD+1` *retains the previous Ritz vector if* $\|r_r\| < 0.1$*. Matrix: west*0479*. Left graph: with correction equation. Right graph: no correction equation.*

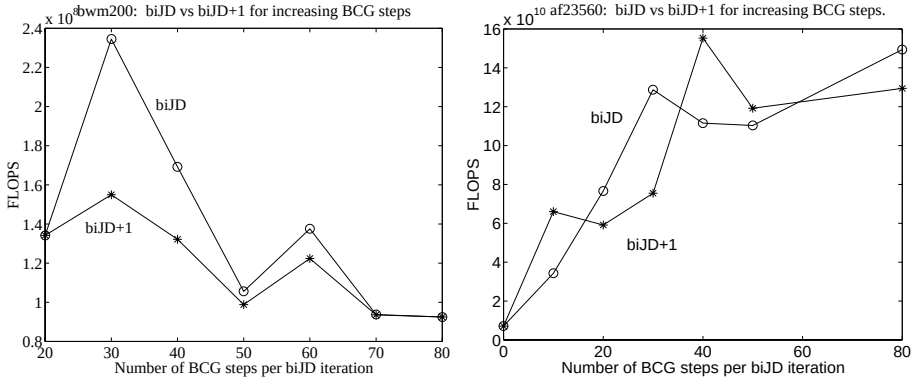


FIG. 5.6. *Floating point operations (flops) required for two problems as a function of inner BCG steps. No preconditioning is used for the bwm*200 *matrix, while a Matlab* `luinc(1e-3)` *preconditioner is used for the af*23560 *matrix. Solving the correction equation more accurately can reduce both execution time and the benefits of the* `biJD+1` *scheme, but neither is guaranteed.*

the number of outer iterations decreases when the inner ones increase, thus making the `biJD+1` scheme less necessary. However, this may not always be the case.

The left graph in Figure 5.6 shows an example of this expected behavior. We use the problem for matrix bwm200 without preconditioning, where the `biJD+1` scheme proved useful (see Figure 5.4). We show the number of floating point operations (flops) required by biJD and `biJD+1` for a range of numbers of inner BCG steps. As expected, the differences between the methods diminish and the overall operation counts decrease with larger numbers of BCG steps. Still, the `biJD+1` scheme is consistently better than or comparable to biJD.

The right graph in Figure 5.6 repeats the same experiment on the matrix af23560, the largest matrix available in the Matrix Market, seeking the eigenpair closest to zero. In this case, each BCG iteration applies also `luinc(A,1e-3)`, a Matlab incomplete LU factorization with threshold preconditioner. Contrary to the previous example, the total operation counts increase with larger numbers of BCG steps. This behavior is akin to inexact Newton solvers, where it does not pay to solve the inner equation too accurately when the approximation is far from the solution. The `biJD+1` restarting scheme is not always better than biJD, but it is competitive.

TABLE 5.1

Gflops (billion flops) required for finding the eigenvalue closest to zero of the matrix af23560. We compare JD, biJD, and the Matlab function `eigs`*. We test* `eigs` *as Arnoldi (no preconditioner) and as Arnoldi on* A^{-1}*. JD and biJD can also use the Matlab incomplete factorization* `luinc`*. The numbers include the factorization costs that are shown separately in the table on the right.*

Method	Preconditioner				
	None	`luinc`(ϵ)			LU
		ϵ=1e-3	ϵ=1e-4	ϵ=1e-6	
`eigs`	137.52	N/A	N/A	N/A	5.56
JD	139.89	404.81	4.67	5.92	7.80
biJD	-	8.00	3.23	10.08	11.86

Factorization	GFLOPS
`luinc`(1e-3)	0.894
`luinc`(1e-4)	1.617
`luinc`(1e-6)	3.000
LU	4.327

Typical heuristic strategies start with a small number of BCG steps and increase it slowly during the iterations as indicated by measured performance. As Figure 5.6 shows, the price of underestimating the number of BCG steps is much smaller (left graph) than the price of overestimating them (right graph). In view of the above, we expect the `biJD+1` restarting to be useful in general.

In our final experiment, we test the viability of the biJD method on the large, more realistic af23560 problem. We compare against JD and the implicitly restarted Arnoldi as implemented in Matlab's `eigs` function. All methods use a basis size of 20 and thick restart with 10 vectors. We vary the quality of the preconditioner, from no preconditioner at all, to incomplete factorizations `luinc(A,1e-3)`, `luinc(A,1e-4)`, and `luinc(A,1e-6)`, and to a complete factorization of the matrix A. Obviously `eigs` can be used only as standard Arnoldi or as shift-and-invert Arnoldi with A^{-1}. For the comparisons to be independent of hardware specifics, Table 5.1 reports the Gflops (10^9 flops) required to perform the factorizations and solve the problem. It also reports the factorization costs in the separate side table. We should mention that because left and right matrix-vector multiplications can be combined for efficient cache reuse, the timings of biJD can be better than its Gflops suggest. In addition, for the same Gflops, biJD takes about half the iterations of JD.

There are two main observations from this experiment: first, biJD provides the fastest possible solution to the problem; second, the performance of biJD is more consistent than that of JD for various preconditioners, and it is a better choice for weaker preconditioners. Because it is relatively inexpensive to LU factorize af23560 (it is close to a banded matrix), the performance of shift-and-invert Arnoldi is also competitive. However, complete factorizations are not possible in general. Without preconditioning Arnoldi is the least expensive method, although JD and biJD can still be advantageous for some problems (see, e.g., Figure 5.4). biJD did not converge in this case, even when the basis size was increased to 200. This suggests that the Galerkin projection is preferable to the Petrov–Galerkin for this problem.

6. Conclusions. The proposed biorthogonal Jacobi–Davidson method incorporates many of the advantages of the nonsymmetric Lanczos and the Jacobi–Davidson methods. We have given an elegant formulation of the algorithm that allows for a host of features and functionalities, including preconditioning, simple resolution of breakdowns, use of harmonic Ritz pairs, thick restarting, and use of left eigenvectors for both eigenvalue approximation and convergence estimation. We have also shown that on today's computers with multiple memory hierarchies, the multiplication of the adjoint of the matrix with a vector can be performed with only one memory access, and thus with minimal additional cost.

The two distinct characteristics of the biJD method that make it competitive

against the JD method are an asymptotically faster correction equation and an efficient restarting strategy. Restarting with a combination of Ritz vectors from the current and previous steps could offer huge convergence improvements to biJD, but not to JD. Although a similar restarting scheme could possibly be developed for CGS-like methods, the additional features and the faster correction equation make biJD a more promising choice.

As confirmed by our experiments, the method often outperforms JD, with and without restarting or correction equation. However, as with the Lanczos and Arnoldi methods, biJD and JD outperform each other in different problems. Moreover, harvesting the huge potential of the restarting scheme is not as easy to tune as in the symmetric case. Overall, however, biJD is a highly competitive algorithm for a difficult problem.

Acknowledgments. The author would like to thank Eric de Sturler for many fruitful discussions and Roland Freund and the anonymous referees for their constructive comments that improved this paper significantly.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods,* Cambridge University Press, New York, 1994.
[2] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM J. Sci. Comput., 20 (1998), pp. 243–269.
[3] Z. Bai, D. Day, J. Demmel, and J. Dongarra, *A Test Matrix Collection for Non-Hermitian Eigenvalue Problems*, Technical report, Department of Mathematics, University of Kentucky, Lexington, KY, 1996.
[4] Z. Bai, D. Day, and Q. Ye, *ABLE: An adaptive block Lanczos method for non-Hermitian eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1060–1082.
[5] 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.
[6] R. Barrett, M. W. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1994.
[7] C. Beattie, *Harmonic Ritz and Lehmann bounds*, Electron. Trans. Numer. Anal., 7 (1998), pp. 18–39.
[8] R. F. Boisvert, R. Pozo, K. Remington, R. Barrett, and J. J. Dongarra, *The Matrix Market: A web resource for test matrix collections*, in Quality of Numerical Software, Assessment and Enhancement, R. F. Boisvert, ed., Chapman & Hall, London, 1997, pp. 125–137.
[9] A. Chapman and Y. Saad, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., 4 (1997), pp. 43–66.
[10] M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.
[11] E. R. Davidson, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, J. Comput. Phys., 17 (1975), pp. 87–94.
[12] E. de Sturler, *Truncation strategies for optimal Krylov subspace methods*, SIAM J. Numer. Anal., 36 (1999), pp. 864–889.
[13] M. Eiermann and O. G. Ernst, *The geometry of Krylov subspace methods*, Acta Numer., 10 (2001), pp. 251–312.
[14] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis (Dundee, 1975), Lecture Notes in Math. 506, G. A. Watson, ed., Springer, Berlin, 1976, pp. 73–89.
[15] D. R. Fokkema, G. L. G. Sleijpen, and H. A. van der Vorst, *Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1998), pp. 94–125.
[16] D. R. Fokkema, *Subspace Methods for Linear, Nonlinear and Eigen Problems*, Ph.D. thesis, Utrecht University, Utrecht, The Netherlands, 1996.
[17] R. W. Freund, *Quasi-kernel polynomials and their use in non-Hermitian matrix iterations*, J. Comput. Appl. Math., 43 (1992), pp. 135–158.

[18] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
[19] M. Genseberger and G. L. G. Sleijpen, *Alternative Correction Equations in the Jacobi-Davidson Method*, Technical Report 1073, Department of Mathematics, University of Utrecht, Utrecht, The Netherlands, 1999.
[20] M. E. Hochstenbach and G. L. G. Sleijpen, *Two-sided and alternating Jacobi-Davidson*, Linear Algebra Appl., to appear.
[21] A. V. Knyazev, *A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace*, in Numerical Treatment of Eigenvalue Problems, Internat. Ser. Numer. Math. 96, Birkhäuser, Basel, 1991, pp. 143–154.
[22] C. Lanczos, *An iterative method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Nur. Stand., 45 (1950), pp. 255–282.
[23] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Nur. Stand., 49 (1952), pp. 33–53.
[24] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[25] R. B. Morgan, *On restarting the Arnoldi method for large nonsymmetric eigenvalue problems*, Math. Comp., 65 (1996), pp. 1213–1230.
[26] R. B. Morgan and D. S. Scott, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 817–825.
[27] C. W. Murray, S. C. Racine, and E. R. Davidson, *Improved algorithms for the lowest eigenvalues and associated eigenvectors of large matrices*, J. Comput. Phys., 103 (1992), pp. 382–389.
[28] C. C. Paige, B. N. Parlett, and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov spaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.
[29] B. N. Parlett, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 567–593.
[30] B. N. Parlett, D. R. Taylor, and Z. A. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comput., 33 (1985), pp. 680–687.
[31] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.
[32] Y. Saad, *SPARSKIT: A Basic Toolkit for Sparse Matrix Computations*, Technical Report 90-20, Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffet Field, CA, 1990. Software currently available at ftp://ftp.cs.umn.edu/dept/sparse/.
[33] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, 1992.
[34] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[35] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, MA, 1996.
[36] G. De Samblanx and A. Bultheel, *Nested Lanczos: Implicitly restarting an unsymmetric Lanczos algorithm*, Numer. Algorithms, 18 (1998), pp. 31–50.
[37] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. van der Vorst, *Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.
[38] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[39] G. L. G. Sleijpen and F. W. Wubs, *Effective Preconditioning Techniques for Eigenvalue Problems*, Technical Report 1117, Department of Mathematics, University of Utrecht, Utrecht, The Netherlands, 1999.
[40] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[41] A. Stathopoulos, *Some insights on restarting symmetric eigenvalue methods with Ritz and harmonic Ritz vectors*, in Iterative Methods in Scientific Computation IV, D. R. Kincaid and A. C. Elster, eds., IMACS, New Brunswick, NJ, 1999, pp. 297–311.
[42] A. Stathopoulos and Y. Saad, *Restarting techniques for (Jacobi-)Davidson symmetric eigenvalue methods*, Electron. Trans. Numer. Algorithms, 7 (1998), pp. 163–181.
[43] A. Stathopoulos, Y. Saad, and K. Wu, *Dynamic thick restarting of the Davidson, and the implicitly restarted Arnoldi methods*, SIAM J. Sci. Comput., 19 (1998), pp. 227–245.
[44] J. H. Van Lenthe and P. Pulay, *A space-saving modification of Davidson's eigenvector algorithm*, J. Comput. Chem., 11 (1990), pp. 1164–1168.
[45] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.
[46] K. Wu, Y. Saad, and A. Stathopoulos, *Inexact Newton preconditioning techniques for eigenvalue problems*, Electron. Trans. Numer. Algorithms, 7 (1998), pp. 202–214.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 260–280

LOW RANK SOLUTION OF LYAPUNOV EQUATIONS*

JING-REBECCA LI† AND JACOB WHITE‡

Abstract. This paper presents the Cholesky factor–alternating direction implicit (CF–ADI) algorithm, which generates a low rank approximation to the solution X of the Lyapunov equation $AX + XA^T = -BB^T$. The coefficient matrix A is assumed to be large, and the rank of the right-hand side $-BB^T$ is assumed to be much smaller than the size of A. The CF–ADI algorithm requires only matrix-vector products and matrix-vector solves by shifts of A. Hence, it enables one to take advantage of any sparsity or structure in A.

This paper also discusses the approximation of the dominant invariant subspace of the solution X. We characterize a group of spanning sets for the range of X. A connection is made between the approximation of the dominant invariant subspace of X and the generation of various low order Krylov and rational Krylov subspaces. It is shown by numerical examples that the rational Krylov subspace generated by the CF–ADI algorithm, where the shifts are obtained as the solution of a rational minimax problem, often gives the best approximation to the dominant invariant subspace of X.

Key words. Lyapunov equation, alternating direction implicit iteration, low rank approximation, dominant invariant subspace, iterative methods

AMS subject classifications. 65F30, 65F10, 15A24, 93C05

PII. S0895479801384937

1. Introduction.

In this paper we present the Cholesky factor–alternating direction implicit (CF–ADI) algorithm, which is well suited to solving large-scale Lyapunov equations whose right-hand sides have low rank. A Lyapunov equation has the form

$$AX + XA^T = -BB^T, \qquad A \in \mathbb{R}^{n\times n},\ X \in \mathbb{R}^{n\times n}. \tag{1.1}$$

The unknown is the matrix X. We assume that the coefficient matrix A is large and stable, $\lambda_i(A) < 0\ \forall i$. Furthermore, we assume that the rank of the right-hand side $-BB^T$ is much smaller than n, or simply, $\operatorname{rank}(B) = r_b \ll n$. When A is stable, the matrix X is symmetric from the uniqueness of the solution to (1.1), and it is positive semidefinite [18]. Such Lyapunov equations occur in the analysis and model reduction of large, linear, time-invariant systems, where the number of inputs and the number of outputs are small compared to the system size.

The first contribution of this paper is the CF–ADI algorithm, which is a reformulation of the alternating direction implicit (ADI) algorithm for Lyapunov equations [5, 8, 23, 38, 39, 40] and gives exactly the same approximation. However, CF–ADI requires only matrix-vector products and matrix-vector solves by shifts of A. Hence, it enables one to take advantage of any sparsity or structure in the coefficient matrix

*Received by the editors February 12, 2001; accepted for publication (in revised form) by V. Mehrmann January 15, 2002; published electronically July 9, 2002.

http://www.siam.org/journals/simax/24-1/38493.html

†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, Room 1118, New York, NY 10012-1185 (jingli@cims.nyu.edu). The research of this author was supported by the National Science Foundation and the Semiconductor Research Corporation/Hewlett-Packard Graduate Fellowships.

‡Research Laboratory of Electronics, Massachusetts Institute of Technology, Room 36-817, Cambridge, MA 02139-4307 (white@mit.edu).

A. The CF–ADI algorithm is intended to be used as a low rank algorithm to provide a low rank approximation to the exact solution matrix X. Frequently the exact solution X itself has low numerical rank [1, 28].

For some applications, it is sufficient to find the dominant invariant subspace of X. The complete knowledge of X is not necessary. For example, in the linear systems setting, the dominant invariant subspace of X may have physical meaning either as the span of the directions most sensitive to input or as the span of the directions to which the output is the most sensitive (see [6, 9, 33]). In fact, knowledge of the dominant invariant subspace of X is enough to produce the balanced truncation reduced model [26, 29] for symmetric systems [20, 21]. Hence, for some applications, approximating the dominant invariant subspace of X is as relevant as approximating X itself. In light of this, the second half of this paper is devoted to the approximation of the dominant invariant subspace of X.

The second contribution of this paper is making the connection between the approximation of the dominant invariant subspace of X and the generation of various low order Krylov and rational Krylov subspaces. It is shown that various methods of generating low rank approximations to X, including the CF–ADI algorithm, involve finding a low order Krylov or rational Krylov subspace to approximate the dominant invariant subspace of X. All these subspaces, when taken to order n, yield the full range of X. We compare the CF–ADI choice of a rational Krylov subspace, where the shifts are obtained by solving a rational minimax problem, with several other natural choices. We show by numerical examples that the subspace generated by CF–ADI often provides the best approximation to the dominant invariant subspace of X.

A preliminary form of the CF–ADI algorithm as applied to the model reduction problem can be found in [20, 21, 22]. In this paper we give details of the CF–ADI algorithm as relevant to the solution of (1.1). We also include complexity analysis, parameter selection procedure, stopping criteria, the use of real arithmetic, and numerical results on convergence, all of which appear for the first time in literature. Some early numerical results on using CF–ADI to approximate the dominant invariant subspace of X can be found in [22].

It has come to the authors' attention that another low rank reformulation of the ADI algorithm was independently proposed in [27]. However, in that version, the work required to produce a rank k approximation to X increases as $O(k^2)$, whereas for the CF–ADI algorithm presented in this paper, the work increases as $O(k)$. In fact, the algorithm in [27] appears as an intermediate step in deriving the final CF–ADI algorithm.

This paper is organized in the following way. Section 2 motivates the solution of the Lyapunov equation and the approximation of the dominant invariant subspace of the solution in the context of linear, time-invariant systems. Section 3 provides background on existing approaches to the solution of (1.1), including the ADI algorithm in some detail. Section 4 develops the CF–ADI algorithm. Section 5 contains a collection of definitions and useful results concerning Krylov and rational Krylov subspaces. Section 6 characterizes spanning sets for a subspace based on A and B. Section 7 shows that these spanning sets also span the range of X and uses that result to prove several properties of CF–ADI. Section 8 makes the connection between the approximation of the dominant invariant subspace of X and the generation of various low order Krylov and rational Krylov subspaces. We also make numerical comparisons of several different Krylov and rational Krylov subspace approximations. Section 9 contains the conclusions.

2. Motivation. Lyapunov equations with a low rank right-hand side occur in the analysis and model reduction of large, linear, time-invariant systems, where the system size is much larger than the number of inputs and the number of outputs. In this paper we focus on systems whose coefficient matrices are large and sparse. Such systems occur in interconnect modeling, solutions of PDEs, and other applications.

A linear, time-invariant system with realization (A, B, C) is characterized by the equations

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t), \tag{2.1}$$

$$y(t) = Cx(t). \tag{2.2}$$

The vector valued function $x(t) : \mathbb{R} \mapsto \mathbb{R}^n$ gives the state at time t and has n components. The input $u(t) : \mathbb{R} \mapsto \mathbb{R}^{r_b}$ and the output $y(t) : \mathbb{R} \mapsto \mathbb{R}^{r_c}$ have r_b and r_c components, respectively. The matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times r_b}$, $C \in \mathbb{R}^{r_c\times n}$ are the system matrix, the input coefficient matrix, and the output coefficient matrix, respectively. For single-input single-output (SISO) systems, $r_b = 1, r_c = 1$. Even for multiple-input, multiple-output (MIMO) systems, r_b and r_c are usually both very small compared to n.

If the system matrix A is stable, i.e., all the eigenvalues of A are in the open left half plane, then the controllability Gramian $P \in \mathbb{R}^{n\times n}$ and the observability Gramian $Q \in \mathbb{R}^{n\times n}$ associated with the system in (2.1)–(2.2) are the unique, symmetric, and positive semidefinite solutions to the following two Lyapunov equations (see, e.g., [6, 9, 18, 33]):

$$AP + PA^T = -BB^T, \tag{2.3}$$

$$A^TQ + QA = -C^TC. \tag{2.4}$$

If the number of inputs r_b is much smaller than the number of state components n, then $\operatorname{rank}(BB^T) = \operatorname{rank}(B) \le r_b \ll n$, and the right-hand side of (2.3) has low rank. Similarly, if the number of outputs r_c is much smaller than n, then the right-hand side of (2.4) has low rank.

The physical importance of the dominant eigenvectors of the Gramians P and Q is that they are the directions most sensitive to the input and the directions to which the output is the most sensitive, respectively (see [6, 9, 33]). In addition, for symmetric systems, where $A = A^T$ and $C = B^T$ in (2.1)–(2.2) and (2.3) and (2.4) are the same, knowledge of the dominant invariant subspace of $P = Q$ is sufficient to produce the balanced truncation reduced model [26, 29] for the system [20, 21]. Therefore, for some applications, approximating the dominant invariant subspace of the solution to (1.1) is as relevant as approximating the solution itself.

3. Previous methods. This section describes several existing methods for finding or approximating the solution X to the Lyapunov equation (1.1). Several of the algorithms described in this paper utilize the Cholesky factors of square matrices, and we give the definition of the Cholesky factor below.

DEFINITION 3.1. *A matrix Z is a* Cholesky factor *of $X \in \mathbb{R}^{n\times n}$ if it satisfies*

$$X = ZZ^T. \tag{3.1}$$

In this paper, the Cholesky factor Z is not required to be a square matrix nor have lower triangular structure.

The Bartels–Stewart method [2] first transforms A to real Schur form and then back solves for the solution of the transformed Lyapunov equation. The solution X is then obtained by a congruence transformation. Reducing a general, possibly sparse matrix to real Schur form requires $O(n^3)$ work, as does the congruence transformation to produce X.

The Hammarling method [12] also first transforms A to Schur form and has $O(n^3)$ complexity. It computes the lower triangular matrix Cholesky factor of the solution X rather than X itself.

The matrix sign function method [3, 30] exploits a simple connection between X and the matrix sign function of the $2n \times 2n$ matrix $\left[\begin{smallmatrix} A^T & 0 \\ BB^T & -A \end{smallmatrix}\right]$. The latter is found by Newton iteration. The complexity of this approach depends on the speed of the convergence of the Newton iteration but is at best $O(n^3)$. Low rank versions of the matrix sign function method can be found in [4, 19].

An approximate power iteration algorithm to determine the dominant invariant subspace of X is contained in [13], where approximations to the matrix-vector products Xv are computed. At each iteration, a Sylvester equation with a large left coefficient matrix and a small right coefficient matrix must be solved.

The low rank Smith(l) method in [27] gives the same approximation as the ADI method with cyclic parameters, and exploits the low rank of the right-hand side of the Lyapunov equation, but it is not as efficient as the CF–ADI algorithm to be derived in section 4. The main reason is that it is dependent on a low rank implementation of the ADI algorithm which is given in this paper in (4.6)–(4.7) and which is only an intermediate step in deriving the final CF–ADI algorithm.

3.1. Alternating direction implicit iteration. The ADI method [5, 39, 40, 41] is an iterative method and is given as Algorithm 1. The parameters $\{p_1, p_2, \ldots, p_J\}$, $\mathrm{Re}\{p_j\} < 0$, are called the ADI parameters. To keep the final ADI approximation

Algorithm 1. Alternating direction implicit algorithm.

INPUT: A, B.

1. If $v \mapsto Av, v \in \mathbb{R}^n$, is not $O(n)$ work, tridiagonalize A.
   a. Find $\tilde{A}$ tridiagonal, such that $\tilde{A} = SAS^{-1}$.
   b. Set $\tilde{B} := SB$.

   Otherwise, set $\tilde{A} := A, \tilde{B} := B$.
2. Choose ADI parameters, $\{p_1, \ldots, p_J\}$, $\mathrm{Re}\{p_i\} < 0$, (real or complex conjugate pairs), according to section 3.1.1 and references, using spectral bounds on $\tilde{A}$.
3. Initial guess,
   $$\tilde{X}_0 = 0_{n\times n}. \tag{3.2}$$
4. FOR $j = 1, 2, \ldots, J$, DO
   $$(\tilde{A} + p_j I)\tilde{X}_{j-\frac{1}{2}} = -BB^T - \tilde{X}_{j-1}(\tilde{A}^T - p_j I), \tag{3.3}$$
   $$(\tilde{A} + p_j I)\tilde{X}_j = -BB^T - \tilde{X}^T_{j-\frac{1}{2}}(\tilde{A}^T - p_j I). \tag{3.4}$$
   END
5. If A was tridiagonalized, recover solution,
   $$X^{adi}_J := S^{-1}\tilde{X}_J S^{-T}. \tag{3.5}$$
   Otherwise, $X^{adi}_J = \tilde{X}_J$.

OUTPUT: $X^{adi}_J \in \mathbb{R}^{n\times n}$, $X^{adi}_J \approx X$.

X_J^{adi} real, it is assumed that in the parameter list $\{p_1, p_2, \ldots, p_J\}$, each parameter is either real or comes as a part of a complex conjugate pair.

A general matrix A must be first reduced to tridiagonal form before proceeding with the ADI iteration in (3.3)–(3.4), to avoid the two full matrix-matrix products and two full matrix-matrix solves. However, it is well known that tridiagonalization of a general nonsymmetric matrix can be unstable (see, e.g., [10]).

The complexity of the ADI algorithm is $O(n^3) + O(Jn^2)$, where J is the total number of ADI iterations [23]. The $O(n^3)$ term comes from the tridiagonalization of a general matrix A, and the transformation in (3.5) to obtain the final ADI approximation. If A is already sparse or structured, there is no need to reduce A to tridiagonal form. In either case, the $O(Jn^2)$ term comes from J iterations of (3.3)–(3.4). In terms of complexity, the ADI method is competitive with the Bartels–Stewart and Hammarling methods, which are also $O(n^3)$ methods. However, the need in the ADI algorithm for the tridiagonalization of a general matrix A can pose a potentially serious problem.

If A is diagonalizable, then the ADI approximation X_J^{adi} has the following error bound [39]:

$$\|X_J^{adi} - X\|_F \leq \|T\|_2^2 \|T^{-1}\|_2^2 k(\mathbf{p})^2 \|X_0^{adi} - X\|_F,$$

$$k(\mathbf{p}) = \max_{x \in spec(A)} \left| \prod_{j=1}^{J} \frac{(p_j - x)}{(p_j + x)} \right|, \tag{3.6}$$

where T is a matrix whose columns are eigenvectors of A and $\mathbf{p} = \{p_1, p_2, \ldots, p_J\}$ are the ADI parameters.

3.1.1. ADI parameter selection. The selection of good parameters is vitally important to the successful application of the ADI algorithm. Optimal ADI parameters $\{p_1, p_2, \ldots, p_J\}$ are a function of J and solve the following rational minimax problem [40]:

$$\min_{p_1, p_2, \ldots, p_J} \max_{x \in \mathcal{R}} \left| \prod_{j=1}^{J} \frac{(p_j - x)}{(p_j + x)} \right|, \tag{3.7}$$

where $\mathcal{R}$ is a region in the open left half plane, and

$$\lambda_1(A), \ldots, \lambda_n(A) \in \mathcal{R} \subset \mathbb{C}^-.$$

If the eigenvalues of A are strictly real, then the solution to (3.7) is known (see [40]). The solution to (3.7) is not known when $\mathcal{R}$ is an arbitrary region in the open left half plane. The problem of finding optimal and near-optimal parameters was investigated in several papers [8, 15, 34, 35, 37, 40].

Here we summarize a parameter selection procedure given in [40]. Define the spectral bounds a, b, and α for the matrix A as

$$a = \min_i(\mathrm{Re}\{\lambda_i\}), \quad b = \max_i(\mathrm{Re}\{\lambda_i\}), \quad \alpha = \tan^{-1} \max_i \left| \frac{\mathrm{Im}\{\lambda_i\}}{\mathrm{Re}\{\lambda_i\}} \right|, \tag{3.8}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $-A$. It is assumed that the spectrum of $-A$ lies entirely inside the "elliptic function domain" determined by a, b, α, as defined

in [40]. If this assumption does not hold, one should try to apply a more general parameter selection algorithm. Let

$$\cos^2\beta = \frac{2}{1+\frac{1}{2}(\frac{a}{b}+\frac{b}{a})},$$

$$m = \frac{2\cos^2\alpha}{\cos^2\beta} - 1.$$

If $m < 1$, the parameters are complex and are given in [8, 40]. If $m \geq 1$, the parameters are real, and we define

$$k' = \frac{1}{m+\sqrt{m^2-1}}, \quad k = \sqrt{1-k'^2}.$$

Note that $k' = \frac{a}{b}$ if all the eigenvalues of A are real. Define the elliptic integrals K and v as

$$F[\psi,k] = \int_0^\psi \frac{dx}{\sqrt{1-k^2\sin^2 x}},$$

$$K = K(k) = F\left[\frac{\pi}{2},k\right], \qquad v = F\left[\sin^{-1}\sqrt{\frac{a}{bk'}},k'\right].$$

The number of ADI iterations required to achieve $k(\mathbf{p})^2 \leq \epsilon_1$ is $J = \lceil \frac{K}{2v\pi}\log\frac{4}{\epsilon_1}\rceil$, and the ADI parameters are given by

$$p_j = -\sqrt{\frac{ab}{k'}}\,dn\left[\frac{(2j-1)K}{2J},k\right], \quad j = 1,2,\ldots,J, \tag{3.9}$$

where $dn(u,k)$ is the elliptic function. It was noted in [23] that for many practical problems ADI converges in a few iterations with these parameters.

3.2. Low rank methods. In [14, 16], low rank approximations to X were proposed which have the form

$$X \approx X_J^{lr} := U_J X_{J\times J} U_J^T, \tag{3.10}$$

where the columns of $U_J \in \mathbb{R}^{n\times r_J}, r_J \leq Jr_b$, form an orthonormal basis for the block Krylov subspace

$$\mathcal{K}_J(A,B) := span\{B, AB, A^2B, \ldots, A^{J-1}B\}.$$

The columns of U_J, as well as the quantities $B_J := (U_J)^T B$ and $A_{J\times J} := U_J^T A U_J$, are obtained via the block Arnoldi process [7, 42].

If $\lambda_i(A_{J\times J}) + \bar{\lambda}_j(A_{J\times J}) \neq 0\ \forall\, i,j$, ensuring that a unique solution to (3.11) exists, then the residual of (1.1),

$$R_J(X_{J\times J}) := A(U_J X_{J\times J} U_J^T) + (U_J X_{J\times J} U_J^T)A^T + BB^T,$$

satisfies the Galerkin condition

$$U_J^T R_J(X_{J\times J}) U_J = 0$$

if and only if $X_{J\times J}$ satisfies

$$A_{J\times J}X_{J\times J} + X_{J\times J}A_{J\times J}^T + B_J B_J^T = 0 \tag{3.11}$$

[14, 16]. The more complicated linear matrix equation that $X_{J\times J}$ must satisfy in order to minimize the Frobenius norm of $R_J(X_{J\times J})$ was also given in [16].

4. CF–ADI. A major contribution of this paper is the development of the CF–ADI algorithm, which is presented in this section. For the low rank right-hand side Lyapunov equation (1.1), CF–ADI produces the same approximation as the ADI method described in section 3 but is much more efficient because it iterates on the Cholesky factor of the ADI approximation rather than on the ADI approximation itself.

For simplicity, all quantities in Algorithm 1 with tildes will be written in this section without the tildes. It is assumed that B has full column rank. Otherwise, we replace B with $\tilde{B}$, where $\tilde{B}$ has full column rank, and $\tilde{B}\tilde{B}^T = BB^T$.

There are two matrix-matrix products and two matrix-matrix solves in (3.3)–(3.4) of Algorithm 1. The need for matrix-matrix operations rather than simply matrix-vector operations at each ADI step makes Algorithm 1 extremely expensive. The first step in developing CF–ADI is to combine (3.3) and (3.4) to obtain

$$\begin{aligned} X_j = &-2p_j(A+p_jI)^{-1}BB^T(A+p_jI)^{-T} \\ &+ (A+p_jI)^{-1}(A-p_jI)X_{j-1}(A-p_jI)^T(A+p_jI)^{-T}. \end{aligned} \tag{4.1}$$

From (4.1) and the fact that $X_0 = 0_{n\times n}$, it can be seen that X_j is symmetric $\forall\ j \in \mathbb{Z}$, and that $\operatorname{rank}(X_j) \le \operatorname{rank}(X_{j-1}) + \operatorname{rank}(B)$. Since iteration begins with the zero matrix initial guess, $\operatorname{rank}(X_j) \le jr_b$, where r_b is the number of columns in B. Therefore, X_j can be represented as an outer product,

$$X_j = Z_j Z_j^T, \tag{4.2}$$

where Z_j has jr_b columns. The matrix Z_j is a Cholesky factor of $X_j \in \mathbb{R}^{n\times n}$.

Replacing X_j with $Z_j {Z_j}^T$ in (4.1) results in

$$Z_0 = 0_{n\times p}, \tag{4.3}$$

$$\begin{aligned} Z_jZ_j^T = &-2p_j\big\{(A+p_jI)^{-1}B\big\}\big\{(A+p_jI)^{-1}B\big\}^T \\ &+ \big\{(A+p_jI)^{-1}(A-p_jI)Z_{j-1}\big\}\big\{(A+p_jI)^{-1}(A-p_jI)Z_{j-1}\big\}^T. \end{aligned} \tag{4.4}$$

The left-hand side of (4.4) is an outer product, and the right-hand side is the sum of two outer products. Thus, Z_j on the left-hand side of (4.4) can be obtained simply by combining the two factors in the two outer products on the right:

$$Z_j = \left[\sqrt{-2p_j}\big\{(A+p_jI)^{-1}B\big\},\ \big\{(A+p_jI)^{-1}(A-p_jI)Z_{j-1}\big\}\right]. \tag{4.5}$$

Thus, the ADI algorithm can be reformulated in terms of the Cholesky factor Z_j of X_j. There is no need to calculate or store X_j at each iteration—only Z_j is needed.

The preliminary form of CF–ADI which iterates on the Cholesky factor Z_j of X_j is

$$Z_1 = \sqrt{-2p_1}(A+p_1I)^{-1}B, \quad Z_1 \in R^{n\times r_b}, \tag{4.6}$$

$$Z_j = \left[\sqrt{-2p_j}(A+p_jI)^{-1}B, (A+p_jI)^{-1}(A-p_jI)Z_{j-1}\right], \quad Z_j \in R^{n\times jr_b}. \tag{4.7}$$

In this formulation, at each iteration, the previous Cholesky factor $Z_{j-1} \in \mathbb{R}^{n\times(j-1)r_b}$ needs to be modified by multiplication on the left by $(A+p_jI)^{-1}(A-p_jI)$. Thus, the number of columns which need to be modified at each iteration increases by r_b. The implementation in (4.6)–(4.7) was independently developed in [27].

In this paper, a further step is taken to keep constant the number of columns modified at each iteration.

The Jr_b columns of Z_J, the Cholesky factor of the Jth ADI approximation, can be written out explicitly:

$$Z_J = \left[S_J\sqrt{-2p_J}B, \quad S_J\left(T_JS_{J-1}\right)\sqrt{-2p_{J-1}}B, \ldots, S_JT_J\cdots S_2\left(T_2S_1\right)\sqrt{-2p_1}B \right],$$

where

$$S_i = (A + p_iI)^{-1}, \quad T_i = (A - p_iI). \tag{4.8}$$

Note that the S_i's and the T_i's commute:

$$S_iS_j = S_jS_i, \quad T_iT_j = T_jT_i, \quad S_iT_j = T_jS_i \quad \forall i, j.$$

The Cholesky factor Z_J then becomes

$$Z_J = [z_J, \quad P_{J-1}(z_J), \quad P_{J-2}(P_{J-1}z_J), \quad \ldots, P_1(P_2\cdots P_{J-1}z_J)], \tag{4.9}$$

where

$$z_J := \left(\sqrt{-2p_J}\right) S_JB = \sqrt{-2p_J}(A + p_JI)^{-1}B, \tag{4.10}$$

$$\begin{aligned} P_l := \left(\frac{\sqrt{-2p_l}}{\sqrt{-2p_{l+1}}}\right) S_lT_{l+1} &= \frac{\sqrt{-2p_l}}{\sqrt{-2p_{l+1}}}(A + p_lI)^{-1}(A - p_{l+1}I) \\ &= \left(\frac{\sqrt{-2p_l}}{\sqrt{-2p_{l+1}}}\right)[I - (p_{l+1} + p_l)\,(A + p_lI)^{-1}]. \end{aligned} \tag{4.11}$$

Since there is no significance to the order in which the ADI parameters appear, the index $1, \ldots, J$ in (4.9) can be reversed. The CF–ADI algorithm which comprises (4.9)–(4.11) with the index reversed is given as Algorithm 2.

Algorithm 2. The CF–ADI algorithm.

INPUT: A, B.

1. Choose CF–ADI parameters, $\{p_1, \ldots, p_{J_{max}}\}$, $\mathrm{Re}\{p_i\} < 0$, (real or complex conjugate pairs).
2. Define: $P_i = \left(\frac{\sqrt{-2p_{i+1}}}{\sqrt{-2p_i}}\right)[I - (p_{i+1} + p_i)(A + p_{i+1}I)^{-1}]$.

   a. $z_1 = \left(\sqrt{-2p_1}\right)(A + p_1I)^{-1}B,$ (4.12)

   b. $Z_1^{cfadi} = \left[\ z_1\ \right].$ (4.13)

3. FOR $j = 2, 3, \ldots, J_{max}$

   a. $z_j = P_{j-1}z_{j-1},$ (4.14)

   b. If ($\|z_j\|_2 > tol_1$ or $\frac{\|z_j\|_2}{\|Z_{j-1}\|_2} > tol_2$) and ($j \le J_{max}$)

   $$Z_j^{cfadi} = \left[\ Z_{j-1}^{cfadi} \quad z_j\ \right].$$ (4.15)

   Otherwise, $J = j - 1$, stop.

   END

OUTPUT: $Z_J^{cfadi} \in \mathbb{C}^{n\times Jr_b}$, $X \approx X_J^{cfadi} := Z_J^{cfadi}(Z_J^{cfadi})^T \in \mathbb{R}^{n\times n}$.

We now show that CF–ADI produces the same approximation as the ADI method.

THEOREM 4.1. *If X_J^{adi} is obtained by running J steps of Algorithm 1 with the ADI parameters $\{p_1, p_2, \ldots, p_J\}$ and Z_J^{cfadi} is obtained by running J steps of Algorithm 2 with the same parameters in any order, then*

$$X_J^{adi} = Z_J^{cfadi}(Z_J^{cfadi})^T. \tag{4.16}$$

Proof. From the derivation of CF–ADI, it is clear that (4.16) is true when the order of the parameters is reversed. The fact that parameter order does not matter in either algorithm is shown by

$$X_j = (A + p_jI)^{-1}(A + p_{j-1}I)^{-1}\Big((A - p_jI)(A - p_{j-1}I)X_{j-2}(A - p_jI)^T(A - p_{j-1}I)^T$$
$$- 2(p_j + p_{j-1})(ABB^TA^T + p_jp_{j-1}BB^T)\Big)(A + p_jI)^{-T}(A + p_{j-1}I)^{-T}.$$

Clearly, this expression does not depend on the order of p_j and p_{j-1}. Any ordering of $\{p_1, \ldots, p_J\}$ can be obtained by exchanging neighboring parameters. ∎

As a matter of notation, define

$$X_J^{cfadi} := Z_J^{cfadi}(Z_J^{cfadi})^T. \tag{4.17}$$

Both X_J^{cfadi} and Z_J^{cfadi} will be referred to as the Jth CF–ADI approximation—which one is meant will be clear from context. The full matrix X_J^{cfadi} is usually not explicitly calculated. It will be used in subsequent sections for analysis purposes only.

4.1. Stopping criteria and parameter selection. The stopping criterion $\|X_j^{cfadi} - X_{j-1}^{cfadi}\|_2 \leq tol^2$ can be implemented as $\|z_j\|_2 \leq tol$, since

$$\|Z_jZ_j^T - Z_{j-1}Z_{j-1}^T\|_2 = \|z_jz_j^T\|_2 = \|z_j\|_2^2.$$

It is not necessarily true that a small z_j implies that all further z_{j+k} will be small, but this has been observed in practice. Relative error can also be used, in which case the stopping criterion is $\frac{\|z_j\|_2}{\|Z_{j-1}\|_2} \leq tol$. The 2-norm of Z_{j-1}, which is also its largest singular value, can be estimated by performing power iterations to estimate the largest eigenvalue of $Z_{j-1}Z_{j-1}^T$, taking advantage of the fact that $j \ll n$. This cost is still high, and this estimate should be used only after each segment of several iterations.

The criterion for picking CF–ADI parameters, $\{p_1, \ldots, p_{J_{max}}\}$, is exactly the same as for ADI parameters, namely, they should solve the rational minimax problem (3.7). Section 3.1.1 gives a parameter selection procedure based on three spectral bounds of A in (3.8). These three bounds for A may be estimated using the power and inverse power iterations, or Gershgorin's circles (see [10]). Power and inverse power iterations can be done at the cost of a few matrix-vector products and solves. A numerical comparison of different choices of parameters is given in section 8.1.

4.2. Complexity. The following definition is helpful when B has more than one column.

DEFINITION 4.2. *An r_b-vector $v \in \mathbb{R}^{n \times r_b}$ is a matrix that has r_b columns.*

The final CF–ADI approximation Z_J^{cfadi} can be obtained from the starting r_b-vector z_1 after $J - 1$ products of the form P_iz_i. The cost of applying P_i to a vector

TABLE 4.1
ADI and CF–ADI complexity comparison when A is sparse.

	CF–ADI	ADI
Sparse A	$O(Jr_bn)$	$O(Jn^2)$
Full A	$O(Jr_bn^2)$	$O(n^3)+O(Jn^2)$

is that of a matrix-vector solve. The starting r_b-vector z_1 is obtained after r_b matrix-vector solves with the columns of $B \in \mathbb{R}^{n\times r_b}$ as the right-hand sides. Each succeeding r_b-vector in Z_J^{cfadi} is obtained from the previous r_b-vector at the cost of r_b matrix-vector solves. Thus, the work per iteration has been reduced from two matrix-matrix products and two matrix-matrix solves in (3.2)–(3.3) in the original ADI method to r_b matrix-vector solves in (4.14) in the CF–ADI algorithm.

The Cholesky factor of the Lyapunov solution is precisely what is needed in the model reduction of linear, time-invariant systems [26, 32, 36]. In general, if Z_J^{cfadi} is available, it is not necessary to calculate $X_J^{cfadi} = Z_J^{cfadi}(Z_J^{cfadi})^T$, whereas if the ADI approximation X_J^{adi} is available, it is often necessary to calculate its Cholesky factor in the subsequent model reduction procedure.

If the matrix A is sparse enough so that $v \mapsto Av$ as well as $v \mapsto (A+p_iI)^{-1}v$ have $O(n)$ complexity, where v is a vector, then Table 4.1 gives the complexity comparison between ADI and CF–ADI. Since r_b, the number of inputs, is by assumption much smaller than n, CF–ADI always results in substantial savings when A is sparse, reducing the work from $O(n^2)$ to $O(n)$.

4.3. Real CF–ADI for complex parameters. Algorithm 2 will result in a complex Cholesky factor $Z_J \in \mathbb{C}^{n\times Jr_b}$ if there are complex ADI parameters, although $Z_JZ_J^T \in \mathbb{R}^{n\times n}$ is guaranteed to be real if the parameters come in complex conjugate pairs.

A version of CF–ADI which uses only operations with real numbers can be implemented by noting that analogous to the matrices associated with a real parameter p_i, given in (4.8), the matrices associated with a complex conjugate pair $\{p_i, \bar{p_i}\}$ are

$$Q_i := (A^2 - \sigma_iA + \tau_iI)^{-1}, \quad R_i := (A^2 + \sigma_iA + \tau_iI), \tag{4.18}$$

$$\sigma_i = 2\mathrm{Re}\{-p_i\}, \quad \tau_i = |p_i|^2, \tag{4.19}$$

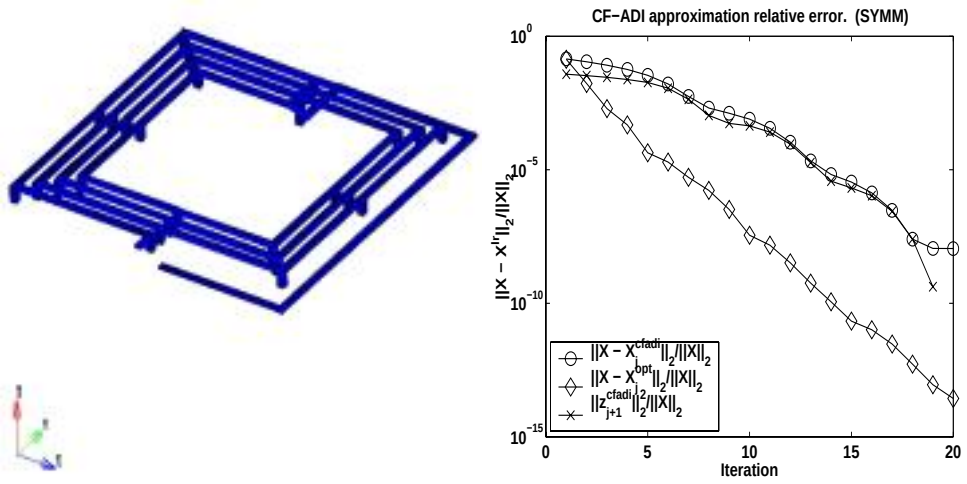
which involve only real quantities.

4.4. Numerical results. This section gives numerical results on the CF–ADI approximation to the solution to (1.1).

The example in Figure 4.1(b) comes from the inductance extraction of an on-chip planar square spiral inductor suspended over a copper plane [17], shown in Figure 4.1(a). The original order 500 system has been symmetrized according to [25]. The matrix A is a symmetric 500×500 matrix, and the input coefficient matrix $B \in \mathbb{R}^n$ has one column.

Because A is symmetric, the eigenvalues of A are real, and good CF–ADI parameters are easy to find. The procedure given in section 3.1.1 was followed. CF–ADI was run to convergence in this example, which took 20 iterations.

Figure 4.1(b) shows the relative 2-norm error of the CF–ADI approximation, $\frac{\|X-X_j^{cfadi}\|_2}{\|X\|_2}$, for $j = 1,\ldots,20$. At $j = 20$, relative error has reached 10^{-8}, which is about the same size as the error of the optimal [10] rank 11 approximation. The error estimate $\|z_{j+1}^{cfadi}\|_2^2$ approximates the actual error $\|X - X_j^{cfadi}\|$ closely $\forall\ j$.



(a) Spiral inductor (b) CF–ADI approximation

FIG. 4.1. *Spiral inductor, a symmetric system.*

5. Krylov and rational Krylov subspace results. This section contains a collection of definitions and results concerning Krylov and rational Krylov subspaces which will be used in subsequent sections.

We begin by giving definitions of Krylov and rational Krylov subspaces.

DEFINITION 5.1. *An* order m *Krylov subspace* $\mathcal{K}_{\mathbf{m}}(\mathbf{A}, \mathbf{z_1})$, $A \in \mathbb{R}^{n\times n}, z_1 \in \mathbb{R}^n$, *is the subspace*

$$\mathcal{K}_m(A, z_1) := span\left\{z_1,\ Az_1,\ A^2z_1, \ldots,\ A^{m-1}z_1\right\}. \tag{5.1}$$

DEFINITION 5.2. *An* order m *rational Krylov subspace* $\mathcal{K}^{\mathbf{rat}}_{\mathbf{m}}(\mathbf{A}, \mathbf{z_1}, \mathbf{p_{m-1}})$, $A \in \mathbb{R}^{n\times n}, z_1 \in \mathbb{R}^n$, $\mathbf{p_{m-1}} = \{p_1, \ldots, p_{m-1}\}$, $p_i \in \mathbb{R}$, *is the subspace*

(5.2)

$$\mathcal{K}^{rat}_m(A, z_1, \mathbf{p_{m-1}})$$
$$:= span\left\{z_1,\ (A+p_1I)^{-1}z_1,\ (A+p_2I)^{-1}(A+p_1I)^{-1}z_1, \ldots,\ \prod_{i=1}^{m-1}(A+p_iI)^{-1}z_1\right\}.$$

Note that for both Krylov and rational Krylov subspaces, the dimension of the subspace may be strictly smaller than the order m. The sets $\{z_1, \ldots, A^{m-1}z_1\}$ and $\{z_1, (A+p_1I)^{-1}z_1, \ldots, \prod_{i=1}^{m-1}(A+p_i)^{-1}z_1\}$ are *spanning sets* for $\mathcal{K}_m(A, z_1)$ and $\mathcal{K}^{rat}_m(A, z_1, \mathbf{p_{m-1}})$, respectively.

The following well-known result can be found in many standard textbooks, including [10].

PROPOSITION 5.3. *If* $m > n$, *then* $\mathcal{K}_m(A, B) = \mathcal{K}_n(A, B)$.

Theorem 5.4 characterizes the rational Krylov subspace $\mathcal{K}^{rat}_m(A, (A+p_1I)^{-1}B, \{p_2, \ldots, p_m\})$ as the direct sum of l rational Krylov subspaces, where l is the number of distinct parameters in the list $\{p_1, \ldots, p_m\}$.

THEOREM 5.4. *Let* $\mathcal{K}_m^{rat}(A,(A+p_1I)^{-1}B,\{p_2,\dots,p_m\})$ *be such that no* $(A+p_iI)$ *is singular. Then*

$$\begin{aligned}
&\mathcal{K}_m^{rat}\left(A,(A+p_1I)^{-1}B,\{p_2,\dots,p_m\}\right)\\
&= span\left\{(A+p_1I)^{-1}B,\dots,\prod_{i=1}^{j}(A+p_iI)^{-1}B,\dots,\prod_{i=1}^{m}(A+p_iI)^{-1}B\right\}\\
&= \sum_{i=1}^{l} span\{(A+p_iI)^{-1}B,\dots,(A+p_iI)^{-s_i}B\}\\
&= \sum_{i=1}^{l}\mathcal{K}_{s_i}^{rat}\left((A+p_iI),\ (A+p_iI)^{-1}B,\ \mathbf{0_{s_i-1}}\right),
\end{aligned}$$

where $s_1+\cdots+s_l=m$, *each* p_i *appears in* $\{p_1,\dots,p_m\}$ *a total of* s_i *times, and the summation sign denotes direct sum of subspaces.*

Proof. If the parameters are distinct, the proof follows from the partial fractions expansion

$$\prod_{i=1}^{j}(A+p_iI)^{-1}=\sum_{i=1}^{j}\left(\prod_{k\neq i}\left(\frac{1}{p_k-p_i}\right)\right)(A+p_iI)^{-1},\quad p_1\neq p_2\neq\cdots\neq p_J.$$

A slightly different expansion taking into account repeated parameters can be calculated to give the general statement of the theorem. ∎

6. Spanning sets of $\mathcal{L}(A,B)$. In this section we prove Theorem 6.1, which shows the equivalence of an infinite number of order n Krylov and rational Krylov subspaces based on A and B. For simplicity we assume B has only one column. Most of the results in this section can be easily generalized to the case when B has more than one column.

THEOREM 6.1. *Let* $A\in\mathbb{R}^{n\times n}$ *be invertible,* $B\in\mathbb{R}^n$, $B\neq 0$, $\mathbf{p}=\{\dots,p_{-2},p_{-1},p_0,p_1,p_2,\dots\}$, $p_i\in\mathbb{R}$, *and define the subspace* $\mathcal{L}(A,B,\mathbf{p})$ *as*

$$\begin{aligned}
&\mathcal{L}(A,B,\mathbf{p})\\
&:= span\Big\{\dots,\ \prod_{i=-j}^{-1}(A+p_iI)^{-1}B,\dots,\ (A+p_{-2}I)^{-1}(A+p_{-1}I)^{-1}B,\\
&\qquad (A+p_{-1}I)^{-1}B,\ B,\ (A+p_0I)B,\\
&\qquad (A+p_1I)(A+p_0I)B,\dots,\ \prod_{i=0}^{j-1}(A+p_iI)B,\dots\Big\}\\
&= span\big\{\dots,v_{-j}(A,B,\mathbf{p}),\dots,v_{-2}(A,B,\mathbf{p}),\ v_{-1}(A,B,\mathbf{p}),\ v_0(A,B,\mathbf{p}),\\
&\qquad v_1(A,B,\mathbf{p}),\ v_2(A,B,\mathbf{p}),\dots,v_j(A,B,\mathbf{p}),\dots\big\},
\end{aligned}\tag{6.1}$$

where

$$v_j(A,B,\mathbf{p})=\begin{cases}B, & j=0,\\ \prod_{i=0}^{j-1}(A+p_iI)B, & j>0,\\ \prod_{i=j}^{-1}(A+p_iI)^{-1}B, & j<0,\end{cases}\tag{6.2}$$

and where all matrix inverses in (6.1) *are well defined. Then* $\forall s \in \mathbb{Z}$, $\forall \mathbf{p}$, $\forall \mathbf{r} = \{\dots, r_{-1}, r_0, r_1, \dots\}$, $\forall \mathbf{q} = \{\dots, q_{-1}, q_0, q_1, \dots\}, r_i, q_i \in \mathbb{R}$,

$$\mathcal{L}(A, B, \mathbf{p}) = span\{v_s(A, B, \mathbf{p}),\ v_{s+1}(A, B, \mathbf{p}), \dots, v_{s+(n-1)}(A, B, \mathbf{p})\} \tag{6.3}$$

$$= span\{B, AB, \dots,\ A^{n-1}B\} \tag{6.4}$$

$$= \mathcal{L}(A, v_s(A, B, \mathbf{r}), \mathbf{q}) \tag{6.5}$$

if all matrix inverses in (6.5) *are well defined.*

REMARK 1. *We refer to* B *in* $\mathcal{L}(A, B, \mathbf{p})$ *as the base vector. Because of* (6.5), $\mathcal{L}(A, B) := \mathcal{L}(A, v_s(A, B, \mathbf{r}), \mathbf{q})$ *may be written without referring to the base vector* $v_s(A, B, \mathbf{r})$ *or the shifts* $\mathbf{q}$.

The proof of Theorem 6.1 needs the following lemmas. The dependence of the v_i's on $A, B, \mathbf{p}$ will be suppressed in the proofs unless needed.

LEMMA 6.2. *Let the* v_j*'s be defined as in* (6.2). *Then*

$$v_l \in span\{v_s,\ v_{s+1},\ v_{s+2}, \dots, v_{s+(n-1)}\} \tag{6.6}$$

whenever $l > s + (n-1)$.

Proof. From (6.2), it can be seen that $v_j = (A + p_{j-1}I)v_{j-1}\ \forall j$; hence,

$$span\{v_{j-1}, v_j\} = span\{v_{j-1}, Av_{j-1}\}$$

and

$$span\{v_s,\ v_{s+1},\ v_{s+2}, \dots, v_l\} = span\{v_s, Av_s, \dots, A^{l-s}v_s\} = \mathcal{K}_{l-s+1}(A, v_s). \tag{6.7}$$

From Proposition 5.3,

$$\begin{aligned} span\{v_s,\ v_{s+1},\ v_{s+2}, \dots, v_l\} &= \mathcal{K}_{l-s+1}(A, v_s) \\ &= \mathcal{K}_n(A, v_s) = span\{v_s,\ v_{s+1},\ v_{s+2}, \dots, v_{s+(n-1)}\}. \end{aligned}$$

The result follows. ∎

LEMMA 6.3. *Let the* v_j*'s be defined as in* (6.2)*; then*

$$v_l \in span\{v_s, v_{s+1}, v_{s+2}, \dots, v_{s+(n-1)}\} \tag{6.8}$$

whenever $l < s$.

Proof. First we show that the lemma is true for $l = s - 1$. Equivalently, because of (6.7), show that

$$(A + p_{s-1}I)^{-1}v_s \in span\{v_s, Av_s, \dots, A^{n-1}v_s\}. \tag{6.9}$$

Shifts can be added in the right-hand side of (6.9),

$$span\{v_s, Av_s, \dots, A^{n-1}v_s\} = span\{v_s, (A + p_{s-1}I)v_s, \dots, (A + p_{s-1}I)^{n-1}v_s\},$$

without affecting its column span. Because $\{v_{s-1}, v_s, \dots, v_{s+(n-1)}\}$ are $n+1$ vectors in $\mathbb{R}^n$, there exist coefficients, $c_0, \dots, c_n$, not all zero, such that

$$c_0 v_s + c_1(A + p_{s-1}I)v_s + \cdots + c_{n-1}(A + p_{s-1}I)^{n-1}v_s + c_n(A + p_{s-1}I)^{-1}v_s = 0. \tag{6.10}$$

If $c_n \neq 0$, (6.9) is proven. Otherwise, since $B \neq 0$, we can choose $0 \leq j < n-1$ such that $c_j \neq 0$ and $c_i = 0\ \forall i < j$. Then multiply (6.10) by $(A + p_{s-1}I)^{-(j+1)}$ to obtain

$$\begin{aligned} &c_j(A + p_{s-1}I)^{-1}v_s + c_{j+1}v_s + \cdots + c_{n-1}(A + p_{s-1}I)^{n-2-j}v_s = 0 \\ &\implies c_j(A + p_{s-1}I)^{-1}v_s = -c_{j+1}v_s - \cdots - c_{n-1}(A + p_{s-1}I)^{n-2-j}v_s. \end{aligned}$$

Thus, (6.9) is proven, and (6.8) holds for $l = s-1$. If $l < s-1$,

$$v_l \in \mathit{span}\{v_{l+1}, v_{l+2}, \ldots, v_{l+n}\} \tag{6.11}$$

$$\subseteq \mathit{span}\{v_{l+2}, \ldots, v_{l+n+1}\} \tag{6.12}$$

$$\vdots \tag{6.13}$$

$$\subseteq \mathit{span}\{v_s, \ldots, v_{s+n-1}\}. \tag{6.14}$$

Relation (6.12) follows because each vector $v_{l+1}, \ldots, v_{l+n}$ is in $\mathit{span}\{v_{l+2}, \ldots, v_{l+n+1}\}$. ∎

Proof of Theorem 6.1. Lemmas 6.2 and 6.3 show that for any $\mathbf{p}$,

$$\mathcal{L}(A, B, \mathbf{p}) = \mathit{span}\{v_s(A, B, \mathbf{p}),\ v_{s+1}(A, B, \mathbf{p}), \ldots, v_{s+(n-1)}(A, B, \mathbf{p})\}$$

holds for any s. Equation (6.4) follows from the fact that for any $\mathbf{p}$, with the choice of $s = 0$,

$$\mathit{span}\{v_0(A, B, \mathbf{p}), v_1(A, B, \mathbf{p}), \ldots, v_{n-1}(A, B, \mathbf{p})\} = \mathit{span}\{B, AB, \ldots, A^{n-1}B\}.$$

Equation (6.5) follows from

$$\begin{aligned} \mathcal{L}(A, B, \mathbf{p}) &= \mathit{span}\{B, AB, \ldots, A^{n-1}B\} = \mathcal{L}(A, B, \mathbf{r}) \\ &= \mathit{span}\{v_s(A, B, \mathbf{r}), v_{s+1}(A, B, \mathbf{r}), \ldots, v_{s+(n-1)}(A, B, \mathbf{r})\} \\ &= \mathit{span}\{v_s(A, B, \mathbf{r}), Av_s(A, B, \mathbf{r}), \ldots, A^{n-1}v_s(A, B, \mathbf{r})\} \\ &= \mathcal{L}(A, v_s(A, B, \mathbf{r}), \mathbf{q}) \quad \forall \mathbf{p}, \quad \forall \mathbf{r}, \quad \forall \mathbf{q}. \end{aligned}$$ ∎

Remark 2. *Special cases of Theorem* 6.1 *can be found in many references, including* [11, 31].

7. Lyapunov solution and rational Krylov subspaces. In this section we characterize the range of the Lyapunov solution as order n Krylov and rational Krylov subspaces with different starting vectors and different sets of shifts. We also state several properties of the CF–ADI approximation.

Proposition 7.1 is a well-known result which makes the connection between the range of the Lyapunov solution X and the Krylov subspace $\mathcal{K}_n(A, B)$ (see [6, 33]).

Proposition 7.1. *Let X be the solution to* (1.1). *Then*

$$\text{range}(X) = \mathit{span}\{B, AB, \ldots, A^{n-1}B\} = \mathcal{K}_n(A, B). \tag{7.1}$$

The following corollary of Theorem 6.1 gives a more complete characterization of the range of X as Krylov and rational Krylov subspaces.

Corollary 7.2. *With the same notation as in Theorem* 6.1,

$$\text{range}(X) = \mathcal{L}(A, v_t(A, B, \mathbf{r}), \mathbf{q}) \quad \forall t \in \mathbb{Z}, \quad \forall \mathbf{r}, \quad \forall \mathbf{q}. \tag{7.2}$$

Theorem 6.1 and Corollary 7.2 together imply that any n consecutive vectors $\{w_s, \ldots, w_{s+n-1}\}$, $s \in \mathbb{Z}$, in the infinite spanning set for $\mathcal{L}(A, v_t(A, B, \mathbf{r}), \mathbf{q})$ are a spanning set for range(X).

We now state some properties of the CF–ADI approximation and omit the proofs.

PROPOSITION 7.3. *Let Z_j^{cfadi} be the jth CF–ADI approximation. Then its column span has the following characterization:*

$$colsp(Z_j^{cfadi}) = \mathcal{K}_j^{rat}(A,\ (A + p_1 I)^{-1} B,\ \{p_2, \ldots, p_j\}). \tag{7.3}$$

PROPOSITION 7.4. *Let $Z_j^{cfadi} = [z_1, \ldots, z_j]$ be the jth CF–ADI approximation, and let $B \in \mathbb{R}^n$. If z_{j+1} is a linear combination of $\{z_1, \ldots, z_j\}$, then z_l is a linear combination of $\{z_1, \ldots, z_j\}$ whenever $l \geq j + 1$.*

PROPOSITION 7.5. *Let $Z_n^{cfadi} = [z_1, \ldots, z_n]$ be the nth CF–ADI approximation. Then*

$$\begin{aligned} colsp(Z_n^{cfadi}) &= \mathcal{K}_n^{rat}(A, (A - p_1 I)^{-1} B, \{p_2, \ldots, p_n\}) \\ &= \text{range}(X). \end{aligned}$$

REMARK 3. *Proposition* 7.5 *states that if CF–ADI is run n steps, the range of X emerges.*

PROPOSITION 7.6. *If z_{j+1} at the $(j+1)$st step of the CF–ADI iteration is a linear combination of the previous iterates, $z_1, \ldots, z_j$, and $B \in \mathbb{R}^n$, then*

$$span\{z_1, \ldots, z_j\} = \text{range}(X).$$

REMARK 4. *If the goal is to find the range of the exact solution X, then iteration can stop when z_{j+1} is a linear combination of the previous columns. If, however, the goal is to approximate X by $Z_j^{cfadi}(Z_j^{cfadi})^T$, then iteration may have to continue, since even if $Z_j^{cfadi}(Z_j^{cfadi})^T$ has the same range as X, they may not be close as matrices.*

8. Rational Krylov subspace approximation to dominant invariant subspace. In this section we are concerned with the approximation of the dominant invariant subspace of the Lyapunov solution. In particular, we make the connection between approximating the dominant invariant subspace of the solution X and the generation of various low order Krylov and rational Krylov subspaces. As described in section 2, for some important applications it is sufficient to find the dominant invariant subspace of X. The complete knowledge of X is not necessary.

Corollary 7.2 in section 7 shows that range$(X) = \mathcal{L}(A, v_t(A, B, \mathbf{r}), \mathbf{q})\ \forall t,\ \forall \mathbf{r},\ \forall \mathbf{q}$. The range of X can also be characterized in terms of its eigenvectors. Let

$$X = [u_1, \ldots, u_n] \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \end{bmatrix} [u_1, \ldots, u_n]^T$$

be the eigenvalue (singular value) decomposition of X, with the eigenvalues ordered so that

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

Then the eigenvectors of X associated with the nonzero eigenvalues, $u_1, \ldots, u_r$, span the range of X,

$$\text{range}(X) = span\,\{u_1, \ldots, u_r\}\,. \tag{8.1}$$

Combining Corollary 7.2 and (8.1) gives spanning sets for the invariant subspace, $span\,\{u_1, \ldots, u_r\}$, of X,

$$span\{u_1, \ldots, u_r\} = span\{w_s, \ldots, w_{s+n-1}\}, \tag{8.2}$$

where w_i, $i = s, \ldots, s+n-1$, are n consecutive vectors in the infinite spanning set for $\mathcal{L}(A, v_t(A, B, \mathbf{r}), \mathbf{q})$.

It is then natural to approximate the J dimensional dominant invariant subspace of X, $span\{u_1, \ldots, u_J\}$, $J \le r \le n$, by $span\{v_1, \ldots, v_J\}$,

$$span\{u_1, \ldots, u_J\} \approx span\{v_1, \ldots, v_J\}, \tag{8.3}$$

where $\{v_1, \ldots, v_J\}$ is a subset of the order n spanning set $\{w_s, \ldots, w_{s+n-1}\}$ for some choice of s, t, $\mathbf{r}$, $\mathbf{q}$. Since only the matrix A and the vector B are given, from practical concerns the subset $\{v_1, \ldots, v_J\}$ should contain consecutive components of $\{w_s, \ldots, w_{s+n-1}\}$. Without loss of generality, we choose

$$\{v_1, \ldots, v_J\} = \{w_s, \ldots, w_{s+J-1}\}. \tag{8.4}$$

The set $\{v_1, \ldots, v_J\}$ may be generated as for an order J Krylov subspace based on the matrix A and the vector v_1 or may be generated in reverse order as for a rational Krylov subspace based on A and v_J.

A basis for any Krylov or rational Krylov subspace choice in (8.4) may be generated stably via the Arnoldi algorithm [7, 42]. The subspace $span\{u_1, \ldots, u_r\}$ will emerge in the same number of Arnoldi steps, which is at most n, for any subspace choice in (8.4). Because it is not practical to run any of these Krylov subspace-based approaches to n Arnoldi steps, we focus on the case when $J \ll n$.

A few examples of the approximation we consider in section 8.1 are $span\{v_1, \ldots, v_J\} =$

$$\text{a. } \mathcal{K}_J(A, B) = span\{B, AB, \ldots, A^{J-1}B\}, \tag{8.5}$$

$$\text{b. } \mathcal{K}_J^{rat}(A,\ A^{-1}B,\ \mathbf{0_{J-1}}) = span\{A^{-1}B, A^{-2}B, \ldots, A^{-J}B\}, \tag{8.6}$$

$$\text{c. } \mathcal{K}_J^{rat}\left(A, (A + p_1 I)^{-1}B, \{p_2, \ldots, p_J\}\right) \quad \text{for any } \{p_1, \ldots, p_J\}. \tag{8.7}$$

The choice in (8.5) was utilized in [14, 16]. If we choose the shifts $\{p_1, \ldots, p_J\}$ to be CF–ADI parameters in (8.7), we obtain the CF–ADI approximation to the dominant invariant subspace of X. Clearly, the shifts in (8.7) may be chosen in other ways. It is also possible to realize the choice in (8.7) as the direct sum of shifted rational Krylov subspaces due to Theorem 5.4.

The answer to the question of which choice in (8.5)–(8.7) best satisfies (8.3) depends on A, B, J, and the shift parameters $\{p_1, \ldots, p_J\}$. However, since there is more freedom in the choice in (8.7) than in (8.5) or (8.6), in general, one expects (8.7) to be a better choice if the shift parameters are chosen well. One answer to how to choose the shifts in (8.7) is to use the CF–ADI parameters, which are the solution of the rational minimax problem (3.7). The justification is that these parameters minimize the norm of the error $\|X - X_J^{cfadi}\|$.

8.1. Numerical results. In this section we give numerical examples of approximating the dominant invariant subspace of X by the Krylov and rational Krylov subspace choices in (8.5)–(8.7), including several natural choices of shifts in (8.7). Some preliminary numerical results on using CF–ADI to approximate the dominant invariant subspace of X can be found in [22], but the subspaces comparisons have not appeared before in literature.

The measure of the closeness of two subspaces is provided by the concept of principal angles between subspaces (see [10]).

DEFINITION 8.1. *Let S^1 and S^2 be two subspaces, of dimension d_1 and d_2, respectively, and assume $d_1 \geq d_2$. Then the d_2* principal angles *between S^1 and S^2 are $\theta_1, \ldots, \theta_{d_2}$ such that*

$$\cos(\theta_j) = \max_{u^1 \in S^1, \|u^1\|=1} \max_{u^2 \in S^2, \|u^2\|=1} (u^1)^T u^2 = (u_j^1)^T u_j^2$$

under the constraints that

$$(u^1)^T u_i^1 = 0, \quad (u^2)^T u_i^2 = 0, \quad i = 1 : j-1.$$

REMARK 5. *If the columns of U^1 are an orthonormal basis for S^1, the columns of U^2 are an orthonormal basis for S^2, and $(U^1)^T U^2$ has singular value decomposition $(U^1)^T U^2 = U \Sigma V^T$, then*

$$\cos(\theta_j) = \Sigma(j,j), \quad u_j^1 = U^1 U(:,j), \quad u_j^2 = U^2 V(:,j).$$

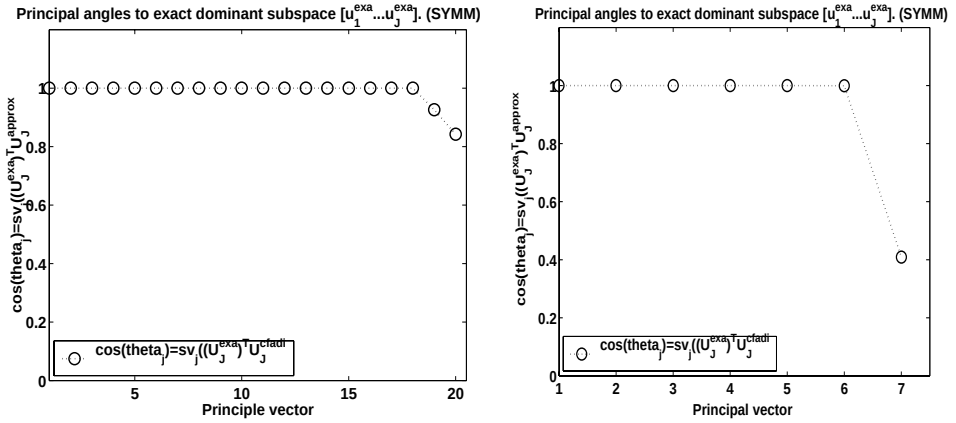
It can be seen that if $S^1 = S^2$, then $\cos(\theta_j) = 1$, $j = 1, \ldots, d_1 = d_2$, and if $S^1 \perp S^2$, then $\cos(\theta_j) = 0$, $j = 1, \ldots, d_2$.

The two bases $\{u_1^1, \ldots, u_{d_2}^1\}$ and $\{u_1^2, \ldots, u_{d_2}^2\}$ are mutually orthogonal, $(u_i^1)^T u_j^2 = 0$, if $i \neq j$. And $(u_i^1)^T u_i^2 = \cos(\theta_i)$ indicates the closeness of u_i^1 and u_i^2. A basis for the intersection of S^1 and S^2 is given by those basis vectors whose principal angle is 0. Thus, the closeness of two subspaces can be measured by how many of their principal angles are close to 0.

The example in Figure 8.1 comes from the spiral inductor problem considered in section 4.4. The matrix A is symmetric, 500×500, and B has one column. CF–ADI was run for 20 iterations and the results are shown in Figure 8.1(a). The relative error after 20 iterations is $\frac{\|X - X_j^{cfadi}\|_2}{\|X\|_2} = 10^{-8}$. The cosines of 18 of the principal angles between the exact invariant subspace and the approximate subspace are 1, and the cosines of the last 2 are above 0.8, indicating close match of all dominant eigenvectors. In contrast, Figure 8.1(b) shows the results after CF–ADI was run for only 7 iterations. The relative error $\frac{\|X - X_7^{cfadi}\|_2}{\|X\|_2}$ is 4.0×10^{-3}. However, it can be seen that the cosines of 6 principal angles are 1. Thus, dominant eigenspace information about X can emerge, even when CF–ADI has not converged.

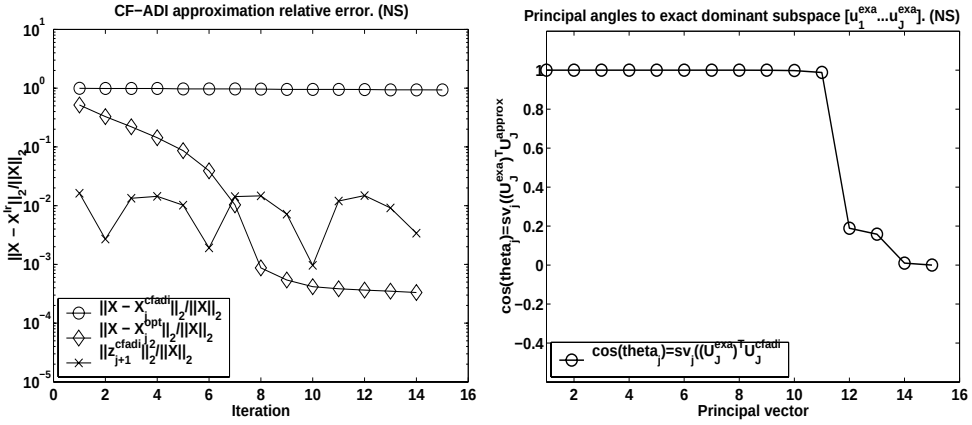
Figure 8.2 shows another example of running CF–ADI only a small number of steps, before convergence occurs. It comes from a discretized transmission line example [24]. The system matrix A is nonsymmetric, 256×256, and the input matrix B has one column. The parameter selection procedure in [40] was followed and the resulting CF–ADI parameters were complex.

Figure 8.2(a) shows that the CF–ADI error was not decreasing at all during the 15 iterations. The relative error stagnates at 1. However, Figure 8.2(b) shows that the intersection of the 15 dimensional exact dominant invariant subspace and the 15 dimensional CF–ADI approximation has dimension 10 (almost 11).



(a) 20 CF–ADI iterations (b) 7 CF–ADI iterations

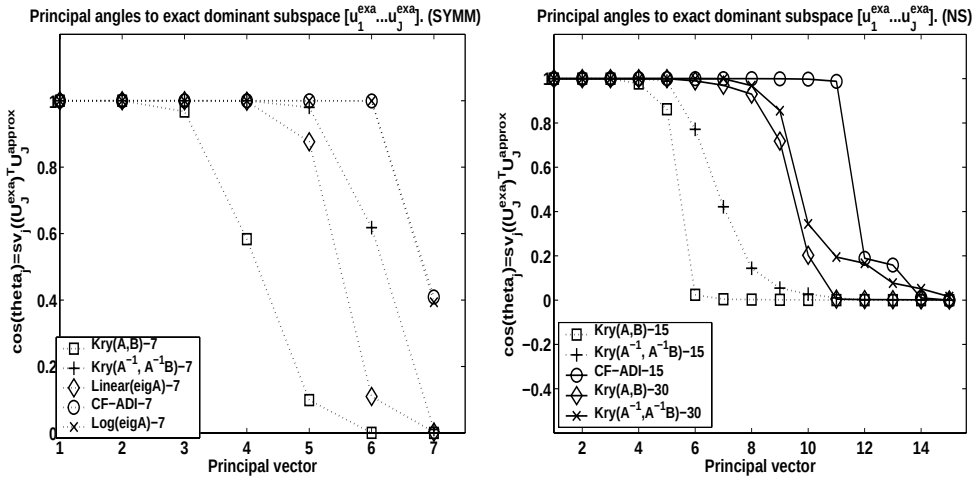
FIG. 8.1. *Symmetric matrix,* $n = 500$. *Principal angles.*



(a) Convergence (b) Principal angle

FIG. 8.2. *Nonsymmetric matrix,* $n = 256$, 15 *CF–ADI iterations, not converged.*

In Figure 8.3 we make comparison of the Krylov and rational Krylov subspace choices in (8.5)–(8.7). In Figure 8.3(a) we compare different rank 7 approximations to the exact dominant invariant subspace for the symmetric spiral inductor example. The shifted rational Krylov subspace is compared with the unshifted Krylov subspace, $\mathcal{K}_J(A, B)$, and the unshifted rational Krylov subspace, $\mathcal{K}_J(A^{-1}, A^{-1}B)$, for $J = 7$. Three choices of shift parameters for the rational Krylov subspace, $\mathcal{K}_J^{rat}(A, (A + p_1 I)^{-1}B, \{p_2, \ldots, p_J\})$, are compared. They are linearly and logarithmicaly spaced points on the eigenvalue interval of A and CF–ADI parameters from the solution of rational minimax problem (3.7). Figure 8.3(a) shows that $\mathcal{K}_7(A, B)$ provides the worst approximation. A better approximation is $\mathcal{K}_7^{rat}(A, (A + p_1 I)^{-1}B, \{p_2, \ldots, p_7\})$, with $\{p_1, \ldots, p_7\}$ linearly spaced points on the eigenvalue interval of A. A better approxi-



(a) Symmetric system

(b) Nonsymmetric system

FIG. 8.3. *Comparison of various low rank approximations to the exact dominant invariant subspace.*

mation than that is the unshifted rational Krylov subspace, $\mathcal{K}_7(A^{-1}, A^{-1}B)$. Finally, for this example, using the CF–ADI parameters and using logarithmically spaced points in $\mathcal{K}_7^{rat}(A, (A+p_1I)^{-1}B, \{p_2, \ldots, p_7\})$ both provide the best approximation.

In Figure 8.3(b) comparison is made for the nonsymmetric transmission line example. Order 15 and 30 unshifted Krylov and rational Krylov subspaces, $\mathcal{K}_J(A, B)$, $\mathcal{K}_J(A^{-1}, A^{-1}B)$, $J = 15, 30$, are compared with the order 15 shifted rational Krylov subspace, $\mathcal{K}_{J_{cfadi}}^{rat}(A, (A+p_1I)^{-1}B, \{p_2, \ldots, p_{J_{cfadi}}\})$, $J_{cfadi} = 15$, where $\{p_1, \ldots, p_{J_{cfadi}}\}$ are an approximate solution to the complex region rational minimax problem (3.7), obtained by the procedure described in [40].

Figure 8.3(b) shows that $\mathcal{K}_{15}(A, B)$ gives the worst approximation, followed by $\mathcal{K}_{15}(A^{-1}, A^{-1}B)$. Finding order 30 unshifted subspaces, $\mathcal{K}_{30}(A, B)$ and $\mathcal{K}_{30}(A^{-1}, A^{-1}B)$, to match the 15 dimensional exact dominant invariant subspace offers improvement. But clearly the order 15 subspace, $\mathcal{K}_{15}^{rat}(A, (A+p_1I)^{-1}B, \{p_2, \ldots, p_{15}\})$, using the CF–ADI parameters, gives the best approximation.

9. Conclusions. In this paper we developed the CF–ADI algorithm to generate a low rank approximation to the solution to the Lyapunov equation. CF–ADI requires only matrix-vector products and linear solves. Hence, it enables one to take advantage of any sparsity or structure in the coefficient matrix. The range of the CF–ADI approximation is a low order shifted rational Krylov subspace, where the shifts are the solution of a rational minimax problem.

We characterized the range of the solution to the Lyapunov equation as order n Krylov and rational Krylov subspaces with various starting vectors and various sets of shifts. A connection is made between the approximation of the dominant invariant subspace of the Lyapunov solution and the generation of low order Krylov and rational Krylov subspaces.

It is shown that the rational Krylov subspace generated by the CF–ADI algorithm

frequently gives the most accurate approximation to the dominant invariant subspace of the exact solution to the Lyapunov equation, which is needed in many engineering applications.

REFERENCES

[1] A. Antoulas, D. Sorensen, and Y. Zhou, *On the Decay Rate of Hankel Singular Values and Related Issues*, Technical report, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2001.

[2] R. H. Bartels and W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.

[3] A. N. Beavers, Jr. and E. D. Denman, *A new solution method for the Lyapunov matrix equation*, SIAM J. Appl. Math., 29 (1975), pp. 416–421.

[4] P. Benner and E. S. Quintana-Ortí, *Solving stable generalized Lyapunov equations with the matrix sign function*, Numer. Algorithms, 20 (1999), pp. 75–100.

[5] G. Birkhoff, R. S. Varga, and D. Young, *Alternating direction implicit methods*, in Advances in Computers, Vol. 3, Academic Press, New York, 1962, pp. 189–273.

[6] P. C. Chandrasekharan, *Robust Control of Linear Dynamical Systems*, Harcourt Brace, London, San Diego, CA, 1996.

[7] I. Elfadel and D. Ling, *A block rational Arnoldi algorithm for multipoint passive model-order reduction of multiport RLC networks*, in Proceedings of the International Conference on Computer-Aided Design, San Jose, CA, 1997, pp. 66–71.

[8] N. S. Ellner and E. L. Wachspress, *Alternating direction implicit iteration for systems with complex spectra*, SIAM J. Numer. Anal., 28 (1991), pp. 859–870.

[9] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

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

[11] E. Grimme, *Krylov Projection Methods for Model Reduction*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 1997.

[12] S. J. Hammarling, *Numerical solution of the stable, nonnegative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303–323.

[13] A. S. Hodel, B. Tenison, and K. R. Poolla, *Numerical solution of the Lyapunov equation by approximate power iteration*, Linear Algebra Appl., 236 (1996), pp. 205–230.

[14] D. Y. Hu and L. Reichel, *Krylov-subspace methods for the Sylvester equation*, Linear Algebra Appl., 172 (1992), pp. 283–313.

[15] M.-P. Istace and J.-P. Thiran, *On the third and fourth Zolotarev problems in the complex plane*, SIAM J. Numer. Anal., 32 (1995), pp. 249–259.

[16] I. M. Jaimoukha and E. M. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM J. Numer. Anal., 31 (1994), pp. 227–251.

[17] M. Kamon, F. Wang, and J. White, *Recent improvements for fast inductance extraction and simulation [packaging]*, in Proceedings of the IEEE 7th Topical Meeting on Electrical Performance of Electronic Packaging, West Point, NY, 1998, pp. 281–284.

[18] P. Lancaster, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 544–566.

[19] V. B. Larin and F. A. Aliev, *Construction of square root factor for solution of the Lyapunov matrix equation*, Systems Control Lett., 20 (1993), pp. 109–112.

[20] J.-R. Li, F. Wang, and J. White, *An efficient Lyapunov equation-based approach for generating reduced-order models of interconnect*, in Proceedings of the 36th Design Automation Conference, New Orleans, LA, 1999, pp. 1–6.

[21] J.-R. Li and J. White, *Efficient model reduction of interconnect via approximate system gramians*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, San Jose, CA, 1999, pp. 380–383.

[22] J.-R. Li and J. White, *Reduction of large circuit models via low rank approximate gramians*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 1151–1171.

[23] A. Lu and E. L. Wachspress, *Solution of Lyapunov equations by alternating direction implicit iteration*, Comput. Math. Appl., 21 (1991), pp. 43–58.

[24] N. Marques, M. Kamon, J. White, and L. Silveira, *A mixed nodal-mesh formulation for efficient extraction and passive reduced-order modeling of 3D interconnects*, in Proceedings of the 35th ACM/IEEE Design Automation Conference, San Francisco, CA, June 1998, pp. 297–302.

[25] L. Miguel Silveira, M. Kamon, I. Elfadel, and J. White, *A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits*, in Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, San Jose, CA, 1996, pp. 288–294.
[26] B. C. Moore, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.
[27] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (2000), pp. 1401–1418.
[28] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: The symmetric case*, Systems Control Lett., 40 (2000), pp. 139–144.
[29] L. Pernebo and L. M. Silverman, *Model reduction via balanced state space representations*, IEEE Trans. Automat. Control, 27 (1982), pp. 382–387.
[30] 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.
[31] A. Ruhe, *The rational Krylov algorithm for nonsymmetric eigenvalue problems.* III. *Complex shifts for real matrices*, BIT, 34 (1994), pp. 165–176.
[32] M. G. Safonov and R. Y. Chiang, *A Schur method for balanced-truncation model reduction*, IEEE Trans. Automat. Control, 34 (1989), pp. 729–733.
[33] E. D. Sontag, *Mathematical Control Theory*, 2nd ed., Springer-Verlag, New York, 1998.
[34] G. Starke, *Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1431–1445.
[35] G. Starke, *Fejér-Walsh points for rational functions and their use in the ADI iterative method*, J. Comput. Appl. Math., 46 (1993), pp. 129–141.
[36] M. S. Tombs and I. Postlethwaite, *Truncated balanced realization of a stable nonminimal state-space system*, Internat. J. Control, 46 (1987), pp. 1319–1330.
[37] E. L. Wachspress, *Optimum alternating-direction-implicit iteration parameters for a model problem*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 339–350.
[38] E. L. Wachspress, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Lett., 1 (1988), pp. 87–90.
[39] E. L. Wachspress, *ADI iterative solution of Lyapunov equations*, in Iterative Methods in Linear Algebra (Brussels, 1991), North-Holland, Amsterdam, 1992, pp. 229–231.
[40] E. L. Wachspress, *The ADI Model Problem*, Self published, Windsor, CA, 1995.
[41] O. B. Widlund, *On the rate of convergence of an alternating direction implicit method in a noncommutative case*, Math. Comp., 20 (1966), pp. 500–515.
[42] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford Sci. Publ., Oxford University Press, New York, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 281–291

QUASI-BIRTH-AND-DEATH PROCESSES WITH LEVEL-GEOMETRIC DISTRIBUTION*

TUĞRUL DAYAR† AND FRANCK QUESSETTE‡

Abstract. A special class of homogeneous continuous-time quasi-birth-and-death (QBD) Markov chains (MCs) which possess level-geometric (LG) stationary distribution is considered. Assuming that the stationary vector is partitioned by levels into subvectors, in an LG distribution all stationary subvectors beyond a finite level number are multiples of each other. Specifically, each pair of stationary subvectors that belong to consecutive levels is related by the same scalar, hence the term level-geometric. Necessary and sufficient conditions are specified for the existence of such a distribution, and the results are elaborated in three examples.

Key words. Markov chains, quasi-birth-and-death processes, geometric distributions

AMS subject classifications. 60J27, 65F50, 65H10, 65F05, 65F10, 65F15

PII. S089547980138914X

1. Introduction.

The continuous-time Markov process on the countable state space $\mathcal{S} = \{(l,i) \ : \ l \geq 0, \ 1 \leq i \leq m\}$ with block tridiagonal infinitesimal generator matrix

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{1}$$

having blocks that are $(m \times m)$ matrices is called a homogeneous continuous-time quasi-birth-and-death (QBD) Markov chain (MC). The row sums of Q are zero, meaning $(B_0 + A_0)e = 0$ and $(A_0 + A_1 + A_2)e = 0$, where e is a column vector of 1's with appropriate length. The matrices A_0 and A_2 are nonnegative, and the matrices B_0 and A_1 have nonnegative off-diagonal elements and strictly negative diagonals. The first component, l, of the state descriptor vector denotes the level and its second component, i, the phase. In homogeneous QBD MCs, the elements of B_0, A_0, A_1, and A_2 do not depend on the level number.

Neuts has done substantial work in the area of matrix analytic methods for such processes and has written two books [11], [12]. An informative resource that discusses the developments in the area since then is the recent book of Latouche and Ramaswami [9]. The most significant application area of these methods at present is the performance evaluation of communication systems. See, for instance, [13] for several case studies covering application areas from asynchronous transfer mode (ATM) networks to World Wide Web traffic and Transmission Control Protocol/Internet Protocol (TCP/IP) networking.

We assume that the homogeneous continuous-time QBD MC at hand is irreducible and positive recurrent, meaning its steady state probability distribution vector, π

*Received by the editors May 10, 2001; accepted for publication (in revised form) February 6, 2002; published electronically July 9, 2002.
http://www.siam.org/journals/simax/24-1/38914.html
†Department of Computer Engineering, Bilkent University, 06533 Bilkent, Ankara, Turkey (tugrul@cs.bilkent.edu.tr).
‡Lab. PRiSM, Université de Versailles, 45 Avenue des États-Unis, 78035 Versailles Cedex, France (qst@prism.uvsq.fr).

(see [14]), exists. Recall that an MC is said to be positive recurrent if the mean time to return to each state for the first time after leaving it is finite [14, p. 9]. In infinite QBD MCs, this requires that the drift to higher level states be smaller than the drift to lower level states [5, pp. 153–154]. Throughout the paper, we adhere to the convention that probability vectors are row vectors. Being a stationary distribution, π satisfies $\pi Q = 0$ and $\pi e = 1$. Now, let π be partitioned by levels into subvectors π_l, $l \geq 0$, where π_l is of length m. Then π also satisfies the matrix-geometric property [9, p. 142]

$$\pi_{l+1} = \pi_l R \quad \text{for} \quad l \geq 0, \tag{2}$$

where the matrix R of order m records the rate of visit to level $(l+1)$ per unit of time spent in level l. Fortunately, the elements of R for homogeneous QBD MCs do not depend on the level number. Quadratically convergent algorithms for solving QBD MCs appear in [8], [4], [1].

In this paper, we consider a special class of homogeneous continuous-time QBD MCs which possess what we call level-geometric (LG) stationary distribution. To the best of our knowledge, this property has not been explicitly defined before, and hence our "level-geometric" designation. An LG distribution is one that satisfies

$$\pi_{l+1} = \alpha \pi_l \quad \text{for} \quad l \geq L, \tag{3}$$

where $\alpha \in (0, 1)$ and L is a finite nonnegative integer. Note that an LG distribution with $L = 0$ is a product-form solution. An LG distribution can be expressed alternatively as

$$\pi_{L+k} = (1 - \alpha)\alpha^k a \quad \text{for} \quad k \geq 0, \tag{4}$$

where a is a positive probability vector of length m, with $ae = 1$ when $L = 0$. In an LG distribution, the level is independent of the phase for level numbers greater than or equal to L, and the marginal probability distribution of the levels are given by $\pi_{L+k}e = (1 - \alpha)\alpha^k ae$ [9, pp. 295–299] for $k \geq 0$. Throughout the paper, we refer to an LG distribution for which L is the smallest possible nonnegative integer that satisfies (3) as an LG distribution with parameter L. Our motivation is to come up with a solution method for this special class of QBD MCs that does not require R to be computed. We remark that if S_ϵ is the number of iterations required to reach an accuracy of ϵ by the successive substitution algorithm [5, p. 160], then the computation of R with quadratically convergent algorithms takes about $O(\log_2 S_\epsilon)$ iterations (hence, the term quadratically convergent), each of which has a time complexity of $O(m^3)$ floating-point operations. The results that we develop can be extended to the homogeneous discrete-time case without difficulty.

In section 2, we provide background information on the solution of QBD MCs with special structure. In section 3, we give three examples of QBD MCs with LG stationary distribution. In section 4, we specify conditions related to such a distribution and show how it can be computed when it exists. In section 5, we reconsider the three examples of section 3 in light of the new results introduced in section 4. We conclude in section 6.

2. Background material. In this section, an overview of some concepts discussed in [9] and relevant propositions are given. Wherever something has been taken from [9], the appropriate reference to the corresponding page(s) is placed.

Due to the fixed pattern of transitions among levels and within each level, it is not difficult to check the irreducibility of Q. The next proposition is about checking the positive recurrence of Q when Q and $A = A_0 + A_1 + A_2$ are both irreducible. When Q is irreducible but A has multiple irreducible classes, one can resort to the theorem in [9, p. 160]. Note that A is an infinitesimal generator matrix.

PROPOSITION 1. *If Q and A are irreducible, then Q is positive recurrent if and only if $\pi_A(A_0 - A_2)e < 0$, where π_A satisfies $\pi_A A = 0$ and $\pi_A e = 1$* [9, p. 158].

Throughout this paper, we assume that the homogeneous continuous-time QBD MC at hand is irreducible and positive recurrent. Now, let $\rho(R)$ denote the spectral radius of R (i.e., $\rho(R) = \max\{|\lambda| \mid \lambda \in \lambda(R)\}$, where $\lambda(R) = \{\lambda \mid Rv = \lambda v, v \neq 0\}$ is its spectrum). Then, $\rho(R) < 1$ [9, p. 133].

The next proposition specifies necessary and sufficient conditions for the existence of an LG distribution with parameter $L = 0$.

PROPOSITION 2. *The stationary distribution of Q is LG with parameter $L = 0$ if and only if there exists a positive vector a with $ae = 1$ and a positive scalar $\alpha = \rho(R)$ with $\alpha < 1$ such that $a(A_0 + \alpha A_1 + \alpha^2 A_2) = 0$ and $a(B_0 + \alpha A_2) = 0$* [9, pp. 297–298].

This proposition, although very concise and to the point, has two shortcomings. First, it does not indicate how to check for an LG distribution with parameter $L \geq 1$. Second, it requires the solution of a nonlinear system of equations.

The following two propositions indicate the improvement that is obtained in the solution when A_2 and/or A_0 are rank-1 matrices.

PROPOSITION 3. *When A_2 is of rank-1, then $R = -A_0(A_1 + A_0 e b^T)^{-1}$, where $A_2 = cb^T$ and $b^T e = 1$* [9, p. 197]. *Furthermore, π_0 can be computed up to a multiplicative constant using $\pi_0(B_0 + A_0 e b^T) = 0$* [9, p. 236].

Hence, it is relatively simple to compute the stationary distribution when A_2 is of rank-1.

PROPOSITION 4. *When A_0 is of rank-1, then $R = c\xi^T$, where $A_0 = cb^T$, $b^T e = 1$, $\xi^T = -b^T(A_1 + \alpha A_2)^{-1}$, and $\alpha = \xi^T c$ with $\alpha = \rho(R)$* [9, p. 198]. *The stationary subvectors satisfy $\pi_0 = \pi_1 C_0$, where $C_0 = -A_2 B_0^{-1}$, and $\pi_l = \pi_{l+1} C_1$ for $l \geq 1$, where $C_1 = -A_2(A_1 + A_2 e b^T)^{-1}$* [9, p. 236].

COROLLARY 1. *When A_0 is of rank-1, then R is also of rank-1, and $R^2 = \alpha R$ thereby implies $\pi_{l+1} = \alpha \pi_l$ for $l \geq 1$. Hence, Q has an LG distribution with parameter $L \leq 1$.*

The next section elaborates these results with three examples.

3. Examples. The following examples all have LG distributions, and they aid in understanding the concepts introduced in section 2 and the concepts to be developed in section 4. In order to compactly describe single queueing stations, we use the so-called Kendall notation, which consists of six identifiers separated by vertical bars [5, pp. 13–14]:

Arrivals|Services|Servers|Buffersize|Population|Scheduling.

Here Arrivals and Services, respectively, characterize the customer arrival and service processes by specifying the interarrival and interservice distributions. For these distributions there are various possibilities, among which are M (i.e., Markovian) for exponential and E_k for k-phase Erlang. Servers gives the number of service-providing entities; Buffersize gives the maximum number of customers in the queueing station, including any in service; Population gives the size of the customer population from which the arrivals are taking place; and Scheduling specifies the employed scheduling strategy. When the Buffersize and/or the Population are omitted, they are assumed

to be infinitely large. When the scheduling strategy is omitted, it is assumed to be first come, first served (FCFS).

3.1. Example 1. The first example we consider is a system of two independent queues, where queue 1 is M|M|1 and queue 2 is M|M|1|$m-1$. Queue $i \in \{1,2\}$ has a Poisson arrival process with rate λ_i and an exponential service distribution with rate μ_i. This system corresponds to a QBD process with the level representing the length of queue 1, which is unbounded, and the phase representing the length of queue 2, which can range between 0 and $(m-1)$. We assume $\lambda_1 < \mu_1$. Letting $d = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$, we have $A_0 = \lambda_1 I$, $A_2 = \mu_1 I$,

$$A_1 = \begin{pmatrix} -(d-\mu_2) & \lambda_2 & & & & \\ \mu_2 & -d & \lambda_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mu_2 & -d & \lambda_2 & \\ & & & \mu_2 & -(d-\lambda_2) & \end{pmatrix},$$

and

$$B_0 = \begin{pmatrix} -(\lambda_1+\lambda_2) & \lambda_2 & & & & \\ \mu_2 & -(d-\mu_1) & \lambda_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mu_2 & -(d-\mu_1) & \lambda_2 & \\ & & & \mu_2 & -(\lambda_1+\mu_2) & \end{pmatrix}.$$

Q is irreducible, and from Proposition 1 we have

$$A = A_0 + A_1 + A_2 = \begin{pmatrix} -\lambda_2 & \lambda_2 & & & \\ \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 \\ & & & \mu_2 & -\mu_2 \end{pmatrix},$$

which is irreducible, and π_A is the truncated geometric distribution with parameter λ_2/μ_2 [5, p. 84]. Hence, $\pi_A(A_0 - A_2)e = \lambda_1 - \mu_1 < 0$ and Q is positive recurrent. For this example, $\alpha = \lambda_1/\mu_1$, $a_k = \nu^k(1-\nu)/(1-\nu^m)$, $0 \leq k \leq m-1$, and $L = 0$, where $\nu = \lambda_2/\mu_2$, turn out to be the parameters in (4) that specify an LG distribution.

Recalling that an MC is said to be lumpable with respect to a given partitioning if each block in the partitioning has equal row sums [7, p. 124], we remark that the QBD MC in this example is lumpable, and the lumped chain represents queue 1.

3.2. Example 2. The second example we consider is the continuous-time equivalent of the discrete-time QBD process discussed in [8, pp. 668–669]. The model has 2 phases at each level (i.e., $m = 2$). Assuming that $0 < p < 1$, the process moves from state $(l,1)$, $l \geq 1$, to $(l,2)$ with rate p, and to $(l-1,1)$ with rate $(1-p)$. The process moves from state $(l,2)$, $l \geq 0$, to $(l,1)$ with rate $2p$, and to $(l+1,2)$ with rate $(1-2p)$. Finally, the process moves from state $(0,1)$ to $(0,2)$ with rate 1. All diagonal elements of Q are -1. Hence, we have

$$A_0 = \begin{pmatrix} 0 & 0 \\ 0 & 1-2p \end{pmatrix},\ A_1 = \begin{pmatrix} -1 & p \\ 2p & -1 \end{pmatrix},\ A_2 = \begin{pmatrix} 1-p & 0 \\ 0 & 0 \end{pmatrix},\ B_0 = \begin{pmatrix} -1 & 1 \\ 2p & -1 \end{pmatrix}.$$

Q is irreducible, and from Proposition 1 we have

$$A = A_0 + A_1 + A_2 = \begin{pmatrix} -p & p \\ 2p & -2p \end{pmatrix},$$

which is irreducible, and $\pi_A = (2/3\ 1/3)$. Hence, $\pi_A(A_0 - A_2)e = -1/3 < 0$ and Q is positive recurrent. For this example, $\alpha = (1-2p)/(1-p)$, $a = (1/2\ 1/2)$, and $L = 0$ turn out to be the parameters in (4) that specify an LG distribution. Direct substitution in $\pi Q = 0$ and $\pi e = 1$ confirms this solution.

In this example, Proposition 3 applies with $c = (1-p)e_1$ and $b = e_1$, where e_i is the ith principal axis vector. Hence, $R = (1-2p)e_2^T e/(1-p)$, and $\rho(R) = \alpha$ as expected. Furthermore, $\pi_0 = (1-\alpha)(1/2\ 1/2)$. Note that in this example, Proposition 4 applies as well. The rate matrix is of rank-1 and $\xi = e/(1-p)$. In section 5, we will argue why this example has an LG distribution with parameter $L = 0$ and not $L = 1$. Finally, we remark that this example is also used as a test case in [1].

3.3. Example 3. The third example we consider is the $E_m|M|1$ FCFS queue which has an exponential service distribution with rate μ and an m-phase Erlang arrival process with rate $m\lambda$ in each phase [9, pp. 206–208]. The expected interarrival time and the expected service time of this queue are, respectively, $1/\lambda$ and $1/\mu$. We assume $\lambda < \mu$. The queue corresponds to a QBD process with the level representing the queue length (including any in service) and the phase representing the state of the Erlang arrival process. Letting $d = m\lambda + \mu$, we have the $(m \times m)$ matrices $A_0 = m\lambda e_m e_1^T$, $A_2 = \mu I$,

$$A_1 = \begin{pmatrix} -d & m\lambda & & \\ & \ddots & \ddots & \\ & & -d & m\lambda \\ & & & -d \end{pmatrix}, \quad B_0 = \begin{pmatrix} -m\lambda & m\lambda & & \\ & \ddots & \ddots & \\ & & -m\lambda & m\lambda \\ & & & -m\lambda \end{pmatrix}.$$

Q is irreducible, and from Proposition 1 we have

$$A = A_0 + A_1 + A_2 = \begin{pmatrix} -m\lambda & m\lambda & & \\ & \ddots & \ddots & \\ & & -m\lambda & m\lambda \\ m\lambda & & & -m\lambda \end{pmatrix},$$

which is irreducible, and $\pi_A = e^T/m$. Hence, $\pi_A(A_0 - A_2)e = \lambda - \mu < 0$ and Q is positive recurrent. Although the $E_m|M|1$ queue does not have an explicit solution, it can be shown by following the formulae in [6, p. 323] that its stationary distribution has an LG distribution with parameter $L = 1$.

In this example, Proposition 4 applies with $c = m\lambda e_m$ and $b = e_1$, implying R is of rank-1, $C_0 = -A_2 B_0^{-1}$, and $C_1 = -A_2(A_1 + \mu e e_1^T)^{-1}$.

The next section builds on the results in section 2 with the aim of coming up with a solution method to compute an LG distribution when it exists.

4. Checking for and computing the LG distribution. The assumption of irreducibility of Q implies that the nonnegative matrix A_0 has at least one positive row sum (see (1)). Since we also have $(B_0 + A_0)e = 0$, it must be that B_0 has nonpositive row sums with at least one negative row sum. Together with the fact that B_0 has nonnegative off-diagonal elements and a strictly negative diagonal, this implies that $-B_0$ is a nonsingular M-matrix and $-B_0^{-1} \geq 0$; see [3].

The next proposition is essential in formulating the results in this section.

PROPOSITION 5. *The sequence of matrices* $D_{l+1} = A_1 - A_2 D_l^{-1} A_0$, $l \geq 0$, *where* $D_0 = B_0$, *is well defined. For* $l \geq 0$, $-D_l$ *is a nonsingular M-matrix,* $-D_l^{-1} \geq 0$, *and* D_l^T *denotes the diagonal block at level* l *after* l *steps of block Gaussian elimination (GE) on* Q^T. *Furthermore,* $\pi_l = \pi_{l+1} C_l$, *where* $C_l = -A_2 D_l^{-1} \geq 0$ *for* $l \geq 0$.

Proof. Since $-D_0$ is a nonsingular M-matrix, let us show that $-D_1$ is too. It is possible to construct the infinitesimal generator

$$\bar{Q} = \begin{pmatrix} D_0 & A_0 & 0 \\ A_2 & A_1 & s \\ 0 & r^T & \delta \end{pmatrix}$$

so that it is irreducible. Here $s = A_0 e$, r is any nonnegative vector that ensures the irreducibility of $\bar{Q}$, and $\delta = -r^T e$. Now let $X = -\bar{Q}$ and consider the partitioning

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} = \left(\begin{array}{c|cc} -D_0 & -A_0 & 0 \\ \hline -A_2 & -A_1 & -s \\ 0 & -r^T & -\delta \end{array} \right).$$

The negated infinitesimal generator X is an irreducible singular M-matrix [3] by its definition. Therefore, the Schur complement [10, p. 123] S of X_{11}, which is given by

$$S = X_{22} - X_{21} X_{11}^{-1} X_{12} = \begin{pmatrix} -A_1 + A_2 D_0^{-1} A_0 & -s \\ -r^T & -\delta \end{pmatrix},$$

is an irreducible singular M-matrix (see Lemma 1 in [2]). All principal submatrices of an irreducible singular M-matrix except itself are nonsingular M-matrices [3, p. 156]. Hence, $-A_1 + A_2 D_0^{-1} A_0$; that is, $-D_1$ is a nonsingular M-matrix and $-D_1^{-1} \geq 0$. One can similarly show that $-D_l$ is a nonsingular M-matrix and $-D_l^{-1} \geq 0$ for $l > 1$.

Since Q^T is a block tridiagonal matrix, block GE on $Q^T \pi^T = 0$ yields $Z^T \pi^T = 0$ (or equivalently $\pi Z = 0$), where

$$Z = \begin{pmatrix} D_0 & & & \\ A_2 & D_1 & & \\ & A_2 & D_2 & \\ & & \ddots & \ddots \end{pmatrix}, \tag{5}$$

$D_0 = B_0$, and $D_{l+1} = A_1 - A_2 D_l^{-1} A_0$ for $l \geq 0$.

Recalling that $\pi = (\pi_0, \pi_1, \ldots)$ and using $\pi Z = 0$, we obtain $\pi_l D_l + \pi_{l+1} A_2 = 0$, which implies $\pi_l = -\pi_{l+1} A_2 D_l^{-1}$ for $l \geq 0$. That $C_l \geq 0$ for $l \geq 0$ follows from $-D_l^{-1} \geq 0$ and $A_2 \geq 0$. ∎

4.1. Checking for the LG distribution. The form of Z in (5) together with Proposition 5 suggests the next lemma.

LEMMA 1. *If* $D_{L+1} = D_L$ *for some finite nonnegative integer* L, *then* $D_l = D_L$ *for* $l > L + 1$, *and* $\pi_L = \pi_{L+k} C_L^k$ *for* $k \geq 0$.

Proof. From Proposition 5 we have $D_{L+1} = A_1 - A_2 D_L^{-1} A_0$ and $D_{L+2} = A_1 - A_2 D_{L+1}^{-1} A_0$. If $D_{L+1} = D_L$, then $D_{L+2} = A_1 - A_2 D_L^{-1} A_0 = D_{L+1} = D_L$. The same argument may be used to show that $D_l = D_L$ for $l > L + 2$. The second part of the lemma follows from its first part and the last part of Proposition 5. ∎

The next theorem states a condition under which one has an LG distribution.

THEOREM 1. *Let L be the smallest finite nonnegative integer for which $D_{L+1} = D_L$. Then the stationary distribution of Q is LG with parameter less than or equal to L.*

Proof. From Lemma 1 and (5), when $D_{L+1} = D_L$, we have

$$Z = \begin{pmatrix} D_0 & & & & \\ A_2 & D_1 & & & \\ & \ddots & \ddots & & \\ & & A_2 & D_{L-1} & \\ & & & Y_L & Z_L \end{pmatrix}, \tag{6}$$

where

$$Y_L = \begin{pmatrix} A_2 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \quad \text{and} \quad Z_L = \begin{pmatrix} D_L & & & \\ A_2 & D_L & & \\ & A_2 & D_L & \\ & & \ddots & \ddots \end{pmatrix}.$$

Since π_l of length m is positive for finite l and unique up to a multiplicative constant with $\lim_{l\to\infty} \pi_l = 0$, the identities $(\pi_L, \pi_{L+1}, \ldots)Z_L = 0$ and $(\pi_{L+1}, \pi_{L+2}, \ldots)Z_L = 0$ obtained from equations $\pi Z = 0$ and (6) together with the recursive structure of Z_L given by

$$Z_L = \begin{pmatrix} D_L & \\ Y_L & Z_L \end{pmatrix}$$

suggest that $\pi_{l+1} = \alpha\pi_l$ for $l \geq L$, where $\alpha \in (0, 1)$. ∎

COROLLARY 2. *When $B_0 = A_1 - A_2B_0^{-1}A_0$, the stationary distribution of Q is LG with parameter $L = 0$.*

Next we state two lemmas, which will be used in checking for an LG distribution.

LEMMA 2. *If A_1 is irreducible and $A_2e > 0$, then D_l is irreducible and $C_l > 0$ for $l \geq 1$.*

Proof. From Proposition 5 we have $D_{l+1} = A_1 + C_lA_0$, where $C_l = -A_2D_l^{-1} \geq 0$ and $l \geq 0$. Since $A_0 \geq 0$ by definition, we obtain $C_lA_0 \geq 0$. Besides, A_1 has nonnegative off-diagonal elements and is assumed to be irreducible. Hence, its sum with the nonnegative C_lA_0 will not change the irreducibility, thereby implying irreducible D_{l+1} for $l \geq 0$. Alternatively, D_l, $l \geq 1$, is irreducible. That $-D_l$ is a nonsingular M-matrix from Proposition 5, together with the fact it is irreducible, implies $-D_l^{-1} > 0$ for $l \geq 1$ [3, p. 141]. Since $A_2 \geq 0$ and is assumed to have a nonzero in each row, its product with $-D_l^{-1}$ is positive. Hence, $C_l > 0$ for $l \geq 1$. ∎

LEMMA 3. *If $e^TA_0 > 0$, $A_2e > 0$, and D_L is irreducible for some finite nonnegative integer L, then D_l is irreducible and $C_l > 0$ for $l \geq L$.*

Proof. When D_L is irreducible and A_2 has a nonzero in each row, we have $C_L > 0$ as in the proof of Lemma 1. Since $A_0 \geq 0$ and is assumed to have a nonzero in each column, we have $C_LA_0 > 0$, thereby implying an irreducible D_{L+1}. The same circle of arguments may be used to show that $C_l > 0$ and D_{l+1} is irreducible for $l > L$. ∎

The next theorem states another condition under which one has an LG distribution.

THEOREM 2. *Let L be the smallest finite nonnegative integer for which C_l is irreducible and $\rho(C_l) = \rho(C_{l+1})$, where $l \geq L$. Then the stationary distribution of Q is LG with parameter L.*

Proof. From Proposition 5 we have $C_l \geq 0$ for $l \geq 0$. If C_l, $l \geq L$, is irreducible, then by the Perron–Frobenius theorem C_l has $\rho(C_l) > 0$ as a simple eigenvalue and a corresponding positive left-hand eigenvector. There are no other linearly independent positive left-hand eigenvectors of C_l [10, p. 673]. From Proposition 5 we also have $\pi_l = \pi_{l+1}C_l$ and $\pi_l > 0$ with $\lim_{l\to\infty} \pi_l = 0$. Multiplying both sides of $\pi_l = \pi_{l+1}C_l$ by $\rho(C_l)$, we obtain $\rho(C_l)\pi_l = (\rho(C_l)\pi_{l+1})C_l$. Since $\rho(C_l)$ is a simple eigenvalue of C_l for $l \geq L$, we must have π_l as its corresponding positive left-hand eigenvector. Therefore, it must also be that $\pi_l = \rho(C_l)\pi_{l+1}$ for $l \geq L$. Since $\rho(C_l) = \rho(C_{l+1})$ for $l \geq L$, we have $\pi_l = \rho(C_L)\pi_{l+1}$, or $\pi_{l+1} = (1/\rho(C_L))\pi_l$ for $l \geq L$. Consequently, Q has an LG distribution with parameter L. ∎

4.2. Computing the LG distribution. The next theorem gives the value of α in (3) and indicates how π_L can be computed up to a multiplicative constant when one has an LG distribution with parameter L.

THEOREM 3. *If the stationary distribution of Q is LG with parameter L, then $\rho(C_L)\pi_L = \pi_L C_L$, where $\alpha = 1/\rho(C_L)$ and $\pi_L > 0$ in (3).*

Proof. Since Q has an LG distribution with parameter L, from (3) we have $\pi_{L+1} = \alpha\pi_L$, where $\alpha \in (0, 1)$, and $\pi_L > 0$ and $\pi_{L+1} > 0$ with $\lim_{l\to\infty} \pi_l = 0$. That is, for finite L, π_{L+1} is a positive multiple of π_L. Furthermore, from Proposition 5 we have $\pi_L = \pi_{L+1}C_L$, where $C_L \geq 0$. Since π_{L+1} is a positive multiple of π_L, π_L is clearly a positive left-hand eigenvector of C_L and therefore corresponds to the eigenvalue $\rho(C_L)$ [3, p. 28]. Combining the two statements, we obtain $\rho(C_L)\pi_L = \pi_L C_L$, where $\alpha = 1/\rho(C_L)$ and $\pi_L > 0$. ∎

COROLLARY 3. *When the stationary distribution of Q is LG with parameter less than or equal to L, where $L > 0$, if $\rho(C_L) \neq \rho(C_{L-1})$, then the parameter is L; otherwise the parameter is less than or equal to $L - 1$.*

5. Examples revisited. In this section, we demonstrate the results of the previous section using the three examples introduced in section 3.

5.1. Example 1. For the first example in section 2, D_l^{-1}, $l \geq 0$, is a full matrix, and we have experimentally shown that $D_{l+1} = D_l$ as l approaches infinity. For the particular case of $m = 2$, we have

$$B_0^{-1} = \frac{-1}{\lambda_1(d-\mu_1)} \begin{pmatrix} \lambda_1+\mu_2 & \lambda_2 \\ \mu_2 & \lambda_1+\lambda_2 \end{pmatrix} \quad \text{and} \quad C_0 = -A_2B_0^{-1} = -\mu_1 B_0^{-1},$$

where $d = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$. The correction to A_1 is given by $C_0A_0 = -\lambda_1\mu_1 B_0^{-1}$, and therefore

$$D_1 = A_1 + C_0A_0 = \begin{pmatrix} -(d-\mu_2) + \frac{\mu_1(\lambda_1+\mu_2)}{d-\mu_1} & \lambda_2 + \frac{\lambda_2\mu_1}{d-\mu_1} \\ \mu_2 + \frac{\mu_1\mu_2}{d-\mu_1} & -(d-\lambda_2) + \frac{\mu_1(\lambda_1+\lambda_2)}{d-\mu_1} \end{pmatrix} \neq B_0.$$

In a similar manner one can show that $D_{l+1} \neq D_l$ for finite values of l. Hence, Theorem 1 does not apply. However, Lemma 3 applies since A_0 and A_2 are of full-rank and D_0 is irreducible, implying irreducible C_l for $l \geq 0$. Consequently, there is reason to guess that the QBD MC has an LG distribution with parameter $L = 0$ from Theorem 2 and to compute the eigenvalue-eigenvector pair $(\rho(C_0),\pi_0)$ using

Theorem 3. Then the guessed solution can be verified in $\pi Q = 0$. Although this approach will sometimes fail, it works in Example 1 and can be recommended for small values of L.

For $m = 2$, it is not difficult to find, using Theorem 3, that $\rho(C_0) = \mu_1/\lambda_1 > 1$, implying $\alpha = \lambda_1/\mu_1$, and

$$\pi_0 = (1-\alpha)\begin{pmatrix} \dfrac{1-\nu}{1-\nu^2} & \dfrac{\nu(1-\nu)}{1-\nu^2} \end{pmatrix},$$

where $\nu = \lambda_2/\mu_2$.

5.2. Example 2. Consider the second example in section 2, for which we have

$$B_0^{-1} = \frac{-1}{1-2p}\begin{pmatrix} 1 & 1 \\ 2p & 1 \end{pmatrix} \quad \text{and} \quad C_0 = -A_2 B_0^{-1} = \frac{1-p}{1-2p}\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Note that C_0 is reducible. The correction to A_1 is given by $C_0 A_0 = (1-p)e_1 e_2^T$, and therefore

$$D_1 = A_1 + C_0 A_0 = \begin{pmatrix} -1 & 1 \\ 2p & -1 \end{pmatrix} = B_0.$$

Hence, in this example, $D_l = D_0$ for $l \geq 1$ from Lemma 1 due to $D_1 = D_0$. From Corollary 2 we conclude that Example 2 has an LG distribution with parameter $L = 0$.

Finally, from Theorem 3 we obtain $\rho(C_0) = (1-p)/(1-2p) > 1$, implying $\alpha = (1-2p)/(1-p)$, and $\pi_0 = (1-\alpha)(1/2\ 1/2)$.

5.3. Example 3. Now consider the third example in section 3, for which we have

$$B_0^{-1} = \frac{-1}{m\lambda}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ & 1 & \cdots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \quad \text{and} \quad C_0 = -A_2 B_0^{-1} = \frac{\mu}{m\lambda}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ & 1 & \cdots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}.$$

Note that C_0 is reducible and $\rho(C_0) = \mu/(m\lambda)$, which is not necessarily greater than 1. The correction to A_1 is given by $C_0 A_0 = \mu e e_1^T$, and therefore

$$D_1 = \begin{pmatrix} -m\lambda & m\lambda & & & \\ \mu & -(m\lambda+\mu) & m\lambda & & \\ \vdots & & \ddots & \ddots & \\ \mu & & & -(m\lambda+\mu) & m\lambda \\ \mu & & & & -(m\lambda+\mu) \end{pmatrix} \neq B_0.$$

Noticing that $D_1 = A_1 + \mu e e_1^T$, in which the correction $\mu e e_1^T$ is of rank-1, the Sherman–Morrison formula [10, p. 124] yields

$$D_1^{-1} = A_1^{-1} - \mu \frac{A_1^{-1} e e_1^T A_1^{-1}}{1 + \mu e_1^T A_1^{-1} e}.$$

Letting $\gamma = m\lambda/(m\lambda+\mu)$, we obtain

$$A_1^{-1} = \frac{-1}{m\lambda+\mu}\begin{pmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{m-1} \\ & 1 & \gamma & \cdots & \gamma^{m-2} \\ & & \ddots & \ddots & \vdots \\ & & & 1 & \gamma \\ & & & & 1 \end{pmatrix}, \quad (1+\mu e_1^T A_1^{-1} e) = \gamma^m,$$

$$\mu(A_1^{-1}e)(e_1^T A_1^{-1}) = \frac{1}{m\lambda+\mu}\begin{pmatrix} 1-\gamma^m & \gamma(1-\gamma^m) & \cdots & \gamma^{m-1}(1-\gamma^m) \\ 1-\gamma^{m-1} & \gamma(1-\gamma^{m-1}) & \cdots & \gamma^{m-1}(1-\gamma^{m-1}) \\ \vdots & \vdots & \ddots & \vdots \\ 1-\gamma & \gamma(1-\gamma) & \cdots & \gamma^{m-1}(1-\gamma) \end{pmatrix},$$

and, after some algebra, $C_1A_0 = \mu e e_1^T$. Hence, $D_2 = A_1 + C_1A_0 = D_1$, implying $D_l = D_1$ for $l \geq 2$ from Lemma 1. From Theorem 1 we have an LG distribution with parameter $L \leq 1$. We also remark that the two matrices C_0 and C_1 introduced in Proposition 4 for QBD processes with rank-1 A_0 matrices are given in this example as $C_0 = -\mu D_0^{-1}$ and $C_1 = -\mu D_1^{-1}$. Since $\rho(C_0)$ may be less than 1 and therefore different than $\rho(C_1)$, from Corollary 3 we conclude Example 2 has an LG distribution with parameter $L = 1$.

Regarding the computation of α, for instance, when $m = 2$

$$C_0 = \eta\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad C_1 = \eta\begin{pmatrix} 1+\eta & 1 \\ \eta & 1 \end{pmatrix},$$

where $\eta = \mu/(2\lambda)$. Hence, we have

$$\rho(C_1) = \eta\left(1 + \frac{1}{2}\eta + \sqrt{\eta\left(1+\frac{1}{4}\eta\right)}\right).$$

Note that $\rho(C_0) \neq \rho(C_1)$. Now, using $\rho(C_1)\pi_1 = \pi_1 C_1$, $\pi_0 = \pi_1 C_0$, and $\pi_1 e/(1-\alpha) + \pi_0 e = 1$, where $\alpha = 1/\rho(C_1)$, we obtain

$$\pi_1 = \left(\frac{(\rho(C_1)-\eta)(\rho(C_1)-1)}{\rho^2(C_1)+\eta(\rho(C_1)-1)(2\rho(C_1)-\eta)} \;\; \frac{\eta(\rho(C_1)-1)}{\rho^2(C_1)+\eta(\rho(C_1)-1)(2\rho(C_1)-\eta)}\right)$$

and

$$\pi_0 = \left(\frac{\eta(\rho(C_1)-\eta)(\rho(C_1)-1)}{\rho^2(C_1)+\eta(\rho(C_1)-1)(2\rho(C_1)-\eta)} \;\; \frac{\eta\rho(C_1)(\rho(C_1)-1)}{\rho^2(C_1)+\eta(\rho(C_1)-1)(2\rho(C_1)-\eta)}\right).$$

Normally the computation would be performed numerically for the given parameters of the problem. For $m \geq 3$, we would first compute C_0 and C_1. Then we would obtain the eigenvalue-eigenvector pair $(\rho(C_1), \pi_1)$ from $\rho(C_1)\pi_1 = \pi_1 C_1$ (see Theorem 3). Next we would compute $\pi_0 = \pi_1 C_0$. Finally we would normalize π_0 and π_1 with $\pi_1 e/(1-\alpha) + \pi_0 e$.

6. Conclusion. This paper introduces necessary and sufficient conditions for a homogeneous continuous-time quasi-birth-and-death (QBD) Markov chain (MC) to possess level-geometric (LG) stationary distribution. Furthermore, it discusses how an LG distribution can be computed when it exists. Results that utilize the matrices

A_0, A_1, A_2, and B_0 are given, showing how one can easily check for and compute an LG distribution with parameter $L \leq 1$. The results are elaborated through three examples. Examples 2 and 3, which have been used in the literature as test cases, are shown to possess LG distributions, respectively, with parameters $L = 0$ and $L = 1$. Since the matrices A_0, A_1, A_2, and B_0 that arise in applications are usually sparse, the results developed in this paper may be used before resorting to quadratically convergent algorithms to compute the rate matrix, R.

Acknowledgments. We thank the anonymous referees and Reinhard Nabben for their remarks, which led to an improved manuscript.

REFERENCES

[1] N. Akar and K. Sohraby, *An invariant subspace approach in M/G/1 and G/M/1 type Markov chains*, Comm. Statist. Stochastic Models, 13 (1997), pp. 381–416.
[2] M. Benzi and M. Tůma, *A parallel solver for large-scale Markov chains*, Appl. Numer. Math., 41 (2002), pp. 135–153.
[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[4] D. Bini and B. Meini, *On the solution of a nonlinear matrix equation arising in queueing problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 906–926.
[5] B. R. Haverkort, *Performance of Computer Communication Systems: A Model-Based Approach*, John Wiley & Sons, Chichester, England, 1998.
[6] K. Kant, *Introduction to Computer System Performance Evaluation*, McGraw-Hill, New York, 1992.
[7] J. R. Kemeny and J. L. Snell, *Finite Markov Chains*, Van Nostrand, New York, 1960.
[8] G. Latouche and V. Ramaswami, *A logarithmic reduction algorithm for quasi-birth-and-processes*, J. Appl. Probab., 30 (1993), pp. 650–674.
[9] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, SIAM, Philadelphia, 1999.
[10] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
[11] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models. An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, MD, 1981.
[12] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York, 1989.
[13] A. Ost, *Performance of Communication Systems: A Model-Based Approach with Matrix-Geometric Methods*, Springer-Verlag, Berlin, 2001.
[14] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 1, pp. 292–301

PRODUCT TRIANGULAR SYSTEMS WITH SHIFT*

CARLA D. MORAVITZ MARTIN† AND CHARLES F. VAN LOAN‡

Abstract. Systems of the form $(R^{(1)} \cdots R^{(p)} - \lambda I)x = b$, where each $R^{(i)}$ is an n-by-n upper triangular matrix, can be solved in $O(pn^3)$ flops if the matrix of coefficients is explicitly formed. We develop a new method for this system that circumvents the explicit product and requires only $O(pn^2)$ flops to execute. The error bounds for the new algorithm are essentially the same as the error bounds for the explicit method. The new algorithm extends readily to the situation when $R^{(1)}$ is upper quasi-triangular.

Key words. back-substitution, matrix products

AMS subject classification. 65F05

PII. S0895479801396051

1. Introduction.

Suppose the matrices $R^{(1)}, R^{(2)}, \ldots, R^{(p)} \in \mathbb{R}^{n\times n}$ are all upper triangular and that we want to solve

$$(R^{(1)} \cdots R^{(p)} - \lambda I)x = b, \tag{1.1}$$

where $\lambda \in \mathbb{R}$, $b \in \mathbb{R}^n$, and the matrix of coefficients is nonsingular. This problem arises in various product eigenvalue problems $(A^{(1)} \cdots A^{(p)})x = \lambda x$. (See [2].) In these settings the A-matrices are reduced to triangular form without the explicit formation of the product. The computation of eigenvectors by back-substitution involves the solution of a product triangular system with shift.

One way to solve (1.1) is to form the upper triangular matrix $(R^{(1)} \cdots R^{(p)} - \lambda I)$ and then use back-substitution. We refer to this as the *explicit method* and note that it is an $O(pn^3)$ procedure because of the matrix-matrix multiplications. In this paper we develop an *implicit method* that carefully engages selected parts of the coefficient matrix during the back-substitution process. The implicit algorithm requires only $O(pn^2)$ flops and has the same backward error properties as the explicit method. Our contribution therefore adds to the set of product-free matrix algorithms that have recently been developed for problems that involve matrix products. See [1] for an overview of this important paradigm and [4] for an example.

To illustrate the main idea without getting bogged down in details we first work through the $p = 2$ case. We then discuss the general algorithm, including a simple extension that can handle the case when $R^{(1)}$ is upper quasi-triangular. An error analysis and some reaffirming numerical results complete the paper.

2. The $p = 2$ case.

Consider the situation when the product in (1.1) involves just two matrices. Assume that $S, T \in \mathbb{R}^{n\times n}$ are upper triangular and that the system

$$(ST - \lambda I)x = b, \qquad \lambda \in \mathbb{R},\ b \in \mathbb{R}^n,$$

*Received by the editors October 3, 2001; accepted for publication (in revised form) by I. C. F. Ipsen February 15, 2002; published electronically July 9, 2002. This work was supported by NSF grant CCR-9901988.

http://www.siam.org/journals/simax/24-1/39605.html

†Center for Applied Mathematics, Cornell University, 657 Rhodes Hall, Ithaca, NY 14853-7510 (carlam@cam.cornell.edu).

‡Department of Computer Science, Cornell University, 4130 Upson Hall, Ithaca, NY 14853-7510 (cv@cs.cornell.edu).

is nonsingular. Suppose $1 \le k \le n-1$. Define

$$\begin{aligned} S_+ &= S(n-k{:}n, n-k{:}n), \\ T_+ &= T(n-k{:}n, n-k{:}n), \\ x_+ &= x(n-k{:}n), \\ b_+ &= b(n-k{:}n) \end{aligned}$$

and observe that

$$(S_+T_+ - \lambda I_{k+1})\, x_+ = b_+$$

is just the trailing $(k+1)$-by-$(k+1)$ portion of $(ST-\lambda I)x = b$. It has the form

$$\left(\begin{bmatrix} \sigma & u^T \\ 0 & S_c \end{bmatrix} \begin{bmatrix} \tau & v^T \\ 0 & T_c \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda I_k \end{bmatrix} \right) \begin{bmatrix} \gamma \\ x_c \end{bmatrix} = \begin{bmatrix} \beta \\ b_c \end{bmatrix},$$

where $\sigma, \tau, \gamma, \beta \in \mathbb{R}$, $u, v, x_c, b_c \in \mathbb{R}^k$, and $S_c, T_c \in \mathbb{R}^{k\times k}$. The two rows in this equation tell us that

$$(S_cT_c - \lambda I_k)x_c = b_c \tag{2.1}$$

and

$$\gamma = \frac{\beta - \sigma v^T x_c - u^T T_c x_c}{\sigma\tau - \lambda}\,. \tag{2.2}$$

The efficiency of "ordinary" back-substitution relies on the fact that at the start of step k the vector x_c is available and that the scalar γ can be obtained in $O(k)$ flops. However, in our product system if we literally use (2.2) to compute γ, then $O(k^2)$ flops are required because of the matrix-vector product T_cx_c. Unless this computation can be rearranged we are headed for an overall algorithm that needs $O(n^3)$ flops.

Fortunately there is a way to do this through a simple recursion that can be used to compute $w_+ = T_+x_+$ (the "next" w) from $w_c = T_cx_c$ (the "current" w). Since

$$w_+ = T_+x_+ = \begin{bmatrix} \tau & v^T \\ 0 & T_c \end{bmatrix} \begin{bmatrix} \gamma \\ x_c \end{bmatrix} = \begin{bmatrix} \tau\gamma + v^Tx_c \\ w_c \end{bmatrix}$$

it follows that we need only compute the scalar $\omega \equiv \tau\gamma + v^Tx_c$ to get w_+ from w_c. Thus, we can carry out each of the transitions $x_c \to x_+$ and $w_c \to w_+$ in $O(k)$ flops, and this renders the following overall procedure.

IMPLICIT METHOD ($p = 2$).

$x_c \leftarrow b_n/(S(n,n)T(n,n) - \lambda)$

$w_c \leftarrow T(n,n)x_c$

for $k = 1{:}n-1$

$\sigma \leftarrow S(n-k, n-k)$; $u \leftarrow S(n-k, n-k+1{:}n)^T$

$\tau \leftarrow T(n-k, n-k)$; $v \leftarrow T(n-k, n-k+1{:}n)^T$

$\gamma \leftarrow (\beta - \sigma(v^Tx_c) - u^Tw_c)/(\sigma\tau - \lambda)$

$\omega \leftarrow \tau\gamma + v^Tx_c$

$x_c \leftarrow \begin{bmatrix} \gamma \\ x_c \end{bmatrix}$; $w_c \leftarrow \begin{bmatrix} \omega \\ w_c \end{bmatrix}$

end

$x \leftarrow x_c$

In step k there are two length-k inner products, i.e., $v^T x_c$ and $u^T w_c$. Thus, the algorithm requires a total of $2n^2$ flops.

3. The general case. We now extend the above algorithm to the case when the coefficient matrix involves the product of p upper triangular matrices:

$$\left(R^{(1)} \cdots R^{(p)} - \lambda I\right) x = b. \tag{3.1}$$

Suppose $1 \le k \le n-1$. Define

$$\begin{aligned} R_+^{(i)} &= R^{(i)}(n-k{:}n, n-k{:}n), \qquad i = 1{:}p, \\ x_+ &= x(n-k{:}n), \\ b_+ &= b(n-k{:}n) \end{aligned}$$

and observe that

$$\left(R_+^{(1)} \cdots R_+^{(p)} - \lambda I_{k+1}\right) x_+ = b_+$$

is just the trailing $(k+1)$-by-$(k+1)$ portion of (3.1). It has the form

$$\left(\begin{bmatrix} \sigma_1 & u_1^T \\ 0 & R_c^{(1)} \end{bmatrix} \cdots \begin{bmatrix} \sigma_p & u_p^T \\ 0 & R_c^{(p)} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda I_k \end{bmatrix} \right) \begin{bmatrix} \gamma \\ x_c \end{bmatrix} = \begin{bmatrix} \beta \\ b_c \end{bmatrix}, \tag{3.2}$$

where $\sigma_i, \gamma, \beta \in \mathbb{R}$, $u_i, x_c, b_c \in \mathbb{R}^k$, and $R_c^{(i)} \in \mathbb{R}^{k \times k}$ for $i = 1{:}p$. In order to develop the necessary recursions for the back-substitution process we need to look more carefully at the product of the partitioned triangular matrices in this equation. It is easy to show by induction on p that

$$\begin{bmatrix} \sigma_1 & u_1^T \\ 0 & R_c^{(1)} \end{bmatrix} \cdots \begin{bmatrix} \sigma_p & u_p^T \\ 0 & R_c^{(p)} \end{bmatrix} = \begin{bmatrix} \sigma_1 \cdots \sigma_p & \displaystyle\sum_{j=1}^{p} (\sigma_1 \cdots \sigma_{j-1}) u_j^T R_c^{(j+1)} \cdots R_c^{(p)} \\ 0 & R_c^{(1)} \cdots R_c^{(p)} \end{bmatrix}.$$

By substituting this into (3.2) we conclude that

$$(R_c^{(1)} \cdots R_c^{(p)} - \lambda I_k) x_c = b_c \tag{3.3}$$

and

$$\gamma = \frac{\beta - \sum_{j=1}^{p} (\sigma_1 \cdots \sigma_{j-1})\, u_j^T\, w_c^{(j)}}{\sigma_1 \cdots \sigma_p - \lambda}, \tag{3.4}$$

where $w_c^{(p)} = x_c$ and $w_c^{(j)} = R_c^{(j+1)} \cdots R_c^{(p)} x_c$ for $j = p-1{:}-1{:}1$. In order to effect an $O(k)$ transition from x_c to x_+ we need to develop $O(k)$ update recipes for the w-vectors. In particular, we need a fast method for computing

$$w_+^{(j)} = R_+^{(j+1)} \cdots R_+^{(p)} x_+, \qquad j = 1{:}p-1, \tag{3.5}$$

assuming that we are in possession of $w_c^{(1)}, \ldots, w_c^{(p)}$. Since the matrices

$$R_+^{(i)} = \begin{bmatrix} \sigma_i & u_i^T \\ 0 & R_c^{(i)} \end{bmatrix}, \qquad i = j{+}1{:}p,$$

are upper triangular it follows that each $w_+^{(j)}$ has the form

$$w_+^{(j)} \;=\; \begin{bmatrix} \omega_j \\ w_c^{(j)} \end{bmatrix},$$

and so we just need a quick way to compute the scalars $\omega_1, \ldots, \omega_p$. Since $w_+^{(p)} = x_+$ we have

$$\begin{bmatrix} \omega_p \\ w_c^{(p)} \end{bmatrix} \;=\; \begin{bmatrix} \gamma \\ x_c \end{bmatrix},$$

and so $\omega_p = \gamma$. Simple formulae for $\omega_{p-1}, \ldots, \omega_1$ can be derived from (3.5). This equation tells us that $w_+^{(j)} = R_+^{(j+1)} w_+^{(j+1)}$, i.e.,

$$\begin{bmatrix} \omega_j \\ w_c^{(j)} \end{bmatrix} \;=\; \begin{bmatrix} \sigma_{j+1} & u_{j+1}^T \\ 0 & R_c^{(j+1)} \end{bmatrix} \begin{bmatrix} \omega_{j+1} \\ w_c^{(j+1)} \end{bmatrix}.$$

Thus,

$$\omega_j \;=\; \sigma_{j+1}\omega_{j+1} + u_{j+1}^T w_c^{(j+1)}$$

for $j = p{-}1{:}\,{-}1{:}1$. Combining all this we obtain the following procedure.

IMPLICIT METHOD (GENERAL p).

$x_c \;\leftarrow\; b_n/(R^{(1)}(n,n)\cdots R^{(p)}(n,n) - \lambda)$
$w_c^{(p)} = x_c$
$w_c^{(j)} \;\leftarrow\; R_c^{(j+1)}(n,n) w_c^{(j+1)} \qquad (j = p{-}1{:}\,{-}1{:}1)$
for $k = 1{:}n-1$
  $\sigma_j \;\leftarrow\; R^{(j)}(n-k, n-k) \qquad (j = 1{:}p)$
  $u_j \;\leftarrow\; R^{(j)}(n-k, n-k+1{:}n)^T \qquad (j = 1{:}p)$
  $\beta \;\leftarrow\; b(n-k)$
  $$\gamma \;=\; \left(\beta \;-\; \sum_{j=1}^{p} (\sigma_1 \cdots \sigma_{j-1})\; u_j^T\; w_c^{(j)} \right) \Big/ \; (\sigma_1 \cdots \sigma_p \;-\; \lambda)$$
  if $k < n-1$
    $\omega_p \;\leftarrow\; \gamma;\; w_c^{(p)} \leftarrow x_c$
    $\omega_j \leftarrow \sigma_{j+1}\omega_{j+1} + u_{j+1}^T w_c^{(j+1)} \qquad (j = p{-}1{:}\,{-}1{:}1)$
    $w_c^{(j)} \;\leftarrow\; \begin{bmatrix} \omega_j \\ w_c^{(j)} \end{bmatrix} \qquad (j = 1{:}p)$
  end
  $x_c \;\leftarrow\; \begin{bmatrix} \gamma \\ x_c \end{bmatrix}$
end
$x \leftarrow x_c$

There are p length-k inner products to compute in step k, i.e., $u_j^T w_c^{(j)}$, $j = 1{:}p$. Thus, the overall algorithm requires pn^2 flops.

4. The quasi-triangular case. In the eigenvector application mentioned in the introduction it is sometimes the case that $R^{(1)}$ is upper quasi-triangular, i.e., block upper triangular with 1-by-1 and 2-by-2 blocks along the diagonal. The 2-by-2 bumps correspond to complex conjugate eigenvalue pairs.

The implicit algorithm generalizes in a straightforward way to handle this situation. To see how to carry out a step that corresponds to a 2-by-2 bump we rewrite (3.2) as follows:

$$\left(\begin{bmatrix} S_1 & U_1^T \\ 0 & R_c^{(1)} \end{bmatrix} \cdots \begin{bmatrix} S_p & U_p^T \\ 0 & R_c^{(p)} \end{bmatrix} - \begin{bmatrix} \lambda I_2 & 0 \\ 0 & \lambda I_k \end{bmatrix} \right) \begin{bmatrix} \gamma \\ x_c \end{bmatrix} = \begin{bmatrix} \beta \\ b_c \end{bmatrix}, \tag{4.1}$$

where $S_i \in \mathbb{R}^{2\times 2}$, $U_i \in \mathbb{R}^{(n-k)\times 2}$, $R_c^{(i)} \in \mathbb{R}^{k\times k}$, $\beta \in \mathbb{R}^2$, $b_c \in \mathbb{R}^k$, and $x_c \in \mathbb{R}^k$ are given. Our goal is to compute efficiently $\gamma \in \mathbb{R}^2$. Following the corresponding discussion in section 3 it can be shown that

$$\begin{bmatrix} S_1 & U_1^T \\ 0 & R_c^{(1)} \end{bmatrix} \cdots \begin{bmatrix} S_p & U_p^T \\ 0 & R_c^{(p)} \end{bmatrix} = \begin{bmatrix} S_1 \cdots S_p & \displaystyle\sum_{j=1}^{p} (S_1 \cdots S_{j-1}) U_j^T R_c^{(j+1)} \cdots R_c^{(p)} \\ 0 & R_c^{(1)} \cdots R_c^{(p)} \end{bmatrix}.$$

From this it follows that $(R_c^{(1)} \cdots R_c^{(p)} - \lambda I_k) x_c = b_c$ and

$$(S_1 \cdots S_p - \lambda I_2)\, \gamma = \beta - \sum_{j=1}^{p} (S_1 \cdots S_{j-1})\, U_j^T\, w_c^{(j)}, \tag{4.2}$$

where $w_c^{(p)} = x_c$ and

$$w_c^{(j)} = R_c^{(j+1)} \cdots R_c^{(p)} x_c, \qquad j = p{-}1{:}\,{-}1{:}1. \tag{4.3}$$

Thus, the next two components of x, i.e., $\gamma = x(n-k-1{:}n-k)$, are found by solving (4.2), a 2-by-2 linear system. The update of the w-vectors is analogous to the update derived in section 3 for the triangular case. Since $w_+^{(p)} = x_+$ we have

$$w_+^{(p)} \equiv \begin{bmatrix} \omega_p \\ w_c^{(p)} \end{bmatrix} = \begin{bmatrix} \gamma \\ x_c \end{bmatrix},$$

and so $\omega_p = \gamma$. From (4.3) $w_+^{(j)} = R_+^{(j+1)} w_+^{(j+1)}$, i.e.,

$$\begin{bmatrix} \omega_j \\ w_c^{(j)} \end{bmatrix} = \begin{bmatrix} S_{j+1} & U_{j+1}^T \\ 0 & R_c^{(j+1)} \end{bmatrix} \begin{bmatrix} \omega_{j+1} \\ w_c^{(j+1)} \end{bmatrix},$$

and so the vectors $\omega_j \in \mathbb{R}^2$ can be found via

$$\omega_j = S_{j+1}\omega_{j+1} + U_{j+1}^T w_c^{(j+1)}$$

for $j = p{-}1{:}\,{-}1{:}1$. Thus, the transition from $\{x_c, w_c^{(1)}, \ldots, w_c^{(p)}\}$ to $\{x_+, w_+^{(1)}, \ldots, w_+^{(p)}\}$ involves $O(k)$ flops even if a 2-by-2 bump is encountered.

5. Backward error analysis. We show that backward error analyses for the explicit and implicit methods are essentially the same. The goal is not to derive the "best possible" results but simply to substantiate observed numerical behavior. In particular, we show that both the explicit and implicit methods produce a computed solution $\hat{x}$ that solves a "nearby" system

$$(R^{(1)} \cdots R^{(p)} - \lambda I + E)\hat{x} \;=\; b, \tag{5.1}$$

where the perturbation matrix E satisfies

$$\| E \| \;=\; O\left(\mathbf{u}\left(\| R^{(1)} \| \cdots \| R^{(p)} \| \;+\; |\lambda|\right)\right) \tag{5.2}$$

with $\mathbf{u}$ being the unit roundoff and $\| \cdot \|$ designating (say) the 2-norm.

For simplicity we assume that $R^{(1)}$ is upper triangular. The analysis for the quasi-triangular case is similar and basically yields the same results.

Consider the explicit method first. It begins with the computation of the matrix of coefficients $A = R^{(1)} \cdots R^{(p)} - \lambda I$:

$A_1 = R^{(1)}$
for $j = 2{:}p$
    $A_j = \text{fl}(A_{j-1}R^{(j)})$
end
$\hat{A} = \text{fl}(A_p - \lambda I)$

Here, $\text{fl}(x \text{ op } y)$ is the floating point version of x op y, where x and y are floating point scalars, vectors, or matrices and "op" is some legitimate operation between them. Applying standard floating point error results that can be found in [2] or [3], it can be shown that

$$A_j \;=\; A_{j-1}R^{(j)} + E_j, \qquad\qquad \| E_j \| \;=\; O\left(\mathbf{u} \| A_{j-1} \| \| R^{(j)} \|\right) \tag{5.3}$$

for $j = 2{:}p$. Taking into account the roundoff error associated with the λ-shift gives

$$\hat{A} \;=\; A_p \;-\; \lambda I \;+\; F_1, \qquad\qquad \| F_1 \| \;=\; O\left(\mathbf{u}\left(\| A_p \| + |\lambda|\right)\right), \tag{5.4}$$

and so by a simple inductive argument we find that

$$\hat{A} \;=\; R^{(1)} \cdots R^{(p)} - \lambda I \;+\; F_1 \;+\; \sum_{j=2}^{p} E_j R^{(j+1)} \cdots R^{(p)}.$$

At this point back-substitution is applied to $\hat{A}x = b$ and produces an $\hat{x}$ that satisfies

$$(\hat{A} + F_2)\hat{x} = b, \qquad\qquad \| F_2 \| \;=\; O\left(\mathbf{u}\| \hat{A} \|\right). \tag{5.5}$$

Combining all of these results, it is not hard to show that (5.1) holds with

$$E = F_1 + \sum_{j=2}^{p} E_j R^{(j+1)} \cdots R^{(p)} + F_2.$$

Taking norms in this equation and simplifying the right-hand side with (5.3), (5.4), and (5.5) confirms (5.2).

To show that (5.1) and (5.2) apply to the implicit method, we proceed by induction on n. The $n = 1$ case holds because the $\hat{x}$ produced by the implicit method is identical to the $\hat{x}$ that is produced by the explicit method.

Assume that $1 \leq k \leq n-1$. Using the notation of section 3 and "hats" to designate computed quantities, the induction argument is complete if we can show that

$$\left(R_+^{(1)} \cdots R_+^{(p)} \; - \; \lambda I_{k+1} \; + \; E_+\right) \hat{x}_+ = b_+ \tag{5.6}$$

with

$$\| \, E_+ \, \| \; = \; O\left(\mathbf{u}\left(\| \, R_+^{(1)} \, \| \cdots \| \, R_+^{(p)} \, \| \; + \; |\lambda|\right)\right), \tag{5.7}$$

given that

$$\left(R_c^{(1)} \cdots R_c^{(p)} \; - \; \lambda I_k \; + \; E_c\right) \hat{x}_c = b_c \tag{5.8}$$

with

$$\| \, E_c \, \| \; = \; O\left(\mathbf{u}\left(\| \, R_c^{(1)} \, \| \cdots \| \, R_c^{(p)} \, \| \; + \; |\lambda|\right)\right). \tag{5.9}$$

To that end partition (5.6)

$$E_+ \; = \; \begin{bmatrix} \epsilon & e^T \\ 0 & E_c \end{bmatrix}$$

conformably with (3.2).

From (5.6) we see that our task is to show that $\hat{\gamma}$ satisfies

$$\left(\begin{bmatrix} \sigma_1 & u_1^T \\ 0 & R_c^{(1)} \end{bmatrix} \cdots \begin{bmatrix} \sigma_p & u_p^T \\ 0 & R_c^{(p)} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda I_k \end{bmatrix} + \begin{bmatrix} \epsilon & e^T \\ 0 & E_c \end{bmatrix}\right) \begin{bmatrix} \hat{\gamma} \\ \hat{x}_c \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1 \cdots \sigma_p - \lambda + \epsilon & \displaystyle\sum_{j=1}^{p} (\sigma_1 \cdots \sigma_{j-1}) u_j^T R_c^{(j+1)} \cdots R_c^{(p)} + e^T \\ 0 & R_c^{(1)} \cdots R_c^{(p)} - \lambda I_k + E_c \end{bmatrix} \begin{bmatrix} \hat{\gamma} \\ \hat{x}_c \end{bmatrix} = \begin{bmatrix} \beta \\ b_c \end{bmatrix}$$

(5.10)

with

$$|\epsilon| \; = \; O\left(\mathbf{u}\left(\| \, R_+^{(1)} \, \| \cdots \| \, R_+^{(p)} \, \| \; + \; |\lambda|\right)\right) \tag{5.11}$$

and

$$\| \, e \, \| \; = \; O\left(\mathbf{u}\left(\| \, R_+^{(1)} \, \| \cdots \| \, R_+^{(p)} \, \| \; + \; |\lambda|\right)\right). \tag{5.12}$$

Using elementary properties of the 2-norm, it can be shown that (5.7) follows from (5.9), (5.11), and (5.12).

Before we set out to verify (5.11) and (5.12) we establish a handy tilde notation that can be used to indicate accuracy to machine precision. If M is a matrix, then $\widetilde{M}$ is an approximation that satisfies $\| M - \widetilde{M} \|/\| M \| = O(\mathbf{u})$. The notation is a useful way to account for the rounding errors in floating point matrix-vector multiplication. Indeed, if M is a floating point matrix and v is a floating point vector, then $\mathrm{fl}(Mv) = \widetilde{M}v$.

When we account for all the rounding errors associated with the evaluation of the right-hand side in (3.4) we find that

$$\hat{\gamma} \;=\; \frac{\beta \;-\; \sum_{j=1}^{p} \left((\sigma_1 \cdots \sigma_{j-1})(1+\mu_j) \tilde{u}_j^T \hat{w}_c^{(j)} \right)}{\sigma_1 \cdots \sigma_p (1+\delta_1) \;-\; \lambda(1+\delta_2)}, \tag{5.13}$$

where $|\delta_i| = O(\mathbf{u})$ for $i = 1, 2$ and $|\mu_j| = O(\mathbf{u})$ for $i = 1{:}p$. We shall establish below that the computed w_c vectors satisfy

$$\hat{w}_c^{(j)} = (R_c^{(j+1)} + F_c^{(j+1)}) \cdots (R_c^{(p)} + F_c^{(p)}) \hat{x}_c, \qquad \| F_c^{(j)} \| \;=\; O\left(\mathbf{u} \| R_c^{(j)} \| \right) \tag{5.14}$$

for $j = 1{:}p$. This says that

$$\hat{w}_c^{(j)} = \left(R_c^{(j+1)} \cdots R_c^{(p)} \;+\; G^{(j)} \right) \hat{x}_c, \tag{5.15}$$

where $\| G^{(j)} \| \;=\; O(\mathbf{u} \| R_c^{(j+1)} \| \cdots \| R_c^{(p)} \|)$. By rearranging (5.13) and substituting (5.15) we get

$$\begin{aligned} \beta &= (\sigma_1 \cdots \sigma_p (1+\delta_1) - \lambda(1+\delta_2))\, \hat{\gamma} \\ &\quad + \left(\sum_{j=1}^{p} (\sigma_1 \cdots \sigma_{j-1})(1+\mu_j) \tilde{u}_j^T \left(R_c^{(j+1)} \cdots R_c^{(p)} \;+\; G^{(j)} \right) \right) \hat{x}_c \\ &= (\sigma_1 \cdots \sigma_p - \lambda + \epsilon)\, \hat{\gamma} \;+\; \left(\sum_{j=1}^{p} (\sigma_1 \cdots \sigma_{j-1}) u_j^T \left(R_c^{(j+1)} \cdots R_c^{(p)} \right) + e^T \right) \hat{x}_c, \end{aligned}$$

which completely specify $\epsilon \in \mathbb{R}$ and $e \in \mathbb{R}^k$. It follows that (5.10) holds for this choice of ϵ and e. Moreover, (5.11) and (5.12) are both satisfied.

The last thing we must do is verify (5.14) for $j = 1{:}p$. This result is certainly correct if $k = 1$ since $\hat{w}_c = \mathrm{fl}(r_{nn}^{(j+1)} \cdots r_{nn}^{(p)})$. Assume that it holds for general k. Our task is to show that

$$\hat{w}_+^{(j)} = \begin{bmatrix} \hat{\omega}_j \\ \hat{w}_c^{(j)} \end{bmatrix} \;=\; (R_+^{(j+1)} \;+\; F_+^{(j+1)}) \cdots (R_+^{(p)} \;+\; F_+^{(p)}) \hat{x}_+, \tag{5.16}$$

where

$$\| F_+^{(j)} \| \;=\; O\left(\mathbf{u} \| R_+^{(j)} \| \right), \qquad j = p{:}-1{:}1. \tag{5.17}$$

In looking at the specification of the implicit algorithm in section 3, we see that (5.16) holds if $j = p$ since

$$\hat{w}_+^{(p)} = \begin{bmatrix} \hat{\gamma} \\ \hat{x}_c \end{bmatrix} \;=\; \hat{x}_+.$$

Assume that (5.16) and (5.17) hold for some general j that satisfies $1 < j \leq p$. From (5.14) and the definition of ω_j we have

$$\hat{w}_+^{(j-1)} = \begin{bmatrix} \hat{\omega}_{j-1} \\ \hat{w}_c^{(j-1)} \end{bmatrix} = \begin{bmatrix} \mathrm{fl}(\sigma_j \hat{\omega}_j + u_j^T \hat{w}_c^{(j)}) \\ (R_c^{(j)} + F_c^{(j)}) \cdots (R_c^{(p)} + F_c^{(p)}) \hat{x}_c \end{bmatrix}.$$

Since

$$\mathrm{fl}(\sigma_j \hat{\omega}_j + u_j^T \hat{w}_c^{(j)}) = \sigma_j \hat{\omega}_j (1+\tau) + \tilde{u}_j^T \hat{w}_c^{(j)},$$

where $|\tau| = O(\mathbf{u})$, we have

$$\hat{w}_+^{(j-1)} = \begin{bmatrix} \sigma_j(1+\tau) & \tilde{u}_j^T \\ 0 & (R_c^{(j)} + F_c^{(j)}) \end{bmatrix} \begin{bmatrix} \hat{\omega}_j \\ \hat{w}_c^{(j)} \end{bmatrix} = (R_+^{(j)} + F_+^{(j)}) \hat{w}_+^{(j)},$$

where

$$F_+^{(j)} = \begin{bmatrix} \tau\sigma_j & (\tilde{u}_j - u_j)^T \\ 0 & F_c^{(j)} \end{bmatrix}.$$

It follows that (5.16) and (5.17) hold for $j = p\colon -1\colon 1$.

This completes the verification that both the explicit and implicit methods produce computed solutions that satisfy (5.1) and (5.2). We mention that if $\lambda = 0$, then we can solve (1.1) via repeated back-substitution. Using standard results about this process it can be shown that

$$(R^{(1)} + E^{(1)}) \cdots (R^{(p)} + E^{(p)}) \hat{x} = b,$$

where $|E^{(j)}| = O\left(\mathbf{u}|R^{(j)}|\right)$ for $j = 1\colon p$. So although we have shown that the error bounds for the implicit and explicit methods are essentially the same, neither result is as strong as that which can be obtained for the $\lambda = 0$ case.

6. Numerical results. MATLAB implementations of the explicit and implicit methods are available at http://www.cs.cornell.edu/cv/ and were tested to see if the preceding inverse error analysis is realistic. Results like (5.1)–(5.2) that claim a computed solution $\hat{x}$ satisfies a "nearby" system $(A + E)\hat{x} = b$ can be affirmed by comparing the 2-norm of

$$\hat{E} = (b - A\hat{x})\hat{x}^T / (\hat{x}^T \hat{x})$$

with the alleged 2-norm error bound. This is because $(A + \hat{E})\hat{x} = b$ and $\hat{E}$ has the smallest 2-norm of all matrices E that satisfy $(A + E)\hat{x} = b$. Indeed,

$$\| \hat{E} \| = \frac{\| b - A\hat{x} \|}{\| \hat{x} \|}.$$

In our case $A = R^{(1)} \cdots R^{(p)} - \lambda I$, and so the issue before us is the size of

$$\phi(\hat{x}) = \frac{\left(\| b - (R^{(1)} \cdots R^{(p)} - \lambda I)\hat{x} \| / \| \hat{x} \| \right)}{\left(\mathbf{u} \left(\| R^{(1)} \| \cdots \| R^{(p)} \| + |\lambda| \right) \right)} \tag{6.1}$$

TABLE 6.1
Lower and upper bounds for $\| \hat{E}_{imp} \| / \| \hat{E}_{exp} \|$.

	$p = 2$	$p = 4$	$p = 6$
$n = 50$	(.10 , 8.4)	(.08 , 14.6)	(.06 , 7.6)
$n = 100$	(.14 , 6.3)	(.07 , 5.5)	(.06 , 6.6)
$n = 150$	(.09 , 6.0)	(.04 , 5.4)	(.16 , 6.6)
$n = 200$	(.19 , 3.4)	(.09 , 8.7)	(.06 , 8.6)

when $\hat{x}$ is the solution obtained via the implicit and explicit methods. Denote these solutions by $\hat{x}_{imp}$ and $\hat{x}_{exp}$, respectively. Note that if

$$\hat{E}_{imp} = (b - A\hat{x}_{imp})\hat{x}_{imp}^T / (\hat{x}_{imp}^T \hat{x}_{imp}),$$

$$\hat{E}_{exp} = (b - A\hat{x}_{exp})\hat{x}_{exp}^T / (\hat{x}_{exp}^T \hat{x}_{exp}),$$

then from (6.1)

$$\frac{\phi(\hat{x}_{imp})}{\phi(\hat{x}_{exp})} = \frac{\| \hat{E}_{imp} \|}{\| \hat{E}_{exp} \|}.$$

For a particular choice of n and p we experimentally determined a lower bound α and an upper bound β for this quotient, i.e., α and β so that

$$\alpha \| \hat{E}_{exp} \| \leq \| \hat{E}_{imp} \| \leq \beta \| \hat{E}_{exp} \|.$$

In Table 6.1 we report the results. Each cell specifies an estimate of α and β based on 100 randomly generated examples. The results substantiate what the error analysis of section 5 says. The inverse error analysis for the implicit method is essentially the same as the inverse error analysis for the explicit method.

REFERENCES

[1] B. DE MOOR AND P. VAN DOOREN, *Generalizations of the singular value and QR decompositions*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 993–1014.
[2] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[3] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[4] C. F. VAN LOAN, *A general matrix eigenvalue algorithm*, SIAM J. Numer. Anal., 12 (1975), pp. 817–834.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 303–319

TWO-SIDED ARNOLDI AND NONSYMMETRIC LANCZOS ALGORITHMS*

JANE CULLUM† AND TONG ZHANG‡

Abstract. We introduce new two-sided Arnoldi recursions and use them to define a model reduction procedure for large, linear, time-invariant, multi-input/multi-output differential algebraic systems. We prove that this procedure has desirable moment matching properties. We define a corresponding model reduction procedure which is based on a band nonsymmetric Lanczos recursion and prove that if the deflation is exact and there are no breakdowns in the recursions, then these two model reduction procedures generate identical reduced-order systems. We prove similar equivalences for corresponding eigenelement procedures. We concentrate on the theoretical properties of the new algorithms.

Key words. two-sided Arnoldi, nonsymmetric Lanczos, equivalences, relationships, iterative methods, reduced-order systems, eigenvalues, model reduction

AMS subject classifications. 15A06, 15A18, 65F, 65L, 93C

PII. S0895479898339013

1. Introduction.

This paper arose out of our work on model reduction algorithms for large multi-input/multi-output (*mi/mo*), time-invariant, delay-differential-algebraic systems of equations which occur in modeling VLSI interconnects (wires, planes, conductors) [7]. The inner loop of the outer/inner loop procedure that was developed in [7] repeatedly exercises an iterative model reduction procedure for time-invariant linear systems [11]. This inner algorithm must be able to handle systems with arbitrary numbers of inputs and outputs.

Single-sided Arnoldi model reduction methods have been proposed which directly use the system block input matrix in the iterative method but do not directly use the system block output matrix. We define two-sided Arnoldi recursions with the capability of directly incorporating both of these matrices into an iterative model reduction procedure.

The two-sided nature of these recursions leads us to a comparison of a model reduction procedure which is based upon the new two-sided Arnoldi recursions and a corresponding procedure which is based upon the nonsymmetric band Lanczos recursion developed in [1]. We prove that these two procedures generate identical reduced-order models. We also prove that the iterates generated by corresponding eigenelement procedures are identical. We focus on the theoretical properties of these procedures.

In section 2 we define a two-sided block Arnoldi recursion which consists of two independent applications of a corresponding one-sided block Arnoldi recursion and a computation which combines the quantities generated by these two applications. One application uses the system matrix with the system block input matrix. The second application uses the transpose of the system matrix with the system block output matrix. We illustrate some properties of this basic two-sided Arnoldi recursion, including

*Received by the editors May 18, 1998; accepted for publication (in revised form) by R. Freund February 24, 2002; published electronically August 5, 2002.
http://www.siam.org/journals/simax/24-2/33901.html
†MS B250, Los Alamos National Laboratory, Los Alamos, NM 87545 (cullumj@lanl.gov). The research of this author was supported by the Department of Energy under contract W-7405-ENG-36 and the Office of Science, MICS, Applied Mathematical Sciences Program under grant KC-07-01-01.
‡IBM T. J. Watson Research Center, Yorktown Heights, NY (tzhang@watson.ibm.com).

the very interesting possibility of recovery from breakdown without any modifications of these recursions. This two-sided block Arnoldi recursion is an extension of the work found in [17].

In section 3 we use the recursions in section 2 to define a two-sided block Arnoldi model reduction procedure for generating approximations to transfer functions of large systems of time-invariant, differential-algebraic equations. We prove that these approximations possess desirable matrix moment matching properties [11]. The proof is interesting because it is self-contained and uses only general properties of two-sided Krylov recursions. For example, it does not use the orthogonality of the associated vectors.

In section 4, we review briefly some of the properties of the band nonsymmetric Lanczos recursions defined in [1]. Assuming exact arithmetic and exact deflation, we derive common properties of the band Lanczos and of the two-sided block Arnoldi recursions. We exploit those properties within the context of corresponding model reduction and eigenelement methods to prove that corresponding methods generate identical approximations. These results complement the earlier work in [4], [3], comparing one-sided Arnoldi methods and nonsymmetric Lanczos methods for solving $Ax = b$ or $Ax = \lambda x$.

Assumptions. Unless it is stated explicitly otherwise, in any discussion of any Arnoldi or nonsymmetric Lanczos-based procedure, we will assume that all of the required quantities are well-defined; no breakdowns occur; the underlying recursions do not terminate prematurely; and the arithmetic is exact.

1.1. Notation. We summarize the notation which is nonstandard.

- *si/so*: single input, single output system
- *mi/mo*: multiple input, multiple output system
- r, l: subscripts (superscripts) to denote quantities associated with right and left vectors which were generated using A and A^T
- $B(:, [i:j])$ $[B([i:j], :)]$: columns [rows] i through j of matrix B
- $I_{[i:j]}$: columns i through j of an identity matrix I_K where K is specified within the local context
- $I_{[xlast]}$, $I_{[xfirst]}$: denote corresponding $I_{[i:j]}$ where the block $[i:j]$ corresponds to the indices in the last (first) block column in an associated block structure and setting $x = r, l(v, w)$ indicates right or left quantities for Arnoldi (Lanczos) recursions
- $H([i:j], [k:l])$: submatrix of H consisting of the intersection of rows i through j with columns k through l
- $[Q_j, \widetilde{Q}_{j+1}]$: equals $[\widetilde{Q}_1, \ldots, \widetilde{Q}_j, \widetilde{Q}_{j+1}]$ for $Q_j \equiv [\widetilde{Q}_1, \ldots, \widetilde{Q}_j]$

2. Arnoldi recursions. Blocks occur naturally in *mi/mo* systems. System inputs and outputs are controlled by matrix blocks, and we measure the quality of our proposed reduced-order model by the number of rectangular block moments of the transfer function of the original system which are matched by corresponding moments of the transfer function of the reduced-order system. Therefore, initially we focus on block, two-sided Arnoldi recursions.

2.1. A block Arnoldi recursion. Given a matrix A, a consistent starting block of vectors X with d_r columns, and a deflation tolerance ϵ_d, we define a one-sided, block Arnoldi recursion [18]. At each iteration of this recursion we are working with a block of vectors which was generated by applying A to a block of vectors which was generated at an earlier iteration and invoking orthogonalization. It is possible and

typical, as the iterations proceed, that one or more vectors within the current block become dependent or nearly dependent upon vectors which have already been generated. In this situation, to preserve the integrity of the procedure, the (nearly) dependent vector(s) must be deflated from the process. Deflation is accomplished implicitly without any explicit permutations of vectors or modifications of the recursions. Example 2.1 provides a concrete illustration of the use of deflation in a block. For more details on deflation, see [5], [8], [9], [1].

In the statement of Algorithm 2.1 we use Q_{j+1} to denote the Arnoldi vectors after j block steps of the recursion, d_j to denote the size of the jth block, $\widetilde{Q}_j$, $s_m \equiv \sum_{j=1}^{m} d_j$ equals the number of columns in Q_m, and e_d is the deflation tolerance.

ALGORITHM 2.1. BLOCK ARNOLDI RECURSION.

Specify A, X, $\epsilon_d \geq 0$.
Decompose $X = \widetilde{Q}_1 S_r + \tilde{\Delta}_r$ *where* $\widetilde{Q}_1^T \widetilde{Q}_1 = I$
  and $\|\tilde{\Delta}_r\|_F \leq \sqrt{(d_r - d_1)}\epsilon_d$ *with* $d_1 = \operatorname{rank}(\widetilde{Q}_1)$.
Set $s = d = d_1$, $Y_1 = \widetilde{Q}_1$, $Q_1 = [\widetilde{Q}_1]$.
for $j = 1 : m$
  $Q(:, [s - d + 1 : s]) = Y_j$.
  $P_j = AY_j$.
  $H([1 : s], [s - d + 1 : s]) = Q_j^T P_j$.
  $P_j = P_j - Q_j H([1 : s], [s - d + 1 : s])$.
  if $j < m$
    Decompose $P_j = \widetilde{Q}_{j+1} S_j + \widetilde{\Delta}_j$ *where* $\widetilde{Q}_{j+1}^T \widetilde{Q}_{j+1} = I$
      and $\|\widetilde{\Delta}_j\|_F < \sqrt{d - d_{j+1}}\epsilon_d$ *with* $d_{j+1} = \operatorname{rank}(\widetilde{Q}_{j+1})$.
    $H([s + 1 : s + d_j], [s - d + 1 : s]) \equiv S_j$.
    $d = d_{j+1}$
    $s = s + d$
    $Q_{j+1} = [Q_j, \widetilde{Q}_{j+1}]$.
  else
    $R_m = P_m$
    $H_m \equiv H([1 : s_m], [1 : s_m])$.
  end
end

H_m denotes the square block upper Hessenberg matrix generated with diagonal blocks of size d_j, $1 \leq j \leq m$. R_m denotes the final $n \times d_m$ block residual matrix. The matrix form for these recursions is

$$AQ_m = Q_m H_m + R_m I_{[last]}^T, \tag{2.1}$$

where $I_{[last]}$ denotes $I_{[s_m - d_m + 1 : s_m]}$. We will also use $I_{[first]}$ to denote $I_{[1:d_1]}$. If Algorithm 2.1 is applied to $\{A, X_r\}$, we use $\{X_r, S_r, Q_r, H_r, R_r, I_{[rlast]}, I_{[rfirst]}\}$. Similarly, for $\{A^T, X_l\}$ we use $\{X_l, S_l, Q_l, H_l, R_l, I_{[llast]}, I_{[lfirst]}\}$.

If for some j, $d_j - d_{j+1} > 0$, then deflation has occurred in that block P_j. Checks for deflation are accomplished by applying modified Gram–Schmidt orthogonalization within each block and deflating any vector with norm smaller than or equal to ϵ_d. Since every right (left) Arnoldi vector generated is explicitly orthogonalized w.r.t. all existing right (left) vectors, modifications in the recursion formulas are not required when deflation occurs.

Example 2.1 illustrates some possible deflation scenarios. In the implementations any vector $\|p_j^{(k)}\| \leq \epsilon_d$ is ignored. Vectors are considered in their natural order and in place. For simplicity, in this example we assume that $\|p_1\| > \epsilon_d$.

Example 2.1. Let $P = [p_1, p_2, p_3]$ be a block of 3 vectors and consider the vectors in order. We have the following possible steps. The γ_{ij} denote the Gram–Schmidt orthogonalization coefficients.

1. $q_1 \equiv p_1/\|p_1\|$; $\quad p_2^{(2)} \equiv p_2 - \gamma_{21} q_1$; $\quad p_3^{(2)} \equiv p_3 - \gamma_{31} q_1$; $\quad$ go to step 2.
2. There are three cases.
   a. If $\|p_2^{(2)}\| \le \epsilon_d$ and $\|p_3^{(2)}\| > \epsilon_d$, set $q_2 \equiv p_3^{(2)}/\|p_3^{(2)}\|$ and terminate.
   b. If $\|p_2^{(2)}\| > \epsilon_d$, set $q_2 \equiv p_2^{(2)}/\|p_2^{(2)}\|$; $p_3^{(3)} = p_3^{(2)} - \gamma_{32} q_2$; go to step 3.
   c. If $\max(\|p_2^{(2)}\|, \|p_3^{(2)}\|) \le \epsilon_d$, terminate.
3. There are two cases.
   a. If $\|p_3^{(3)}\| \le \epsilon_d$, terminate.
   b. If $\|p_3^{(3)}\| > \epsilon_d$, set $q_3 \equiv p_3^{(3)}/\|p_3^{(3)}\|$ and terminate.

The following combinations of steps and matrix block relationships are possible.

A. $\{1, 2a\} \Rightarrow [p_1, p_2, p_3] = [q_1, q_2] \begin{bmatrix} \|p_1\| & \gamma_{21} & \gamma_{31} \\ 0 & 0 & \|p_3^{(2)}\| \end{bmatrix} + [0, p_2^{(2)}, 0]$.

B. $\{1, 2b, 3a\} \Rightarrow [p_1, p_2, p_3] = [q_1, q_2] \begin{bmatrix} \|p_1\| & \gamma_{21} & \gamma_{31} \\ 0 & \|p_2^{(2)}\| & \gamma_{32} \end{bmatrix} + [0, 0, p_3^{(3)}]$.

C. $\{1, 2b, 3b\} \Rightarrow [p_1, p_2, p_3] = [q_1, q_2, q_3] \begin{bmatrix} \|p_1\| & \gamma_{21} & \gamma_{31} \\ 0 & \|p_2^{(2)}\| & \gamma_{32} \\ 0 & 0 & \|p_3^{(3)}\| \end{bmatrix}$.

D. $\{1, 2c\} \Rightarrow [p_1, p_2, p_3] = [q_1] \begin{bmatrix} \|p_1\| & \gamma_{21} & \gamma_{31} \end{bmatrix} + [0, p_2^{(2)}, p_3^{(2)}]$.

Since the vectors within a candidate block P_j are considered in the natural order, and modified Gram–Schmidt orthogonalization is applied successively to each of these vectors, the first d_{j+1} columns of the subblock below each jth diagonal block in H_m form an upper triangular matrix. Therefore, we can truncate this matrix interior to a $(j+1)$st block and the truncated matrix $\widehat{H}$ retains the block upper Hessenberg structure. The corresponding block residual matrix $\widehat{R}$ for $\widehat{H}$ will have the same number of columns as the residual corresponding to the H matrix with $(j+1)$ complete blocks. The indices of the columns corresponding to $\widehat{R}$ are obtained by shifting the column indices of the $(j+1)$st block left by the number of columns truncated from that block. We will exercise this ability to truncate and retain structure in section 4 in our comparisons of methods which are based upon a two-sided block Arnoldi recursion with corresponding methods which are based upon the nonsymmetric band Lanczos recursion in [1].

2.2. A two-sided block Arnoldi recursion. We construct a two-sided block Arnoldi recursion by combining two independent applications of Algorithm 2.1, to $\{A, X_r\}$ and to $\{A^T, X_l\}$, with an appropriate *vector merge* of the resulting left and right Arnoldi vectors, $\{Q_l, Q_r\}$. The merge creates a modified right (left) residual matrix that is biorthogonal to the left (right) Arnoldi vectors. To maintain the equalities, the modification to the residual matrix must also be applied to the corresponding H matrix.

ALGORITHM 2.2. TWO-SIDED BLOCK ARNOLDI RECURSION.

Specify A, X_r, X_l, ϵ_d, m_r, m_l.
Apply Algorithm 2.1 *to* $\{A, X_r\}$ *for* m_r *block steps to generate* Q_r.
Apply Algorithm 2.1 *to* $\{A^T, X_l\}$ *for* m_l *block steps to generate* Q_l.
Compute $\overline{Q} = Q_l^T Q_r$.
If $\operatorname{rank}(Q_l) \le \operatorname{rank}(Q_r)$
  Solve $\overline{Q} Z = Q_l^T R_r$.
  Set $\overline{H} \equiv [\overline{H}]_r = H_r + Z I_{[rlast]}^T$.

else
Solve $\overline{Q}^T Z = Q_r^T R_l$.
Set $\overline{H} \equiv [\overline{H}]_l = (H_l + Z I_{[llast]}^T)^T$.
end

2.3. Properties and breakdown. A two-sided Arnoldi recursion possesses several important properties. It is easy to implement. Each one-sided block recursion is well-defined for blocks of any size, and deflation can occur independently in either or both single-sided recursions at any point in the computations. There is no requirement that rank (Q_r) = rank (Q_l) and typically they are not equal.

The vector merge computations require $\overline{Q} \equiv Q_l^T Q_r$ to have full rank. The column block Z which is generated in a merge is incorporated into the appropriate H matrix. If the right recursions are used, this modified matrix equals $\overline{H}$. If the left recursions are used, the transpose of this modified matrix equals $\overline{H}$. $\overline{H}$ is a representation of A, and these matrices can be used to define two-sided iterative methods involving A.

If $\overline{Q} \equiv Q_l^T Q_r$ does not have full rank, then the merge operation cannot be accomplished and $\overline{H}$ cannot be generated. Breakdown in Algorithm 2.2 occurs. A near-breakdown will exhibit itself as a nearly rank deficient $\overline{Q}$. However, as illustrated in Example 2.2, even if breakdown occurs, there is the possibility of recovery from breakdown without modifications of the recursions. The recursions can simply be continued until the corresponding $\overline{Q}$ has full rank.

Example 2.2. Apply Algorithm 2.2 to

$$A = \begin{bmatrix} 5 & 12 & 38 & -21 \\ 3 & 8 & 24 & -13 \\ -2 & -6 & -19 & 12 \\ -1 & -4 & -12 & 8 \end{bmatrix}, \quad l \equiv X_l = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix}, \quad r \equiv X_r = \begin{bmatrix} 7 \\ 4 \\ -3 \\ -2 \end{bmatrix}.$$

After two steps of the recursion, we obtain the left and right vectors:

$$Q_l \equiv \begin{bmatrix} l/\sqrt{3}, & [2,1,6,-1]^T/\sqrt{42}, & [-4,19,2,23]^T/\sqrt{910} \end{bmatrix}$$
$$Q_r \equiv \begin{bmatrix} r/\sqrt{78}, & [-4,7,-2,3]^T/\sqrt{78}, & [1,1,3,1]^T/\sqrt{12} \end{bmatrix}.$$

If we stop after one step of the recursion and attempt a merge, we encounter breakdown. However, if we ignore the breakdown and continue the recursions one more step, then $\overline{Q}$ has full rank.

The merge operation in Algorithm 2.2 can accept any number of left and right vectors. As indicated earlier the H matrices generated by a one-sided band Arnoldi recursion can be truncated and still retain the block upper Hessenberg structure in the truncated H-matrix and the block structure for the corresponding residual. Therefore, we can restate Algorithm 2.2 to allow for such truncation. We will use this flexibility in section 4, where we prove the equivalence of methods based upon the nonsymmetric band Lanczos recursion in [1] and Algorithm 2.3, which is a truncated version of Algorithm 2.2.

ALGORITHM 2.3. TRUNCATED TWO-SIDED BLOCK ARNOLDI RECURSION.
Specify A, X_r, X_l, ϵ_d, s.
Apply Algorithm 2.1 *to* $\{A, X_r\}$ *to generate* Q_r *with* rank $(Q_r) \geq s$.
Apply Algorithm 2.1 *to* $\{A^T, X_l\}$ *to generate* Q_l *with* rank $(Q_l) \geq s$.
Set $\widehat{Q}_r = Q_r(:, [1:s])$, $\widehat{Q}_l = Q_l(:, [1:s])$,
$\widehat{H}_r = H_r([1:s], [1:s])$, $\widehat{H}_l = H_l([1:s], [1:s])$.
Define $\widehat{R}_l, \widehat{R}_r$ *as the residual blocks needed to preserve equality in the recursions.*

Compute $\overline{\widehat{Q}} = \widehat{Q}_l^T \widehat{Q}_r$.
Solve $\overline{\widehat{Q}} \widehat{Z}_r = \widehat{Q}_l^T \widehat{R}_r$.
Solve $\overline{\widehat{Q}}^T \widehat{Z}_l = \widehat{Q}_r^T \widehat{R}_l$.
Set $[\overline{\widehat{H}}]_r = \widehat{H}_r + \widehat{Z}_r I_{[rlast]}^T$.
Set $[\overline{\widehat{H}}]_l = (\widehat{H}_l + \widehat{Z}_l I_{[llast]}^T)^T$.
end

If we exercise Algorithm 2.3, Lemma 2.3 states that the resulting $\overline{\widehat{H}}$ matrices are Petrov–Galerkin projections of A [19]. There is the implicit assumption that $\overline{\widehat{Q}}$ is nonsingular.

LEMMA 2.3. *Let* $\widehat{Q}_l$, $\widehat{Q}_r$, $[\overline{\widehat{H}}]_r$, *and* $[\overline{\widehat{H}}]_l$ *be generated by Algorithm* 2.3; *then*

$$
\begin{aligned}
[\overline{\widehat{H}}]_r &\equiv \overline{\widehat{Q}}^{-1} \widehat{Q}_l^T A \widehat{Q}_r, \\
[\overline{\widehat{H}}]_l &\equiv \widehat{Q}_l^T A \widehat{Q}_r \overline{\widehat{Q}}^{-1}, \\
[\overline{\widehat{H}}]_l &= \overline{\widehat{Q}} [\overline{\widehat{H}}]_r \overline{\widehat{Q}}^{-1}.
\end{aligned} \tag{2.2}
$$

$[\overline{\widehat{H}}]_r$ *and* $[\overline{\widehat{H}}]_l$ *are oblique projections of* A *onto* $\operatorname{span}\{\widehat{Q}_r\}$ *and along* $\operatorname{span}\{\widehat{Q}_l\}$.

Proof. The proof is a direct consequence of the fact that

$$
\widehat{Q}_l^T \overline{\widehat{R}}_r = 0 \quad \text{and} \quad \widehat{Q}_r^T \overline{\widehat{R}}_l = 0. \qquad \square
$$

2.4. Related work. Ruhe [17] introduced the two-sided Arnoldi method specifically as a method for computing approximations to left eigenvectors of a matrix A. In [17] a one-sided Arnoldi method is applied to A to obtain converged approximations $\{\theta, x_r\}$ to an eigenvalue and right eigenvector of A. The quantities generated are used to compute an approximation, x_l, to a corresponding left eigenvector of A. A second application of the one-sided Arnoldi method is applied to $\{A^T, x_l\}$ to generate a better approximation $\tilde{x}_l$ to the left eigenvector. The two applications of the Arnoldi method produce different eigenvalue approximations. To obtain a consistent triplet for approximations to an eigenvalue and to the corresponding right and left eigenvector of A, additional computations are introduced which correspond to the merge operations.

The proposed two-sided Arnoldi recursions are generalizations of the algorithm in Ruhe [17]. The recursions in Algorithm 2.2 can handle starting blocks with any number of vectors. The two applications of a one-sided Arnoldi method are exercised independently. The merge computations can handle any number of left and right vectors, as long as the corresponding matrix $\overline{Q}$ has full rank, and the resulting $\overline{H}$ matrices can be used to define a variety of iterative methods, including model reduction and eigenelement methods.

3. Two-sided block Arnoldi model reduction. We are interested in iterative methods for computing reduced-order models of large linear systems of time-invariant, differential-algebraic equations,

$$
C\dot{x} = Gx + Bu, \qquad y = E^T x. \tag{3.1}
$$

C and G are $n \times n$ matrices, where n is the order of the system. The block input matrix B is $n \times q$, where q is the number of input variables. The block output matrix

E^T is $o \times n$, where o is the number of output variables of the system. The behavior of such a system is encapsulated in the system transfer function $\mathcal{T}(s)$, which maps the Laplace transform of the input functions u to the transform of the output functions y [16].

$$y(s) = \widetilde{T}(s)u(s) \equiv E^T(sC - G)^{-1}Bu(s). \tag{3.2}$$

If $q = o = 1$, then the system is *si/so* and $\widetilde{T}(s)$ is a rational function of s. In general, a system is *mi/mo* and $\widetilde{T}(s)$ is an $o \times q$ matrix of rational functions. Each entry in $\widetilde{T}(s)$ is a *si/so* transfer function for one of the possible input/output combinations.

Typically, C is not invertible, and the matrix $(Cs - G)^{-1}B$ is replaced by a matrix $(I + \sigma F)^{-1}R$, where $F \equiv (Cs_0 - G)^{-1}C$, $R \equiv (Cs_0 - G)^{-1}B$, s_0 is some well-chosen expansion point, and $\sigma = s - s_0$. An iterative model reduction method can then be applied to the system

$$F\dot{x} = -x + Ru, \quad y = E^T x, \tag{3.3}$$

defined by $\{F, R, E\}$ to obtain smaller systems $\{\overline{F}, \overline{R}, \overline{E}\}$. The original and the reduced systems have the same number of inputs and outputs. The approximation of the smaller system to the larger system is expressed as relationships between the transfer function of the original system and the transfer function of the smaller system. We have used R as the input block for a system and also as the residual matrix in a recursion. The reader should be able to deduce which use is intended from the local context.

We use Algorithm 2.2 to define a model reduction algorithm, Algorithm 3.1, for (3.3). Formally, we can expand the transfer function $\widetilde{T}(s)$ of the original system in terms of the *moments*, $E^T F^j R$.

$$\widetilde{T}(s) \equiv E^T(I + \sigma F)^{-1}R = \sum_{j=0}^{\infty} (-1)^j E^T F^j R \sigma^j. \tag{3.4}$$

The performance of a model reduction procedure is typically measured by the number of moments of the transfer function of the reduced-order system which match the corresponding moments of the transfer function of the original system. See, for example, [10], [11]. For some $0 \le k \le M$, the moments of the reduced-order system $\{\overline{F}, \overline{R}, \overline{E}\}$ satisfy

$$\overline{E}^T \overline{F}^k \overline{R} = E^T F^k R. \tag{3.5}$$

ALGORITHM 3.1. TWO-SIDED BLOCK ARNOLDI MODEL REDUCTION.
Specify F, R, E, $\epsilon_d \ge 0$.
Apply Algorithm 2.2 *to* $\{F, R, E\}$ *for* m_r, m_l *steps to generate* $Q_r, Q_l, \overline{H}$.
Set $\overline{F} = \overline{H}$, $\quad \overline{Q} = Q_l^T Q_r$.
If rank $(Q_l) \le$ rank (Q_r)
  $\overline{R} = I_{[rfirst]} S_r$, $\quad \overline{E} = \overline{Q}^T I_{[lfirst]} S_l$.
else
  $\overline{R} = \overline{Q} I_{[rfirst]} S_r$, $\quad \overline{E} = I_{[lfirst]} S_l$.
end

There is an implicit assumption in Algorithm 3.1 that the corresponding $\overline{Q}$ has full rank. The following theorem states that if the deflation tolerance $\epsilon_d = 0$, then the

transfer function of a reduced order system, $\{\overline{F}, \overline{R}, \overline{E}\}$, obtained using Algorithm 3.1 achieves the maximum number of block moment matches to the transfer function of the original system $\{F, R, E\}$.

THEOREM 3.1. *Apply Algorithm* 3.1 *with* $\epsilon_d = 0$ *to the system* $\{F, R, E\}$ *to generate a reduced-order system* $\{\overline{F}, \overline{R}, \overline{E}\}$. *Then the first* $0 \le k \le m_l + m_r - 1$ *block moments of the reduced system match the corresponding block moments of the original system:*

$$\overline{E}^T \overline{F}^k \overline{R} = E^T F^k R \quad \textit{for} \quad 0 \le k \le m_l + m_r - 1. \tag{3.6}$$

The proof of Theorem 3.1 uses only the basic form of the recursions, for example, $AQ_r = Q_r \overline{H} + \overline{R}_r I_{[rlast]}^T$, the relationship between the modified residual matrix in one recursion and the vectors generated in the other, $Q_l^T \overline{R}_r = 0$, and the block upper Hessenberg shape of the projection matrices $\overline{H}$. The proof invokes Lemmas 3.2, 3.3, and 3.6.

Lemma 3.2 states that for small $\epsilon_d > 0$, deflation introduces correspondingly small perturbations in the one-sided Arnoldi recursions. This lemma is a direct consequence of the constructions in Algorithm 2.1. The proof of (3.7) is by induction. Equation (3.8) is an immediate consequence of the fact that for any block upper Hessenberg matrix H with diagonal block sizes $d_1, \ldots, d_m$ that for any $0 \le k < m - 1$, $\text{span}\{H_m^k I_{[first]}\}$ is contained in $\text{span}\{e_1, \ldots, e_K\}$, where $K = \sum_{j=1}^{k+1} d_j$ and e_l denotes the lth coordinate vector.

LEMMA 3.2. *After* m *steps of Algorithm* 2.1,

$$\begin{aligned} AQ_m &= Q_m H_m + R_m I_{[last]}^T + \Delta_m, \\ X &= Q_m I_{[first]} S + \widetilde{\Delta}, \end{aligned} \tag{3.7}$$

where $\|\widetilde{\Delta}\|_F \le \sqrt{d_r - d_1}\epsilon_d$ *and* $\|\widetilde{\Delta}_m\|_F \le \sqrt{d_1 - d_m}\epsilon_d$. *For* $0 \le k < m - 1$,

$$I_{[last]}^T H_m^k I_{[first]} = 0. \tag{3.8}$$

Lemma 3.3 relates the action of powers of a matrix A as applied to a starting block X to powers of a corresponding reduced-order matrix H operating on the corresponding reduced starting block. The proof is by induction and uses Lemma 3.2. This relationship will be used to prove that block moments of the transfer functions of the reduced-order systems approximate block moments of the transfer function of the original system.

LEMMA 3.3. *Let* $\{Q_m, H_m, R_m, S\}$ *be generated by applying* m *steps of Algorithm* 2.1 *to* $\{A, X\}$. *For* $0 \le k < m$,

$$\begin{aligned} A^k X &= Q_m H_m^k I_{[first]} S + (A^k \widetilde{\Delta} + \textstyle\sum_{\ell=0}^{k-1} A^\ell \Delta_m H_m^{k-1-\ell} I_{[first]} S), \\ A^m X &= Q_m H_m^m I_{[first]} S + R_m I_{[last]}^T H_m^{m-1} I_{[first]} S \\ &\quad + (A^m \widetilde{\Delta} + \textstyle\sum_{\ell=0}^{m-1} A^\ell \Delta_m H_m^{m-1-\ell} I_{[first]} S). \end{aligned} \tag{3.9}$$

COROLLARY 3.4. *Apply Algorithm* 2.2 *to* $\{A, X_r, A^T, X_l\}$. *Assume* $\text{rank}\,(Q_r) \ge \text{rank}\,(Q_l)$. *By construction* $\overline{H}$ *is block upper Hessenberg. If* $\epsilon_d = 0$, *then for* $0 \le k < m_r$,

$$\begin{aligned} A^k X_r &= Q_r \overline{H}^k I_{[rfirst]} S_r, \\ A^{m_r} X_r &= Q_r \overline{H}^{m_r} I_{[rfirst]} S + \overline{R}_r I_{[rlast]}^T \overline{H}^{m_r - 1} I_{[rfirst]} S_r. \end{aligned} \tag{3.10}$$

Corollary 3.4 follows directly from Lemma 3.3 and the fact that $\overline{H}$ retains the block upper Hessenberg form of H_r. We now consider the two-sided block Arnoldi recursion, Algorithm 2.2, with $\epsilon_d = 0$. $[\overline{H}]_r$ and $[\overline{H}]_l$ denote, respectively, $\overline{H}$ matrices which correspond to $Q_l^T \overline{R}_r = 0$ and $Q_r^T \overline{R}_l = 0$.

LEMMA 3.5. *Assume that* $Q_r, [\overline{H}]_r, \overline{R}_r, Q_l, H_l, R_l$ *satisfy*

$$\begin{aligned} AQ_r &= Q_r[\overline{H}]_r + \overline{R}_r I_{[rlast]}^T, \\ A^T Q_l &= Q_l H_l + R_l I_{[llast]}^T, \\ Q_l^T \overline{R}_r &= 0. \end{aligned}$$

For any Y_l *such that* $I_{[llast]}^T Y_l = 0$,

$$Y_l^T H_l^T Q_l^T Q_r = Y_l^T Q_l^T Q_r [\overline{H}]_r. \tag{3.11}$$

A similar statement is valid if the roles of the right and the left Arnoldi vectors are reversed. For any Y_r *such that* $I_{[rlast]}^T Y_r = 0$,

$$Y_r^T H_r^T Q_r^T Q_l = Y_r^T Q_r^T Q_l [\overline{H}]_l. \tag{3.12}$$

Proof. We prove (3.11).

$$\begin{aligned} Y_l^T H_l^T Q_l^T Q_r &= Y_l^T (Q_l H_l)^T Q_r \\ &= Y_l^T (A^T Q_l - R_l I_{[llast]}^T)^T Q_r \\ &= Y_l^T (A^T Q_l)^T Q_r \\ &= Y_l^T Q_l^T (A Q_r) \\ &= Y_l^T Q_l^T Q_r [\overline{H}]_r. \qquad \square \end{aligned}$$

LEMMA 3.6. *Let the hypotheses of Lemma* 3.5 *corresponding to* $[\overline{H}]_r$ *be satisfied. Assume that for some* X_k,

$$A^k Q_r I_{[rfirst]} = Q_r [\overline{H}]_r^k I_{[rfirst]} + \overline{R}_r X_k \qquad (0 \le k \le m_r), \tag{3.13}$$

$$(A^T)^k Q_l I_{[lfirst]} = Q_l H_l^k I_{[lfirst]} \qquad (0 \le k < m_l), \tag{3.14}$$

and that $I_{[lfirst]}^T (H_l^T)^k I_{[llast]} = 0$ *for* $0 \le k < m_l - 1$. *Then for* $0 \le k < m_l + m_r$,

$$(Q_l I_{[lfirst]})^T A^k (Q_r I_{[rfirst]}) = I_{[lfirst]}^T Q_l^T Q_r [\overline{H}]_r^k I_{[rfirst]}. \tag{3.15}$$

A similar statement is valid if the roles of the right and the left Arnoldi vectors are reversed. Assume that for some X_k,

$$(A^T)^k Q_l I_{[lfirst]} = Q_l ([\overline{H}]_l^T)^k I_{[lfirst]} + \overline{R}_l X_k \qquad (0 \le k \le m_l), \tag{3.16}$$

$$A^k Q_r I_{[rfirst]} = Q_r H_r^k I_{[rfirst]} \qquad (0 \le k < m_r), \tag{3.17}$$

and that $I_{[rfirst]}^T (H_r^T)^k I_{[rlast]} = 0$ *for* $0 \le k < m_r - 1$; *then for* $0 \le k < m_l + m_r$,

$$(Q_l I_{[lfirst]})^T A^k (Q_r I_{[rfirst]}) = I_{[lfirst]}^T [\overline{H}]_l^k Q_l^T Q_r I_{[rfirst]}. \tag{3.18}$$

Proof. We prove (3.15). By construction for $0 \le k < m_r$,

$$\begin{aligned}(Q_l I_{[lfirst]})^T A^k (Q_r I_{[rfirst]}) &= (Q_l I_{[lfirst]})^T (Q_r [\overline{H}]_r^k I_{[rfirst]} + \overline{R}_r X_k) \\ &= (Q_r^T Q_l I_{[lfirst]})^T [\overline{H}]_r^k I_{[rfirst]}.\end{aligned}$$

For $m_r \le k < m_l + m_r$, let $k = m_r + k_l$. Clearly, $0 \le k_l < m_l$ and

$$\begin{aligned}(Q_l I_{[lfirst]})^T A^k (Q_r I_{[rfirst]}) &= ((A^T)^{k_l} Q_l I_{[lfirst]})^T (A^{m_r} Q_r I_{[rfirst]}) \\ &= (Q_l H_l^{k_l} I_{[lfirst]})^T (Q_r [\overline{H}]_r^{m_r} I_{[rfirst]} + R_r X_{m_r}) \\ &= (Q_l H_l^{k_l} I_{[lfirst]})^T (Q_r [\overline{H}]_r^{m_r} I_{[rfirst]}) \\ &= I_{[lfirst]}^T (H_l^T)^{k_l} (Q_l^T Q_r) [\overline{H}]_r^{m_r} I_{[rfirst]}.\end{aligned}$$

For any $p < k_l$, define $Y_l^p \equiv H_l^p I_{[lfirst]}$. By assumption, $I_{[llast]}^T Y_l^p = 0$. If we apply Lemma 3.5 recursively for $p = 0, \ldots, k_l - 1$, we obtain $I_{[lfirst]}^T (H_l^T)^{k_l} (Q_l^T Q_r) = I_{[lfirst]}^T (Q_l^T Q_r) [\overline{H}]_r^{k_l}$. Therefore,

$$\begin{aligned}(Q_l I_{[lfirst]})^T A^k (Q_r I_{[rfirst]}) &= I_{[lfirst]}^T (H_l^T)^{k_l} (Q_l^T Q_r) [\overline{H}]_r^{m_r} I_{[rfirst]} \\ &= I_{[lfirst]}^T (Q_l^T Q_r) [\overline{H}]_r^{k_l} [\overline{H}]_r^{m_r} I_{[rfirst]} \\ &= (Q_r^T Q_l I_{[lfirst]})^T [\overline{H}]_r^k I_{[rfirst]}. \qquad \square\end{aligned}$$

Proof of Theorem 3.1. We consider the case $\overline{H} = [\overline{H}]_r$. An analogous proof applies when $\overline{H} = [\overline{H}]_l$. By construction, $FQ_r = Q_r \overline{H} + \overline{R}_r I_{[rlast]}^T$ and $Q_l^T \overline{R}_r = 0$. Therefore, from Lemma 3.3,

$$\begin{aligned}F^k Q_r I_{[rfirst]} &= Q_r \overline{H}^k I_{[rfirst]} + \overline{R}_r X_k \qquad (0 \le k \le m_r), \\ (F^T)^\ell Q_l I_{[lfirst]} &= Q_l H_l^\ell I_{[lfirst]} \qquad (0 \le \ell < m_l)\end{aligned}$$

for some consistent X_k. From Lemma 3.2, $I_{[lfirst]}^T (H_l^T)^k I_{[llast]} = 0$ for $0 \le k < m_l - 1$. From Lemma 3.6, for $0 \le k \le m_l + m_r - 1$,

$$(Q_l I_{[lfirst]})^T F^k (Q_r I_{[rfirst]}) = (Q_r^T Q_l I_{[lfirst]})^T \overline{H}^k I_{[rfirst]}.$$

However, $E = Q_l I_{[lfirst]} S_l$, $R = Q_r I_{[rfirst]} S_r$, $\overline{E} = Q_r^T Q_l I_{[lfirst]} S_l$, $\overline{R} = I_{[rfirst]} S_r$, and $\overline{F} = \overline{H}$. Therefore, for $0 \le k \le m_l + m_r - 1$,

$$\begin{aligned}E^T F^k R &= (Q_l I_{[lfirst]} S_l)^T F^k (Q_r I_{[rfirst]} S_r) \\ &= (Q_r^T Q_l I_{[lfirst]} S_l)^T \overline{H}^k I_{[rfirst]} S_r = \overline{E}^T \overline{F}^k \overline{R}. \qquad \square\end{aligned}$$

Reference [13] also uses a two-sided Arnoldi method to obtain approximations to transfer functions of control systems. The focus in [13] is on *si/so* systems, and the emphasis is on approximating Lyapunov functions [16]. Moment matching connections, as presented in this section, are not discussed. Connections with nonsymmetric Lanczos methods are mentioned but not developed. The statement is made that the results extend to block methods, but deflation is not discussed and the infeasibility of a nonsymmetric block Lanczos recursion is not acknowledged.

4. Lanczos recursions. Reference [4] focuses on relationships between nonsymmetric Lanczos and one-sided Arnoldi methods for solving $Ax = b$. In [4] it is proved that any residual norm behavior resulting from the application of the Lanczos-based biconjugate gradient method (BiCG) [19] to $Ax = b$ can be replicated by the application of the one-sided Arnoldi-based, full orthogonal method (FOM), but to a different problem: $Cy = d$. The applications $\{BiCG, Ax = b\}$ and $\{FOM, Cy = d\}$ generate identical residual norms. Reference [3] focuses on relationships between corresponding eigenelement methods.

The two-sided nature of Algorithm 2.2 leads us to ask whether or not we can prove much stronger relationships between iterative methods which are based upon it and corresponding methods which are based upon the nonsymmetric band Lanczos recursion in [1]. In this section we explore that question. We prove that corresponding two-sided iterative methods generate identical iterates. Therefore, they are simply different implementations of the same iterative methods.

The nonsymmetric band Lanczos recursions in [1] generate sets of right vectors, V_s, and left vectors, W_s, which are biorthogonal. For each s, $W_s^T V_s = D_s$ with D_s a diagonal matrix. The vectors V_s and W_s are bases for corresponding right and left subspaces spanned by sets of Krylov vectors. Typically, as the recursion accumulates information about the original problem, global biorthogonality is lost [2].

Procedures based upon nonsymmetric Lanczos recursions have been used successfully in a variety of applications. See, for example, [10], [6], [14]. However, the basic nonsymmetric Lanczos recursions may encounter *breakdown*. If there is no mismatch in the left and the right starting Lanczos vectors [15], breakdown can be circumvented by invoking *look-ahead* ideas [15], [12]. Incorporating look-ahead requires modifications in the basic Lanczos recursions.

Attempts have been made to construct nonsymmetric block Lanczos algorithms. However, it is now recognized that it is not feasible to construct nonsymmetric Lanczos recursions which are based upon explicit blocks of vectors. This difficulty is a consequence of the facts that the left and the right Lanczos vectors are biorthogonal (not independently orthogonal) so at each stage must be generated in pairs, and that as the recursions proceed, vectors within a w-block or a v-block can become dependent upon vectors generated earlier and must be deflated. Deflation does not, however, have to occur in (v, w) pairs. There can be deflations in the left (right) vector block without similar deflations in the right (left) block. If at some point in the recursions the sizes of the left and of the right blocks are not equal, then the corresponding equations which determine the biorthogonalization coefficients are overdetermined, and the recursions cannot be continued. This problem does not occur in the two-sided block Arnoldi recursion because the left and the right vectors are generated independently.

In the band nonsymmetric Lanczos recursion in [1], this problem is resolved by generating individual pairs of (v, w) vectors, one vector at a time, by alternating back and forth between the generation of a v-vector and the generation of a w-vector. One complete iteration corresponds to the generation of one v-vector and one w-vector. Therefore, at the completion of each iteration there are equal numbers of left and right Lanczos vectors.

At each iteration possible candidates for the next v-vector are drawn from an *implicit block* of vectors. The first implicit $v-$block is $\widetilde{V}_1$ obtained from the starting block $X_v = \widetilde{V}_1 S_v + \widetilde{\Delta}_v$, and similarly for $\widetilde{W}_1$ from X_w, where the S_v, S_w are constructed so that $\widetilde{V}_1$ and $\widetilde{W}_1$ are biorthogonal. The second implicit v-block consists of those v-vectors which were generated from the v-candidate vectors obtained by applying A to $\tilde{V}_1$ and invoking appropriate Gram–Schmidt biorthogonalization w.r.t.

w-vectors. The jth implicit block consists of those v-vectors which were generated from the v-candidate vectors which were obtained by applying A to $\tilde{V}_{j-1}$ and invoking biorthogonalization. If deflation occurs during the construction of some implicit block $\tilde{V}_j$, then $\text{rank}\,(\tilde{V}_j) < \text{rank}\,(\tilde{V}_{j-1})$. If no suitable candidates are found for some such block, then the recursions terminate. Analogous statements hold for the w-vector implicit blocks with A replaced by A^T.

We use d_j^v (d_j^w) to denote the number of vectors in the current implicit $\tilde{V}_j, \tilde{W}_j$ blocks. Since deflation occurs independently in the right and the left vectors, d_j^v need not equal d_j^w. If the deflation tolerance $\epsilon_d > 0$ and deflation occurs, then the recursions must be modified to include explicit biorthogonalization of each new v-vector (w-vector) w.r.t. the parents of deflated w-candidates (v-candidates). For example, if a v-candidate vector which was generated from some Av_j is deflated, then the left recursions must be modified to include explicit biorthogonalization of each new w-vector w.r.t. the parent of this candidate, v_j. If $\epsilon_d = 0$, the equalities are unaffected by any deflation and no modifications are needed.

Thus, the nonsymmetric band Lanczos algorithm is analogous to a corresponding block algorithm where the vectors within a given block are constructed one by one and this one by one construction alternates between the construction of a v-vector and a w-vector. The alternation is required to maintain the feasibility of the biorthogonalization.

In our comparisons of Lanczos-based and Arnoldi-based methods, we will assume that the deflation is exact ($\epsilon_d = 0$), that no breakdown occurs, and that the ranks of the biorthogonal left and right starting blocks are equal. If these ranks differ, then the initial phase of the band Lanczos recursions has to be modified to generate enough right or left vectors to make the number of left and right vectors equal. See [1] for details.

The banded nonsymmetric Lanczos recursion has the following matrix form:

$$\begin{aligned} AV_s &= V_s T_v + R_v I_{[vlast]}^T, \\ A^T W_s &= W_s T_w + R_w I_{([wlast])}^T. \end{aligned} \tag{4.1}$$

vlast $= [s - d_v + 1 : s]$, wlast $= [s - d_w + 1 : s]$, and d_v (d_w) denote the number of columns in the final implicit v-block (w-block). R_v and R_w are, respectively, residual blocks of vectors with d_v and d_w columns. A merge operation is not necessary since the block residuals generated satisfy $W_s^T R_v = 0$ and $V_s^T R_w = 0$. The Lanczos matrices T_v (T_w) are $s \times s$ banded matrices with maximum upper bandwidth of d_1^w (d_1^v) and maximum lower bandwidth of d_1^v (d_1^w). Typically, the bandwidths decrease as the iterations proceed.

We develop relationships between methods based upon the band nonsymmetric Lanczos recursion and methods based upon the (truncated) two-sided block Arnoldi recursion, Algorithm 2.3.

Example 4.1. Apply the nonsymmetric Lanczos recursion to the triplet $\{A, r, l\}$ defined in Example 2.2. The first two $\{v, w\}$ Lanczos pairs are

$$\begin{aligned} v_1 &= r/\sqrt{78}, & w_1 &= l/\sqrt{3}, \\ v_2 &= [4, 3, -2, -1]^T/\sqrt{30}, & w_2 &= [0, 1, 2, -1]/\sqrt{6}. \end{aligned}$$

Since $w_2^T v_2 = 0$, breakdown occurs at step 2. This coincides with the observed breakdown in Example 2.2 in the two-sided Arnoldi recursion.

For the two-sided block Arnoldi recursion, breakdown did not result in modifications in the Arnoldi recursions. Those recursions were simply continued until the merge matrix $Q_l^T Q_r$ had full rank. For the Lanczos recursions, however, breakdown does necessitate modifications in the recursions. See [1]. Breakdown is a function of the starting blocks and the associated Krylov subspaces. In Lemma 4.4 we prove that if breakdown occurs in either the Lanczos or the two-sided block Arnoldi recursions, it must occur at corresponding points in these recursions.

Lemma 4.2 states that until deflation occurs, the band nonsymmetric Lanczos recursion and the two-sided block Arnoldi recursion are generating bases for the same subspaces. Lemma 4.2 can be proved using mathematical induction with the fact that within each block of the one-sided Arnoldi recursions in the two-sided block Arnoldi recursion, candidate vectors are considered in order and one vector at a time. As defined by Algorithm 2.3, the Arnoldi recursions can be truncated at any intermediate vectors.

LEMMA 4.2. *For some s, apply the band nonsymmetric Lanczos recursion and the truncated two-sided block Arnoldi recursion, Algorithm* 2.3, *to $\{A, X_r, X_l\}$ to generate V_s, W_s, $\widehat{Q}_l, \widehat{Q}_r$. Assume no breakdown, no deflation, and exact arithmetic. Then* $\operatorname{span}(\widehat{Q}_r) = \operatorname{span}(V_s)$ *and* $\operatorname{span}(\widehat{Q}_l) = \operatorname{span}(W_s)$.

Deflation occurs only if a candidate vector is dependent upon previously generated vectors. Since at each stage, each recursion is generating vectors which span the same subspaces, if the deflation is exact, $\epsilon_d = 0$, then any deflation must occur simultaneously in both recursions.

LEMMA 4.3. *Under the hypotheses of Lemma* 4.2 *allow exact deflation, $\epsilon_d = 0$. Assume no breakdown. If deflation of some right (left) candidate vector corresponding to some Av_i ($A^T w_i$) occurs in the band Lanczos recursion, then the corresponding right (left) candidate vector corresponding to Aq_{ri} ($A^T q_{li}$) in the right (left) one-sided block Arnoldi recursion must also be deflated, and vice versa.*

Thus, with exact deflation, the corresponding subspaces generated using either the band nonsymmetric Lanczos recursion or the two-sided block Arnoldi recursion are identical. Therefore, breakdown, if it occurs, must occur simultaneously in both recursions.

LEMMA 4.4. *Under the hypotheses of Lemma* 4.3, *if breakdown occurs at step $s+1$ in the nonsymmetric band Lanczos recursion, $w_{s+1}^T v_{s+1} = 0$, and we extend the truncated two-sided block Arnoldi recursion to $s+1$ vectors, the corresponding $\widehat{Q}_l^T \widehat{Q}_r$ is singular. Similarly, if we extend the truncated two-sided block Arnoldi recursion to $s+1$ vectors and the corresponding $\widehat{Q}_l^T \widehat{Q}_r$ is singular, then extending the nonsymmetric band Lanczos recursion yields $w_{s+1}^T v_{s+1} = 0$.*

Proof. If $w_{s+1}^T v_{s+1} = 0$, then the diagonal matrix $W_{s+1}^T V_{s+1}$ is singular and the Lanczos recursions cannot be continued. By Lemmas 4.2 and 4.3, the two recursions generate bases for the same subspaces. Therefore, there exist nonsingular matrices B and C such that $W_{s+1} = \widehat{Q}_l C$ and $V_{s+1} = \widehat{Q}_r B$, and $\widehat{Q}_l^T \widehat{Q}_r = C^{-T} W_{s+1}^T V_{s+1} B$ must be singular. The argument is easily reversed. ∎

We can use either recursion, nonsymmetric band Lanczos or the truncated two-sided block Arnoldi, to construct oblique projections of the matrix A. Lemma 4.5 relates the oblique projection matrices generated by these two recursions. The corresponding matrix recursions are (4.1) and the Arnoldi recursions,

$$\begin{aligned} A\widehat{Q}_r &= \widehat{Q}_r [\overline{\widehat{H}}]_r + \overline{\widehat{R}}_r I^T_{\widehat{[rlast]}}, \\ A^T \widehat{Q}_l &= \widehat{Q}_l [\overline{\widehat{H}}]_l + \overline{\widehat{R}}_l I^T_{\widehat{[llast]}}. \end{aligned} \tag{4.2}$$

$[\widehat{rlast}]$ and $[\widehat{llast}]$ denote the indices of the columns which contain the residual matrix corresponding to the truncated right and left block Arnoldi recursions. By construction, $\widehat{Q}_l^T\overline{\widehat{R}}_r = 0$, $\widehat{Q}_r^T\overline{\widehat{R}}_l = 0$, $W_s^T R_v = 0$, and $V_s^T R_w = 0$.

LEMMA 4.5. *Apply the nonsymmetric band Lanczos recursion and the truncated two-sided block Arnoldi recursion to* $\{A, X_r, X_l\}$ *using exact deflation. Define corresponding projection matrices* T_v, T_w *and* $\{[\overline{\widehat{H}}]_r, [\overline{\widetilde{H}}]_l\}$ *as defined in* (4.2). *Then there exists nonsingular matrices* B, C *such that* $T_v = B^{-1}[\overline{\widetilde{H}}]_r B$ *and* ${T_w}^T = C^T[\overline{\widetilde{H}}]_l C^{-T}$.

Iterative methods based upon these recursions compute approximations to quantities associated with the original problem by solving reduced-order problems associated with these projection matrices. We use Lemma 4.5 to prove that eigenelement and model reduction methods defined using these recursions generate identical iterates. Therefore, they are different implementations of the same methods.

4.1. Computing eigenvalues/eigenvectors. We will use $[\theta^L, z_r^L, z_l^L]$ and $[\theta^A, z_r^A, z_l^A]$ to denote approximations to eigenvalues and to corresponding right and left eigenvectors of A generated by a Lanczos or an Arnoldi procedure. The two-sided block Arnoldi methods can be defined using either Algorithm 2.2 or Algorithm 2.3. In the comparisons we need to work with the same number of left and right vectors in the Lanczos and in the Arnoldi methods, so we use Algorithm 2.3.

ALGORITHM 4.1. TWO-SIDED BLOCK ARNOLDI EIGENELEMENT ALGORITHM.
Specify A, X_r, X_l, ϵ_d, s.
Apply Algorithm 2.3 *to* $\{A, X_r, X_l\}$.
Compute $[\overline{\widehat{H}}]_r u_r = \theta u_r$ *and* $[\overline{\widehat{H}}]_l^T u_l = \theta u_l$.
Compute $z_r \equiv \widehat{Q}_r u_r$, $z_l \equiv \widehat{Q}_l u_l$.
Compute error estimates $\epsilon_r = \overline{\widehat{R}}_r u_r([\widehat{rlast}])$ *and* $\epsilon_l = \overline{\widehat{R}}_l u_l([\widehat{rlast}])$.

Lemma 4.5 tells us that the eigenvalues computed using $[\overline{\widetilde{H}}]_r$ and $[\overline{\widetilde{H}}]_l^T$ are identical. We define a corresponding band Lanczos eigenelement algorithm.

ALGORITHM 4.2. TWO-SIDED BAND LANCZOS EIGENELEMENT ALGORITHM.
Specify A, X_r, X_l, ϵ_d, s.
Apply the nonsymmetric band Lanczos recursions to $\{A, X_r, X_l\}$.
Compute $T_v u_r = \theta u_r$ *and* $T_w u_l = \theta u_l$.
Compute $z_r \equiv V u_r$, $z_l \equiv W u_l$.
Compute unnormalized error estimates $\epsilon_r = R_v u_r([vlast])$, $\epsilon_l = R_w u_l([wlast])$.

4.2. Model reduction. Similarly, we can define methods for model reduction of linear systems. See (3.1). Algorithm 3.1 specifies a two-sided block Arnoldi model reduction method. This definition maps directly onto a corresponding model reduction method which is based upon Algorithm 2.3.

ALGORITHM 4.3. BAND NONSYMMETRIC LANCZOS MODEL REDUCTION.
Specify F, R, E, $\epsilon_d \geq 0$, s.
Apply the nonsymmetric band Lanczos recursions to generate V_s, W_s, T_v, T_w.
Set $\overline{F} = T_v$, $\quad \overline{R} = I_{[vfirst]} S_v^L$, $\quad \overline{E} = (V_s^T W_s) I_{[wfirst]} S_w^L$.

4.3. Equivalences between two-sided Arnoldi and Lanczos methods.

THEOREM 4.6. *Set* $\epsilon_d = 0$. *Apply Algorithm* 4.2 *to* $\{A, X_r, X_l\}$ *to generate eigenelement approximations* $\{\theta_j^L, z_{rj}^L, z_{lj}^L\}$. *Apply Algorithm* 4.1 *to* $\{A, X_r, X_l\}$ *to generate eigenelement approximations* $\{\theta_j^A, z_{rj}^A, z_{lj}^A\}$. *Then the eigenvalue approximations and the corresponding left and right Ritz vectors generated by these two algorithms are identical.*

Proof. By Lemma 4.2 there exist nonsingular B, C such that $V_s = \widehat{Q}_r B$ and $W_s = \widehat{Q}_l C$. By Lemma 4.5, $T_v = B^{-1}[\overline{\widehat{H}}]_r B$ and $T_w = C^{-1}[\overline{\widehat{H}}]_l^T C$. Furthermore, $T_w^T = (W_s^T V_s) T_v (W_s^T V_s)^{-1}$. Therefore, these two procedures generate identical eigenvalue approximations. Moreover, each $u_{rj}^A = B u_{rj}^L$. Therefore, $z_{rj}^A = \widehat{Q}_r u_{rj}^A = \widehat{Q}_r B u_{rj}^L = V_s u_{rj}^L = z_{rj}^L$, and similarly for z_{lj}^A, z_{lj}^L. ∎

Theorem 4.7 states that corresponding Lanczos and Arnoldi model reduction algorithms generate identical approximations to the transfer function of the original system.

THEOREM 4.7. *Let $\{F, R, E\}$ be a mi/mo system defined by* (3.3). *Apply the band nonsymmetric Lanczos model reduction procedure, Algorithm* 4.3, *to $\{F, R, E\}$ to obtain the reduced-order system $\{\overline{F}^L, \overline{R}^L, \overline{E}^L\}$, where $\overline{F}^L \equiv \overline{T}^L \equiv T_v$. Apply the truncated two-sided block Arnoldi model reduction procedure to $\{F, R, E\}$ to obtain the reduced-order system $\{\overline{F}^A, \overline{R}^A, \overline{E}^A\}$ of the same size, where $\overline{F}^A \equiv \overline{\widehat{H}}^A \equiv [\overline{\widehat{H}}]_r$. The transfer functions of the Arnoldi and of the Lanczos reduced-order systems are equal to*

$$\widetilde{\mathcal{T}}^A(\sigma) \equiv [\overline{E}^A]^T (I + \sigma \overline{\widehat{H}}^A)^{-1} \overline{R}^A, \tag{4.3}$$

$$\widetilde{\mathcal{T}}^L(\sigma) \equiv [\overline{E}^L]^T (I + \sigma \overline{T}^L)^{-1} \overline{R}^L. \tag{4.4}$$

For all complex σ,

$$\widetilde{\mathcal{T}}^A(\sigma) = \widetilde{\mathcal{T}}^L(\sigma). \tag{4.5}$$

Proof. By construction,

$$\widetilde{\mathcal{T}}^A(\sigma) \equiv (S_l^A)^T I_{[\widehat{lfirst}]}^T \overline{\widehat{Q}} (I + \sigma \overline{\widehat{H}}^A)^{-1} I_{[\widehat{rfirst}]} S_r^A. \tag{4.6}$$

From Lemmas 4.2 and 4.5, there exist nonsingular, upper triangular matrices B, C such that

$$V_s = \widehat{Q}_r B, \quad W_s = \widehat{Q}_l C, \quad \overline{\widehat{H}}^A = B \overline{T}^L B^{-1}. \tag{4.7}$$

Let

$$C_1 = I_{[\widehat{lfirst}]}^T C I_{[\widehat{lfirst}]}, \quad B_1 = I_{[\widehat{rfirst}]}^T B I_{[\widehat{rfirst}]}.$$

Since C and B are upper triangular, $C_1^{-1} = (C^{-1})_1$ and $B_1^{-1} = (B^{-1})_1$. Therefore, by (4.7),

$$\begin{aligned} I_{[\widehat{lfirst}]}^T \overline{\widehat{Q}} &= (I_{[\widehat{lfirst}]}^T \widehat{Q}_l^T) \widehat{Q}_r = I_{[\widehat{lfirst}]}^T (W_s C^{-1})^T V_s B^{-1} \\ &= (W_s I_{[\widehat{lfirst}]} C_1^{-1})^T V_s B^{-1} = C_1^{-T} I_{[\widehat{lfirst}]}^T W_s^T V_s B^{-1}. \end{aligned} \tag{4.8}$$

By construction,

$$\begin{aligned} X_r &= \widehat{Q}_r I_{[\widehat{rfirst}]} S_r^A = V_s I_{[vfirst]} S_v^L, \\ X_l &= \widehat{Q}_l I_{[\widehat{lfirst}]} S_l^A = W_s I_{[wfirst]} S_w^L. \end{aligned}$$

Therefore,

$$
\begin{aligned}
S_r^A &= I_{[\widehat{rfirst}]} Q_r^T V_s I_{[vfirst]} S_v^L = I_{[\widehat{rfirst}]} B I_{[vfirst]} S_v^L = B_1 S_v^L, \\
S_l^A &= I_{[\widehat{lfirst}]} Q_l^T W_s I_{[wfirst]} S_w^L = C_1 S_w^L.
\end{aligned} \tag{4.9}
$$

Using (4.8), (4.9) in (4.6), we obtain

$$
\begin{aligned}
\widetilde{\mathcal{T}}_A(\sigma) &= (S_l^A)^T C_1^{-T} I_{[\widehat{lfirst}]}^T W_s^T V_s B^{-1} (I + \sigma B \overline{T}^L B^{-1})^{-1} I_{[\widehat{rfirst}]} S_r^A \\
&= (S_w^L)^T I_{[\widehat{lfirst}]}^T W_s^T V_s (I + \sigma \overline{T}^L)^{-1} I_{[\widehat{rfirst}]} S_v^L = \widetilde{\mathcal{T}}^L(\sigma). \qquad \square
\end{aligned}
$$

5. Summary. We have proposed new two-sided block Arnoldi recursions which are extensions of the work in [17] for use in iterative methods. Iterative methods which are based upon these recursions have the advantage that any breakdown is centered in a vector merge matrix, and that breakdown can be handled without requiring modifications to the recursions. We used these two-sided block Arnoldi recursions to define a model reduction procedure which was proved to have maximum block moment matching properties. In comparisons of eigenelement and model reduction algorithms based upon these two-sided Arnoldi recursions and the band nonsymmetric Lanczos recursion in [1], we proved that the corresponding methods produce identical iterates. Therefore, they are different implementations of the same method.

Acknowledgments. The authors would like to thank the referees and the editor for their helpful comments.

REFERENCES

[1] J. I. Aliaga, D. L. Boley, R. W. Freund, and V. A. Hernández, *A Lanczos-type method for multiple starting vectors*, Math. Comp., 69 (2000), pp. 1577–1601.

[2] Z. Bai, *Error analysis of the Lanczos algorithm for the nonsymmetric eigenvalue problem*, Math. Comp., 62 (1994), pp. 209–226.

[3] J. Cullum, *Arnoldi versus nonsymmetric Lanczos algorithms for solving matrix eigenvalue problems*, BIT, 36 (1996), pp. 470–493.

[4] J. Cullum, *Iterative methods for solving $Ax = b$, GMRES/FOM versus QMR/BiCG*, Adv. Comp. Math., 6 (1996), pp. 1–24.

[5] J. Cullum and W. E. Donath, *A block Lanczos algorithm for computing the q algebraically-largest eigenvalues and a corresponding eigenspace for large, sparse symmetric matrices*, in Proceedings of the 1974 IEEE Conference on Decision and Control, IEEE Press, New York, 1974, pp. 505–509.

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

[7] J. Cullum, A. Ruehli, and T. Zhang, *A method for reduced-order modeling and simulation of large interconnect circuits and its application to PEEC models with retardation*, IEEE Trans. Circuits and Systems II, 47 (2000), pp. 261–273.

[8] J. Cullum and R. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations. Vol. 1, Theory*, Birkhäuser, Basel, Switzerland, 1985.

[9] J. Cullum and R. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations. Vol. 2, Programs*, Birkhäuser, Basel, Switzerland, 1985.

[10] P. Feldman and R. W. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Trans. Computer-Aided Design, 14 (1995), pp. 639–649.

[11] R. W. Freund, *Computation of matrix Padé approximations of transfer functions via a Lanczos-type process*, in Approximation Theory VIII, Vol. 1, Approximation and Interpolation, C. K. Chui and L. L. Schumaker, eds., World Scientific, River Edge, NJ, 1995, pp. 215–222.

[12] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
[13] I. M. Jaimoukha and E. M. Kasenally, *Oblique projection methods for large scale model reduction*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 602–627.
[14] W. Kerner, *Large-scale complex eigenvalue problems*, J. Comput. Phys., 85 (1989), pp. 1–85.
[15] B. N. Parlett, D. R. Taylor, and Z. Liu, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[16] R. V. Patel, A. J. Laub, and P. M. van Dooren, *Numerical Linear Algebra Techniques for Systems and Control*, IEEE Press, Piscataway, NJ, 1994.
[17] A. Ruhe, *The two-sided Arnoldi algorithm for nonsymmetric eigenvalue problems*, in Matrix Pencils, Lecture Notes in Math. 973, B. Kågström and A. Ruhe, eds., Springer-Verlag, New York, 1983, pp. 104–120.
[18] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.
[19] Y. Saad, *Iterative Methods for Sparse Linear System*, PWS Publishing, Boston, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 320–333

ON "SLUGGISH TRANSIENTS" IN MARKOV CHAINS*

COLM O'CINNEIDE†

Abstract. The distance between the powers P^m of an aperiodic stochastic matrix and their limit behaves roughly like ρ^m as $m \to \infty$, where ρ is the maximum modulus of the eigenvalues whose moduli are less than 1. G. W. Stewart noted (see [*Stochastic Models*, 31 (1997), pp. 85–94]) that when there are defective eigenvalues that are close to 1 in modulus, the powers of P may initially display slower convergence than might be expected based on the magnitudes of the eigenvalues alone. Stewart introduced a quantity σ that has a bearing on the strength of this effect. Numerical experimentation led him to suggest that σ cannot be too large. We derive upper bounds on σ which help to explain Stewart's empirical observations.

Key words. Markov chain, rate of convergence, Jordan block

AMS subject classifications. 60J10, 15A18, 15A51

PII. S0895479899355359

1. Introduction.

This paper concerns "sluggish transients" as described by Stewart [25]. To explain this idea, suppose P is an order-n aperiodic, irreducible stochastic matrix with steady-state distribution π defined by $\pi P = \pi, \sum \pi_i = 1$. Let e denote a column vector of 1's. The entries of the powers P^m differ from their limits πe by a sum of terms of the form $m^k \lambda^m$, where $k < n$ is a nonnegative integer and λ is an eigenvalue of P with modulus less than 1. If there is such a term with $k > 0$ (which implies that the corresponding eigenvalue λ is defective) and $|\lambda|$ close to 1, the factor m^k may get quite large before λ^m gets small. The result is that convergence to steady state may initially appear much slower than the crude estimate ρ^m, where

$$\rho = \mathrm{sp}(P - e\pi),$$

which might be guessed based on the magnitudes of the eigenvalues alone. (Here, ρ may be described as the maximum of the moduli of the eigenvalues of modulus less than 1.) This slow convergence, arising from a defective eigenvalue, is what is meant here by a sluggish transient. A quantity σ, defined in (3) below for the 2-norm, measures the strength of this effect in the case of a defective eigenvalue of order 2 ($k = 1$ above). A numerical investigation led Stewart to suggest that σ cannot be too large. In section 3 we place an upper bound on σ for a general matrix A, and in section 4 we generalize this bound to defective eigenvalues of order ≥ 2. In the last two sections we present a discussion of *invariant polytopes* and use them to derive more refined bounds in the case of a stochastic matrix P, taking into account the order of the matrix. These results help to explain Stewart's empirical observations.

There is substantial literature on convergence to steady state of Markov chains to be found in various probability and theoretical computer science journals. See [16] for a canonical example and [23] for an overview and references to the literature. Much of this literature is directed toward the goal of identifying provably polynomial-time Markov-chain Monte Carlo algorithms for important computational problems. The

*Received by the editors April 26, 1999; accepted for publication (in revised form) by C.D. Meyer July 11, 2001; published electronically August 5, 2002. This research was supported in part by NSF grant DMI-9713730.

http://www.siam.org/journals/simax/24-2/35535.html

†School of Industrial Engineering, Purdue University, West Lafayette, IN 47907-1287 (colm@ecn.purdue.edu).

chains on which these algorithms are based are usually constructed to be *reversible*, which is to say the transition matrix P is self-adjoint with respect to $L^2(\pi)$ (with norm $||x||_{2,\pi} = \sum \pi_i x_i^2$ [7]). The eigenvalues of such a chain are never degenerate, and yet, even in this case, the inadequacy of eigenvalue information to capture important features of convergence to steady state is well documented. The "cutoff phenomenon" [3, 5, 6] illustrates some of the subtleties of convergence to steady state. In this, the total variation distance to steady state remains close to 1 for some time, and then, at a "cutoff point" (which is strictly a cutoff *point* only after a limit in which "dimension" goes to infinity), the exponential decay prescribed by the second eigenvalue takes over. See [13] for insights into the cutoff phenomenon through pseudospectra.

The literature on convergence rates of Markov chains focuses on the reversible case, primarily for two reasons. First, many useful Markov-chain Monte Carlo algorithms are based on reversible chains. Second, the mathematical framework is substantially simpler in the reversible case. The paper [10] is an exception, treating nonreversible chains by exploiting a relation between the 2-norm distance to steady state and singular values (with respect to $L^2(\pi)$) of the transition matrix. A good illustration of the complexities of the nonreversible case is Aldous's observation [1, 2] that, without reversibility, bounding ρ away from 1 does not imply good mixing properties, in the sense that, even in steady state, strong "correlations" may persist at long time lags. Thus the simple eigenvalue bound of [1, Proposition 4.1] on the variance of a sample average for a reversible chain in steady state has no analogue for more general chains.

Stewart's sluggish transients concern this nonreversible case, since reversible chains do not have degenerate eigenvalues. His paper [25] indicates a natural direction to explore in trying to understand the approach to steady state in this more general setting. The issue at hand, in the view of the present author, is to understand how the behavior of $P^n - \pi e$ is different from that of the corresponding quantity in which P is replaced with a general matrix A which shares with P the property of having 1 as a simple dominant eigenvalue. The goal here is to take a first step in this direction. The invariant polytope techniques of sections 5 and 6 allow us to take this first step, which involves a detailed analysis of the action of P on the maximal invariant subspace associated with some degenerate eigenvalue. The analysis hints that there is much structure to be exploited. Perhaps this is a good first step, in light of the many extremality properties of the Jordan block [4, 19, 22], which characterizes the action of P on an irreducible invariant subspace. To date, however, the invariant polytope methods have proved unwieldy for eigenspaces of (real) dimension more than 2.

2. Preliminaries. We review some basic facts about invariant subspaces, which may be found, for example, in [11, Chapter XIII.3]. Let A be an order-n square matrix, viewed as a linear transformation $x \to Ax$ on $\boldsymbol{C}^n$. We call an invariant subspace of A *irreducible* if it is not expressible as a direct sum of nontrivial invariant subspaces. Any irreducible invariant subspace W of A is *cyclic*, and so has a basis of column vectors $w_0, w_1, \ldots, w_k$ for which

$$Aw_{i-1} = \lambda w_{i-1} + w_i, \quad i = 1, 2, \ldots, k; \qquad Aw_k = \lambda w_k, \tag{1}$$

where $0 \le k \le n-1$ and λ is a scalar. Thus w_k is a right eigenvector of A with eigenvalue λ.

Suppose now that the space W is *maximal* in the sense that it is not contained in any larger irreducible invariant subspace. Then there is a complementary invariant subspace W', so that $W \oplus W' = \boldsymbol{C}^n$. Corresponding to the w_i's, there is then a

unique collection of row vectors $v_0, v_1, \ldots, v_k$ orthogonal to W' ($v_i w' = 0$ for $w' \in W'$) such that

$$v_i w_j = \delta_{ij},$$

δ_{ij} being the Kronecker delta. Moreover,

$$v_0 A = \lambda v_0; \qquad v_i A = \lambda v_i + v_{i-1}, \quad i = 1, 2, \ldots, k. \tag{2}$$

The question of sluggish transients has to do with how fast the norms of the w_i's can grow as i increases. For the $k = 1$ case, the issue is how large the norm of w_1 can be when w_0 is scaled to unit norm, or, considering the 2-norm as in Stewart [25], to bound the quantity

$$\sigma_2 \equiv \frac{||w_1||_2}{||w_0||_2}. \tag{3}$$

(We consider this quantity with respect to various norms, identifying the norm in question with a subscript in the obvious way.) Based on numerical experimentation, Stewart posed the question of whether, for A *stochastic*, values of σ_2 greater than $2(1 - |\lambda|)$ were possible. We prove in the next section that, in the stochastic case, $\sigma_2 \leq 6.3(1 - |\lambda|)$. In section 4 we give a general bound on the growth of the norms of the w_i's as i increases. In section 6 we give an improved bound in the $k = 1$ case assuming that the matrix A is stochastic and that λ is real and positive. This bound depends on the order n of the matrix.

Remark. Our definition of σ differs in a minor way from Stewart's. In fact, our σ is Stewart's multiplied by $|\lambda|$, and so is a little smaller in the stochastic case. This seems to make the analysis and the final result slightly more natural. Of course, in the stochastic case it is the large eigenvalues ($|\lambda| \approx 1$) that most interest us, and for these there is little difference between the two σ's.

For a stochastic matrix P, we need the fact that $||P|| \leq 1$ for certain norms, for example, the (p, π)-norm

$$||w||_{p,\pi} \equiv \left(\sum_{i=1}^{n} \pi_i w_i^p \right)^{1/p}.$$

The next proposition follows easily from definitions and Jensen's inequality.

PROPOSITION 1. *$||P||_\infty = 1$ for any stochastic matrix P, and if P is irreducible with steady-state distribution π, then $||P||_{p,\pi} \leq 1$ for $p \geq 1$.*

3. A bound on σ in the case $k = 1$. Throughout this section and the next, we have a fixed but arbitrary vector norm, and of course our matrix norm is the operator norm that it induces. Let A be a matrix and suppose that (1) holds with $k = 1$. We extract a natural bound on $\sigma \equiv ||w_1||/||w_0||$. We emphasize that A is not necessarily stochastic in the following.

It is convenient to work with the matrix $B = (\bar{\lambda}/|\lambda|)A$ in place of A. (Here and below, take $\bar{\lambda}/|\lambda|$ to be 1 if $\lambda = 0$ so that then $B = A$.) Then B has $|\lambda|$ as a defective eigenvalue, and (1) holds with B in place of A and with the w_i's replaced by

$$\hat{w}_i \equiv \left(\frac{\bar{\lambda}}{|\lambda|} \right)^i w_i, \quad i = 0, 1, 2, \ldots, k.$$

In the case $k = 1$ we have

$$B^m \hat{w}_0 = |\lambda|^m \hat{w}_0 + m|\lambda|^{m-1}\hat{w}_1 \text{ for } m = 0, 1, 2, \ldots.$$

Defining e^{tB} by its Taylor series for t real, it follows that

$$e^{tB}\hat{w}_0 = e^{|\lambda|t}(\hat{w}_0 + t\hat{w}_1).$$

Let us write a for $||A|| = ||B||$. Then $||e^{tB}|| \leq e^{at}$ for $t > 0$, and we argue that

$$e^{at}||\hat{w}_0|| \geq ||e^{tB}\hat{w}_0|| = e^{|\lambda|t}\,||\hat{w}_0 + t\hat{w}_1||.$$

Thus, as $||\hat{w}_i|| = ||w_i||$,

$$e^{(a-|\lambda|)t}\,||w_0|| \geq t||w_1|| - ||w_0||$$

or

$$(1 + e^{(a-|\lambda|)t})||w_0|| \geq t||w_1||.$$

As this is true for all positive t, we must have $|\lambda| < a$. We define $y = (a - |\lambda|)t > 0$ and continue with

$$(a - |\lambda|)\frac{1 + e^y}{y}||w_0|| \geq ||w_1|| \text{ for all } y > 0.$$

Defining

$$\mathcal{C}_1 = \min_{y>0} \frac{1 + e^y}{y} \approx 3.67 \leq 1 + e,$$

it follows that

$$\mathcal{C}_1(a - |\lambda|)||w_0|| \geq ||w_1||.$$

Recalling that $a \equiv ||A||$, we have proved the following.

THEOREM 1. *For a square matrix A and nonzero vectors w_0 and w_1 satisfying $Aw_0 = \lambda w_0 + w_1$ and $Aw_1 = \lambda w_1$, we have*

$$\sigma \equiv \frac{||w_1||}{||w_0||} \leq \mathcal{C}_1(||A|| - |\lambda|).$$

Specializing to the case $A = P$ stochastic, and taking the norm to be the ∞-norm or the $(2, \pi)$-norm of Proposition 1 so that $||P|| \leq 1$, we have also proved the following.

COROLLARY. *For an irreducible stochastic matrix P and nonzero vectors w_0 and w_1 satisfying $Pw_0 = \lambda w_0 + w_1, Pw_1 = \lambda w_1$, we have*

$$\sigma_\infty \equiv \frac{||w_1||_\infty}{||w_0||_\infty} \leq \mathcal{C}_1(1 - |\lambda|) \quad \textit{and} \quad \sigma_{2,\pi} \equiv \frac{||w_1||_{2,\pi}}{||w_0||_{2,\pi}} \leq \mathcal{C}_1(1 - |\lambda|).$$

Working from the first of these, we get a bound on σ_2 addressing Stewart's numerical examples. Since, for any n-vector v,

$$\frac{1}{\sqrt{n}}||v||_2 \leq ||v||_\infty \leq ||v||_2,$$

we conclude from the corollary that for P stochastic

$$\sqrt{n}\mathcal{C}_1(1-|\lambda|)||w_0||_2 \geq ||w_1||_2.$$

Thus

$$\sigma_2 \leq \sqrt{n}\mathcal{C}_1(1-|\lambda|),$$

and in Stewart's 3×3 examples ($n=3$) we are assured that

$$\sigma_2 \leq \sqrt{3}\mathcal{C}_1(1-|\lambda|) \leq 6.3(1-|\lambda|). \tag{4}$$

The "critical value" of σ_2 proposed by Stewart is $2(1-|\lambda|)$, and the bound shows that it cannot exceed this critical value by more than a modest multiple.

4. A generalization ($k \geq 1$). We return to the general setting of section 2, where we have a square matrix A and an irreducible invariant space W spanned by the vectors in (1). Again we replace A by the matrix $B = A\bar{\lambda}/|\lambda|$ and set $\hat{w}_i = (\bar{\lambda}/|\lambda|)^i w_i$. The following formulas give in essence the powers and the exponential of a Jordan block:

$$B^m w_i = \sum_{j=i}^{k} |\lambda|^{m-j+i} \binom{m}{j-i} \hat{w}_j \text{ for } i = 0,1,\ldots,k,$$

$$e^{tB}\hat{w}_i = e^{|\lambda|t}\left(\hat{w}_i + t\hat{w}_{i+1} + \frac{t^2}{2!}\hat{w}_{i+2} + \cdots + \frac{t^{k-i}}{(k-i)!}\hat{w}_k\right) \text{ for } i = 0,1,\ldots,k. \tag{5}$$

In this section we prove the following.

THEOREM 2. *Let A be a square matrix and let $w_0, w_1, \ldots, w_k$ denote vectors for which* (1) *holds. There are positive absolute constants $\mathcal{C}_i, i = 1,2,\ldots,$ such that*

$$\mathcal{C}_{k-j}(||A|| - |\lambda|)||w_j|| \geq ||w_{j+1}||, \quad j = 0,1,2,\ldots,k-1; \tag{6}$$

$$\mathcal{D}_{j,\ell}(||A|| - |\lambda|)^{\ell-j}||w_j|| \geq ||w_\ell||, \quad 0 \leq j \leq \ell \leq k, \text{ where } \mathcal{D}_{j,\ell} = \prod_{h=j}^{\ell-1} \mathcal{C}_{k-h}. \tag{7}$$

Proof. Note first that (7) is a direct consequence of (6), because

$$\begin{aligned}\mathcal{D}_{j,\ell}(||A|| - |\lambda|)^{\ell-j}||w_j|| &= \mathcal{D}_{j+1,\ell}(||A|| - |\lambda|)^{\ell-j-1}\{(||A|| - |\lambda|)\mathcal{C}_{k-j}||w_j||\} \\ &\geq \mathcal{D}_{j+1,\ell}(||A|| - |\lambda|)^{\ell-j-1}||w_{j+1}|| = \cdots \geq ||w_\ell||,\end{aligned}$$

applying (6) repeatedly. It remains only to prove (6).

Again, let a denote $||A|| = ||B||$. Since $||e^{tB}|| \leq e^{at}$, we have from (5)

$$e^{at}||\hat{w}_i|| \geq e^{|\lambda|t}\left|\left|\hat{w}_i + t\hat{w}_{i+1} + \frac{t^2}{2!}\hat{w}_{i+2} + \cdots + \frac{t^{k-i}}{(k-i)!}\hat{w}_k\right|\right|,$$

which, using $||w_i|| = ||\hat{w}_i||$, at once gives

$$e^{t(||A||-|\lambda|)}||w_i|| \geq t||w_{i+1}|| - ||w_i|| - \frac{t^2}{2!}||w_{i+2}|| - \cdots - \frac{t^{k-i}}{(k-i)!}||w_k||. \tag{8}$$

To begin an induction argument, Theorem 1 gives (6) for $j = k$, with $\mathcal{C}_1$ as defined in section 3. Now suppose that we have defined the constants $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{k-i-1}$ and

that (6) is true for $j \geq i+1$. Then (7) is also true for $j \geq i+1$ by the argument of the first paragraph of this proof. We will next deduce that (6) holds for $j = i$, with a suitable choice of C_{k-i}, and this will complete the proof. We define $y \equiv t(a - |\lambda|)$ as before. By first reorganizing terms in (8) and multiplying by $a - |\lambda|$, and then using (7) for $j \geq i+1$, we deduce that

$$\begin{aligned}
&(e^y + 1)(a - |\lambda|)||w_i|| \\
&\quad \geq (a - |\lambda|)t||w_{i+1}|| - (a - |\lambda|)\left(\frac{t^2}{2!}||w_{i+2}|| + \cdots + \frac{t^{k-i}}{(k-i)!}||w_k||\right) \\
&\quad \geq (a - |\lambda|)t||w_{i+1}|| \\
&\qquad - \left(\mathcal{D}_{i+1,i+2}\frac{(a - |\lambda|)^2 t^2}{2!} + \cdots + \mathcal{D}_{i+1,k}\frac{(a - |\lambda|)^{k-i} t^{k-i}}{(k-i)!}\right)||w_{i+1}|| \\
&\quad \geq y||w_{i+1}|| - \left(\mathcal{D}_{i+1,i+2}\frac{y^2}{2!} + \cdots + \mathcal{D}_{i+1,k}\frac{y^{k-i}}{(k-i)!}\right)||w_{i+1}||.
\end{aligned}$$

Rewriting, we have

$$\frac{e^y + 1}{y - \left(\mathcal{D}_{i+1,i+2}\dfrac{y^2}{2!} + \cdots + \mathcal{D}_{i+1,k}\dfrac{y^{k-i}}{(k-i)!}\right)}\{(a - |\lambda|)||w_i||\} \geq ||w_{i+1}||$$

for any value of $y > 0$ for which the denominator on the left is positive. Now define $\mathcal{C}_{k-i}$ as the minimum of the coefficient of $(a - ||\lambda||)||w_i||$ on the left side, as y ranges over positive values for which the denominator polynomial is positive. (The set of such values is nonempty, as it includes small positive values of y.) Then (6) holds for $j = i$. This completes the induction argument. □

5. Invariant polytopes. Henceforth our matrix A is assumed to be stochastic and so is denoted by P. In this section we discuss the idea of an *invariant polytope* [18, 19] (see [12] for some related results), which is used in the next section to refine Theorem 1, giving a bound that depends on n, the order of the matrix P. That analysis may be viewed as an extension of the Dmitriev–Dynkin bounds [8] on the eigenvalues of a stochastic matrix, at least insofar as the arguments are based on a similar geometric idea.

Let U be a finite-dimensional vector space and T a linear transformation on U. Let C be a *polytope* (by which we mean the convex hull of a finite number of points) in U. Then C is an *invariant polytope* for T if

$$TC \subset C.$$

Invariant polytopes arise naturally in the context of Markov chains, stochastic matrices, and generators [8, 9, 12, 19] and played a key role in the proof of the *characterization of phase-type distributions* [18, 21] (see the earlier [14, 24] for the "discrete case," shown in [17] to be equivalent to the "continuous case").

Now consider a stochastic matrix P acting as a linear transformation through multiplication on the right ($x \to xP$) on $\boldsymbol{C}^n$. Suppose V is a maximal irreducible invariant subspace of P with dimension $k+1$, spanned by the $v_0, v_1, \ldots, v_k$ of (2). If $k \geq 1$, then λ is a defective eigenvalue. Let V' be the invariant subspace complementary to V. Let Π denote the operator projecting $\boldsymbol{C}^n$ onto V parallel to V'. That is, for any vector $x \in \boldsymbol{C}^n$, there is a unique decomposition $x = v + v', v \in V, v' \in V'$, and $x\Pi$ is defined as v. Note that $P\Pi = \Pi P$.

Let e_i denote the ith unit (row) vector. Then P has as a natural invariant polytope the unit simplex

$$S = \mathrm{co}\{e_1, e_2, \ldots, e_n\}$$

(here, co denotes *convex hull*) in the sense that $SP \subset S$. This is because the rows of P are in S. The projection of S under Π, denoted by $C = S\Pi$, is also an invariant polytope under P because

$$CP = (S\Pi)P = S(\Pi P) = S(P\Pi) = (SP)\Pi \subset S\Pi = C.$$

The next-to-last step is because S is an invariant polytope for P. We have $C \subset V$, and if we represent the action of P on V as a matrix with respect to the basis $v_k, v_{k-1}, \ldots, v_0$ (taken in this order) of V, the result is the Jordan block

$$J = \begin{pmatrix} \lambda & 1 & 0 & \ldots & 0 & 0 \\ 0 & \lambda & 1 & \ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & \lambda & 1 \\ 0 & 0 & 0 & \ldots & 0 & \lambda \end{pmatrix}, \quad \text{a } (k+1) \times (k+1) \text{ matrix.}$$

With respect to the basis $v_k, v_{k-1}, \ldots, v_0$, the extreme points of C have coordinates given by the *rows* of the matrix

$$\tilde{W} \equiv (w_k; w_{k-1}; \ldots; w_0),$$

whose columns are the vectors of (1). Now the relation

$$\tilde{W}J = P\tilde{W}$$

shows that the action of J on the rows of W results in vectors that are convex combinations of the rows of W (and so are in C). This means that when we embed the set C in the $(k+1)$-dimensional complex space $\boldsymbol{C}^{k+1}$ via the basis $v_k, v_{k-1}, \ldots, v_0$, the resulting polytope is invariant under right multiplication by the Jordan block J. This is the manner in which we use invariance in the next section.

There are other invariant polytopes to consider besides C. Let Π_i denote the projection of $\boldsymbol{C}^n$ onto $V_i \equiv \mathrm{span}\{v_{k-i}, v_{i+2}, \ldots, v_k\}$ parallel to $V' \cup \mathrm{span}\{v_0, v_1, \ldots, v_{k-i-1}\}$. In particular, $\Pi_k = \Pi$ and $V_k = V$. V_i is *not* invariant under P, but we do have $\Pi_i P \Pi_i = P\Pi_i$. Now the polytope $C_i \equiv S\Pi_i$ is invariant under $P_i \equiv P\Pi_i$, because

$$C_i P_i = (S\Pi_i)(P\Pi_i) = S(\Pi_i P \Pi_i) = SP\Pi_i \subset S\Pi_i = C_i.$$

By (2), we see that, with respect to the basis $v_k, v_{k-1}, \ldots, v_{k-i+1}$ of V_i, P_i has the representation

$$J_i = \begin{pmatrix} \lambda & 1 & 0 & \ldots & 0 & 0 \\ 0 & \lambda & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda & 1 \\ 0 & 0 & 0 & \ldots & 0 & \lambda \end{pmatrix}, \quad \text{an } (i+1) \times (i+1) \text{ matrix.}$$

Also, the coordinates of the extreme points of C_i with respect to this basis are the rows of the matrix

$$\tilde{W}_i \equiv (w_k; w_{k-1}; \ldots; w_{k-i}), \text{ for which } \tilde{W}_i J_i = P\tilde{W}_i.$$

So for each $i \leq k$ we have identified a polytope C_i having no more than n extreme points which, when viewed as subsets of $\boldsymbol{C}^{i+1}$ in a natural way, is invariant under a Jordan block J_i.

6. Exploiting invariant polytopes. This section is devoted to the proof of the following theorem, using the invariant polytopes of the previous section. This result refines Theorem 1 for the case of the ∞-norm with A stochastic and λ real and positive by replacing the constant $\mathcal{C}_1$ by a smaller quantity depending on the order n of the stochastic matrix P. There is a satisfying consistency between the two results in that the constant of Theorem 1 arises here as the limit of the $\mathcal{C}^{(n)}$'s.

THEOREM 3. *For a stochastic matrix P of order $n \geq 3$ and nonzero real vectors x and y satisfying $Py = \lambda y + x$ and $Px = \lambda x$, for some λ, $0 \leq \lambda < 1$, we have*

$$\mathcal{C}^{(n)}(1-\lambda)||y||_\infty \geq ||x||_\infty,$$

where $1/\mathcal{C}^{(n)} = s_{n-2}$ is the unique solution to (17) *below with $N = n-2$. Moreover, we have*

$$\mathcal{C}^{(1)} < \mathcal{C}^{(2)} < \cdots, \text{ and } \lim_{n\to\infty} \mathcal{C}^{(n)} = \mathcal{C}_1,$$

where $\mathcal{C}_1$ is as defined in section 3.

The hypotheses of the theorem imply that λ is a defective eigenvalue of P, and that (1) holds for some $k \geq 1$ and certain vectors w_i with $y = w_{k-1}$ and $x = w_k$ where $W = \text{span}\{w_0, w_1, \ldots, w_k\}$ is a *maximal* irreducible invariant subspace. Recalling the notation of the previous section, let us consider the action of $P_1 \equiv P\Pi_1$ (multiplying on the right $x \to xP_1$) on the subspace V_1 spanned by v_k, v_{k-1}. Expressed with respect to the natural basis $\{v_k, v_{k-1}\}$, the invariant polytope C_1 defined in the previous section is given by

$$C_1 = \text{co}\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

in which the n points listed are the rows of the $n \times 2$ matrix $\tilde{W}_1 \equiv (w_k, w_{k-1}) = (x; y)$. C_1 is then a polytope with no more than n extreme points (it could have fewer) in $\mathbb{R}^2$. C_1 is invariant under multiplication on the right by

$$J_1 = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

For $\epsilon \geq 0$, let

$$\Phi_\epsilon \equiv I + \frac{\epsilon}{1-\lambda}(J_1 - I) = \begin{pmatrix} 1-\epsilon & \epsilon\theta \\ 0 & 1-\epsilon \end{pmatrix}, \text{ where } \theta \equiv \frac{1}{1-\lambda}. \tag{9}$$

For a point $z \in \mathbb{R}^2$, we define

$$z\Phi_{\mathbb{R}^+} \equiv \{z\Phi_\epsilon, \epsilon \geq 0\},$$

which is the ray originating at z and going through zJ_1. This ray may be thought of as *the direction in which J_1 "points" from z*, and is a key to the geometry of invariant polytopes for J_1. The manner in which we exploit the invariance of C_1 under J_1 is the following simple result.

PROPOSITION 2. *For $z \in C_1$, $z\Phi_\epsilon$ is in C_1 for $0 \leq \epsilon \leq 1-\lambda$.*

This is a direct consequence of the invariance property of C_1, along with its convexity. This result is expressed by saying that J_1 *points inward to* C_1 *at each point of* C_1. (See the "invariant polytope lemma" of [18], to which the inward-pointing property is central; [12] discusses related ideas.)

H denotes the right half plane and L denotes the ray of slope θ extending in the positive direction from the origin:

$$H \equiv \{(x, y) \mid x \geq 0\}; \quad L \equiv \{(x, y) \mid y = \theta x, x \geq 0\}.$$

We now enumerate some further elementary properties of C_1 and the transformations Φ_ϵ that are needed in what follows.

PROPOSITION 3. *The origin is in* C_1.

This is because C_1 is invariant under J_1 and $J_1^i \to 0$ as $i \to \infty$.

PROPOSITION 4. *Let* $z' = (x', y') = z\Phi_\epsilon$ *where* $z = (x, y), x \neq 0$, *and* $0 \leq \epsilon \neq 1$, *and let* $r = y/x$ *and* $r' = y'/x'$. *Then*

$$r' = r + \frac{\epsilon}{1 - \epsilon}\theta. \tag{10}$$

This is a simple calculation. Thus Φ_ϵ has the effect of increasing the ratio $r = y/x$ for a point $z = (x, y)$ by a positive constant for $\epsilon < 1$.

PROPOSITION 5. *For* $z = (x, y)$, *the point at which the ray* $z\Phi_{\mathbb{R}^+}$ *intersects the y-axis is*

$$z\Phi_1 = (x, y)\Phi_1 = (0, \theta x). \tag{11}$$

This is independent of the y-coordinate of z. Thus

(a) *if the point* $z = (x, y) \in H$ *lies on the ray* L $(y = \theta x)$, *then the ray* $z\Phi_{\mathbb{R}^+}$ *points horizontally to the left:* $(x, \theta x)\Phi_1 = (0, \theta x)$;

(b) *for points* $z \in H$ *below the ray* L $(y < \theta x)$, *the ray* $z\Phi_{\mathbb{R}^+}$ *points "northwest" (in the direction of decreasing x and increasing y)*;

(c) *for points* $z \in H$ *above the ray* L $(y > \theta x)$, *the ray* $z\Phi_{\mathbb{R}^+}$ *points "southwest" (in the direction of decreasing x and decreasing y)*.

These again follow by simple calculation.

Identify an extreme point $e_1 = (\chi_1, \eta_1)$ of C_1 whose x-coordinate is maximal in magnitude. Of course, e_1 is among the points (x_i, y_i), $i = 1, 2, \ldots, n$, and

$$|\chi_1| = ||x||_\infty.$$

This is nonzero by hypothesis. Now by replacing C_1 by $-C_1$ if necessary, another polytope invariant under J_1, we may suppose without loss of generality that

$$\chi_1 = ||x||_\infty > 0,$$

and, in particular, that $e_1 \in H$. Suppose for a moment that

$$|\eta_1| \geq \theta\chi_1, \text{ which implies that } ||y||_\infty \geq \theta||x||_\infty.$$

We shall see that the $\mathcal{C}^{(n)}$'s of Theorem 3 are ≥ 1, and so this supposition implies the conclusion of the theorem. Therefore *it is enough to prove the theorem under the condition that*

$$|\eta_1| \leq \theta\chi_1, \tag{12}$$

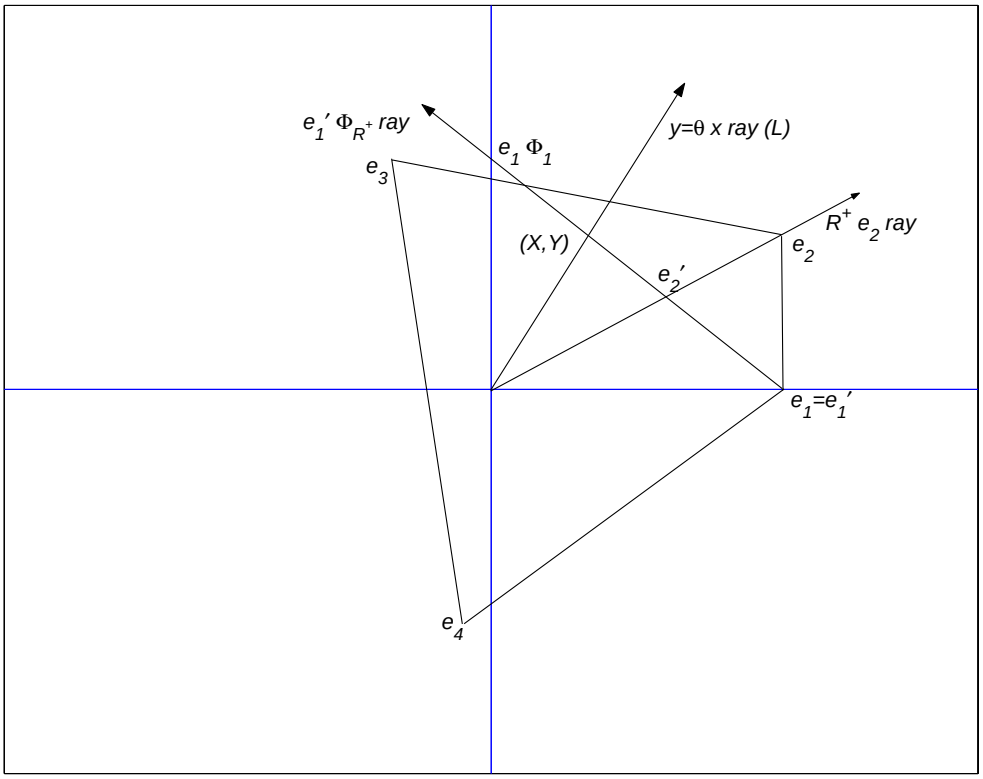


FIG. 1. *The invariant polytope* $C_1 \subset \mathbb{R}^2$ (*with* $N = 2$).

which we henceforth assume to hold.

Let us enumerate the extreme points of C_1, starting at e_1 and proceeding counterclockwise about the origin, but stopping just before we cross the $y = \theta x$ ray L. (See Figure 1.) Let N be the number of such points. At (12) we have assumed that $N \geq 1$. We denote these N extreme points by $e_1, e_2, \ldots, e_N$, again in the counterclockwise order.

These N extreme points lie in the wedge formed by the rays $y = \pm\theta x, x \geq 0$. Under (12) we must have $N \leq n - 2$. This may be seen as follows. C_1 must have at least one extreme point f on or above the rays L and $e_N\Phi_{\mathbb{R}^+}$, for otherwise the ray $e_N\Phi_{\mathbb{R}^+}$ does not point inward to C_1 as it must by Proposition 2. C_1 must have yet another extreme point outside of the wedge between the rays $y = \pm\theta x, x \geq 0$, because for any point, such as f, above the rays L and $e_N\Phi_{\mathbb{R}^+}$ (in fact, for any point in the right half plane above L or any point in the upper left quadrant), *the ray* $f\Phi_{\mathbb{R}^+}$ *does not intersect this wedge,* and so $\Phi_{\mathbb{R}^+}$ cannot point inward to C_1 from f, as it must by Proposition 2, unless there is another extreme point.

We set $e_1' = (\chi_1', \eta_1') \equiv e_1$. For $i = 2, 3, \ldots, N$, we define $e_i' = (\chi_i', \eta_i')$ to be the point of intersection of the ray $e_{i-1}'\Phi_{\mathbb{R}^+}$ and the ray $\mathbb{R}^+e_i \equiv \{\epsilon e_i \mid \epsilon \geq 0\}$ from the origin through e_i (see Figure 1). Proposition 5(b) ensures that these two rays do intersect. Thus e_i' is a positive multiple (≤ 1) of e_i. Since $e_i' \in H$, by (11) there are $\epsilon_i, i = 1, 2, \ldots, N-1$, $0 < \epsilon_i < 1$, with

$$e_{i+1}' = e_i'\Phi_{\epsilon_i}.$$

These points e_i' are in C_1 because C_1 is convex and so contains any point on the line segment from 0 to e_i by Proposition 3.

As e_N is the last extreme point encountered before we cross L as we move counterclockwise around the boundary of C_2 from e_1, the ray $e_N'\Phi_{\mathbb{R}^+}$ is within C_1 at least until it reaches the ray L, which it does at the point

$$(X,Y) \equiv \left(\frac{1}{2-r_N/\theta}\right)(\chi_N', \theta\chi_N'), \text{ where } r_i \equiv \frac{\eta_i}{\chi_i} = \frac{\eta_i'}{\chi_i'}, \quad i = 1,2,\ldots,N, \tag{13}$$

by simple algebra using (9). As $(X,Y) \in C_1$, Y is a lower bound on the largest y-coordinate in magnitude in C_1:

$$||y||_\infty \geq Y. \tag{14}$$

The r_i's are increasing as i increases, by Proposition 4. Moreover, we have $r_N \leq \theta$, because each of the e_i's is below the $y = \theta x$ ray L in the right half plane. We also have

$$e_N' = e_1' \prod_{i=1}^{N-1} \Phi_{\epsilon_i},$$

which we can calculate directly using

$$\prod_{i=1}^{N-1} \Phi_{\epsilon_i} = \left(\prod_{i=1}^{N-1}(1-\epsilon_i)\right)\begin{pmatrix} 1 & \theta\sum_1^n \epsilon_i/(1-\epsilon_i) \\ 0 & 1 \end{pmatrix} = \frac{1}{\prod_{i=1}^{N-1}(1+\alpha_i)}\begin{pmatrix} 1 & \theta\sum_1^n \alpha_i \\ 0 & 1 \end{pmatrix},$$

where $\alpha_i = \epsilon_i/(1-\epsilon_i)$. Thus $e_N' = (\chi_N', \eta_N')$ is given by

$$\chi_N' = \frac{\chi_1}{\prod_{i=1}^{N-1}(1+\alpha_i)} \quad \text{and} \quad \eta_N' = \frac{1}{\prod_{i=1}^{N-1}(1+\alpha_i)}\left[\chi_1'\left(\theta\sum_1^n \alpha_i\right) + \eta_1'\right],$$

and so

$$r_N \equiv \eta_N'/\chi_N' = r_1 + \theta\sum_1^{N-1} \alpha_i. \tag{15}$$

(This is also a consequence of (10).) This allows us to identify Y using (13) as

$$Y = \frac{\theta}{2-r_N/\theta}\chi_N = \left(\frac{\theta}{2-r_N/\theta}\right)\frac{\chi_1}{\prod_{i=1}^{N-1}(1+\alpha_i)} = \left(\frac{\theta}{2-r_N/\theta}\right)\frac{||x||_\infty}{\prod_{i=1}^{N-1}(1+\alpha_i)}. \tag{16}$$

This is a lower bound on $||y||_\infty$ by (14). However, it depends on the unknown ϵ_i's. To remove this dependency, we simply *choose the* $\epsilon_i, i = 1,2,\ldots,N-1$, *so as to minimize* Y *subject to the constraint that* $r_N \leq \theta$. We first solve this minimization problem with r_N constrained to take a fixed value $\leq \theta$. This leads to the following maximization problem:

$$\begin{aligned} &\text{Maximize } \prod_{i=1}^{N-1}(1+\alpha_i) \\ &\text{subject to } \sum_1^{N-1} \alpha_i = \frac{r_N - r_1}{\theta} \text{ and } \alpha_i \geq 0,\ i = 1,2,\ldots,N-1. \end{aligned}$$

By (15), the constraint here says simply that as the α_i's are allowed to vary, r_N remains fixed. Lagrange multipliers give the solution

$$\alpha_i = \frac{r_N - r_1}{(N-1)\theta}, \quad i = 1, 2, \ldots, N-1,$$

leading via (16) to the following lower bound on Y:

$$Y \geq \theta ||x||_\infty \left(\frac{1}{2 - r_N/\theta}\right) \left(\frac{(N-1)\theta}{(N-1)\theta + r_N - r_1}\right)^{N-1} \cdots.$$

We choose r_N now to minimize this, leading to

$$r_N = r_N^* = \frac{\theta(N-1) + r_1}{N},$$

which we note is in the range $[r_1, \theta] \subset [-\theta, \theta]$. Substituting this into the right side of the previous expression above allows us to continue

$$\begin{aligned} \cdots &\geq \theta ||x||_\infty \left(1 + \frac{1 - r_1/\theta}{N}\right)^{-N} \\ &= \theta ||x||_\infty \left(1 + \frac{1-s}{N}\right)^{-N}, \text{ where } s \equiv r_1/\theta = \eta_1/(\theta ||x||_\infty). \end{aligned}$$

This becomes

$$Y \geq \theta \phi_N(s) ||x||_\infty, \text{ where } \phi_N(s) \equiv \left(1 + \frac{1-s}{N}\right)^{-N}.$$

Recalling (14) we deduce that

$$||y||_\infty \geq \theta \phi_N(s) ||x||_\infty.$$

Since $s \equiv \eta_1/(||x||_\infty \theta)$, it follows that $\theta |s|\, ||x||_\infty = |\eta_1|$, but, since $(\chi_1, \eta_1) \in C_1$, $|\eta_1|$ is a lower bound on $||y||_\infty$ and so

$$||y||_\infty \geq \theta |s|\, ||x||_\infty.$$

The last two inequalities combine to give

$$||y||_\infty \geq \theta \min\{|s|, \phi_N(s)\} ||x||_\infty.$$

Now $-1 \leq s \leq 1$ by the assumption (12), and the minimum of $\max\{|t|, \phi_N(t)\}$ over the range $-1 \leq t \leq 1$ occurs at the point $-s_N < 0$, where s_N is the unique solution to

$$\phi_N(-t) = t. \tag{17}$$

This is because $\phi_N(t)$ is increasing in t and $0 < \phi_N(-1) < 1$. (See Figure 2.) Thus

$$||y||_\infty \geq \theta s_N ||x||_\infty.$$

The ϕ_N's decrease in N at each argument value, and so the s_N's decrease also. Since $N \leq n-2$, we finally get that

$$||y||_\infty \geq \theta s_{n-2} ||x||_\infty.$$

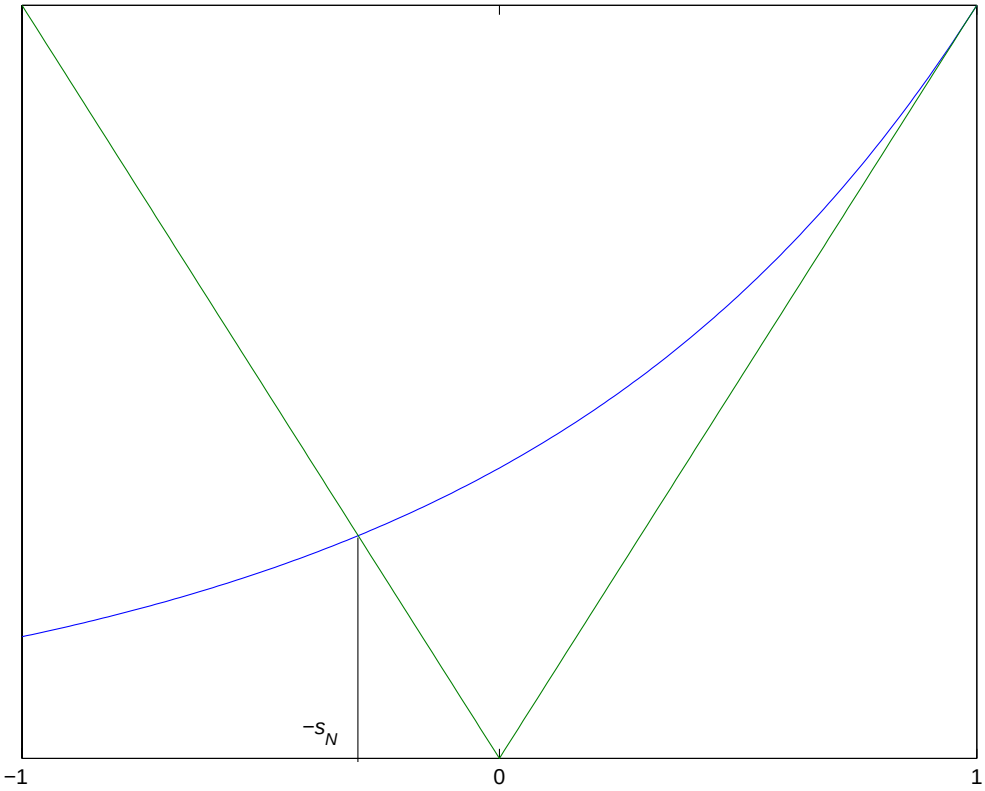


FIG. 2. *The graphs of* $t \to \phi_N(t)$ *and* $t \to |t|$.

This is the inequality of the theorem.

The s_N's decrease toward a limit s_∞ which is the solution to

$$\phi_\infty(-s) = s, \text{ where } \phi_\infty(s) \equiv \lim_{N\to\infty} \phi_N(s) = e^{s-1}.$$

It is elementary that $s_\infty = \mathcal{C}_1^{-1}$. This completes the proof of Theorem 3. □

Finally, let us compute s_1, which relates to the $n = 3$ case. This is the solution to

$$\phi_1(-s) = s \quad \text{or} \quad \frac{1}{2+s} = s,$$

which is $s_1 = \sqrt{2}-1 \approx .41$. Thus in Stewart's 3×3 case, arguing as just before (4) and making the assumption that λ is real and positive so that Theorem 3 is applicable, we get

$$\sigma_2 \le \frac{\sqrt{3}}{\sqrt{2}-1}(1-\lambda) \le 4.2(1-\lambda),$$

which improves (4), but at the expense of assuming that the matrix in question is stochastic.

If one tries to carry out the analysis here with a complex λ, the main difficulty that arises is that one has to deal with a four-dimensional (real) vector space rather than a two-dimensional one, and the analysis is bound to become much more technical.

7. Concluding remark. Much is known about the eigenvalues of stochastic matrices, but there are some simple questions awaiting answers. See [15, section 4], for an especially intriguing one. The geometric idea of an invariant polytope gives certain insights into such questions, as we saw here; for further examples, see [8, 9, 19, 20, 21, 22]. In all of these, the order n plays a central role.

REFERENCES

[1] D. Aldous, *On the Markov chain method for uniform combinatorial distributions and simulated annealing*, Probab. Engrg. Inform. Sci., 1 (1987), pp. 33–46.

[2] D. Aldous, *Finite-time implications of relaxation times for stochastically monotone processes*, Probab. Theory Related Fields, 77 (1988), pp. 137–145.

[3] D. Aldous and P. Diaconis, *Shuffling cards and stopping times*, Amer. Math. Monthly, 93 (1986), pp. 333–348.

[4] J.V. Burke, A.S. Lewis, and M.L. Overton, *Optimal stability and eigenvalue multiplicity*, Found. Comput. Math., 1 (2001), pp. 205–225.

[5] P. Diaconis, *The cutoff phenomenon in finite Markov chains*, Proc. Natl. Acad. Sci., 93 (1996), pp. 1659–1664.

[6] P. Diaconis and L. Saloff-Coste, *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Probab., 6 (1996), pp. 695–750.

[7] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–41.

[8] N. Dmitriev and E.B. Dynkin, *On the characteristic numbers of a stochastic matrix*, C. R. (Doklady) Acad. Sci. URSS (N.S.), 49 (1945), pp. 159–162 (in Russian).

[9] N. Dmitriev and E.B. Dynkin, *On the characteristic numbers of stochastic matrices*, Bull. Acad. Sci. URSS. Sér. Math. [Izvestia Akad. Nauk SSSR], 10 (1946), pp. 167–184 (in Russian with English summary).

[10] J.A. Fill, *Eigenvalue bounds on convergence to stationarity for non-reversible Markov chains, with applications to the exclusion process*, Ann. Appl. Probab., 1 (1991), pp. 64–87.

[11] W.H. Greub, *Linear Algebra*, Springer-Verlag, New York, 1975.

[12] A. Heller, *On stochastic processes derived from Markov chains*, Ann. Math. Statist., 36 (1965), pp. 1286–1291.

[13] G.J. Jonsson and L.N. Trefethen, *A numerical analyst's look at the cutoff phenomenon in card shuffling and other Markov chains*, in Numerical Anaysis 1997, D.F. Griffiths, D.J. Higham, and G.A. Watson, eds., Longman, Harlow, UK, 1998.

[14] T. Katayama, M. Okamoto, and H. Enomoto, *Characterization of the structure-generating functions of regular sets and the DOL growth condition*, Inform. and Control, 36 (1978), pp. 85–101.

[15] J.F.C. Kingman, *Three unsolved problems in discrete Markov theory*, in Stochastic Analysis: A Tribute to the Memory of Rollo Davidson, John Wiley, London, 1973, pp. 180–191.

[16] L. Lovász and M. Simonovits, *Random walks in a convex body and an improved volume algorithm*, Random Structures Algorithms, 4 (1993), pp. 359–412.

[17] R.S. Maier, *The algebraic construction of phase-type distributions*, Stochastic Models, 7 (1991), pp. 573–602.

[18] C.A. O'Cinneide, *Characterization of phase-type distributions*, Stochastic Models, 6 (1990), pp. 1–57.

[19] C.A. O'Cinneide, *Phase-type distributions and invariant polytopes*, Adv. Appl. Probab., 23 (1991), pp. 515–535.

[20] C.A. O'Cinneide, *Phase-type distributions and majorization*, Ann. Appl. Probab., 1 (1991), pp. 219–227.

[21] C.A. O'Cinneide, *Triangular order of triangular phase-type distributions*, Stochastic Models, 9 (1993), pp. 507–530.

[22] C.A. O'Cinneide, *Phase-type distributions: Open problems and a few properties*, Stochastic Models, 15 (1999), pp. 731–757.

[23] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, New York, 1995.

[24] M. Soittola, *Positive rational sequences*, Theoret. Comput. Sci., 2 (1976), pp. 317–322.

[25] G.W. Stewart, *On Markov chains with sluggish transients*, Stochastic Models, 31 (1997), pp. 85–94.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 334–355

CHARACTERIZATION OF STATIONARY DISCRETE-TIME GAUSSIAN RECIPROCAL PROCESSES OVER A FINITE INTERVAL*

BERNARD C. LEVY† AND AUGUSTO FERRANTE‡

Abstract. This paper examines the class of stationary discrete-time multivariate Gaussian reciprocal processes defined over a finite interval $[0, N]$. The matrix covariance function of such processes obeys a second-order self-adjoint difference equation whose structure is described by a symplectic matrix pencil. The canonical form of symplectic matrix pencils obtained in [Ferrante and Levy, *Linear Algebra Appl.*, 274 (1998), pp. 259–300] is employed to characterize and classify stationary Gaussian reciprocal processes. It is shown that each class of n-dimensional reciprocal processes with fixed reciprocal dynamics is parametrized by n real parameters.

Key words. Gaussian reciprocal processes, symplectic matrix pencils, covariance matrices

AMS subject classifications. 60G10, 60G12, 60G15, 15A21, 15A22, 15A57

PII. S0895479801368622

1. Introduction.

The class of reciprocal (or quasi-Markov) processes is a natural generalization of Markov processes and is particularly useful for modeling random signals indexed by space instead of time. It was first introduced by Bernstein [2] to "restore the symmetry between the past and the future" in an attempt at modeling quantum mechanics phenomena without imposing a preferred time direction.

Recall that a stochastic process $x(t)$ defined on a linearly ordered time interval I is *Markov* if, for any $t_0 \in I$, the past and the future (with respect to t_0) of the process are conditionally independent given $x(t_0)$. The same process is said to be *reciprocal* if, for any $t_0, t_1 \in I$ with $t_0 < t_1$, the process in the interior of the interval (t_0, t_1) and the process in the exterior of the same interval are conditionally independent given the boundary values $x(t_0)$ and $x(t_1)$. We refer to [13] for a more precise mathematical formulation. Observe that Markov processes are necessarily reciprocal, but there exists [12, 5, 3, 14] reciprocal processes that are not Markov. The class of reciprocal processes is thus larger than the Markov class, and it naturally extends to the multidimensional case where the parameter set of the process is not linearly ordered. In fact multidimensional Markov random fields [22] reduce in one dimension to reciprocal processes, not Markov processes.

During the last fifteen years, significant progress has been made in characterizing the properties, dynamics, and conservation laws satisfied by reciprocal processes; see, for example, [14, 9, 16, 18, 19, 1, 15, 4, 21, 6, 7] and the references therein. The first results concerning the classification of scalar stationary continuous-time Gaussian reciprocal processes were obtained by Jamison [12] and were later completed by Chay [5] and Carmichael, Massé, and Theodorescu [3]. For all processes appearing in this classification, second-order stochastic differential equations with Dirichlet boundary conditions were obtained in [16]. In this respect, it is worth noting that once the

*Received by the editors August 22, 2001; accepted for publication (in revised form) by U. Helmke April 29, 2002; published electronically October 18, 2002.
http://www.siam.org/journals/simax/24-2/36862.html

†Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 (levy@ece.ucdavis.edu).

‡Dipartimento di Elettronica e Informatica, Università di Padova, via Gradenigo 6/A, 35131 Padova, Italy (augusto@dei.unipd.it). The research of this author was partially performed at the Institute of Theoretical Dynamics, University of California, Davis, with support provided by a CNR fellowship.

equation satisfied by a scalar Gaussian reciprocal diffusion (GRD) is fixed, the class of stationary diffusions is parametrized by a single parameter that can be used to fix the lifetime of the process. Jamison's classification was extended recently to multivariate stationary GRDs in [17], where it was shown that n-dimensional stationary GRDs with fixed reciprocal dynamics are parametrized by n real parameters. The results of [17] rely heavily on the Hamiltonian structure of the matrix describing the dynamics and conservation laws of GRDs.

The results of this paper represent the discrete-time counterpart of those presented in [17]; specifically, we obtain a characterization and classification of discrete-time multivariable stationary Gaussian reciprocal processes. However, the extension of results from the continuous-time to the discrete-time is not straightforward, since it relies on a detailed characterization of the structure of symplectic matrix pencils, which until the work of [25] and [8] was not fully understood. The main difficulty is that discrete-time reciprocal dynamics may include modes at zero and infinity that cannot arise in the continuous-time case, thus making the structure of discrete-time reciprocal processes more complex. We use as a starting point the second-order self-adjoint difference equation derived in [18] for the covariance of a discrete-time Gaussian reciprocal process. This equation can be rewritten as a first-order descriptor system of twice the dimension of the original system. The matrix pencil corresponding to this descriptor system has a symplectic structure. The canonical form of symplectic matrix pencils presented in [8] proves to be a convenient tool for parametrizing the solutions of the covariance equation. It allows the derivation of a parametric form for the solutions obeying the self-adjointness constraint $R(t) = R^T(-t)$, where T denotes the matrix transpose. As a consequence, the class of covariances corresponding to n-dimensional stationary processes with fixed reciprocal dynamics is shown to be parametrized by only n real parameters. We believe that the results derived here constitute a useful first step towards the development of a reciprocal stochastic realization theory similar to the one presented in [20] for the Markov case.

The paper is organized as follows: The second-order dynamics of reciprocal processes are reviewed in section 2, where they are converted into an equivalent first-order descriptor system. The matrix pencil describing the descriptor system has a symplectic structure, and the canonical form of symplectic pencils obtained in [8] is presented in section 3. This canonical form is then employed in section 4 to characterize and parametrize the covariances of multivariable stationary reciprocal processes with fixed reciprocal dynamics. Our results are illustrated by examples in section 5. Finally, some concluding remarks and issues requiring further research are presented in section 6.

2. Reciprocal dynamics. Let $x(t) \in \mathbb{R}^n$ be a zero-mean, discrete-time, Gaussian reciprocal process defined over the interval $[0, N]$, with covariance $R(k, s) = E[x(k)x^T(s)]$, where $E[\cdot]$ denotes mathematical expectation. If the process x is *nonsingular*, the covariance matrix

$$\mathcal{R}_N := E\begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N) \end{bmatrix}\begin{bmatrix} x(0)^T & x(1)^T & \dots & x(N)^T \end{bmatrix} \tag{2.1}$$

is invertible, and it is shown in [18] that there exist matrices $M_0(k)$ and $M(k)$ such that $R(k, s)$ satisfies the second-order difference equation

$$M^T(k)R(k-1,s) + M_0(k)R(k,s) + M(k+1)R(k+1,s) = I\delta(k-s) \tag{2.2}$$

for $t = k-s = 0, \pm 1, \pm 2, \ldots, \pm(N-1)$, with $\delta(\cdot)$ being the *Kronecker* function defined by

$$\delta(t) = \begin{cases} 1, & t = 0, \\ 0, & t \neq 0. \end{cases} \tag{2.3}$$

The goal of this paper is to characterize the covariances $R(k,s)$ corresponding to *stationary* processes, for which

$$R(k,s) = R(k-s) \tag{2.4}$$

depends only on the difference $k-s$. It is easy to check that if $x(t)$ is stationary, the matrices $M(k)$ and $M_0(k)$ are constant. Specifically, setting $s = k-1$, $s = k$, and $s = k+1$ in (2.2) and taking into account (2.4), we easily get the identity

$$[M^T(k) \;\; M_0(k) \;\; M(k+1)]\mathcal{R}_2 = [0 \;\; I \;\; 0], \quad k \in [-N, N], \tag{2.5}$$

where $\mathcal{R}_2$ is defined as in (2.1). This implies

$$[M^T(k) \mid M_0(k) \mid M(k+1)] = [M^T \mid M_0 \mid M] = [0 \mid I \mid 0]\mathcal{R}_2^{-1}, \tag{2.6}$$

so $M_0(k) = M_0$ and $M(k) = M$ do not depend on k.

Thus in the stationary case the covariance $R(t)$ must satisfy the recursion

$$MR(t+1) + M_0R(t) + M^TR(t-1) = I\delta(t) \tag{2.7}$$

for $0 \leq |t| \leq N-1$. Observe from (2.6), in light of the nonsingularity assumption, that M_0 must be positive definite and hence nonsingular. Therefore by performing the transformation

$$\bar{x}(t) = M_0^{1/2}x(t), \tag{2.8}$$

we can assume $M_0 = I$ without any loss of generality. Our goal is to characterize the covariance functions $R(t)$ satisfying

$$MR(t+1) + R(t) + M^TR(t-1) = I\delta(t) \tag{2.9}$$

for $0 \leq |t| \leq N-1$. The recursion (2.9) specifies what we call the second-order reciprocal dynamics of the process $x(t)$. With M fixed, different covariances satisfying this equation are said to belong to the same reciprocal class. Covariances in the same class differ only by the selection of Dirichlet boundary conditions specified by $R(N)$ and $R(-N) = R^T(N)$.

The classification of multivariate stationary Gaussian reciprocal processes that we present proceeds in two phases. First, we characterize the structure of the reciprocal dynamics (2.9). Then, assuming that the dynamics are fixed, we classify all stationary processes in the same reciprocal class. It turns out that stationary covariances in a given reciprocal class are parametrized by n real parameters, where n denotes the dimension of the process $x(t)$. As a benchmark, each reciprocal class contains only one stationary Markov process, and in the scalar case, the classification obtained by Jamison and others depends on a single scalar real parameter.

To characterize the structure of the second-order recursion (2.9) it is convenient to rewrite it as the first-order descriptor system

$$\begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} R(t+1) \\ R_1(t+1) \end{bmatrix} = \begin{bmatrix} -I & -M^T \\ I & 0 \end{bmatrix} \begin{bmatrix} R(t) \\ R_1(t) \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} \delta(t)\,, \tag{2.10}$$

with $0 \le |t| \le N-1$, where we denote $R_1(t) := R(t-1)$.

We now analyze the eigenstructure of the pencil $sE - tA$ with

$$E := \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} \quad \text{and} \quad A := \begin{bmatrix} -I & -M^T \\ I & 0 \end{bmatrix} \tag{2.11}$$

specifying the descriptor dynamics (2.10). This structure will be used to parametrize the solutions $R(t)$ of (2.10) or, equivalently, of (2.9).

We start by observing that the pencil $sE - tA$ is *symplectic*, since by denoting

$$K := \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}, \tag{2.12}$$

where I_n denotes the identity of dimension n, the relation

$$E^T K E = A^T K A \tag{2.13}$$

holds. For more information on symplectic matrix pencils, see [25, 8] and the references therein.

Furthermore, if the interval length N is sufficiently large, the pencil $sE - tA$ must be *regular*, i.e., its determinant does not vanish identically. To prove this fact, assume by contradiction that $\det(sE - tA) \equiv 0$ or, equivalently, that $\det M(z) \equiv 0$, where we define $M(z) := z^2 M^T + zI + M$. Then there exists [10] a vector polynomial $p(z) = p_0 + p_1 z + \cdots + p_r z^r$ with $p_0 \neq 0$ such that $M(z)p(z) \equiv 0$. Equating to zero the coefficients of z^i with $i = 1, 2, \ldots, r+1$ gives

$$\begin{bmatrix} I & M & 0 & 0 & \dots & 0 \\ M^T & I & M & 0 & \dots & 0 \\ 0 & M^T & I & M & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & M^T & I & M \\ 0 & \dots & 0 & 0 & M^T & I \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_r \end{bmatrix} = 0\,, \tag{2.14}$$

so that the $r \times r$ block matrix

$$\mathcal{M}_r := \begin{bmatrix} I & M & 0 & 0 & \dots & 0 \\ M^T & I & M & 0 & \dots & 0 \\ 0 & M^T & I & M & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & M^T & I & M \\ 0 & \dots & 0 & 0 & M^T & I \end{bmatrix} \tag{2.15}$$

must be singular. This represents a contradiction, since (2.9) implies $\mathcal{M}_r = \mathcal{R}_r^{-1}$, which by assumption is positive definite for all $r \le N$.

In the next section we describe a canonical form of regular symplectic matrix pencils obtained in [8] which plays an important role in our analysis.

3. Canonical form of symplectic matrix pencils. We start by establishing some notation. Given two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$, if a_{ij} denotes the (i,j)th element of the matrix A, the *Kronecker product* $A \otimes B$ of A and B is the $np \times mq$ matrix defined by

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1m}B \\ a_{21}B & a_{22}B & \dots & a_{2m}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1}B & a_{n2}B & \dots & a_{nm}B \end{bmatrix}. \tag{3.1}$$

The reader is referred to [11] for a discussion of the properties of the matrix Kronecker product. We denote by $A \oplus B$ the block diagonal matrix

$$A \oplus B := \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} . \tag{3.2}$$

We denote also by Z_r and Σ_r the following $r \times r$ matrices:

$$Z_r = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}, \tag{3.3}$$

$$\Sigma_r = \begin{bmatrix} 0 & \dots & 0 & 0 & (-1)^{r-1} \\ 0 & \dots & 0 & (-1)^{r-2} & 0 \\ 0 & \iddots & \iddots & \iddots & 0 \\ 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix} . \tag{3.4}$$

Finally, for a complex number $a = \sigma + j\omega$ (where j stands for the imaginary unit), $J_r(a)$ denotes a Jordan block of size r with eigenvalue a, i.e.,

$$J_r(a) = aI_r + Z_r . \tag{3.5}$$

Similarly, $J_{2r}(a, a^*)$ represents the $2r \times 2r$ real Jordan block obtained by pairing the complex Jordan blocks of size r associated with a and a^*:

$$J_{2r}(a, a^*) = I_r \otimes \begin{bmatrix} \sigma & -\omega \\ \omega & \sigma \end{bmatrix} + Z_r \otimes I_2 . \tag{3.6}$$

The following result was proved in [8].

THEOREM 1. *Given a regular symplectic matrix pencil $sE - tA$, there exist two real nonsingular matrices V and W such that*

$$(sE - tA)V = W(\oplus_{i=1}^{l}(sE_i - tA_i)) \tag{3.7}$$

and

$$W^T K W = \oplus_{i=1}^{l}(K_i), \tag{3.8}$$

where K is given by (2.12) *and the blocks E_i, A_i, and K_i are of four possible types.*

Type 1. *The blocks E_i, A_i, and K_i corresponding to a real eigenvalue pair (a_i, a_i^{-1}), with $|a_i| < 1$, or to the pair $(0, \infty)$ take the form*

$$E_i = \oplus_{k=1}^{p_i} \begin{bmatrix} I_{r_k} & 0 \\ 0 & J_{r_k}^T(a_i) \end{bmatrix} , \tag{3.9a}$$

$$A_i = \oplus_{k=1}^{p_i} \begin{bmatrix} J_{r_k}(a_i) & 0 \\ 0 & I_{r_k} \end{bmatrix} , \tag{3.9b}$$

$$K_i = \oplus_{k=1}^{p_i} \begin{bmatrix} 0 & -I_{r_k} \\ I_{r_k} & 0 \end{bmatrix} . \tag{3.9c}$$

Type 2. *The blocks E_i, A_i, and K_i corresponding to a complex eigenvalue quadruple $(a_i, a_i^*, a_i^{-1}, a_i^{-*})$ such that $a_i = \sigma_i + j\omega_i$ with $\omega_i > 0$ and $|a_i| < 1$ admit the structure*

$$E_i = \oplus_{k=1}^{p_i} \begin{bmatrix} I_{2r_k} & 0 \\ 0 & J_{2r_k}^T(a_i, a_i^*) \end{bmatrix}, \tag{3.10a}$$

$$A_i = \oplus_{k=1}^{p_i} \begin{bmatrix} J_{2r_k}(a_i, a_i^*) & 0 \\ 0 & I_{2r_k} \end{bmatrix}, \tag{3.10b}$$

$$K_i = \oplus_{k=1}^{p_i} \begin{bmatrix} 0 & -I_{2r_k} \\ I_{2r_k} & 0 \end{bmatrix}. \tag{3.10c}$$

Type 3. *The blocks E_i, A_i, and K_i corresponding to a complex eigenvalue pair $(e^{j\theta_i}, e^{-j\theta_i})$ on the unit circle, with $0 < \theta_i < \pi$, admit the structure*

$$E_i = \oplus_{k=1}^{p_i}(I_{2r_k} - J_{2r_k}(jb_i, -jb_i)), \tag{3.11a}$$

$$A_i = \oplus_{k=1}^{p_i}(I_{2r_k} + J_{2r_k}(jb_i, -jb_i)), \tag{3.11b}$$

$$K_i = \oplus_{k=1}^{p_i}\kappa_k(\Sigma_{r_k} \otimes \Sigma_2), \tag{3.11c}$$

with $\kappa_k = \pm 1$ and $b_i = \tan(\theta_i/2)$.

Type 4. *The blocks E_i, A_i, and K_i corresponding to eigenvalues located at $\varepsilon_i = \pm 1$ take the form*

$$E_i = \oplus_{k=1}^{p_{ei}}(I_{2r_k} - Z_{2r_k}) \oplus \left(\oplus_{k=1}^{p_{oi}} \begin{bmatrix} I_{2r'_k+1} - Z_{2r'_k+1} & 0 \\ 0 & (I_{2r'_k+1} + Z_{2r'_k+1})^T \end{bmatrix}\right), \tag{3.12a}$$

$$A_i = \varepsilon_i \Bigg[\oplus_{k=1}^{p_{ei}} (I_{2r_k} + Z_{2r_k}) \oplus \left(\oplus_{k=1}^{p_{oi}} \begin{bmatrix} I_{2r'_k+1} + Z_{2r'_k+1} & 0 \\ 0 & (I_{2r'_k+1} - Z_{2r'_k+1})^T \end{bmatrix}\right)\Bigg], \tag{3.12b}$$

$$K_i = (\oplus_{k=1}^{p_{ei}}\kappa_k\Sigma_{2r_k}) \oplus \left(\oplus_{k=1}^{p_{oi}} \begin{bmatrix} 0 & -I_{2r'_k+1} \\ I_{2r'_k+1} & 0 \end{bmatrix}\right), \tag{3.12c}$$

with $\kappa_k = \pm 1$, and where each p_{ei} is even and each p_{oi} is odd.

The following lemma, whose proof is straightforward, describes some features of the blocks E_i, A_i, and K_i.

LEMMA 1. *The blocks E_i, A_i, and K_i have the following properties:*

1. *$sE_i - tA_i$ is a regular pencil;*
2. *E_i and A_i commute;*
3. *K_i is skew-symmetric of even size and orthogonal: $K_i^T = -K_i = K_i^{-1}$;*
4. *the following relation holds:*

$$A_i^T = e_i K_i E_i K_i^T, \tag{3.13}$$

   where $e_i = 1$ for blocks of Type $1, 2, 3$ and $e_i = \varepsilon_i = \pm 1$ for blocks of Type 4;
5. *the pencil $sE_i - tA_i$ is K_i-symplectic, i.e., $E_i^T K_i E_i = A_i^T K_i A_i$.*

Observe that (3.13) implies

$$E_i^T = e_i K_i A_i K_i^T, \tag{3.14}$$

so we have the identity

$$(sA_i - tE_i)^T = e_i K_i (sE_i - tA_i) K_i^T \tag{3.15}$$

or, equivalently,

$$E_i^T K_i = e_i K_i A_i \,, \tag{3.16a}$$

$$A_i^T K_i = e_i K_i E_i \,. \tag{3.16b}$$

4. Covariance characterization. We are now ready to characterize the stationary Gaussian reciprocal processes over the interval $[0, N]$. Partition the matrices V and W of Theorem 1 as

$$V = [V_1 \mid V_2 \mid \ldots \mid V_l] \,, \quad W = [W_1 \mid W_2 \mid \ldots \mid W_l] \tag{4.1}$$

so that V_i and W_i have the same number of columns as A_i and E_i. The decomposition (3.7) can then be rewritten as

$$(sE - tA)V_i = W_i(sE_i - tA_i), \tag{4.2}$$

which yields

$$EV_iA_i = AV_iE_i \,, \tag{4.3}$$

where V_i has full column rank, and where we have used the fact that E_i and A_i commute. Now partition each W_i as

$$W_i = \begin{bmatrix} Y_i \\ X_i \end{bmatrix} , \tag{4.4}$$

where X_i and Y_i have n rows. Then, taking into account the definitions of E and A, (4.2) implies

$$V_i = \begin{bmatrix} X_iA_i \\ X_iE_i \end{bmatrix} . \tag{4.5}$$

The following lemma specifies the form of the matrix functions $R(t)$ obeying the reciprocal dynamics (2.9).

LEMMA 2. *The $n \times n$ matrix valued function $R(t)$ satisfies equation (2.9) for $1 \leq |t| \leq N-1$ if and only if it admits the form*

$$R(t) = \begin{cases} \displaystyle\sum_{i=1}^{l} X_i E_i^{N-t} A_i^t C_i, & t \geq 0, \\ \displaystyle\sum_{i=1}^{l} X_i E_i^{-t} A_i^{N+t} D_i, & t \leq 0, \end{cases} \tag{4.6}$$

where C_i and D_i are arbitrary matrices with n columns and as many rows as A_i and E_i.

Proof. Without loss of generality, we consider the case where $1 \le t \le N$. As in the previous section, we write $R_1(t) := R(t-1)$ and let

$$\tilde{R}(t) := \begin{bmatrix} R(t) \\ R_1(t) \end{bmatrix}, \tag{4.7}$$

with $1 \le t \le N$. The function $R(t)$ satisfies (2.9) if and only if the function $\tilde{R}(t)$ obeys (2.10) for $1 \le t \le N-1$, which, after left multiplication by W^{-1}, can be rewritten as

$$(\oplus_{i=1}^{l} E_i)\bar{R}(t+1) = (\oplus_{i=1}^{l} A_i)\bar{R}(t) \tag{4.8}$$

for $1 \le t \le N-1$, with $\bar{R}(t) := V^{-1}\tilde{R}(t)$, where we have used relation (3.7). Since $(\oplus_{i=1}^{l} E_i)$ and $(\oplus_{i=1}^{l} A_i)$ have a block diagonal form, we can compute the rows of $\bar{R}(t)$ corresponding to nonsingular blocks of $(\oplus_{i=1}^{l} E_i)$ by propagating forward in time the corresponding part of (4.8) and considering as parameters the entries of the corresponding rows of $\bar{R}(1)$. Since the singular blocks of $(\oplus_{i=1}^{l} E_i)$ correspond to nonsingular blocks of $(\oplus_{i=1}^{l} A_i)$, we can compute the remaining rows by propagating backward in time the corresponding part of (4.8) while using as parameters the entries of the corresponding rows of $\bar{R}(N)$. Then, taking into account the commutativity of E_i and A_i, it is not difficult to check that any solution of (4.8) has the form

$$\bar{R}(t) = \begin{bmatrix} E_1^{N-t}A_1^{t-1}C_1 \\ E_2^{N-t}A_2^{t-1}C_2 \\ \vdots \\ E_l^{N-t}A_l^{t-1}C_l \end{bmatrix} \tag{4.9}$$

for $1 \le t \le N$, and hence, taking into account (4.5),

$$\begin{bmatrix} R(t) \\ R_1(t) \end{bmatrix} = \tilde{R}(t) = V\bar{R}(t) = \begin{bmatrix} \sum_{i=1}^{l} X_i E_i^{N-t} A_i^t C_i \\ \sum_{i=1}^{l} X_i E_i^{N+1-t} A_i^{t-1} C_i \end{bmatrix} \tag{4.10}$$

for $1 \le t \le N$, where the matrix blocks C_i are free parameters. Observing that $R_1(t) = R(t-1)$, the last equation implies (4.6) for $1 \le t \le N$. ∎

To complete the parametrization of $R(t)$ we must find which restriction imposed to C_i and D_i will satisfy the following requirements:

1. $R(t)$ is the covariance of a real valued process;
2. $R(t)$ satisfies (2.9) for $t = 0$, and the two values of $R(t)$ given by (4.6) for $t = 0$ coincide.

The first requirement implies that $R(t)$ is real. Since X_i, E_i, and A_i are real valued, if the matrices C_i and D_i are complex valued, their imaginary parts do not affect the value of $R(t)$, so without loss of generality we assume that C_i and D_i are real. To deal with the self-adjointness condition

$$R(t) = R^T(-t) \tag{4.11}$$

we assume that the interval $[-N, N]$ is sufficiently large so that (4.11) implies

$$X_i E_i^{N-t} A_i^t C_i = D_i^T (A_i^T)^{N-t} (E_i^T)^t X_i^T, \quad 0 \le t \le N, \quad 1 \le i \le l. \tag{4.12}$$

An intuitive proof of identity (4.12) is the following: each subpencil $sE_i - tA_i$ corresponds to a different eigenvalue—and hence a different mode—of $sE - tA$. Since

the interval $[-N, N]$ is assumed to be large, each component of the sum (4.6) for $t \geq 0$ must be matched by the component corresponding to the same eigenvalue for $t \leq 0$. But according to (3.15), the subpencil $sE_i - tA_i$ has the same eigenvalues as $sA_i^T - tE_i^T$. Matching components of (4.11) therefore leads to (4.12). A formal proof can be found in the appendix.

Identity (4.12), taking into account (3.13), may be rewritten as

$$X_i E_i^{N-t} A_i^t C_i = e_i^N D_i^T K_i E_i^{N-t} A_i^t K_i^T X_i^T \tag{4.13}$$

for $1 \leq t \leq N$ or, equivalently, as

$$\begin{aligned} &\underbrace{\begin{bmatrix} X_i A_i \\ X_i E_i \end{bmatrix}}_{V_i} E_i^{N-(t+1)} A_i^{t-1} [E_i C_i \mid A_i C_i] \\ &= e_i^N \begin{bmatrix} D_i^T K_i A_i \\ D_i^T K_i E_i \end{bmatrix} E_i^{N-(t+1)} A_i^{t-1} [E_i K_i^T X_i^T \mid A_i K_i^T X_i^T], \end{aligned} \tag{4.14}$$

with $1 \leq t \leq N$. Substituting in (4.14) the identities

$$\begin{bmatrix} D_i^T K_i A_i \\ D_i^T K_i E_i \end{bmatrix} = e_i \begin{bmatrix} D_i^T E_i^T \\ D_i^T A_i^T \end{bmatrix} K_i, \tag{4.15a}$$

$$[E_i K_i^T X_i^T \mid A_i K_i^T X_i^T] = e_i K_i^T \underbrace{[A_i^T X_i^T \mid E_i^T X_i^T]}_{V_i^T} \tag{4.15b}$$

obtained from (3.13), we get

$$V_i E_i^{N-(t+1)} A_i^{t-1} [E_i C_i \mid A_i C_i] = e_i^N \begin{bmatrix} D_i^T E_i^T \\ D_i^T A_i^T \end{bmatrix} K_i E_i^{N-(t+1)} A_i^{t-1} K_i^T V_i^T , \tag{4.16}$$

with $1 \leq t \leq N-1$.

Since V_i has full column rank, it admits a left inverse V_i^{-L}, and V_i^T admits a right inverse $(V_i^T)^{-R} = (V_i^{-L})^T$. Let

$$S_i := [E_i C_i \mid A_i C_i](V_i^{-L})^T K_i, \tag{4.17a}$$

$$T_i := e_i^N V_i^{-L} \begin{bmatrix} D_i^T E_i^T \\ D_i^T A_i^T \end{bmatrix} K_i \, . \tag{4.17b}$$

Pre- and postmultiplying (4.16) by V_i^{-L} and $(V_i^{-L})^T K_i$, respectively, and taking into account the definitions (4.17) gives

$$E_i^{N-(t+1)} A_i^{t-1} S_i = T_i E_i^{N-(t+1)} A_i^{t-1} \tag{4.18}$$

for $1 \leq t \leq N-1$.

Equation (4.13) also implies

$$V_i E_i^{N-t} A_i^{t-1} C_i = e_i^N \begin{bmatrix} D_i^T K_i A_i \\ D_i^T K_i E_i \end{bmatrix} E_i^{N-t} A_i^{t-1} K_i^T X_i^T \, , \quad 1 \leq t \leq N, \tag{4.19a}$$

$$V_i E_i^{N-(t+1)} A_i^t C_i = e_i^N \begin{bmatrix} D_i^T K_i A_i \\ D_i^T K_i E_i \end{bmatrix} E_i^{N-(t+1)} A_i^t K_i^T X_i^T \, , \quad 0 \leq t \leq N-1. \tag{4.19b}$$

By premultiplying (4.19a) and (4.19b) by V_i^{-L} and taking into account the identity (4.15b) and the definition (4.17b) of T_i, we obtain

$$E_i^{N-t}A_i^{t-1}C_i = e_iT_iE_i^{N-t}A_i^{t-1}K_i^TX_i^T\,, \quad 1 \le t \le N, \tag{4.20a}$$

$$E_i^{N-(t+1)}A_i^tC_i = e_iT_iE_i^{N-(t+1)}A_i^tK_i^TX_i^T\,, \quad 0 \le t \le N-1. \tag{4.20b}$$

Similarly, (4.13) yields

$$X_iE_i^{N-(t+1)}A_i^t\begin{bmatrix} E_iC_i & A_iC_i \end{bmatrix} = e_i^{N+1}D_i^TK_iE_i^{N-(t+1)}A_i^tK_i^TV_i^T\,, \quad 0 \le t \le N-1, \tag{4.21a}$$

$$X_iE_i^{N-t}A_i^{t-1}\begin{bmatrix} E_iC_i & A_iC_i \end{bmatrix} = e_i^{N+1}D_i^TK_iE_i^{N-t}A_i^{t-1}K_i^TV_i^T\,, \quad 1 \le t \le N. \tag{4.21b}$$

By postmultiplying (4.19a) and (4.19b) with $(V_i^{-L})^TK_i$ and taking into account the identities (4.5) and (3.16) and the definition (4.17a) of S_i, we find

$$D_i^T(A_i^T)^{N-(t+1)}(E_i^T)^t = X_iE_i^{N-(t+1)}A_i^tS_iK_i^T, \quad 0 \le t \le N-1, \tag{4.22a}$$

$$D_i^T(A_i^T)^{N-t}(E_i^T)^{t-1} = X_iE_i^{N-t}A_i^{t-1}S_iK_i^T, \quad 1 \le t \le N. \tag{4.22b}$$

Structure of the blocks S_i and T_i. Expressions (4.20) and (4.22) allow a parametrization of $R(t)$ in terms of the matrices T_i and/or S_i. To derive this parametrization, we first examine the constraints that are imposed on the matrices S_i and T_i by the identity (4.18). In this context, it is convenient to make the additional assumption that the pencil $sE - tA$ is *nonderogatory*, i.e., that its Kronecker decomposition admits only one Jordan block for each eigenvalue. This implies that the subpencils $sE_i - tA_i$ are also nonderogatory, in which case the structure of the matrices T_i and S_i obeying (4.18) can be characterized as follows.

Type 1. The matrices E_i and A_i have the structure

$$E_i = \begin{bmatrix} I_r & 0 \\ 0 & J_r^T(a) \end{bmatrix}\,, \quad A_i = \begin{bmatrix} J_r(a) & 0 \\ 0 & I_r \end{bmatrix} \tag{4.23}$$

so that

$$E_i^{N-(t+1)}A_i^{t-1} = \begin{bmatrix} (J_r(a))^{t-1} & 0 \\ 0 & (J_r^T(a))^{N-(t+1)} \end{bmatrix}\,. \tag{4.24}$$

Partitioning S_i and T_i as

$$S_i = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix}, \quad T_i = \begin{bmatrix} T_1 & T_2 \\ T_3 & T_4 \end{bmatrix}, \tag{4.25}$$

we find that (4.18) implies

$$J_r^{t-1}(a)S_1 = T_1J_r^{t-1}(a), \tag{4.26a}$$

$$J_r^{t-1}(a)S_2 = T_2(J_r^T(a))^{N-(t+1)}, \tag{4.26b}$$

$$(J_r^T(a))^{N-(t+1)}S_3 = T_3J_r^{t-1}(a), \tag{4.26c}$$

$$(J_r^T(a))^{N-(t+1)}S_4 = T_4(J_r^T(a))^{N-(t+1)} \tag{4.26d}$$

for $1 \le t \le N-1$. Setting $t=1$ and $t=2$ in (4.26a) we find that $S_1 = T_1$ commutes with $J_r(a)$, so that

$$S_1 = T_1 = t_1(Z_r), \tag{4.27}$$

where $t_1(x) = \sum_{i=0}^{r-1} t_{1i} x^i$ is a polynomial of degree $r-1$. Similarly, setting $t = N-1$ and $t = N-2$ in (4.26d) gives

$$S_4 = T_4 = t_4(Z_r^T), \tag{4.28}$$

where $t_4(x) = \sum_{i=0}^{r-1} t_{4i} x^i$ is a polynomial of degree $r-1$. Setting $t=1$ and $t = N-1$ in (4.26b) we find

$$S_2 = T_2 (J_r^T(a))^{N-2}\,, \quad (J_r(a))^{N-2} S_2 = T_2 \tag{4.29}$$

so that

$$S_2 - (J_r(a))^{N-2} S_2 (J_r^T(a))^{N-2} = 0\,. \tag{4.30}$$

Since $a < 1$, the matrix $(J_r(a))^{N-2}$ is stable, and hence the above Lyapunov equation admits $S_2 = 0$ as its unique solution. This, in turn, implies $T_2 = 0$. Similarly, we find $S_3 = T_3 = 0$.

In conclusion, we have proved that $S_i = T_i$ is block diagonal and commutes with E_i and A_i.

Type 2. In this case

$$E_i = \begin{bmatrix} I_r & 0 \\ 0 & J_{2r}^T(a, a^*) \end{bmatrix}\,, \quad A_i = \begin{bmatrix} J_{2r}(a, a^*) & 0 \\ 0 & I_r \end{bmatrix}\,. \tag{4.31}$$

Proceeding as in the Type 1 case, we find

$$S_i = T_i = \begin{bmatrix} T_1 & 0 \\ 0 & T_4 \end{bmatrix}\,, \tag{4.32}$$

where T_1 commutes with $J_{2r}(a, a^*)$ and hence has the structure

$$T_1 = t_{1R}(Z_r) \otimes I_2 + t_{1I}(Z_r) \otimes \Sigma_2\,, \tag{4.33}$$

where $t_{1R}(x)$ and $t_{1I}(x)$ are polynomials of degree $r-1$. Similarly, T_4 commutes with $(J_{2r}(a, a^*))^T$ and admits the structure

$$T_4 = [t_{4R}(Z_r) \otimes I_2 + t_{4I}(Z_r) \otimes \Sigma_2]^T\,, \tag{4.34}$$

where $t_{4R}(x)$ and $t_{4I}(x)$ are polynomials of degree $r-1$.

Type 3. In this case, we have

$$E_i = I_{2r} - B, \quad A_i = I_{2r} + B\,, \tag{4.35}$$

where $B := J_{2r}(jb, -jb)$. Then (4.18) takes the form

$$[I_{2r} - B]^{N-(t+1)} [I_{2r} + B]^{t-1} S_i = T_i [I_{2r} - B]^{N-(t+1)} [I_{2r} + B]^{t-1}\,. \tag{4.36}$$

Observing that

$$I_{2r} = \left[\frac{1}{2}(I_{2r} - B) + \frac{1}{2}(I_{2r} + B)\right]^{N-2}$$

$$= \left(\frac{1}{2}\right)^{N-2} \sum_{t=1}^{N-1} \begin{pmatrix} N-2 \\ t-1 \end{pmatrix} (I_{2r} - B)^{N-(t+1)} (I_{2r} + B)^{t-1} , \tag{4.37}$$

equation (4.36) implies $S_i = T_i$. Then, again using (4.36), T_i commutes with B, so that it has the structure

$$T_i = t_R(Z_r) \otimes I_2 + t_I(Z_r) \otimes \Sigma_2 , \tag{4.38}$$

where $t_R(x)$ and $t_I(x)$ are polynomials of degree $r - 1$.

Type 4. In this case

$$E_i = I_{2r} - Z_{2r} , \quad A_i = \varepsilon(I_{2r} + Z_{2r}) , \tag{4.39}$$

and proceeding as in Type 3, it is easy to check that

$$S_i = T_i = t(Z_{2r}) , \tag{4.40}$$

where $t(x)$ is a polynomial of degree $2r - 1$.

The above argument shows that $T_i = S_i$ commutes with A_i and E_i, and if $2r$ denotes the size of T_i, each T_i is parametrized by polynomials with $2r$ real coefficients. Employing the expression for $E_i^{N-t} A_i^{t-1} C_i$ given by (4.20a) and taking into account the commutativity of T_i with A_i and E_i, equation (4.6) yields

$$R(t) = \sum_{i=1}^{l} X_i A_i E_i^{N-t} A_i^{t-1} C_i = \sum_{i=1}^{l} e_i X_i T_i E_i^{N-t} A_i^t K_i^T X_i^T \tag{4.41}$$

for $1 \le t \le N$. If we repeat the same argument by factoring out $X_i E_i$ instead of $X_i A_i$ and use (4.20b) instead of (4.20a), we again obtain (4.41), but for $0 \le t \le N - 1$, so that

$$R(t) = \sum_{i=1}^{l} e_i X_i T_i E_i^{N-t} A_i^t K_i^T X_i^T \tag{4.42}$$

holds for $0 \le t \le N$. Also, transposing (4.22) and employing identities (3.16) and the commutativity of $T_i = S_i$ with A_i and E_i, we find

$$E_i^t A_i^{N-(t+1)} D_i = e_i^{N-1} E_i^t A_i^{N-(t+1)} K_i T_i^T X_i^T , \quad t = 0, 1, \ldots, N-1, \tag{4.43a}$$

$$E_i^{t-1} A_i^{N-t} D_i = e_i^{N-1} E_i^{t-1} A_i^{N-t} K_i T_i^T X_i^T , \quad t = 1, 2, \ldots, N, \tag{4.43b}$$

so that the argument that led to (4.42) allows us to conclude that

$$R(t) = \sum_{i=1}^{l} e_i^{N-1} X_i E_i^{-t} A_i^{N+t} K_i T_i^T X_i^T \tag{4.44}$$

for $-N \le t \le 0$.

It is easy to verify that the converse is also true: namely, if $R(t)$ is given by (4.42) and (4.44), where T_i has the appropriate structure for each block type, then $R(t) = R^T(-t)$ for $0 \le t \le N$.

The final step is to require that the expressions (4.42) and (4.44) for $R(0)$ coincide and that

$$MR(1) + R(0) + M^T R(-1) = I_n. \tag{4.45}$$

This will result in further restrictions on the matrices T_i. Equation (4.45) and

$$\sum_{i=1}^{l} e_i X_i E_i^N T_i K_i^T X_i^T = \sum_{i=1}^{l} e_i^{N-1} X_i A_i^N K_i T_i^T X_i^T \tag{4.46}$$

may be combined as

$$\underbrace{\begin{bmatrix} M & 0 \\ 0 & I_n \end{bmatrix}}_{E} \sum_{i=1}^{l} \begin{bmatrix} X_i A_i \\ X_i E_i \end{bmatrix} e_i E_i^{N-1} T_i K_i^T X_i^T$$

$$- \underbrace{\begin{bmatrix} -I_n & -M^T \\ I_n & 0 \end{bmatrix}}_{A} \sum_{i=1}^{l} \begin{bmatrix} X_i A_i \\ X_i E_i \end{bmatrix} e_i^{N-1} A_i^{N-1} K_i T_i^T X_i^T = \begin{bmatrix} I_n \\ 0 \end{bmatrix}. \tag{4.47}$$

Defining

$$X := [X_1 \mid X_2 \mid \ldots \mid X_l], \tag{4.48}$$

we can write the last identity as

$$\begin{aligned} &EV(\oplus_{i=1}^{l} e_i E_i^{N-1} T_i K_i^T) X^T \\ &- AV(\oplus_{i=1}^{l} e_i^{N-1} A_i^{N-1} K_i T_i^T) X^T = \begin{bmatrix} I_n \\ 0 \end{bmatrix}. \end{aligned} \tag{4.49}$$

At this point, substituting the equivalence relation (3.7), we find

$$W \left[\oplus_{i=1}^{l} e_i (E_i^N T_i K_i^T - e_i^N A_i^N K_i T_i^T) X^T\right] = \begin{bmatrix} I_n \\ 0 \end{bmatrix}. \tag{4.50}$$

Premultiplying (4.50) by $W^T K$ and taking (3.8) into account gives

$$\oplus_{i=1}^{l} e_i (K_i E_i^N T_i K_i^T - e_i^N K_i A_i^N K_i T_i^T) X^T = X^T. \tag{4.51}$$

Employing (3.16) and the related identity $(E_i^T)^N = -e_i^N K_i A_i^N K_i$, and premultiplying (4.51) first by $\oplus_{i=1}^{l} A_i^T$ and then by $\oplus_{i=1}^{l} E_i^T$, we get

$$(\oplus_{i=1}^{l} (e_i [K_i E_i^N T_i K_i^T + (E_i^T)^N T_i^T] - I_{r_i})) \underbrace{\begin{bmatrix} A_1^T X_1^T & E_1^T X_1^T \\ A_2^T X_2^T & E_2^T X_2^T \\ \vdots & \vdots \\ A_l^T X_l^T & E_l^T X_l^T \end{bmatrix}}_{V^T} = 0. \tag{4.52}$$

Since V is nonsingular, this implies

$$(E_i^N T_i)^T = e_i I_{r_i} + K_i E_i^N T_i K_i \tag{4.53}$$

for $i = 1, 2, \ldots, l$.

Polynomial parametrization of the blocks T_i. Equation (4.53) further restricts the structure of T_i, which, depending on the block type (see Theorem 1), takes the following form.

Type 1. For blocks of this type, the identity (4.53) becomes

$$\begin{bmatrix} t_1^T(Z_r) & 0 \\ 0 & t_4(Z_r)(J_r(a_i))^N \end{bmatrix} = I_{2r} - \begin{bmatrix} t_4^T(Z_r)(J_r^T(a_i))^N & 0 \\ 0 & t_1^T(Z_r) \end{bmatrix}, \tag{4.54}$$

which, employing the fact that $J_r(a_i) = a_i I_r + Z_r$, gives

$$t_1(x) = 1 - t_4(x)(a+x)^N \mod x^r. \tag{4.55}$$

Thus, once $t_4(x)$ is fixed, $t_1(x)$ is also determined.

Type 2. Proceeding as in the case of Type 1 blocks, we get

$$t_1(x) = 1 - t_4(x)(a+x)^N \mod x^r, \tag{4.56}$$

where t_1 and t_4 are defined by

$$t_1(x) = t_{1R}(x) + jt_{1I}(x), \tag{4.57a}$$

$$t_4(x) = t_{4R}(x) + jt_{4I}(x)\,. \tag{4.57b}$$

Thus, once $t_{4R}(x)$ and $t_{4I}(x)$ are selected, $t_{1R}(x)$ and $t_{1I}(x)$ are also determined.

Type 3. In this case, taking into account the identity

$$Z_r^k \Sigma_r = (-1)^k \Sigma_r (Z_r^T)^k \tag{4.58}$$

and (3.13), we find

$$\begin{aligned} K_i E_i^N T_i K_i &= K_i E_i^N K_i [t_R(-Z_r^T) \otimes I_2 - t_I(-Z_r^T) \otimes \Sigma_2^T] \\ &= -(A_i^T)^N [t_R(-Z_r^T) \otimes I_2 - t_I(-Z_r^T) \otimes \Sigma_2^T]\,. \end{aligned} \tag{4.59}$$

Thus the identity (4.53) may be written as

$$[1 - (jb+x)]^N t(x) = 1 - [1 + jb + x]^N t^*(-x) \mod x^r\,, \tag{4.60}$$

where we have defined

$$t(x) = t_R(x) + jt_I(x)\,. \tag{4.61}$$

Let $\lambda = 1 - jb \neq 0$. Denoting

$$f(x) = (\lambda - x)^N t(x) \mod x^r \tag{4.62}$$

so that $f(x)$ is a complex polynomial of degree less than or equal to $r-1$, equation (4.60) may be rewritten as

$$f(x) + f^*(-x) = 1\,. \tag{4.63}$$

Let $f_{ER}(x)$ be the even real part of $f(x)$, i.e., if $f(x) = \sum_{i=0}^{r-1} f_i x^i$, $f_{ER}(x) := \sum_{i=0}^{\lfloor \frac{r-1}{2} \rfloor} \mathbb{R}\mathrm{e}(f_{2i}) x^{2i}$. Let $f_{OI}(x)$ be the odd imaginary part of $f(x)$ defined similarly. Equation (4.63) fixes $f_{ER}(x)$ to be $f_{ER}(x) = 1/2$ and $f_{OI}(x)$ to be $f_{OI}(x) = 0$, so that $f(x)$ is parametrized only by r real coefficients. Then, since $\lambda \neq 0$, the polynomials

$(\lambda - x)^N$ and x^r are coprime, so that by employing the Bezout identity for these two polynomials, it is easy to construct a unique polynomial $t(x)$ of degree less than or equal to $r-1$ obeying (4.62). The polynomial $t(x)$ is parametrized by only r real coefficients.

Type 4. Using the same approach as for Type 3 blocks, we have

$$K_i E_i^N T_i K_i = -[A_i^T]^N t(-Z_{2r}^T)\,, \tag{4.64}$$

and hence identity (4.53) may be written as

$$(1-x)^N t(x) + (1+x)^N t(-x) = \varepsilon_i \mod x^{2r}. \tag{4.65}$$

Equation (4.65) places r linear constraints on the $2r$ parameters of the polynomials $t_R(x)$ and $t_I(x)$ so that, once again, the polynomial $t(x)$ is parametrized by only r real coefficients.

The above polynomial parametrization of the matrices T_i shows that each block $T_i \in \mathbb{R}^{2r_i \times 2r_i}$ depends on r_i real parameters, and hence the expression (4.42), (4.44) for $R(t)$ requires only $\sum_{i=1}^l r_i = n$ real parameters. We shall denote these parameters by p_k, $k = 1, 2, \ldots, n$. The following theorem summarizes the structure of stationary reciprocal covariances.

THEOREM 2. *Consider a stationary Gaussian reciprocal process over* $[0, N]$ *with reciprocal dynamics* (2.9) *such that the pencil* $sE - tA$ *given by* (2.11) *is nonderogatory. Then if the symplectic canonical form of* $(sE - tA, K)$ *is given by* $(\oplus_{i=1}^l (sE_i - tA_i), \oplus_{i=1}^l K_i)$, *and the blocks* X_i *are given by* (4.4), *the matrix covariance function* $R(t) = R^T(-t)$ *of the process takes the form*

$$R(t) = \begin{cases} \displaystyle\sum_{i=1}^{l} e_i X_i T_i E_i^{N-t} A_i^t K_i^T X_i^T, & 0 \le t \le N, \\ \displaystyle\sum_{i=1}^{l} e_i^{N-1} X_i E_i^{-t} A_i^{N+t} K_i T_i^T X_i^T, & -N \le t \le 0\,, \end{cases} \tag{4.66}$$

where, depending on the block type (see Theorem 1*), the matrices* T_i *have the following structure.*

Type 1.

$$T_i = \begin{bmatrix} t_1(Z_r) & 0 \\ 0 & t_4(Z_r^T) \end{bmatrix}, \tag{4.67}$$

where $t_1(x)$ *and* $t_4(x)$ *are real polynomials of degree* $r-1$ *such that*

$$t_1(x) = 1 - t_4(x)(a+x)^N \mod x^r\,. \tag{4.68}$$

Type 2.

$$T_i = \begin{bmatrix} T_1 & 0 \\ 0 & T_4 \end{bmatrix}, \tag{4.69}$$

with

$$T_1 = t_{1R}(Z_r) \otimes I_2 + t_{1I}(Z_r) \otimes \Sigma_2\,, \quad T_4 = [t_{4R}(Z_r) \otimes I_2 + t_{4I}(Z_r) \otimes \Sigma_2]^T\,, \tag{4.70}$$

where $t_{1R}(x)$, $t_{1I}(x)$, $t_{4R}(x)$, *and* $t_{4I}(x)$ *are real polynomials of degree* $r-1$ *such that*

$$t_{1R}(x) + jt_{1I}(x) = 1 - [t_{4R}(x) + jt_{4I}(x)](a+x)^N \mod x^r \,. \tag{4.71}$$

Type 3.

$$T_i = t_R(Z_r) \otimes I_2 + t_I(Z_r) \otimes \Sigma_2 \,, \tag{4.72}$$

where $t_R(x)$ *and* $t_I(x)$ *are real polynomials of degree* $r-1$ *such that*

$$(1-jb-x)^N [t_R(x) + jt_I(x)] + (1+jb+x)^N [t_R(-x) - jt_I(-x)] = 1 \mod x^r. \tag{4.73}$$

Type 4.

$$T_i = t(Z_{2r}), \tag{4.74}$$

where $t(x)$ *is a polynomial of degree* $2r-1$ *such that*

$$(1-x)^N t(x) + (1+x)^N t(-x) = \varepsilon_i \mod x^{2r} \,. \tag{4.75}$$

We finally observe that additional constraints must be imposed on the parameters of $R(t)$ to ensure that the covariance function $R(t-s)$ is nonnegative definite. These constraints take the form of inequalities, say for all the principal minors of the covariance matrix $\mathcal{R}_N$, so that they do not affect the number of degrees of freedom of $R(t)$, which remains equal to n.

5. Examples. Before considering multivariate examples, it is useful to adapt quickly Jamison's classification of continuous-time scalar stationary reciprocal processes to the discrete-time case. This will allow us to ignore higher-dimensional processes that can be decomposed into scalar components.

Assume that $M \in \mathbb{R}$ so that $R(t)$ is parametrized by a single parameter p. If $M = 0$, the reciprocal process is just white noise with covariance $R(t) = \delta(t)$. Hence, we consider the nontrivial case with $M \neq 0$. Let us assume $M < 0$, since the case of $M > 0$ is similar. In this case the pencil (2.11) has the two eigenvalues a and a^{-1} with $a = \frac{-1+\sqrt{1-4M^2}}{2M}$. At this point we have several cases.

Case 1. $M > -1/2$. In this case a is real and $0 < a < 1$. The covariance takes the form

$$R(t) = R(0) \left[(1-\mu) a^{-|t|} + \mu a^{|t|} \right] \,, \tag{5.1}$$

where

$$R(0) = \frac{1+a^2}{1-a^2}[1 - 2pa^N], \quad \mu = -\frac{pa^N}{1-2pa^N} \,. \tag{5.2}$$

To ensure that $R(t)$ is a nonnegative definite covariance function, the parameter p must be restricted to the interval

$$-\frac{1}{1-a^N} \leq p \leq \frac{1}{1+a^N}. \tag{5.3}$$

The case $p = 0$ corresponds to the only first-order Markov process of the reciprocal class.

The two extreme values $p = -\frac{1}{1-a^N}$ and $p = \frac{1}{1+a^N}$ correspond to the covariances

$$R_c(t) = \frac{a^{\left(\frac{N}{2}-|t|\right)} + a^{-\left(\frac{N}{2}-|t|\right)}}{a^{\frac{N}{2}} + a^{-\frac{N}{2}}}, \tag{5.4}$$

$$R_s(t) = \frac{a^{\left(\frac{N}{2}-|t|\right)} - a^{-\left(\frac{N}{2}-|t|\right)}}{a^{a^{\frac{N}{2}}-\frac{N}{2}}} \tag{5.5}$$

which represent the discrete-time versions of the *hyperbolic cosine* and *hyperbolic sine* processes of [14]. Specifically, with the substitution $\alpha := \ln a$, these covariances can be rewritten as

$$R_c(t) = \frac{\cosh\left[\alpha\left(\frac{N}{2} - |t|\right)\right]}{\cosh\left[\frac{\alpha N}{2}\right]}, \tag{5.6}$$

$$R_s(t) = \frac{\sinh\left[\alpha\left(\frac{N}{2} - |t|\right)\right]}{\sinh\left[\frac{\alpha N}{2}\right]}\,. \tag{5.7}$$

Case 2. $M = -1/2$. In this case $a = a^{-1} = 1$ is a double root of the equation $\det(\lambda E - A) = 0$. The covariance takes the form

$$R(t) = N - p - 2|t| = R(0)\left[1 - \mu|t|\right]\,, \tag{5.8}$$

where

$$R(0) = N - p, \quad \mu = 2R(0)^{-1} = \frac{2}{N-p}. \tag{5.9}$$

To guarantee that $R(t)$ is a covariance we require $R(0) > 0$ and $0 \le \mu \le 2/N$. Equivalently the parameter p must satisfy

$$p \le 0\,. \tag{5.10}$$

The covariance corresponding to the extreme value $p = 0$ takes the form

$$R(t) = R(0)\left[1 - \frac{2}{N}|t|\right]. \tag{5.11}$$

It may be viewed as the covariance of the discrete-time version of the *Slepian process* [24].

At the other extreme, when $p \to -\infty$, we obtain the *constant process* with covariance $R(t) = R(0)$ for all t. It satisfies $x(t) = x(0)$ and is therefore a *purely deterministic* process.

Case 3. $M < -1/2$. In this case a and a^{-1} are complex conjugate and lay on the unit circle. Accordingly, we can write $a = e^{-j\theta}$ with $0 < \theta < \frac{\pi}{2}$.

The covariance takes the form

$$R(t) = \frac{2\left[p\cos\left(\theta|t| - \frac{N\theta}{2}\right) + \frac{\left(\cos\frac{\theta}{2}\right)^N \sin\left(\theta|t| - \frac{N\theta}{2}\right)}{2\cos\left(\frac{N\theta}{2}\right)} - p\tan\left(\frac{N\theta}{2}\right)\sin\left(\theta|t| - \frac{N\theta}{2}\right)\right]}{\tan\theta\left(\cos\frac{\theta}{2}\right)^N}, \tag{5.12}$$

which may be expressed as

$$R(t) = R(0)\frac{\cos(\theta|t| + \beta)}{\cos\beta}, \tag{5.13}$$

where the parameter β depends on p in a fairly complicated way. The conditions

$$0 < \theta N < \pi, \tag{5.14a}$$

$$0 \le \beta \le \frac{\pi}{2} - \frac{\theta N}{2} \tag{5.14b}$$

guarantee that the function $R(t)$ is a covariance.

When $\beta = 0$ we have the discrete-time version of the *cosine process* [14], whose covariance is given by

$$R(t) = R(0)\cos(\theta|t|). \tag{5.15}$$

The latter is the covariance of a purely deterministic process, in the sense that the process $x(t)$ is completely specified by the boundary conditions $x(0)$ and $x(N)$.

At the other extreme, when $\beta = \frac{\pi}{2} - \frac{\theta N}{2}$, we have the discrete-time version of the *shifted sine process* [14], whose covariance is given by

$$R(t) = R(0)\frac{\sin\left[\theta\left(\frac{N}{2} - |t|\right)\right]}{\sin\left(\frac{\theta N}{2}\right)}. \tag{5.16}$$

In the case of processes of dimension $n > 1$ it is useful to observe that premultiplying the process $x(t)$ by an orthonormal matrix V does not affect the normalization $M_0 = I$ obtained by the change of basis (2.8). Under such a transformation, the matrix M becomes $V^T M V$ so that we may assume without loss of generality that M is upper triangular (or semitriangular in the case of complex eigenvalues).

Consider for example the case of 2-dimensional processes with

$$M = \begin{bmatrix} m_1 & m_2 \\ 0 & m_3 \end{bmatrix}. \tag{5.17}$$

In this case

$$\begin{aligned}\det[sE - tA] = {} & m_1 m_3 s^4 + (m_1 + m_3)\, s^3 t + \left(1 - m_2^2 + 2m_1 m_3\right) s^2 t^2 \\ & + (m_1 + m_3)\, s t^3 + m_1 m_3 t^4,\end{aligned} \tag{5.18}$$

and it is not difficult to check that by suitable choice of the parameters m_1, m_2, and m_3, the pencil $sE - tA$ may have any 4-tuple of eigenvalues, provided that they satisfy the symplectic symmetry (each eigenvalue must be paired with its reciprocal, and eigenvalues located at ± 1 must have even multiplicity) and the complex conjugation symmetry (each eigenvalue must be paired with its complex conjugate). Accordingly we may obtain any combination (summing up to dimension 4) of blocks of Types 1, 2, 3, and 4. For the case of two distinct pairs of reciprocal eigenvalues, the covariance dynamics may be decoupled in two parts, each one corresponding to one reciprocal pair of eigenvalues. The more interesting case occurs, therefore, when the canonical form of the pencil $sE - tA$ is formed by a single block of dimension 4. The general analysis then proceeds along the same lines as the continuous-time case discussed in [17, sect. 5]. However the discrete-time case includes a class of covariance functions

that cannot arise in the continuous time due to the existence of blocks of Type 1 associated to the eigenvalue pair $(0,\infty)$ which have no continuous-time counterpart.

To illustrate covariances of this type, consider the case

$$M = \begin{bmatrix} 0 & m \\ 0 & 0 \end{bmatrix} \tag{5.19}$$

corresponding to a single block of Type 1 and dimension 4 associated to an eigenvalue pair of multiplicity 2 at 0 and ∞. The parameter m is restricted to the interval

$$-1 \le m \le 1, \tag{5.20}$$

which is a necessary and sufficient condition for the matrix (2.15) to be positive definite. As was shown earlier, in this case the matrices $R(t)$ solving (2.9) form a class parametrized by two real parameters p_0 and p_1. A cumbersome but straightforward calculation yields

$$R(t) = \begin{cases} \frac{1}{1-m^2}\left[(-mZ_2^T)^t + (p_1 mZ_2 - p_0 I_2)Z^{N-t}\right], & t \ge 0, \\ \frac{1}{1-m^2}\left[(-mZ_2)^{-t} + (p_1 mZ_2^T - p_0 I_2)(Z^T)^{N+t}\right], & t < 0, \end{cases} \tag{5.21}$$

or, equivalently,

$$R(t) = \frac{1}{1-m^2}\bar{R}(t) \tag{5.22}$$

with

$$\bar{R}(t) = \begin{cases} I_2, & t = 0, \\ -mZ_2^T, & t = 1, \\ -mZ_2, & t = -1, \\ 0_2, & 1 < |t| < N-1, \\ p_0 mZ_2, & t = N-1, \\ p_0 mZ_2^T, & t = -(N-1), \\ -p_0 I_2 + p_1 mZ_2, & t = N, \\ -p_0 I_2 + p_1 mZ_2^T, & t = -N. \end{cases} \tag{5.23}$$

It is not difficult to check that the function $R(t)$ given by (5.22) is a covariance, i.e., is positive semidefinite, if and only if the parameters p_0 and p_1 satisfy

$$-1 \le p_0 \le 1\,, \tag{5.24a}$$

$$\frac{p_0^2 - 1}{m} \le p_1 \le \frac{1 - p_0^2}{m}\,. \tag{5.24b}$$

For example, in the extreme case when $p_0 = -1$ and $p_1 = 0$, $x(N) = x(0) \sim \mathcal{N}(0, \frac{I_2}{1-m^2})$.

6. Conclusions. The main contribution of this paper is Theorem 2, which characterizes the covariances of stationary Gaussian reciprocal processes. More precisely, the covariances corresponding to a fixed set of reciprocal dynamics are parametrized by n real parameters, where n represents the process dimension. In the Markov case, a similar characterization plays an important role in stochastic realization theory [20]. The stochastic realization problem seeks to model a stationary Gaussian process as a partially observed Gauss–Markov process in noise. However, Gauss–Markov models

have a preferred time direction, giving rise to forward and backward causal Markov realizations. To eliminate the issue of time direction, it would be natural to develop a reciprocal stochastic realization theory (see [23] for a solution in the special case of periodic processes), and the authors believe that the results presented here constitute a useful first step towards this objective.

Appendix. Proof of (4.12). We now prove (4.12), assuming N to be sufficiently large. In view of (4.6) we may write (4.11) as

$$\sum_{i=0}^{l} X_i E_i^{N-t} A_i^t C_i = \sum_{i=0}^{l} D_i^T (A_i^T)^{N-t} (E_i^T)^t X_i^T, \quad 0 \le t \le N. \tag{A.1}$$

If the pair $(0,\infty)$ appears among the pairs of eigenvalues of the pencil $sE - tA$, let i_0 be the corresponding index. In this case observe that $E_{i_0}^{N-t} A_{i_0}^t = (A_{i_0}^T)^{N-t} (E_{i_0}^T)^t = 0$ for $\nu < t < N - \nu$, with ν being the size of the largest Jordan block $J_r(0)$ in A_{i_0}. Thus, from (A.1) we get

$$\sum_{i=0, i\neq i_0}^{l} X_i E_i^{N-t} A_i^t C_i = \sum_{i=0, i\neq i_0}^{l} D_i^T (A_i^T)^{N-t} (E_i^T)^t X_i^T, \quad \nu < t < N - \nu. \tag{A.2}$$

Taking into account that all the E_i and A_i, $i \neq i_0$, are nonsingular, and taking into account the commutativity of E_i and A_i, (A.2) may be written as

$$\sum_{i=0, i\neq i_0}^{l} \Xi_i F_i^t C_i = \sum_{i=0, i\neq i_0}^{l} D_i^T (A_i^T)^N (F_i^{-T})^t X_i^T, \quad \nu < t < N - \nu, \tag{A.3}$$

with $\Xi_i := X_i E_i^N$ and $F_i := E_i^{-1} A_i$. The latter, in turn, may be rewritten as

$$\sum_{i=0, i\neq i_0}^{l} \Xi_i F_i^t C_i = \sum_{i=0, i\neq i_0}^{l} \Delta_i F_i^t \Theta_i, \quad \nu < t < N - \nu, \tag{A.4}$$

with $\Delta_i := D_i^T (A_i^T)^N K_i$, $\Theta_i := K_i^T X_i^T$. Identity (A.4) may be rewritten in matrix form as

$$\Xi F^t \Gamma = \Delta F^t \Theta, \quad \nu < t < N - \nu, \tag{A.5}$$

where $F := \oplus_{i\neq i_0} F_i$, Ξ is the matrix obtained by stacking together in the same row all the Ξ_i, Γ is the matrix obtained by stacking together in the same column all the C_i, and similarly for Δ and Θ. Since N is assumed to be large, using the Cayley–Hamilton theorem [10, p. 86], it is immediate to show that the identity in (A.5) holds for any $t > \nu$. Moreover, since F is nonsingular, using again the Cayley–Hamilton result, we may conclude that

$$\Xi F^t \Gamma = \Delta F^t \Theta, \quad t \ge 0, \tag{A.6}$$

or, equivalently,

$$\sum_{i=0, i\neq i_0}^{l} \Xi_i F_i^t C_i = \sum_{i=0, i\neq i_0}^{l} \Delta_i F_i^t \Theta_i, \quad t \ge 0. \tag{A.7}$$

Hence, for $z \in \mathbb{C}$, we have

$$\sum_{t=0}^{\infty} z^{-t-1} \sum_{i=0, i\neq i_0}^{l} \Xi_i F_i^t C_i = \sum_{t=0}^{\infty} z^{-t-1} \sum_{i=0, i\neq i_0}^{l} \Delta_i F_i^t \Theta_i, \tag{A.8}$$

which gives

$$\sum_{i=0, i\neq i_0}^{l} \Xi_i (zI - F_i)^{-1} C_i = \sum_{i=0, i\neq i_0}^{l} \Delta_i (zI - F_i)^{-1} \Theta_i. \tag{A.9}$$

The left-hand side and the right-hand side of (A.9) may be viewed as partial fraction expansions of the same rational function, so that, taking into account that for any pair F_i, F_j, $i \neq j$, F_i and F_j have disjoint spectra, we can conclude

$$\Xi_i (zI - F_i)^{-1} C_i = \Delta_i (zI - F_i)^{-1} \Theta_i, \quad i \neq i_0, \tag{A.10}$$

from which (4.12) may be directly obtained (except for $i = i_0$) by retracing the route leading from (A.2) to (A.4) in the reverse direction. Identity (4.12) for the case when $i = i_0$ may be obtained by subtraction. □

REFERENCES

[1] A. Beghi, *Continuous-time Gauss-Markov processes with fixed reciprocal dynamics*, J. Math. Systems, Estimation, and Control, 7 (1997), pp. 343–367.

[2] S. Bernstein, *Sur les liaisons entre les grandeurs aléatoires*, in Proceedings of the International Congress of Mathematicians, Zürich, Switzerland, 1932, pp. 288–309.

[3] J. P. Carmichael, J. C. Massé, and R. Theodorescu, *Processus Gaussiens stationnaires réciproques sur un intervalle*, C. R. Acad. Sci. Paris Sér. I Math., 295 (1982), pp. 291–293.

[4] J. P. Carmichael, J. C. Massé, and R. Theodorescu, *Reciprocal covariance solutions of some matrix differential equations*, Stochastic Process. Appl., 37 (1991), pp. 45–60.

[5] S. C. Chay, *On quasi-Markov random fields*, J. Multivariate Anal., 2 (1972), pp. 14–76.

[6] J. Coleman, *Gaussian reciprocal processes and their associated conservation laws*, Stochastics Stochastics Rep., 51 (1994), pp. 301–314.

[7] J. M. Coleman, B. C. Levy, and A. J. Krener, *Gaussian reciprocal diffusions and positive definite Sturm–Liouville operators*, Stochastics Stochastics Rep., 55 (1995), pp. 279–313.

[8] A. Ferrante and B. Levy, *Canonical form for symplectic matrix pencils*, Linear Algebra Appl., 274 (1998), pp. 259–300.

[9] R. Frezza, *Models of Higher-Order and Mixed-Order Gaussian Reciprocal Processes with Application to the Smoothing Problem*, Ph.D. dissertation, Applied Mathematics Program, University of California, Davis, CA, 1990.

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

[11] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[12] B. Jamison, *Reciprocal processes: The stationary Gaussian case*, Ann. Math. Statist., 41 (1970), pp. 1624–1630.

[13] B. Jamison, *Reciprocal processes*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 30 (1974), pp. 65–86.

[14] A. J. Krener, *Reciprocal diffusions and stochastic differential equations of second order*, Stochastics, 24 (1988), pp. 393–422.

[15] A. J. Krener, *Reciprocal diffusions in flat space*, Probab. Theory Related Fields, 107 (1997), pp. 243–281.

[16] A. J. Krener, R. Frezza, and B. C. Levy, *Gaussian reciprocal processes and self-adjoint stochastic differential equations of second order*, Stochastics Stochastics Rep., 34 (1991), pp. 29–56.

[17] B. C. Levy, *Characterization of multivariate stationary Gaussian reciprocal diffusions*, J. Multivariate Anal., 62 (1997), pp. 74–99.

[18] B. C. Levy, R. Frezza, and A. J. Krener, *Modeling and estimation of discrete-time Gaussian reciprocal processes*, IEEE Trans. Automat. Control, 35 (1990), pp. 1013–1023.
[19] B. C. Levy and A. J. Krener, *Kinematics and dynamics of reciprocal diffusions*, J. Math. Phys., 34 (1993), pp. 1846–1875.
[20] A. Lindquist and G. Picci, *Realization theory of multivariate stationary Gaussian processes*, SIAM J. Control Optim., 23 (1985), pp. 809–857.
[21] R. Recoules, *Gaussian reciprocal processes revisited*, Statist. Probab. Lett., 12 (1991), pp. 297–303.
[22] Yu. A. Rozanov, *Markov Random Fields*, Springer-Verlag, New York, 1982.
[23] J.-A. Sand, *Reciprocal realizations on the circle*, SIAM J. Control Optim., 34 (1996), pp. 507–520.
[24] D. Slepian, *First passage time for a particular Gaussian process*, Ann. Math. Statist., 32 (1961), pp. 610–612.
[25] H. Wimmer, *Normal forms of symplectic pencils and the discrete-time algebraic Riccati equation*, Linear Algebra Appl., 147 (1991), pp. 411–440.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 356–367

THE DISTANCE OF POTENTIALLY STABLE SIGN PATTERNS TO THE UNSTABLE MATRICES*

QING LIN†, D. D. OLESKY†, AND P. VAN DEN DRIESSCHE‡

Abstract. A measure $\delta_{us}(A)$ of the relative distance to the unstable matrices for a stable n-by-n matrix A is defined and extended to a potentially stable sign pattern. It is shown that $\delta_{us}(A) \in (0,1]$ and $|\delta_{us}(A) - \delta_{us}(B)|$ is bounded above in terms of $\|A - B\|_2$. For $n = 2$, there is a unique (up to permutation and signature similarity) minimally potentially stable sign pattern, and an optimal stable matrix (i.e., having maximal relative distance to the unstable matrices) is determined analytically. For $n = 3$ and 4 a complete list of all minimally potentially stable tree sign patterns is given and an approximation to an optimal matrix is found for each of these patterns. Such a matrix is also computed for some rooted trees with $n = 5$, and these results can be applied to estimate the distance to the unstable matrices for more general potentially stable sign patterns.

Key words. complex stability radius, distance to unstable matrices, eigenvalues, potential stability, rooted tree, sign pattern

AMS subject classifications. 15A18, 65A15

PII. S089547980139087X

1. Introduction.

In some qualitative matrix problems, the signs rather than the values of matrix entries are specified. These problems arise in such areas as economics, biology, chemistry, and social sciences, where the exact values for entries in a connection matrix may be unknown, but there is information on the signs of entries; see, for example, [4]. A *sign pattern (matrix)* $\mathcal{A} = [\alpha_{ij}]$ has $\alpha_{ij} \in \{+, 0, -\}$, and a real (numerical) matrix $A = [a_{ij}]$ belongs to this sign pattern (written $A \in \mathcal{A}$) if $\text{sign}(a_{ij}) = \alpha_{ij}$ for all i, j. Matrix A (possibly complex) is *stable* if all of its eigenvalues have negative real parts; otherwise A is *unstable*. A sign pattern $\mathcal{A}$ is *potentially stable* if there is a stable matrix $A \in \mathcal{A}$ and is *sign stable* if all matrices $A \in \mathcal{A}$ are stable. Sign stable patterns have been characterized (see, e.g., [7]), but the characterization of potentially stable sign patterns is much more difficult and remains open.

A sufficient condition for potential stability can be given by considering the signs of a nested sequence of principal minors. Letting $B[\{1, \ldots, k\}]$ denote the principal submatrix of a real matrix B lying in rows and columns $1, \ldots, k$, matrix A has a *properly signed nest* if there exists a permutation matrix P such that

$$\text{sign} \ \det(P^T A P[\{1, \ldots, k\}]) = (-1)^k \quad \text{for} \ \ k = 1, \ldots, n.$$

A theorem proved by Fisher and Fuller [3] and by Ballantine [1] states that if A has a properly signed nest, then there exists a positive diagonal matrix D so that DA is stable. Since A and DA have the same sign pattern, this theorem shows that if $\mathcal{A}$ is a sign pattern that allows a properly signed nest, then it is potentially stable [8, Theorem 2.1]. A proof of the Fisher–Fuller–Ballantine theorem determines D so that

*Received by the editors June 14, 2001; accepted for publication (in revised form) by N. J. Higham March 5, 2002; published electronically October 18, 2002. The research of the second and third authors was supported in part by an NSERC research grant.

http://www.siam.org/journals/simax/24-2/39087.html

†Department of Computer Science, University of Victoria, Victoria, BC, V8W 3P6 Canada (qlin@csr.uvic.ca, dolesky@csr.uvic.ca).

‡Department of Mathematics and Statistics, University of Victoria, Victoria, BC, V8W 3P4 Canada (pvdd@math.uvic.ca).

DA has simple real eigenvalues. However, this construction results in matrices that are very close to being unstable, and so are not of practical value. Our interest is in maximizing the relative distance of $A \in \mathcal{A}$ to the unstable matrices, and thus we do not rely on this construction.

In section 2, measures of the relative distance to the unstable matrices for a fixed stable matrix and for a fixed sign pattern are defined and some properties given. In section 3, minimally potentially stable sign patterns are discussed and their graphical structures explored, especially those corresponding to rooted trees. For the unique minimally potentially stable sign pattern with $n = 2$, an optimal stable matrix (i.e., having maximum relative distance to the unstable matrices) is determined analytically in section 4. In section 5, a "good" stable matrix is found numerically for each minimally potentially stable tree sign pattern with $n = 3$ and 4 (most of which have a properly signed nest) and for some rooted trees with $n = 5$. Finally, in section 6, we briefly consider more general patterns and suggest aspects for further research.

2. Relative distance to the unstable matrices. Let $\alpha(A)$ denote the spectral abscissa of an n-by-n complex matrix A, namely,

$$\alpha(A) = \max\{\mathrm{Re}(\lambda) : \lambda \text{ is an eigenvalue of } A\}.$$

In [12], Van Loan points out that $|\alpha(A)|$ is not an adequate measure of the distance of a given complex stable matrix A to the unstable matrices. He proposes the following alternative measure of stability [12, p. 246], which parallels the usual measure of distance from singularity. If A is a fixed stable matrix, then *the distance of A to the unstable matrices*, denoted by $d_{us}(A)$ as in [5], is defined as

$$d_{us}(A) = \min_{E}\{\|A - E\|_2 \ : \alpha(E) = 0\}.$$

In the literature, $d_{us}(A)$ is sometimes called the *complex stability radius* of A, and in the case that E is restricted to be real, the corresponding distance is called the real stability radius; see, for example, [5] and the references therein.

Van Loan uses the Frobenius norm to define $d_{us}(A)$, but (see Byers [2] and Higham [6]) either the Frobenius norm or the 2-norm ($\|A\|_2 = \sqrt{\rho(A^*A)}$ with $\rho(A)$ denoting the spectral radius of A) can be used. If A is a stable matrix, then taking $E = A - \alpha(A)I_n$, where I_n is the n-by-n identity matrix, gives $d_{us}(A) \le |\alpha(A)|$. If A is a real stable n-by-n matrix, then this bound leads to

$$d_{us}(A) \le |\operatorname{tr}(A)|/n, \tag{2.1}$$

which is easily calculated.

Assuming that $\alpha(E) = 0$ and E has an eigenvalue $i\mu$, matrix $G = E - \mu i I_n$ is singular. Thus $d_{us}(A) = \min_G\{\|A - G - \mu i I_n\|_2 : G \text{ is singular}\}$. This gives

$$d_{us}(A) = \min_{\mu \in \mathcal{R}}\{\sigma_{\min}(A - \mu i I_n)\}, \tag{2.2}$$

where $\sigma_{\min}(A)$ denotes the minimum singular value of A; see, for example, [12], [13].

Observe that if A is stable, then so is pA for p any positive constant, but $d_{us}(pA) = pd_{us}(A)$ can be arbitrarily large. Thus we define the following normalized measure. *The relative distance of a stable matrix A to the unstable matrices*, denoted by $\delta_{us}(A)$, is defined as

$$\delta_{us}(A) = \frac{d_{us}(A)}{\|A\|_2}. \tag{2.3}$$

This relative distance is invariant under multiplication by a positive constant (i.e., $\delta_{us}(pA) = \delta_{us}(A)$ for $p > 0$) and also is invariant under unitary similarity. The following result gives bounds for this relative distance.

LEMMA 2.1. *If A is any stable matrix, then $0 < \delta_{us}(A) \leq 1$, with the upper bound attained by $-pI_n$, where p is any positive constant.*

Proof. With $\mu = 0$ in (2.2), $d_{us}(A) \leq \sigma_{\min}(A) = \lambda_{\min}^{1/2}(A^*A) \leq \sqrt{\rho(A^*A)} = \|A\|_2$. Thus $\delta_{us}(A) \leq 1$, and it obviously must be positive. If $A = -I_n$, then $\sigma_{\min}(-(1+\mu i)I_n) = \sqrt{1+\mu^2}$, giving $d_{us}(-I_n) = 1$ by (2.2). Since $\|-I_n\|_2 = 1$, it follows that $\delta_{us}(-I_n) = 1$. Hence $\delta_{us}(-pI_n)$ attains the upper bound of 1. ∎

It is well known that the relative distance from a nonsingular matrix A to the nearest singular matrix is equal to the reciprocal of the condition number, i.e., $\frac{1}{\kappa_2(A)} = \frac{1}{\|A\|_2\|A^{-1}\|_2}$; see, for example, [13, Corollary 7.3.10]. It follows that $\delta_{us}(A) \leq 1/\kappa_2(A)$.

For a potentially stable sign pattern $\mathcal{A}$, *the distance of $\mathcal{A}$ to the unstable matrices*, denoted by $\delta_{us}(\mathcal{A})$, is defined as

$$\delta_{us}(\mathcal{A}) = \sup_A\{\delta_{us}(A) : A \in \mathcal{A} \text{ is stable}\}.$$

By compactness and continuity, there exists a stable matrix $A_0 \in \mathcal{A}$ that has $\|A_0\|_2 = 1$ and $\delta_{us}(\mathcal{A}) = \delta_{us}(A_0)$. Such a matrix cA_0 for any $c > 0$ is called an *optimal stable matrix* in the sign pattern $\mathcal{A}$. For example, $-pI_n$ with $p > 0$ is an optimal stable matrix in the n-by-n sign pattern $\operatorname{diag}(-,\ldots,-)$. It is well known that the computation of $d_{us}(A)$ is in general difficult. Therefore, finding an optimal stable matrix in a potentially stable sign pattern is often extremely difficult (see section 4 for $n = 2$). For certain sign patterns $\mathcal{A}$, bounds on $\delta_{us}(\mathcal{A})$ may be obtained. Before presenting such an example we give some perturbation results.

LEMMA 2.2. *For any two stable n-by-n matrices A and B,*

$$|d_{us}(A) - d_{us}(B)| \leq \|A - B\|_2.$$

This result holds for any distance function; see [6, equation (1.2)].

THEOREM 2.3. *For any two stable n-by-n matrices A and B,*

$$|\delta_{us}(A) - \delta_{us}(B)| \leq \frac{2\|A-B\|_2}{\max(\|A\|_2, \|B\|_2)}.$$

Proof. Without loss of generality assume that $\|A\|_2 \geq \|B\|_2$. Then

$$\begin{aligned}|\delta_{us}(A) - \delta_{us}(B)| &\leq \frac{|d_{us}(A) - d_{us}(B)|}{\|A\|_2} + \frac{d_{us}(B)}{\|B\|_2}\ \frac{(\|A\|_2 - \|B\|_2)}{\|A\|_2} \\ &\leq \frac{2\|A-B\|_2}{\|A\|_2}\end{aligned}$$

by Lemmas 2.1 and 2.2. ∎

Example 2.4. Let $\mathcal{A}_n = [\alpha_{ij}]$ be the tridiagonal n-by-n sign stable pattern with $\alpha_{11} = -$, $\alpha_{i,i+1} = +$, $\alpha_{i+1,i} = -$ for $i = 1,\ldots,n-1$, and all other entries 0.

(i) Any matrix $A = [a_{ij}] \in \mathcal{A}_n$ can be normalized so that its unique nonzero diagonal entry is $a_{11} = -1$. Thus $|\operatorname{tr}(A)| = 1$, giving $d_{us}(A) \leq 1/n$. If e_1 denotes the column vector with 1 in the first entry and all other entries 0, then $\|A\|_2 \geq \|Ae_1\|_2 = \sqrt{1+a_{21}^2} > 1$. Hence $\delta_{us}(\mathcal{A}_n) < 1/n$.

(ii) Let $\hat{A} \in \mathcal{A}_n$ have $\hat{a}_{11} = -1$, $\hat{a}_{i,i+1} = a > 0$, $\hat{a}_{i+1,i} = -a$ for $i = 1,\ldots,n-1$. Take E to be the unstable skew-symmetric matrix with $e_{i,i+1} = 1 = -e_{i+1,i}$ for

$i = 1, \ldots, n-1$, and all other entries 0. Then $\|\hat{A}/a - E\|_2 = 1/a$, and since $\|\hat{A}/a\|_2 > 1$, it follows that $\delta_{us}(\hat{A}/a) = \delta_{us}(\hat{A}) < 1/a$, which can be arbitrarily close to zero for any n.

(iii) To illustrate Theorem 2.3, let $A \in \mathcal{A}_n$ have $a_{11} = -1$, $a_{i,i+1} = a_i = -a_{i+1,i}$, and let $B \in \mathcal{A}_n$ have $b_{11} = -1$, $b_{i,i+1} = b_i = -b_{i+1,i}$ for $i = 1, \ldots, n-1$, where $a_i, b_i > 0$. With $c_i = a_i - b_i$ and $\|A - B\|_2^2 = \max_{\|x\|_2=1} \|(A-B)x\|_2^2 = c_1^2 x_2^2 + \sum_{i=1}^{n-2}(c_{i+1}x_{i+2} - c_i x_i)^2 + c_{n-1}^2 x_{n-1}^2 \le 4 \max_i c_i^2 \sum_{i=1}^n x_i^2 = 4 \max_i c_i^2$, it follows that

$$|\delta_{us}(A) - \delta_{us}(B)| \le \frac{4 \max_i(|a_i - b_i|)}{\max(\|A\|_2, \|B\|_2)}.$$

As previously remarked, the computation of $d_{us}(A)$, and thus of $\delta_{us}(A)$, is difficult. One of the most feasible methods for real A uses the associated $2n$-by-$2n$ matrix $H(\alpha)$ defined by

$$H(\alpha) = \begin{bmatrix} A & -\alpha I_n \\ \alpha I_n & -A^T \end{bmatrix},$$

where $\alpha \ge 0$. The first part of the following result is [2, Theorem 1], while the second part is a restatement in terms of $\delta_{us}(A)$.

THEOREM 2.5. *If A is a stable matrix, then the associated matrix $H(\alpha)$ has an eigenvalue with real part zero if and only if $\alpha \ge d_{us}(A)$; equivalently $H(\alpha\|A\|_2)$ has an eigenvalue with real part zero if and only if $\alpha \ge \delta_{us}(A)$.*

For a fixed potentially stable sign pattern $\mathcal{A}$ and fixed $\alpha > 0$, this theorem suggests the following method for attempting to show that $\delta_{us}(\mathcal{A}) > \alpha$. Let S be a subset of the stable matrices $A \in \mathcal{A}$ such that $0 < \beta \le \|A\|_2 \le \gamma$ for some fixed constants β, γ. A finite sequence of real stable matrices $A_k \in S$ is selected that is sufficiently dense in S, i.e., for all $A = [a_{ij}] \in S$ and some sufficiently small $\varepsilon > 0$, there exists $A_k = [a_{ij}^{(k)}]$ such that $|a_{ij} - a_{ij}^{(k)}| < \varepsilon$ for all i, j. Theorem 2.5 is applied to each matrix A_k in an attempt to determine numerically whether or not the associated matrix $H(\alpha\|A_k\|_2)$ has an eigenvalue with real part zero. If for some $A_k \in S$, the associated matrix $H(\alpha\|A_k\|_2)$ has an eigenvalue with real part zero (e.g., if $|\operatorname{Re}\lambda_j(H(\alpha\|A_k\|_2))| < 10^{-10}$), then $\alpha \ge \delta_{us}(A_k)$. If no such matrix A_k can be found, and if the set S of matrices A_k is sufficiently dense in $\mathcal{A}$ (i.e., ε is sufficiently small), then likely $\delta_{us}(\mathcal{A}) > \alpha$. This method is used in an algorithm to find the numerical matrices A and bounds on $\delta_{us}(A)$ given in section 5.

Matrix $H(\alpha)$ is also used in our analytical result for $n = 2$ (section 4), where we need its characteristic polynomial. This can be expressed as

$$\det(H(\alpha) - \lambda I_{2n}) = \det((\alpha^2 + \lambda^2)I_n - AA^T + \lambda(A^T - A)),$$

which is an even polynomial in λ.

3. Minimally potentially stable sign patterns. A sign pattern is *minimally potentially stable* if it is potentially stable, irreducible, and if replacing any + or − entry by 0 results in a pattern that is not potentially stable. For example, the sign pattern $\mathcal{A}_n$ in Example 2.4 is minimally potentially stable. The minimal patterns are the "atoms" of the potentially stable sign patterns, since if $\mathcal{A}$ is potentially stable and $\mathcal{A}$ is a subpattern of $\tilde{\mathcal{A}}$, then $\tilde{\mathcal{A}}$ is also potentially stable [9, Theorem 3].

The sign pattern $\mathcal{A}_n$ in Example 2.4 is a *tree sign pattern* (t.s.p.); see [8], [9]. As $\alpha_{ij} \ne 0$ whenever $\alpha_{ji} \ne 0$, this sign pattern can be represented by a *signed tree* with

vertex 1 signed negative to agree with α_{11}, and the edge between i and $i+1$ signed negative to agree with the product $\alpha_{i,i+1}\alpha_{i+1,i}$ for $i = 1, \ldots, n-1$. If a signed tree has a unique nonzero vertex, then we call it a *rooted tree*, with the nonzero vertex as the root.

LEMMA 3.1. *A potentially stable rooted t.s.p. is minimally potentially stable.*

The proof of the above statement is clear, since replacing any nonzero entry by zero results in a pattern that either has every diagonal entry equal to zero or has such an irreducible component and thus is not potentially stable. Note that not all minimally potentially stable sign patterns are represented by rooted trees; see, e.g., $A_{3.2}$ in section 5. Before giving the main result of this section, we need two more definitions. A *complete matching* in a signed tree is a set of disjoint edges and nonzero vertices that cover the vertex set of the tree. A complete matching exists if and only if there exists a nonsingular matrix belonging to this t.s.p. If a rooted t.s.p. has its root and each edge signed negative, then we call it a *canonical t.s.p.* Thus $\mathcal{A}_n$ of Example 2.4 is a canonical t.s.p. rooted at vertex 1.

THEOREM 3.2. *A t.s.p. represented by a rooted tree is minimally potentially stable if and only if it is a canonical t.s.p. with a complete matching.*

Proof. Assume that the rooted t.s.p. is canonical and has a complete matching. Since the t.s.p. has only one nonzero diagonal entry and it is negative, as in the proof of [8, Corollary 3.7], the t.s.p. allows a properly signed nest. By the Fisher–Fuller–Ballantine theorem [1], [3] (see the introduction), the t.s.p. is potentially stable and is minimal by Lemma 3.1. For the converse, assume that the t.s.p. is minimally potentially stable. By the proof of [8, Theorem 4.2], each edge of the t.s.p. is negative. Thus it is a canonical t.s.p., and potential stability implies that it has a complete matching. ∎

Note that in the statement of [8, Corollary 3.7] the necessary hypothesis that the t.s.p. has no positive diagonal entry was inadvertently omitted.

4. An optimal stable matrix of order 2. In this section we consider 2-by-2 minimally potentially stable sign patterns. From Theorem 3.2, the canonical t.s.p. pattern

$$\mathcal{A} = \begin{bmatrix} - & + \\ - & 0 \end{bmatrix}$$

is minimally potentially stable (and is in fact sign stable). It is readily checked that up to signature and permutation similarity, this is the only minimally potentially stable pattern of order 2. Since $\delta_{us}(A) = \delta_{us}(pA)$ for $p > 0$, without loss of generality assume throughout this section that

$$A = \begin{bmatrix} -1 & a \\ -b & 0 \end{bmatrix} \in \mathcal{A},$$

with $a, b > 0$. Thus, we need only to determine a, b so that $\delta_{us}(A) = \delta_{us}(\mathcal{A})$ gives an optimal stable matrix of order 2.

To determine $d_{us}(A)$, we introduce the following notation:

$$c = \operatorname{tr}(A^T A) = 1 + a^2 + b^2, \qquad d = c^2 - 4a^2b^2.$$

THEOREM 4.1. *For the 2-by-2 matrix A given above,*

$$d_{us}^2(A) = \begin{cases} \frac{4ab-1}{4(a+b)^2} & \text{when } d < (a+b)^4, \\ \frac{1}{2}(c - \sqrt{d}) & \text{otherwise.} \end{cases}$$

Proof. To use (2.2), let $f(\mu) = \sigma_{\min}(A - \mu i I)$, where A is the 2-by-2 matrix given above and $I = I_2$. Thus

$$f^2(\mu) = \text{min eigenvalue of } [(A - \mu i I)^*(A - \mu i I)].$$

It is readily verified that

$$\begin{aligned} \text{tr}[(A - \mu i I)^*(A - \mu i I)] &= c + 2\mu^2, \\ \det[(A - \mu i I)^*(A - \mu i I)] &= (ab - \mu^2)^2 + \mu^2, \end{aligned}$$

giving

$$f^2(\mu) = \frac{1}{2}(c + 2\mu^2 - \sqrt{d + 4\mu^2(a+b)^2}).$$

To determine $d_{us}(A)$, note that $d(f^2(\mu))/d\mu = 0$ if and only if either $\mu = 0$ or $(a+b)^4 = d + 4\mu^2(a+b)^2$, and note that $f^2(\mu)$ is even and $\to \infty$ as $\mu \to \pm\infty$. There are two cases to consider.

Case 1. Assume that $d < (a+b)^4$.

In this case $\mu = 0$ and $\mu^2 = \frac{1}{4}((a+b)^4 - d)/(a+b)^2$ specify 3 critical points. The value $\mu = 0$ gives a local maximum, with the other 2 values giving local minima. Thus

$$d_{us}^2(A) = \min_{\mu \in \mathcal{R}}\{f^2(\mu)\} = \frac{1}{2}\left(c + \frac{(a+b)^4 - d}{2(a+b)^2} - (a+b)^2\right) = \frac{(4ab-1)}{4(a+b)^2}.$$

Note that $d < (a+b)^4$ is equivalent to $(2 - 4ab)(a+b)^2 < 4ab - 1$. If $4ab - 1 \le 0$, then $2 - 4ab < 0$, which gives a contradiction; thus $4ab - 1 > 0$.

Case 2. Assume that $d \ge (a+b)^4$.

In this case $\mu = 0$ is the only critical point of $f^2(\mu)$, and thus $d_{us}^2(A) = \sigma_{\min}^2(A) = f^2(0) = (c - \sqrt{d})/2$. Note that $d_{us}^2(A)$ is continuous at $d = (a+b)^4$. ∎

Since the eigenvalues of A depend on the product ab (not on a and b individually), we use matrix B given by

$$B = \begin{bmatrix} -1 & t \\ -t & 0 \end{bmatrix} \in \mathcal{A}$$

in the remainder of this section. Theorem 4.1 readily gives $d_{us}^2(B)$ as follows.

COROLLARY 4.2. *For the 2-by-2 matrix B given above,*

$$d_{us}^2(B) = \begin{cases} \frac{4t^2-1}{16t^2} & \text{when } 1 + 4t^2 < 16t^4, \\ \frac{1}{2}(1 + 2t^2 - \sqrt{1+4t^2}) & \text{otherwise.} \end{cases}$$

With $t = \sqrt{ab}$, we now prove that $\delta_{us}(B)$ is at least as large as $\delta_{us}(A)$.

THEOREM 4.3. *For the 2-by-2 matrices A and B with $t = \sqrt{ab}$ defined above, $\delta_{us}(A) \le \delta_{us}(B)$, with equality if and only if $A = B$.*

Proof. We consider cases based on the values of d in Theorem 4.1.

Case 1. Assume that $d < (a+b)^4$.

This inequality is equivalent to $(2 - 4ab)(a+b)^2 < 4ab - 1$. Using the arithmetic-geometric mean inequality (AG) in the form $(a+b)^2 \ge (2\sqrt{ab})^2$, this implies that

$2 - 4ab < (4ab-1)/4ab$, which is equivalent to $1 + 4t^2 < 16t^4$. From Corollary 4.2, and Case 1 of Theorem 4.1,

$$d_{us}^2(B) = \frac{4t^2-1}{16t^2} \geq \frac{4ab-1}{4(a+b)^2} = d_{us}^2(A),$$

in which the inequality comes from using AG again.

Case 2. Assume that $d \geq (a+b)^4$. This case must be further subdivided into two cases based on the values of t in Corollary 4.2.

Case 2(i). Assume also that $1 + 4t^2 \geq 16t^4$. Then from Case 2 of Theorem 4.1, $d_{us}^2(A) = (c - \sqrt{d})/2 = 2a^2b^2/(c + \sqrt{d})$. Using $a^2 + b^2 \geq 2ab$ and $t^2 = ab$, this gives $d_{us}^2(A) \leq 2t^4/(1 + 2t^2 + \sqrt{1+4t^2}) = \frac{1}{2}(1 + 2t^2 - \sqrt{1+4t^2}) = d_{us}^2(B)$.

Case 2(ii). Assume that $1 + 4t^2 < 16t^4$. From Theorem 2.5,

$$d_{us}^2(A) = \min\{\alpha^2 : H(\alpha) \text{ has an eigenvalue with real part zero}\}.$$

For our 2-by-2 matrix A, the characteristic polynomial given at the end of section 2 for the 4-by-4 matrix $H(\alpha)$ becomes

$$\lambda^4 + E_2\lambda^2 + E_4 = 0,$$

in which $E_2 = 2ab + 2\alpha^2 - 1$ and $E_4 = \alpha^4 - c\alpha^2 + a^2b^2$. Setting $x = -\lambda^2$, $H(\alpha)$ has an eigenvalue λ with real part zero if and only if $x^2 - E_2x + E_4 = 0$ has a solution $x \geq 0$ (then $\lambda = \pm i\sqrt{x}$). This occurs if and only if $E_4 \leq 0$ or ($E_4 > 0$; $E_2 > 0$; and $E_2^2 \geq 4E_4$). This second condition is equivalent to ($\alpha^2 > (c + \sqrt{d})/2$ or $\alpha^2 < (c - \sqrt{d})/2$; $\alpha^2 > (1 - 2ab)/2$; and $\alpha^2 \geq (4ab-1)/4(a+b)^2$). If $(c - \sqrt{d})/2 \leq \alpha^2 \leq (c + \sqrt{d})/2$, then $E_4 \leq 0$, and $H(\alpha)$ has an eigenvalue with real part zero. Thus $d_{us}^2(A) \leq (c - \sqrt{d})/2$. Moreover, equality occurs if and only if either $(1 - 2ab)/2 \geq d_{us}^2(A)$ or $(4ab-1)/4(a+b)^2 \geq d_{us}^2(A)$. If the first inequality holds, then since $1 + 4t^2 < 16t^4$ is equivalent to $(1 - 2t^2)/2 < (4t^2-1)/16t^2$,

$$d_{us}^2(A) \leq \frac{1-2ab}{2} < \frac{4t^2-1}{16t^2} = d_{us}^2(B).$$

If the second inequality holds, then

$$d_{us}^2(B) = \frac{4t^2-1}{4(2t)^2} \geq \frac{4ab-1}{4(a+b)^2} \geq d_{us}^2(A),$$

in which the first inequality comes from AG. If equality holds here, then $A = B$, but this contradicts the conditions of Case 2(ii).

Thus in each case, $d_{us}^2(A) \leq d_{us}^2(B)$, and it is easily seen from Cases 1 and 2(i) that equality implies that $A = B$. By direct computation,

$$\|A\|_2^2 = \frac{1}{2}(c + \sqrt{d}) \geq \frac{1}{2}(1 + 2t^2 + \sqrt{1+4t^2}) = \|B\|_2^2,$$

with equality if and only if $A = B$. Taking positive square roots and using (2.3), gives $\delta_{us}(A) \leq \delta_{us}(B)$ with equality if and only if $A = B$. □

Consequently, from the above theorem, for the determination of an optimal stable matrix in the sign pattern class $\mathcal{A}$, it suffices to consider matrices of the form B with $t > 0$. The following result identifies all optimal minimally potentially stable matrices of order 2.

THEOREM 4.4. *For the sign pattern* $\mathcal{A} = \left[\begin{smallmatrix} - & + \\ - & 0 \end{smallmatrix}\right]$, $\delta_{us}(\mathcal{A}) = \sqrt{2}/(3\sqrt{3}) \approx 0.27217$. *Moreover this optimal value is achieved only at a matrix* pB, *where* $B = \left[\begin{smallmatrix} -1 & t \\ -t & 0 \end{smallmatrix}\right] \in \mathcal{A}$ *with* $t = \sqrt{3}/2$ *and* $p > 0$.

Proof. Since $\|B\|_2^2 = (1 + 2t^2 + \sqrt{1+4t^2})/2$, Corollary 4.2 can be restated with $x = t^2$ as

$$\delta_{us}^2(B_x) = \begin{cases} \frac{4x-1}{8x(1+2x+\sqrt{1+4x})} & \text{when } x > \frac{1+\sqrt{5}}{8}, \\ \frac{4x^2}{(1+2x+\sqrt{1+4x})^2} & \text{otherwise.} \end{cases}$$

This formula is used to determine the absolute maximum of $\delta_{us}(B_x)$ where the notation emphasizes the dependence of $\delta_{us}(B)$ on x. First, consider $x \le (1+\sqrt{5})/8$. The function $g(x) = 2x/(1+2x+\sqrt{1+4x})$ is positive and increasing for $x > 0$, and thus the maximum value of $\delta_{us}^2(B_x)$ on this interval occurs at the end point $x_1 = (1+\sqrt{5})/8$. To determine this maximum value, notice that $(1+\sqrt{5})/2$ is a solution of $y^2 - y - 1 = 0$. Thus $y = \sqrt{y+1}$, i.e., $(1+\sqrt{5})/2 = \sqrt{1+4x_1}$. Substituting this into the formula above gives

$$\delta_{us}(B_x) \le \delta_{us}(B_{x_1}) = \frac{1+\sqrt{5}}{7+3\sqrt{5}} \approx 0.23607$$

if $x \le (1+\sqrt{5})/8$. Second, consider $x > (1+\sqrt{5})/8$ and note that $\delta_{us}(B_x) \to 0$ as $x \to \infty$. Let $h(x) = (8x(1+2x+\sqrt{1+4x}))/(4x-1)$, and consider the minimum for x on this range. By differentiating, $h'(x) = 0$ implies that $64x^3 - 64x^2 + 8x + 3 = 0$, with the only solution in this range being at $x_2 = 3/4$. Thus using the formula above gives $\delta_{us}(B_{x_2}) = \sqrt{2}/(3\sqrt{3}) \approx 0.27217$. Since $\delta_{us}(B_{x_1}) < \delta_{us}(B_{x_2})$, it follows that $\delta_{us}(B)$ takes its maximum value when $t = \sqrt{3}/2$. The result then follows from Theorem 4.3 and the invariance of the relative distance under multiplication by a positive constant. ∎

For any normalized matrix $A = \left[\begin{smallmatrix} -1 & a \\ -b & 0 \end{smallmatrix}\right] \in \mathcal{A}$ (the sign pattern in Theorem 4.4), a lower bound on $\delta_{us}(A)$ can be found from Theorems 2.3 and 4.4 as

$$\delta_{us}(A) \ge \sqrt{2}/(3\sqrt{3}) - 2\max\{|a - \sqrt{3}/2|, |b - \sqrt{3}/2|\}/\|A\|_2.$$

5. Good stable matrices of higher orders. Even for $n = 2$, the explicit analytic determination of an optimal stable matrix in the minimally potentially stable sign pattern (given in section 4) is complicated, so we now turn our attention to "good" stable matrices (rather than optimal stable matrices) for $n \ge 3$. For a fixed sign pattern $\mathcal{A}$ with $\delta_{us}(\mathcal{A}) < M$, a stable matrix $A \in \mathcal{A}$ is called a *good* stable matrix if there is a positive number m such that $m < \delta_{us}(A)$ and the ratio m/M is not too small; in the following we use $m/M > 0.01$. For example, the tridiagonal pattern $\mathcal{A}_n$ of Example 2.4 has $\delta_{us}(\mathcal{A}_n) < 1/n = M$. For $n \ge 3$, let $A \in \mathcal{A}_n$ have $a_{11} = -1$ and $a_{i,i+1} = 0.8 - (i-1)s_n = -a_{i+1,i}$ for $i = 1, \ldots, n-1$, with $s_n = 0.6/(n-2)$. Numerically (by a procedure like that described after Theorem 2.5), for $3 \le n \le 550$, $\delta_{us}(A) > 1/(8n) = m$. Based on these computations for this range of n, A is a good stable matrix in the sign pattern $\mathcal{A}_n$, and furthermore $\delta_{us}(\mathcal{A}_n) = \mathcal{O}(\frac{1}{n})$. We believe that this statement holds for all $n \ge 3$. For $n = 3, 4, 5$, we are able to improve the bound m for the relative distance of this tridiagonal sign pattern; see $A_{3.1}, A_{4.1}, A_{5.1}$ below.

We now consider the numerical determination of a good stable matrix A in a minimally potentially stable t.s.p. $\mathcal{A}$ with $n \ge 3$. For matrix $A \in \mathcal{A}$, we assume

without loss of generality that $a_{11} = -1$ giving $||A||_2 \geq 1$, and we may choose A to have all of its strictly upper triangular entries nonnegative. Any other nonzero a_{ii} are chosen to have magnitude less than one. Thus by (2.1), for each pattern $\mathcal{A}$, the bound M is taken as the number of negative diagonal entries of $\mathcal{A}$ divided by n. Numerically we found that good stable matrices $A \in \mathcal{A}$ tend to have $|a_{ij}| = |a_{ji}|$ for all $i \neq j$, and those with a properly signed nest tend to have each nonzero $|a_{ij}|$ in the range $(0.1, 1]$. For patterns allowing a properly signed nest, we use an algorithm based on the ideas described at the end of section 2, with values $\varepsilon = 10^{-r}$ and sets $S = \{A_k\}$ given by $a_{ij}^{(k)} \in \{0.d_1 d_2 \dots d_r : d_t \text{ integer}, 0 \leq d_t \leq 9, d_1 \neq 0\}$. Details of the algorithm can be found in [10]. For patterns not allowing a properly signed nest, a systematic search over a large number of stable matrices is used to find a good stable matrix in the pattern.

The results of numerical experiments are now recorded by giving a good stable matrix in each (up to signature and permutation similarity) minimally potential t.s.p. for $n = 3$ and $n = 4$. With each such matrix, tight bounds for its relative distance to the unstable matrices are given. These give a lower bound for $\delta_{us}(\mathcal{A})$ in each sign pattern.

5.1. Good stable matrices for $n = 3$. All 3-by-3 potentially stable tree sign patterns are given in [9, Figure 2]. Of these, only two sign patterns are minimal: $A_{3.1}$ is represented by a canonical t.s.p. and $A_{3.2}$ has a properly signed nest.

$$A_{3.1} = \begin{bmatrix} -1 & 0.96 & 0 \\ -0.96 & 0 & 0.62 \\ 0 & -0.62 & 0 \end{bmatrix}, \qquad 0.15545 < \delta_{us}(A_{3.1}) \leq 0.15546,$$

$$A_{3.2} = \begin{bmatrix} -1 & 0.94 & 0.38 \\ -0.94 & 0.4 & 0 \\ 0.38 & 0 & 0 \end{bmatrix}, \qquad 0.044969 < \delta_{us}(A_{3.2}) \leq 0.044970.$$

5.2. Good stable matrices for $n = 4$. Potentially stable tree sign patterns of order 4 are given in [9, Figures 3 and 4, Table 1]. All (up to signature and permutation similarity) minimally potentially stable 4-by-4 tree sign patterns are listed below by representative good stable matrices. Two of these are canonical tree sign patterns, represented by $A_{4.1}, A_{4.2}$, seven more have a properly signed nest ($A_{4.3}$ to $A_{4.9}$), but the remaining two do not have this property ($A_{4.10}, A_{4.11}$).

$$A_{4.1} = \begin{bmatrix} -1 & 0.98 & 0 & 0 \\ -0.98 & 0 & 0.72 & 0 \\ 0 & -0.72 & 0 & 0.52 \\ 0 & 0 & -0.52 & 0 \end{bmatrix}, \qquad 0.10859 < \delta_{us}(A_{4.1}) \leq 0.10860,$$

$$A_{4.2} = \begin{bmatrix} -1 & 0.78 & 0 & 0.57 \\ -0.78 & 0 & 0.87 & 0 \\ 0 & -0.87 & 0 & 0 \\ -0.57 & 0 & 0 & 0 \end{bmatrix}, \qquad 0.10864 < \delta_{us}(A_{4.2}) \leq 0.10865,$$

$$A_{4.3} = \begin{bmatrix} -1 & 0.7 & 0.91 & 0 \\ 0.7 & -0.23 & 0 & 0 \\ -0.91 & 0 & 0 & 0.22 \\ 0 & 0 & 0.22 & 0 \end{bmatrix}, \qquad 0.023805 < \delta_{us}(A_{4.3}) \leq 0.023806,$$

$$A_{4.4} = \begin{bmatrix} -1 & 1 & 0 & 0.37 \\ -1 & 0 & 0.65 & 0 \\ 0 & 0.65 & -0.99 & 0 \\ 0.37 & 0 & 0 & 0 \end{bmatrix}, \qquad 0.03828 < \delta_{us}(A_{4.4}) \le 0.03829,$$

$$A_{4.5} = \begin{bmatrix} -1 & 1 & 0.46 & 0 \\ -1 & 0.4 & 0 & 0 \\ 0.46 & 0 & 0 & 0.13 \\ 0 & 0 & -0.13 & 0 \end{bmatrix}, \qquad 0.03174 < \delta_{us}(A_{4.5}) \le 0.03175,$$

$$A_{4.6} = \begin{bmatrix} -1 & 0.9 & 0 & 0 \\ -0.9 & 0 & 0.85 & 0.39 \\ 0 & -0.85 & 0.35 & 0 \\ 0 & 0.39 & 0 & 0 \end{bmatrix}, \qquad 0.04432 < \delta_{us}(A_{4.6}) \le 0.04433,$$

$$A_{4.7} = \begin{bmatrix} -1 & 0.95 & 0.4 & 0.18 \\ -0.95 & 0.54 & 0 & 0 \\ 0.4 & 0 & 0.1 & 0 \\ -0.18 & 0 & 0 & 0 \end{bmatrix}, \qquad 0.0117 < \delta_{us}(A_{4.7}) \le 0.0118,$$

$$A_{4.8} = \begin{bmatrix} -1 & 1 & 0.6 & 0 \\ -1 & 0.27 & 0 & 0.13 \\ 0.6 & 0 & 0 & 0 \\ 0 & 0.13 & 0 & 0 \end{bmatrix}, \qquad 0.0211 < \delta_{us}(A_{4.8}) \le 0.0212,$$

$$A_{4.9} = \begin{bmatrix} -1 & 0.46 & 0.466 & 0 \\ 0.46 & -0.593 & 0 & 0.102 \\ -0.466 & 0 & 0.27 & 0 \\ 0 & 0.102 & 0 & 0 \end{bmatrix}, \qquad 0.00835 < \delta_{us}(A_{4.9}) \le 0.00836,$$

$$A_{4.10} = \begin{bmatrix} -1 & 0.62 & 0 & 0 \\ -0.62 & 0.3595 & 0.7125 & 0 \\ 0 & 0.7125 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \qquad 0.007 < \delta_{us}(A_{4.10}) \le 0.008,$$

$$A_{4.11} = \begin{bmatrix} -1 & 0.9166 & 0 & 0 \\ 0.9166 & 0 & 4.8375 & 0 \\ 0 & -4.8375 & 0 & 3.6931 \\ 0 & 0 & 3.6931 & 0.249 \end{bmatrix}, \qquad 0.005 < \delta_{us}(A_{4.11}) \le 0.006.$$

Note that the minimally potentially stable sign patterns represented by $A_{4.8}$ to $A_{4.11}$ are not contained in [9]. Sign pattern $A_{4.11}$ is given in [8], whereas sign patterns $A_{4.8}$ to $A_{4.10}$ are new (having been discovered through a systematic search by Pang [11]). Canonical tree sign patterns give the largest relative distance to the unstable matrices, whereas the two sign patterns with three nonzero diagonal entries (represented by $A_{4.7}$ and $A_{4.9}$) and those patterns not allowing a properly signed nest (represented by $A_{4.10}$ and $A_{4.11}$) give small distances. These properties seem to persist in numerical results for larger n.

5.3. Good stable matrices with canonical tree sign patterns for $n = 5$. By using Theorem 3.2, it can be seen that (up to signature and permutation similarity) there are three minimally potentially stable canonical tree sign patterns for $n = 5$.

Good stable matrices are listed below.

$$A_{5.1} = \begin{bmatrix} -1 & 0.95 & 0 & 0 & 0 \\ -0.95 & 0 & 0.72 & 0 & 0 \\ 0 & -0.72 & 0 & 0.6 & 0 \\ 0 & 0 & -0.6 & 0 & 0.44 \\ 0 & 0 & 0 & -0.44 & 0 \end{bmatrix}, \qquad 0.08305 < \delta_{us}(A_{5.1}) \leq 0.08306,$$

$$A_{5.2} = \begin{bmatrix} -1 & 0.64 & 0 & 0.69 & 0 \\ -0.64 & 0 & 0.96 & 0 & 0 \\ 0 & -0.96 & 0 & 0 & 0 \\ -0.69 & 0 & 0 & 0 & 0.32 \\ 0 & 0 & 0 & -0.32 & 0 \end{bmatrix}, \qquad 0.08308 < \delta_{us}(A_{5.2}) \leq 0.08309,$$

$$A_{5.3} = \begin{bmatrix} -1 & 0.98 & 0 & 0 & 0 \\ -0.98 & 0 & 0.58 & 0.44 & 0 \\ 0 & -0.58 & 0 & 0 & 0.78 \\ 0 & -0.44 & 0 & 0 & 0 \\ 0 & 0 & -0.78 & 0 & 0 \end{bmatrix}, \qquad 0.083083 < \delta_{us}(A_{5.3}) \leq 0.083084.$$

6. Good stable matrices in general patterns. Our heuristic algorithm used to determine the good stable matrices in section 5 can be used for patterns with larger values of n that contain a properly signed nest. However, this is limited by the lack of both a complete list of minimally potentially stable tree sign patterns for $n \geq 5$ and a good upper bound M for $\delta_{us}(\mathcal{A})$.

For potentially stable sign patterns that are not minimal, we use the minimally potentially stable patterns as "atoms" to estimate their relative distance from the unstable matrices. For example, if

$$\mathcal{A} = \begin{bmatrix} - & + & + \\ + & + & + \\ + & - & - \end{bmatrix},$$

then it can be split into the $(1,1)$ entry and the lower 2-by-2 block. Thus a good stable matrix in this sign pattern is

$$A = \begin{bmatrix} -1 & \epsilon & \epsilon \\ \epsilon & \epsilon & \sqrt{3}/2 \\ \epsilon & -\sqrt{3}/2 & -1 \end{bmatrix},$$

with $0 < \epsilon \ll 1$, giving $\delta_{us}(\mathcal{A}) > 0.27$ (see Theorem 2.3 and section 4). Similarly, our results can be used to give a good stable matrix in a potentially stable sign pattern of larger order that contains a spanning forest of minimally potentially stable tree sign patterns of order 3, 4, or 5. Effective ways of splitting general potentially stable sign patterns into minimally potentially stable blocks remain to be explored.

Acknowledgments. We thank Charles R. Johnson for discussions that led to the problem addressed here, John H. Drew and James Pang for discussions on minimally potentially stable patterns, and the referees and editor for constructive suggestions.

REFERENCES

[1] C. S. Ballantine, *Stabilization by a diagonal matrix*, Proc. Amer. Math. Soc., 25 (1970), pp. 729–734.

[2] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.

[3] 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.

[4] D. Hale, G. Lady, J. Maybee, and J. Quirk, *Nonparametric Comparative Statics and Stability*, Princeton University Press, Princeton, NJ, 1999.

[5] C. He and G. A. Watson, *An algorithm for computing the distance to instability,* SIAM J. Matrix Anal. Appl., 20 (1998), pp. 101–116.

[6] N. J. Higham, *Matrix nearness problems and applications*, in Applications of Matrix Theory, M. J. C. Gover and S. Barnett, eds., Oxford University Press, London, 1989, pp. 1–27,

[7] C. Jeffries, V. Klee, and P. van den Driessche, *Qualitative stability of linear systems*, Linear Algebra Appl., 87 (1987), pp. 1–48.

[8] C. R. Johnson, J. S. Maybee, D. D. Olesky, and P. van den Driessche, *Nested sequences of principal minors and potential stability*, Linear Algebra Appl., 262 (1997), pp. 243–257.

[9] C. R. Johnson and T. S. Summers, *The potentially stable tree sign patterns for dimensions less than five*, Linear Algebra Appl., 126 (1989), pp. 1–13.

[10] Q. Lin, *The Distance of Potentially Stable Sign Patterns to the Unstable Matrices,* M.Sc. thesis, Department of Computer Science, University of Victoria, BC, Canada, 2001.

[11] J. C. Pang, *Potential Stability of Tree Sign Patterns of Dimension Less than Five*, Work term report, Department of Mathematics and Statistics, University of Victoria, BC, Canada, 1996.

[12] C. Van Loan, *How near is a stable matrix to an unstable matrix?*, in Linear Algebra and Its Role in Systems Theory, Contemp. Math. 47, AMS, Providence, RI, 1985, pp. 465–478.

[13] D. S. Watkins, *Fundamentals of Matrix Computations*, Wiley, New York, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 368–391

KRYLOV SUBSPACE METHODS FOR SADDLE POINT PROBLEMS WITH INDEFINITE PRECONDITIONING*

M. ROZLOŽNÍK† AND V. SIMONCINI‡

Abstract. In this paper we analyze the null-space projection (constraint) indefinite preconditioner applied to the solution of large-scale saddle point problems. Nonsymmetric Krylov subspace solvers are analyzed; moreover, it is shown that the behavior of short-term recurrence methods can be related to the behavior of preconditioned conjugate gradient method (PCG). Theoretical properties of PCG are studied in detail and simple procedures for correcting possible misconvergence are proposed. The numerical behavior of the scheme on a real application problem is discussed and the maximum attainable accuracy of the approximate solution computed in finite precision arithmetic is estimated.

Key words. saddle point problems, preconditioning, indefinite linear systems, finite precision arithmetic, conjugate gradients

AMS subject classification. 65F10

PII. S0895479800375540

1. Introduction.

We consider the symmetric indefinite system of linear equations

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.1}$$

where the $n \times n$ matrix block A is symmetric positive definite and the $n \times m$ block B has full column rank. We denote by M the coefficient matrix and for the system (1.1) we also use the notation $Mt = b$ with $t = [x; y]$ and $b = [f; g]$.

Systems of the form (1.1) arise in many application problems such as mixed or mixed-hybrid finite element discretization of partial differential equations and quadratic or nonlinear programming with equality constraints; see [30, 29, 28, 21, 10, 20] and their references.

For large two-dimensional and general three-dimensional problems, sparse direct methods are often unsuitable to solve the indefinite system (1.1); see, e.g., [28, 29]. On the other hand, due to the high sparsity of the coefficient matrix, the linear system (1.1) may be efficiently solved using iterative schemes. In order to improve the efficiency of standard iterative solvers, some preconditioning technique is commonly employed, such as simple diagonal scaling, incomplete factorization of the system matrix or its inverse, up to problem dependent preconditioning [32, 2, 31, 5, 25, 3, 34]. Block matrices such as that in (1.1) naturally lead to the implementation of ad-hoc algebraic preconditioning strategies that aim to exploit the block structure of the

*Received by the editors July 20, 2000; accepted for publication (in revised form) by G. H. Golub April 1, 2002; published electronically October 18, 2002. This work was carried out within the framework of the CNR Italy–Academy of Sciences of the Czech Republic bilateral contract CNR/AVCR, 1998–2000.

http://www.siam.org/journals/simax/24-2/37554.html

†Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic (miro@cs.cas.cz). Part of this author's work was supported by the Grant Agency of the Czech Republic under grants 101/00/1035 and 201/00/0080.

‡Dipartimento di Matematica, Università di Bologna, Piazza di Porta S., Donato, 5, 40127 Bologna, Italy, and Istituto di Matematica Applicata e Tecnologie Informatiche del CNR, I-27100 Pavia, Italy (valeria@dm.unibo.it).

original system; see, for instance, [28, 13, 6]. Especially attractive is positive definite preconditioning where symmetric solvers are regularly applicable [30, 34].

In our paper we concentrate on the use of the symmetric but indefinite preconditioner

$$P = \begin{bmatrix} I & B \\ B^T & 0 \end{bmatrix}. \tag{1.2}$$

This choice has been shown to be particularly effective on problems associated with constrained nonlinear programming [28, 20, 21]. More precisely, it has been shown that this preconditioner projects the problem onto the kernel of the constraint operator and that the constraint equation is exactly satisfied [29, 21]. In the partial differential equation context, preconditioner (1.2) is attractive when A (B) corresponds to a zero (one) order operator, as is the case, e.g., in mixed formulations of elliptic problems. Indeed, in such a setting the preconditioner in (1.2) is optimal in the sense that the number of iterations of the preconditioned solver to converge to a fixed tolerance does not depend on the problem dimension; we refer to [28] for a detailed analysis.

Due to the indefiniteness of the preconditioning matrix P, the preconditioned system is naturally nonsymmetric so that nonsymmetric solvers must be applied. Although this fact could be considered as a practical drawback, experience on real problems has demonstrated good performance of this approach [28, 20, 7, 21, 3, 4]. The computationally expensive generalized minimum residual (GMRES) method [33] can be applied on the preconditioned system; in practice, however, simplified versions of short-term recurrence methods such as nonsymmetric biconjugate gradient (BiCG) or quasi-minimum residual (QMR) [8] methods can also be used.

A thorough analysis of the preconditioner P for a class of magnetostatic problems, together with implementation considerations, can be found in [28]. In this paper we instead concentrate on algebraic properties of the preconditioned iteration process. We give general convergence results for the long-term recurrence method GMRES, and we derive a connection between short-term recurrence methods and the preconditioned conjugate gradient (PCG) approach. This analysis is motivated by the theoretical as well as numerical results in [21, 14], where CG and the conjugate residual method were successfully applied to the indefinite system (1.1) preconditioned by the indefinite preconditioner (1.2) for $g = 0$. We show the equivalence between CG and simplified BiCG when right-preconditioning is applied; the convergence analysis of preconditioned CG leads to the development of safeguard strategies to avoid possible misconvergence of the indefinite CG iteration. We also show that round-off may considerably influence the performance of the applied method, and we provide theoretical results on the behavior of the approximate solution in finite precision arithmetic. As a general result, we derive that the motivation for applying a diagonal prescaling of the block matrix A is threefold: (i) together with indefinite preconditioning it leads to independence of the problem size of the iterative solver [28]; (ii) it ensures convergence of the CG method in most cases; and (iii) it preserves numerical stability of the scheme in finite precision arithmetic.

The outline of the paper is as follows. In section 2 we study some theoretical properties of a general (nonsymmetric) Krylov subspace method applied to the preconditioned system and the setting for the subsequent sections is described. In section 3 several possible solution methods are discussed and related to previous works. The residual norm minimizing GMRES is studied in detail in section 4, and the related results are compared in subsequent sections with those of short-term recurrence

methods. The analysis of the case $g = 0$ starts in section 5. In the subsequent section it is shown that the (theoretical) rate of convergence of the preconditioned GMRES method, up to a small factor, depends only on the spectral distribution of the preconditioned matrix, making this computationally expensive method interesting from a theoretical point of view. The equivalence between simplified BiCG and CG is shown in section 7, so that in the subsequent sections the CG method is analyzed in detail. More precisely, in section 8 we prove that for the PCG method the indefinite M-inner product of the error decreases monotonically, whereas the residual norm can show completely different convergence history and it may even diverge unless special measures (correction or suitable scaling) are used to avoid this difficulty. In section 10 it is shown that not only the theoretical rate of convergence (measured by the easily computable residual norm) but also the maximum attainable accuracy level of the approximate solution computed in finite precision arithmetic depends on the scaling of the matrix block A. The use of the CG method applied to the suitably scaled symmetric indefinite system (1.1) together with indefinite preconditioning (1.2) and $g = 0$ is thus theoretically well justified. Numerical experiments also on a real application problem confirm the described theoretical results. In section 11 we draw our conclusions.

The notation used in this paper is as follows. MATLAB notation is always used when possible. Vectors corresponding to the large system will be usually split as $v = [v^{(1)}; v^{(2)}]$ with $v^{(1)} \in \mathbb{R}^n$ and $v^{(2)} \in \mathbb{R}^m$, unless different letters are given to the two block vectors. Given $x \in \mathbb{R}^n$, x^T denotes the transpose vector; the 2-norm of x is defined as $\|x\|^2 = x^T x = \sum_{i=1}^n x_i^2$ and the H-inner product as $\langle x, x\rangle_H = x^T H x$. The norm induced by the vector 2-norm is used for matrices. $\mathbb{P}_k$ indicates the set of polynomials of degree at most k. Finally, $\mathcal{N}(X)$ and $\mathrm{span}\{X\}$ indicate the null- and range-spaces of the matrix X, respectively.

2. Indefinite preconditioning. Given a starting approximation t_0 and the associated residual $r_0 = b - Mt_0$, the indefinite preconditioner P may be applied either from the right, yielding the system

$$MP^{-1}\hat{t} = r_0, \qquad t = P^{-1}\hat{t}, \tag{2.1}$$

or from the left, so that the system to be solved becomes

$$P^{-1}Mt = P^{-1}r_0, \tag{2.2}$$

(left-right preconditioning will not be considered in this paper, although it does not entail major consequences in the analysis). When standard nonsymmetric systems are preconditioned, the difference between the two approaches in (2.1) and (2.2) is that the former monitors the convergence of the true residual and preconditioned solution, whereas the latter monitors the preconditioned residual and the approximate solution to the original problem. We will see that for our particular problem there may be a close connection between the true residual and the preconditioned residual from the right and left preconditioned method, respectively, and their corresponding approximate solutions may even coincide for certain methods when carefully implemented.

The eigenvalues of $P^{-1}M$ and MP^{-1} are equal; therefore general spectral results can be given in terms of any of the two formulations. We first recall the following result [28, 21, 20]. Here and in the following, $\Pi = B(B^TB)^{-1}B^T$ denotes the orthogonal projector onto $\mathrm{span}\{B\}$.

PROPOSITION 2.1. *Let λ be an eigenvalue of MP^{-1}. Then either $\lambda = 1$ or λ is a nonzero eigenvalue of $(I - \Pi)A(I - \Pi)$.*

Due to the positive definiteness of A, the eigenvalues of MP^{-1} are thus all real and positive. Unfortunately, the matrix MP^{-1} is not diagonalizable and the standard analysis on the convergence rate of residual minimizing methods (see [16]) cannot be directly applied.

The inverse of the preconditioner P can be written as

$$P^{-1} = \begin{bmatrix} I - \Pi & B(B^TB)^{-1} \\ (B^TB)^{-1}B^T & -(B^TB)^{-1} \end{bmatrix}, \tag{2.3}$$

so that

$$MP^{-1} = \begin{bmatrix} A(I-\Pi)+\Pi & (A-I)B(B^TB)^{-1} \\ 0 & I \end{bmatrix}. \tag{2.4}$$

For brevity, we shall also use the notation

$$MP^{-1} = \begin{bmatrix} G & S \\ 0 & I \end{bmatrix} \tag{2.5}$$

with obvious meaning of G and S. Proposition 2.1 completely describes the eigenvalue distribution of matrix MP^{-1}. In particular, the eigenvalues of G are either unit or are eigenvalues of the symmetric matrix $(I-\Pi)A(I-\Pi)$ (see also [28]). One can show that the nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$ are contained in the smallest interval including the eigenvalues of A (see also section 9).

Due to the symmetry of the matrices M and P, the coefficient matrix in the left preconditioned system is partitioned as

$$P^{-1}M = (MP^{-1})^T = \begin{bmatrix} G^T & 0 \\ S^T & I \end{bmatrix}.$$

We would like to emphasize that neither P^{-1} nor the preconditioned matrix MP^{-1} are formed explicitly in practical implementations. Instead, the following factorization of P is exploited:

$$P = \begin{bmatrix} I & 0 \\ B^T & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -B^TB \end{bmatrix} \begin{bmatrix} I & B \\ 0 & I \end{bmatrix},$$

which yields a convenient factorization for its inverse,

$$P^{-1} = \begin{bmatrix} I & -B \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -(B^TB)^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -B^T & I \end{bmatrix} \equiv L^{-1}D^{-1}L^{-T}.$$

Therefore, a matrix-vector product $P^{-1}v$ is carried out as $P^{-1}v = L^{-1}(D^{-1}(L^{-T}v))$. Note that in general the highest computational cost is due to the system solution with B^TB, as would be the case if block diagonal preconditioners were applied (see, e.g., [34]); performance comparisons between these two approaches can be found in [28] for matrices stemming from a magnetostatic problem.

When solving the right preconditioned system with a Krylov subspace method,[1] the subspace $K_k(MP^{-1}, r_0)$ is computed, while left preconditioning computes the subspace $K_k(P^{-1}M, P^{-1}r_0)$. Vectors belonging to Krylov subspaces can be written

[1] Given a matrix H and a vector v, a Krylov subspace of at most dimension k is the space spanned by $\{v, Hv, \ldots, H^{k-1}v\}$ and is denoted by $K_k(H, v)$.

in terms of polynomials; therefore, if $v \in K_{k+1}(M, r_0)$, then $v = \phi(M)r_0$ for some polynomial $\phi \in \mathbb{P}_k$ [32].

We next show that vectors in $K_{k+1}(MP^{-1}, r_0)$ and in $K_{k+1}(P^{-1}M, P^{-1}r_0)$ can in fact be written in terms of polynomials in the matrix G defined in (2.5). These results will be used in the next sections to describe the residual behavior of selected Krylov subspace methods.

LEMMA 2.2. *A vector $v \in K_{k+1}(MP^{-1}, r_0)$ can be written as*

$$v = \phi_k(MP^{-1})r_0 = \begin{bmatrix} \phi_k(G)r_0^{(1)} + \psi_{k-1}(G)Sr_0^{(2)} \\ \phi_k(1)r_0^{(2)} \end{bmatrix}, \quad \phi_k \in \mathbb{P}_k, \tag{2.6}$$

where the polynomial ψ_{k-1} is of degree at most $k-1$ and is defined as

$$\psi_{k-1}(\lambda) = \begin{cases} \phi_k'(\lambda), & \lambda = 1, \\ \frac{\phi_k(\lambda)-\phi_k(1)}{\lambda-1}, & \lambda \neq 1. \end{cases} \tag{2.7}$$

Proof. By explicitly writing the polynomial we see that the vector v satisfies

$$v^{(1)} = \phi_k(MP^{-1})r_0|_{1:n} \quad v^{(2)} = \phi_k(1)r_0^{(2)}.$$

Moreover, since $(MP^{-1})^k r_0|_{1:n} = G^k r_0^{(1)} + G^{k-1}Sr_0^{(2)} + G^{k-2}Sr_0^{(2)} + \cdots + Sr_0^{(2)}$, we obtain for the polynomial $\phi_k(\lambda) = \sum_{i=0}^m \alpha_i \lambda^i$,

$$\begin{aligned} \phi_k(MP^{-1})r_0|_{1:n} &= \alpha_0 r_0^{(1)} + \alpha_1(Gr_0^{(1)} + Sr_0^{(2)}) + \alpha_2(G^2 r_0^{(1)} + GSr_0^{(2)} + Sr_0^{(2)}) \\ &\quad + \alpha_3(G^3 r_0^{(1)} + G^2 Sr_0^{(2)} + GSr_0^{(2)} + Sr_0^{(2)}) + \cdots \\ &= \phi_k(G)r_0^{(1)} + \psi_{k-1}(G)Sr_0^{(2)}. \end{aligned}$$

The polynomial ψ is defined as

$$\psi_{k-1}(\lambda) = \alpha_1 + (1+\lambda)\alpha_2 + (1+\lambda+\lambda^2)\alpha_3 + \cdots + (1+\lambda+\cdots+\lambda^{k-1})\alpha_k.$$

For $\lambda = 1$, $\psi_{k-1}(1) = \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k = \phi_k'(1)$. For $\lambda \neq 1$ we can write $(1+\lambda+\cdots+\lambda^{k-1}) = (1-\lambda^k)(1-\lambda)^{-1}$ so that

$$\begin{aligned} \psi_{k-1}(\lambda) &= (1-\lambda)^{-1}\left((1-\lambda)\alpha_1 + (1-\lambda^2)\alpha_2 + \cdots + (1-\lambda^k)\alpha_k\right) \\ &= (1-\lambda)^{-1}(\phi(1) - \phi(\lambda)). \end{aligned}$$

∎

More comments on the role of ϕ_k and ψ_{k-1} will be given in the next sections.

We next show that a similar relation for the Krylov subspace generated with the left preconditioned matrix can be obtained. We also observe that a polynomial description of an element $w \in K_{k+1}(P^{-1}M, P^{-1}r_0)$ could also be obtained directly from the previous result as $w = P^{-1}\phi_k(MP^{-1})r_0$, yielding, however, a less insightful relation, at least for general r_0 (cf. section 5 for the case $r_0 = [r_0^{(1)}; 0]$).

LEMMA 2.3. *A vector $w \in K_{k+1}(P^{-1}M, \tilde{r}_0)$ with $\tilde{r}_0 = P^{-1}r_0$ can be written as*

$$w = \phi_k(P^{-1}M)P^{-1}r_0 = \begin{bmatrix} \phi_k(G^T)\tilde{r}_0^{(1)} \\ S^T\psi_{k-1}(G^T)\tilde{r}_0^{(1)} + \phi_k(1)\tilde{r}_0^{(2)} \end{bmatrix}, \quad \phi_k \in \mathbb{P}_k, \tag{2.8}$$

with ψ_{k-1} as in (2.7).

Although left and right preconditioning in general generate different spaces in which an approximate solution is computed, the first block of the approximate solution to the original problem (1.1) always belongs to the same space, regardless of the side the preconditioner is employed. This is shown in the following proposition.

PROPOSITION 2.4. *Let $t_k = [x_k; y_k]$ be the approximate solution to* (1.1) *either in $K_k(MP^{-1}, r_0)$ or in $K_k(P^{-1}M, P^{-1}r_0)$. Then $x_k = \phi(G^T)\tilde{r}_0^{(1)}$ for some $\phi \in \mathbb{P}_{k-1}$, where $\tilde{r}_0 = P^{-1}r_0 = [\tilde{r}_0^{(1)}; \tilde{r}_0^{(2)}]$. (The polynomial may not be the same for the two spaces.)*

Proof. We first show that t_k belongs to $K_k(P^{-1}M, P^{-1}r_0)$ for both right and left preconditioning. For left preconditioning, the result follows from Lemma 2.3.

Let V_k be a basis of $K_k(MP^{-1}, r_0)$ satisfying

$$MP^{-1}V_k = V_{k+1}H_k \tag{2.9}$$

and $H_k \in \mathbb{R}^{(k+1)\times k}$ upper Hessenberg. It can be shown that $Q_k = P^{-1}V_k$ is a basis of $K_k(P^{-1}M, P^{-1}r_0)$. Let $\hat{t}_k = V_k z_k \in K_k(MP^{-1}, r_0)$ be an approximate solution to the right preconditioned system $MP^{-1}\hat{t} = r_0$. Then the approximate solution t_k to the unpreconditioned system $Mt = r_0$ is computed as $t_k = P^{-1}\hat{t}_k = P^{-1}V_k z_k = Q_k z_k$ so that $t_k \in K_k(P^{-1}M, P^{-1}r_0)$.

Using (2.9), we obtain $P^{-1}MQ_k = Q_{k+1}H_k$, so that the basis $Q_k = [Q_k^{(1)}; Q_k^{(2)}]$ satisfies

$$\begin{bmatrix} G^T & 0 \\ S^T & I \end{bmatrix} \begin{bmatrix} Q_k^{(1)} \\ Q_k^{(2)} \end{bmatrix} = \begin{bmatrix} Q_{k+1}^{(1)} \\ Q_{k+1}^{(2)} \end{bmatrix} H_k$$

and, in particular, $G^T Q_k^{(1)} = Q_{k+1}^{(1)} H_k$. Therefore, $\operatorname{span}\{Q_k^{(1)}\} = K_k(G^T, q_1^{(1)})$, where $q_1^{(1)}$ is the first vector of the matrix $Q_k^{(1)}$. Recalling from $t_k = Q_k z_k$ that $x_k = Q_k^{(1)} z_k$, the result follows. ∎

The proposition above shows that the convergence to the first block of the solution may depend only on the properties of the matrix G.

3. Solution methods. The preconditioned coefficient matrix is nonsymmetric, therefore nonsymmetric solvers seem to be required. Preconditioned GMRES determines an approximate solution in the generated Krylov subspace so as to minimize its residual 2-norm. This optimality condition is obtained by explicitly constructing an orthogonal basis of the computed Krylov subspace [32]. Due to the high computational cost per iteration, GMRES in its original implementation is discarded in practical situations. Cheaper methods are preferred: these give up optimality by omitting the generation of the full orthogonal basis (e.g., restarted GMRES, BiCG, BiCGSTAB).

Classical two-sided Lanczos-type approaches such as BiCG employ short-term recurrences to generate the subspace by imposing a biorthogonality condition between the basis elements of two distinct subspaces. The computational cost grows only linearly with the number of iterations, while quasi-monotonic behavior of the residual norm may be obtained by employing a smoothing procedure [8, 36]. Given the starting residual r_0 and an auxiliary vector $\tilde{r}_0$, the two Krylov subspaces $K_k(MP^{-1}, r_0)$ and $K_k((MP^{-1})^T, \tilde{r}_0)$ are constructed if right preconditioning is used; the two spaces are usually called right and left Krylov subspaces. Analogously, if left preconditioning is considered, the right and left generated spaces are $K_k(P^{-1}M, P^{-1}r_0)$ and $K_k((P^{-1}M)^T, \tilde{r}_0)$. By comparing the two preconditioning approaches, it is clear

that right preconditioning with $\tilde{r}_0 = P^{-1}r_0$ exactly corresponds to reversing the role of right and left spaces in the left preconditioning with $\tilde{r}_0 = r_0$. This consideration, together with the result of Proposition 2.4, shows that left and right preconditionings of the indefinite problem provide similar information, at least for the first block of the approximate solution vector. We will see that care must be taken in the approximation of the second block when right or left preconditioning is applied. We should remark, however, that often the second block vector refers to terms that do not have physical meaning and therefore are discarded in real applications.

Because of the symmetry of P and M, a lot of redundant information is generated when constructing the right and left spaces. This is clearly seen when choosing $\tilde{r}_0 = P^{-1}r_0$ as auxiliary vector in right preconditioning. Indeed, in this case,

$$((MP^{-1})^T)^k\tilde{r}_0 = (P^{-1}M)^k\tilde{r}_0 = P^{-1}(MP^{-1})^k r_0 \quad \forall k \geq 0,$$

so that vectors in the left space $K_k((MP^{-1})^T, \tilde{r}_0)$ can be simply obtained by premultiplying by P^{-1} vectors in the right space $K_k(MP^{-1}, r_0)$.

This is a special case of the more general *J-symmetry* property. A matrix H is called J-symmetric if there exists a nonsingular matrix J such that $H^TJ = JH$, that is, H is (real) symmetric with respect to J. It was shown in [18] and later developed in [9] that J-symmetry can be exploited so as to decrease the computational cost of nonsymmetric Lanczos processes. In summary, when the coefficient matrix is J-symmetric, the auxiliary Lanczos recurrence that is used to generate the left space is obtained at low cost from the computed right basis vectors. For right preconditioning, $H = MP^{-1}$ and $J = P^{-1}$, while for left preconditioning, $H = P^{-1}M$ and $J = P$. We refer to [9] for implementation issues concerning J-symmetry. J-symmetry of the preconditioned matrix was used in [28, 29] to enhance the efficiency of iterative solvers on real application problems.

It already appears from the results given so far that if nonsymmetric short-term recurrence methods are applied, the analysis and the experimental results will substantially differ depending on the choice of the auxiliary vector. In this paper we shall focus on the special choice $\tilde{r}_0 = P^{-1}r_0$ for right preconditioning and $\tilde{r}_0 = r_0$ for left preconditioning, which lead to convenient computational savings as shown above. Moreover, we shall see that these choices of auxiliary vector $\tilde{r}_0$ also entail fundamental theoretical considerations.

4. General convergence results. General convergence results are not easily derived due to the nontrivial Jordan structure of the coefficient matrix MP^{-1}. This, however, turns out to be unnecessary, since the block form introduced in (2.4) allows us to write the residual norm in terms of polynomials in G. From these, upper bounds for the residual norm can be readily obtained. More insightful relations can be written when the right-hand side of the system (1.1) is of the form $[f; 0]$, that is, when $g = 0$. We anticipate that setting $g = 0$ is not restrictive, since the starting approximate solution can be chosen so as to fall in such a framework. We shall focus on the general case in this section, while the rest of the paper will be devoted to the analysis for $g = 0$.

For a diagonalizable coefficient matrix $C \in \mathbb{R}^{n\times n}$, a bound on the GMRES residual norm can be given as (see [16])

$$\|r_k^{\mathrm{GMRES}}\| \leq \|r_0\|\kappa_2(Q) \min_{\substack{\phi\in\mathbb{P}_k \\ \phi(0)=1}} \max_{i=1,\ldots,n} |\phi(\lambda_i)|,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of C and $\kappa_2(Q) := \|Q\| \, \|Q^{-1}\|$ is the condition number of its eigenvector basis Q. Although MP^{-1} does not have a full system of eigenvectors, using the notation and the result of Lemma 2.2, a bound on the convergence of the GMRES residual can be written in terms of polynomials in the matrix $G = A(I - \Pi) + \Pi$. Indeed, the right preconditioned GMRES residual satisfies

$$\begin{aligned} \|r_k^{\mathrm{GMRES}}\|^2 &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \|\phi(MP^{-1}) r_0\|^2 \\ &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \left(\|\phi(G) r_0^{(1)} + \psi(G) S r_0^{(2)}\|^2 + |\phi(1)|^2 \|r_0^{(2)}\|^2 \right), \end{aligned} \tag{4.1}$$

where the polynomial ψ is of degree at most $k-1$ and is defined through ϕ as in (2.7).

The presence of ψ in (4.1) shows that ϕ is chosen so as to have small derivative at the unit value, which seems to suggest that ϕ will grow only slowly in the neighborhood of one.

Analogously, using (2.8) with $\tilde{r}_0 = P^{-1} r_0$, left preconditioning gives

$$\begin{aligned} \|r_k^{\mathrm{GMRES}}\|^2 &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \|\phi(P^{-1}M) \tilde{r}_0\|^2 \\ &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \left(\|\phi(G^T) \tilde{r}_0^{(1)}\|^2 + \|S^T \psi(G^T) \tilde{r}_0^{(1)} + \phi(1) \tilde{r}_0^{(2)}\|^2 \right). \end{aligned} \tag{4.2}$$

5. The case $g = 0$. This section serves as introduction to the following sections, where we shall focus on the case in which the original problem satisfies $g = 0$. We note that even though $g \neq 0$, the starting approximate solution t_0 can be chosen so that the starting residual has the form $r_0 = [s_0; 0]$, yielding in practice an equivalent setting as if g were equal to the zero vector. For this reason, we shall assume throughout this and the following sections that $g = 0$ and $t_0 = 0$, so that $r_0 = [f; 0]$.

We start by analyzing right preconditioning, which provides the most unexpected results in practical circumstances. We will show that for $g = 0$ the convergence analysis of GMRES can be carried out by employing only the upper left block matrix G in (2.4). Moreover, we show that simplified BiCG behaves very differently than expected, and that its convergence is strictly related to that of preconditioned CG on the indefinite problem.

If left preconditioning is used, then the condition $g = 0$ may not lead to significant changes in the generation of the Krylov subspace basis. Indeed, the vector generating the Krylov subspace in such case is $\tilde{r}_0 = [(I - \Pi) r_0^{(1)}; (B^T B)^{-1} B^T r_0^{(1)}]$, which in general will not have zero blocks. We shall see later that this fact does not represent a serious difficulty for Lanczos-type methods.

In our analysis we will take advantage of some basic properties of matrices P^{-1} and MP^{-1} when applied to a vector $[v; 0]$. Namely, it follows that

$$P^{-1} \begin{pmatrix} v \\ 0 \end{pmatrix} = \begin{pmatrix} (I - \Pi) v \\ (B^T B)^{-1} B^T v \end{pmatrix}, \quad MP^{-1} \begin{pmatrix} v \\ 0 \end{pmatrix} = \begin{pmatrix} Gv \\ 0 \end{pmatrix}. \tag{5.1}$$

Actually, there is a connection to the solution of the linear least squares problem associated with the matrix B and the right-hand side vector v: while the vector $(I - \Pi)v$ is the least squares residual, the vector $(B^T B)^{-1} B^T v$ is the least squares solution.

As pointed out by one referee, the case $g = 0$ can also be formulated as a (general) weighted least squares problem: the augmented linear system (1.1) is equivalent to

the system of normal equations $B^TA^{-1}By = B^TA^{-1}f$ (see [1, Chap. 4]). Various algebraic techniques could be considered for its solution and their efficiency evaluated on this type of problem. We refer here to [26, 27] and references therein; see also [35].

6. The GMRES method. By writing the GMRES residual as $r_k^{\mathrm{GMRES}} = \phi_k(MP^{-1})r_0$, where ϕ_k is the optimal GMRES residual polynomial, the optimality of the residual can be expressed only in terms of the matrix G; see also [28].

COROLLARY 6.1. *With the notation of Lemma* 2.2 *and for* $r_0 = [s_0; 0]$*, the right preconditioned GMRES residual satisfies*

$$\|r_k^{\mathrm{GMRES}}\| = \min_{\substack{\phi\in\mathbb{P}_k \\ \phi(0)=1}} \|\phi(G)s_0\|.$$

Assuming $G \equiv A(I-\Pi)+\Pi$ diagonalizable, we obtain

$$\|r_k^{\mathrm{GMRES}}\| \le \|r_0\|\kappa_2(Z) \min_{\substack{\phi\in\mathbb{P}_k \\ \phi(0)=1}} \max_{i=1,\ldots,n} |\phi(\lambda_i)|, \tag{6.1}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of G and Z is its eigenvector matrix [28]. Consequently, although the system matrix MP^{-1} is nondiagonalizable, the rate of convergence of preconditioned GMRES depends only on the eigenvalue distribution of the block $A(I-\Pi)+\Pi$ and on the conditioning of its eigenvector basis. In the following proposition, we show that the matrix $A(I-\Pi)+\Pi$ does have a full set of eigenvectors and give a bound for its condition number.

PROPOSITION 6.2. *Let us assume that the matrix* $A(I-\pi)+\pi$ *has* $n-m$ *nonunit eigenvalues* λ_i, $i = 1, \ldots n-m$. *Let* Z_2 *be an orthogonal basis of span*$\{\pi\}$ *and let the columns of* $Y_1 \in \mathbb{R}^{n\times(n-m)}$ *be eigenvectors of* $(I-\pi)A(I-\pi)$ *corresponding to all its nonzero eigenvalues. Then there exists an eigenvector matrix in the form* $Z = [Z_1, Z_2]$ *of* $A(I-\pi)+\pi$ *such that*

$$\kappa(Z) \le (1+\|\gamma\|)^2 \qquad \textit{with} \quad \|\gamma\| \le \frac{\|A\|}{\min_i |\lambda_i - 1|},$$

where $\Lambda = diag(\lambda_i)$ *and* $\gamma = Z_2^TAY_1(\Lambda - I)^{-1}$.

Proof. It is clear that the columns of the matrix Z_2 are eigenvectors of $A(I-\pi)+\pi$ corresponding to the unit eigenvalue. Moreover, the matrix Y_1 can be chosen so that it forms an orthogonal basis of $\mathcal{N}(B^T)$ satisfying $\pi Y_1 = 0$. Thus $[Y_1, Z_2]$ forms an orthogonal basis of $\mathbb{R}^n$. We next show that the matrix $\hat{Z} := (Y_1 + Z_2\gamma)(\Lambda - I)$, $\hat{Z} \in \mathbb{R}^{n\times(n-m)}$, is an eigenvector matrix of $A(I-\pi)+\pi$, such that $(A(I-\pi)+\pi)\hat{Z} = \hat{Z}\Lambda$. It follows from the definition of γ and Y_1 that $\hat{Z} = Y_1(\Lambda - I) + \pi AY_1$. Then

$$\begin{aligned}(A(I-\pi)+\pi)\hat{Z} &= AY_1\Lambda - AY_1 + \pi AY_1 \\ &= AY_1\Lambda - Y_1\Lambda = \pi AY_1\Lambda + (I-\pi)AY_1\Lambda - Y_1\Lambda \\ &= \pi AY_1\Lambda + Y_1\Lambda^2 - Y_1\Lambda = \hat{Z}\Lambda.\end{aligned}$$

Hence, the columns of $\hat{Z}$ are eigenvectors of $A(I-\pi)+\pi$. Since $\Lambda - I$ is diagonal, the matrix $Z_1 = Y_1 + Z_2\gamma$ is obtained from $\hat{Z}$ by column scaling, therefore, the columns of Z_1 are also eigenvectors of $A(I-\pi)+\pi$. Let us write $Z = [Z_1, Z_2]$. By using the orthogonal basis $[Y_1, Z_2]$, we next construct its inverse Z^{-1}. Let

$$Y_2 = -Y_1\gamma^T + Z_2.$$

The matrix $Y^T := [Y_1, Y_2]^T$ is the inverse of Z. Indeed, using the definition of Y_2 and Z_1, and recalling that $[Y_1, Z_2]$ is a square orthogonal matrix, we can write

$$Y^T Z = \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix} [Z_1, Z_2] = \begin{bmatrix} Y_1^T Z_1 & 0 \\ Y_2^T Z_1 & I \end{bmatrix},$$

and $Y_1^T Z_1 = Y_1^T (Y_1 + Z_2\gamma) = I$, $Y_2^T Z_1 = (-\gamma Y_1^T + Z_2^T)(Y_1 + Z_2\gamma) = -\gamma + \gamma = 0$. Moreover,

$$\begin{aligned} ZY^T &= [Z_1, Z_2] \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix} = Z_1 Y_1^T + Z_2 Y_2^T \\ &= (Y_1 + Z_2\gamma) Y_1^T + Z_2 (Z_2^T - \gamma Y_1^T) = Y_1 Y_1^T + Z_2 Z_2^T = I. \end{aligned}$$

This proves that Y^T is the inverse of Z. Using

$$[Y_1, Y_2] = [Y_1, Z_2] \begin{bmatrix} I & -\gamma^T \\ 0 & I \end{bmatrix}, \qquad [Z_1, Z_2] = [Y_1, Z_2] \begin{bmatrix} I & 0 \\ \gamma & I \end{bmatrix}$$

we obtain $\kappa(Z) = \|Y\| \, \|Z\| \le (1 + \|\gamma\|)^2$. The bound for $\|\gamma\|$ is immediate from its definition. ∎

Proposition 6.2 explicitly constructs an eigenvector basis Z. We note that the bound on $\kappa(Z)$ depends only on the (spectrum of the) matrix A. The matrix A can be scaled so that the nonunit eigenvalues of $(I - \Pi)A(I - \Pi)$ are sufficiently far from the unit eigenvalue, although such scaling can lead to an increase in the norm of A. In general the convergence behavior of GMRES thus depends only on the nonzero eigenvalue distribution of the symmetric matrix $(I - \Pi)A(I - \Pi)$ and on the norm of the matrix A.

Using standard results on Chebyshev polynomials to bound the polynomial min-max problem [16], we also obtain

$$\frac{\|r_k^{\mathrm{GMRES}}\|}{\|r_0\|} \le 2\zeta \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k,$$

where $\kappa = \lambda_{max}/\lambda_{min}$ stands for the ratio of the extremal (real) eigenvalues of matrix $A(I - \Pi) + \Pi$ (and so it is not its condition number!) and where ζ is a constant that bounds $\kappa_2(Z)$.

An analogous result is well known to hold for the M-inner product of the relative PCG error, with $\zeta = 1$ and when M and P are positive definite.

If using left preconditioning, fewer simplifications take place. Using Lemma 2.3, the following relation for the left preconditioned GMRES residual can be simply obtained. The minimization problem (6.3) follows from (5.1) with

$$P^{-1} \begin{bmatrix} \phi(G) s_0 \\ 0 \end{bmatrix} = \begin{bmatrix} (I - \Pi)\phi(G) s_0 \\ (B^T B)^{-1} B^T \phi(G) s_0 \end{bmatrix}. \tag{6.2}$$

Corollary 6.3. *The left preconditioned GMRES residual norm with $r_0 = [s_0; 0]$ can be written as*

$$\begin{aligned} \|r_k^{\mathrm{GMRES}}\|^2 &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0) = 1}} \|\phi(P^{-1}M) P^{-1} r_0\|^2 \\ &= \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0) = 1}} \left(\|(I - \Pi)\phi(G) s_0\|^2 + \|(B^T B)^{-1} B^T \phi(G) s_0\|^2 \right). \end{aligned} \tag{6.3}$$

More directly, from (6.2) we also obtain

$$\begin{aligned}\|r_k^{\mathrm{GMRES}}\| &\le \|P^{-1}\| \min_{\substack{\phi\in\mathbb{P}_k\\ \phi(0)=1}} \|\phi(G)s_0\| \\ &\le \|P^{-1}\|\|r_0\|\kappa_2(Z) \min_{\substack{\phi\in\mathbb{P}_k\\ \phi(0)=1}} \max_{i=1,\dots,n} |\phi(\lambda_i)|,\end{aligned}$$

where λ_i's are the eigenvalues of G with corresponding eigenvector matrix Z. The norm of P^{-1} is bounded as (cf., e.g., [30])

$$\|P^{-1}\| \le \max\left\{\frac{2}{\sqrt{1+4\sigma_{min}(B)^2}-1}, 1\right\},$$

where $\sigma_{min}(B)$ is the smallest singular value of B.

7. Other nonsymmetric solvers. In this section we briefly discuss nonsymmetric solvers that employ short-term recurrences and which can be used for solving our preconditioned system for $g = 0$.

Motivated by the considerations in section 3, we consider the implementation of simplified BiCG as a short-term recurrence approach. We next show that for $r_0 = [s_0; 0]$ and provided that $\tilde{r}_0 = P^{-1}r_0$, simplified BiCG is equivalent to the CG method applied to the system $Mt = b$ with preconditioner P.

We start by recalling the classical right preconditioned BiCG recurrence: given $r_0, \tilde{r}_0$ and setting $p_0 = r_0$ and $\tilde{p}_0 = \tilde{r}_0$, for $k = 0, 1, \dots$ we have

$$\begin{aligned}&\alpha_k = (\tilde{r}_k, r_k)/(\tilde{p}_k, MP^{-1}p_k),\\ &t_{k+1} = t_k + \alpha_k p_k,\\ &r_{k+1} = r_k - \alpha_k MP^{-1}p_k, \qquad \tilde{r}_{k+1} = \tilde{r}_k - \alpha_k P^{-1}M\tilde{p}_k,\\ &\beta_k = (\tilde{r}_{k+1}, r_{k+1})/(\tilde{r}_k, r_k),\\ &p_{k+1} = r_{k+1} + \beta_k p_k, \qquad \tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_k \tilde{p}_k.\end{aligned}$$

Using J-symmetry (with $J = P^{-1}$) and by setting $\tilde{r}_0 = P^{-1}r_0$ we obtain $\tilde{r}_k = P^{-1}r_k$ for all subsequent $k > 0$, and analogously for $\tilde{p}_k$. Therefore, the iterates $\tilde{r}_k, \tilde{p}_k$ can be computed explicitly from r_k, p_k, and the auxiliary "tilde" recurrence can be omitted. The resulting algorithm is nothing but the usual implementation of the CG method preconditioned with the indefinite matrix P [11]. In Figure 1 we report the obtained J-symmetric BiCG recurrence versus the PCG recurrence for the choice $r_0 = [s_0; 0]$. If we look at the formulae of both algorithms in the figure, it is clear that $\hat{\alpha}_k = \alpha_k$ and $\hat{\beta}_k = \beta_k$ and both algorithms are equivalent for $t_k = [x_k; y_k]$, and if $r_k = [s_k; 0]$, $p_k = P^{-1}[u_k; v_k]$. This condition can be easily proved. Indeed, if $r_0 = [s_0; 0]$ and due to (5.1), the vector $p_0 = P^{-1}r_0 = [p_0^{(1)}; p_0^{(2)}]$ satisfies $B^T p_0^{(1)} = 0$ which gives $Mp_0 = [Ap_0^{(1)} + Bp_0^{(2)}; 0]$. Using induction, one can show for all $j = 0, 1, \dots$ the properties $B^T p_{j+1}^{(1)} = 0$ and $Mp_{j+1} = [Ap_{j+1}^{(1)} + Bp_{j+1}^{(2)}; 0]$, which imply that r_{j+1} can be written in the form $r_{j+1} = [s_{j+1}; 0]$.

Equivalence can also be shown in the case of left preconditioning. Indeed, the P-symmetric BiCG applied to the preconditioned system with coefficient matrix $P^{-1}M$ and auxiliary vector $\tilde{r}_0 = r_0$, is equivalent to PCG. More precisely, the quantities t_k and p_k coincide, while the left preconditioned BiCG residual corresponds to the preconditioned residual iterates $P^{-1}r_k$.

We note that simplified QMR can also be viewed (at least in exact arithmetic) as simplified BiCG method with the QMR residual smoothing procedure applied on its top; cf., for instance, [17, 36].

P^{-1}-symmetric BiCG(MP^{-1})	PCG(M)
$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$	$t_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$
$b - M\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$	$r_0 = b - Mt_0 = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$
$\begin{pmatrix} u_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} s_0 \\ 0 \end{pmatrix}$	$p_0 = P^{-1}r_0$
$k = 0, 1, \ldots$	$k = 0, 1, \ldots$
$\hat{\alpha}_k = \dfrac{\left(\begin{pmatrix} s_k \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_k \\ 0 \end{pmatrix}\right)}{\left(MP^{-1}\begin{pmatrix} u_k \\ p_k \end{pmatrix}, P^{-1}\begin{pmatrix} u_k \\ p_k \end{pmatrix}\right)}$	$\alpha_k = \frac{(r_k, P^{-1}r_k)}{(p_k, Mp_k)}$
$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \hat{\alpha}_k P^{-1}\begin{pmatrix} u_k \\ v_k \end{pmatrix}$	$t_{k+1} = t_k + \alpha_k p_k$
$\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix} = \begin{pmatrix} s_k \\ 0 \end{pmatrix} - \hat{\alpha}_k MP^{-1}\begin{pmatrix} u_k \\ v_k \end{pmatrix}$	$r_{k+1} = r_k - \alpha_k Mp_k$
$\hat{\beta}_k = \dfrac{\left(\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix}\right)}{\left(\begin{pmatrix} s_k \\ 0 \end{pmatrix}, P^{-1}\begin{pmatrix} s_k \\ 0 \end{pmatrix}\right)}$	$\beta_k = \frac{(r_{k+1}, P^{-1}r_{k+1})}{(r_k, P^{-1}r_k)}$
$P^{-1}\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = P^{-1}\begin{pmatrix} s_{k+1} \\ 0 \end{pmatrix} + \hat{\beta}_k P^{-1}\begin{pmatrix} u_k \\ v_k \end{pmatrix}$	$p_{k+1} = P^{-1}r_{k+1} + \beta_k p_k$

FIG. 1. *Equivalence of right preconditioned BiCG and PCG for* $r_0 = [s_0; 0]$ *and* $\tilde{r}_0 = P^{-1}r_0$.

8. Preconditioned CG. In light of the considerations of the previous section, we see that simplified BiCG for $g = 0$ reduces to standard PCG applied on (1.1) with preconditioner P. Clearly, the indefiniteness of both M and P does not make the algorithm robust, and breakdown may occur, as observed in [21, 22]; however, in [21] safeguard strategies were suggested to overcome possible breakdown. In this section we give a closer look at the behavior of CG on the indefinite system (1.1) and give explicit formulae describing the possible (mis)convergence of the method.

Given the linear system $Mt = b$, initial guess t_0 with $r_0 = b - Mt_0$, and the preconditioner P, the PCG algorithm generates iterates t_k with residuals $r_k = b - Mt_k$ and preconditioned residuals $z_k = P^{-1}r_k$, $k = 1, 2, \ldots$ such that the error $e_k = t - t_k$ satisfies

$$e_k \in e_0 + \{z_0, \ldots, z_k\}, \quad e_k^T M z_j = e_k^T MP^{-1}Me_j = 0, \quad j = 0, \ldots, k-1.$$

If P and M were positive definite, then the M-inner product of e_k would be minimized over $e_0 + \{z_0, \ldots, z_k\}$. The error e_k can then be written in the form $e_k = \phi_k(P^{-1}M)e_0$, where ϕ_k is the CG polynomial of degree k such that $\phi_k(0) = 1$. The residual vector r_k satisfies $r_k = M\phi_k(P^{-1}M)e_0$ and

$$r_k \perp \{z_0, \ldots, z_k\}.$$

We have already shown that since $r_0 = [s_0; 0]$, then all subsequent r_j's have second block component equal to zero, that is, $r_j = [s_j; 0]$, $j = 0, 1, \ldots, k$. In particular, this implies that the approximate solution $[x_k; y_k]$ satisfies $B^T x_k = 0$ or, equivalently, $B^T e_k^{(1)} = 0$. The preconditioned residuals z_j, $j = 0, 1, \ldots, k$, then satisfy the relation

$Mz_j = [Gs_j; 0]$, so that the M-orthogonality of the error $e_k = [e_k^{(1)}; e_k^{(2)}]$ gives

$$0 = e_k^T M z_j = (e_k^{(1)})^T G s_j \qquad j = 0, \ldots, k-1.$$

Therefore, the condition on the error is only imposed on the first block component. Moreover, since $B^T e_k^{(1)} = 0$ for $k = 0, \ldots,$ then

$$e_k^T M e_k = (e_k^{(1)})^T A e_k^{(1)} > 0 \qquad \forall e_k^{(1)} \neq 0 \tag{8.1}$$

so that $\langle e_k, Me_k \rangle$ is always nonnegative. The next result follows from the properties of the preconditioned residual in the PCG method. To simplify the notation and using (8.1), from now on we shall write $\|e_k\|_M = \sqrt{\langle e_k, Me_k \rangle}$, even though M is not positive definite.

PROPOSITION 8.1. *Let $e_0 = [e_0^{(1)}; e_0^{(2)}]$ be the starting error of PCG. Then*

$$\|\phi_k(P^{-1}M)e_0\|_M = \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \|\phi(P^{-1}M)e_0\|_M = \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \|\phi((I-\Pi)A(I-\Pi))e_0^{(1)}\|_A.$$

Proof. We have to prove for every polynomial ϕ that

$$\|\phi(P^{-1}M)e_0\|_M = \|\phi((I-\Pi)A(I-\Pi))e_0^{(1)}\|_A.$$

Since $B^T e_0^{(1)} = B^T x = 0$, we have that $Me_0 = [Ae_0^{(1)} + Be_0^{(2)}; 0]$ and therefore $P^{-1}Me_0 = [(I-\Pi)Ae_0^{(1)}; \star]$. It also follows that $\phi(P^{-1}M)e_0 = [\phi((I-\Pi)A)e_0^{(1)}; \star]$. Since $e_0^{(1)} = (I-\Pi)e_0^{(1)}$, and using a similar approach as in (8.1), we obtain

$$\|\phi(P^{-1}M)e_0\|_M^2 = \|\phi((I-\Pi)A(I-\Pi))e_0^{(1)}\|_A^2.$$

Moreover, from the condition $(e_k^{(1)})^T G s_j = 0$, $j = 0, \ldots, k-1$, it follows that $e_k^{(1)} = \phi_k((I-\Pi)A(I-\Pi))e_0^{(1)} \perp (I-\Pi)A(I-\Pi)e_j^{(1)} = (I-\Pi)A(I-\Pi)\phi_j((I-\Pi)A(I-\Pi))e_0^{(1)}$, $j = 0, \ldots k-1$. Thus x_k is identical to the approximate solution after k iterations of the CG method applied to the system $(I-\Pi)A(I-\Pi)x = (I-\Pi)f$, and the statement follows. ∎

Since $e_0^{(1)} = (I-\Pi)e_0^{(1)}$, the indefinite M-inner product of the error $e_k = \phi_k(P^{-1}M)e_0$ is minimized only over the set of nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$. We thus have the following bound

$$\min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \|\phi(P^{-1}M)e_0\|_M \leq \|e_0^{(1)}\|_A \min_{\substack{\phi \in \mathbb{P}_k \\ \phi(0)=1}} \max_{\lambda \in [\alpha, \beta]} |\phi(\lambda)|, \tag{8.2}$$

where $[\alpha, \beta]$ is the smallest interval containing the nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$. Using once more standard Chebyshev polynomial results, we see that the indefinite M-inner product of the error decreases asymptotically at least as the optimal Chebyshev polynomial on $[\alpha, \beta]$. On the other hand, the residual norm of PCG (both the preconditioned residual and the true residual) does not obey the corresponding asymptotic rule, and the convergence curve may differ dramatically. This is due to the fact that the quantity $\|e_k\|_M$ may be zero for nonzero e_k, with $e_k^{(1)} = 0$ and $e_k^{(2)} \neq 0$ (cf. (8.1)), showing that $\|\cdot\|_M$ is not a definite norm. We next show that this is the reason why the energy norm (the indefinite M-inner product) fails to describe the convergence of the PCG residual on this problem. The residual $r_k = b - Mt_k$ satisfies

$$r_k = M\phi_k(P^{-1}M)e_0 = \begin{bmatrix} \phi_k(A(I-\Pi)+\Pi)s_0 \\ 0 \end{bmatrix}.$$

Let $A(I-\Pi)+\Pi = Z\Lambda Z^{-1}$ be the spectral decomposition of $A(I-\Pi)+\Pi$. Then

$$\begin{aligned}\|r_k\| &= \|\phi_k(A(I-\Pi)+\Pi)s_0\| \\ &\le \kappa_2(Z)\,\|s_0\|\,\|\phi_k(\Lambda)\| \le \kappa_2(Z)\,\|s_0\|\,\max\{\phi_{max}, |\phi_k(1)|\},\end{aligned} \tag{8.3}$$

where $\phi_{max} = \max_{\lambda\in[\alpha,\beta]}|\phi_k(\lambda)|$ and ϕ_k is the optimal PCG polynomial. While ϕ_{max} decreases as expected, $|\phi_k(1)|$ might not decrease (if it does at all). Therefore, the rate at which the bound of $\|r_k\|$ decreases depends on the value of the PCG polynomial ϕ_k at $\lambda = 1$. A similar dependence was already observed for GMRES. However, it is more crucial for PCG, since the optimal polynomial ϕ_k is minimized over the set of nonzero eigenvalues of $(I-\Pi)A(I-\Pi)$, which might not contain the value 1. Assuming, however, that 1 is an eigenvalue of $(I-\Pi)A(I-\Pi)$ and using the standard result on Chebyshev polynomials in (8.3) (see, e.g., [16]), the following estimate holds for the relative residual norm of PCG:

$$\frac{\|r_k\|}{\|r_0\|} \le 2\zeta\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \tag{8.4}$$

where ζ is a constant that bounds $\kappa_2(Z)$ (see Proposition 6.2) and κ stands for the ratio of extremal nonzero eigenvalues of the symmetric positive semidefinite matrix $(I-\Pi)A(I-\Pi)$, which can be bounded further by $\kappa_2(A)$. At first sight, this result may sound unexpected. Nevertheless, the convergence of PCG becomes natural when recalling the equivalence between indefinite preconditioning and the null-space method (cf. [29]).

We shall see in the next section that the problem can be scaled so that the condition $1 \in [\alpha,\beta]$ is satisfied. Provided that some eigenvalue of $(I-\Pi)A(I-\Pi)$ is reasonably close to 1 and that the polynomial ϕ_k does not pathologically blow up at $\lambda = 1$, then we can expect that the bound (8.4) holds.

The typical situation when $1 \notin [\alpha,\beta]$ is the occurrence of breakdown before the residual has dropped below the required (sufficiently small) tolerance. Nevertheless, there is a remedy how to avoid the unsuccessful termination of the PCG method. Since the first part of the error $e_k^{(1)}$ converges to zero, in exact arithmetic the computation terminates with the breakdown $(r_k, P^{-1}r_k) = (e_k, Me_k) = 0$ which results in $e_k^{(1)} = 0$. Then using $s_k = Ae_k^{(1)} + Be_k^{(2)} = Be_k^{(2)}$ we can correct the approximate solution (see [22]) as

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{pmatrix} e_k^{(1)} \\ e_k^{(2)} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{pmatrix} 0 \\ (B^TB)^{-1}B^Ts_k \end{pmatrix}. \tag{8.5}$$

In particular, this shows that checking the residual norm may be misleading and may lead to pessimistic expectation on the obtained approximation. In the lack of better knowledge of estimates on the error norm (cf., for instance, [12]), it is clearly desirable that this correction step be avoided and that the method terminate successfully on both components of the error. This is discussed in the next section.

The correction step in (8.5) suggests another useful strategy that attempts to avoid the difficulty with the erratic residual norm behavior. Instead of computing the second component of the approximate solution t_{k+1} via the CG iteration, Braess, Deuflhard, and Lipikov [3] compute the vector y_{k+1} using the minimum residual direction as

$$y_{k+1} = y_k + (B^TB)^{-1}B^Ts_k,$$

which locally minimizes at each step $k+1$ the residual norm

$$\|f - Ax_{k+1} - By_{k+1}\| = \min_{y \in \mathbb{R}^m} \|f - Ax_{k+1} - By\|;$$

cf. also the residual update strategy presented in [14].

9. Conjugate gradients and diagonal scaling. In the previous section we have shown that while the indefinite M-inner product of the PCG error must necessarily decrease, the 2-norm of the residual may not decrease at the same rate as the iteration proceeds, or it may not converge at all. The rate of convergence, when measured by the norm of residual, strongly depends on the value of the PCG polynomial at the eigenvalue 1, which may be outside the interval that contains the nonzero eigenvalues of $(I - \Pi)A(I - \Pi)$. This problem, however, can be easily overcome by prescaling the original coefficient matrix as described below.

If A is symmetric positive definite and $D = \text{diag}(A)$, then the eigenvalues of the matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ either are all ones or are contained in a nontrivial interval $[\alpha, \beta]$ strictly including the unit value.[2] However, this fact does not necessarily imply that the spectral interval of the projected matrix $(I - \Pi)D^{-\frac{1}{2}}AD^{-\frac{1}{2}}(I - \Pi)$ also includes the unit value, although this is usually the case. Standard theory only ensures that the nonzero eigenvalues of $(I-\Pi)D^{-\frac{1}{2}}AD^{-\frac{1}{2}}(I-\Pi)$ are contained in a subset of $[\alpha, \beta]$, which may or may not include the unit value. Nevertheless, this problem can be solved by means of a simple scalar scaling of A as follows. Let $v \in \mathbb{R}^n$ be any vector with unit norm such that $v = (I - \Pi)v$ and let $\chi = v^T Av > 0$. Then the smallest interval containing the nonzero eigenvalues of the matrix $F_\chi = (I-\Pi)(\chi^{-1}A)(I-\Pi)$ includes the unit value. Indeed, let $\lambda_{min}, \lambda_{max}$ be the nonzero smallest and largest eigenvalues of F_χ, respectively. Then

$$\lambda_{max} = \max_{0 \neq x} \frac{x^T F_\chi x}{x^T x} \geq v^T F_\chi v = 1$$

and, using standard variational arguments (see, e.g., [11]),

$$\lambda_{min} = \min_{0 \neq x \perp \text{span}\{B\}} \frac{x^T F_\chi x}{x^T x} \leq v^T F_\chi v = 1.$$

In terms of the quantities in the original problem, the theory above is recovered by simply rescaling the saddle point problem as

$$D_\chi^{-\frac{1}{2}} M D_\chi^{-\frac{1}{2}} \hat{t} = D_\chi^{-\frac{1}{2}} b, \qquad \hat{t} = D_\chi^{-\frac{1}{2}} t, \qquad D_\chi = \text{diag}(\chi I, \chi^{-1} I),$$

and then using the corresponding indefinite preconditioner. It should also be mentioned that scaling with D_χ does not affect the constraint matrix B. Moreover, different scaling of A and B is in general harmless, since B appears only in the analysis in a projector. Independent scaling of the columns in the matrix B may have, however, an effect on the accuracy of the least squares solution with matrix B in (5.1).

As a general implementation rule, we suggest to first scale A by its diagonal, which in several applications makes the preconditioned system independent of the

[2] This can be shown in a number of ways. Martin Gutknecht proposed the following: *The $n \times n$ matrix $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ has trace n. Since the trace is the sum of its (positive) eigenvalues, then either all eigenvalues are equal to* 1 *or there exist at least one eigenvalue less than* 1 *and one eigenvalue which is greater than* 1*, that is,* $1 \in]\alpha, \beta[$.

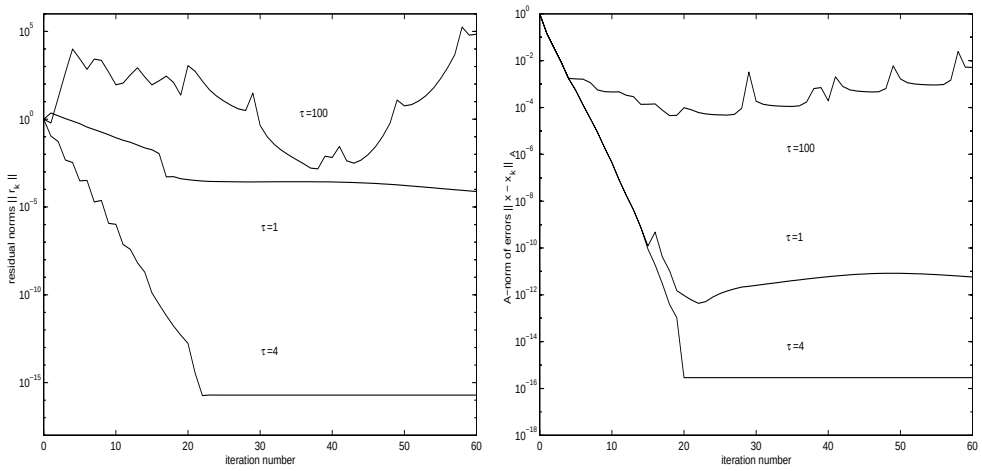


FIG. 2. *Residual norm (left) and error indefinite M-inner product (right) history of PCG for various values of τ.*

problem dimension, and then employ the additional scaling matrix D_χ to ensure a convenient location of the spectrum.

In the following examples we show the behavior of PCG with respect to the location of the interval $[\alpha, \beta]$. We emphasize that analogous results could be obtained by using simplified BiCG with right preconditioning.

We consider the following setting: $n = 25$, $s = 5$,

$$A = \text{tridiag}(1, \underline{4}, 1) \in \mathbb{R}^{n\times n}, \quad B = \text{rand}(n, s), \quad f = \text{rand}(n, 1), \quad g = 0,$$

where A is a tridiagonal matrix with constant diagonal elements equal to 4. The nonzero eigenvalues of $(I - \Pi)A(I - \Pi)$ are in the interval $[\alpha, \beta] = [2.1268, 5.8275]$. We consider two diagonal scalings of A that provide matrices $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ whose spectral interval is shifted. Since the diagonal of A is constant, this simply amounts to considering matrices of the form $D = \tau I$. We shall denote by $[\alpha_\tau, \beta_\tau]$ the corresponding eigenvalue interval. Clearly, $\tau = 1$ gives the original matrix, while $\tau > 1$ shifts $[\alpha_\tau, \beta_\tau]$ towards zero. The value $\tau = 4$ is optimal in the sense that it corresponds to the choice $D = \text{diag}(A)$. No scaling with χ, as described in section 9, is carried out.

In Figure 2 (left) the exact residual norm history of PCG for $\tau = 1, 4, 100$ is reported, while Figure 2 (right) shows the corresponding indefinite M-inner product of the error. Both residual and $\|e_k\|_M$ fall to machine precision level with the prescribed asymptotic convergence behavior for $\tau = 4$. For $\tau = 1$, $[\alpha_1, \beta_1] = [2.1268, 5.8275]$, and the residual norm does not decrease at the same rate as the indefinite M-inner product of the error, since the residual polynomial might not be small at the unit value. This is clearly observed in the figures. It should be mentioned, however, that we do not expect the residual to grow unboundedly because of the constraint $\phi(0) = 1$ (cf., e.g., Proposition 8.1). Mitigating effects on the residual norm (cf. Figure 2 (left)) no longer take place for $\tau = 100$, since $\alpha_\tau < \beta_\tau < 1$ and $\phi(1)$ may be substantially larger than one. Surprisingly, complete failure of the method is reported for $\tau = 100$, where at least the A-norm of the error should converge to zero, in exact arithmetic. In fact, finite precision arithmetic computation is responsible for this failure. The behavior of PCG on the indefinite problem in finite precision arithmetic is discussed

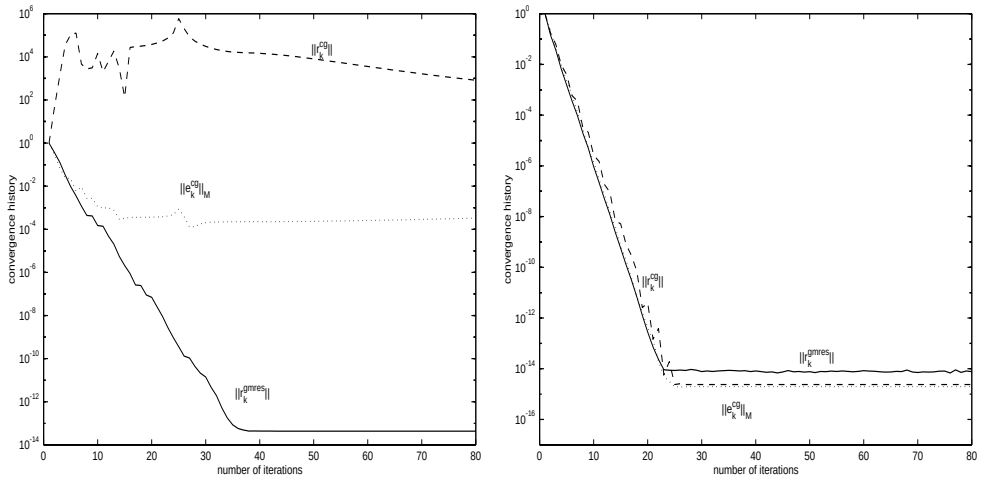


FIG. 3. *Convergence history of PCG and preconditioned GMRES on real application problem. Left: original problem; right: scaled problem.*

in section 10.

We next show the same kind of behavior on a real application problem. We consider the potential fluid flow problem in a rectangular domain with homogeneous Neumann conditions and Dirichlet conditions imposed on a part of the boundary [23, 25]. General prismatic discretization of the domain is used and a mixed-hybrid finite element formulation is considered [19, 23]. The lowest order Raviart–Thomas finite element approximation to the problem leads to the symmetric indefinite system of the form (1.1) of total dimension 868. Such small problem size was chosen for convenience; analogous results are obtained on much larger problems (see, e.g., [29, 28]). The positive definite block A represents a discrete form of the tensor in the Darcy law describing the physical properties (hydraulic permeability) of the porous medium in the domain. The off-diagonal block B describes the geometry of the domain and the fulfillment of Neumann boundary conditions. The dependence of the spectrum of M on the discretization parameter (mesh size) was analyzed in [24] and the rate of convergence of the unpreconditioned minimal residual (MINRES) method applied to the indefinite system (1.1) was estimated. The eigenvalues of the matrix $(I - \Pi)A(I - \Pi)$ are contained in $[4 \cdot 10^{-3}, 8 \cdot 10^{-2}]$. In Figure 3 we report the convergence history of preconditioned CG and GMRES on the unscaled (left plot) and scaled (right plot) problems. Scaling with χ was not necessary on this problem. The reported residual is the true residual given by the current approximate solution. In Figure 3 (left), the GMRES residual norm converges towards its maximum accuracy with the expected asymptotic slope. The spectral distribution explains the divergence of the CG residual, while the indefinite M-inner product of the CG error converges to its final accuracy after few iterations. The connection between the behavior of the error and the residual of PCG in finite precision arithmetic is discussed in detail in the next section.

Figure 3 (right) confirms that scaling optimally cures the problem, and maximum accuracy is obtained with both methods.

10. Behavior in finite precision arithmetic. We have experimentally observed in the previous section that round-off may cause convergence difficulties for

PCG on the indefinite problem. In this section we discuss the maximum attainable accuracy of the preconditioned CG scheme, measured in terms of the A-norm of the error $x - \bar{x}_k$, where $\bar{x}_k$ is the first part of the approximate solution $\bar{t}_k$ computed in finite precision arithmetic. Computed quantities will be identified by upper bar. For the A-norm of the error $x - \bar{x}_k$, the following bound holds in our case.

PROPOSITION 10.1. *The A-norm of the error $x - \bar{x}_k$ can be bounded as*

$$\|x - \bar{x}_k\|_A \leq \gamma_1\gamma_2\|\Pi(x - \bar{x}_k)\| + \gamma_3\|(I - \Pi)A(I - \Pi)(x - \bar{x}_k)\|,$$

where $\gamma_1 = \|A\|^{1/2}$, $\gamma_2 = (1 + (\kappa(A))^{1/2})$, and $\gamma_3 = \|A^{-1}\|^{1/2}$.

Proof. Since $\Pi x = 0$, the A-norm of the error $\bar{e}_k = x - \bar{x}_k$ can be written as

$$\begin{aligned}(10.1)\quad \|\bar{e}_k\|_A^2 &= (\Pi A\bar{e}_k, \Pi\bar{e}_k) + ((I - \Pi)A\bar{e}_k, (I - \Pi)\bar{e}_k) \\ &= (A\bar{e}_k, \Pi\bar{e}_k) + ((I - \Pi)A(I - \Pi)\bar{e}_k, \bar{e}_k) + ((I - \Pi)A\Pi\bar{e}_k, \bar{e}_k).\end{aligned}$$

Using some manipulation, we get

$$\begin{aligned}\|\bar{e}_k\|_A^2 &\leq \|A\bar{e}_k\|\|\Pi\bar{e}_k\| + \|(I - \Pi)A(I - \Pi)\bar{e}_k\|\|\bar{e}_k\| + \|(I - \Pi)A\Pi\bar{e}_k\|\|\bar{e}_k\| \\ &\leq \|A\|^{1/2}\|\bar{e}_k\|_A\|\Pi\bar{e}_k\| + \|(I - \Pi)A(I - \Pi)\bar{e}_k\|\|A^{-1}\|^{1/2}\|\bar{e}_k\|_A \\ &\quad + \|I - \Pi\|\,\|A\|\,\|\Pi\bar{e}_k\|\,\|A^{-1}\|^{1/2}\,\|\bar{e}_k\|_A,\end{aligned}$$

and the result follows. ∎

The first term on the right-hand side should be zero in exact arithmetic and it describes the departure of the computed iterate $\bar{x}_k$ from the null-space of B^T. The second term will converge to zero in exact arithmetic (see Proposition 8.1). By using a small modification of the proof in Proposition 10.1 we can get from (10.1) the following statement.

COROLLARY 10.2. *The A-norm of the error $x - \bar{x}_k$ can be bounded as*

$$(10.2)\qquad \|x - \bar{x}_k\|_A \leq \gamma_1\|\Pi(x - \bar{x}_k)\| + \gamma_3\|(I - \Pi)(f - A\bar{x}_k - B\bar{y}_k)\|.$$

The bound on $\|x - \bar{x}_k\|_A$ consists of two parts, the first of which is related to the departure of $\bar{x}_k$ from the null-space of B^T; the second part is related to the projection of the residual $f - A\bar{x}_k - B\bar{y}_k$ onto $\mathcal{N}(B^T)$. We next give some computable bounds for the A-norm of the error in terms of the gap between the true and updated residuals during the actual iteration process. In exact arithmetic the second part of the residual $r_k = [s_k; 0]$ should be zero. For the recursively updated residual vector $\bar{r}_k = [\bar{s}_k^{(1)}; \bar{s}_k^{(2)}]$ this is no longer the case, and we have

$$(10.3)\qquad (b - M\bar{t}_k) - \bar{r}_k = \begin{bmatrix} (f - A\bar{x}_k - B\bar{y}_k) - \bar{s}_k^{(1)} \\ B^T(x - \bar{x}_k) - \bar{s}_k^{(2)} \end{bmatrix}.$$

In finite precision arithmetic this quantity is no longer zero. In addition, it is a well-known fact that there is a limitation in the accuracy of the (true) residual vector obtained directly from the computed iterates $\bar{t}_k$. Namely, the quantity $\|b - M\bar{t}_k\|$ cannot decrease below a certain level, which is called the maximum attainable accuracy of the scheme. Using the theory of Greenbaum and after slight modification of Theorem 1 given in [15] we can formulate the following proposition.

PROPOSITION 10.3. *Assuming that the initial residual r_0 is computed exactly, the gap between the true residual $b - M\bar{t}_k$ and the recursively computed residual $\bar{r}_k$ can be bounded as*

$$(10.4)\quad \|(b - M\bar{t}_k) - \bar{r}_k\| \leq \varepsilon k\|M\|\left(\|t\| + (6 + 2\mu(n + m)^{1/2}) \max_{j=0,\ldots,k}\|t - \bar{t}_j\|\right),$$

where μ stands for the maximum number of nonzeros per row in the matrix M, and ε denotes the machine precision.

If we assume that the method converges, we can expect that even the norm of the recursively computed residual $\bar{r}_k$ will decrease far below the machine precision level. Consequently, from the bound for the gap we receive the bound for the maximum attainable accuracy level (measured by the true residual norm) which depends on the largest error norm during the whole process of convergence. It was shown by Greenbaum (see [15]) that the growth in the norm does not occur for the error or residual norm minimizing methods (with respect to any positive definite norm). Unfortunately, since in our case the "M-inner product" of the error is minimized, and since M is indefinite and does not induce a norm, these results cannot be applied directly to our scheme. The right-hand side of (10.4) can be further bounded in terms of the residual norm using $\varepsilon\|t - \bar{t}_j\| \leq \varepsilon\|M^{-1}\|\,\|\bar{r}_j\| + O(\varepsilon^2)$, therefore the bound on $\|(b - M\bar{t}_k) - \bar{r}_k\|$ depends in general on the maximum residual norm during the iteration steps $j = 0, \ldots, k$. We assume, however, that our problem is well-scaled and that the norm of the computed residual $\|\bar{r}_k\|$ converges far below machine precision. Under these assumptions, convergence is usually monotonic or nearly monotonic. Thus the maximum attainable accuracy, measured by the true residual norm, can be assumed to be at the level $p(k, \mu, n+m)\varepsilon\kappa(M)\|\bar{r}_0\|$, which is the level one gets for the standard CG algorithm (see [15]). Here, the term $p(k, \mu, n+m)$ stands for a low degree polynomial in k, μ, and $n+m$, and it does not play an important role in our considerations. The fact that the numerical behavior of this scheme depends heavily on the size of computed residuals is already known and it was analyzed in [14], where iterative refinement techniques and other residual update strategies were proposed in order to reduce the errors caused by large residuals; see also [3].

From Proposition 10.3 it also follows that the residual $\bar{s}_k^{(1)}$ is a good approximation to the true one $f - A\bar{x}_k - B\bar{y}_k$, provided we are above the limiting accuracy level given by the bound (10.4). This implies that the second term in the right-hand side of (10.2) is close to the computable quantity $\|(I - \Pi)\bar{s}_k^{(1)}\|$. For the first term in (10.2) we can write

$$\|\Pi(x - \bar{x}_k)\| \leq \delta_1 \|B^T(x - \bar{x}_k)\|, \tag{10.5}$$

where $\delta_1 = (\sigma_{min}(B))^{-1}$. It immediately follows from (10.3) that

$$\|B^T(x - \bar{x}_k) - \bar{s}_k^{(2)}\| \leq \|(b - M\bar{t}_k) - \bar{r}_k\| \tag{10.6}$$

and, again, provided that the residuals are above the level of maximum attainable accuracy, the second part of the updated residual $\bar{s}_k^{(2)}$ is a good approximation to the quantity $B^T(x - \bar{x}_k)$. So we can use (10.5) to obtain the bound in terms of $\|\bar{s}_k^{(2)}\|$ which is also easily computable. The A-norm of the error $x - \bar{x}_k$ is thus well-approximated (from above) by the maximum between the quantities $\gamma_1\delta_1\|\bar{s}_k^{(2)}\|$ and $\gamma_3\|(I - \Pi)\bar{s}_k^{(1)}\|$. In the case when the recursively computed residual $\bar{r}_k$ converges ultimately below the machine precision level, then $\|(I - \Pi)\bar{s}_k^{(1)}\|$ and $\|\bar{s}_k^{(2)}\|$ also converge below the machine precision level, and the quantities $B^T(x - \bar{x}_k)$ in (10.6) and $f - A\bar{x}_k - B\bar{y}_k$ in Corollary 10.2 can be bounded using Proposition 10.3. As a consequence, we obtain a bound on the level of maximum attainable accuracy of the method, measured by $\|x - \bar{x}_k\|_A$. On the other hand, if the system is badly scaled so that its unit eigenvalue is at the exterior of the spectral interval of $(I - \Pi)A(I - \Pi)$, then the quantities $\gamma_3\|(I - \Pi)\bar{s}_k^{(1)}\|$ and $\gamma_1\delta_1\|\bar{s}_k^{(2)}\|$ may remain at a much higher level. This leads to

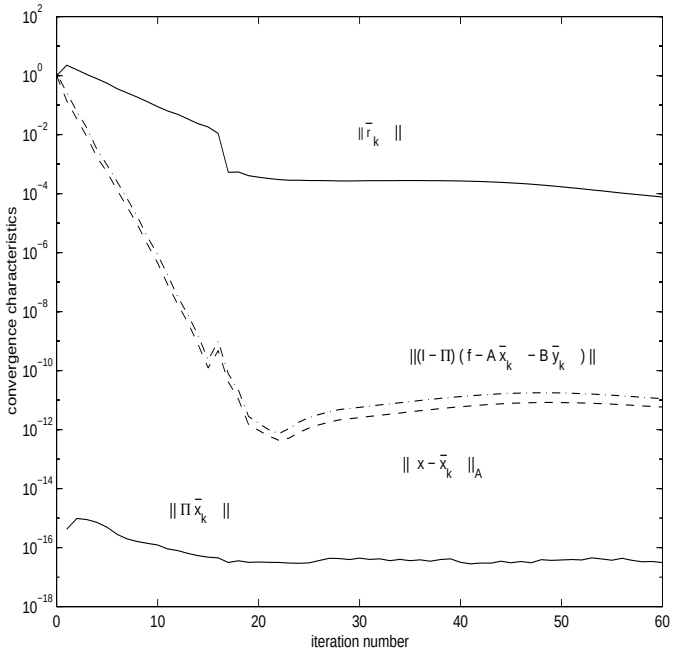


FIG. 4. *Behavior in finite precision arithmetic. Original problem.*

low accuracy of the computed $\bar{x}_k$, which is reflected in large $\|x - \bar{x}_k\|_A$. We can summarize the considerations above by saying that a proper scaling not only ensures the convergence of the residual norm in exact arithmetic but also allows us to obtain a satisfactory level of maximum attainable accuracy of the computed approximate solution $\bar{x}_k$.

We have already noticed at the end of section 8 that, in the general case, y_k may not converge to the solution y at all, so one can hardly expect some accuracy in the computed approximate solution $\bar{y}_k$, unless the correction step (8.5) is used. Nevertheless, assuming that the problem is well-scaled, y_k does converge, further considerations based on Proposition 10.3 can be made, and the accuracy of the computed second block $\bar{y}_k$ can be estimated. Indeed, we have

$$\|B(y - \bar{y}_k)\| \leq \|f - A\bar{x}_k - B\bar{y}_k\| + \|A(x - \bar{x}_k)\|. \tag{10.7}$$

Considering (10.7) and using the inequality $\|A(x - \bar{x}_k)\| \leq \|A\|^{1/2}\|x - \bar{x}_k\|_A$ we get the bound on $\|y - \bar{y}_k\|$

$$\|y - \bar{y}_k\| \leq \delta_1 \left(\|f - A\bar{x}_k - B\bar{y}_k\| + \gamma_1\|x - \bar{x}_k\|_A\right). \tag{10.8}$$

Considering the inequality from (10.3)

$$\|(f - A\bar{x}_k - B\bar{y}_k) - \bar{s}_k^{(1)}\| \leq \|(b - M\bar{t}_k) - \bar{r}_k\| \tag{10.9}$$

and assuming further that $\|\bar{r}_k\|$ is beyond the level of machine precision, the first term in (10.8) can be bounded using Proposition 10.3. Together with the bounds on $\|x - \bar{x}_k\|_A$, this gives us the level of maximum attainable accuracy of the scheme,

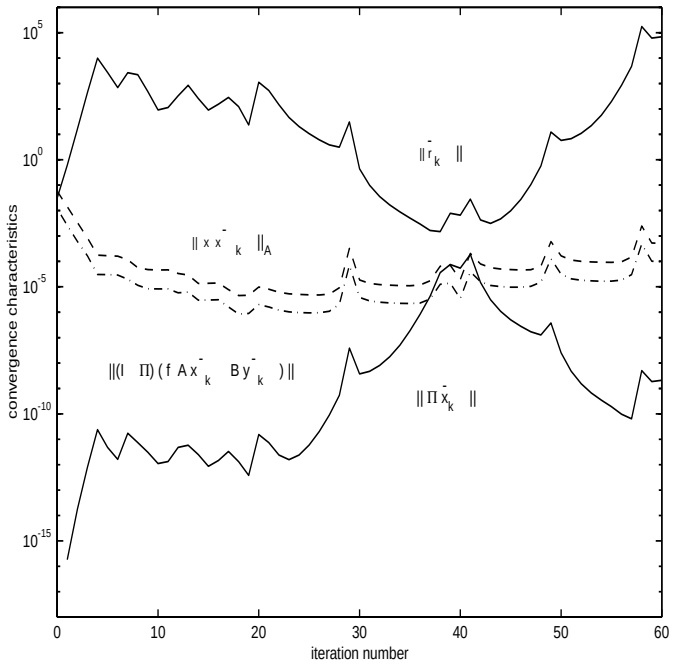


FIG. 5. *Behavior in finite precision arithmetic. Diagonal scaling $D = \tau I$ with $\tau = 100$.*

measured by $\|y - \bar{y}_k\|$. In the case the residual $\bar{r}_k$ is above its level of maximum attainable accuracy, the norm $\|y-\bar{y}_k\|$ is well-approximated by the maximum between the quantities $\delta_1\|\bar{s}_k^{(1)}\|$, $(\gamma_2 - 1)\delta_1\|(I - \Pi)\bar{s}_k^{(1)}\|$, and $\gamma_1\delta_1^2\|\bar{s}_k^{(2)}\|$.

In the following we report numerical experiments on the finite arithmetic behavior of the computed quantities generated during the CG recurrence. We consider the same 30×30 example as before and solve the system scaled by τ, for $\tau = 100, 4, 1$. In Figure 4 the true residual norm of PCG for $\tau = 1$ is reported (upper solid line). Since the method does not converge to the high accuracy level on the original problem, the solid line coincides fully with the norm of the updated residual vector $\|\bar{r}_k\|$. The norm of the departure from $\mathcal{N}(B^T)$, measured by $\|\Pi\bar{x}_k\|$ (lower solid line), remains close to the level of machine precision and is well-approximated by the term $\gamma_1\delta_1\|\bar{s}_k^{(2)}\|$ (not reported in the plot).

It is immediately clear from Figure 4 that the error $\|x - \bar{x}_k\|_A$ (dashed line) is determined by the second term of the bound (10.2) in Corollary 10.2. Due to the poor convergence of the residual norm, the quantity $\|(I-\Pi)(f - A\bar{x}_k - B\bar{y}_k)\|$ (dash-dotted line) coincides with $\|(I - \Pi)\bar{s}_k^{(1)}\|$. It is clear that in the case $\tau = 1$ this term determines the level of accuracy of the computed approximate solution $\bar{x}_k$.

Figure 5 shows the same quantities as Figure 4 for $\tau = 100$. For $\tau = 100$, the problem becomes even more badly scaled and the residual norm (of either the true or the updated residual—their difference is almost invisible) does not converge at all. Moreover, the departure from $\mathcal{N}(B^T)$ is no longer close to the level of machine precision and actually reaches the level of $\|x - \bar{x}_k\|_A$. This indicates that for very irregular residual behavior (or, in other words, very badly scaled problems) the first term in (10.2) may play an important role.

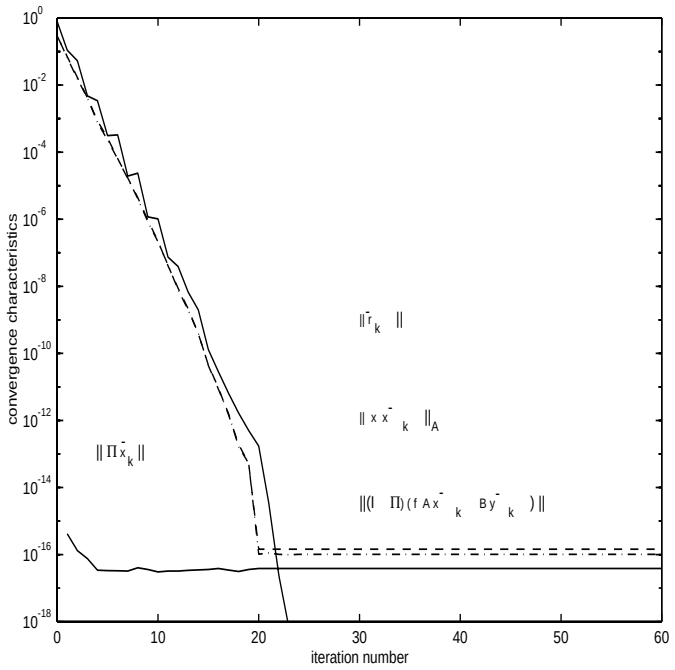


FIG. 6. *Behavior in finite precision arithmetic. Diagonal scaling* $D = \tau I$ *with* $\tau = 4$.

Figure 6 illustrates the behavior of PCG on the problem with optimal scaling $\tau = 4$. Both norms of the true and updated residual converge almost monotonically; while the true residual norm remains stagnating at machine precision level, the quantity $\|\bar{r}_k\|$ (upper solid line) converges even far beyond this level. Consequently the terms $\|\Pi\bar{x}_k\|$ and $\|(I-\Pi)(f-A\bar{x}_k-B\bar{y}_k)\|$ remain close to machine precision leading to a very accurate (whole) approximate solution $\bar{t}_k$.

11. Conclusions. Indefinite preconditioning has recently shown to be particularly attractive for solving saddle point problems arising from constrained nonlinear programming. Short-term recurrence nonsymmetric methods are applicable, at a cost comparable to that of symmetric solvers. However, numerical experience indicated that convergence was not always guaranteed (cf. [21, 22] for the indefinite CG method).

In this paper we have shown that there is a tight connection between short-term recurrence methods such as BiCG and the indefinite CG method used in [21]. More precisely, they are equivalent for a special choice of auxiliary vector, with which BiCG simplifies. Moreover, we have proved that the convergence of preconditioned CG strongly depends on the location of the unit eigenvalue with respect to the rest of the spectrum, so that if 1 is properly located, then convergence of preconditioned CG on the indefinite problem is usually achieved. We have shown that this condition is not restrictive, as it can be easily satisfied by scaling the original matrix. Scaling turns out to be fundamental also for the stability of the method.

In spite of its indefiniteness, we have thus shown that the scaled problem can be efficiently solved using CG with indefinite preconditioning approximately at the same asymptotic convergence rate as that given by preconditioned CG on a related positive definite problem.

Finally, it is interesting to note that numerical experiments related to the work in [28] showed that similar considerations with respect to the behavior of PCG seem to also hold for the problem

$$\begin{bmatrix} A & B \\ B^T & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

with C positive semidefinite, $\mathcal{N}(B) \neq \emptyset$, and $C+B^TB$ positive definite, which includes a wider class of problems than that treated in this paper.

Acknowledgments. The authors thank L. Lukšan for enlightening discussions on [21] and [22]. We thank M. Gutknecht, D. Braess, and M. Hanke for fruitful conversations. We thank one of the referees for insightful comments.

REFERENCES

[1] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[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, P. Deuflhard, and K. Lipikov, *A Subspace Cascadic Multigrid Method for Mortar Elements*, Preprint SC-99-07, Konrad-Zuse-Zentrum, Berlin, 1999.
[4] D. Braess and R. Sarazin, *An efficient smoother for the Stokes problem*, Appl. Numer. Math., 23 (1997), pp. 3–20.
[5] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.
[6] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Iterative methods for problems in computational fluid dynamics*, in Iterative Methods in Scientific Computing, R. H. Chan, C. T. Chan, and G. H. Golub, eds., Springer-Verlag, Singapore, 1997, pp. 271–327.
[7] 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-Verlag, Berlin, 1990, pp. 28–43.
[8] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[9] R. W. Freund and N. M. Nachtigal, *Software for simplified Lanczos and QMR algorithms*, Appl. Numer. Math., 19 (1995), pp. 319–341.
[10] P. E. Gill, W. Murray, D. B. Ponceleón, and M. A. Saunders, *Preconditioners for indefinite systems arising in optimization*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 292–311.
[11] G. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[12] G. H. Golub and G. Meurant, *Matrices, moments and quadrature.* II. *How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.
[13] G. H. Golub and A. J. Wathen, *An iteration for indefinite systems and its application to the Navier–Stokes equations*, SIAM J. Sci. Comput., 19 (1998), pp. 530–539.
[14] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1376–1395.
[15] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.
[16] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[17] M. Gutknecht and M. Rozložník, *Residual smoothing techniques: Do they improve the limiting accuracy of iterative solvers?*, BIT, 41 (2001), pp. 86–114.
[18] K. C. Jea and D. M. Young, *On the simplification of generalized conjugate-gradient methods for nonsymmetrizable linear systems*, Linear Algebra Appl., 52/53 (1983), pp. 399–417.
[19] E. F. Kaasschieter and A. J. M. Huijben, *Mixed-hybrid finite elements and streamline computation for the potential flow problem*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 221–266.
[20] C. Keller, N. I. M. Gould, and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.

[21] L. Lukšan and J. Vlček, *Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 219–247.

[22] L. Lukšan and J. Vlček, *Conjugate gradient methods for saddle point systems*, in Proceedings of the 13th Summer School on Software and Algorithms of Numerical Mathematics, I. Marek, ed., Nečtiny, Czech Republic, 1999, pp. 223–230.

[23] J. Maryška, M. Rozložník, and M. Tůma, *Mixed-hybrid finite element approximation of the potential fluid flow problem*, J. Comput. Appl. Math., 63 (1995), pp. 383–392.

[24] J. Maryška, M. Rozložník, and M. Tůma, *The potential fluid flow problem and the convergence rate of the minimal residual method*, Numer. Linear Algebra Appl., 3 (1996), pp. 525–542.

[25] J. Maryška, M. Rozložník, and M. Tůma, *Schur complement systems in the mixed-hybrid finite element approximation of the potential fluid flow problem*, SIAM J. Sci. Comput., 22 (2000), pp. 704–723.

[26] C. C. Paige, *Computer solution and perturbation analysis of generalized linear least squares problems*, Math. Comp., 33 (1979), pp. 171–184.

[27] C. C. Paige, *Fast numerically stable computations for generalized least squares problems*, SIAM J. Numer. Anal., 16 (1979), pp. 165–171.

[28] I. Perugia and V. Simoncini, *Block–diagonal and indefinite symmetric preconditioners for mixed finite element formulations*, Numer. Linear Algebra Appl., 7 (2000), pp. 585–616.

[29] I. Perugia, V. Simoncini, and M. Arioli, *Linear algebra methods in a mixed approximation of magnetostatic problems*, SIAM J. Sci. Comput., 21 (1999), pp. 1085–1101.

[30] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.

[31] T. Rusten and R. Winther, *Substructure preconditioners for elliptic saddle point problems*, Math. Comp., 60 (1993), pp. 23–48.

[32] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.

[33] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[34] 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.

[35] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.

[36] L. Zhou and H. F. Walker, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1994), pp. 297–312.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 392–410

A JACOBI–DAVIDSON TYPE METHOD FOR A RIGHT DEFINITE TWO-PARAMETER EIGENVALUE PROBLEM*

MICHIEL E. HOCHSTENBACH† AND BOR PLESTENJAK‡

Abstract. We present a new numerical iterative method for computing selected eigenpairs of a right definite two-parameter eigenvalue problem. The method works even without good initial approximations and is able to tackle large problems that are too expensive for existing methods. The new method is similar to the Jacobi–Davidson method for the eigenvalue problem. In each step, we first compute Ritz pairs of a small projected right definite two-parameter eigenvalue problem and then expand the search spaces using approximate solutions of appropriate correction equations. We present two alternatives for the correction equations, introduce a selection technique that makes it possible to compute more than one eigenpair, and give some numerical results.

Key words. right definite two-parameter eigenvalue problem, subspace method, Jacobi–Davidson method, correction equation, Ritz pair, inexact Newton method

AMS subject classifications. 65F15, 15A18, 15A69

PII. S0895479801395264

1. Introduction.

We are interested in computing one or more eigenpairs of a right definite two-parameter eigenvalue problem

$$\begin{aligned} A_1 x &= \lambda B_1 x + \mu C_1 x, \\ A_2 y &= \lambda B_2 y + \mu C_2 y, \end{aligned} \tag{1.1}$$

where A_i, B_i, and C_i are given real symmetric $n_i \times n_i$ matrices for $i = 1, 2$ and $\lambda, \mu \in \mathbb{R}$, $x \in \mathbb{R}^{n_1}$, $y \in \mathbb{R}^{n_2}$. A pair (λ, μ) is called an eigenvalue if it satisfies (1.1) for nonzero vectors x, y. The tensor product $x \otimes y$ is the corresponding eigenvector. The condition for right definiteness is that the determinant

$$\begin{vmatrix} x^T B_1 x & x^T C_1 x \\ y^T B_2 y & y^T C_2 y \end{vmatrix} \tag{1.2}$$

is strictly positive for all nonzero vectors $x \in \mathbb{R}^{n_1}$, $y \in \mathbb{R}^{n_2}$. Right definiteness and symmetry of matrices A_i, B_i, and C_i imply that there exist $n_1 n_2$ linearly independent eigenvectors for the problem (1.1) [2].

Multiparameter eigenvalue problems of this kind arise in a variety of applications [1], particularly in mathematical physics when the method of separation of variables is used to solve boundary value problems [22].

*Received by the editors September 17, 2001; accepted for publication (in revised form) by Z. Strakoš May 13, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/simax/24-2/39526.html
†Mathematical Institute, Utrecht University, P.O. Box 80 010, 3508 TA Utrecht, The Netherlands (hochstenbach@math.uu.nl).
‡IMFM/TCS, University of Ljubljana, Jadranska 19, SI-1000 Ljubljana, Slovenia (bor.plestenjak@fmf.uni-lj.si). This author's research was supported in part by the Ministry of Education, Science, and Sport of Slovenia (Research Project Z1-3136).

Two-parameter problems can be expressed as two coupled generalized eigenvalue problems. On the tensor product space $S := \mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2}$ of the dimension $N := n_1 n_2$, we define matrices

$$\begin{aligned} \Delta_0 &= B_1 \otimes C_2 - C_1 \otimes B_2, \\ \Delta_1 &= A_1 \otimes C_2 - C_1 \otimes A_2, \\ \Delta_2 &= B_1 \otimes A_2 - A_1 \otimes B_2 \end{aligned} \tag{1.3}$$

(for details on the tensor product, see, for example, [2]). Since the tensor product of symmetric matrices is symmetric, Δ_i is a symmetric matrix for $i = 0, 1, 2$. Atkinson [2, Theorem 7.8.2] proves that right definiteness of (1.1) is equivalent to the condition that Δ_0 is positive definite. He also shows that matrices $\Delta_0^{-1}\Delta_1$ and $\Delta_0^{-1}\Delta_2$ commute and that the problem (1.1) is equivalent to the associated problem

$$\begin{aligned} \Delta_1 z &= \lambda \Delta_0 z, \\ \Delta_2 z &= \mu \Delta_0 z \end{aligned} \tag{1.4}$$

for decomposable tensors $z \in S$, $z = x \otimes y$. The eigenvectors of (1.1) are Δ_0-orthogonal; i.e., if $x_1 \otimes y_1$ and $x_2 \otimes y_2$ are eigenvectors of (1.1) corresponding to different eigenvalues, then

$$(x_1 \otimes y_1)^T \Delta_0 (x_2 \otimes y_2) = \begin{vmatrix} x_1^T B_1 x_2 & x_1^T C_1 x_2 \\ y_1^T B_2 y_2 & y_1^T C_2 y_2 \end{vmatrix} = 0. \tag{1.5}$$

Decomposable tensors $x_i \otimes y_i$ for $i = 1, \ldots, N$ form a complete basis for S.

There exist numerical methods for right definite two-parameter eigenvalue problems. First, the associated problem (1.4) can be transformed in such a way that it can be solved by numerical methods for simultaneous diagonalization of commutative symmetric matrices [9, 14, 21]. This is only feasible for problems of low dimension as the size of the matrices of the associated problem is $N \times N$. Among other methods, we mention those based on Newton's method [7], the gradient method [5, 6, 8], and the minimal residual quotient iteration [4]. A deficiency of these methods is that they require initial approximations close enough to the solution in order to avoid misconvergence.

The continuation method [16, 17] overcomes problems with initial approximations, but, since the ordering of the eigenvalues is not necessarily preserved in a continuation step, we have to compute all eigenvalues even if we are interested only in a small portion. In this paper, we introduce a new numerical method which is similar to the Jacobi–Davidson method for the one-parameter eigenvalue problem [20]. The method can be used to compute selected eigenpairs and does not need good initial approximations.

Our method computes the exterior eigenvalue (λ, μ) of (1.1), which has the maximum value of $\lambda \cos\alpha + \mu \sin\alpha$ for a given α. We also present a version that computes the interior eigenpair closest to a given pair (λ_0, μ_0), i.e., the one with minimum $(\lambda - \lambda_0)^2 + (\mu - \mu_0)^2$.

The outline of the paper is as follows. We generalize the Rayleigh–Ritz approach to right definite two-parameter eigenvalue problems in section 2. In section 3, we present a Jacobi–Davidson type method for right definite two-parameter eigenvalue problems and introduce two alternatives for the correction equations. We discuss how the method can be used for exterior and interior eigenvalues in section 4. In section 5,

we present a selection technique that allows us to compute more than one eigenpair. The time complexity is given in section 6, and some numerical examples are presented in section 7. Conclusions are summarized in section 8.

2. Subspace methods and Ritz pairs. The Jacobi–Davidson method [20] is one of the subspace methods that may be used for the numerical solution of one-parameter eigenvalue problems. (For an overview of subspace methods, see, for example, [3].) The common principle of subspace methods is to compute accurate eigenpairs from low-dimensional subspaces. This approach reduces computational time and memory usage and thus enables us to tackle larger problems that are too expensive for methods that work in the entire space.

A subspace method works as follows. We start with a given search subspace from which approximations to eigenpairs are computed (*extraction*). In the extraction, we usually have to solve a smaller eigenvalue problem of the same type as the original one. After each step, we expand the subspace by a new direction (*expansion*). The idea is that, as the search subspace grows, the eigenpair approximations will converge to an eigenpair of the original problem. In order to keep computation costs low, we usually do not expand the search space to the whole space. If the process does not converge in a certain number of iterations, then the method is restarted with a few selected approximations as the basis of a new search space. In this section, we discuss the extraction, and, in the next section, we discuss the algorithm and the expansion.

The Rayleigh–Ritz approach defines approximations to the eigenpairs that can be extracted from the given subspace (see, for instance, [15]). We generalize the Rayleigh–Ritz approach for the two-parameter eigenvalue problem as follows. Suppose that the k-dimensional search subspaces $\mathcal{U}_k$ of $\mathbb{R}^{n_1}$ and $\mathcal{V}_k$ of $\mathbb{R}^{n_2}$ are represented by matrices $U_k \in \mathbb{R}^{n_1 \times k}$ and $V_k \in \mathbb{R}^{n_2 \times k}$ with orthonormal columns, respectively. The Ritz–Galerkin conditions

$$(A_1 - \sigma B_1 - \tau C_1)u \perp \mathcal{U}_k,$$
$$(A_2 - \sigma B_2 - \tau C_2)v \perp \mathcal{V}_k,$$

where $u \in \mathcal{U}_k \backslash \{0\}$ and $v \in \mathcal{V}_k \backslash \{0\}$, lead to the smaller projected right definite two-parameter problem

$$\begin{aligned} U_k^T A_1 U_k c &= \sigma U_k^T B_1 U_k c + \tau U_k^T C_1 U_k c, \\ V_k^T A_2 V_k d &= \sigma V_k^T B_2 V_k d + \tau V_k^T C_2 V_k d, \end{aligned} \tag{2.1}$$

where $u = U_k c \neq 0$, $v = V_k d \neq 0$, $c, d \in \mathbb{R}^k$, and $\sigma, \tau \in \mathbb{R}$.

We say that an eigenvalue (σ, τ) of (2.1) is a *Ritz value* for the two-parameter eigenvalue problem (1.1) and subspaces $\mathcal{U}_k, \mathcal{V}_k$. If (σ, τ) is an eigenvalue of (2.1) and $c \otimes d$ is the corresponding eigenvector, then $u \otimes v$ is a *Ritz vector*, where $u = U_k c$ and $v = V_k d$. Altogether, we obtain k^2 *Ritz pairs* that are approximations to the eigenpairs of (1.1). It is easy to check that, if $u \otimes v$ is a Ritz vector corresponding to the Ritz value (σ, τ), then σ and τ are equal to the tensor Rayleigh quotients [16]

$$\sigma = \rho_1(u,v) = \frac{(u \otimes v)^T \Delta_1 (u \otimes v)}{(u \otimes v)^T \Delta_0 (u \otimes v)} = \frac{(u^T A_1 u)(v^T C_2 v) - (u^T C_1 u)(v^T A_2 v)}{(u^T B_1 u)(v^T C_2 v) - (u^T C_1 u)(v^T B_2 v)},$$
$$\tau = \rho_2(u,v) = \frac{(u \otimes v)^T \Delta_2 (u \otimes v)}{(u \otimes v)^T \Delta_0 (u \otimes v)} = \frac{(u^T B_1 u)(v^T A_2 v) - (u^T A_1 u)(v^T B_2 v)}{(u^T B_1 u)(v^T C_2 v) - (u^T C_1 u)(v^T B_2 v)}.$$

In order to obtain Ritz values, we have to solve small right definite two-parameter eigenvalue problems. For this purpose, one of the available numerical methods that

computes all eigenpairs of a small right definite two-parameter eigenvalue problem can be used. For instance, the associated problem (1.4) can be solved using methods for simultaneous diagonalization of two commutative symmetric matrices [9, 14, 21].

3. Jacobi–Davidson method. The Jacobi–Davidson method [20] is a subspace method where approximate solutions of certain correction equations are used to expand the search space. Jacobi–Davidson type methods restrict the search for a new direction to the subspace that is orthogonal or oblique to the last chosen Ritz vector.

Jacobi–Davidson type methods have been successfully applied to the eigenvalue problem [20, 13], to the generalized eigenvalue problem [18], and to the singular value problem [12]. In this paper, we show that a Jacobi–Davidson type method can be applied to the right definite two-parameter problem as well.

A brief sketch of the Jacobi–Davidson type method for the right definite two-parameter problem is presented in Algorithm 3.1. In step 2(b), we have to decide which Ritz pair to select. We give details of this step in section 4, where we discuss how to deal with exterior and interior eigenvalues. In step 2(e), we have to find new search directions in order to expand the search subspaces. We will discuss two possible correction equations for step 2(e) later in this section.

ALGORITHM 3.1. A Jacobi–Davidson type method for a right definite two-parameter eigenvalue problem.

1. **Start.** Choose initial nontrivial vectors u and v.
   (a) Compute $u_1 = u/\|u\|$, $v_1 = v/\|v\|$, and set $U_1 = [u_1]$, $V_1 = [v_1]$.
   (b) Set $k = 1$.
2. **Iterate.** Until convergence or $k > k_{\max}$ do:
   (a) Solve the projected right definite two-parameter eigenvalue problem

$$\begin{aligned} U_k^T A_1 U_k c &= \sigma U_k^T B_1 U_k c + \tau U_k^T C_1 U_k c, \\ V_k^T A_2 V_k d &= \sigma V_k^T B_2 V_k d + \tau V_k^T C_2 V_k d. \end{aligned} \tag{3.1}$$

   (b) Select an appropriate Ritz value (σ, τ) and the corresponding Ritz vector $u \otimes v$, where $u = U_k c$, $v = V_k d$.
   (c) Compute the residuals

$$\begin{aligned} r_1 &= (A_1 - \sigma B_1 - \tau C_1)u, \\ r_2 &= (A_2 - \sigma B_2 - \tau C_2)v. \end{aligned} \tag{3.2}$$

   (d) Stop if $\rho_k \le \epsilon$, where

$$\rho_k = (\|r_1\|^2 + \|r_2\|^2)^{1/2}. \tag{3.3}$$

   (e) Compute new search directions s and t.
   (f) Expand the search subspaces. Set

$$\begin{aligned} U_{k+1} &= \mathrm{RGS}(U_k, s), \\ V_{k+1} &= \mathrm{RGS}(V_k, t), \end{aligned}$$

   where RGS denotes the repeated Gram–Schmidt orthonormalization.
   (g) Set $k = k + 1$.
   (h) Restart. If the dimension of U_k and V_k exceeds $l_{\max}$, then replace U_k, V_k with new orthonormal bases of dimension $l_{\min}$.

To apply this algorithm, we need to specify a tolerance ϵ, a maximum number of steps $k_{\max}$, a maximum dimension of the search subspaces $l_{\max}$, and a number $l_{\min} < l_{\max}$ that specifies the dimension of the search subspaces after a restart.

A larger search space involves a larger projected problem (2.1). The existing methods are able to solve only low-dimensional two-parameter problems in a reasonable time. Therefore, we expand search spaces up to the preselected dimension $l_{\max}$ and then restart the algorithm. For a restart, we take the most promising $l_{\min}$ eigenvector approximations as a basis for the initial search space.

Suppose that we have computed new directions s and t for the search spaces $\mathcal{U}_{k+1}$ and $\mathcal{V}_{k+1}$, respectively. We expand the search spaces simply by adding new columns to the matrices U_k and V_k. For reasons of efficiency and stability, we want orthonormal columns, and, therefore, we orthonormalize s against U_k and t against V_k by a stable form of the Gram–Schmidt orthonormalization.

The next theorem expresses that, if the residuals (3.2) are small, then the Ritz value (σ, τ) is a good approximation to an eigenvalue of (1.1). This justifies the criterion in step 2(d).

THEOREM 3.2. *If (σ, τ) is a Ritz value and r_1, r_2 are the residuals* (3.2)*, then there exists an eigenvalue (λ, μ) of the right definite two-parameter problem* (1.1) *such that*

$$(\lambda-\sigma)^2+(\mu-\tau)^2 \le \|\Delta_0^{-1}\|^2\Big[(\|B_1\|\|r_2\|+\|B_2\|\|r_1\|)^2 + (\|C_1\|\|r_2\|+\|C_2\|\|r_1\|)^2\Big]. \tag{3.4}$$

Proof. In order to prove (3.4), we consider the associated problem (1.4). First, we derive a relation between the residuals (3.2) and the residuals of the associated problem. We denote

$$\begin{aligned} p_1 &= \Delta_1(u\otimes v) - \sigma\Delta_0(u\otimes v),\\ p_2 &= \Delta_2(u\otimes v) - \tau\Delta_0(u\otimes v), \end{aligned} \tag{3.5}$$

where u, v are the normalized Ritz vectors from step 2(b). From (1.3) and (3.2), it follows that

$$\begin{aligned} p_1 &= -C_1u\otimes r_2 + r_1\otimes C_2v,\\ p_2 &= B_1u\otimes r_2 - r_1\otimes B_2v, \end{aligned}$$

and we have the bounds

$$\begin{aligned} \|p_1\| &\le \|C_1\|\|r_2\| + \|C_2\|\|r_1\|,\\ \|p_2\| &\le \|B_1\|\|r_2\| + \|B_2\|\|r_1\|. \end{aligned} \tag{3.6}$$

Now we return to the residuals (3.5). As Δ_0 is a symmetric positive definite matrix, we can transform (3.5) into

$$\begin{aligned} \Delta_0^{-1/2}p_1 &= G_1w - \sigma w,\\ \Delta_0^{-1/2}p_2 &= G_2w - \tau w, \end{aligned} \tag{3.7}$$

where $w = \Delta_0^{1/2}(u\otimes v)$ and $G_i = \Delta_0^{-1/2}\Delta_i\Delta_0^{-1/2}$ for $i = 1, 2$. The matrices G_1 and G_2 are symmetric and commute because the matrices $\Delta_0^{-1}\Delta_1$ and $\Delta_0^{-1}\Delta_2$ commute.

As a result, there exists a common orthonormal basis of eigenvectors $w_1, \ldots, w_N$ such that

$$\begin{aligned} G_1 w_i &= \lambda_i w_i, \\ G_2 w_i &= \mu_i w_i, \end{aligned} \tag{3.8}$$

where (λ_i, μ_i), $i = 1, \ldots, N$, are the eigenvalues of (1.1). In the eigenvector basis, we can decompose w as $w = \sum_{j=1}^{N} \alpha_j w_j$. From (3.7) and (3.8), we get

$$\begin{aligned} \Delta_0^{-1/2} p_1 &= \sum_{j=1}^{N} \alpha_j (\lambda_j - \sigma) w_j, \\ \Delta_0^{-1/2} p_2 &= \sum_{j=1}^{N} \alpha_j (\mu_j - \mu) w_j, \end{aligned} \tag{3.9}$$

and

$$\|\Delta_0^{-1/2} p_1\|^2 + \|\Delta_0^{-1/2} p_2\|^2 = \sum_{j=1}^{N} \alpha_j^2 ((\lambda_j - \sigma)^2 + (\mu_j - \tau)^2).$$

Since $\sum_{j=1}^{N} \alpha_j^2 \geq \|\Delta_0^{-1}\|^{-1}$, it follows that

$$\begin{aligned} \min_{j=1,\ldots,N} ((\lambda_j - \sigma)^2 + (\mu_j - \tau)^2) &\leq \|\Delta_0^{-1}\| (\|\Delta_0^{-1/2} p_1\|^2 + \|\Delta_0^{-1/2} p_2\|^2) \\ &\leq \|\Delta_0^{-1}\|^2 (\|p_1\|^2 + \|p_2\|^2). \end{aligned} \tag{3.10}$$

Finally, when we insert (3.6) into (3.10), we obtain (3.4). ∎

In the next theorem, we show that, if the Ritz vector $u \otimes v$ is close to an eigenvector $x \otimes y$ of problem (1.1), then the residuals r_1 and r_2 from (3.2) are of order $\mathcal{O}(\|u - x\|)$ and $\mathcal{O}(\|v - y\|)$, respectively. This shows that the criterion in step 2(d) will be fulfilled if the Ritz vector $u \otimes v$ approximates an eigenvector of (1.1) well enough.

THEOREM 3.3. *Let (σ, τ) be a Ritz value of* (1.1) *with the corresponding Ritz vector $u \otimes v$, where u and v are normalized. If $(u + s) \otimes (v + t)$ is an eigenvector of* (1.1) *with the corresponding eigenvalue (λ, μ), then we can bound the error of (σ, τ) as*

$$\sqrt{(\lambda - \sigma)^2 + (\mu - \tau)^2} = \mathcal{O}(\|s\|^2 + \|t\|^2) \tag{3.11}$$

and the norm of the residuals r_1, r_2 from (3.2) *as*

$$\begin{aligned} \|r_1\| &\leq \|A_1 - \lambda B_1 - \mu C_1\| \|s\| + \mathcal{O}(\|s\|^2 + \|t\|^2), \\ \|r_2\| &\leq \|A_2 - \lambda B_2 - \mu C_2\| \|t\| + \mathcal{O}(\|s\|^2 + \|t\|^2). \end{aligned} \tag{3.12}$$

Proof. We write the residuals (3.2) as

$$\begin{aligned} r_1 &= -(A_1 - \lambda B_1 - \mu C_1) s + (\lambda - \sigma) B_1 u + (\mu - \tau) C_1 u, \\ r_2 &= -(A_2 - \lambda B_2 - \mu C_2) t + (\lambda - \sigma) B_2 v + (\mu - \tau) C_2 v. \end{aligned} \tag{3.13}$$

When we multiply (3.13) by u^T and v^T, respectively, and take into account that $u^T r_1 = v^T r_2 = 0$, then we obtain

$$\begin{bmatrix} u^T B_1 u & u^T C_1 u \\ v^T B_2 v & v^T C_2 v \end{bmatrix} \begin{bmatrix} \lambda - \sigma \\ \mu - \tau \end{bmatrix} = - \begin{bmatrix} s^T (A_1 - \lambda B_1 - \mu C_1) s \\ t^T (A_2 - \lambda B_2 - \mu C_2) t \end{bmatrix}. \tag{3.14}$$

The system (3.14) is nonsingular because of right definiteness. From (3.14), it follows that

$$\left\| \begin{bmatrix} \lambda-\sigma \\ \mu-\tau \end{bmatrix} \right\| = \left\| \begin{bmatrix} u^TB_1u & u^TC_1u \\ v^TB_2v & v^TC_2v \end{bmatrix}^{-1} \begin{bmatrix} s^T(A_1-\lambda B_1-\mu C_1)s \\ t^T(A_2-\lambda B_2-\mu C_2)t \end{bmatrix} \right\| = \mathcal{O}(\|s\|^2+\|t\|^2),$$

and we get (3.11). The bound (3.12) is now a result of (3.13) and (3.11). ∎

In the following two subsections, the expansion for our Jacobi–Davidson method is discussed. We present two alternatives for the correction equations for the right definite two-parameter eigenvalue problem. Let (σ,τ) be a Ritz value that approximates the eigenvalue (λ,μ) of (1.1), and let $u\otimes v$ be its corresponding Ritz vector. Let us assume that u and v are normalized.

3.1. Correction equations with orthogonal projections. The first alternative for the correction equations is a generalization of the approach used in [20] for the one-parameter eigenvalue problem. We are searching for orthogonal improvements of the vectors u and v of the form

$$A_1(u+s) = \lambda B_1(u+s) + \mu C_1(u+s), \tag{3.15}$$

$$A_2(v+t) = \lambda B_2(v+t) + \mu C_2(v+t), \tag{3.16}$$

where $s \perp u$ and $t \perp v$.

Let

$$r_1 = (A_1-\sigma B_1-\tau C_1)u,$$
$$r_2 = (A_2-\sigma B_2-\tau C_2)v$$

be the residuals of Ritz vector $u\otimes v$ and Ritz value (σ,τ). We can rewrite (3.15) and (3.16) as

$$\begin{aligned}(A_1-\sigma B_1-\tau C_1)s = {} & -r_1 + (\lambda-\sigma)B_1u + (\mu-\tau)C_1u \\ & + (\lambda-\sigma)B_1s + (\mu-\tau)C_1s,\end{aligned} \tag{3.17}$$

$$\begin{aligned}(A_2-\sigma B_2-\tau C_2)t = {} & -r_2 + (\lambda-\sigma)B_2v + (\mu-\tau)C_2v \\ & + (\lambda-\sigma)B_2t + (\mu-\tau)C_2t.\end{aligned} \tag{3.18}$$

In this subsection, we treat (3.17) and (3.18) separately. From Theorem 3.3, it follows that $\|(\lambda-\sigma)B_1u + (\mu-\tau)C_1u\| = \mathcal{O}(\|s\|^2+\|t\|^2)$. Asymptotically (i.e., when $u\otimes v$ is close to an eigenvector of (1.1)), s and t are first order corrections and $(\lambda-\sigma)B_1u+(\mu-\tau)C_1u$ represents some second order correction. In the same sense, the term $(\lambda-\sigma)B_1s+(\mu-\tau)C_1s$ can be interpreted as a third order correction.

If we ignore second and higher order terms in (3.17), then we obtain the equation

$$(A_1-\sigma B_1-\tau C_1)s = -r_1. \tag{3.19}$$

Because r_1 and s are orthogonal to u, we can multiply (3.19) with the orthogonal projection $(I-uu^T)$ and write $(I-uu^T)s$ instead of s. Thus we obtain the correction equation for the vector u:

$$(I-uu^T)(A_1-\sigma B_1-\tau C_1)(I-uu^T)s = -r_1. \tag{3.20}$$

In a similar way, we obtain from (3.18) the correction equation for the vector v:

$$(I-vv^T)(A_2-\sigma B_2-\tau C_2)(I-vv^T)t = -r_2. \tag{3.21}$$

From (3.20) and (3.21), it is clear that the orthogonal projections preserve the symmetry of the matrices. Another advantage of orthogonal projections is that they are stable and easy to implement. The systems (3.20) and (3.21) for s and t are not of full rank, but they are consistent. We solve them only approximately with a Krylov subspace method with initial guess 0, for instance, by a few steps of MINRES. If we do just one step of MINRES, then s and t are scalar multiples of r_1 and r_2, respectively, and then, in the sense that we expand the search spaces by the residuals, we have an Arnoldi-like method, similar to the situation for the standard eigenproblem [20].

3.2. Correction equation with oblique projections. As in the correction equations with orthogonal projections, we start with (3.17) and (3.18). We neglect the third order correction terms $(\lambda-\sigma)B_1s+(\mu-\tau)C_1s$ and $(\lambda-\sigma)B_2t+(\mu-\tau)C_2t$, but, rather than neglecting the second order terms $(\lambda-\sigma)B_1u+(\mu-\tau)C_1u$ and $(\lambda-\sigma)B_2v+(\mu-\tau)C_2v$, we project them to 0 using an oblique projection.

If we define

$$M = \begin{bmatrix} A_1-\sigma B_1-\tau C_1 & 0 \\ 0 & A_2-\sigma B_2-\tau C_2 \end{bmatrix}$$

and

$$r = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix},$$

then we can reformulate (3.17) and (3.18) (without the neglected third order correction terms) as

$$M\begin{bmatrix} s \\ t \end{bmatrix} = -r + (\lambda-\sigma)\begin{bmatrix} B_1u \\ B_2v \end{bmatrix} + (\mu-\tau)\begin{bmatrix} C_1u \\ C_2v \end{bmatrix}. \tag{3.22}$$

Let $V \in \mathbb{R}^{(n_1+n_2)\times 2}$ be a matrix with columns (for reasons of stability, preferably orthonormal) such that

$$\operatorname{span}(V) = \operatorname{span}\left(\begin{bmatrix} B_1u \\ B_2v \end{bmatrix}, \begin{bmatrix} C_1u \\ C_2v \end{bmatrix}\right),$$

and let $W \in \mathbb{R}^{(n_1+n_2)\times 2}$ be

$$W = \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix}.$$

With the oblique projection

$$P = I - V(W^TV)^{-1}W^T$$

onto $\operatorname{span}(V)^\perp$ along $\operatorname{span}(W)$, it follows that

$$Pr = r \quad \text{and} \quad P\begin{bmatrix} B_1u \\ B_2v \end{bmatrix} = P\begin{bmatrix} C_1u \\ C_2v \end{bmatrix} = 0. \tag{3.23}$$

Therefore, from multiplying (3.22) by P, we obtain

$$PM\begin{bmatrix} s \\ t \end{bmatrix} = -r. \tag{3.24}$$

Furthermore, since $s \perp u$ and $t \perp v$, it follows that

$$P \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix}, \tag{3.25}$$

and the result is the correction equation

$$PMP \begin{bmatrix} s \\ t \end{bmatrix} = -r \tag{3.26}$$

for $s \perp u$ and $t \perp v$.

The correction equation (3.26) is again not of full rank but consistent, and it is often sufficient to solve it only approximately (e.g., by a few steps of GMRES). As before, if we do one step of GMRES, then s and t are scalar multiples of r_1 and r_2, respectively.

The Jacobi–Davidson method for the one-parameter problem can be viewed as an accelerated inexact Newton scheme [19]. In a similar manner, we now show that there is a connection between the Jacobi–Davidson correction equation (3.26) and the Newton method for the right definite two-parameter eigenvalue problem in [16].

Eigenpairs of the two-parameter problem (1.1) are solutions of the equation

$$G(x, y, \lambda, \mu) := \begin{bmatrix} A_1 x - \lambda B_1 x - \mu C_1 x \\ A_2 y - \lambda B_2 y - \mu C_2 y \\ \frac{1}{2}(x^T x - 1) \\ \frac{1}{2}(y^T y - 1) \end{bmatrix} = 0. \tag{3.27}$$

If we apply Newton's method to (3.27) and use u, v, σ, τ with $\|u\| = \|v\| = 1$ as an initial approximation, then, in order to obtain the improved approximation $u + s, v + t, \lambda, \tau$, we have to solve the system

(3.28)
$$\begin{bmatrix} A_1 - \sigma B_1 - \tau C_1 & 0 & -B_1 u & -C_1 u \\ 0 & A_2 - \sigma B_2 - \tau C_2 & -B_2 v & -C_2 v \\ u^T & 0 & 0 & 0 \\ 0 & v^T & 0 & 0 \end{bmatrix} \begin{bmatrix} s \\ t \\ \lambda - \sigma \\ \mu - \tau \end{bmatrix} = \begin{bmatrix} -r_1 \\ -r_2 \\ 0 \\ 0 \end{bmatrix}.$$

LEMMA 3.4. *The Jacobi–Davidson correction equation* (3.26)*, where $s \perp u$ and $t \perp v$, is equivalent to Newton's equation* (3.28)*. That is, if (s, t) is a solution of* (3.26)*, then there exist unique λ, μ such that $(s, t, \lambda - \sigma, \mu - \tau)$ is a solution of* (3.28)*, and, if $(s, t, \lambda - \sigma, \mu - \tau)$ is a solution of* (3.28)*, then (s, t) is a solution of* (3.26)*.*

Proof. We can rewrite (3.28) as

$$M \begin{bmatrix} s \\ t \end{bmatrix} = -r + (\lambda - \sigma) \begin{bmatrix} B_1 u \\ B_2 v \end{bmatrix} + (\mu - \tau) \begin{bmatrix} C_1 u \\ C_2 v \end{bmatrix},$$

and $s \perp u$, $t \perp v$, which is exactly the equation (3.22) that appears in the derivation of the Jacobi–Davidson correction equation (3.26). The proof now follows from the relations (3.23) and (3.25) and the fact that $\mathrm{Ker}(P) = \mathrm{span}(V)$. ∎

This shows that the Jacobi–Davidson type method with the correction equation (3.26) is a Newton scheme, accelerated by the projection of (1.1) onto the subspace of all previous approximations. Therefore, we expect locally at least quadratic convergence of the Jacobi–Davidson method when the correction equations are solved exactly.

4. Selection of Ritz values. In this section, we present different options for the selection of Ritz values in step 2(b) of Algorithm 3.1.

4.1. Exterior eigenvalues. First, we discuss how to obtain the eigenvalue (λ, μ) of (1.1) with the maximum value of λ. We denote such an eigenvalue by $(\lambda_{\max}, \mu_{\max})$. We show that, if we select the Ritz value (σ, τ) with the maximum value of σ in each step 2(b) of Algorithm 3.1, then the Ritz pairs will converge monotonically to an eigenpair of (1.1).

LEMMA 4.1. *Let (σ, τ) be the Ritz value for problem* (1.1) *and subspaces $\mathcal{U}, \mathcal{V}$ with the maximum value of σ. Then*

$$\sigma = \max_{\substack{u \in \mathcal{U},\ v \in \mathcal{V} \\ u, v \neq 0}} \frac{(u \otimes v)^T \Delta_1 (u \otimes v)}{(u \otimes v)^T \Delta_0 (u \otimes v)}. \tag{4.1}$$

Proof. Let the columns of U and V be orthonormal bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. It follows from (1.1), (1.4), and (2.1) that, if (σ, τ) is a Ritz pair, then σ is an eigenvalue of a symmetric definite pencil

$$(U \otimes V)^T \Delta_1 (U \otimes V) - \sigma (U \otimes V)^T \Delta_0 (U \otimes V). \tag{4.2}$$

From the minimax theorem [11, p. 411], it follows that

$$\sigma = \max_{\substack{w \in \mathcal{U} \otimes \mathcal{V} \\ w \neq 0}} \frac{w^T \Delta_1 w}{w^T \Delta_0 w}.$$

Since pencil (4.2) is related to the two-parameter problem (2.1), we can restrict w to a decomposable tensor $w = u \otimes v$, where $u \in \mathcal{U}$ and $v \in \mathcal{V}$. From this, (4.1) follows. ∎

If we select the Ritz value (σ_k, τ_k) in step 2(b) of Algorithm 3.1 with the maximum σ_k, then it follows from Lemma 4.1 that

$$\sigma_k \leq \sigma_{k+1} \leq \lambda_{\max}.$$

We cannot guarantee that the eigenvalue (λ, μ) of (1.1) to which (σ_k, τ_k) converges is equal to $(\lambda_{\max}, \mu_{\max})$, but convergence to a local optimum also may happen in the Jacobi–Davidson method for the symmetric eigenproblem and in all projection methods. Our numerical examples indicate that we usually do obtain the eigenvalue with the largest value of λ.

We can use the algorithm to obtain the eigenvalue (λ, μ) of (1.1) with the maximum value of $\lambda \cos\alpha + \mu \sin\alpha$ for a given parameter α if we apply the orthogonal linear substitution

$$\begin{aligned} \lambda &= \lambda' \cos\alpha - \mu' \sin\alpha, \\ \mu &= \lambda' \sin\alpha + \mu' \cos\alpha \end{aligned}$$

to the problem (1.1). The associated two-parameter eigenproblem with this substitution is now

$$\begin{aligned} A_1 x &= \lambda'(\cos\alpha B_1 + \sin\alpha C_1)x + \mu'(-\sin\alpha B_1 + \cos\alpha C_1)x, \\ A_2 y &= \lambda'(\cos\alpha B_2 + \sin\alpha C_2)y + \mu'(-\sin\alpha B_2 + \cos\alpha C_2)y. \end{aligned} \tag{4.3}$$

The operator determinant Δ_0 remains unchanged, and the substituted problem (4.3) is right definite as well. Using orthogonal linear substitutions, we can thus obtain exterior eigenvalues of (1.1) in chosen directions in the (λ, μ)-plane.

4.2. Interior eigenvalues. Suppose that we are interested in the eigenvalue (λ, μ) of (1.1) closest to a specific target (λ_0, μ_0). Let us denote such an eigenvalue as $(\lambda_{\text{int}}, \mu_{\text{int}})$.

Similar to the algorithm for exterior eigenvalues, we decide to select the Ritz value nearest to the target in step 2(b) of Algorithm 3.1. The convergence for interior Ritz values is not as favorable as for the exterior Ritz values. If a Ritz value (σ, τ) is close enough to $(\lambda_{\max}, \mu_{\max})$, then the Ritz vector corresponding to (σ, τ) is a good approximation to the eigenvector corresponding to $(\lambda_{\max}, \mu_{\max})$. On the contrary, if (σ, τ) is close to $(\lambda_{\text{int}}, \mu_{\text{int}})$, then the Ritz vector corresponding to (σ, τ) may be a poor approximation to the eigenvector corresponding to $(\lambda_{\text{int}}, \mu_{\text{int}})$, just as in the real symmetric eigenproblem.

Numerical examples in section 7 show that, although the convergence is very irregular, the method can still be used to compute the eigenvalue closest to the target. It turns out that, for interior eigenvalues, good approximations for new search directions which may be obtained with more GMRES steps for the correction equations are needed. The number of GMRES steps is of large influence. The more steps of GMRES we take, the better updates for the approximate eigenvectors will be added to the search spaces. If we take too many steps, then the method often converges to an eigenvalue $(\lambda, \mu) \neq (\lambda_{\text{int}}, \mu_{\text{int}})$. On the other hand, if we take too few GMRES steps, then we need many outer iterations or we have no convergence at all.

If we are interested in interior eigenvalues of a symmetric eigenproblem $Ax = \lambda x$, then one of the possible tools are harmonic Ritz values. The question remains how to generalize harmonic Ritz values to a right definite two-parameter eigenvalue problem. We believe that any progress on this subject might lead to better methods for interior eigenvalues.

Remark 4.2. It is easy to see that step 2(b) of Algorithm 3.1 can be modified in a similar manner if we are interested in the eigenvalue (λ, μ) of (1.1) with the maximum value of $\lambda^2 + \mu^2$.

5. Computing more eigenpairs. Suppose that we are interested in $p > 1$ eigenpairs of (1.1). In a one-parameter problem, various deflation techniques can be applied in order to compute more than one eigenpair. In this section, we first show difficulties that are met when we try to translate standard deflation ideas from one-parameter problems to two-parameter problems. We then propose a selection method for Ritz vectors that makes it possible to obtain more than one eigenpair for two-parameter problems.

If (ξ, z) is an eigenpair of a symmetric matrix A, then all other eigenpairs can be computed from the projection of A onto the subspace $z^{\perp}$. Similarly, if (λ, μ) is an eigenvalue of (1.1) and $x \otimes y$ is the corresponding eigenvector, then all other eigenvectors lie in the subspace

$$(x \otimes y)^{\perp_{\Delta_0}} := \{z \in S : z^T \Delta_0 (x \otimes y) = 0\}$$

of the dimension $n_1 n_2 - 1$. By comparing the dimensions, it is clear that the subspace $(x \otimes y)^{\perp_{\Delta_0}}$ cannot be written as $\mathcal{U} \otimes \mathcal{V}$, where $\mathcal{U} \subset \mathbb{R}^{n_1}$ and $\mathcal{V} \subset \mathbb{R}^{n_2}$. Therefore, this kind of deflation cannot be applied to Algorithm 3.1.

Another way of deflation of a symmetric matrix A is to shift the eigenvalue to an unwanted part of the spectrum using the matrix $A' = A - (\xi - \widetilde{\xi}) z z^T$. Matrix A' has the same eigenvalues as matrix A except for ξ, which is transformed into $\widetilde{\xi}$. A generalization of this approach would be to transform the two-parameter problem (1.1) into a two-parameter problem with the same eigenvalues as (1.1) except for the

eigenvalue (λ, μ), which should be transformed into $(\widetilde{\lambda}, \widetilde{\mu})$. Since, in a two-parameter problem, there can exist eigenvalues (λ, μ) and (λ', μ') with eigenvectors $x \otimes y$ and $x' \otimes y'$, respectively, such that $(\lambda, \mu) \neq (\lambda', \mu')$ and $x = x'$, this approach would again work only if we apply the associated problem (1.4) in the tensor product space S. However, then we have to work with large Δ_i matrices, and this is too expensive.

We propose the following approach. Suppose that we have already found p eigenvalues (λ_i, μ_i) and eigenvectors $x_i \otimes y_i$, $i = 1, \ldots, p$. Based on the fact that eigenvectors are Δ_0-orthogonal (see (1.5)), we adjust Algorithm 3.1 so that, in step 2(b), we consider only those Ritz vectors $u \otimes v$ which satisfy

$$|(u \otimes v)^T \Delta_0 (x_i \otimes y_i)| < \eta \text{ for } i = 1, \ldots, p, \tag{5.1}$$

for an $\eta > 0$. Suppose that we are interested in eigenvalues with the maximum values of λ. Then, in step 2(b), we first order Ritz pairs $(\sigma_i, \tau_i), u_i \otimes v_i$ by their σ values so that $\sigma_i \geq \sigma_j$ for $i < j$, and then we select the Ritz pair that satisfies (5.1) and has the minimal index. In the case of interior eigenvalues, a different ordering is used.

If none of the Ritz pairs meet (5.1), then we take the Ritz pair with index 1, but, in this case, the algorithm is not allowed to stop. This is achieved by a change of the stopping criterion in step 2(d), where, in addition to a small residual norm (3.3), we now also require that the Ritz vector $u \otimes v$ satisfies (5.1). This guarantees that the method does not converge to the already computed eigenpairs.

The bound η should not be taken too small in order to avoid that none of the Ritz vectors are sufficiently Δ_0-orthogonal to the set of already computed eigenvectors. In numerical experiments in section 7, we use

$$\eta = \frac{1}{2} \min_{i=1,\ldots,p} |(x_i \otimes y_i)^T \Delta_0 (x_i \otimes y_i)|,$$

and that value successfully prevents the method from converging to the already computed eigenpairs.

All other steps of Algorithm 3.1 remain unchanged. Numerical results in section 7 show that this approach enables us to compute more than one eigenpair.

6. Time complexity. We examine the time complexity of one outer iteration step of Algorithm 3.1. Let $n = n_1 = n_2$, let k be the dimension of the search spaces, and let m be the number of GMRES (MINRES) steps for a correction equation. The two steps that largely determine the time complexity are steps 2(a) and 2(e). In step 2(a), we first construct the smaller projected problem (3.1). We need to compute only the last row (and column) of matrices in (3.1). In the second part of step 2(a), we solve (3.1) by solving its associated problem with matrices of size k^2, and thus we need $\mathcal{O}(k^6)$ [9].

First, we assume that matrices A_i, B_i, and C_i are sparse. This is true in many applications—for instance, when two-parameter Sturm–Liouville problems [10] are discretized. Because MINRES and GMRES are methods intended for sparse matrices, the Jacobi–Davidson type method can in principle handle very large sparse problems. For such problems, the time complexities of steps 2(a) and 2(e) can be expressed as 6 MV + $\mathcal{O}(k^6)$ and $6m$ MV, respectively, where MV stands for a matrix-vector multiplication with an $n \times n$ matrix.

The analysis for dense matrices A_i, B_i, and C_i is as follows. In step 2(a), we need $\mathcal{O}(n^2)$ for the construction of the smaller problem (3.1) and additional $\mathcal{O}(k^6)$ for the solution of (3.1). As, in practice, only very small values of k are used, we can assume that $k = \mathcal{O}(n^{1/3})$, and thus the time complexity of step 2(a) is $\mathcal{O}(n^2)$. If

we use correction equations (3.20), (3.21) with orthogonal projections and perform m steps of MINRES, then the time complexity of step 2(e) is $\mathcal{O}(mn^2)$ when we perform m matrix-vector multiplications. We obtain the same time complexity for step 2(e) when we use the correction equation (3.26) with oblique projections and do m steps of GMRES. The only difference is that we are working with one matrix of size $2n$, while we are working with two matrices of size n if we use orthogonal projections.

Based on the above assumptions, the time complexity of one outer step of Algorithm 3.1 for dense matrices is $\mathcal{O}(mn^2)$. Also important is the storage requirement. If an algorithm works with matrices A_i, B_i, and C_i as Algorithm 3.1 does, then it requires $\mathcal{O}(n^2)$ memory. The methods that work with the associated system (1.4) need $\mathcal{O}(n^4)$ memory, which may exceed memory rapidly, even for modest values of n.

7. Numerical examples. We present some numerical examples obtained with Matlab 5.3. If the dimension of the matrices is $n = n_1 = n_2 = 100$, then none of the existing methods that work in the tensor product space are able to compute all eigenpairs in a reasonable time [16]. Therefore, we construct right definite two-parameter examples where the exact eigenpairs are known, which enables us to check the obtained results.

We construct our right definite two-parameter examples in the following way. We take matrices

$$A_i = Q_i F_i Q_i^T, \quad B_i = Q_i G_i Q_i^T, \quad C_i = Q_i H_i Q_i^T, \tag{7.1}$$

where F_i, G_i, and H_i are diagonal matrices and Q_i is a random orthogonal matrix for $i = 1, 2$. We select diagonal elements of matrices F_1, F_2, G_2, and H_1 as uniformly distributed random numbers from the interval $(0, 1)$ and diagonal elements of G_1 and H_2 as uniformly distributed random numbers from the interval $(1, 2)$. The determinant (1.2) is clearly strictly positive for nonzero x, y, and the obtained two-parameter problem is right definite. All matrices are of dimension $n \times n$.

Let us denote $F_i = \operatorname{diag}(f_{i1}, \ldots, f_{in})$, $G_i = \operatorname{diag}(g_{i1}, \ldots, g_{in})$, and $H_i = \operatorname{diag}(h_{i1}, \ldots, h_{in})$. It is easy to see that eigenvalues of the two-parameter problem (1.1) are solutions of linear systems

$$\begin{aligned} f_{1i} &= \lambda g_{1i} + \mu h_{1i}, \\ f_{2j} &= \lambda g_{2j} + \mu h_{2j} \end{aligned}$$

for $i, j = 1, \ldots, n$. This enables us to compute all of the eigenvalues from the diagonal elements of F_i, G_i, H_i for $i = 1, 2$. In order to construct a two-parameter problem that has the point $(0, 0)$ in the interior of the convex hull of all the eigenvalues, we take the shifted problem

$$\begin{aligned} (A_1 - \lambda_0 B_1 - \mu_0 C_1)x &= (\lambda - \lambda_0)B_1 x + (\mu - \mu_0)C_1 x, \\ (A_2 - \lambda_0 B_2 - \mu_0 C_2)y &= (\lambda - \lambda_0)B_2 y + (\mu - \mu_0)C_2 y, \end{aligned}$$

where the shift (λ_0, μ_0) is the arithmetic mean of all of the eigenvalues. Figure 1 shows the distribution of eigenvalues obtained for $n = 100$.

For the following numerical examples, we use GMRES instead of MINRES in the correction equation with orthogonal projections because MINRES is not standardly available in Matlab 5.3.

Example 7.1. In the first example, we use the Jacobi–Davidson type method for the exterior eigenvalues. Our goal is to compute the eigenvalue $(\lambda_{\max}, \mu_{\max})$ with

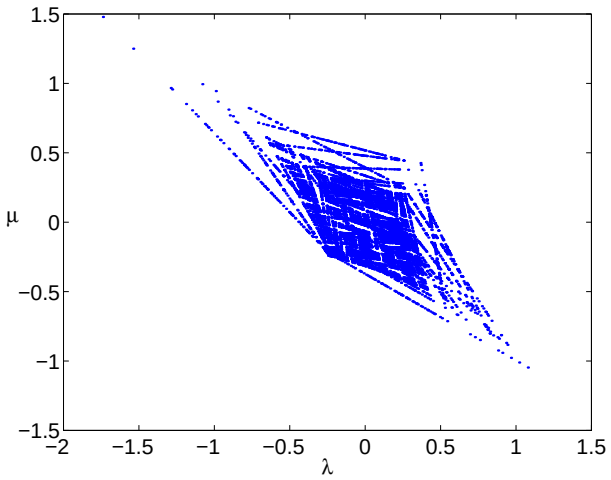


FIG. 1. *Distribution of eigenvalues for a right definite two-parameter problem of size* $n = 100$.

TABLE 1

Statistics of the Jacobi–Davidson type method for the eigenvalue $(\lambda_{\max}, \mu_{\max})$ *using different correction equations and number of GMRES steps for right definite two-parameter problems of size* $n = 100$ *and* $n = 200$: *average number of outer iterations, percentage of convergence to* $(\lambda_{\max}, \mu_{\max})$, *and average number of flops over* 250 *trials with different random initial vectors. Correction equations: JO(m) stands for orthogonal projections and m steps of GMRES; JS(m) stands for oblique projections and m steps of GMRES.*

Correction equation	$n = 100$			$n = 200$		
	Iterations	Percentage	Flops	Iterations	Percentage	Flops
JO(1)=JS(1)	105.4	100.0 %	$4.6 \cdot 10^8$	68.9	100.0 %	$3.4 \cdot 10^8$
JO(2)	50.0	100.0 %	$2.2 \cdot 10^8$	35.6	100.0 %	$2.0 \cdot 10^8$
JO(4)	26.7	100.0 %	$1.1 \cdot 10^8$	25.7	100.0 %	$1.6 \cdot 10^8$
JO(8)	23.3	99.2 %	$1.1 \cdot 10^8$	27.7	99.2 %	$2.1 \cdot 10^8$
JO(16)	25.4	30.0 %	$1.4 \cdot 10^8$	34.0	48.4 %	$3.6 \cdot 10^8$
JO(32)	29.8	38.0 %	$2.2 \cdot 10^8$	42.8	10.4 %	$7.2 \cdot 10^8$
JO(64)	33.1	28.0 %	$4.0 \cdot 10^8$	51.6	9.6 %	$16.0 \cdot 10^8$
JS(2)	96.4	100.0 %	$4.6 \cdot 10^8$	94.4	100.0 %	$6.1 \cdot 10^8$
JS(4)	99.9	100.0 %	$5.0 \cdot 10^8$	92.9	100.0 %	$6.6 \cdot 10^8$
JS(8)	63.9	100.0 %	$3.3 \cdot 10^8$	62.4	100.0 %	$5.2 \cdot 10^8$
JS(16)	45.2	94.0 %	$2.6 \cdot 10^8$	53.5	98.4 %	$6.0 \cdot 10^8$
JS(32)	41.9	82.4 %	$3.2 \cdot 10^8$	55.4	70.8 %	$9.6 \cdot 10^8$
JS(64)	39.7	66.0 %	$4.9 \cdot 10^8$	56.0	35.6 %	$17.6 \cdot 10^8$

the maximum value of λ. We are interested in the number of iterations that the Jacobi–Davidson method needs for sufficiently accurate approximations and also in the percentage of the convergence to the eigenvalue $(\lambda_{\max}, \mu_{\max})$ for a test set of 250 different initial vectors.

We test both alternatives for the correction equations using various numbers of GMRES steps. Each combination is tested on the same set of 250 random initial vectors. The algorithm is restarted after every 10 iterations with the current eigenvector approximation, so $l_{\max} = 10$ and $l_{\min} = 1$. The value $\epsilon = 10^{-8}$ is used for the test of convergence, and flop counts in Matlab are used for a measure of time complexity.

Table 1 contains results obtained for $n = 100$ and $n = 200$. JO(m) and JS(m) denote that m steps of GMRES are used for the correction equation with orthogonal

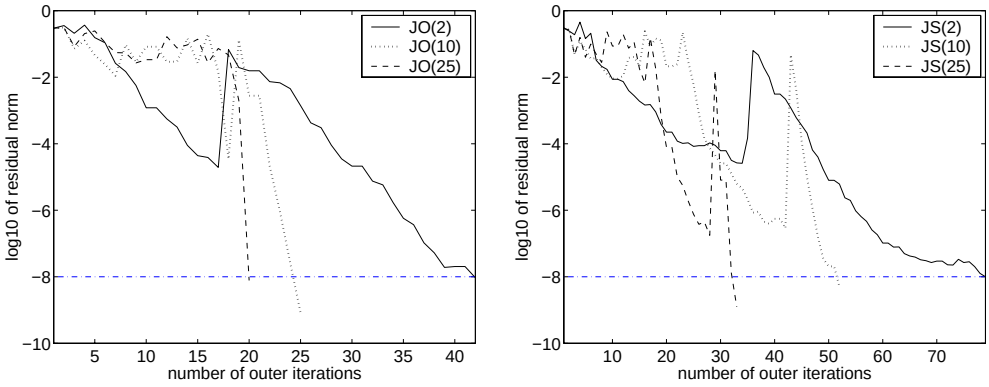


FIG. 2. *Convergence plot for the exterior eigenvalue* $(\lambda_{\max}, \mu_{\max})$ *for* $n = 100$ *and* $u = v = [1 \ \cdots \ 1]^T$. *The plots show the* $\log_{10}$ *of the residual norm* ρ_k (3.3) *versus the outer iteration number* k *for the Jacobi–Davidson type method for the eigenvalue* $(\lambda_{\max}, \mu_{\max})$ *using* 2 *(solid line),* 10 *(dotted line), and* 25 *(dashed line) GMRES steps to solve the correction equation with orthogonal projections (left plot) and oblique projections (right plot), respectively.*

projections or with oblique projections, respectively. For each combination, we list the average number of outer iterations for convergence, the percentage of eigenvalues that converged to the eigenvalue $(\lambda_{\max}, \mu_{\max})$, and the average number of flops in Matlab, all obtained on the same set of 250 different initial vectors.

The results in Table 1 indicate that the method is likely to converge to an unwanted eigenvalue if we solve the correction equation too accurately, i.e., if too many GMRES steps are used to solve the correction equation. A comparison of the flops suggests that the best approach is to do a few steps of GMRES. We also see that, for larger n, the number of GMRES steps has more impact on the time complexity than the number of outer iterations. The reason is that, for larger n, the factor k^6 becomes relatively smaller compared to mn^2.

The correction equations with orthogonal projections behave similarly to the one with oblique projections but require fewer operations. The experiments suggest using the correction equations with orthogonal projections in combination with a small number of GMRES steps in each outer iteration for $(\lambda_{\max}, \mu_{\max})$.

Example 7.2. In the second example, the convergence to the exterior eigenvalue for the two-parameter problem of dimension $n = 100$ and initial vectors $u = v = [1 \ \cdots \ 1]^T$ is examined. We compare the convergence for 2, 10, and 25 GMRES steps per iteration for the correction equation with orthogonal and the one with oblique projections, respectively. Figure 2 shows the $\log_{10}$ plot of residual norm ρ_k (3.3) versus the outer iteration number k. In all six cases, the Ritz values converge to the eigenvalue $(\lambda_{\max}, \mu_{\max})$.

It is clear from Figure 2 that convergence near the solution is faster if more GMRES steps are used. Experiments indicate that, if only a few steps of GMRES are applied, then the convergence near the solution is about linear; this is similar to the Jacobi–Davidson method for the standard eigenvalue problem [20, p. 419].

Example 7.3. In this example, we examine the convergence of the Jacobi–Davidson type method for the interior eigenvalues. We look for the eigenvalue closest to $(0, 0)$. We use the same $n = 100$ two-parameter problem as in Example 7.1 and again test both correction equations with a different number of GMRES steps on a set of 250 different initial vectors. The algorithm is restarted after every 10 itera-

TABLE 2

Statistics of the Jacobi–Davidson type method for the eigenvalue closest to $(0,0)$ *using different correction equations and different inner iteration processes for a right definite two-parameter problem of size* $n = 100$*: average number of iterations, percentage of convergence to the eigenvalue closest to* $(0,0)$*, and average number of flops over* 250 *trials with different random initial vectors. Correction equations: JO(m) stands for orthogonal projections and m steps of GMRES; JS(m) stands for oblique projections and m steps of GMRES.*

Correction equation	Iterations	Percentage	Flops
JO(90)	15.2	80.8 %	$2.4 \cdot 10^8$
JO(80)	15.9	89.2 %	$2.2 \cdot 10^8$
JO(70)	18.9	90.0 %	$2.4 \cdot 10^8$
JO(60)	23.3	91.2 %	$2.5 \cdot 10^8$
JO(50)	32.8	79.6 %	$3.2 \cdot 10^8$
JO(40)	41.4	81.6 %	$3.5 \cdot 10^8$
JO(30)	76.5	72.8 %	$5.8 \cdot 10^8$
JO(20)	219.2	63.2 %	$14.4 \cdot 10^8$
JS(90)	20.2	92.4 %	$4.7 \cdot 10^8$
JS(80)	21.1	96.4 %	$4.3 \cdot 10^8$
JS(70)	24.2	95.6 %	$4.4 \cdot 10^8$
JS(60)	29.0	94.4 %	$4.7 \cdot 10^8$
JS(50)	38.1	93.2 %	$5.4 \cdot 10^8$
JS(40)	47.0	93.2 %	$5.7 \cdot 10^8$
JS(30)	82.9	94.0 %	$8.5 \cdot 10^8$
JS(20)	239.7	84.0 %	$20.5 \cdot 10^8$

tions with the current eigenvector approximation. For the convergence test, we take $\epsilon = 10^{-6}$. The reason for a more relaxed criterion is an irregular convergence of the interior eigenvalues (see the peaks in Figure 3).

The results, presented in Table 2, show that the method may also be used effectively for interior eigenvalues. In contrast to Example 7.1, more GMRES steps are required for one outer iteration step. If too many steps are applied, then the process converges to an unwanted eigenvalue, similar to Example 7.1. On the other hand, if we do not take enough GMRES steps, then we need many outer iteration steps, and the results may be worse. This is different from Example 7.1, where the process converges in reasonable time even if only one GMRES step is applied per Jacobi–Davidson iteration step. The correction equation with oblique projections is more effective than the one with orthogonal projections. It is more expensive, but the probability of coming close to the eigenvalue closest to $(0,0)$ is higher.

Example 7.4. We examine the convergence to the eigenvalue closest to $(0,0)$ for the two-parameter problem of size $n = 100$ and initial vectors $u = v = [1 \ \cdots \ 1]^T$. Figure 3 shows the $\log_{10}$ plot of residual norms ρ_k (3.3) versus the outer iteration number k. We compare 40, 60, and 80 GMRES steps for the correction equation with orthogonal and with oblique projections, respectively. In all six cases, the Ritz values converge to the eigenvalue closest to $(0,0)$. We observe that the more GMRES steps are taken, the fewer iteration steps are needed. The convergence is not as smooth as in Figure 2 for Example 7.2, but the algorithm is clearly useful for interior eigenvalues.

Example 7.5. In the last example, we test the selection technique from section 5 for computing more eigenpairs for the two-parameter problem of dimension $n = 100$. With 5 GMRES steps for the correction equation with orthogonal projections, we try to compute 30 successive eigenvalues with the maximum value of λ. Figure 4 shows how well the first 15 and all 30 computed eigenvalues agree with the desired eigenvalues, respectively.

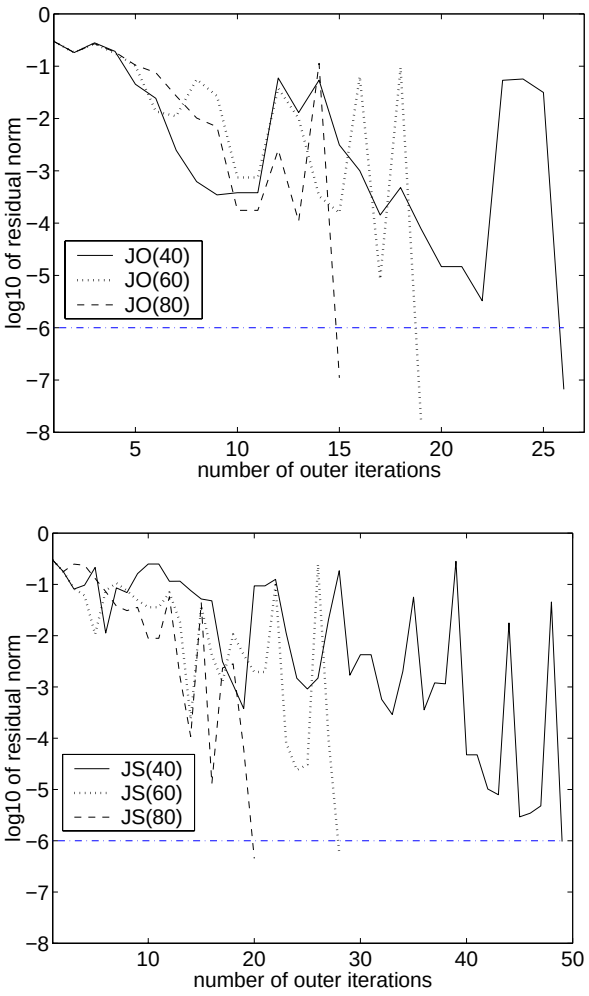


FIG. 3. *Convergence plot for the eigenvalue closest to* $(0,0)$ *for* $n = 100$ *and* $u = v = [1 \ \cdots \ 1]^T$. *The plots show the* $\log_{10}$ *of the residual norm* ρ_k (3.3) *versus the outer iteration number* k *for the Jacobi–Davidson type method for the eigenvalue closest to* $(0,0)$ *using* 40 *(solid line),* 60 *(dotted line), and* 80 *(dashed line) GMRES steps to solve the correction equation with orthogonal projections (left plot) and oblique projections (right plot), respectively.*

The eigenvalues are not necessarily computed in the same order as their λ values. This explains the situation in Figure 4, where some eigenvalues that are in the top 30 by their λ values are not among the 30 computed eigenvalues. In order to obtain the top k eigenvalues with high probability, it is therefore advisable to always compute more than k eigenvalues.

8. Conclusions. We have presented a new Jacobi–Davidson type method for a right definite two-parameter eigenvalue problem. It has several advantages over the existing methods. It can compute selected eigenpairs, and it does not require good initial approximations. Probably the most important advantage is that it can tackle very large two-parameter problems, especially if the matrices A_i, B_i, and C_i are sparse.

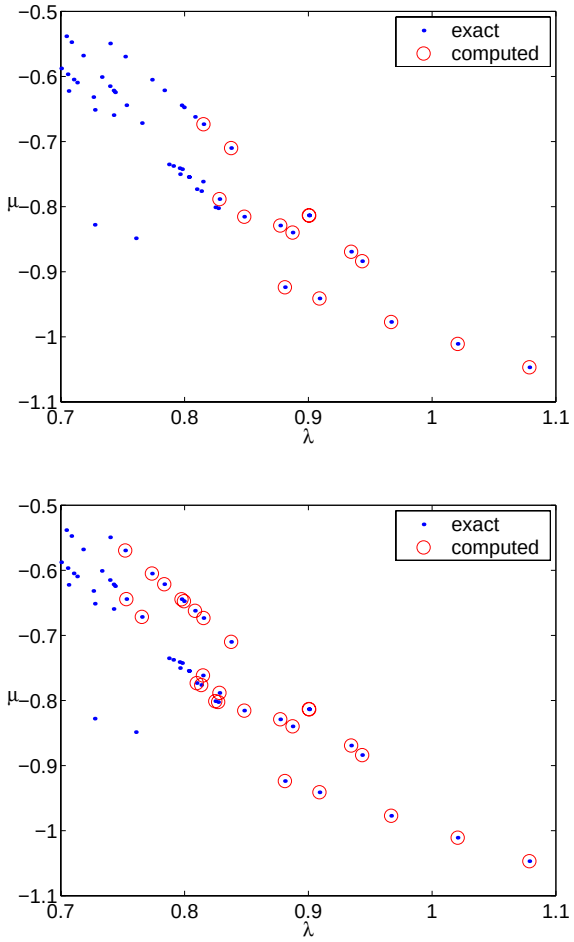


FIG. 4. *First* 15 *(left plot) and first* 30 *(right plot) computed eigenvalues with maximum value of* λ *for a two-parameter problem of size* $n = 100$ *computed using selection for Ritz vectors. The Jacobi–Davidson type method used* 5 *GMRES steps for the correction equation with orthogonal projections.*

We have proposed two correction equations. On one hand, orthogonal projections are generally more stable than oblique projections, and, in addition, orthogonal projections preserve symmetry. On the other hand, the correction equation with oblique projections can be viewed as an inexact Newton scheme which guarantees asymptotically quadratic convergence. Numerical results indicate that the correction equation with oblique projections is more reliable but more expensive. It is therefore more suitable for the interior eigenvalues, while the one with orthogonal projections may be used for the exterior eigenvalues.

Numerical results indicate that the probability of misconvergence is low when parameters are optimal. The number of GMRES steps is important. Experiments suggest taking up to 5 GMRES steps for exterior eigenvalues and more GMRES steps for interior eigenvalues. Restarts also impact the behavior of the method. In our experiments, we restart the method after every 10 iterations with the current

eigenvector approximations, but a different setting may further improve the method.

Because standard deflation techniques for a one-parameter problem cannot be applied to two-parameter problems, we came up with a new selection technique for Ritz vectors.

Acknowledgments. The authors are grateful to Gerard Sleijpen and Henk van der Vorst for suggestions that improved the paper.

REFERENCES

[1] F. V. Atkinson, *Multiparameter spectral theory*, Bull. Amer. Math. Soc., 74 (1968), pp. 1–27.
[2] F. V. Atkinson, *Multiparameter Eigenvalue Problems*, Academic Press, New York, 1972.
[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.
[4] E. K. Blum and A. F. Chang, *A numerical method for the solution of the double eigenvalue problem*, J. Inst. Math. Appl., 22 (1978), pp. 29–42.
[5] E. K. Blum and A. R. Curtis, *A convergent gradient method for matrix eigenvector-eigentuple problems*, Numer. Math., 31 (1978), pp. 247–263.
[6] E. K. Blum and P. B. Geltner, *Numerical solution of eigentuple-eigenvector problems in Hilbert spaces by a gradient method*, Numer. Math., 31 (1978), pp. 231–246.
[7] Z. Bohte, *Numerical solution of some two-parameter eigenvalue problems*, in Anton Kuhelj Memorial Volume, Slovenian Academy of Science and Art, Ljubljana, Slovenia, 1982, pp. 17–28.
[8] P. J. Browne and B. D. Sleeman, *A numerical technique for multiparameter eigenvalue problems*, IMA J. Numer. Anal., 2 (1982), pp. 451–457.
[9] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *Numerical methods for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 927–949.
[10] M. Faierman, *Two-Parameter Eigenvalue Problems in Ordinary Differential Equations*, Pitman Res. Notes in Math. 205, Longman Scientific and Technical, Harlow, UK, 1991.
[11] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.
[12] M. E. Hochstenbach, *A Jacobi–Davidson type SVD method*, SIAM J. Sci. Comput., 23 (2001), pp. 606–628.
[13] M. E. Hochstenbach and G. L. G. Sleijpen, *Two-Sided and Alternating Jacobi–Davidson*, Preprint 1196, Department of Mathematics, Utrecht University, Utrecht, The Netherlands, 2001. Linear Algebra Appl., to appear.
[14] X. Ji, *Numerical solution of joint eigenpairs of a family of commutative matrices*, Appl. Math. Lett., 4 (1991), pp. 57–60.
[15] B. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Appl. Math. 20, SIAM, Philadelphia, 1997.
[16] B. Plestenjak, *A continuation method for a right definite two-parameter eigenvalue problem*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1163–1184.
[17] M. Shimasaki, *Homotopy algorithm for two-parameter eigenvalue problems*, Z. Angew. Math. Mech., 76 (1996), pp. 675–676.
[18] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. van der Vorst, *Jacobi–Davidson type method for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.
[19] G. L. G. Sleijpen and H. A. van der Vorst, *The Jacobi–Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes*, in Iterative Methods in Linear Algebra II, S. D. Margenov and P. S. Vassilevski, eds., IMACS Series in Computational and Applied Mathematics 3, IMACS, New Brunswick, NJ, 1996, pp. 377–389.
[20] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[21] T. Slivnik and G. Tomšič, *A numerical method for the solution of two-parameter eigenvalue problems*, J. Comput. Appl. Math., 15 (1986), pp. 109–115.
[22] H. Volkmer, *Multiparameter Problems and Expansion Theorems*, Lecture Notes in Math. 1356, Springer-Verlag, New York, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 411–438

PERTURBATION ANALYSIS OF THE PERIODIC DISCRETE-TIME ALGEBRAIC RICCATI EQUATION*

WEN-WEI LIN† AND JI-GUANG SUN‡

Abstract. This paper is devoted to the perturbation analysis for the periodic discrete-time algebraic Riccati equations (P-DAREs). Perturbation bounds and condition numbers of the Hermitian positive semidefinite solution set to the P-DAREs are obtained. The results are illustrated by numerical examples.

Key words. periodic Riccati equation, periodic Hermitian positive semidefinite solution set, perturbation bound, condition number

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

PII. S0895479801384391

1. Introduction.

We consider the periodic discrete-time algebraic Riccati equation (P-DARE) with period $p \geq 1$,

$$\begin{aligned} X_{j-1} &= A_j^H X_j A_j - A_j^H X_j B_j (R_j + B_j^H X_j B_j)^{-1} B_j^H X_j A_j + H_j \\ &= A_j^H X_j (I + G_j X_j)^{-1} A_j + H_j, \end{aligned} \tag{1.1}$$

where, for all j, $A_j = A_{j+p}, H_j = H_{j+p}$, and $X_j = X_{j+p}$ are $n \times n$ matrices, $B_j = B_{j+p}$ are $n \times m$ matrices, and $R_j = R_{j+p}$ are $m \times m$ matrices; B_j is of full column rank, R_j is Hermitian positive definite $(R_j > 0), G_j \equiv B_j R_j^{-1} B_j^H = G_{j+p}$, and H_j is Hermitian positive semidefinite (p.s.d.) with $H_j = C_j^H C_j \geq 0$, a full rank decomposition (f.r.d.). Note that the second equation of (1.1) is obtained by the Sherman–Morrison–Woodbury formula (see, e.g., [9, p. 50]) provided that $(I + G_j X_j)^{-1}$ exists. In this paper, the indices j for all periodic coefficient matrices are chosen in $\{1, \ldots, p\}$ modulo p without ambiguity.

Appropriate assumptions on the periodic coefficient matrices will be made in the following to guarantee the existence and uniqueness of the Hermitian p.s.d. solution set $\{X_j\}_{j=1}^p$ to the P-DARE (1.1). The equation (1.1) arises frequently in solving the periodic discrete-time linear optimal control problem

$$\begin{aligned} &\text{Minimize} \quad \mathcal{J} = \tfrac{1}{2} \sum_{j=1}^{\infty} [x_j^H H_j x_j + u_j^H R_j u_j], \\ &\text{subject to} \quad x_{j+1} = A_j x_j + B_j u_j. \end{aligned} \tag{1.2}$$

The periodic optimal feedback vector $u_j^\star$ for (1.2) is given by [2]

$$u_j^\star = -(R_j + B_j^H X_j B_j)^{-1} B_j^H X_j A_j x_j \tag{1.3}$$

for $j = 1, \ldots, p$, where $\{X_j\}_{j=1}^p$ is the Hermitian p.s.d. solution set to (1.1). The real case, i.e., to find the real symmetric p.s.d. solution set $\{X_j\}_{j=1}^p$ to the P-DARE (1.1)

*Received by the editors February 1, 2001; accepted for publication (in revised form) by V. Mehrmann March 15, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/simax/24-2/38439.html

†Department of Mathematics, National Tsinghua University, Hsinchu, 300, Taiwan (wwlin@am.nthu.edu.tw). The research of this author was supported by the National Science Council.

‡Department of Computing Science, Umeå University, S-90187, Umeå, Sweden (jisun@cs.umu.se). The research of this author was supported by the Faculty of Science and Technology and the Department of Computing Science, Umeå University.

when all of the periodic coefficient matrices are real, is essentially important in many applications. We consider here the real case as well as the general, that is, complex, case.

The P-DARE can be regarded as an extension of the time-invariant case. For $p = 1$, the P-DARE becomes the usual discrete-time algebraic Riccati equation (DARE) by setting $X_j = X_{j-1}$ in (1.1). There are many contributions in the literature on the perturbation theory and numerical methods of the DARE (see, e.g., [13], [20], [21], [11], [22], [17]). In the case of $p > 1$, many research efforts have been devoted to the existence of different types of solution sets to the P-DARE under variant assumptions [1], [2], [5], [7], [12], [15], [18], [23]. In this paper, we study the perturbation theory for the P-DARE. This work, as a generalization of the results given by [20] and [21], derives perturbation bounds and condition numbers of the Hermitian p.s.d. solution set $\{X_j\}_{j=1}^p$ to the P-DARE (1.1). The interest in this topic is motivated by the fact that the P-DARE is usually subject to perturbation in the coefficient matrices, reflecting various errors in the formulation of the problem and in its solution by a computer. (See, e.g., [3], [6] for numerical methods for solving the P-DARE.)

Throughout this paper, we denote by $\mathcal{H}_n(S_n)$ and $\mathbb{C}_n(\mathbb{R}_n)$ the sets of $n \times n$ Hermitian (real symmetric) and $n \times n$ complex (real) matrices, respectively, and we denote by $\mathcal{H}_n^p$ and $\mathbb{C}_n^p$ the p-tuple product spaces $\mathcal{H}_n \times \cdots \times \mathcal{H}_n$ and $\mathbb{C}_n \times \cdots \times \mathbb{C}_n$, respectively. $\overline{A}$ denotes the conjugate of a matrix A. $A^\top$ denotes the transpose of A, and $A^H = \overline{A}^\top$. I stands for the identity matrix, I_n is the identity matrix of order n, and 0 is the null matrix. The set of all eigenvalues of $A \in \mathbb{C}_n$ is denoted by $\lambda(A)$. The spectral radius $\rho(A)$ is defined by $\rho(A) = \max\{|\lambda_i| : \lambda_i \in \lambda(A)\}$. The symbol $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_2$ is the spectral norm and the Euclidean vector norm. For $A = (a_1, \ldots, a_n) = (a_{ij}) \in \mathbb{C}_n$ and a matrix B, $A \otimes B = (a_{ij}B)$ is a Kronecker product, and $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A) = (a_1^\top, \ldots, a_n^\top)^\top$. An $n \times n$ matrix Φ is said to be d-stable if $\lambda(\Phi) \subset \mathcal{D}$, where $\mathcal{D} \equiv \{z \in \mathbb{C} : |z| < 1\}$. In order to save the space of the matrix representation, we also use the following notation:

$$\mathrm{diag}\{N_j\}_{j=1}^p = \begin{bmatrix} N_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & N_p \end{bmatrix}, \quad \mathrm{cyc}\{N_j\}_{j=1}^p = \begin{bmatrix} 0 & \cdots & 0 & N_1 \\ N_2 & \ddots & & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & N_p & 0 \end{bmatrix}.$$

DEFINITION 1.1. *Let $\Phi_1, \ldots, \Phi_p \in \mathbb{C}_n$. If there are complex numbers $\alpha_1, \ldots, \alpha_p$ such that*

$$\det\left[\mathrm{diag}\{\alpha_j \mathrm{I}\}_{j=1}^p - \mathrm{cyc}\{\Phi_j\}_{j=1}^p\right] = 0,$$

then $\alpha_1\alpha_2\cdots\alpha_p$ is an eigenvalue of the periodic matrix set $\{\Phi_j\}_{j=1}^p$.

The set of all eigenvalues of $\{\Phi_j\}_{j=1}^p$ is denoted by $\lambda(\{\Phi_j\}_{j=1}^p)$. Note that it is easily seen that $\lambda(\{\Phi_j\}_{j=1}^p) = \lambda(\Phi_p\Phi_{p-1}\cdots\Phi_1)$, and so $\rho\left(\{\Phi_j\}_{j=1}^p\right) = \rho\left(\Phi_p\Phi_{p-1}\cdots\Phi_1\right)$.

DEFINITION 1.2. *Let $\Phi_1, \ldots, \Phi_p \in \mathbb{C}_n$. The periodic matrix set $\{\Phi_j\}_{j=1}^p$ is said to be pd-stable if the matrix $\Phi_p\Phi_{p-1}\cdots\Phi_1$ is d-stable, i.e., $\lambda(\Phi_p\Phi_{p-1}\cdots\Phi_1) \subset \mathcal{D}$.*

From Definition 1.2, we see that, if $\{\Phi_j\}_{j=1}^p$ is pd-stable, then $\lambda(\Phi_{p-1}\cdots\Phi_1\Phi_p), \ldots, \lambda(\Phi_1\Phi_p\cdots\Phi_2) \subset \mathcal{D}$.

DEFINITION 1.3 (see [2]). *The periodic matrix pair sets $\{(A_j, B_j)\}_{j=1}^p$ and $\{(A_j, C_j)\}_{j=1}^p$ are said to be pd-stabilizable and pd-detectable, respectively, if the pairs $(\mathcal{A}_j, \mathcal{B}_j)$ and $(\mathcal{A}_j, \mathcal{C}_j)$ are d-stabilizable and d-detectable, respectively, for $j = 1, \ldots p$,*

where

$$\begin{aligned}
\mathcal{A}_j &= A_{\pi_j(p)} \cdots A_{\pi_j(1)},\\
\mathcal{B}_j &= [A_{\pi_j(p)} \cdots A_{\pi_j(2)} B_{\pi_j(1)} | A_{\pi_j(p)} \cdots A_{\pi_j(3)} B_{\pi_j(2)} | \cdots | A_{\pi_j(p)} B_{\pi_j(p-1)} | B_{\pi_j(p)}],\\
\mathcal{C}_j &= [C^{\top}_{\pi_j(1)} | A^{\top}_{\pi_j(1)} C^{\top}_{\pi_j(2)} | A^{\top}_{\pi_j(1)} A^{\top}_{\pi_j(2)} C^{\top}_{\pi_j(3)} | \cdots | A^{\top}_{\pi_j(1)} \cdots A^{\top}_{\pi_j(p-1)} C^{\top}_{\pi_j(p)}]^{\top},
\end{aligned}$$

and $\pi_j(\cdot)$ *is a permutation defined by*

$$\pi_j(k) = \begin{cases} k-j+1+p & \text{for} \quad k = 1, \ldots, j-1,\\ k-j+1 & \text{for} \quad k = j, \ldots, p. \end{cases}$$

Note that the pair (A, B) is d-stabilizable if $w^H B = 0$, $w^H A = \lambda w^H$ for some constant λ implies $|\lambda| < 1$ or $w = 0$, and the pair (A, C) is d-detectable if (A^H, C^H) is d-stabilizable.

Throughout this paper, the periodic matrix pair sets $\{(A_j, B_j)\}_{j=1}^p$ and $\{(A_j, C_j)\}_{j=1}^p$ of (1.1) are assumed to be pd-stabilizable and pd-detectable, respectively. The existence and uniqueness of the Hermitian p.s.d. solution set to the P-DARE (1.1) are studied in [1] and [2].

THEOREM 1.1 (see [1],[2]). *For the P-DARE* (1.1), *if* $\{(A_j, B_j)\}_{j=1}^p$ *and* $\{(A_j, C_j)\}_{j=1}^p$ *are pd-stabilizable and pd-detectable, respectively, then there is a unique Hermitian p.s.d. solution set* $\{X_j\}_{j=1}^p$ *to the P-DARE* (1.1). *Moreover, the periodic matrix set* $\{(I + G_j X_j)^{-1} A_j\}_{j=1}^p$ *is pd-stable.*

Let

$$\widetilde{X}_{j-1} = \widetilde{A}_j^H \widetilde{X}_j (I + \widetilde{G}_j \widetilde{X}_j)^{-1} \widetilde{A}_j + \widetilde{H}_j \tag{1.4}$$

for $j = 1, \ldots, p$ be a perturbed P-DARE of (1.1). Based on the technique described in [20], we shall construct an easily treated system of periodic equations of $\Delta X_j \equiv \widetilde{X}_j - X_j$ for deriving sharp upper bounds for $\|\widetilde{X}_j - X_j\|_F (j = 1, \ldots, p)$ and find some reasonable restrictions on the perturbations in the periodic coefficient matrices of the P-DARE (1.1) such that the perturbed P-DARE (1.4) has a unique Hermitian p.s.d. solution set $\{\widetilde{X}_j\}_{j=1}^p$. Moreover, applying the theory of condition developed by Rice [19], we define a condition number of the Hermitian p.s.d. solution set to the P-DARE (1.1), and, by using the techniques described in [4] and [14], we derive explicit expressions of the condition number.

This paper is organized as follows. In section 2, we prove some lemmas. In section 3, we first construct a perturbation equation for the P-DARE and then derive perturbation bounds for the Hermitian p.s.d. solution set. In section 4, we define a condition number of the Hermitian p.s.d. solution set and derive explicit expressions of the condition number. In section 5, we use a numerical example to illustrate our results.

2. Lemmas. In this section, we prove some preliminary lemmas which are used in sections 3 and 4.

Let $\Phi_j \in \mathbb{C}_n, j = 1, \ldots, p$. Define the linear operator $\mathbf{L} : \mathcal{H}_n^p \to \mathcal{H}_n^p$ by

$$\mathbf{L}(W_1, \ldots, W_p) = (W_1 - \Phi_2^H W_2 \Phi_2, \ldots, W_{p-1} - \Phi_p^H W_p \Phi_p, W_p - \Phi_1^H W_1 \Phi_1), \tag{2.1}$$

where $W_j \in \mathcal{H}_n$ for $j = 1, \ldots, p$.

LEMMA 2.1. *The linear operator* $\mathbf{L}$ *defined by* (2.1) *is singular provided that there is an eigenvalue* $\hat{\lambda} \in \lambda(\{\Phi_j\}_{j=1}^p)$ *with* $|\hat{\lambda}| = 1$.

Proof. By the periodic Schur theorem [3], there is a unitary matrix set $\{U_j\}_{j=1}^p$ such that

$$U_j^H \Phi_j U_{j-1} = \begin{bmatrix} \phi_j & 0 \\ * & * \end{bmatrix} \tag{2.2}$$

for $j = 1, \ldots, p$, and $\hat{\lambda} = \phi_p \phi_{p-1} \cdots \phi_1$. Without loss of generality, we may assume first that Φ_j has the lower triangular form as in (2.2) for $j = 1, \ldots, p$. Taking

$$W_p = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad W_j = \begin{bmatrix} |\phi_{j+1}|^2 \cdots |\phi_p|^2 & 0 \\ 0 & 0 \end{bmatrix}, \quad j = p-1, \ldots, 1, \tag{2.3}$$

and substituting (2.3) into (2.1), we have

$$W_{j-1} - \Phi_j^H W_j \Phi_j = 0, \quad j = 1, \ldots, p, \tag{2.4}$$

and, by assumption, $|\phi_1|^2 \cdots |\phi_p|^2 = 1$. Setting $W_j^* = U_j^H W_j U_j$ for $j = 1, \ldots, p$, there is a nonzero element $(W_1^*, \ldots, W_p^*) \in \mathcal{H}_n^p$ such that $\mathbf{L}(W_1^*, \ldots, W_p^*) = (0, \ldots, 0)$, which implies the assertion. ∎

LEMMA 2.2. *Let $\Phi = \mathrm{cyc}\{\Phi_j\}_{j=1}^p$, where $\Phi_j \in \mathbb{C}_n, j = 1, \ldots, p$. If $\{\Phi_j\}_{j=1}^p$ is pd-stable, then Φ is d-stable.*

Proof. Suppose that $\lambda \in \lambda(\Phi)$. Then there are $n \times 1$ vectors $x_1, \ldots, x_p$ with $(x_1^\top, \ldots, x_p^\top)^\top \neq 0$ such that

$$\mathrm{cyc}\{\Phi_j\}_{j=1}^p \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}. \tag{2.5}$$

Suppose that $x_j \neq 0$ for some j. Comparing the two sides of (2.5), we have

$$(\Phi_j \cdots \Phi_1 \Phi_p \cdots \Phi_{j+1}) x_j = \lambda^p x_j.$$

By the assumption of the pd-stability for $\{\Phi_j\}_{j=1}^p$ (see Definition 1.2), we have $|\lambda| < 1$. ∎

Let

$$L = I_{pn^2} - \begin{bmatrix} 0 & \Phi_2^\top \otimes \Phi_2^H & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \Phi_p^\top \otimes \Phi_p^H \\ \Phi_1^\top \otimes \Phi_1^H & \cdots & \cdots & 0 \end{bmatrix}. \tag{2.6}$$

Then L is a matrix representation of $\mathbf{L}$ on

$$\mathcal{H}^{pn^2} \equiv \{[w_1^\top, \ldots, w_p^\top]^\top | w_j = \mathrm{vec}(W_j), W_j \in \mathcal{H}_n, j = 1, \ldots, p\}.$$

Assume that $\{\Phi_j\}_{j=1}^p$ is pd-stable. By Lemma 2.2, the matrix L defined by (2.6) is nonsingular, and thus $\mathbf{L}^{-1}$ exists. In such a case, we define the quantity ℓ by

$$\ell = \|\mathbf{L}^{-1}\|^{-1}, \tag{2.7}$$

where the operator norm $\|\cdot\|$ for $\mathbf{L}^{-1}$ is induced by the Frobenius norm $\|\cdot\|_F$ on $\mathbb{C}_n^p$. In Appendix B, we shall prove that the quantity ℓ can be expressed by $\ell = \|L^{-1}\|_2^{-1}$.

For the pd-stable periodic matrix set $\{\Phi_j\}_{j=1}^p$, we define the quantity s by

$$s = \min\left\{ \max_{1\le j\le p} \|E_j\|_2 : \rho(\{\Phi_j + E_j\}_{j=1}^p) = 1, E_j \in \mathbb{C}_n \right\}. \tag{2.8}$$

The quantity s measures the smallest $\max_{1\le j\le p}\|E_j\|_2$ such that $\{\Phi_j + E_j\}_{j=1}^p$ has an eigenvalue on the unit circle.

LEMMA 2.3. *Let $\{\Phi_j\}_{j=1}^p$ be pd-stable, and let $\mathbf{L}$ be the linear operator defined by* (2.1) *with L of* (2.6) *as its matrix representation. Let $\varphi = \max_{1\le j\le p}\|\Phi_j\|_2$, $\ell = \|L^{-1}\|_2^{-1}$, and s be given by* (2.8). *Then*

$$\frac{\ell}{\varphi + \sqrt{\varphi^2 + \ell}} \le s. \tag{2.9}$$

Proof. Let $E_j^* \in \mathbb{C}_n$ $(j = 1, \dots, p)$ satisfy

$$s = \max_{1\le j\le p} \|E_j^*\|_2 \quad \text{with} \quad \rho(\{\Phi_j + E_j^*\}_{j=1}^p) = 1. \tag{2.10}$$

By Lemma 2.1, the transformation

$$\begin{bmatrix} W_1 \\ \vdots \\ W_p \end{bmatrix} \longmapsto \begin{bmatrix} W_1 - (\Phi_2 + E_2^*)^H W_2 (\Phi_2 + E_2^*) \\ \vdots \\ W_{p-1} - (\Phi_p + E_p^*)^H W_p (\Phi_p + E_p^*) \\ W_p - (\Phi_1 + E_1^*)^H W_1 (\Phi_1 + E_1^*) \end{bmatrix} \tag{2.11}$$

is singular, where $W_j \in \mathcal{H}_n, j = 1, \dots, p$. Thus there are Hermitian matrices $W_1^*, \dots, W_p^*$ with $W_j^* \neq 0$ for some $j \in \{1, \dots, p\}$ such that

$$\mathbf{L}\begin{bmatrix} W_1^* \\ W_2^* \\ \vdots \\ W_p^* \end{bmatrix} = \begin{bmatrix} \Phi_2^H W_2^* E_2^* + E_2^{*H} W_2^* \Phi_2 + E_2^{*H} W_2^* E_2^* \\ \vdots \\ \Phi_p^H W_p^* E_p^* + E_p^{*H} W_p^* \Phi_p + E_p^{*H} W_p^* E_p^* \\ \Phi_1^H W_1^* E_1^* + E_1^{*H} W_1^* \Phi_1 + E_1^{*H} W_1^* E_1^* \end{bmatrix}, \tag{2.12}$$

or, equivalently, by letting $\mathrm{vec}(W_j^*) = w_j^*$, we have

$$L\begin{bmatrix} w_1^* \\ \vdots \\ w_p^* \end{bmatrix} = \left(\mathrm{cyc}\{\Omega_j^\top\}_{j=1}^p\right)^\top \begin{bmatrix} w_1^* \\ \vdots \\ w_p^* \end{bmatrix}, \tag{2.13}$$

where $\Omega_j \equiv E_j^{*\top} \otimes \Phi_j^H + \Phi_j^\top \otimes E_j^{*H} + E_j^{*\top} \otimes E_j^{*H}$ for $j = 1, \dots, p$. Inverting L and taking the 2-norm of (2.13), we get $s^2 + 2\varphi s - \ell \ge 0$, which implies the inequality (2.9). ∎

The following lemma is an immediate corollary of Lemma 2.3.

LEMMA 2.4. *Let $\{\Phi_j\}_{j=1}^p$ be pd-stable, and let $\mathbf{L}$ be the linear operator defined by* (2.1) *with L in* (2.6) *as its matrix representation. Let $\varphi = \max_{1\le j\le p}\|\Phi_j\|_2$ and $\ell = \|L^{-1}\|_2^{-1}$. If $E_j \in \mathbb{C}_n (j = 1, \dots, p)$ satisfy*

$$\max_{1\le j\le p} \|E_j\|_2 \le \frac{\ell}{\varphi + \sqrt{\varphi^2 + \ell}},$$

then $\{(\Phi_j + E_j)\}_{j=1}^p$ is pd-stable. ∎

3. Perturbation results for the P-DARE. In this section, we present perturbation bounds for the Hermitian p.s.d. solution set to the P-DARE (1.1).

Consider the P-DARE (1.1),

$$X_{j-1} = A_j^H X_j (I + G_j X_j)^{-1} A_j + H_j, \quad j = 1, \dots, p,$$

and a perturbed P-DARE (1.4),

$$\widetilde{X}_{j-1} = \widetilde{A}_j^H \widetilde{X}_j (I + \widetilde{G}_j \widetilde{X}_j)^{-1} \widetilde{A}_j + \widetilde{H}_j, \quad j = 1, \dots, p.$$

For simplicity, we now consider the case of $p = 3$. Define

$$F = (I + G_j X_j)^{-1}, \quad \Phi_j = (I + G_j X_j)^{-1} A_j, \tag{3.1}$$

$$\Psi_j = X_j (I + G_j X_j)^{-1}, \quad K_j = X_j (I + G_j X_j)^{-1} A_j, \tag{3.2}$$

and define

$$\Delta X_j = \widetilde{X}_j - X_j, \quad \Delta A_j = \widetilde{A}_j - A_j, \quad \Delta G_j = \widetilde{G}_j - G_j, \quad \Delta H_j = \widetilde{H}_j - H_j \tag{3.3}$$

for $j = 1, 2, 3$.

Recall the linear operator $\mathbf{L} : \mathcal{H}_n^3 \to \mathcal{H}_n^3$ defined by (2.1), that is,

$$\mathbf{L}(W_1, W_2, W_3) = (W_1, W_2, W_3) - (\Phi_2^H W_2 \Phi_2, \Phi_3^H W_3 \Phi_3, \Phi_1^H W_1 \Phi_1), \tag{3.4}$$

where $W_j \in \mathcal{H}_n$ for $j = 1, 2, 3$, and recall its matrix representation L given by (2.6). It is easily seen that

$$\lambda(L) = \{1 - \lambda \mid \lambda^3 \in \lambda((\Phi_3 \Phi_2 \Phi_1)^\top \otimes (\Phi_3 \Phi_2 \Phi_1)^H)\}. \tag{3.5}$$

From (3.1) and Theorem 1.1, it follows that $|\lambda| < 1$, $k = 1, \dots, n$. Hence L, and thus $\mathbf{L}$, are invertible.

Further, we define the operator $\mathbf{P} : \mathbb{C}_n^3 \to \mathcal{H}_n^3$ and the linear operator $\mathbf{Q} : \mathcal{H}_n^3 \to \mathcal{H}_n^3$ by

$$\mathbf{P}(N_1, N_2, N_3) = \mathbf{L}^{-1}(K_2^H N_2 + N_2^H K_2, K_3^H N_3 + N_3^H K_3, K_1^H N_1 + N_1^H K_1) \tag{3.6}$$

and

$$\mathbf{Q}(M_1, M_2, M_3) = \mathbf{L}^{-1}(K_2^H M_2 K_2, K_3^H M_3 K_3, K_1^H M_1 K_1), \tag{3.7}$$

respectively, where $N_1, N_2, N_3 \in \mathbb{C}_n$ and $M_1, M_2, M_3 \in \mathcal{H}_n$.

The main result of this section is the following theorem.

THEOREM 3.1. *Let $\{X_j\}_{j=1}^p$ be the unique Hermitian p.s.d. solution set to the P-DARE* (1.1)*, and let $\widetilde{A}_j = A_j + \Delta A_j, \widetilde{G}_j = G_j + \Delta G_j, \widetilde{H}_j = H_j + \Delta H_j$ $(j = 1, \dots, p)$ be the coefficient matrices of the perturbed P-DARE* (1.4)*. Define the operators $\mathbf{L}, \mathbf{P}$, and $\mathbf{Q}$ by* (3.4), (3.6)*, and* (3.7)*, respectively. Let*

$$\ell = \|\mathbf{L}^{-1}\|^{-1}, \quad p_d = \|\mathbf{P}\|, \quad q_d = \|\mathbf{Q}\|, \tag{3.8}$$

where $\|\cdot\|$ denotes the operator norm induced by $\|\cdot\|_F$. Moreover, let

$$\varphi = \max_{1 \le j \le p} \|\Phi_j\|_2, \tag{3.9}$$

$$\alpha = \max_{1\le j\le p}\left\{\frac{\|F_j\|_2(\|A_j\|_2+\|\Delta A_j\|_2)}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right\}, \tag{3.10}$$

$$\gamma = \max_{1\le j\le p}\left\{\frac{\|F_j\|_2(\|G_j\|_2+\|\Delta G_j\|_2)}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right\}, \tag{3.11}$$

$$\delta_j = \frac{\|\Delta A_j\|_2+\|K_j\|_2\|\Delta G_j\|_2}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}, \quad j=1,\dots,p, \tag{3.12}$$

$$\zeta = \max_{1\le j\le p}\{\|F_j\|_2\delta_j(2\varphi+\|F_j\|_2\delta_j)\}, \tag{3.13}$$

$$\begin{aligned}\epsilon_1 = {}& \frac{1}{\ell}\|(\Delta H_1,\dots,\Delta H_p)\|_F + p_d\|(\Delta A_1,\dots,\Delta A_p)\|_F \\ & + q_d\|(\Delta G_1,\dots,\Delta G_p)\|_F,\end{aligned} \tag{3.14}$$

$$\epsilon = \epsilon_1 + \frac{1}{\ell}\left\{\sum_{j=1}^{p}\|\Psi_j\|_2^2\delta_j^2(\|\Delta A_j\|_F+\|K_j\|_2\|\Delta G_j\|_F)^2\right\}^{\frac{1}{2}}, \tag{3.15}$$

and

$$\xi_* = \frac{2\ell\epsilon}{\ell-\zeta+\ell\gamma\epsilon+\sqrt{(\ell-\zeta+\ell\gamma\epsilon)^2-4\ell\gamma(\ell-\zeta+\alpha^2)\epsilon}}. \tag{3.16}$$

If $\widetilde{G}_j, \widetilde{H}_j \ge 0$ $(j=1,\dots,p)$, *and if*

$$1-\|\Psi_j\|_2\|\Delta G_j\|_2 > 0 \quad (j=1,\dots,p), \tag{3.17}$$

$$1-\gamma\xi_* > 0, \tag{3.18}$$

$$\max_{1\le j\le p}\left\{\frac{\|F_j\|_2\delta_j+\|\Phi_j\|_2\gamma\xi_*}{1-\gamma\xi_*}\right\} < \frac{\ell}{\varphi+\sqrt{\varphi^2+\ell}}, \tag{3.19}$$

and

$$\epsilon \le \frac{(\ell-\zeta)^2}{\ell\gamma(\ell-\zeta+2\alpha+\sqrt{(\ell-\zeta+2\alpha)^2-(\ell-\zeta)^2})}, \tag{3.20}$$

then the perturbed P-DARE (1.4) *has a unique Hermitian p.s.d. solution set* $\{\widetilde{X}_j\}_{j=1}^p$, *and*

$$\|(\widetilde{X}_1-X_1,\dots,\widetilde{X}_p-X_p)\|_F \le \xi_*. \tag{3.21}$$

See Appendix A for a proof of Theorem 3.1.

Let

$$\delta_{A,G,H} = \sqrt{\sum_{j=1}^{p}\|(\Delta A_j,\Delta G_j,\Delta H_j)\|_F^2}.$$

According to the definitions of ξ_*, δ_j $(j=1,\dots,p)$, and ϵ, the conditions (3.17)–(3.20) simply mean that the quantity $\delta_{A,G,H}$ should be sufficiently small. Theorem 3.1 concludes that, if $\delta_{A,G,H}$ is so small that the conditions (3.17)–(3.20) are satisfied and $\widetilde{G}_j, \widetilde{H}_j \ge 0$ $(j=1,\dots,p)$, then the perturbed P-DARE (1.4) has a unique Hermitian p.s.d. solution set $\{\widetilde{X}_j\}_{j=1}^p$ with the estimate (3.21).

Remark 3.1. *First order estimates.* Obviously, the estimate (3.21) can be expressed by

$$\|(\Delta X_1, \ldots, \Delta X_p)\|_F = \mathcal{O}(\delta_{A,G,H}), \quad \text{as } \delta_{A,G,H} \to 0,$$

where $\Delta X_j = \widetilde{X}_j - X_j$ for $j = 1, \ldots, p$. Moreover, by the proof of Theorem 3.1 (see Appendix A), we have the first order perturbation expansion of $(\widetilde{X}_1, \ldots, \widetilde{X}_p)$ at $(X_1, \ldots, X_p)$,

$$\begin{aligned}(\widetilde{X}_1, \ldots, \widetilde{X}_p) = (X_1, \ldots, X_p) + \mathbf{L}^{-1}(\Delta H_2, \ldots, \Delta H_p, \Delta H_1) + \mathbf{P}(\Delta A_2, \ldots, \Delta A_p, \Delta A_1) \\ -\mathbf{Q}(\Delta G_2, \ldots, \Delta G_p, \Delta G_1) + \mathcal{O}(\delta_{A,G,H}^2),\end{aligned} \tag{3.22}$$

as $\delta_{A,G,H}^2 \to 0$, and thus we get the first order perturbation bound for the solution set $\{X_j\}_{j=1}^p$:

$$\begin{aligned}&\|(\widetilde{X}_1 - X_1, \ldots, \widetilde{X}_p - X_p)\|_F \\ &\le \frac{1}{\ell}\|(\Delta H_1, \ldots, \Delta H_p)\|_F + p_d\|(\Delta A_1, \ldots, \Delta A_p)\|_F + q_d\|(\Delta G_1, \ldots, \Delta G_p)\|_F \\ &= \epsilon_1 \quad \text{as} \quad \delta_{A,G,H} \to 0.\end{aligned} \tag{3.23}$$

Remark 3.2. *Outline of the proof of Theorem* 3.1. We prove Theorem 3.1 by three steps (see Appendix A for the details).

Step 1. From the P-DARE (1.1) and the perturbed P-DARE (1.4), we get an equation for $(\Delta X_1, \ldots, \Delta X_p)$, i.e., a perturbation equation.

Step 2. According to the perturbation equation, we define a continuous mapping $\mathcal{M} : \mathcal{H}_n^p \to \mathcal{H}_n^p$ so that any fixed point of $\mathcal{M}$ is a solution of the equation. Thus the problem of finding a perturbation bound for the Hermitian p.s.d. solution set $\{X_j\}_{j=1}^p$ to the P-DARE (1.1) reduces to the problem of showing the existence of a fixed point $(\Delta X_1^*, \ldots, \Delta X_p^*)$ of $\mathcal{M}$ and determining a bound on its size. This can be done by applying the Schauder fixed-point theorem under certain assumptions on the perturbations in A_j, G_j, and H_j for $j = 1, \ldots, p$.

Step 3. We prove that $(X_1 + \Delta X_1^*, \ldots, X_p + \Delta X_p^*)$ is the unique Hermitian p.s.d. solution set to the perturbed P-DARE (1.4).

Remark 3.3. *The nonperiodic case ($p = 1$).* The DARE is in the form

$$X = A^H X(I + GX)^{-1}A + H,$$

where $A \in \mathbb{C}_n, G, H \in \mathcal{H}_n$, and $G, H \ge 0$. Appropriate assumptions on the coefficient matrices guarantee the existence and uniqueness of a Hermitian p.s.d. solution X. Set $p = 1$ in Theorem 3.1; then we obtain a perturbation result for the DARE which just coincides with [20, Theorem 4.1], but the operator $\mathbf{L}$ is defined by

$$\mathbf{L}W = W - \Phi^H W \Phi, \quad W \in \mathcal{H}_n,$$

where $\Phi = (I + GX)^{-1}A$ is d-stable, and the operators $\mathbf{P}$ and $\mathbf{Q}$ are defined by

$$\mathbf{P}N = \mathbf{L}^{-1}(K^H N + N^H K), \quad N \in \mathbb{C}_n,$$

and

$$\mathbf{Q}M = \mathbf{L}^{-1}(K^H M K), \quad M \in \mathcal{H}_n,$$

respectively, in which $K = X(I + GX)^{-1}A$.

Remark 3.4. *Expression of* ℓ, p_d, *and* q_d. Let $\mathbf{L}, \mathbf{P}$, and $\mathbf{Q}$ be the operators defined by (2.1), (3.6), and (3.7), respectively, and let ℓ, p_d, and q_d be the quantities defined by (3.8). Let

$$L = I_{pn^2} - [\mathrm{cyc}\{\Phi_j \otimes \overline{\Phi}_j\}_{j=1}^p]^\top,$$

as in (2.6), where $\{\Phi_j\}_{j=1}^p$ is pd-stable. Let

$$\begin{aligned} L^{-1}[\mathrm{cyc}\{I \otimes \overline{K}_j\}_{j=1}^p]^\top &= \Omega_1 + i\Omega_2, \\ L^{-1}[\mathrm{cyc}\{\Pi^\top (K_j \otimes I)\}_{j=1}^p]^\top &= \Theta_1 + i\Theta_2, \end{aligned}$$

where $\Omega_1, \Omega_2, \Theta_1$, and Θ_2 are real matrices, Π is the vec-permutation matrix [10, pp. 32–34], and $K_j = X_j(I + G_j X_j)^{-1}A_j$ for $j = 1, \ldots, p$, as in (3.5). Then

$$\ell = \|L^{-1}\|_2^{-1}, \tag{3.24}$$

$$p_d = \left\| \begin{bmatrix} \Omega_1 + \Theta_1 & \Theta_2 - \Omega_2 \\ \Omega_2 + \Theta_2 & \Omega_1 - \Theta_1 \end{bmatrix} \right\|_2 \tag{3.25}$$

for the real case, and, especially,

$$p_d = \|L^{-1}[\mathrm{cyc}\{I \otimes K_j + \Pi^\top (K_j \otimes I)\}_{j=1}^p]^\top\|_2 \tag{3.26}$$

and

$$q_d = \|L^{-1}[\mathrm{cyc}\{K_j \otimes \overline{K}_j\}_{j=1}^p]^\top\|_2. \tag{3.27}$$

See Appendix B for a proof of the formulae (3.24)–(3.27).

4. Condition number of $\{X_j\}_{j=1}^p$. In this section, we define a condition number of the Hermitian p.s.d. solution set $\{X_j\}_{j=1}^p$ to the P-DARE (1.1) and derive explicit expressions of the condition number.

For simplicity, we first consider $p = 3$. From Theorem 3.1 and (3.22), we see that, if $G_j + \Delta G_j \geq 0$ and $H_j + \Delta H_j \geq 0$ for all j, then

$$\begin{aligned} (\Delta X_1, \Delta X_2, \Delta X_3) &= \mathbf{L}^{-1}(\Delta H_2, \Delta H_3, \Delta H_1) + \mathbf{P}(\Delta A_2, \Delta A_3, \Delta A_1) \\ &\quad -\mathbf{Q}(\Delta G_2, \Delta G_3, \Delta G_1) + \mathcal{O}(\delta_{A,G,H}^2) \\ &= \mathbf{L}^{-1}[(\Delta H_2, \Delta H_3, \Delta H_1) + (K_2^H \Delta A_2 + \Delta A_2^H K_2, K_3^H \Delta A_3 \\ &\quad + \Delta A_3^H K_3, K_1^H \Delta A_1 + \Delta A_1^H K_1) - (K_2^H \Delta G_2 K_2, K_3^H \Delta G_3 K_3, \\ &\quad K_1^H \Delta G_1 K_1)] + \mathcal{O}(\delta_{A,G,H}^2), \quad \text{as} \quad \delta_{A,G,H} \to 0, \end{aligned} \tag{4.1}$$

where $\delta_{A,G,H} = \sqrt{\sum_{j=1}^3 \|(\Delta H_j, \Delta A_j, \Delta G_j)\|_F^2}$, $\Delta H_j, \Delta G_j \in \mathcal{H}_n, \Delta A_j \in \mathbb{C}_n$ for $j = 1, 2, 3$. Let

$$\rho = \left\| \left(\frac{\Delta H_1}{\eta_1}, \frac{\Delta H_2}{\eta_2}, \frac{\Delta H_3}{\eta_3}; \frac{\Delta A_1}{\alpha_1}, \frac{\Delta A_2}{\alpha_2}, \frac{\Delta A_3}{\alpha_3}; \frac{\Delta G_1}{\gamma_1}, \frac{\Delta G_2}{\gamma_2}, \frac{\Delta G_3}{\gamma_3} \right) \right\|_F, \tag{4.2}$$

where $\eta_j, \alpha_j, \gamma_j$ are positive parameters. By the theory of condition developed by Rice [19], we define the condition number $c(X_1, X_2, X_3)$ by

$$c(X_1, X_2, X_3) = \lim_{\delta \to 0} \sup_{\substack{\rho \le \delta \\ G_j + \Delta G_j \ge 0 \\ H_j + \Delta H_j \ge 0 \ \forall j}} \frac{\left\| \left(\frac{\Delta X_1}{\xi_1}, \frac{\Delta X_2}{\xi_2}, \frac{\Delta X_3}{\xi_3} \right) \right\|_F}{\delta}, \tag{4.3a}$$

where ξ_1, ξ_2, ξ_3 are positive parameters.

By using the technique described by [21], we need only to derive an explicit expansion of $c(X_1, X_2, X_3)$ in the case of $G_j + \Delta G_j > 0$ and $H_j + \Delta H_j > 0$ for all j; and in such a case, the definition (4.3a) can be written

$$c(X_1, X_2, X_3) = \lim_{\delta \to 0} \sup_{\rho \le \delta} \frac{\left\| \left(\frac{\Delta X_1}{\xi_1}, \frac{\Delta X_2}{\xi_2}, \frac{\Delta X_3}{\xi_3} \right) \right\|_F}{\delta}. \tag{4.3b}$$

Define the operator $\mathbf{V} : \mathcal{H}_n^3 \times \mathbb{C}_n^3 \times \mathcal{H}_n^3 \to \mathcal{H}_n^3$ by

$$\begin{aligned} &\mathbf{V}(M_1, M_2, M_3, D_1, D_2, D_3, Q_1, Q_2, Q_3) \\ &= \mathbf{L}^{-1}[(M_2, M_3, M_1)H^{(2)} + (K_2^H D_2 + D_2^H K_2, K_3^H D_3 + D_3^H K_3, K_1^H D_1 + D_1^H K_1)A^{(2)} \\ &\quad -(K_2^H Q_2 K_2, K_3^H Q_3 K_3, K_1^H Q_1 K_1)\Gamma^{(2)}]\Xi^{-1}, \end{aligned} \tag{4.4}$$

where $M_j, Q_j \in \mathcal{H}_n, D_j \in \mathbb{C}_n$ for $j = 1, 2, 3$, and

$$\begin{aligned} &H^{(2)} = \operatorname{diag}(\eta_2 I_n, \eta_3 I_n, \eta_1 I_n), \quad A^{(2)} = \operatorname{diag}(\alpha_2 I_n, \alpha_3 I_n, \alpha_1 I_n), \\ &\Gamma^{(2)} = \operatorname{diag}(\gamma_2 I_n, \gamma_3 I_n, \gamma_1 I_n), \quad \Xi = \operatorname{diag}(\xi_1 I_n, \xi_2 I_n, \xi_3 I_n). \end{aligned} \tag{4.5}$$

Substituting (4.1) into (4.3b) gives

$$c(X_1, X_2, X_3) = \sup_{\substack{(M_1, \dots; D_1, \dots; Q_1, \dots) \ne 0 \\ M_j, Q_j \in \mathcal{H}_n, D_j \in \mathbb{C}_n, \forall j}} \frac{\|\mathbf{V}(M_1, M_2, M_3, D_1, D_2, D_3, Q_1, Q_2, Q_3)\|_F}{\|(M_1, M_2, M_3, D_1, D_2, D_3, Q_1, Q_2, Q_3)\|_F}. \tag{4.6}$$

Further, we define the operator $\widehat{\mathbf{V}} : \mathbb{C}_n^3 \times \mathbb{C}_n^3 \times \mathbb{C}_n^3 \to \mathbb{C}_n^3$ by

$$\begin{aligned} &\widehat{\mathbf{V}}(N_1, N_2, N_3, E_1, E_2, E_3, R_1, R_2, R_3) \\ &= \widehat{\mathbf{L}}^{-1}[(N_2, N_3, N_1)H^{(2)} + (K_2^H E_2 + E_2^H K_2, K_3^H E_3 + E_3^H K_3, K_1^H E_1 + E_1^H K_1)A^{(2)} \\ &\qquad - (K_2^H R_2 K_2, K_3^H R_3 K_3, K_1^H R_1 K_1)\Gamma^{(2)}]\Xi^{-1}, \end{aligned} \tag{4.7}$$

where $N_j, E_j, R_j \in \mathbb{C}_n$ for $j = 1, 2, 3$, and $\widehat{\mathbf{L}}$ is a natural extension of $\mathbf{L}$ on $\mathbb{C}_n^3$. From the definitions (4.4) and (4.7), we know that

$$\sup_{\substack{(M_1, \dots; D_1, \dots; Q_1, \dots) \ne 0 \\ M_j, Q_j \in \mathcal{H}_n, D_j \in \mathbb{C}_n, \forall j}} \frac{\|\mathbf{V}(M_1, M_2, M_3, D_1, D_2, D_3, Q_1, Q_2, Q_3)\|_F}{\|(M_1, M_2, M_3, D_1, D_2, D_3, Q_1, Q_2, Q_3)\|_F} \tag{4.8a}$$

$$\le \sup_{\substack{(N_1, \dots; E_1, \dots; R_1, \dots) \ne 0 \\ N_j, E_j, R_j \in \mathbb{C}_n, \forall j}} \frac{\|\widehat{\mathbf{V}}(N_1, N_2, N_3, E_1, E_2, E_3, R_1, R_2, R_3)\|_F}{\|(N_1, N_2, N_3, E_1, E_2, E_3, R_1, R_2, R_3)\|_F}. \tag{4.8b}$$

We now prove that the equality in (4.8) holds. Let $(N_1^*, N_2^*, N_3^*, E_1^*, E_2^*, E_3^*, R_1^*, R_2^*, R_3^*)$ be the singular "vector" of $\widehat{\mathbf{V}}$ corresponding to the maximal singular value; that is, the right-hand side of (4.8b) equals

$$\frac{\|\widehat{\mathbf{V}}(N_1^*, N_2^*, N_3^*, E_1^*, E_2^*, E_3^*, R_1^*, R_2^*, R_3^*)\|_F}{\|(N_1^*, N_2^*, N_3^*, E_1^*, E_2^*, E_3^*, R_1^*, R_2^*, R_3^*)\|_F}. \tag{4.9}$$

Let

$$(Z_1^*, Z_2^*, Z_3^*)\Xi^{-1} = \widehat{\mathbf{V}}(N_1^*, N_2^*, N_3^*, E_1^*, E_2^*, E_3^*, R_1^*, R_2^*, R_3^*) \in \mathbb{C}_n^3. \tag{4.10}$$

Then, by the definition (4.7) and the definition of $\widehat{\mathbf{L}}$, we have

$$\begin{aligned}&(Z_1^*, Z_2^*, Z_3^*) - (\Phi_2^H Z_2^* \Phi_2, \Phi_3^H Z_3^* \Phi_3, \Phi_1^H Z_1^* \Phi_1) = \widehat{\mathbf{L}}(Z_1^*, Z_2^*, Z_3^*)\\ &= (N_2^*, N_3^*, N_1^*)H^{(2)} + (K_2^H E_2^* + E_2^{*H} K_2, K_3^H E_3^* + E_3^{*H} K_3, K_1^H E_1^* + E_1^{*H} K_1)A^{(2)}\\ &-(K_2^H R_2^* K_2, K_3^H R_3^* K_3, K_1^H R_1^* K_1)\Gamma^{(2)}.\end{aligned} \tag{4.11}$$

Further, from (4.11),

$$\begin{aligned}(Z_1^{*H}, Z_2^{*H}, Z_3^{*H}) = \widehat{\mathbf{L}}^{-1}[&(N_2^{*H}, N_3^{*H}, N_1^{*H})H^{(2)} + (K_2^H E_2^{*H} + E_2^{*H} K_2, K_3^H E_3^{*H} + E_3^{*H} K_3,\\ &K_1^H E_1^{*H} + E_1^{*H} K_1)A^{(2)} - (K_2^H R_2^{*H} K_2, K_3^H R_3^{*H} K_3, K_1^H R_1^{*H} K_1)\Gamma^{(2)}].\end{aligned} \tag{4.12}$$

Since $\|(Z_1^{*H}, Z_2^{*H}, Z_3^{*H})\Xi^{-1}\|_F = \|(Z_1^*, Z_2^*, Z_3^*)\Xi^{-1}\|_F$, from(4.9), (4.10), and (4.12), it follows that the right-hand side of (4.8b) equals

$$\frac{\|\widehat{\mathbf{V}}(N_1^{*H}, N_2^{*H}, N_3^{*H}, E_1^{*H}, E_2^{*H}, E_3^{*H}, R_1^{*H}, R_2^{*H}, R_3^{*H})\|_F}{\|(N_1^{*H}, N_2^{*H}, N_3^{*H}, E_1^{*H}, E_2^{*H}, E_3^{*H}, R_1^{*H}, R_2^{*H}, R_3^{*H})\|_F}; \tag{4.13}$$

that is, $(N_1^{*H}, N_2^{*H}, N_3^{*H}, E_1^{*H}, E_2^{*H}, E_3^{*H}, R_1^{*H}, R_2^{*H}, R_3^{*H})$ is also a singular "vector" of $\widehat{\mathbf{V}}$ corresponding to the maximal singular value.

Let

$$M_j^* = N_j^* + N_j^{*H} \in \mathcal{H}_n, \quad D_j^* = 2E_j^* \in \mathbb{C}_n, \quad Q_j^* = R_j^* + R_j^{*H} \in \mathcal{H}_n \tag{4.14}$$

for $j = 1, 2, 3$. If $(M_1^*, M_2^*, M_3^*, D_1^*, D_2^*, D_3^*, Q_1^*, Q_2^*, Q_3^*) \neq 0$, then it is also a singular "vector" of $\widehat{\mathbf{V}}$ corresponding to the maximal singular value. By (4.4), (4.7), and the pd-stability of $\{\Phi_j\}_{j=1}^p$, the right-hand side of (4.8b) equals

$$\begin{aligned}&\frac{\|\widehat{\mathbf{V}}(M_1^*, M_2^*, M_3^*, D_1^*, D_2^*, D_3^*, Q_1^*, Q_2^*, Q_3^*)\|_F}{\|(M_1^*, M_2^*, M_3^*, D_1^*, D_2^*, D_3^*, Q_1^*, Q_2^*, Q_3^*)\|_F}\\ &= \frac{\|\mathbf{V}(M_1^*, M_2^*, M_3^*, D_1^*, D_2^*, D_3^*, Q_1^*, Q_2^*, Q_3^*)\|_F}{\|(M_1^*, M_2^*, M_3^*, D_1^*, D_2^*, D_3^*, Q_1^*, Q_2^*, Q_3^*)\|_F}.\end{aligned} \tag{4.15}$$

If $(M_1^*, M_2^*, M_3^*, D_1^*, D_2^*, D_3^*, Q_1^*, Q_2^*, Q_3^*) = 0$, then we set

$$M_j^o = iN_j^* \in \mathcal{H}_n, \quad D_j^o = 0 \in \mathbb{C}_n, \quad Q_j^o = iR_j^* \in \mathcal{H}_n \tag{4.16}$$

for $j = 1, 2, 3$. In such a case, $(M_1^o, M_2^o, M_3^o, D_1^o, D_2^o, D_3^o, Q_1^o, Q_2^o, Q_3^o)$ is also a singular "vector" of $\widehat{\mathbf{V}}$ corresponding to the maximal singular value. Hence the right-hand side of (4.8b) equals

$$\begin{aligned}&\frac{\|\widehat{\mathbf{V}}(M_1^o, M_2^o, M_3^o, D_1^o, D_2^o, D_3^o, Q_1^o, Q_2^o, Q_3^o)\|_F}{\|(M_1^o, M_2^o, M_3^o, D_1^o, D_2^o, D_3^o, Q_1^o, Q_2^o, Q_3^o)\|_F}\\ &= \frac{\|\mathbf{V}(M_1^o, M_2^o, M_3^o, D_1^o, D_2^o, D_3^o, Q_1^o, Q_2^o, Q_3^o)\|_F}{\|(M_1^o, M_2^o, M_3^o, D_1^o, D_2^o, D_3^o, Q_1^o, Q_2^o, Q_3^o)\|_F}.\end{aligned} \tag{4.17}$$

Therefore, from (4.15) and (4.17), it follows that the equality in (4.8) holds. Combining this result with (4.4), (4.6), and (4.7), we obtain

$$c(X_1, X_2, X_3) = \sup_{\substack{(N_1, \dots\,; E_1, \dots\,; R_1, \dots\,) \neq 0 \\ N_j,\, E_j,\, R_j \,\in\, \mathbb{C}_n,\, \forall j}} \frac{\|C(N_1, N_2, N_3, E_1, E_2, E_3, R_1, R_2, R_3)\|_F}{\|(N_1, N_2, N_3, E_1, E_2, E_3, R_1, R_2, R_3)\|_F}, \tag{4.18}$$

where

$$\begin{aligned}&C(N_1, N_2, N_3, E_1, E_2, E_3, R_1, R_2, R_3)\\ &= \widehat{\mathbf{L}}^{-1}\big[(N_2, N_3, N_1)H^{(2)} + (K_2^H E_2 + E_2^H K_2, K_3^H E_3 + E_3^H K_3, K_1^H E_1 + E_1^H K_1)A^{(2)}\\ &\quad - (K_2^H R_2 K_2, K_3^H R_3 K_3, K_1^H R_1 K_1)\Gamma^{(2)}\big]\Xi^{-1}.\end{aligned} \tag{4.19}$$

For the general case of an arbitrary $p \geq 2$, we have a similar formula to (4.18) and (4.19).

Let $z_j = \mathrm{vec}(N_j), w_j = \mathrm{vec}(E_j), c_j = \mathrm{vec}(R_j)$ for $j = 1, \dots, p$. Then, from (4.18) and (4.19), we see that $c(X_1, \dots, X_p)$ can be written as

$$c(X_1, \dots, X_p) = \sup_{\substack{[z_1^\top, \dots\,; w_1^\top, \dots\,; c_1^\top, \dots\,]^\top \neq 0 \\ z_j,\, w_j,\, c_j \,\in\, \mathbb{C}^n,\, \forall j}} \left\{ \frac{\left\| \widehat{\Xi}^{-1} \left(L^{-1}\widehat{H}^{(2)} \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} + L^{-1}\widehat{A}^{(2)} \begin{bmatrix} (I \otimes K_2^H)w_2 + (K_2^\top \otimes I)\Pi\overline{w}_2 \\ \vdots \\ (I \otimes K_p^H)w_p + (K_p^\top \otimes I)\Pi\overline{w}_p \end{bmatrix} - L^{-1}\widehat{\Gamma}^{(2)} \begin{bmatrix} (K_2^\top \otimes K_2^H)c_2 \\ \vdots \\ (K_p^\top \otimes K_p^H)c_p \\ (K_1^\top \otimes K_1^H)c_1 \end{bmatrix} \right) \right\|_2}{\sqrt{\sum_{j=1}^p (\|z_j\|_2^2 + \|w_j\|_2^2 + \|c_j\|_2^2)}} \right\} \tag{4.20}$$

$$= \sup_{\substack{[z_1^\top, \dots\,; w_1^\top, \dots\,; \\ c_1^\top, \dots\,]^\top \neq 0}} \frac{\left\| \widehat{\Xi}^{-1} \left(L^{-1}\widehat{H}^{(2)} \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} + P_1 \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} + P_2 \begin{bmatrix} \overline{w}_1 \\ \vdots \\ \overline{w}_p \end{bmatrix} - Q \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} \right) \right\|_2}{\sqrt{\sum_{j=1}^p (\|z_j\|_2^2 + \|w_j\|_2^2 + \|c_j\|_2^2)}},$$

where

$$\begin{aligned}&\widehat{H}^{(2)} = \mathrm{diag}(\eta_2 I_{n^2}, \dots, \eta_p I_{n^2}, \eta_1 I_{n^2}), \quad \widehat{A}^{(2)} = \mathrm{diag}(\alpha_2 I_{n^2}, \dots, \alpha_p I_{n^2}, \alpha_1 I_{n^2}),\\ &\widehat{\Gamma}^{(2)} = \mathrm{diag}(\gamma_2 I_{n^2}, \dots, \gamma_p I_{n^2}, \gamma_1 I_{n^2}), \quad \widehat{\Xi} = \mathrm{diag}(\xi_1 I_{n^2}, \dots, \xi_p I_{n^2})\end{aligned} \tag{4.21}$$

and

$$P_1 = L^{-1}[\mathrm{cyc}\{\alpha_j I \otimes \overline{K}_j\}_{j=1}^p]^\top,\ P_2 = L^{-1}[\mathrm{cyc}\{\alpha_j \Pi^\top (K_j \otimes I)\}_{j=1}^p]^\top,$$
$$Q = L^{-1}[\mathrm{cyc}\{\gamma_j K_j \otimes \overline{K}_j\}_{j=1}^p]^\top.$$

Denote

$$z_j = x_j + iy_j, \quad w_j = u_j + iv_j, \quad c_j = a_j + ib_j,$$

where $x_j, y_j, u_j, v_j, a_j, b_j \in \mathbb{R}^{n^2}$ for $j = 1, \ldots, p$, and

$$x = [x_1^\top, \ldots, x_p^\top]^\top, \quad y = [y_1^\top, \ldots, y_p^\top]^\top,$$
$$u = [u_1^\top, \ldots, u_p^\top]^\top, \quad v = [v_1^\top, \ldots, v_p^\top]^\top,$$
$$a = [a_1^\top, \ldots, a_p^\top]^\top, \quad b = [b_1^\top, \ldots, b_p^\top]^\top$$

as well as

$$\begin{aligned} &L^{-1}\widehat{H}^{(2)} = \Omega_1 + i\Omega_2,\ P_1 = U_1 + iU_2,\ P_2 = V_1 + iV_2, \\ &Q = \Theta + i\Theta_2, \end{aligned} \tag{4.22}$$

where $\Omega_k, U_k, V_k, \Theta_k$ are real $pn^2 \times pn^2 matrices (k = 1, 2)$. By a technique given by [14], substituting (4.22) into (4.21), we get

$$\begin{aligned} &c(X_1, \ldots, X_p) \\ &= \sup_{\substack{[x^\top, y^\top, u^\top, \\ v^\top, a^\top, b^\top] \neq 0}} \frac{\left\| \left[\begin{array}{cc} \widehat{\Xi}^{-1} & 0 \\ 0 & \widehat{\Xi}^{-1} \end{array}\right] \left[\begin{array}{c|c|c|c|c|c} \Omega_1 & -\Omega_2 & U_1+V_1 & V_2-U_2 & -\Theta_1 & \Theta_2 \\ \Omega_2 & \Omega_1 & U_2+V_2 & U_1-V_1 & -\Theta_2 & -\Theta_1 \end{array}\right] \left[\begin{array}{c} x \\ y \\ u \\ v \\ a \\ b \end{array}\right] \right\|_2}{\sqrt{\|x\|_2^2 + \|y\|_2^2 + \|u\|_2^2 + \|v\|_2^2 + \|a\|_2^2 + \|b\|_2^2}}. \\ &= \left\| \left[\begin{array}{cc} \widehat{\Xi}^{-1} & 0 \\ 0 & \widehat{\Xi}^{-1} \end{array}\right] \left[\begin{array}{c|c|c|c|c|c} \Omega_1 & -\Omega_2 & U_1+V_1 & V_2-U_2 & -\Theta_1 & \Theta_2 \\ \Omega_2 & \Omega_1 & U_2+V_2 & U_1-V_1 & -\Theta_2 & -\Theta_1 \end{array}\right] \right\|_2. \end{aligned} \tag{4.23}$$

Taking

$$\xi_j = \eta_j = \alpha_j = \gamma_j = 1 \quad (j = 1, \ldots, p),$$

we get the absolute condition number $c_{abs}(X_1, \ldots, X_p)$; and taking

$$\xi_j = \|X_j\|_F, \eta_j = \|H_j\|_F, \alpha_j = \|A_j\|_F, \gamma_j = \|G_j\|_F \quad (j = 1, \ldots, p),$$

we get the relative condition number $c_{rel}(X_1, \ldots, X_p)$.

For the real case, we can prove that the equality in (4.9) also holds. Consequently, from (4.21), the condition number $c(X_1, \ldots, X_p)$ can be explicitly expressed as follows:

$$\begin{aligned} c(X_1, \ldots, X_p) &= \sup_{\substack{[z^\top, w^\top, c^\top]^\top \neq 0 \\ z, w, c \in \mathbb{R}^{pn^2}}} \frac{\|L^{-1}\widehat{H}^{(2)} z + P_1 w + P_2 w - Qc\|_2}{\sqrt{\|z\|_2^2 + \|w\|_2^2 + \|c\|_2^2}} \\ &= \left\| \widehat{\Xi}^{-1} L^{-1} \left(\widehat{H}^{(2)}, \left[\mathrm{cyc}\left\{\alpha_j \left[I \otimes K_j + \Pi^\top (K_j \otimes I)\right]\right\}_{j=1}^p\right]^\top, \left[\mathrm{cyc}\left\{\gamma_j \left(K_j \otimes K_j\right)\right\}_{j=1}^p\right]^\top \right) \right\|_2. \end{aligned} \tag{4.24}$$

The absolute condition number $c_{abs}(X_1, \ldots, X_p)$ and the relative condition number $c_{rel}(X_1, \ldots, X_p)$ for the real case can be obtained by evaluating ξ_j, η_j, α_j, and γ_j as above.

5. A numerical example. In this section, we use numerical examples to illustrate our perturbation results given in sections 3 and 4. All computations were performed using MATLAB version 5.3 on a Compaq/DS20 workstation. The machine precision is 2.22×10^{-16}.

Example 5.1 (see [13] for $p = 1$). Consider the P-DARE (1.1) with $n = 3$ and $p = 3$. Let

$$H_j^{(0)} = \operatorname{diag}\left(\frac{1}{j}10^m, j, j \times 10^{-m}\right), \quad G_j^{(0)} = \operatorname{diag}\left(\frac{1}{j}10^{-m}, \frac{1}{j}10^{-m}, j \times 10^{-m}\right), \quad j = 1, 2, 3,$$
$$A_1^{(0)} = \operatorname{diag}(0, 10^{-m}, 1), \quad A_2^{(0)} = \operatorname{diag}(10^{-9}, 10^{-m}, 1 + 10^{-3}),$$
$$A_3^{(0)} = \operatorname{diag}(10^{-3}, 10^{-m+1}, 0.5),$$

and

$$V = I - 2vv^\top, \quad v = \frac{1}{\sqrt{3}}[1, 1, 1]^\top.$$

The coefficient matrices of (1.1) are constructed by

$$H_j = V^\top H_j^{(0)} V, \quad A_j = V^\top A_j^{(0)} V, \quad G_j = V^\top G_j^{(0)} V \equiv B_j B_j^\top, \quad j = 1, 2, 3.$$

The unique symmetric p.s.d. solution set $\{X_j\}_{j=1}^3$ to (1.1) is given by $X_j = V^\top X_j^{(0)} V$ for $j = 1, 2, 3$ with

$$X_j^{(0)} = 2\Big[P_j^{(0)} + \Big(P_j^{(0)^2} + 4\widehat{H}_j^{(0)}\widehat{G}_j^{(0)}\Big)^{1/2}\Big]\widehat{G}_j^{(0)^{-1}} \text{ (diagonal)}$$

and

$$P_j^{(0)} = \widehat{A}_j^{(0)^2} + \widehat{H}_j^{(0)}\widehat{G}_j^{(0)} - I_3 \text{ (diagonal)},$$

where

$$\widehat{A}_j^{(0)} = A_j^{(0)}(I_3 + \widehat{G}_{j-1}^{(0)} H_j^{(0)})^{-1}\widehat{A}_{j-1}^{(0)},$$
$$\widehat{G}_j^{(0)} = G_j^{(0)} + A_j^{(0)}\widehat{G}_{j-1}^{(0)}(I_3 + H_j^{(0)}\widehat{G}_{j-1}^{(0)})^{-1} A_j^{(0)^\top},$$
$$\widehat{H}_j^{(0)} = \widehat{H}_{j-1}^{(0)} + \widehat{A}_{j-1}^{(0)^\top}(I_3 + H_j^{(0)}\widehat{G}_{j-1}^{(0)})^{-1} H_j^{(0)}\widehat{A}_{j-1}^{(0)},$$

and

$$\widehat{A}_{j-1}^{(0)} = A_{j-1}^{(0)}(I_3 + G_{j-2}^{(0)} H_{j-1}^{(0)})^{-1} A_{j-2}^{(0)},$$
$$\widehat{G}_{j-1}^{(0)} = G_{j-1}^{(0)} + A_{j-1}^{(0)} G_{j-2}^{(0)}(I_3 + H_{j-1}^{(0)} G_{j-2}^{(0)})^{-1} A_{j-1}^{(0)^\top},$$
$$\widehat{H}_{j-1}^{(0)} = H_{j-2}^{(0)} + A_{j-2}^{(0)^\top}(I_3 + H_{j-1}^{(0)} G_{j-2}^{(0)})^{-1} H_{j-1}^{(0)} A_{j-2}^{(0)}.$$

Let

$$\Delta H_j^{(0)} = \begin{bmatrix} \frac{1}{j}10^m & -5j & 7 \\ -5j & j & 3 \\ 7 & 3 & j\times 10^{-m} \end{bmatrix} \times 10^{-k}, \quad \Delta A_j^{(0)} = \begin{bmatrix} 3j & -\frac{4}{j} & \frac{8}{j} \\ -6j & \frac{2}{j} & -\frac{9}{j} \\ 2j & 7j & \frac{5}{j} \end{bmatrix} \times 10^{-k},$$

$$\Delta G_j^{(0)} = \begin{bmatrix} \frac{1}{j}10^{-m} & -\frac{1}{j}10^{-m} & \frac{2}{j}10^{-m} \\ -\frac{1}{j}10^{-m} & \frac{5}{j}10^{-m} & -j\times 10^{-m} \\ \frac{2}{j}10^{-m} & -j\times 10^{-m} & 3j\times 10^{-m} \end{bmatrix} \times 10^{-k}, \quad j=1,2,3.$$

The coefficient matrices of the perturbed P-DARE (1.4) are given by

$$\widetilde{H}_j = H_j + V^\top(\Delta H_j^{(0)})V, \quad \widetilde{A}_j = A_j + V^\top(\Delta A_j^{(0)})V,$$
$$\widetilde{G}_j = G_j + V^\top(\Delta G_j^{(0)})V \equiv \widetilde{B}_j\widetilde{B}_j^\top.$$

By applying the file "DARE" of Control System Toolbox in MATLAB, one can compute the unique symmetric p.s.d. solution set $\{\widetilde{X}_j\}_{j=1}^3$ to (1.4). Denote

$$\delta_x = \|(\widetilde{X}_1 - X_1, \widetilde{X}_2 - X_2, \widetilde{X}_3 - X_3)\|_F, \quad n_x = \|(X_1, X_2, X_3)\|_F.$$

Let ϵ_1 be the quantity defined by (A.21), where l, p_d, and q_d are given by (3.24), (3.26), and (3.27), respectively, and let

$$\delta_h = \|(\Delta H_1^{(0)}, \Delta H_2^{(0)}, \Delta H_3^{(0)})\|_F, \quad \delta_a = \|(\Delta A_1^{(0)}, \Delta A_2^{(0)}, \Delta A_3^{(0)})\|_F,$$
$$\delta_g = \|(\Delta G_1^{(0)}, \Delta G_2^{(0)}, \Delta G_3^{(0)})\|_F.$$

From (3.23), (3.26), and (3.27), we have an immediate upper bound for ϵ_1:

$$\epsilon_1 \le \widehat{\epsilon}_1 \equiv \frac{1}{\ell}\left(\delta_h + 2\max_{1\le j\le 3}\{\|K_j\|_2\}\delta_a + \max_{1\le j\le 3}\{\|K_j\|_2^2\}\delta_g\right). \tag{5.1}$$

Some numerical results on relative and absolute perturbation bounds are listed in Tables 5.1–5.3, where the bounds $\epsilon_1, \widehat{\epsilon}_1$, and ξ_* are defined by (A.21), (5.1), and (A.32). The relative condition number $c_{rel} \equiv c_{rel}(X_1, X_2, X_3)$ and the absolute condition number $c_{abs} \equiv c_{abs}(X_1, X_2, X_3)$ are computed by (4.24). The cases when the conditions of Theorem 3.1 are violated are denoted by asterisks.

The results listed in Tables 5.1–5.3 show that the relative perturbation bounds ϵ_1/n_x and ξ_*/n_x, as well as the absolute perturbation bounds ϵ_1 and ξ_*, are fairly sharp. The immediate upper bound $\widehat{\epsilon}_1$ for ϵ_1 has almost the same order as ϵ_1 in this example, which might be used to estimate ϵ_1 economically when the size of the system is too large.

Appendix A. Proof of Theorem 3.1. For simplicity, we now consider the case of $p = 3$.

Step 1. *Perturbation equation.* Let

$$X = \mathrm{diag}\{X_j\}_{j=1}^3, \quad A = \mathrm{cyc}\{A_j\}_{j=1}^3,$$
$$G = \mathrm{diag}\{G_j\}_{j=1}^3, \quad H = \mathrm{diag}(H_2, H_3, H_1). \tag{A.1}$$

TABLE 5.1
$(m = 2)$.

k	δ_x/n_x	ϵ_1/n_x	$\widehat{\epsilon}_1/n_x$	ξ_*/n_x	c_{rel}
4	1.0095e-03	1.0737e-03	1.0792e-03	1.0984e-03	2.7245e+00
5	1.0102e-04	1.0737e-04	1.0792e-04	1.0761e-04	2.7245e+00
6	1.0103e-05	1.0737e-05	1.0792e-05	1.0739e-05	2.7245e+00
7	1.0103e-06	1.0737e-06	1.0792e-06	1.0737e-06	2.7245e+00
8	1.0103e-07	1.0737e-07	1.0792e-07	1.0737e-07	2.7245e+00
9	1.0103e-08	1.0737e-08	1.0792e-08	1.0737e-08	2.7245e+00
10	1.0103e-09	1.0737e-09	1.0792e-09	1.0737e-09	2.7245e+00

TABLE 5.2
$(m = 2)$.

k	δ_x	ϵ_1	$\widehat{\epsilon}_1$	ξ_*	c_{abs}
4	1.1783e-01	1.2532e-01	1.2596e-01	1.2820e-01	3.1886e+00
5	1.1791e-02	1.2532e-02	1.2596e-02	1.2560e-02	3.1886e+00
6	1.1792e-03	1.2532e-03	1.2596e-03	1.2535e-03	3.1886e+00
7	1.1792e-04	1.2532e-04	1.2596e-04	1.2532e-04	3.1886e+00
8	1.1792e-05	1.2532e-05	1.2596e-05	1.2532e-05	3.1886e+00
9	1.1792e-06	1.2532e-06	1.2596e-06	1.2532e-06	3.1886e+00
10	1.1792e-07	1.2532e-07	1.2596e-07	1.2532e-07	3.1886e+00

TABLE 5.3
$(k = 10)$.

m	δ_x	ϵ_1	$\widehat{\epsilon}_1$	ξ_*	c_{abs}
0	2.3182e-08	4.2905e-07	8.9422e-06	*	1.5058e+02
1	1.0716e-08	1.5768e-08	2.1472e-08	*	3.0917e+00
2	1.1792e-07	1.2532e-07	1.2596e-07	1.2532e-07	3.1886e+00
3	1.1824e-06	1.2522e-06	1.2587e-06	1.2522e-06	3.1873e+00
4	1.1824e-05	1.2522e-05	1.2588e-05	1.2522e-05	5.8252e+01
5	1.1824e-04	1.2522e-04	1.2588e-04	1.2522e-04	5.6298e+03
6	1.1824e-03	1.2522e-03	1.2588e-03	1.2522e-03	5.6278e+05
7	1.1824e-02	1.2522e-02	1.2588e-02	1.2522e-02	5.6278e+07

Then (1.1) can be expressed by

$$X = A^H X(I + GX)^{-1}A + H. \tag{A.2}$$

Let

$$F = \operatorname{diag}\{(I + G_jX_j)^{-1}\}_{j=1}^3 \equiv \operatorname{diag}\{F_j\}_{j=1}^3, \tag{A.3}$$

$$\Phi = \operatorname{cyc}\{(I + G_jX_j)^{-1}A_j\}_{j=1}^3 \equiv \operatorname{cyc}\{\Phi_j\}_{j=1}^3. \tag{A.4}$$

Moreover, let

$$\Psi = \operatorname{diag}\{X_j(I + G_jX_j)^{-1}\}_{j=1}^3 \equiv \operatorname{diag}\{\Psi_j\}_{j=1}^3, \tag{A.5}$$

$$K = \operatorname{cyc}\{X_j(I + G_jX_j)^{-1}A_j\}_{j=1}^3 \equiv \operatorname{cyc}\{K_j\}_{j=1}^3, \tag{A.6}$$

$$\begin{aligned}\Theta &= \operatorname{diag}\{(I + G_jX_j)^{-1}(I + \Delta G_jX_j(I + G_jX_j)^{-1})^{-1}\}_{j=1}^3 \\ &\equiv \operatorname{diag}\{\Theta_j\}_{j=1}^3.\end{aligned} \tag{A.7}$$

By [20, (4.7)–(4.12)], we have the perturbation equation

$$\Delta X - \Phi^H \Delta X \Phi = E_1 + E_2 + h_1(\Delta X) + h_2(\Delta X), \tag{A.8}$$

where

$$\Delta X - \Phi^H \Delta X \Phi = \operatorname{diag}(\Delta X_1 - \Phi_2^H \Delta X_2 \Phi_2, \Delta X_2 - \Phi_3^H \Delta X_3 \Phi_3, \Delta X_3 - \Phi_1^H \Delta X_1 \Phi_1), \tag{A.9}$$

$$E_1 = \Delta H + K^H \Delta A + \Delta A^H K - K^H \Delta G K = \operatorname{diag}(E_{12}, E_{13}, E_{11}) \tag{A.10a}$$

with

$$E_{1j} = \Delta H_j + K_j^H \Delta A_j + \Delta A_j^H K_j - K_j \Delta G_j K_j \in \mathcal{H}_n \quad \text{for } j = 1, 2, 3\,; \tag{A.10b}$$

$$\begin{aligned} E_2 &= \Delta A^H \Psi \Delta A + K^H \Delta G \Psi (I + \Delta G \Psi)^{-1} \Delta G K \\ &\quad - K^H \Delta G \Psi (I + \Delta G \Psi)^{-1} \Delta A - \Delta A^H \Psi (I + \Delta G \Psi)^{-1} \Delta G (K + \Psi \Delta A) \\ &= \operatorname{diag}(E_{22}, E_{23}, E_{21}) \end{aligned} \tag{A.11a}$$

with

$$\begin{aligned} E_{2j} &= \Delta A_j^H \Psi_j \Delta A_j + K_j^H \Delta G_j \Psi_j (I + \Delta G_j \Psi_j)^{-1} \Delta G_j K_j \\ &\quad - K_j^H \Delta G_j \Psi_j (I + \Delta G_j \Psi_j)^{-1} \Delta A_j \\ &\quad - \Delta A_j^H \Psi_j (I + \Delta G_j \Psi_j)^{-1} \Delta G_j (K_j + \Psi_j \Delta A_j) \\ &\quad \in \mathcal{H}_n, \quad \text{for } j = 1, 2, 3; \end{aligned} \tag{A.11b}$$

$$h_1(\Delta X) = \Delta \Phi^H \Delta X \Phi + \Phi^H \Delta X \Delta \Phi + \Delta \Phi^H \Delta X \Delta \Phi$$

with

$$\Delta \Phi = F(I + \Delta G \Psi)^{-1} (\Delta A - \Delta G K) \equiv \operatorname{cyc}\{\Delta \Phi_j\}_{j=1}^3, \tag{A.12a}$$

in which

$$\Delta \Phi_j = F_j (I + \Delta G_j \Psi_j)^{-1} (\Delta A_j - \Delta G_j K_j) \quad \text{for } j = 1, 2, 3; \tag{A.12b}$$

so we have

$$h_1(\Delta X) = \operatorname{diag}(h_{12}(\Delta X), h_{13}(\Delta X), h_{11}(\Delta X)) \tag{A.13a}$$

with

$$h_{1j}(\Delta X) = \Delta \Phi_j^H \Delta X_j \Phi_j + \Phi_j^H \Delta X_j \Delta \Phi_j + \Delta \Phi_j^H \Delta X_j \Delta \Phi_j \in \mathcal{H}_n, \quad \text{for } j = 1, 2, 3 \tag{A.13b}$$

and

$$\begin{aligned} h_2(\Delta X) &= -(A + \Delta A)^H \Theta^H \Delta X \Theta (G + \Delta G) \Delta X \Theta [I + (G + \Delta G) \Delta X \Theta]^{-1} (A + \Delta A) \\ &= \operatorname{diag}(h_{22}(\Delta X), h_{23}(\Delta X), h_{21}(\Delta X)) \end{aligned} \tag{A.14a}$$

with

$$\begin{aligned} h_{2j}(\Delta X) = &-(A_j+\Delta A_j)^H\Theta_j^H\Delta X_j\Theta_j(G_j+\Delta G_j)\Delta X_j\Theta_j \\ &\cdot[I+(G_j+\Delta G_j)\Delta X_j\Theta_j]^{-1}(A_j+\Delta A_j) \\ &\in \mathcal{H}_n \quad \text{for } j=1,2,3. \end{aligned} \tag{A.14b}$$

Consequently, from (A.8), (A.9), (A.10a), (A.11a), (A.13a), and (A.14a), we obtain the perturbation equation

$$\Delta X_{j-1} - \Phi_j^H\Delta X_j\Phi_j = E_{1j} + E_{2j} + h_{1j}(\Delta X) + h_{2j}(\Delta X) \tag{A.15}$$

for $j=1,2,3$, where $E_{1j}, E_{2j}, h_{1j}(\Delta X)$, and $h_{2j}(\Delta X)$ are defined by (A.10b), (A.11b), (A.13b), and (A.14b), respectively.

By using the operators $\mathbf{L}, \mathbf{P}$, and $\mathbf{Q}$ (see (3.4), (3.6), and (3.7)), the perturbation equation (A.15) can be expressed by

$$\begin{aligned} (\Delta X_1, \Delta X_2, \Delta X_3) = &\mathbf{L}^{-1}(E_{12}, E_{13}, E_{11}) + \mathbf{L}^{-1}(E_{22}, E_{23}, E_{21}) \\ &+ \mathbf{L}^{-1}(h_{12}(\Delta X), h_{13}(\Delta X), h_{11}(\Delta X)) \\ &+ \mathbf{L}^{-1}(h_{22}(\Delta X), h_{23}(\Delta X), h_{21}(\Delta X)), \end{aligned} \tag{A.16}$$

where

$$\begin{aligned} \mathbf{L}^{-1}(E_{12}, E_{13}, E_{11}) = &\mathbf{L}^{-1}(\Delta H_2, \Delta H_3, \Delta H_1) + \mathbf{P}(\Delta A_2, \Delta A_3, \Delta A_1) \\ &- \mathbf{Q}(\Delta G_2, \Delta G_3, \Delta G_1). \end{aligned} \tag{A.17}$$

Define the function $\mu(\Delta X_1, \Delta X_2, \Delta X_3)$ on $\mathcal{H}_n^3$ by

$$\begin{aligned} \mu(\Delta X_1, \Delta X_2, \Delta X_3) = &\mathbf{L}^{-1}(E_{12}, E_{13}, E_{11}) + \mathbf{L}^{-1}(E_{22}, E_{23}, E_{21}) \\ &+ \mathbf{L}^{-1}(h_{12}(\Delta X), h_{13}(\Delta X), h_{11}(\Delta X)) \\ &+ \mathbf{L}^{-1}(h_{22}(\Delta X), h_{23}(\Delta X), h_{21}(\Delta X)), \end{aligned} \tag{A.18}$$

which can be regarded as a continuous mapping $\mathcal{M} : \mathcal{H}_n^3 \to \mathcal{H}_n^3$, and the set of solutions to (A.16) is just the set of fixed points of the mapping $\mathcal{M}$.

Step 2. *Estimates of some fixed points of* $\mathcal{M}$. Define ℓ, p_d, q_d and φ by (3.8) and (3.9), respectively. Note that the quantity ℓ can be equivalently defined by

$$\ell = \min_{\substack{(W_1,W_2,W_3)\in\mathcal{H}_n^3 \\ \|(W_1,W_2,W_3)\|_F=1}} \|\mathbf{L}(W_1,W_2,W_3)\|_F.$$

Moreover,we define

$$\delta_j = \frac{\|\Delta A_j\|_2 + \|K_j\|_2\|\Delta G_j\|_2}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}, \quad j=1,2,3, \tag{A.19}$$

where $\Delta A_j, \Delta G_j, K_j, \Psi_j$ are defined by (3.3) and (3.2). Here we assume that $\|\Delta G_j\|_2$ satisfy

$$1-\|\Psi_j\|_2\|\Delta G_j\|_2 > 0, \quad j=1,2,3. \tag{A.20}$$

Observe the following facts.

(I) By(A.17), we have

$$\|\mathbf{L}^{-1}(E_{12},E_{13},E_{11})\|_F \le \frac{1}{\ell}\|(\Delta H_1,\Delta H_2,\Delta H_3)\|_F + p_d\|(\Delta A_1,\Delta A_2,\Delta A_3)\|_F + q_d\|(\Delta G_1,\Delta G_2,\Delta G_3)\|_F \equiv \epsilon_1. \tag{A.21}$$

(II) By (A.11b), we have

$$\begin{aligned}
&\|(E_{22},E_{23},E_{21})\|_F = \sqrt{\sum_{j=1}^{3}\|E_{2j}\|_F^2}\\
&\le \left\{\sum_{j=1}^{3}\Big[\|\Psi_j\|_2\|\Delta A_j\|_2\|\Delta A_j\|_F + \frac{\|K_j\|_2^2\|\Psi_j\|_2\|\Delta G_j\|_2\|\Delta G_j\|_F}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right.\\
&\quad + \frac{\|K_j\|_2\|\Psi_j\|_2\|\Delta A_j\|_2\|\Delta G_j\|_F}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\\
&\quad \left.+ \frac{\|\Psi_j\|_2(\|K_j\|_2+\|\Psi_j\|_2\|\Delta A_j\|_2)\|\Delta G_j\|_2\|\Delta A_j\|_F}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\Big]^2\right\}^{\frac12}\\
&= \left\{\sum_{j=1}^{3}\|\Psi_j\|_2^2\left[\frac{(\|\Delta A_j\|_2+\|K_j\|_2\|\Delta G_j\|_2)(\|\Delta A_j\|_F+\|K_j\|_2\|\Delta G_j\|_F)}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right]^2\right\}^{\frac12}\\
&= \left\{\sum_{j=1}^{3}\|\Psi_j\|_2^2\delta_j^2(\|\Delta A_j\|_F+\|K_j\|_2\|\Delta G_j\|_F)^2\right\}^{\frac12} \equiv \epsilon_2.
\end{aligned} \tag{A.22}$$

(III) By (A.13b), we have

$$\|(h_{12}(\Delta X),h_{13}(\Delta X),h_{11}(\Delta X))\|_F \le \left\{\sum_{j=1}^{3}(2\|\Phi_j\|_2\|\Delta\Phi_j\|_2+\|\Delta\Phi_j\|_2^2)\|\Delta X_j\|_F^2\right\}^{\frac12},$$

and, by (A.12b), we have

$$\|\Delta\Phi_j\|_2 \le \frac{\|F_j\|_2(\|\Delta A_j\|_2+\|K_j\|_2\|\Delta G_j\|_2)}{1-\|\Psi_j\|_2\|\Delta G_j\|_2} = \|F_j\|_2\delta_j,$$

where δ_j $(j=1,2,3)$ are defined by (A.19). Thus we have

$$\|(h_{12}(\Delta X),h_{13}(\Delta X),h_{11}(\Delta X))\|_F \le \left\{\sum_{j=1}^{3}\|F_j\|_2^2\delta_j^2(2\|\Phi_j\|_2+\|F_j\|_2\delta_j)^2\|\Delta X_j\|_F^2\right\}^{\frac12}. \tag{A.23}$$

(IV) By (A.14b), we have

$$\begin{aligned}
&\|(h_{22}(\Delta X),h_{23}(\Delta X),h_{21}(\Delta X))\|_F\\
&\le \left\{\sum_{j=1}^{3}\left[\frac{\|\Theta_j\|_2^3(\|A_j\|_2+\|\Delta A_j\|_2)^2(\|G_j\|_2+\|\Delta G_j\|_2)\|\Delta X_j\|_F^2}{1-(\|G_j\|_2+\|\Delta G_j\|_2)\|\Theta_j\|_2\|\Delta X_j\|_F}\right]^2\right\}^{\frac12}.
\end{aligned}$$

Observe that, by (A.7),

$$\|\Theta_j\|_2 \le \frac{\|F_j\|_2}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}, \quad j=1,2,3.$$

Hence we have

$$\begin{aligned}
&\|(h_{22}(\Delta X), h_{23}(\Delta X), h_{21}(\Delta X))\|_F \\
&\le \left\{\sum_{j=1}^{3}\left[\frac{\|F_j\|_2^3(\|A_j\|_2+\|\Delta A_j\|_2)^2(\|G_j\|_2+\|\Delta G_j\|_2)\|\Delta X_j\|_F^2}{(1-\|\Psi_j\|_2\|\Delta G_j\|_2)^3\left[1-\frac{\|1-\|F_j\|_2(\|G_j\|_2+\|\Delta G_j\|_2)\|\Delta X_j\|_F}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right]}\right]^2\right\}^{\frac{1}{2}} \\
&\le \left\{\sum_{j=1}^{3}\left(\frac{\alpha^2\gamma\|\Delta X_j\|_F^2}{1-\gamma\|\Delta X_j\|_F}\right)^2\right\}^{\frac{1}{2}}, \qquad \text{(A.24)}
\end{aligned}$$

where α and γ are defined by

(A.25)
$$\alpha = \max_{1\le j\le 3}\left\{\frac{\|F_j\|_2(\|A_j\|_2+\|\Delta A_j\|_2)}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right\}, \quad \gamma = \max_{1\le j\le 3}\left\{\frac{\|F_j\|_2(\|G_j\|_2+\|\Delta G_j\|_2)}{1-\|\Psi_j\|_2\|\Delta G_j\|_2}\right\},$$

and it is assumed that

$$1-\gamma\|\Delta X_j\|_F > 0, \quad j=1,2,3. \qquad \text{(A.26)}$$

Consequently, from (A.18), (A.21)–(A.24), the function $\mu(\Delta X_1, \Delta X_2, \Delta X_3)$ satisfies

$$\begin{aligned}
\|\mu(\Delta X_1,\Delta X_2,\Delta X_3)\|_F \le \epsilon_1 + \frac{\epsilon_2}{\ell} &+ \frac{1}{\ell}\left\{\sum_{j=1}^{3}(\|F_j\|_2^2\delta_j^2(2\|\Phi_j\|_2+\|F_j\|_2\delta_j)^2\|\Delta X_j\|_F^2\right\}^{\frac{1}{2}} \\
&+ \frac{1}{\ell}\left\{\sum_{j=1}^{3}\left(\frac{\alpha^2\gamma\|\Delta X_j\|_F^2}{1-\gamma\|\Delta X_j\|_F}\right)^2\right\}^{\frac{1}{2}}. \qquad \text{(A.27)}
\end{aligned}$$

Let

$$\epsilon = \epsilon_1 + \frac{\epsilon_2}{\ell}, \quad \zeta = \max_{1\le j\le 3}\{\|F_j\|_2\delta_j(2\varphi+\|F_j\|_2\delta_j)\}. \qquad \text{(A.28)}$$

Then, from (A.27) and (A.28), we have

$$\begin{aligned}
&\|\mu(\Delta X_1,\Delta X_2,\Delta X_3)\|_F \\
&\le \epsilon + \frac{1}{\ell}\left(\zeta\|(\Delta X_1,\Delta X_2,\Delta X_3)\|_F + \frac{\alpha^2\gamma\|(\Delta X_1,\Delta X_2,\Delta X_3)\|_F^2}{1-\gamma\|(\Delta X_1,\Delta X_2,\Delta X_3)\|_F}\right). \qquad \text{(A.29)}
\end{aligned}$$

Consider the equation

$$\xi = \epsilon + \frac{1}{\ell}\left(\zeta\xi + \frac{\alpha^2\gamma\xi^2}{1-\gamma\xi}\right),$$

that is,

$$\gamma(\ell-\zeta+\alpha^2)\xi^2 - (\ell-\zeta+\ell\gamma\epsilon)\xi + \ell\epsilon = 0. \qquad \text{(A.30)}$$

It can be verified that, if ϵ satisfies

$$\epsilon \le \frac{(\ell-\zeta)^2}{\ell\gamma\left(\ell-\zeta+2\alpha+\sqrt{(\ell-\zeta+2\alpha)^2-(\ell-\zeta)^2}\right)}, \tag{A.31}$$

then the positive scalar ξ_* expressed by

$$\xi_* = \frac{2\ell\epsilon}{\ell-\zeta+\ell\gamma\epsilon+\sqrt{(\ell-\zeta+\ell\gamma\epsilon)^2-4\ell\gamma(\ell-\zeta+\alpha^2)\epsilon}} \tag{A.32}$$

is a solution of (A.30).

It is known that the product space $\mathcal{H}_n^3$ with the Frobenius norm $\|\cdot\|_F$ is a Banach space. We now consider the set $S_{\xi_*} \subset \mathcal{H}_n^3$ defined by

$$S_{\xi_*} = \left\{(\Delta X_1, \Delta X_2, \Delta X_3) \in \mathcal{H}_n^3 : \|(\Delta X_1, \Delta X_2, \Delta X_3))\|_F \le \xi_*\right\}.$$

S_{ξ_*} is obviously a bounded closed convex set of $\mathcal{H}_n^3$. Moreover, from (A.29) and the fact that ξ_* is a solution of (A.30), we see that, if $(\Delta X_1, \Delta X_2, \Delta X_3) \in S_{\xi_*}$, then

$$\|\mu(\Delta X_1, \Delta X_2, \Delta X_3)\|_F \le \xi_*,$$

which shows that the continuous mapping $\mathcal{M}$ expressed by (A.18) maps S_{ξ_*} into S_{ξ_*}. Thus, by the Schauder fixed-point theorem, the mapping $\mathcal{M}$ has a fixed point $(\Delta X_1^*, \Delta X_2^*, \Delta X_3^*) \in S_{\xi_*}$. Note that, if the scalar ζ defined by (A.28) satisfies

$$\ell - \zeta > 0, \tag{A.33}$$

then any $(\Delta X_1, \Delta X_2, \Delta X_3) \in S_{\xi_*}$ satisfies the condition (A.26). In fact, for any $(\Delta X_1, \Delta X_2, \Delta X_3) \in S_{\xi_*}$, we have

$$\begin{aligned}
1-\gamma\|\Delta X_j\|_F &\ge 1-\gamma\|(\Delta X_1, \Delta X_2, \Delta X_3)\|_F \\
&\ge 1-\gamma\xi_* \ge 1-\gamma\frac{2\ell\epsilon}{\ell-\zeta+\ell\gamma\epsilon} \qquad \text{(by (A.32))} \\
&= \frac{\ell-\zeta-\ell\gamma\epsilon}{\ell-\zeta+\ell\gamma\epsilon} \ge \frac{\ell-\zeta-\dfrac{(\ell-\zeta)^2}{\ell-\zeta+2\alpha}}{\ell-\zeta+\ell\gamma\epsilon} \qquad \text{(by (A.31))} \\
&= \frac{2\alpha(\ell-\zeta)}{(\ell-\zeta+\ell\gamma\epsilon)(\ell-\zeta+2\alpha)} > 0.
\end{aligned}$$

Step 3. *The periodic matrix set* $\{X_j + \Delta X_j^*\}_{j=1}^3$. Let $\{X_j\}_{j=1}^3$ be the unique Hermitian p.s.d. solution set to the P-DARE (1.1), and let $(\Delta X_1^*, \Delta X_2^*, \Delta X_3^*) \in \mathcal{S}_{\xi^*}$ be the fixed point of the mapping $\mathcal{M}$ by (A.18). Let

$$\begin{aligned}
&X = \operatorname{diag}\{X_j\}_{j=1}^3, \quad \Delta X = \operatorname{diag}\{\Delta X_j\}_{j=1}^3, \\
&Y = X + \Delta X^* \equiv \operatorname{diag}\{Y_j\}_{j=1}^3.
\end{aligned} \tag{A.34}$$

Then the Hermitian matrix Y satisfies

$$Y - \widetilde{A}^H Y(I+\widetilde{G}Y)^{-1}\widetilde{A} - \widetilde{H} = 0. \tag{A.35}$$

We now rewrite (A.35) as

$$\begin{aligned}
&Y - \left[(I+\widetilde{G}Y)^{-1}\widetilde{A}\right]^H Y(I+\widetilde{G}Y)^{-1}\widetilde{A} \\
&= \widetilde{H} + \left[Y(I+\widetilde{G}Y)^{-1}\widetilde{A}\right]^H \widetilde{G}Y(I+\widetilde{G}Y)^{-1}\widetilde{A}.
\end{aligned} \tag{A.36}$$

Observe that

$$\begin{aligned}(I+\widetilde{G}Y)^{-1}\widetilde{A} &= [I+(G+\Delta G)(X+\Delta X^*)]^{-1}(A+\Delta A)\\ &= \Phi+\Phi_1,\end{aligned} \tag{A.37}$$

where $\Phi=(I+GX)^{-1}A$ is d-stable (by Lemma 2.2 and Theorem 1.1), and Φ_1 can be expressed by

$$\Phi_1 = F\left[\Delta A-\Omega(I+\Omega)^{-1}(A+\Delta A)\right] \tag{A.38}$$

with

$$\Omega = \Delta G\Psi + G\Delta X^*F + \Delta G\Delta X^*F. \tag{A.39}$$

A simple calculation gives $\Phi_1=\mathrm{cyc}\{\Phi_{1j}\}_{j=1}^3$, where

$$\begin{aligned}\Phi_{1j} &= F_j(I+\Delta G_j\|\Psi_j\|_2+G_j\Delta X_j^*F_j+\Delta G_j\Delta X_j^*F_j)^{-1}\\ &\quad\cdot(\Delta A_j-\Delta G_jK_j-G_j\Delta X_j^*\Phi_j-\Delta G_j\Delta X_j^*\Phi_j)\end{aligned} \tag{A.40}$$

for $j=1,2,3$, and

$$\begin{aligned}\|\Phi_{1j}\|_2 &\le \frac{\|F_j\|_2\left[\|\Delta A_j\|_2+\|K_j\|_2\|\Delta G_j\|_2+\|\Phi_j\|_2(\|G_j\|_2\|\Delta G_j\|_2)\xi_*\right]}{1-\left[\|\Psi_j\|_2\|\Delta G_j\|_2+\|F_j\|_2(\|G_j\|_2+\|\Delta G_j\|_2)\xi_*\right]}\\ &\le \frac{\|F_j\|_2\delta_j+\|\Phi_j\|_2\gamma\xi_*}{1-\gamma\xi_*}\quad\text{(by (A.19) and (A.25))},\end{aligned} \tag{A.41}$$

where it is assumed that

$$1-\gamma\xi_*>0. \tag{A.42}$$

Hence, by (A.41) and Lemma 2.4, if

$$\max_{1\le j\le p}\left\{\frac{\|F_j\|_2\delta_j+\|\Phi_j\|_2\gamma\xi_*}{1-\gamma\xi_*}\right\}<\frac{\ell}{\varphi+\sqrt{\varphi^2+\ell}}, \tag{A.43}$$

then $\{\Phi_j+\Phi_{1j}\}_{j=1}^3$ is pd-stable. Further, by Lemma 2.2, $\Phi+\Phi_1$ (i.e., $(I+\widetilde{G}Y)^{-1}\widetilde{A}$ by (A.37)) is d-stable.

We now consider the DARE (A.35) in the classical case. The matrix Y is a d-stabilizing solution to the DARE. By [8], the solution Y is unique. Moreover, the Hermitian matrix Y, as a solution to (A.36), is p.s.d. [8]. Thus we have proved that under the conditions (A.20), (A.31), (A.33), (A.42), and (A.43), there is a unique Hermitian p.s.d. solution set $\{X_j+\Delta X_j^*\}_{j=1}^3$ to the perturbed P-DARE (1.4). Note that the condition (A.33) can be deduced from the condition (A.43). In fact, from the inequality (A.43),

$$\|F_j\|_2\delta_j<\frac{\ell}{\varphi+\sqrt{\varphi^2+\ell}},\quad j=1,2,3,$$

which implies

$$2\varphi\|F_j\|_2\delta_j+(\|F_j\|_2\delta_j)^2<\ell \tag{A.44}$$

for $j=1,2,3$. Observe that, by (A.28) and (A.44), we obtain $\zeta<\ell$, that is, the condition (A.33).

Appendix B. (I) Proof of (3.24).

Let $\widehat{\mathbf{L}}$ be a natural extension of the linear operator $\mathbf{L}$ on $\mathbb{C}_n^p$, that is,

$$\widehat{\mathbf{L}}(Z_1,\dots,Z_p) = (Z_1 - \Phi_2^H Z_2 \Phi_2, \dots, Z_{p-1} - \Phi_p^H Z_p \Phi_p, Z_p - \Phi_1^H Z_1 \Phi_1), \tag{B.1}$$

where $Z_j \in \mathbb{C}_n$ for $j = 1,\dots,p$. Then the matrix L defined by (2.6) is a matrix representation of $\widehat{\mathbf{L}}$ on $\mathbb{C}^{pn^2}$. By the definition of the operator norm, we have

$$\begin{aligned} \| \mathbf{L}^{-1} \|^{-1} &= \min_{\substack{W_j \in \mathcal{H}_n \\ j=1,\dots,p}} \frac{\left(\sum_{j=1}^{p} \| W_{j-1} - \Phi_j^H W_j \Phi_j \|_F^2 \right)^{1/2}}{\left(\sum_{j=1}^{p} \| W_j \|_F^2 \right)^{1/2}} \\ &\geq \min_{\substack{Z_j \in \mathbb{C}_n \\ j=1,\dots,p}} \frac{\left(\sum_{j=1}^{p} \| Z_{j-1} - \Phi_j^H Z_j \Phi_j \|_F^2 \right)^{1/2}}{\left(\sum_{j=1}^{p} \| Z_j \|_F^2 \right)^{1/2}} \\ &= \|\widehat{\mathbf{L}}^{-1}\|^{-1} = \|L^{-1}\|_2^{-1}. \end{aligned} \tag{B.2}$$

We shall prove that the equality in (B.2) holds. Let $(Z_1^*,\dots,Z_p^*) \in \mathbb{C}_n^p$ be a singular "vector" such that

$$\|L^{-1}\|_2^{-1} = \frac{\left(\sum_{j=1}^{p} \| Z_{j-1}^* - \Phi_j^H Z_j^* \Phi_j \|_F^2 \right)^{1/2}}{\left(\sum_{j=1}^{p} \| Z_j^* \|_F^2 \right)^{1/2}}. \tag{B.3}$$

Then $(Z_1^{*H},\dots,Z_p^{*H})$ is also a singular "vector" satisfying (B.3). Let $W_j^* = Z_j^* + Z_j^{*H}$ for $j = 1,\dots,p$. Obviously, $W_1^*,\dots,W_p^*$ are Hermitian. Consequently, if $W_j^* \neq 0$ for some $j \in \{1,\dots,p\}$, then we have

$$\|L^{-1}\|_2^{-1} = \frac{\left(\sum_{j=1}^{p} \| W_{j-1}^* - \Phi_j^H W_j^* \Phi_j \|_F^2 \right)^{1/2}}{\left(\sum_{j=1}^{p} \| W_j^* \|_F^2 \right)^{1/2}} = \|\mathbf{L}^{-1}\|^{-1}. \tag{B.4}$$

If $W_j^* = 0$ for all $j = 1,\dots,p$, then $Z_j^* = -(Z_j^*)^H$. In such a case, $iZ_1^*,\dots,iZ_p^*$ are Hermitian, and we also have

$$\|L^{-1}\|_2^{-1} = \frac{\left(\sum_{j=1}^{p} \| iZ_{j-1}^* - \Phi_j^H (iZ_j^*) \Phi_j \|_F^2 \right)^{1/2}}{\left(\sum_{j=1}^{p} \| iZ_j^* \|_F^2 \right)^{1/2}} = \|\mathbf{L}^{-1}\|^{-1}. \tag{B.5}$$

For the real case, i.e., the case when all coefficient matrices are real, the relations (B.1)–(B.4) still hold, where we need only to replace $\mathcal{H}_n$, $\mathbb{C}_n$, and the superscript "H" by $\mathcal{S}_n$, $\mathbb{R}_n$, and the superscript "T," respectively. However, the equality (B.5) no longer holds because the matrix iZ_j^* is Hermitian but not real symmetric. In order to remedy this defect, we first prove the following lemma.

LEMMA B.1. *Suppose that $\Phi_j \in \mathbb{R}_n (j = 1, \ldots, p)$, $\{\Phi_j\}_{j=1}^p$ is pd-stable, and $B_j, C_j \in \mathcal{S}_n$ (or $\mathcal{H}_n$) with $B_j \geq C_j$ (i.e., $B_j - C_j \geq 0$) for $j = 1, \ldots, p$. If X_j and Y_j satisfy*

$$X_{j-1} - \Phi_j^\top X_j \Phi_j = B_j, \quad j = 1, \ldots, p, \tag{B.6}$$

$$Y_{j-1} - \Phi_j^\top Y_j \Phi_j = C_j, \quad j = 1, \ldots, p, \tag{B.7}$$

respectively, then X_j, Y_j are real symmetric (or Hermitian) and $X_j \geq Y_j$ for $j = 1, \ldots, p$.

Proof. Let $\mathrm{vec}(X_j) = x_j, \mathrm{vec}(B_j) = b_j, \mathrm{vec}(Y_j) = y_j, \mathrm{vec}(C_j) = c_j$ for $j = 1, \ldots, p$. Then (B.6) and (B.7) can be written as

$$Lx = b \quad \text{and} \quad Ly = c, \tag{B.8}$$

where L is defined by (2.6), $x = (x_1^\top, \ldots, x_p^\top)^\top$, $b = (b_1^\top, \ldots, b_p^\top)^\top$, $y = (y_1^\top, \ldots, y_p^\top)^\top$, and $c = (c_1^\top, \ldots, c_p^\top)^\top$. By the assumption, $\lambda(L) \subset \mathcal{D}$, and L is invertible. For any $b_j \in \mathbb{R}^{n^2}$ (or $\mathbb{C}^{n^2}$) so that $B_j = \mathrm{unvec}(b_j) \in \mathcal{S}_n$ (or $\mathcal{H}_n$) (here unvec denotes the inverse operator of vec) for $j = 1, \ldots, p$, the solution $x = (x_1^\top, \ldots, x_p^\top)^\top$ in (B.8) is uniquely solvable. Let $X_j = \mathrm{unvec}(x_j)$, for $j = 1, \ldots, p$. Then $\{X_j\}_{j=1}^p$ satisfies (B.6). Taking the transpose (or conjugate transpose) of (B.6), it follows that $\{X_j^\top\}_{j=1}^p$ (or $\{X_j^H\}_{j=1}^p$) is also a solution set of (B.6). By the uniqueness, the solution set $\{X_j\}_{j=1}^p$ of (B.6) satisfies $X_j = X_j^\top \in \mathcal{S}_n$ (or $X_j = X_j^H \in \mathcal{H}_n$) for $j = 1, \ldots, p$. Similarly, the solution set $\{Y_j\}_{j=1}^p$ of (B.7) satisfies $Y_j = Y_j^\top \in \mathcal{S}_n$ (or $Y_j = Y_j^H \in \mathcal{H}_n$) for $j = 1, \ldots, p$. Denote

$$\begin{aligned} &\Phi = \mathrm{cyc}\{\Phi_j\}_{j=1}^p, \quad X = \mathrm{diag}\{X_j\}_{j=1}^p, \quad Y = \mathrm{diag}\{Y_j\}_{j=1}^p, \\ &B = \mathrm{diag}\{B_j\}_{j=1}^p, \quad \text{and} \quad C = \mathrm{diag}\{C_j\}_{j=1}^p. \end{aligned} \tag{B.9}$$

Subtracting (B.6) from (B.7), we have

$$(X - Y) - \Phi^\top (X - Y)\Phi = B - C \geq 0. \tag{B.10}$$

Applying Proposition 2.1 of [8] to (B.10), we obtain $X_j \geq Y_j$ for $j = 1, \ldots, p$. ∎

Now suppose that $(Z_1^*, \ldots, Z_p^*)$ is the singular "vector" satisfying (B.3), where $Z_j^* = -Z_j^{*\top}$ are $n \times n$ real skew-symmetric matrices for $j = 1, \ldots, p$. Let

$$N_j = Z_{j-1}^* - \Phi_j^\top Z_j^* \Phi_j \quad \text{(real skew-symmetric)} \tag{B.11}$$

for $j = 1, \ldots, p$, and let $N_j = U_j D_j U_j^\top$ be the orthogonal spectral decomposition such that D_j is block diagonal with 1×1-zero blocks and 2×2-blocks

$$D_{j,ii} = \begin{bmatrix} 0 & \lambda_i \\ -\lambda_i & 0 \end{bmatrix}$$

for $j = 1, \ldots, p$. According to a technique developed by Byers and Nash [4], we construct the symmetric matrices

$$M_j = U_j E_j U_j^\top, \quad j = 1, \ldots, p, \tag{B.12}$$

where E_j is the block diagonal matrix with the same 1×1-zero block as D_j and

$$E_{j,ii} = \begin{bmatrix} \lambda_i & 0 \\ 0 & \lambda_i \end{bmatrix}$$

provided that $D_{j,ii}$ is of the form

$$\begin{bmatrix} 0 & \lambda_i \\ -\lambda_i & 0 \end{bmatrix}.$$

It is easy to see that

$$M_j \geq iN_j \geq -M_j, \quad j = 1, \ldots, p. \tag{B.13}$$

Let $\{W_j\}_{j=1}^p$ ($W_j \in \mathcal{S}_n$) be the symmetric solution set satisfying

$$W_{j-1}^* - \Phi_j^\top W_j^* \Phi_j = M_j, \quad j = 1, \ldots, p. \tag{B.14}$$

Applying Lemma B.1 to (B.14) and

$$iZ_{j-1}^* - \Phi_j^\top (iZ_j^*)\Phi_j = iN_j \quad \text{(by (B.11))}$$

and by using (B.13), we obtain $W_j^* \geq iZ_j^* \geq -W_j^*$. Hence, by Lemma 7 of [4], we get

$$\|W_j^*\|_F \geq \|Z_j^*\|_F, \quad j = 1, \ldots, p. \tag{B.15}$$

By (B.3), (B.11), and (B.15),

$$\begin{aligned} \|L^{-1}\|_2^{-1} &\geq \frac{\left(\sum_{j=1}^p \|N_j\|_F^2\right)^{1/2}}{\left(\sum_{j=1}^p \|W_j^*\|_F^2\right)^{1/2}} = \frac{\left(\sum_{j=1}^p \|M_j\|_F^2\right)^{1/2}}{\left(\sum_{j=1}^p \|W_j^*\|_F^2\right)^{1/2}} = \frac{\left(\sum_{j=1}^p \|W_{j-1}^* - \Phi_j^\top W_j^* \Phi_j\|_F^2\right)^{1/2}}{\left(\sum_{j=1}^p \|W_j^*\|_F^2\right)^{1/2}} \\ &= \|\mathbf{L}^{-1}\|^{-1} \quad \text{(by (B.12) and (B.14))}, \end{aligned}$$

which shows that, in the real case and when $W_j^* = 0$ for all $j = 1, \ldots, p$, the equality $\|\mathbf{L}^{-1}\|^{-1} = \|L^{-1}\|_2^{-1}$ also holds.

(II) Proof of (3.25) and (3.26).

The complex case. Since $K_j^H N_j + N_j^H K_j \in \mathcal{H}_n$ for any $N_j \in \mathbb{C}_n$, we have $\mathbf{L}^{-1}(H_2, \ldots, H_p, H_1) = \widehat{\mathbf{L}}^{-1}(H_2, \ldots, H_p, H_1)$, where $H_j = K_j^H N_j + N_j^H K_j \in \mathcal{H}_n$. By the definition of the operator norm, we have

$$\begin{aligned} \|\mathbf{P}\| &= \max_{\substack{N_j \in \mathbb{C}_n \\ (N_1, \ldots, N_p) \neq 0}} \frac{\|\mathbf{P}(N_1, \ldots, N_p)\|_F}{\|(N_1, \ldots, N_p)\|_F} \\ &= \max_{\substack{N_j \in \mathbb{C}_n \\ (N_1, \ldots, N_p) \neq 0}} \frac{\|\widehat{\mathbf{L}}^{-1}(K_2^H N_2 + N_2^H K_2, \ldots, K_p^H N_p + N_p^H K_p, K_1^H N_1 + N_1^H K_1)\|_F}{\|(N_1, \ldots, N_p)\|_F} \\ &= \max_{\substack{\mathrm{vec}(N_j) = z_j \in \mathbb{C}^{n^2} \\ (z_1^\top, \ldots, z_p^\top)^\top \neq 0}} \frac{\left\| L^{-1} \begin{bmatrix} (I \otimes K_2^H) z_2 + (K_2^\top \otimes I)\Pi \overline{z}_2 \\ \vdots \\ (I \otimes K_p^H) z_p + (K_p^\top \otimes I)\Pi \overline{z}_p \\ (I \otimes K_1^H) z_1 + (K_1^\top \otimes I)\Pi \overline{z}_1 \end{bmatrix} \right\|_2}{\sqrt{\sum_{j=1}^p \|z_j\|_2^2}}. \end{aligned} \tag{B.16}$$

Denote $z_j = x_j + iy_j, x_j, y_j \in \mathbb{R}^{n^2}$, for $j = 1, \ldots, p$, and $x = (x_1^\top, \ldots, x_p^\top)^\top, y = (y_1^\top, \ldots, y_p^\top)^\top$. Let

$$L^{-1}\big[\mathrm{cyc}\big\{(I \otimes K_j^H)^\top\big\}_{j=1}^p\big]^\top = \Omega_1 + i\Omega_2, \tag{B.17}$$

$$L^{-1}\big[\mathrm{cyc}\big\{\big[(K_j^\top \otimes I)\Pi\big]^\top\big\}_{j=1}^p\big]^\top = \Theta_1 + i\Theta_2, \tag{B.18}$$

where $\Omega_1, \Omega_2, \Theta_1$, and Θ_2 are real matrices. By (B.16), (B.17), and the technique proposed by [14], we have

$$\begin{aligned}\|\mathbf{P}\| &= \max_{[x^\top, y^\top]^\top \neq 0} \frac{\|(\Omega_1 + i\Omega_2)(x+iy) + (\Theta_1 + i\Theta_2)(x-iy)\|_2}{\sqrt{\|x\|_2^2 + \|y\|_2^2}} \\ &= \max_{[x^\top, y^\top]^\top \neq 0} \frac{\left\| \begin{bmatrix} \Omega_1 + \Theta_1 & \Theta_2 - \Omega_2 \\ \Omega_2 + \Theta_2 & \Omega_1 - \Theta_1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2}{\sqrt{\|x\|_2^2 + \|y\|_2^2}} = \left\| \begin{bmatrix} \Omega_1 + \Theta_1 & \Theta_2 - \Omega_2 \\ \Omega_2 + \Theta_2 & \Omega_1 - \Theta_1 \end{bmatrix} \right\|_2.\end{aligned}$$

The real case. By replacing $\mathcal{H}_n, \mathbb{C}_n$ and the superscript "H" by $\mathbf{S}_n, \mathbb{R}_n$, and the superscript "$T$," respectively, (B.16) becomes

$$\|\mathbf{P}\| = \left\| L^{-1}\Big[\mathrm{cyc}\Big\{\Big[I \otimes K_j^\top + (K_j^\top \otimes I)\Pi\Big]^\top\Big\}_{j=1}^p\Big]^\top \right\|_2.$$

(III) Proof of (3.27).
Obviously,

$$\begin{aligned}\|\mathbf{Q}\| &= \max_{\substack{M_j \in \mathcal{H}_n \\ (M_1, \ldots, M_p) \neq 0}} \frac{\|\mathbf{L}^{-1}(K_2^H M_2 K_2, \ldots, K_p^H M_p K_p, K_1^H M_1 K_1)\|_F}{\|(M_1, \ldots, M_p)\|_F} \\ &\leq \max_{\substack{N_j \in \mathbb{C}_n \\ (N_1, \ldots, N_p) \neq 0}} \frac{\|\widehat{\mathbf{L}}^{-1}(K_2^H N_2 K_2, \ldots, K_p^H N_p K_p, K_1^H N_1 K_1)\|_F}{\|(N_1, \ldots, N_p)\|_F} \\ &= \frac{\|\widehat{\mathbf{L}}^{-1}(K_2^H N_2^* K_2, \ldots, K_p^H N_p^* K_p, K_1^H N_1^* K_1)\|_F}{\|(N_1^*, \ldots, N_p^*)\|_F}. \qquad \text{(B.19)}\end{aligned}$$

Let $(Z_1^*, \ldots, Z_p^*) = \widehat{\mathbf{L}}^{-1}(K_2^H N_2^* K_2^*, \ldots, K_p^H N_p^* K_p, K_1^H N_1^* K_1)$. By the definition (B.1) of $\widehat{\mathbf{L}}$, we have

$$\begin{aligned}&(Z_1^*, \ldots, Z_p^*) - (\Phi_2^H Z_2^* \Phi_2, \ldots, \Phi_p^H Z_p^* \Phi_p, \Phi_1^H Z_1^* \Phi_1) \\ &= (K_2^H N_2^* K_2, \ldots, K_p^H N_p^* K_p, K_1^H N_1^* K_1),\end{aligned}$$

which implies

$$\begin{aligned}&(Z_1^{*H}, \ldots, Z_p^{*H}) - (\Phi_2^H Z_2^{*H} \Phi_2, \ldots, \Phi_p^H Z_p^{*H} \Phi_p, \Phi_1^H Z_1^{*H} \Phi_1) \\ &= (K_2^H N_2^{*H} K_2, \ldots, K_p^H N_p^{*H} K_p, K_1^H N_1^{*H} K_1).\end{aligned}$$

Thus we have

$$(Z_1^{*H}, \ldots, Z_p^{*H}) = \widehat{\mathbf{L}}^{-1}(K_2^H N_2^{*H} K_2, \ldots, K_p^H N_p^{*H} K_p, K_1^H N_1^{*H} K_1).$$

By the fact that $\|(Z_1^*, \dots, Z_p^*)\|_F = \|(Z_1^{*H}, \dots, Z_p^{*H})\|_F$, we obtain

$$\begin{aligned}&\|\widehat{\mathbf{L}}^{-1}(K_2^H N_2{}^H K_2, \dots, K_p^H N_p{}^H K_p, K_1^H N_1{}^H K_1)\|_F\\ &= \|\widehat{\mathbf{L}}^{-1}(K_2^H N_2^{*H} K_2, \dots, K_p^H N_p^{*H} K_p, K_1^H N_1^{*H} K_1)\|_F.\end{aligned}$$

We now define the operator $\widehat{\mathbf{Q}}$ on $\mathbb{C}_n^p$ by

$$\widehat{\mathbf{Q}}(N_1, \dots, N_p) = \widehat{\mathbf{L}}^{-1}(K_2^H N_2 K_2, \dots, K_p^H N_p K_p, K_1^H N_1 K_1). \tag{B.20}$$

It is easy to see that the matrix

$$Q = L^{-1}\big[\mathrm{cyc}\big\{K_j \otimes \overline{K}_j\big\}_{j=1}^p\big]^\top$$

is a matrix representation of $\widehat{\mathbf{Q}}$. Combining (B.19) with (B.20) shows that $(N_1{}^*, \dots, N_p{}^*)$ and $(N_1^{*H}, \dots, N_p^{*H})$ are the singular "vectors" of $\widehat{\mathbf{Q}}$ corresponding to its largest singular value. Let

$$W_j^* = Z_j^* + Z_j^{*H} \quad \text{for} \quad j = 1, \dots, p.$$

If $W_j^* \neq 0$ for some $j \in \{1, \dots, p\}$, then $(W_1^*, \dots, W_p^*)$ is also a singular "vector" of $\widehat{\mathbf{Q}}$ corresponding to its largest singular value. Hence we have

$$\begin{aligned}\|Q\|_2 &= \frac{\|\widehat{\mathbf{L}}^{-1}(K_2^H W_2^* K_2, \dots, K_p^H W_p^* K_p, K_1^H W_1^* K_1)\|_F}{\|(W_1^*, \dots, W_p^*)\|_F}\\ &= \frac{\|\mathbf{L}^{-1}(K_2^H W_2^* K_2, \dots, K_p^H W_p^* K_p, K_1^H W_1^* K_1)\|_F}{\|(W_1^*, \dots, W_p^*)\|_F} = \|\mathbf{Q}\|.\end{aligned} \tag{B.21}$$

If $W_j^* = 0$ for all $j = 1, \dots, p$, then set $H_j^* \equiv iZ_j^* \in \mathcal{H}_n$ for $j = 1, \dots, p$, and thus $(H_1^*, \dots, H_p^*)$ is also a singular "vector" of $\widehat{\mathbf{Q}}$ corresponding to $\|Q\|_2$. Hence, we have

$$\begin{aligned}\|Q\|_2 &= \frac{\|\widehat{\mathbf{L}}^{-1}(K_2^H H_2^* K_2, \dots, K_p^H H_p^* K_p, K_1^H H_1^* K_1)\|_F}{\|(H_1^*, \dots, H_p^*)\|_F}\\ &= \frac{\|\mathbf{L}^{-1}(K_2^H H_2^* K_2, \dots, K_p^H H_p^* K_p, K_1^H H_1^* K_1)\|_F}{\|(H_1^*, \dots, H_p^*)\|_F} = \|\mathbf{Q}\|.\end{aligned} \tag{B.22}$$

From (B.21) and (B.22) we obtain $\|\mathbf{Q}\| = \|Q\|_2$.

Similarly, we can prove the expression (3.27) in the real case.

REFERENCES

[1] 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.

[2] S. Bittanti, P. Colaneri, and G. de Nicolao, *The periodic Riccati equation*, in The Riccati Equation, S. Bittanti, A. Laub, and J. Willems, eds., Springer-Verlag, Berlin, 1991, pp. 127–162.

[3] A. Bojanczyk, G. Golub, and P. Van Dooren, *The periodic Schur decomposition, algorithm and applications,* in Proceedings of the SPIE Conference San Diego, Vol. 1770, SPIE, Bellingham, WA, 1992, pp. 31–42.

[4] R. Byers and S. Nash, *On the singular "vectors" of the Lyapunov operator*, SIAM J. Alg. Disc. Meth., 8 (1987), pp. 59–66.

[5] C. E. De Souza, *Periodic strong solution for the optimal filtering problem of liner discrete-time periodic systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 333–357.
[6] W. R. Ferng, W. W. Lin, and C. S. Wang, *On Computing the Stable Invariant Subspace of Cyclic Symplectic Pairs*, Technical report TR9606, Department of Applied Mathematics, Tsinghua University, Taiwan, 1996; Linear Algebra Appl., to appear.
[7] B. Francis and T. T. Georgiou, *Stability theory for linear time-invariant plants with periodic digital controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 820–832.
[8] P. M. Gahinet, A. J. Laub, C. S. Kenney, and G. A. Hewer, *Sensitivity of the stable discrete-time Lyapunov equation*, IEEE Trans. Automat. Control, 35 (1990), pp. 1209–1217.
[9] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., John Hopkins University Press, Baltimore, London, 1996.
[10] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.
[11] T. Gudmundsson, C. Kenney, and A. J. Laub, *Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method*, IEEE Trans. Automat. Control, 37 (1992), pp. 513–518.
[12] H. Kano and T. Nishimura, *Periodic solutions of matrix Riccati equations with detectability and stabilizability*, Internat. J. Control, 29 (1979), pp. 1197–1210.
[13] M. Konstantinov, P. Petkov, and N. D. Christov, *Perturbation analysis of the discrete Riccati equation*, Kybernetika (Prague), 29 (1993), pp. 18–29.
[14] M. Konstantinov and P. Petkov, *Note on "Perturbation theory for algebraic Riccati equations,"* SIAM J. Matrix Anal. Appl., 21 (1999), p. 327.
[15] P. Misra, *Time invariant representation of discrete periodic systems*, Automatica J. IFAC, 32 (1996), pp. 267–272.
[16] C. Paige and C. Van Loan, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 41 (1981), pp. 11–32.
[17] T. Pappas, A. J. Laub, and N. R. Sandell, Jr., *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, 25 (1980), pp. 631–641.
[18] B. Park and E. Verriest, *Canonical forms for discrete linear periodically time-varying systems and application to control*, in Proceedings of the 28th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1989, pp. 1220–1225.
[19] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.
[20] J.-G. Sun, *Perturbation theory for algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 39–65.
[21] J.-G. Sun, *Condition numbers of algebraic Riccati equations in the Frobenius norm*, Linear Algebra Appl., 350 (2002), pp. 237–261.
[22] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[23] P. Van Dooren and J. Sreedhar, *When is a periodic discrete-time system equivalent to a time invariant one?*, Linear Algebra Appl., 162 (1992), pp. 685–708.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 439–453

N ROOTS OF THE SECULAR EQUATION IN $O(N)$ OPERATIONS*

OREN E. LIVNE† AND ACHI BRANDT†

Abstract. We present a novel multilevel algorithm which computes N roots of the secular equation in $O(CN)$ computer operations, where C depends on the desired accuracy. Since current methods of solution require $O(N^2)$ operations, this algorithm can drastically reduce the computational effort in various applications, including updating the singular value decomposition and symmetric eigenvalue problems and solving constrained least-squares problems. The algorithm is based on the multilevel approach for fast evaluation of integral transforms. It has been adapted for the efficient solution of the secular equation. We have also incorporated discontinuous kernel softening, a technique which improves the implementation of multilevel summation algorithms toward theoretical optimality. We present and discuss numerical results, parallelization, and other related applications of the multilevel approach, including a possible substitute for current symmetric tridiagonal eigenbasis solvers (such as the divide and conquer method).

Key words. secular equation, fast multilevel summation, root-search

AMS subject classifications. 15A18, 65F15, 65H17, 65R10, 65R20, 65Y05, 65Y20, 68Q25

PII. S0895479801383695

1. The secular equation.

We consider the computational task of finding all of the roots $\{\lambda_k^*\}_{k=1}^N$ of the secular equation

$$f(\lambda) := 1 + \sigma v(\lambda) = 0, \qquad v(\lambda) := \sum_{k=1}^{N} \frac{u_k}{d_k - \lambda}, \tag{1.1}$$

which are strictly separated by the values $\{d_k\}_{k=1}^N$, namely [20, 25, 32],

$$d_1 < \lambda_1^* < d_2 < \lambda_2^* < \cdots < d_N < \lambda_N^* < d_N + \sigma \sum_{k=1}^{N} u_k^2, \tag{1.2}$$

assuming $d_1 < d_2 < \cdots < d_N$ are real, $\sigma > 0$, and $u_k > 0$ for all k. This problem has various applications in numerical linear algebra, such as

1. updating the singular value decomposition of matrices [1, 10],
2. modifying the symmetric eigenvalue problem [11, 14, 15, 21, 24, 25, 27],
3. solving constrained least-squares-type problems [13, 17, 19, 20, 23, 28, 36, 37, 44],
4. computing the eigenvalues of a matrix using the escalator method [18], and
5. invariant subspace computations [16].

A thorough literature survey may be found in [32, 33, 34, 35].

1.1. Current methods of solution.

Secular equations are often a "subproblem of a larger one" [34], as in the divide and conquer method [26, 21]. Consequently, they "typically have to be solved to high accuracy many times, which requires fast

*Received by the editors January 13, 2001; accepted for publication (in revised form) by G. H. Golub April 5, 2002; published electronically November 6, 2002. This research was supported by grant 696/97 from the Israel Science Foundation, by AFOSR contract F33615-97-D5405, and by the Carl F. Gauss Minerva Center for Scientific Computation at The Weizmann Institute of Science.
http://www.siam.org/journals/simax/24-2/38369.html

†Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel (livneo@wisdom.weizmann.ac.il, achi@wisdom.weizmann.ac.il).

and stable methods" [34]. Many root-searching algorithms for solving (1.1) have been extensively studied and developed; among these are the following:

1. the quadratic BNS methods [10, 11, 38], based on a rational interpolation,
2. Melman's methods [32, 33, 34, 35], which use a change of coordinates transforming the original equation into an equivalent problem for which Newton's method exhibits global quadratic convergence,
3. Gragg's third order zero-finder [24] and other high order methods [34, 35].

These methods (e.g., Melman's) can compute any root of (1.1) to machine accuracy using a small number of direct evaluations of v and its derivative ($O(\log(\log(1/\varepsilon)))$ iterations are needed to obtain an ε-accuracy). Since each such evaluation costs $O(N)$ operations, N roots are computed in $O(N^2)$.

1.2. Objectives. Our goal is to design a linear complexity algorithm for computing N roots of (1.1) in only $O(CN)$ operations, where C depends on the desired accuracy ε, $C = O((\log(1/\varepsilon))^q)$ for some small $q \in \mathbb{R}_+$. This is achieved in a two-stage procedure:

(a) designing an algorithm for *evaluating* v at N values of λ in $O(N)$ operations,
(b) adapting this fast evaluation to the solution of (1.1) in $O(N)$, using any of the root-search methods mentioned in section 1.1.

Both stages are handled efficiently and naturally by the multilevel approach presented in [5]. In section 2, we present our fast multilevel evaluation algorithm (stage (a)) for uniformly dense $\{d_k\}_{k=1}^N$. Section 3 discusses the fast solution of (1.1) (stage (b)). We conclude in section 4 by discussing nonuniform density, generalizations, parallelization, and other related applications of the multilevel approach, including a possible substitute for current symmetric tridiagonal eigenbasis solvers (such as the divide and conquer method [14, 21]).

2. Fast evaluation of $v(\lambda)$. A necessary stage toward the fast solution of (1.1) is the fast evaluation of v. Let $\{\lambda_j\}_{j=1}^N$ be any sequence satisfying (1.2) (e.g., approximations to $\{\lambda_j^*\}_{j=1}^N$ at a certain root-searching step); we wish to calculate

$$v(\lambda_j) = \sum_{k=1}^{N} G(d_k - \lambda_j)u(d_k), \quad j = 1, \dots N, \quad u(d_k) := u_k, \quad G(r) := \frac{1}{r} \tag{2.1}$$

in $O(N)$ operations. The algorithm for computing $\{v'(\lambda_j)\}_{j=1}^N$, if desired, is discussed in section 3.2. For simplicity, let us first assume that $\{d_k\}_{k=1}^N$ have a uniform density α; i.e., it is possible to place a uniform grid $\{D_K^1\}_{K=1}^{N_1}$ with meshsize H over $[d_1, d_N]$ so that, in each interval $[D_K^1, D_{K+1}^1]$, there lies a uniformly bounded number (about $\alpha H =: m$) of d_ks. The interlacement property (1.2) implies that $\{\lambda_j\}_{j=1}^N$ are also uniformly dense (for nonuniform densities, see section 4.1).

Our algorithm is a straightforward application of the general multilevel approach for fast evaluation of integral transforms with asymptotically smooth kernels, which is described in detail in [5, 7, 8, 9]. We also incorporate a technical modification (discontinuous softening) that improves the work-accuracy relation of multilevel summation algorithms toward optimality. This may be of interest in practical implementations.

2.1. Kernel softening. The kernel $G(r) = 1/r$ is *asymptotically smooth*, that is, increasingly smooth for larger r. As in [5, 7, 9], it can be decomposed into

$$G(r) = G_S(r) + G_{\text{local}}(r) \tag{2.2}$$

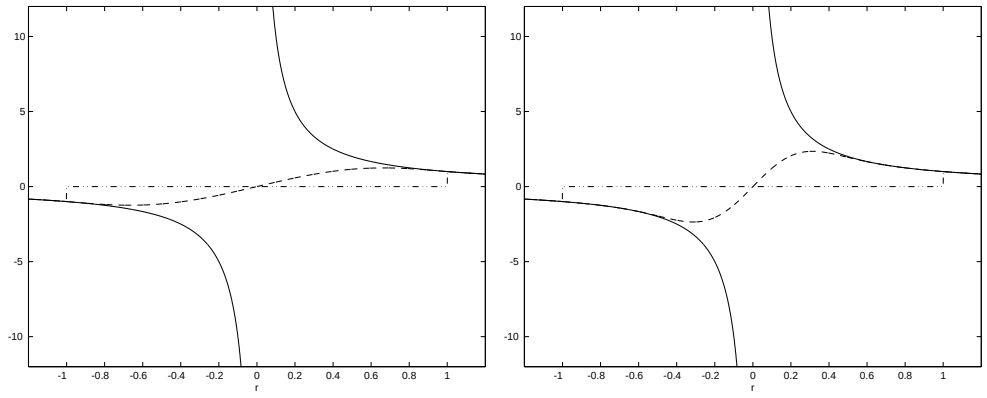


FIG. 1. *The kernel $G(r) = 1/r$ (solid line) and its softenings, $G^*(r)$ (dashed line) and $\tilde{G}(r)$ (dashed-dotted line), for $p = 2$ (left picture) and $p = 12$ (right picture).*

so that

(i) $G_S(r) = G(r)$ (or $G_{\text{local}}(r) = 0$) for all $|r| \geq S$,

(ii) G_S is *suitably smooth* on the scale S; namely, for any $\varepsilon > 0$, there exists $p = O(\log(1/\varepsilon)) \in \mathbb{N}$ such that G_S can be uniformly approximated to an accuracy ε by a p-order interpolation from its values on any uniform grid with a meshsize comparable with S [9].

Traditional multilevel algorithms [5, 8, 9, 40] used a polynomial softened kernel

$$G_S(r) = \frac{1}{S} G^*\left(\frac{r}{S}\right), \qquad G^*(r) := \begin{cases} \sum_{n=0}^{2p-1} a_n r^n, & |r| \leq 1, \\ G(r), & |r| \geq 1, \end{cases} \tag{2.3}$$

which fits $G, G', \ldots, G^{(p)}$ at $r = \pm S$. In this paper, we propose a novel *piecewise* smooth kernel softening in the form

$$G_S(r) = \frac{1}{S} \tilde{G}\left(\frac{r}{S}\right), \qquad \tilde{G}(r) := \begin{cases} 0, & |r| \leq 1, \\ G(r), & |r| > 1, \end{cases} \tag{2.4}$$

which is suitably smooth only for $r \in \mathbb{R}\backslash\{-S, S\}$. Nevertheless, the discontinuous softening (2.4) has the following advantages over the continuous softening (2.3).

1. The derivative $\tilde{G}^{(p)}(r)$ vanishes for $|r| \leq 1$; hence its magnitude is certainly less than $(G^*)^{(p)}(r)$ for all $r \neq \pm 1$. Moreover, $(G^*)^{(p)}$ may have a large magnitude (typically, $(G^*)^{(p)} \sim O((p!)^{1+\nu}) \sim O(p! p^{\nu p})$ for some $\nu > 0$). This is observed especially in kernels that fully depend on r, rather than on $|r|$ only (in [8, 9, 40], $G = G(|r|)$). For instance, in the secular problem, G's sign flip across $r = 0$ causes a "fold" in G^* (see Figure 1), consequently causing a large $\|(G^*)^{(p)}\|_{L^\infty(\mathbb{R})}$ (see Table 1). The relative error ε_I in approximating the scale-S softened kernel $G_S(r) := \tilde{G}(r/S)/S$ by a p-order central interpolation from its values on a meshsize-H uniform grid (when the discontinuities are not straddled by the interpolation interval) satisfies

$$\varepsilon_I \lesssim 2\left(\frac{pH}{2eS}\right)^p, \tag{2.5}$$

as explained in the appendix.

TABLE 1
The powers ν corresponding to the magnitudes of the p-order derivatives of $G^*, \tilde{G}$, which determine the interpolation error ε_I, versus p.

p	$\log(\|(G^*)^{(p)}\|_{L^\infty(\mathbb{R})}/p!)/(p\log(p))$	$\log(\|\tilde{G}^{(p)}\|_{L^\infty(\mathbb{R}\setminus\{-S,S\})}/p!)/(p\log(p))$
2	0.8582	0
4	0.4275	0
6	0.3357	0
8	0.2934	0
10	0.2681	0

2. An evaluation of G^* costs $O(p)$ operations, versus none per $\tilde{G}$-evaluation.

3. Computing $\{a_n\}_{n=0}^{2p}$ of (2.3) requires $O(p^3)$ operations, whereas $\tilde{G}$ requires no "preparatory work." This is usually a preprocessing step, but, if the softened kernel needs to be repeatedly updated, this would mean a major saving of work.

On the other hand, $\tilde{G}$'s jumps at $r = \pm 1$ require additional correction steps, which are described in section 2.2. Overall, the cost-efficiency of the multilevel summation algorithm is improved by using (2.4) instead of (2.3) because of the first two advantages. This is shown in section 2.4 for the secular equation and in section 4.3 for general integral transforms in higher dimensions.

2.2. Derivation of the algorithm. Following the terminology of [5, sections 3–4], observe that

$$v(\lambda_j) = v_S^0(\lambda_j) + v_{\text{local}}^0(\lambda_j), \qquad j = 1, \ldots, N, \tag{2.6}$$

where

$$v_S^0(\lambda_j) := \sum_{k=1}^{N} G_S(d_k - \lambda_j) u(d_k), \qquad j = 1, \ldots, N, \tag{2.7}$$

and

$$v_{\text{local}}^0(\lambda_j) := \sum_{k:|d_k-\lambda_j|\le S} G(d_k - \lambda_j) u(d_k), \qquad j = 1, \ldots, N. \tag{2.8}$$

The sum (2.8) extends over $O(s)$ points d_k if we choose $S = sH$. The softened kernel can be represented as

$$G_S(d_k - \lambda_j) = \sum_{K\in\sigma_k} \omega_{kK}^{1,0} G_S(D_K^1 - \lambda_j) + O(\varepsilon_I), \tag{2.9}$$

where $\sigma_k := \{K : |D_K^1 - d_k| < pH/2\}$, $\omega_{kK}^{1,0}$ are the weights of interpolation from the gridpoints D_K^1 to d_k, and ε_I is bounded by (2.5). The grid $\{D_K^1\}_{K=1}^{N_1}$ may include $O(p)$ points to the left of d_1 and to the right of d_n to keep the interpolation central; from now on, p is assumed to be even. In fact, for a given j, (2.9) holds for all $k = 1, \ldots, N$ except the set

$$\Omega_j^{\text{bad}} := \{k : \exists K, K{+}1 \in \sigma_k, b \in \{-1, 1\}, \quad \operatorname{sgn}(D_K^1 - \lambda_j - bS) \neq \operatorname{sgn}(D_{K+1}^1 - \lambda_j - bS)\},$$

since $G_S(d_k - \cdot)$ is not continuous in the interpolation stencil for $k \in \Omega_j^{\text{bad}}$. Neglecting $O(\varepsilon_I)$ terms, it follows that

$$v_S^0(\lambda_j) = \sum_{k=1}^{N} \sum_{K\in\sigma_k} \omega_{kK}^{1,0} G_S(D_K^1 - \lambda_j) u(d_k) + \omega^0(\lambda_j) = V_S^0(\lambda_j) + \omega^0(\lambda_j), \tag{2.10}$$

where

$$V_S^0(\lambda_j) := \sum_{K=1}^{N_1} G_S(D_K^1 - \lambda_j) U^1(D_K^1), \qquad j = 1, \dots N, \tag{2.11}$$

$$U^1(D_K^1) := \sum_{k \in \tau_K} \omega_{kK}^{1,0} u(d_k), \quad \tau_K := \{k : K \in \sigma_k\}, \quad K = 1, \dots, N_1, \tag{2.12}$$

$$\omega^0(\lambda_j) := \sum_{k \in \Omega_j^{\text{bad}}} G_S(d_k - \lambda_j) u(d_k) - \sum_{K \in \Omega_j^{\text{BAD}}} G_S(D_K^1 - \lambda_j) \tilde{U}_j^1(D_K^1), \tag{2.13}$$

$$\tilde{U}_j^1(D_K^1) := \sum_{k \in \tilde{\tau}_{jK}} \omega_{kK}^{1,0} u(d_k), \quad \tilde{\tau}_{jK} := \Omega_j^{\text{bad}} \cap \tau_K, \quad K \in \Omega_j^{\text{BAD}}, \tag{2.14}$$

$\Omega_j^{\text{BAD}} := \{K : \tilde{\tau}_{jK} \neq \emptyset\}$, and (2.13), (2.14) are defined for all $j = 1, \dots, N$. Note that the sums in (2.12), (2.13), and (2.14) extend over $O(p)$ points; hence they are *local*. $\{U_K^1\}_K$ is the "aggregation" of $\{u_k\}_k$ from the nonuniform fine locations $\{d_k\}_k$ (denoted "level $l = 0$") to the uniform coarse locations $\{D_K\}_K$ (denoted "level $l = 1$"), a procedure referred to as *anterpolation* in [5] since it is the adjoint of interpolation. Similarly, we can use the smoothness of $G_S(d - \lambda)$ in λ to write

$$G_S(D_K^1 - \lambda_j) = \sum_{J \in \bar{\sigma}_j} \bar{\omega}_{jJ}^{1,0} G_S(D_K^1 - \Lambda_J^1) + O(\varepsilon_I), \quad j = 1, \dots, N, \tag{2.15}$$

for all $K = 1, \dots, N_1$ except the set

$$\bar{\Omega}_j^{\text{BAD}} := \{K : \exists J, J+1 \in \bar{\sigma}_j, b \in \{-1, 1\}, \operatorname{sgn}(D_K^1 - \Lambda_J^1 - bS) \neq \operatorname{sgn}(D_K^1 - \Lambda_{J+1}^1 - bS)\},$$

where $\bar{\sigma}_j := \{J : |\Lambda_J^1 - \lambda_j| < pH/2\}$, $\bar{\omega}_{jJ}^{1,0}$ are the λ-interpolation weights, and $\{\Lambda_J^1\}_{J=1}^{\bar{N}_1}$ is a uniform grid with meshsize H over $[\lambda_1, \lambda_N]$ (again including $O(p)$ points to the left of λ_1 and to the right of λ_N), from which we can use p-order central interpolation to all points $\lambda_1, \dots, \lambda_N$. Up to an $O(\varepsilon_I)$ error,

$$V_S^0(\lambda_j) = \bar{V}_S^0(\lambda_j) + z^0(\lambda_j), \qquad j = 1, \dots, N, \tag{2.16}$$

where

$$\bar{V}_S^0(\lambda_j) := \sum_{J \in \bar{\sigma}_j} \bar{\omega}_{jJ}^{1,0} V_S^1(\Lambda_J^1), \tag{2.17}$$

$$V_S^1(\Lambda_J^1) := \sum_{K=1}^{\bar{N}_1} G_S(D_K^1 - \Lambda_J^1) U^1(D_K^1), \qquad J = 1, \dots, \bar{N}_1, \tag{2.18}$$

$$z^0(\lambda_j) := \sum_{K \in \bar{\Omega}_j^{\text{BAD}}} G_S(D_K^1 - \lambda_j) U^1(D_K^1) - \sum_{J \in \bar{\sigma}_j} \bar{\omega}_{jJ}^{1,0} \tilde{V}_j^1(\Lambda_J^1), \tag{2.19}$$

and

$$\tilde{V}_j^1(\Lambda_J^1) := \sum_{K\in\bar{\Omega}_j^{\mathrm{BAD}}} G_S(D_K^1 - \Lambda_J^1)U^1(D_K^1), \qquad J = 1,\dots,\bar{N}_1. \tag{2.20}$$

The sums in (2.17), (2.19) are over local sets and are defined for all $j = 1,\dots,N$; (2.18) is a uniform coarser version of (2.1). We have reduced the original evaluation of v at the nonuniform fine level ($l = 0$) to the evaluation of V_S^1 at the uniform coarse level ($l = 1$). In order to keep the evaluation of (2.8) inexpensive, the coarsening ratio m cannot be too large (e.g., $m = 2$ [9]) and s should not increase with N. To sum up, the multisummation (2.1) is replaced by the following.

(i) *Anterpolation.* Calculate the "aggregated" $\{U_K^1\}_K$ from (2.12).
(ii) *Coarse grid summation.* Carry out the task (2.18).
(iii) *Interpolation.* Interpolate $\{V_S^1(\Lambda_J^1)\}_J$ to $\{\bar{V}_S^0(\lambda_j)\}_j$ using (2.17).
(iv) *Local corrections.* Add the local correction $v_{\mathrm{local}}(\lambda_j)$ defined by (2.8) to $\bar{V}_S^0$.
(v) *w-correction.* Compute w^0 from (2.13),(2.14) and add it to $\bar{V}_S^0$.
(vi) *z-correction.* Compute z^0 from (2.19), (2.20) and add it to $\bar{V}_S^0$.

The number of nodes at level 1 is roughly $N/2$, which may still be too large to calculate directly. Instead, the task (2.18) can be further reduced to summation at level $l = 2$ on twice as coarse (meshsize $2H$) λ- and d-grids, using the same algorithm ((i)–(vi)): decomposition of G_S into G_{2S} plus a local part, anterpolation of U^1 to level 2, level 2 summation, interpolation of V_{2S}^2 to level 1, and addition of the three local corrections. The above-described procedure can be repeated recursively until a grid is reached at which direct summation can be done in at most $O(N)$ operations.

2.3. Computational cost and evaluation error. The local correction (iv) costs $O(sN)$ operations since each G-evaluation costs $O(1)$. However, it is less obvious to implement the w-correction in $O(pN)$ operations. It may seem that, for any given j, it takes $O(p)$ points to compute every $\tilde{U}_j^1(D_K), K \in \bar{\Omega}^{\mathrm{BAD}}$; hence $O(p^2N)$ operations are required for evaluating the right-hand term in the right-hand side of (2.13). Instead, we can use a "sliding window" approach (see, for example, [41, 42, 45]): $\{\tilde{U}_1^1(D_K^1)\}_K$ are calculated in $O(p^2)$ and then are repeatedly updated in $O(p)$ operations to obtain $\{\tilde{U}_2^1(D_K^1)\}_K$, and so on. This is possible since the sets $(\tau_{jK}\cup\tau_{j+1,K})\backslash(\tau_{jK}\cap\tau_{j+1,K})$ contain only $O(1)$ points for every $j = 1,\dots,N-1$. The same approach can be applied to the z-correction, interpolations, and anterpolations. Thus the total computational complexity of steps (i),(iii)–(vi) is $W = O((p+s)N)$, which is smaller than the $O(psN)$ cost of the multilevel summation with the "traditional" softening [5, 8, 9, 40]. Generally, if the order of anterpolation/interpolation from/to level l to/from $l-1$ is denoted by p_l and the softening scale is denoted by $S_l := 2^{l-1}Hs_{l-1}$, the total work W per fine gridpoint in evaluating (2.1) (omitting some constants and neglecting the direct evaluation at the coarsest level) is given by

$$\frac{W}{N} = \sum_{l=0}^{t-1} 2^{-l}(p_l + As_l), \tag{2.21}$$

where $t = O(\log N)$ is the number of levels and $A > 0$ is a constant. The error ε_v in evaluating v satisfies (as implied by (2.5))

$$\varepsilon_v \lesssim 2\sum_{l=0}^{t-1}\left(\frac{p_l}{2es_l}\right)^{p_l}.$$

2.4. Parameter optimization. The values of s, p at each of the levels $l = 0, \ldots, t-1$ should be determined to minimize the computational work under the constraint of a controlled evaluation error, $\varepsilon_v = \varepsilon$.

2.4.1. Two-level parameter optimization. Let us first consider the case $t = 1$. Discarding the coarse level summation portion of the work and omitting constants, the constrained minimization problem for $p := p_0, s := s_0$ is

$$\begin{cases} W/N & \propto p(1 + A\kappa/(2e)) & \longrightarrow \min., \qquad \kappa := 2es/p, \\ \varepsilon_v & \propto \kappa^{-p} & = \varepsilon. \end{cases}$$

The optimum is attained if and only if

$$\left[\frac{d}{d\kappa}\left(\frac{1 + A\kappa/(2e)}{\log(\kappa)}\right)\right]_{k=k_{\text{opt}}} = 0, \quad p_{\text{opt}} = \frac{\log(1/\varepsilon)}{\log(\kappa_{\text{opt}})}.$$

This implies

$$\kappa_{\text{opt}}(\log(\kappa_{\text{opt}}) - 1) = (2e)/A \qquad \Longrightarrow \qquad \text{(e.g.)} \quad \begin{array}{ll} \kappa_{\text{opt}} \approx 6.376, & A = 1, \\ \kappa_{\text{opt}} \approx 9.045, & A = 0.5. \end{array}$$

Thus

$$p_{\text{opt}} = p_{\text{opt}}(\varepsilon) = K_1 \log\left(\frac{1}{\varepsilon}\right), \quad s_{\text{opt}} = s_{\text{opt}}(\varepsilon) = K_2 \log\left(\frac{1}{\varepsilon}\right), \tag{2.22}$$

where, for instance, $K_1 \approx 0.54, K_2 \approx 0.63$ for $A = 1$ and $K_1 \approx 0.45, K_2 \approx 0.75$ for $A = 0.5$. Consequently, the computational complexity of evaluating (2.1) to accuracy ε is

$$W = (K_1 + AK_2)N \log(1/\varepsilon) =: KN \log(1/\varepsilon). \tag{2.23}$$

2.4.2. Multilevel parameter optimization. Clearly, if we use $p_l = p_{\text{opt}}(\varepsilon)$, $s_l := s_{\text{opt}}(\varepsilon)$ at all levels $l = 0, \ldots, t-1$, the error ε_v would be $t\varepsilon$. Instead, we use

$$p_l = p_{\text{opt}}(2^{-l-1}\varepsilon), \quad s_l = s_{\text{opt}}(2^{-l-1}\varepsilon), \qquad l = 0, \ldots, t-1, \tag{2.24}$$

so that

$$\varepsilon_v = \sum_{l=0}^{t-1} 2^{-l-1}\varepsilon \le \varepsilon$$

and

$$W \le KN \sum_{l=0}^{t-1} 2^{-l} \log\left(\frac{2^{l+1}}{\varepsilon}\right) \le 2KN\left(\log\left(\frac{1}{\varepsilon}\right) + 4\log(2)\right), \tag{2.25}$$

using (2.21). This cost is smaller than the total cost of the multilevel summation algorithm with continuous softening. Indeed, with the latter, we get $W = O(N(\log(1/\varepsilon))^2)$.

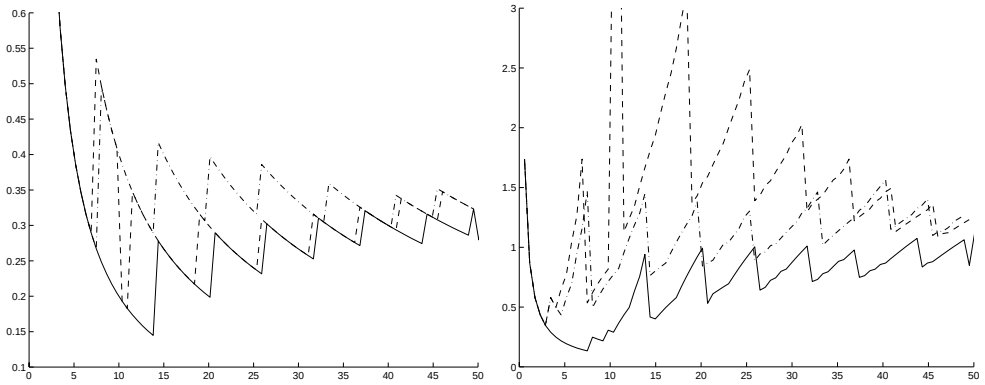


FIG. 2. *The optimal interpolation order* ($p_{\mathrm{opt}}(\varepsilon)/\log(1/\varepsilon)$, *left picture) and the optimal softening distance* ($s_{\mathrm{opt}}(\varepsilon)/\log(1/\varepsilon)$, *right picture) versus* $\log(1/\varepsilon)$ *for* $N = 64$ *(dashed line),* $N = 256$ *(dashed-dotted line), and* $N = 1024$ *(solid line). For small* ε, $p_{\mathrm{opt}} \approx 0.3\log(1/\varepsilon)$ *and* $s_{\mathrm{opt}} \approx 1.1\log(1/\varepsilon)$.

TABLE 2

The computational cost $W/(N\log(1/\tilde{\varepsilon}))$ *versus* N *and* ε. *Each column (starting from the second) corresponds to a different* $\log_{10}(\varepsilon)$ *value, that is,* $\varepsilon = 10^{-2}$ *through* 10^{-12}; W *is the arithmetic operations count. It can be observed that* $W/(N\log(1/\tilde{\varepsilon}))$ *is practically uniformly bounded, as claimed by* (2.25).

N	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
64	36.0	33.1	31.7	30.0	26.9	21.6	8.1	5.5	5.5	5.5	5.5
128	41.0	40.4	37.0	40.3	45.3	51.0	37.4	42.3	38.2	33.6	30.1
256	46.8	46.0	40.4	43.1	59.6	61.3	57.0	60.3	64.6	62.4	60.4
512	46.2	44.0	40.6	40.4	51.6	62.8	61.4	65.3	72.7	71.4	71.5
1024	42.5	41.1	40.4	38.7	56.5	49.4	62.9	64.0	68.2	75.0	74.2
2048	44.1	43.2	42.3	39.8	58.3	48.1	64.9	64.6	69.2	77.7	77.1
4096	43.0	41.8	40.8	36.8	51.1	45.2	63.2	64.0	67.6	77.3	76.9

2.5. Numerical results. First, we performed two-level ($t = 1$) evaluation experiments of (2.1) for different values of N to show that the optimal p, s indeed satisfy (2.22). The pair $(p_{\mathrm{opt}}(\varepsilon), s_{\mathrm{opt}}(\varepsilon))$ corresponding to the minimal W (out of all $0 \le p \le 16, 0 \le s \le 64$) was computed for various ε values and stored in a table. The values of W were averaged over 20 experiments, each using a uniformly random sequence pair $\{d_k\}_{k=1}^{N}, \{\lambda_j\}_{j=1}^{N} \subset [0,1]$ that satisfied (1.2). Figure 2 shows that $p_{\mathrm{opt}}(\varepsilon)/\log(1/\varepsilon), s_{\mathrm{opt}}(\varepsilon)/\log(1/\varepsilon)$ are indeed bounded independently of N.

Second, we performed the multilevel evaluation of (2.1) for various N and ε values ($t = O(\log N)$ being the maximum possible so that level $l = t$ grids contained $O(p_t)$ points) using $\{p_l, s_l\}_{l=0}^{t-1}$, which were computed using the table generated at the two-level stage and (2.24). Table 2 summarizes the computational cost of evaluating v in these experiments; each experiment was averaged over 20 uniformly random sequence pairs $\{d_k\}_{k=1}^{N}, \{\lambda_j\}_{j=1}^{N} \subset [0,1]$ satisfying (1.2). (This was a sufficiently large sample.) The l_∞ error $\tilde{\varepsilon}$ of the differences between the directly computed v values and the values computed using the fast evaluation algorithm was always less than the desired ε. It can be observed that W behaves according to the desired (2.25).

3. Fast solution of $f(\lambda) = 0$. The fast evaluation algorithm presented in section 2 can be naturally adapted to any root-search method for solving $f(\lambda) = 0$. For demonstration purposes, we used Melman's improved Newton method [32]. Let

$1 \le j < N$. (For simplicity, we avoid the case $j = N$, which is also treated in [32].) The iterations take the form

$$\lambda_j^{(n+1)} = d_j + \frac{(\lambda_j^{(n)} - d_j)^2 f'(\lambda_j^{(n)})}{f(\lambda_j^{(n)}) + (\lambda_j^{(n)} - d_j) f'(\lambda_j^{(n)})}, \qquad n = 0, 1, 2, \ldots. \tag{3.1}$$

It was proved that these iterations converge quadratically to λ_j^*, provided that the starting point is

$$\lambda_j^{(0)} = d_j + \frac{2A}{-B - \sqrt{B^2 - 4AC}}, \tag{3.2}$$

where

$$\begin{array}{lll} a := 1 + \Delta_j, & \delta := d_{j+1} - d_j, & \\ A := -\frac{u_j}{\delta}, & B := a\delta + u_j, & C := \frac{u_{j+1}}{\delta} - a, \end{array}$$

and

$$\Delta_j := \sum_{k:k\neq j} \frac{u_k}{d_k - d_j}, \qquad j = 1, \ldots, N. \tag{3.3}$$

Our algorithm for finding the roots $\{\lambda_j^*\}_{j=1}^{n-1}$ to an accuracy ε consists of the following steps:

(i) Compute $\{\Delta_j\}_{j=1}^{N-1}$ of (3.3) using the fast evaluation algorithm (section 2).
(ii) For $j = 1$ to $N - 1$, set λ_j to the expression of (3.2).
(iii) Compute $\{V_S^1(\Lambda_J^1)\}_{J=1}^{N_1}$ using the fast evaluation algorithm (section 2).
(iv) For $j = 1$ to $N - 1$, do steps (v)–(viii).
(v) While (STOP–CRITERION$_j$ = FALSE) do steps (vi)–(viii).
(vi) Compute $f(\lambda_j)$ (see section 3.1).
(vii) Compute $f'(\lambda_j)$ (see section 3.2).
(viii) Set $\lambda_j \leftarrow d_j + ((\lambda_j - d_j)^2 f'(\lambda_j))/(f(\lambda_j) + (\lambda_j - d_j) f'(\lambda_j))$.

Step (i) is executed in $O(N \log(1/\varepsilon))$, using the fast evaluation algorithm of section 2 for computing $\{v(\lambda_j)\}_{j=1}^N$ to accuracy ε, with one modification: the kernel $G(r)$ is defined to be 0 at $r = 0$. Here we can accept a low accuracy since we provide only initial conditions for the roots.

The initialization of $\{\lambda_j\}_j$ (step (ii)) requires $O(N)$ operations.

Step (iii), using an accuracy ε, is a preparatory step for the fast evaluation of f, f' in steps (vi) and (vii) (see sections 3.1 and 3.2). We execute the algorithm for evaluating $\{v(\lambda_j)\}_{j=1}^N$ to accuracy ε, excepting the last four steps (i.e., the steps before interpolating $\{V_S^1(\Lambda_J^1)\}_J$ to level 0). This takes $O(N \log(1/\varepsilon))$ operations.

The stopping criterion may be chosen in different ways. We use the criterion

$$|\lambda_j^{(n+1)} - \lambda_j^{(n)}| \le \varepsilon |d_{j+1} - d_j|.$$

Provided that each evaluation of f or f' at steps (vi) and (vii) costs $O(\log(1/\varepsilon))$ operations (see sections 3.1 and 3.2), the total cost of the algorithm (i)–(viii) is $O(N \log(1/\varepsilon))$. The numerical stability of algorithm (i)–(viii) depends solely on the stability of the root-search methods; the fast evaluation introduces, in addition, central interpolation, which is a numerically stable process.

3.1. $O(\log(1/\varepsilon))$-evaluation of v. Once $\{V_S^1(\Lambda_J^1)\}_J$ is computed and stored (step (ii)), a value $f(\lambda)$ for a given λ may be calculated in additional $O(\log(1/\varepsilon))$ operations, using the last four steps of the evaluation algorithm: interpolation of V_S^1 to the point λ, followed by the three local corrections to $v_s(\lambda)$. Here a "sliding window" (see section 2.3) is used (for every j) to update $\tilde{U}_j^1(D_K^1), \tilde{V}_j^1(\Lambda_J^1)$ from their values at the previous root-search step. Since the approximations $\lambda_j^{(n)}$ to the jth root λ_j^* remain in the interval $[d_j, d_{j+1}]$ for all n (see [32]), the interpolation stencils in the w- and z-corrections can "move" only by at most $O(1)$ meshsizes in every Newton step. Hence each correction costs only $O(p = \log(1/\varepsilon))$ operations per Newton step for a single root.

3.2. $O(\log(1/\varepsilon))$-evaluation of v'. If we also want to evaluate f', we can again use the precomputed values of $\{V_S^1(\Lambda_J^1)\}_J$. As in section 3.1, we perform the last four steps of the evaluation algorithm with two modifications.

1. In the interpolation step, we use different interpolation coefficients $\{\xi_J^{1,0}(\lambda)\}_J$ for interpolating V_S^1 from $\{\Lambda_J^1\}_{J\in\bar{\sigma}_j}$ to λ, instead of $\{\bar{\omega}_J^{1,0}(\lambda)\}_J$ (used for interpolating V_S^1 from $\{\Lambda_J^1\}_{J\in\bar{\sigma}_j}$ to λ in the v-evaluation step; $\{\bar{\omega}_J^{1,0}(\lambda_j)\}_J = \{\bar{\omega}_{jJ}^{1,0}\}_J$). These coefficients are computed from differentiating the interpolation polynomial for $G_S(D_K^1 - \cdot)$ (see also [40]) so that (except when discontinuities are straddled by the interpolation stencil)

$$-G_S'(D_K^1 - \lambda) = \sum_{J\in\bar{\sigma}_j} \xi_J^{1,0}(\lambda)\left[-G_S(D_K^1 - \Lambda_J^1)\right] + O(\varepsilon_I). \tag{3.4}$$

2. The three local corrections are executed with the kernel $-G'$ instead of G. (Note that $(d/d\lambda)[G_S(d-\cdot)] = -G'(d-\cdot)$.)

We remark that we can evaluate v' to a lower accuracy than the one required for v without spoiling the convergence of the Newton iterations (3.1). In fact, we can avoid computing the derivative by switching to the secant root-search method, thereby reducing the overall computing time by a factor of 1.8.

3.3. Numerical results. Table 3 compares the computational cost of evaluating the roots $\{\lambda_j^*\}_{j=1}^N$ of (1.1), using a direct evaluation of v (with $\varepsilon = 10^{-10}$) versus a fast evaluation of v with $\varepsilon = 10^{-20}, 10^{-10}$. The results were averaged over 20 uniformly random sequence pairs $\{d_k\}_{k=1}^N, \{\lambda_j\}_{j=1}^N \subset [0,1]$ satisfying (1.2). (This was a sufficiently large sample.) Indeed, the average cost per root for the direct evaluation method increases linearly with N, whereas it remains constant for our proposed method, as desired. The cross-over (using direct evaluation versus fast evaluation, the roots being computed to the same accuracy ε) was detected at $N \approx 200$ for $\varepsilon = 10^{-10}$ and at $N \approx 450$ for $\varepsilon = 10^{-20}$ (for $\varepsilon = 10^{-5}$ at $N \approx 70$).

4. Concluding remarks. In the previous sections, we described the basic elements of the fast evaluation of v and the fast solution of (1.1) for uniformly dense $\{d_k\}_k$. (N roots are computed in only $O(N)$ operations.) The following are some important insights and generalizations of these algorithms that can be further explored in future research.

4.1. Nonuniform d-density. Recursive local grid refinement (see [2]) is essential to maintain the above work-accuracy relationship wherever the number of d_k-points per meshsize is large, including pathologically high concentrations (for instance, $d_k = 1/k, k = 1, \ldots, N$).

TABLE 3

The computational cost (number of arithmetic operations) of the proposed novel algorithm versus current algorithms: The fourth column is the number of Newton steps (3.1) *in the algorithm* (i)–(viii) *for* $\varepsilon = 10^{-10}$*, and its cost per root (number of arithmetic operations for computing a single root) is given in column* 5*. Columns* 2 *and* 3 *are the corresponding measurements when* f, f' *in the algorithm* (i)–(viii) *are directly computed from* (1.1), (2.1) *to accuracy* $\varepsilon = 10^{-10}$*. Columns* 6 *and* 7 *are the corresponding values to columns* 4 *and* 5 *for* $\varepsilon = 10^{-20}$*.*

N	Direct (−10) # iter.	Direct (−10) cost/N	Fast (−10) # iter.	Fast (−10) cost/N	Fast (−20) # iter.	Fast (−20) cost/N
64	5.43	$2.66 \cdot 10^3$	5.50	$4.01 \cdot 10^3$	6.52	$4.70 \cdot 10^3$
128	5.59	$5.44 \cdot 10^3$	5.53	$8.29 \cdot 10^3$	6.62	$1.16 \cdot 10^4$
256	5.58	$1.08 \cdot 10^4$	5.55	$8.37 \cdot 10^3$	6.66	$2.09 \cdot 10^4$
512	5.56	$2.15 \cdot 10^4$	5.66	$8.09 \cdot 10^3$	6.71	$2.36 \cdot 10^4$
1024	5.57	$4.30 \cdot 10^4$	5.69	$7.34 \cdot 10^3$	6.73	$2.50 \cdot 10^4$
2048	5.56	$8.59 \cdot 10^4$	5.68	$7.49 \cdot 10^3$	6.76	$2.61 \cdot 10^4$
4096	5.56	$1.72 \cdot 10^5$	5.68	$7.45 \cdot 10^3$	6.75	$2.65 \cdot 10^4$

Importantly, the algorithm will be based on patches of *uniform* grids; therefore, interpolations are highly efficient compared with those involving nonuniform meshsizes. In the rest of this section, we first explain where these patches should be introduced, and then we discuss the adaptation of the evaluation algorithm to such patches.

4.1.1. Refinement strategy. Since, in the secular problem, the local average density of $\{\lambda_j\}_j$ is the same as the $\{d_k\}_k$'s density, local refinements are introduced in the same regions for both d and λ spaces; in general, we may need to construct different patches for the λ's (see section 4.3).

A direct application of the evaluation algorithm described in section 2.2 does not efficiently address the v-evaluation task on nonuniform sets $\{d_k\}_k, \{\lambda_j\}_j$. In regions where the number α of d-points per meshsize H of the finest grid employed in the evaluation algorithm is large, the work involved in the local corrections increases like $O(\alpha^2)$. To avoid this, we introduce a patch of a twice finer grid, defined only over these regions. It is possible to construct an optimal quantitative criterion to decide where to introduce such a local refinement, based on its cost effectiveness. For example, a twice finer patch should be introduced in any region with length of at least s cells (where $S = sH$ is the softening distance—see section 2.1) that includes more than $\alpha_c S$ d-points, where α_c is a small integer whose optimal value can be determined experimentally. Clearly, if two close regions need to be locally refined, it is more efficient to unify them into just one patch.

If yet more dense regions exist within the twice finer patches, we create yet finer patches within the former patches, using the same criterion. This is recursively repeated until no further refinement is needed.

4.1.2. The evaluation algorithm with patches. Here we start the algorithm on the finest patches, where we anterpolate u from the original $\{d_k\}_k$, which lie within the region of these patches and have the highest local density, to the equally spaced gridpoints of the finest patches. Thus we have eliminated the regions of the highest density from the original evaluation task. By recursively anterpolating u to yet coarser and larger patches, we finally arrive at the original everywhere uniform grid covering the full domain of the original $\{d_k\}_k$, where the algorithm of section 2.2 can be directly applied.

Note that the local refinement creates intermediate levels with kernels $G(d, \lambda)$ which no longer depend only on $d - \lambda$; but this changes only their local part, G_{local}, and the local corrections still cost $O(s)$ per λ-point.

4.2. General kernels. The multilevel approach for evaluating (2.1) provides the same efficiency for computing integral transforms involving any asymptotically smooth kernel $G(r)$, as shown in [5]. In particular, other secular equations such as

$$1+\sigma\sum_{k=1}^{N}\frac{u_k}{(d_k-\lambda_j)^{\zeta}}=0$$

can be solved in linear time (N roots in $O(N)$ operations) for any $\{u_k\}_k \subset \mathbb{R}$ and $\zeta > 0$. Importantly, the same multilevel approach can be used to address other multisummation tasks with other types of kernels (e.g., oscillatory) [5].

4.3. Higher dimensions. The secular equation (1.1) does not admit a higher dimensional analogue; nevertheless, it may still be interesting to extend the discontinuous softening technique to the multilevel summation of the transform

$$v(\underline{\lambda_j})=\sum_{k=1}^{N}G(\underline{d_k},\underline{\lambda_j})u(\underline{d_k}),\quad \{\underline{d_k}\}_{k=1}^{N},\{\underline{\lambda_j}\}_{j=1}^{\bar{N}}\subset\mathbb{R}^d, \tag{4.1}$$

with an asymptotically smooth $G(\underline{x},\underline{y})$. The discontinuous softening (2.4) can be extended to this case, specifically,

$$\tilde{G}(\underline{x},\underline{y})=\begin{cases}0, & |\underline{x}-\underline{y}|<1,\\ G(\underline{x},\underline{y}), & |\underline{x}-\underline{y}|>1,\end{cases}\qquad |\underline{x}-\underline{y}|:=\max_{\mu}|x_\mu-y_\mu|. \tag{4.2}$$

This kernel is singular on the sphere $|\underline{x}-\underline{y}|=S$; hence the w- and z-corrections involve points in a high dimensional ring including $|\underline{x}-\underline{y}|=S$ (for kernels $G=G(|\underline{x}-\underline{y}|_2)$, it might be more efficient to use $|\underline{x}-\underline{y}|_2:=\sqrt{\sum_{\mu=1}^{d}(x_\mu-y_\mu)^2}$ in $\tilde{G}$ instead of $|x-y|$). It can be shown that they can be implemented in $O(p^d)$ operations per λ_j, again using the "sliding window" technique (see section 2.3). In addition, the local correction costs here only $O(s^d)$ per λ_j, versus $O(s^d p)$ for the usual softening; see, e.g., (2.3). Substituting the optimal $p=s=O(\log(1/\varepsilon))$, the total work amounts to

$$W=O\left(N\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{d}\right),$$

whereas it is $O(N(\log(1/\varepsilon))^q)$ when using everywhere smooth softened kernels, with $q=d+1+\eta$, $\eta\geq 0$, depending on the magnitude of the p-order derivatives of the softened kernel for $r\leq S$ (see section 2.1).

Importantly, discontinuous softening cannot replace the original smooth softening of [5] because of the following.

1. The original smooth softening is much simpler to implement, and the cost per node (for a fixed accuracy) may be smaller. It is also easier to adapt local refinements to it (see 4.1) than to discontinuous softening.
2. In many physical problems (e.g., evaluation of potential or dipole fields [8, 9, 40]), η is small and $d=2,3$; hence the relative loss $(q-d)$ is not too large.
3. An important advantage of continuous softening is that it gives the kind of multiscale description of the interactions needed in dynamic situations, where the particles carrying these interactions participate in multiscale movements, as in molecular dynamics [3, 5, 6, 39].

However, the cost-efficiency of the multilevel summation algorithms may be significantly improved by discontinuous softening when

1. the dimension d is low (in the extreme case, when $d = 1$);
2. the kernel G is inherently "hard-to-soften," as in the secular problem or in some problems where G is not a function of $r := |\underline{x} - \underline{y}|$ (i.e., unlike the cases of [8, 9, 40], where $G(r) = \log(r)$ or $1/r$).

The power d of $\log(1/\varepsilon)$ in the total work seems to be generally the smallest possible, since the local corrections involve $O(p^d)$ points per evaluation; in this sense, discontinuous softening leads to the optimal work-accuracy relation.

4.4. Parallelization. Our presented algorithms can be efficiently parallelized. The number of unparallelizable steps is theoretically only $O(\log(N))$ since we mostly rely on interpolations and local corrections, which can be fully parallelized (see also [5]).

4.5. Eigenbasis computation. The divide and conquer method [14, 21, 26] for finding the full spectrum (eigenvalues and eigenvectors) of an $N \times N$ symmetric tridiagonal matrix requires $O(N^2)$ computer operations. (Although it runs in $O(N \log N)$ for some very special matrices, the storage always increases as $O(N^2)$.) Even if we incorporate our fast evaluation algorithm for the secular equation, the computational cost remains the same. Instead, the recently developed approach of *multiscale eigenbasis* (MEB) [6, 29, 30, 31] seems more reasonable and efficient in addressing such eigenbasis computations. For instance, this method can be directly applied for computing N eigenvectors and eigenvalues of the symmetric tridiagonal eigenproblem, as well as many other sparse eigenproblems, in $O(N \log N)$ operations and storage. This cost can be reduced to $O(N)$ in special cases, e.g., for discretizations of constant coefficient differential operators. Of course, singular cases (e.g., when the ratio of two adjacent diagonal elements is very large) need to be treated (e.g., deflated), as demonstrated in [10, 11, 14, 26] for $O(N^2)$ eigenbasis solvers.

Appendix. The error of a pth order central interpolation of $\tilde{G}$ defined by (2.4) from its values on a grid of meshsize H satisfies [43]

$$\varepsilon_I \le \frac{1}{p!} \|\tilde{G}^{(p)}\|_{L^\infty(\mathbb{R})} \cdot \left(\frac{H}{2} \cdot \frac{3H}{2} \cdot \cdots \cdot \frac{(p-1)H}{2} \right)^2 = \frac{1}{p!} p! H^p \left(\frac{p!}{2^p (\frac{p}{2})!} \right)^2$$

for any positive even p. Simplifying the expression and using Stirling's formula, we get

$$\varepsilon_I \lesssim H^p \left(\frac{\sqrt{2\pi p} (\frac{p}{e})^p}{2^p \sqrt{\pi p} (\frac{p}{2e})^{\frac{p}{2}}} \right)^2 = 2 \left(\frac{pH}{2e} \right)^p .$$

When we scale $G_S(r) = \tilde{G}(r/S)/S$, the pth derivative is scaled by S^{-p-1} and the relative error by S^{-p}. This implies (2.5).

Acknowledgments. The authors are thankful to the referees of this paper. In particular, they brought to our attention that the idea of evaluating the secular equation in $O(N)$ operations is not new. In [26, section 5], Gu and Eisenstat describe how the fast multipole method [12] can be used to compute all of the roots of the secular equation in $O(N)$ operations. The resulting divide and conquer algorithm for the symmetric tridiagonal eigenproblem computes all of the eigenvalues in $O(N \log N)$ operations and all of the eigenvectors in $O(N^2 \log N)$ operations.

We can add that the idea of fast multilevel evaluation of integral transforms with very general kernels, on which the method presented here is based, appears already in [4, section 8.6], although the specific case of the secular equation is not mentioned there. Also, the recent MEB approach (see section 4.5) shows how to reduce the $O(N^2 \log N)$ cost of calculating the entire eigenbasis to only $O(N \log N)$ operations.

REFERENCES

[1] P. Arbenz and G. H. Golub, *On the spectral decomposition of Hermitian matrices modified by low rank perturbations with applications*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 40–58.

[2] D. Bai and A. Brandt, *Local mesh refinement multilevel techniques*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 109–134.

[3] D. Bai and A. Brandt, *Multiscale computation of polymer models*, in Multiscale Computational Methods in Chemistry and Physics, A. Brandt, J. Bernholc, and K. Binder, eds., NATO Science Series III: Computer and Systems Sciences, NATO Science Series 177, IOS Press, The Netherlands, 2001, pp. 250–266.

[4] A. Brandt, *Guide to multigrid development*, in Multigrid Methods (Cologne, 1981), W. Hackbusch and U. Trottenberg, eds., Lecture Notes in Math. 960, Springer-Verlag, Berlin, 1982, pp. 220–312.

[5] A. Brandt, *Multilevel computations of integral transforms and particle interactions with oscillatory kernels*, Comput. Phys. Comm., 65 (1991), pp. 471–476.

[6] A. Brandt, *Multiscale scientific computation: Review* 2001, in Multiscale and Multiresolution Methods: Theory and Applications, T. J. Barth, T. F. Chan, and R. Haimes, eds., Lect. Notes Comput. Sci. Eng. 20, Springer-Verlag, Heidelberg, 2001, pp. 3–96.

[7] A. Brandt and A. A. Lubrecht, *Multilevel matrix multiplication and the fast solution of integral equations*, J. Comput. Phys., 90 (1990), pp. 348–370.

[8] A. Brandt and C. H. Venner, *Multilevel Evaluation of Integral Transforms on Adaptive Grids*, Report WI/GC–5, Weizmann Institute of Science, Rehovot, Israel, 1996.

[9] A. Brandt and C. H. Venner, *Multilevel evaluation of integral transforms with asymptotically smooth kernels*, SIAM J. Sci. Comput., 19 (1998), pp. 468–492.

[10] J. R. Bunch and C. P. Nielsen, *Updating the singular value decomposition*, Numer. Math., 31 (1978), pp. 111–129.

[11] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[12] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 669–686.

[13] T. F. Chan, J. A. Olkin, and D. W. Cooley, *Solving quadratically constrained least squares using black box solvers*, BIT, 32 (1992), pp. 481–495.

[14] J. J. M. Cuppen, *A divide and conquer method for the tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[15] J. J. Dongarra and D. C. Sorensen, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. S139–S154.

[16] V. N. Faddeeva, *Computational Methods of Linear Algebra*, Dover, New York, 1959.

[17] G. E. Forsythe and G. H. Golub, *On the stationary values of a second-degree polynomial on the unit sphere*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 1050–1068.

[18] D. R. Fuhrmann, *An algorithm for subspace computation with applications in signal processing*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 213–220.

[19] W. Gander, *Least squares with a quadratic constraint*, Numer. Math., 36 (1981), pp. 291–307.

[20] W. Gander, G. H. Golub, and U. von Matt, *A constrained eigenvalue problem*, Linear Algebra Appl., 114/115 (1989), pp. 815–839.

[21] D. Gill and E. Tadmor, *An $O(N^2)$ method for computing the eigensystem of $N \times N$ symmetric tridiagonal matrices by the divide and conquer approach*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 161–173.

[22] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973) pp. 318–334.

[23] G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.

[24] W. B. Gragg and L. Reichel, *A divide and conquer method for unitary and orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.

[25] 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.
[26] 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.
[27] S. L. Handy and J. L. Barlow, *Numerical solution of the eigenproblem for banded, symmetric Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 205–214.
[28] R. Hanson and J. Phillips, *An adaptive numerical method for solving linear Fredholm integral equations of the first kind*, Numer. Math., 24 (1975), pp. 291–307.
[29] O. E. Livne, *Multiscale Eigenbasis Algorithms*, Ph. D. Thesis, Weizmann Institute of Science, Rehovot, Israel, 2000.
[30] O. E. Livne and A. Brandt, *$O(N \log N)$ multilevel calculation of N eigenfunctions*, in Multiscale Computational Methods in Chemistry and Physics, A. Brandt, J. Bernholc, and K. Binder, eds., NATO Science Series III: Computer and Systems Sciences, NATO Science Series 177, IOS Press, The Netherlands, 2001, pp. 123–148.
[31] O. E. Livne and A. Brandt, *Multiscale eigenbasis calculations: N eigenfunctions in $O(N \log N)$*, in Multiscale and Multiresolution Methods: Theory and Applications, T. J. Barth, T. F. Chan, and R. Haimes, eds., Lect. Notes Comput. Sci. Eng. 20, Springer-Verlag, Heidelberg, 2001, pp. 347–358.
[32] A. Melman, *Numerical solution of a secular equation*, Numer. Math., 69 (1995), pp. 483–493.
[33] A. Melman, *A unifying convergence analysis of second-order methods for secular equations*, Math. Comp., 66 (1997), pp. 333–344.
[34] A. Melman, *A numerical comparison of methods for solving secular equations*, J. Comput. Appl. Math., 86 (1997), pp. 237–249.
[35] A. Melman, *Analysis of third-order methods for secular equations*, Math. Comp., 67 (1998), pp. 271–286.
[36] C. H. Reinsch, *Smoothing by spline functions*, Numer. Math., 10 (1967), pp. 167–183.
[37] C. H. Reinsch, *Smoothing by spline functions* II, Numer. Math., 16 (1971), pp. 451–454.
[38] R. C. Li, *Solving the Secular Equation Stably and Efficiently*, Technical report, Department of Mathematics, University of California, Berkeley, CA, 1993. LAPACK Working Note 89.
[39] D. Ron and A. Brandt, *Renormalization multigrid (RMG): Coarse-to-fine Monte Carlo acceleration and optimal derivation of macroscopic actions*, in Multiscale Computational Methods in Chemistry and Physics, A. Brandt, J. Bernholc, and K. Binder, eds., NATO Science Series III: Computer and Systems Sciences, NATO Science Series 177, IOS Press, The Netherlands, 2001, pp. 163–186.
[40] B. Sandak and A. Brandt, *Multiscale fast summation of long range charge and dipolar interactions*, J. Comput. Chem., to appear.
[41] M. I. Skolnik, ed., *Radar Handbook*, McGraw–Hill, New York, 1970.
[42] M. I. Skolnik, ed., *Introduction to Radar Systems*, 2nd ed., McGraw–Hill, Tokyo, 1983.
[43] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, translation of *Einführung in die Numerische*, Springer-Verlag, New York, 1980.
[44] U. von Matt, *Large Constrained Quadratic Problems*, Verlag dr Fachvereine, Zürich, 1993.
[45] M. Weiss, *Analysis of some modified cell-averaging CFAR processes in multiple-target situations*, IEEE Trans. Aerospace Electron. Systems, 18 (1982), pp. 102–114.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 454–464

ON THE ELASTICITY OF THE PERRON ROOT OF A NONNEGATIVE MATRIX*

S. J. KIRKLAND†, M. NEUMANN‡, N. ORMES‡, AND J. XU‡

Abstract. Let $A = (a_{i,j})$ be an $n \times n$ nonnegative irreducible matrix whose Perron root is λ. The quantity $e_{i,j} = \frac{a_{i,j}}{\lambda} \frac{\partial \lambda}{\partial a_{i,j}}$ is known as the elasticity of λ with respect to $a_{i,j}$. In this paper, we give two proofs of the fact that $\frac{\partial e_{i,j}}{\partial a_{i,j}} \geq 0$ so that $e_{i,j}$ is increasing as a function of $a_{i,j}$. One proof uses ideas from symbolic dynamics, while the other, which is matrix theoretic, also yields a characterization of the case when $\frac{\partial e_{i,j}}{\partial a_{i,j}} = 0$. We discuss a resulting connection between the elements of A and the elements of the group inverse of $\lambda I - A$.

Key words. elasticity, population models, nonnegative matrices

AMS subject classifications. 15A09, 15A18, 15A48, 92D25

PII. S0895479801398244

1. Introduction.

A large class of models in mathematical population biology (and in other areas) has the following common structure: we are given an $n \times n$ matrix A with nonnegative entries, frequently called the *projection matrix of the model*, and an initial population vector $x_0 \in \mathbb{R}^n_+$, and, for each $k \in \mathbb{N}$, we define $x_k = Ax_{k-1}$. In the case that the Perron root of A, say, λ, is a simple dominant eigenvalue, it follows that x_k/λ^k converges to an appropriate scalar multiple of the right Perron vector for A. Thus λ can be thought of as the asymptotic growth rate for the population being modeled.

A specific example and almost the simplest population model in mathematical biology is the Leslie model. In this model, individuals can live up to the age of n, and the projection matrix of this model is given by

$$A = \begin{pmatrix} F_1 & F_2 & \dots & \dots & F_{n-1} & F_n \\ T_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & T_2 & 0 & 0 & 0 & 0 \\ \vdots & \dots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \dots & \dots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & T_{n-1} & 0 \end{pmatrix}, \tag{1.1}$$

where the F_i's signify the *birth rate* (or *fecundity*) at age i, $i = 1, \dots, n$, while the T_i's signify the *survival rate* from age i to age $i+1$, $i = 1, \dots, n-1$. In the literature on mathematical models for population growth, the birth and survival rates together are often referred to as the *vital rates*. We refer the reader to Caswell [5] for a comprehensive introduction and a reference source on matrix population models.

*Received by the editors November 13, 2001; accepted for publication (in revised form) by R. Bhatia April 29, 2002; published electronically November 6, 2002.

http://www.siam.org/journals/simax/24-2/39824.html

†Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada S4S 0A2 (kirkland@math.uregina.ca). The research of this author was supported in part by NSERC under grant OGP0138251.

‡Department of Mathematics, University of Connecticut, Storrs, CT 06269–3009 (neumann@math.uconn.edu, ormes@math.du.edu, jhxu@math.uconn.edu). The work of the second author was supported in part by NSF grant DMS9973247.

While the perturbation analysis of the Perron root λ of a nonnegative and irreducible matrix A occurs in a variety of applications of nonnegative matrices (see, for example, Berman and Plemmons [3] and Varga [13], where further background material on nonnegative and M-matrices can also be found), the usual sensitivity analysis neglects inherent restrictions on the magnitudes of the various entries of A. Returning to the example of the Leslie population model given above, an entry of A in (1.1) corresponding to a birth rate may exceed 1, while an entry corresponding to a survival rate must necessarily be at most 1. Consequently, in order to take these magnitude restrictions into account, the notion of the *elasticity of* λ with respect to an entry or vital rate in A has been introduced in the mathematical biology literature (see [5] and De Kroon, Plaisier, van Groenendael, and Caswell [6]). Here is its formal definition.

DEFINITION 1.1. *Let $A = (a_{i,j})$ be a nonnegative matrix, and suppose that its Perron root λ is a simple eigenvalue. The elasticity of λ with respect to the (i,j)th entry of A is given by*

$$e_{i,j} = \frac{a_{i,j}}{\lambda}\frac{\partial\lambda}{\partial a_{i,j}}, \quad i,j = 1,\dots,n. \tag{1.2}$$

If we regard $\partial\lambda/\partial a_{i,j}$ as the measure of the *sensitivity* of λ to a change in $a_{i,j}$, then we can view the elasticity with respect to the (i,j)th entry as the *proportional sensitivity of λ to a change in $a_{i,j}$*. We note that, from (1.2), $e_{i,j}$ also admits the representation as

$$e_{i,j} = \frac{\partial \log\lambda}{\partial \log a_{i,j}}, \quad i,j = 1,\dots,n.$$

Thus the elasticity can be thought of as measuring the multiplicative change in λ due to a multiplicative change in $a_{i,j}$, while the sensitivity measures the additive effect on λ arising from an additive change in $a_{i,j}$. Finally, we note that, in [6], it is shown that

$$\sum_{i,j=1}^{n} e_{i,j} = 1$$

so that

$$\sum_{i,j=1}^{n} e_{i,j}\lambda = \lambda.$$

In this way, the elasticities $e_{i,j}$ provide a quantification of the contribution of $a_{i,j}$ to the size of λ.

Throughout this paper, we will focus on the fundamental case that A is irreducible. This case is of both practical and theoretical interest, and it is well known that, in this case, the Perron root of A is simple so that, for each entry in A, the corresponding elasticity is well defined.

In [5, 9.7.1], Caswell discusses the sensitivity of the elasticities to changes in the vital rates and deduces from (1.2) that

$$\frac{\partial e_{i,j}}{\partial a_{k,\ell}} = \frac{a_{i,j}}{\lambda}\frac{\partial^2\lambda}{\partial a_{i,j}\partial a_{k,\ell}} - \frac{a_{i,j}}{\lambda^2}\frac{\partial\lambda}{\partial a_{k,\ell}}\frac{\partial\lambda}{\partial a_{i,j}} + \frac{\delta_{i,k}\delta_{j,\ell}}{\lambda}\frac{\partial\lambda}{\partial a_{i,j}}, \quad i,j,k,\ell = 1,\dots,n, \tag{1.3}$$

where $\delta_{p,q}$ is 1 or 0 according to whether $p = q$.

Fortunately, formulae are available for the various partial derivatives that appear in (1.3). For an $n \times n$ nonnegative and irreducible matrix A with Perron root λ and right and left Perron vectors x and w^T, respectively, normalized so that $w^T x = 1$, it is well known that

$$\frac{\partial \lambda}{\partial a_{i,j}} = w_i x_j, \quad i,j = 1, \ldots, n \tag{1.4}$$

(see Wilkinson [15] or Stewart [12], for example). Further, in [7], it is shown that

$$\frac{\partial^2 \lambda}{\partial a_{i,j} \partial a_{k,\ell}} = \frac{\partial \lambda}{\partial a_{i,\ell}} Q^{\#}_{j,k} + \frac{\partial \lambda}{\partial a_{k,j}} Q^{\#}_{\ell,i}, \quad i,j,k,\ell = 1, \ldots, n, \tag{1.5}$$

where $Q^{\#}$ is the group inverse of the singular irreducible M-matrix $Q = \lambda I - A$. (See Ben-Israel and Greville [2] and Campbell and Meyer [4] for background material on generalized inverses.) Substituting (1.4) and (1.5) into (1.3), we see that, in particular,

$$\frac{\partial e_{i,j}}{\partial a_{i,j}} = \frac{1}{\lambda} w_i x_j \left(2 a_{i,j} Q^{\#}_{j,i} - a_{i,j} \frac{1}{\lambda} w_i x_j + 1 \right). \tag{1.6}$$

Thus we find that the elasticity of the Perron root with respect to the (i,j)th entry is increasing as a function of the (i,j)th entry if and only if

$$2 a_{i,j} Q^{\#}_{j,i} - a_{i,j} \frac{1}{\lambda} w_i x_j + 1 \geq 0. \tag{1.7}$$

In this paper, we give two different proofs of the fact that, for each pair of indices i, j, the quantity $e_{i,j}$ is increasing as a function of $a_{i,j}$. The first proof, developed in section 2, is matrix theoretic, and it also yields a characterization of the case of equality in (1.7). The second proof, developed in section 3, relies on techniques from the theory of symbolic dynamics. In section 4, we give some closing remarks. For convenience, we now state our main result.

THEOREM 1.2. *Let $A = (a_{i,j})$ be an irreducible nonnegative matrix of order n with Perron root λ and right Perron vector $x = (x_1, \ldots, x_n)^T$. Then, for each $1 \leq i, j \leq n$, $e_{i,j}$ is an increasing function of $a_{i,j}$. Specifically,*

$$\frac{\partial e_{i,j}}{\partial a_{i,j}} \geq 0. \tag{1.8}$$

Moreover,

$$\frac{\partial e_{i,j}}{\partial a_{i,j}} = 0 \tag{1.9}$$

if and only if A is permutationally similar to the matrix $\lambda D \tilde{A} D^{-1}$, where D is the diagonal matrix whose ith diagonal entry is $x_i, i = 1, \ldots, n$, and where stochastic matrix $\tilde{A}$ is periodic and has the form

$$\tilde{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & X_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & X_2 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X_p \\ \mathbf{1} & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{1.10}$$

where $\mathbf{1}$ denotes the all-ones vector, where the ith row of A corresponds to the first row of $\tilde{A}$, and where X_1 has just one row, which corresponds to row j of A.

2. A matrix theoretic proof of Theorem 1.2. In this section, we give a matrix theoretic proof of all of the conclusions of Theorem 1.2.

Let $A = (a_{i,j})$ be an $n \times n$ irreducible nonnegative matrix whose Perron root is λ. If $x = (x_1, \ldots, x_n)^T$ is a right Perron vector of A and D is the diagonal matrix whose ith diagonal entry is x_i, $i = 1, \ldots, n$, then it is well known (see [3, Theorem 2.5.4]) that the matrix $\hat{A} := \frac{1}{\lambda} D^{-1} A D$ is stochastic. Let $Q = \lambda I - A$ and $\hat{Q} = I - \hat{A}$. Then $\hat{Q}^{\#} = \lambda D^{-1} Q^{\#} D$, and it is not difficult to show, from (1.6), that, if $E = (e_{i,j})$ and $\hat{E} = (\hat{e}_{i,j})$ are the matrices of elasticities arising from A and $\hat{A}$, respectively, then

$$\frac{\partial e_{i,j}}{\partial a_{i,j}} \;=\; \frac{1}{\lambda} \frac{x_j}{x_i} \frac{\partial \hat{e}_{i,j}}{\partial \hat{a}_{i,j}}, \quad i, j = 1, \ldots, n.$$

We thus conclude that

$$\operatorname{sign}\left(\frac{\partial e_{i,j}}{\partial a_{i,j}}\right) \;=\; \operatorname{sign}\left(\frac{\partial \hat{e}_{i,j}}{\partial \hat{a}_{i,j}}\right), \quad i, j = 1, \ldots, n,$$

and so, for our purposes in this section, it will suffice to consider the case that our original $n \times n$ irreducible nonnegative matrix A is stochastic.

We begin with the following lemma.

LEMMA 2.1. *Let B be a substochastic matrix of order $n \geq 2$ whose spectral radius is less than 1. Fix an index j, with $1 \leq j \leq n$, and, for each $l \in \mathbb{N}$, let $\alpha_l = e_j^T B^l \mathbf{1}$. Then*

$$\sum_{l=1}^{\infty} \alpha_l^2 + 2 \sum_{l=1}^{\infty} \sum_{m=l+1}^{\infty} \alpha_l \alpha_m \;\leq\; \sum_{l=1}^{\infty} \alpha_l + 2 \sum_{l=1}^{\infty} (l-1) \alpha_l. \tag{2.1}$$

Suppose further that each vertex in the digraph of B can be reached from j by some walk. Then, if equality holds in (2.1)*, there is a $p \in \mathbb{N}$ such that B can be permuted to the form*

$$\begin{bmatrix} 0 & X_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & X_2 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X_p \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $X_i \mathbf{1} = \mathbf{1}$ for $i = 1, \ldots, p$, and where X_1 has only one row, which corresponds to index j.

Proof. For each $l \in \mathbb{N}$, $0 \leq \alpha_l \leq 1$, so we see that $\sum_{l=1}^{\infty} \alpha_l^2 \leq \sum_{l=1}^{\infty} \alpha_l$. Also,

$$\sum_{l=1}^{\infty} \sum_{m=l+1}^{\infty} \alpha_l \alpha_m \;=\; \sum_{m=2}^{\infty} \sum_{l=1}^{m-1} \alpha_l \alpha_m \;\leq\; \sum_{m=2}^{\infty} (m-1) \alpha_m.$$

The inequality (2.1) now follows readily.

Suppose now that equality holds in (2.1). Then, in particular, we must have that $\alpha_l = \alpha_l^2$ for each l so that α_l is either 1 or 0 for each l. Note that, since $B^l \to 0$ as $l \to \infty$, we see that $\alpha_p = 0$ for some p. However, that implies that, in the digraph of B, there is no walk of length p starting from vertex j and hence no walk of length longer than p starting from j. (Note, in particular, that the digraph has no cycles.) We conclude that, for some p, we have $\alpha_l = 1$ if $l \leq p$ and $\alpha_l = 0$ if $l \geq p + 1$.

We claim that this last condition implies that the vertices in the digraph of B which are distinct from j can be partitioned into sets $S_1, \ldots, S_p$ such that, for each i, S_i is the set of vertices v such that the distance from j to v is i. We prove the claim by induction and note that, for the case when $p = 1$, each vertex distinct from j must be in the outset of j, giving the desired partitioning. Next, suppose that the claim holds for some $p \geq 1$ and that we have that $\alpha_l = 1$ if $l \leq p+1$ and $\alpha_l = 0$ if $l \geq p+2$. Let S_1 be the outset of j, and note that, for each $l \geq 2$, $\alpha_l = \sum_{i=1}^{n} b_{j,i}\alpha_{i,l-1}$, where $\alpha_{i,l-1} = e_i^T B^{l-1}\mathbf{1}$. It follows that $\alpha_{i,l} = 1$ for $1 \leq l \leq p$ and $\alpha_{i,l} = 0$ for $l \geq p+1$. Thus, for each vertex $i \in S_1$, the induction hypothesis applies to those vertices reachable from i, yielding a corresponding partitioning of the vertex set. However, a vertex at a distance d from i is necessarily at a distance $d+1$ from j, and the desired partitioning follows, completing the induction step.

From the above claim, it now follows that we can write B in the form

$$\begin{bmatrix} 0 & X_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & X_2 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X_p \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Finally, the fact that $X_i\mathbf{1} = \mathbf{1}$ for $i = 1, \ldots, p$ now follows since $\alpha_l = 1$ for $1 \leq l \leq p$. ∎

Lemma 2.1 yields the following corollary.

COROLLARY 2.2. *Let B be as in Lemma* 2.1, *and fix an index j. Then*

$$e_j^T(I-B)^{-1}\mathbf{1} + \left[e_j^T(I-B)^{-1}\mathbf{1}\right]^2 \leq 2e_j^T(I-B)^{-2}\mathbf{1}. \tag{2.2}$$

Suppose also that each vertex in the digraph of B can be reached from j by some walk. If equality holds in (2.2), *then there is a $p \in \mathbb{N}$ such that B can be written as*

$$\begin{bmatrix} 0 & X_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & X_2 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X_p \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $X_i\mathbf{1} = \mathbf{1}$ for $i = 1, \ldots, p$ and where X_1 has only one row, which corresponds to index j.

Proof. As in Lemma 2.1, we let $\alpha_l = e_j^T B^l \mathbf{1}$ for each $l \in \mathbb{N}$. Note that $(I-B)^{-1} = \sum_{l=0}^{\infty} B^l$ and that $(I-B)^{-2} = \sum_{l=1}^{\infty} lB^{l-1}$. Thus we see that $2e_j^T(I-B)^{-2}\mathbf{1} = 2 + 2\sum_{l=2}^{\infty} l\alpha_{l-1}$, while $e_j^T(I-B)^{-1}\mathbf{1} = 1 + \sum_{l=1}^{\infty}\alpha_l$. Consequently, the inequality

$$2e_j^T(I-B)^{-2}\mathbf{1} \geq e_j^T(I-B)^{-1}\mathbf{1} + \left[e_j^T(I-B)^{-1}\mathbf{1}\right]^2$$

is equivalent to the inequality

$$2 + 2\sum_{l=2}^{\infty} l\alpha_{l-1} \geq 1 + \sum_{l=1}^{\infty}\alpha_l + \left(1 + \sum_{l=1}^{\infty}\alpha_l\right)^2.$$

However, this last inequality is easily seen to simplify to

$$\sum_{l=1}^{\infty} \alpha_l + 2\sum_{l=1}^{\infty}(l-1)\alpha_l \;\geq\; \sum_{l=1}^{\infty}\alpha_l^2 + 2\sum_{l=1}^{\infty}\sum_{m=l+1}^{\infty}\alpha_l\alpha_m.$$

The results, including the equality case, now follow from Lemma 2.1. ∎

Consider the case when $i = j$ in (1.7) for a stochastic matrix A with left Perron vector w^T, normalized so that its entries sum to 1. In that situation, the left side of (1.7) becomes $2a_{i,i}Q^{\#}_{i,i} - a_{i,i}w_i + 1$. It is shown in [7] that $Q^{\#}_{i,i} > 0$ for each $i = 1, \ldots, n$, and it follows readily then that $\frac{\partial e_{i,i}}{\partial a_{i,i}} \geq w_i(1 - a_{i,i}w_i) > 0$. Thus, in order to establish Theorem 1.2, we need only consider the case when $i \neq j$. That case is (essentially) considered in the following proposition.

PROPOSITION 2.3. *Let A be an irreducible stochastic matrix of order $n \geq 3$, written as*

$$A = \left[\begin{array}{c|ccc} m_0 & m_1 & \cdots & m_{n-1} \\ \hline y & & B & \end{array}\right].$$

Then, for each $1 \leq j \leq n-1$,

$$\frac{\partial e_{1,j+1}}{\partial a_{1,j+1}} \;\geq\; 0. \tag{2.3}$$

Furthermore,

$$\frac{\partial e_{1,j+1}}{\partial a_{1,j+1}} \;=\; 0 \tag{2.4}$$

if and only if A is permutationally similar to a matrix $\tilde{A}$ of the form

$$\tilde{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & X_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & X_2 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X_p \\ \mathbf{1} & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{2.5}$$

where the first row of A corresponds to the first row of $\tilde{A}$ and where X_1 has just one row, which corresponds to row $j+1$ of A.

Proof. Let $m^T = (m_1, \ldots, m_{n-1}) \in \mathbb{R}^{1,n-1}$, and let w^T be the left Perron vector for A whose entries sum to 1. Note that $w_1 = 1/[1 + m^T(I-B)^{-1}\mathbf{1}]$. Also, if $Q = I - A$, then, for $j = 1, \ldots, n-1$, we have, using Meyer [10, (5.1)] in conjunction with a permutation similarity, that

$$Q^{\#}_{1,j+1} \;=\; w_1^2 m^T(I-B)^{-2}\mathbf{1} - w_1 e_j^T(I-B)^{-1}\mathbf{1}.$$

It now follows, from (1.6), that

$$\begin{aligned} \frac{1}{w_1^3}\,\frac{\partial e_{1,j+1}}{\partial a_{1,j+1}} \;&=\; 2m_j m^T(I-B)^{-2}\mathbf{1} \\ &\quad - 2m_j e_j^T(I-B)^{-1}\mathbf{1}\left[1 + m^T(I-B)^{-1}\mathbf{1}\right] \\ &\quad - m_j\left[1 + m^T(I-B)^{-1}\mathbf{1}\right] \;+\; \left[1 + m^T(I-B)^{-1}\mathbf{1}\right]^2 \;\equiv\; f. \end{aligned} \tag{2.6}$$

Suppose that $i \neq 0, j$, and note that

$$\begin{aligned}\frac{\partial f}{\partial m_i} &= m_j \left[2e_i^T(I-B)^{-2}\mathbf{1} - e_i^T(I-B)^{-1}\mathbf{1}\right] \\ &\quad + 2e_i^T(I-B)^{-1}\mathbf{1}\left[1 + m^T(I-B)^{-1}\mathbf{1} - m_j e_j^T(I-B)^{-1}\mathbf{1}\right] \geq 0.\end{aligned}$$

Thus, in order to show that f is nonnegative, it suffices to show that fact in the case when, for some $t \in [0,1]$, $m_j = t$ and $m_0 = 1-t$. However, in that instance, f reduces to

$$1 - t + t^2\left[2e_j^T(I-B)^{-2}\mathbf{1} - e_j^T(I-B)^{-1}\mathbf{1} - (e_j^T(I-B)^{-1}\mathbf{1})^2\right].$$

Appealing to Corollary 2.2, we see that $f \geq 0$.

Next suppose that $f = 0$. We find, from the above, that necessarily $m_j = 1$ and

$$2e_j^T(I-B)^{-2}\mathbf{1} = e_j^T(I-B)^{-1}\mathbf{1} + \left[e_j^T(I-B)^{-1}\mathbf{1}^2\right].$$

Since A is irreducible, the spectral radius of B is less than 1. Also, since each vertex in the digraph of A can be reached by a walk starting from vertex 1 and since each walk starting from 1 must pass immediately through $j+1$, we see that each vertex distinct from 1 and $j+1$ is reachable by a walk from $j+1$ which is contained in the digraph of B. Thus we see that all of the hypotheses of Corollary 2.2 apply to B. As a result, there is a p such that we may write B as

$$\begin{bmatrix} 0 & X_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & X_2 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & X_p \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $X_i\mathbf{1} = \mathbf{1}$ for $i = 1, \ldots, p$ and where the X_1 has only one row which corresponds to index $j+1$. Using the irreducibility and stochasticity of A, we deduce that A is permutationally similar to a matrix $\tilde{A}$ having the form of (2.5).

Finally, if A is permutationally similar to the matrix $\tilde{A}$ of (2.5), then it follows from results in [9] that $Q^{\#}_{j+1,1} = -(p+1)/(2p+4)$, while $w_1 = 1/(p+2)$ so that, by (1.6),

$$\frac{\partial e_{1,j+1}}{\partial a_{1,j+1}} = 0. \qquad \square$$

3. A proof of (1.8) via symbolic dynamics. Our second proof of (1.8) relies on a well-known principle from the theory of symbolic dynamical systems, the variational principle for pressure. An adapted form of that principle is given below and is stated in the language of nonnegative matrices. A more authentic form of this principle can be found in Walters [14] and in Arnold, Gundlach, and Demetrius [1].

THEOREM 3.1 (variational principle for pressure, restated). *Let $A = (a_{i,j})$ be an $n \times n$ irreducible nonnegative matrix with Perron root λ_A and right Perron vector x. Let $\mathcal{M}_A$ be the collection of all $n \times n$ stochastic nonnegative matrices $P = (p_{i,j})$ such that $a_{i,j} = 0 \Leftrightarrow p_{i,j} = 0$. For each $P \in \mathcal{M}_A$, let r_P be the left Perron vector of P whose entries sum to 1. Then*

$$\log \lambda_A = \sup_{P \in \mathcal{M}_A} \left\{ -\sum_{i,j=1}^{n} (r_P)_i p_{i,j} \log p_{i,j} + \sum_{i,j=1}^{n} (r_P)_i p_{i,j} \log a_{i,j} \right\}.$$

Furthermore, the supremum is achieved at the stochastic matrix P_A such that, for each $1 \le i, j \le n$,

$$(P_A)_{i,j} \;=\; \frac{a_{i,j}x_j}{\lambda_A x_i}. \tag{3.1}$$

In our next result, we apply the characterization of the Perron root given in Theorem 3.1 in order to describe the elasticity with respect to a particular entry of A.

PROPOSITION 3.2. *The elasticity $e_{k,\ell}$ evaluated at the matrix $A = (a_{i,j})$ is equal to $(r_{P_A})_k (P_A)_{k,\ell}$.*

Proof. Let w and x be positive vectors satisfying $Ax = \lambda_A x$ and $w^T A = \lambda_A w^T$, and note that

$$e_{k,\ell} \;=\; \frac{a_{k,\ell}}{\lambda_A}\frac{w_k x_\ell}{w^T x}.$$

Notice also that $r_{P_A}^T = (1/w^T x)(x_1 w_1, x_2 w_2, \ldots, x_n w_n)$ since, for each $1 \le j \le n$, we have

$$\begin{aligned}\sum_{i=1}^{n} \frac{1}{w^T x} x_i w_i (P_A)_{i,j} \;&=\; \sum_{i=1}^{n} \frac{1}{w^T x} x_i w_i \frac{a_{i,j}x_j}{\lambda_A x_i} \\ &=\; \sum_{i=1}^{n} \frac{1}{w^T x}\frac{w_i a_{i,j} x_j}{\lambda_A} \\ &=\; \frac{1}{w^T x} x_j w_j,\end{aligned} \tag{3.2}$$

while clearly

$$\left\| \frac{1}{w^T x}(x_1 w_1, x_2 w_2, \ldots, x_n w_n) \right\|_1 \;=\; 1.$$

We thus conclude that

$$(r_{P_A})_k (P_A)_{k,\ell} \;=\; \frac{1}{w^T x} w_k x_k \frac{a_{k,\ell}x_\ell}{\lambda_A x_k} \;=\; \frac{a_{k,\ell}}{\lambda_A}\frac{w_k x_\ell}{w^T x} \;=\; e_{k,\ell}. \qquad \square$$

With Theorem 3.1 and Proposition 3.2 in mind, fix an ordered pair (k, ℓ), $1 \le k, \ell \le n$, and let $B = (b_{i,j})$ be a nonnegative matrix whose entries are as follows: $b_{i,j} = a_{i,j}$ for all $(i, j) \ne (k, \ell)$ and $b_{k,\ell} > a_{k,\ell}$. Denoting $e_{k,\ell}$ evaluated at A and at B by $e_{k,\ell}|_A$ and $e_{k,\ell}|_B$, respectively, we see that, if $a_{k,\ell} = 0$, then $e_{k,\ell}|_A = 0$ while $e_{k,\ell}|_B > 0$ so that $e_{k,\ell}|_B > e_{k,\ell}|_A$. Thus, if $a_{k,\ell} = 0$, then $e_{k,\ell}$ is increasing in $a_{k,\ell}$.

Next assume that $a_{k,\ell} > 0$, and let λ_B be the Perron root of B. Since $B \ge A$, but A and B differ only in the (k, ℓ) position (where each has a positive entry), we see that $\mathcal{M}_A = \mathcal{M}_B$. Then, by Theorem 3.1,

$$\begin{aligned}\log \lambda_B \;&\ge\; -\sum_{i,j} (r_{P_A})_i (P_A)_{i,j} \log (P_A)_{i,j} + \sum_{i,j} (r_{P_A})_i (P_A)_{i,j} \log b_{i,j} \\ &=\; \log \lambda_A + (r_{P_A})_k (P_A)_{k,\ell} \left[\log b_{k,\ell} - \log a_{k,\ell}\right].\end{aligned} \tag{3.3}$$

From (3.3) and Proposition 3.2, we see that

$$\frac{\log \lambda_B - \log \lambda_A}{\log b_{k,\ell} - \log a_{k,\ell}} \;\ge\; e_{k,\ell}|_A. \tag{3.4}$$

Similarly, we also find from Theorem 3.1 that

$$\begin{aligned}\log \lambda_A &\geq -\sum_{i,j}(r_{P_B})_i(P_B)_{i,j}\log(P_B)_{i,j} + \sum_{i,j}(r_{P_B})_i(P_B)_{i,j}\log a_{i,j} \\ &= \log\lambda_B + (r_{P_B})_k(P_B)_{k,\ell}\left[\log a_{k,\ell} - \log b_{k,\ell}\right].\end{aligned} \tag{3.5}$$

Applying Proposition 3.2 to the matrix B, it follows from (3.5) that

$$e_{k,\ell}|_B \geq \frac{\log\lambda_B - \log\lambda_A}{\log b_{k,\ell} - \log a_{k,\ell}}. \tag{3.6}$$

Consequently, from (3.4) and (3.6), we find that

$$e_{k,\ell}|_B \;\geq\; e_{k,\ell}|_A$$

so that $e_{k,\ell}$ is nondecreasing in the (k,ℓ) entry of A. In particular, (1.8) follows readily.

4. Examples and remarks. We begin this section with an example illustrating the case of equality in Theorem 1.2. Let

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

Then calculations show that $\lambda = \sqrt{2}$, and the matrix of elasticities is given by

$$E = (e_{i,j}) = \begin{bmatrix} 0 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 1/8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/16 & 1/16 \\ 0 & 0 & 0 & 0 & 1/16 & 1/16 \\ 1/8 & 0 & 0 & 0 & 0 & 0 \\ 1/8 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

From Theorem 1.2, we anticipate that $\partial e_{i,j}/\partial a_{i,j}$ is 0 only for $i = 1$ and $j = 2$, while the remaining quantities are positive. This is indeed the case; our computations yield

$$\left(\frac{\partial e_{i,j}}{\partial a_{i,j}}\right) = \begin{bmatrix} 0.17678 & \boxed{0} & 0.17678 & 0.17678 & 0.12500 & 0.12500 \\ 0.12500 & 0.17678 & 0.062500 & 0.062500 & 0.088388 & 0.088388 \\ 0.088388 & 0.12500 & 0.088388 & 0.088388 & 0.046875 & 0.046875 \\ 0.088388 & 0.12500 & 0.088388 & 0.088388 & 0.046875 & 0.046875 \\ 0.062500 & 0.17678 & 0.12500 & 0.12500 & 0.088388 & 0.088388 \\ 0.062500 & 0.17678 & 0.12500 & 0.12500 & 0.088388 & 0.088388 \end{bmatrix}.$$

Theorem 1.2 shows that, for an $n \times n$ nonnegative irreducible matrix $A = (a_{i,j})$, $\partial e_{i,j}/\partial a_{i,j}$ is bounded below by 0; the following example shows that these derivatives are not bounded from above. Let J be the $n \times n$ all-ones matrix, let $\alpha \in (0,1)$, and let $A = (\alpha/n)J$ so that the Perron root of A is α. Let $Q = \alpha I - A$. It follows readily that $Q^{\#} = (1/\alpha^2)Q$. In this case, $a_{i,i} = \alpha/n$, $Q^{\#}_{i,i} = (1/\alpha)(1 - 1/n)$, and $\partial\lambda/\partial a_{i,i} = 1/n$ for all $1 \le i \le n$. Substituting these three expressions in (1.6), we obtain that, for all $1 \le i \le n$,

$$\frac{\partial e_{i,i}}{\partial a_{i,i}} \;=\; \frac{n^2 + 2n - 3}{\alpha n^3}$$

and, similarly, that, for distinct indices i, j with $1 \le i, j \le n$, we have

$$\frac{\partial e_{i,j}}{\partial a_{i,j}} \;=\; \frac{n^2 - 3}{\alpha n^3}.$$

Observe that each of these quantities can be made arbitrarily large by choosing the positive parameter α sufficiently close to 0.

We close with a consequence of (1.8). Suppose that we have an irreducible stochastic matrix A of order n, and let w^T denote its left Perron vector, normalized so that its entries sum to 1. (In particular, A can be thought of as the transition matrix of a Markov chain with stationary distribution vector w.) Letting $Q = I - A$, it turns out that the modulii of the entries in $Q^{\#}$ can be used to measure the stability of the computation of w^T. Specifically, Funderlic and Meyer [8] propose $\max_{i,j=1,\ldots,n} |Q^{\#}_{i,j}|$ as a condition number for the Markov chain, while Meyer [11] suggests $||Q^{\#}||_\infty$ as a condition number for the chain. From (1.8), we find that, for each pair of indices $1 \le i, j \le n$,

$$2a_{i,j}Q^{\#}_{j,i} - a_{i,j}w_i + 1 \;\ge\; 0. \tag{4.1}$$

Since $w^T Q^{\#} = 0^T$ and $Q^{\#}\mathbf{1} = 0$ and since the diagonal entries of $Q^{\#}$ are positive, it follows that $Q^{\#}$ has at least one negative entry in each row and column. Suppose now that $Q^{\#}_{j,i} < 0$ and that $a_{i,j} > 0$. Then, from (4.1), we see that

$$|Q^{\#}_{j,i}| \;=\; -Q^{\#}_{j,i} \;\le\; \frac{1 - a_{i,j}w_i}{2a_{i,j}} < \frac{1}{2a_{i,j}}. \tag{4.2}$$

Thus we find that (1.8) can be used to provide an upper bound on the modulii of some of the negative entries of $Q^{\#}$ in terms of the entries in A and w^T. This observation may be useful in discussing the condition numbers mentioned above.

REFERENCES

[1] L. Arnold, V. Gundlach, and L. Demetrius, *Evolutionary formalism for products of positive random matrices*, Ann. Appl. Probab., 4 (1994), pp. 859–901.

[2] A. Ben-Israel and T. N. Greville, *Generalized Inverses: Theory and Applications*, Academic Press, New York, 1973.

[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics in Appl. Math. 9, SIAM, Philadelphia, 1994.

[4] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover, New York, 1991.

[5] H. Caswell, *Matrix Population Models: Construction, Analysis, and Interpretation*, 2nd ed., Sinauer, Sunderland, MA, 2001.

[6] H. De Kroon, A. Plaisier, J. van Groenendael, and H. Caswell, *Elasticity: The relative contribution of demographic parameters to population growth rate*, Ecology, 65 (1986), pp. 1427–1431.
[7] E. Deutsch and M. Neumann, *Derivatives of the Perron root at an essentially nonnegative matrix and the group inverse of an M-matrix*, J. Math. Anal. Appl., 102 (1984), pp. 1–29.
[8] R. E. Funderlic and C. D. Meyer, Jr., *Sensitivity of the stationary distribution vector for an ergodic Markov chain*, Linear Algebra Appl., 76 (1986), pp. 1–17.
[9] S. J. Kirkland, *The group inverse associated with an irreducible periodic nonnegative matrix*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1127–1134.
[10] C. D. Meyer, Jr., *The role of the group generalized inverse in the theory of finite Markov chains*, SIAM Rev., 17 (1975), pp. 443–464.
[11] C. D. Meyer, Jr., *The condition of a finite Markov chain and perturbations bounds for the limiting probabilities*, SIAM J. Algebraic Discrete Methods, 1 (1980), pp. 273–283.
[12] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[13] R. S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[14] P. Walters, *An Introduction to Ergodic Theory*, Springer-Verlag, New York, 1982.
[15] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 465–474

A TANDEM QUEUE WITH A MOVABLE SERVER: AN EIGENVALUE APPROACH*

WINFRIED K. GRASSMANN† AND JAVAD TAVAKOLI‡

Abstract. In this paper, we analyze a two station tandem queue with Poisson arrivals and exponential service times. All arrivals occur at the first line, and, after receiving service at the first station, they proceed to the second line. There is only a finite buffer between the stations, and, as soon as the buffer is full, any job completed by the first server is lost. To reduce customer loss, the first server can move to the second station and help the second server, thereby increasing its service rate. Once the work at station two is complete, the job leaves the system. The problem will be solved by using eigenvalues which can be obtained in explicit form. It is shown that this method is substantially faster than matrix analytic methods.

Key words. tandem queues, movable servers, eigenvalues, quasi-birth-and-death (QBD) processes

AMS subject classifications. 60K25, 90B22, 15A22, 65F15

PII. S0895479801394088

1. Introduction.

In this paper, we investigate a model with Poisson arrivals, exponential service times, and two queues in tandem. All arrivals first join the first queue, and, once served at the first station, they proceed to the second station. They depart after having received service at the second station. There is a finite buffer between the stations, and, as soon as the buffer is full, a job completed by the first server is either lost or else leaves the system. To reduce customer loss, the server at the first station can move to help the server of the second one, thereby increasing the service rate of the second station.

Queueing problems with movable servers have attracted some interest recently. In particular, in apparel manufacture, Bischak [3] found that movable servers increase flexibility and reduce assembly line imbalance while at the same time increasing worker satisfaction. Moreover, as pointed out by Andradottir, Ayhan, and Down [1], throughput can be increased if an idle worker can move to help another worker.

Tandem queues with finite buffers and loss but without movable servers have a product form solution [13]. However, this product form solution is lost when servers can move between stations, and more complex methods must be employed. The method used here is based on eigenvalues, and this method is extremely efficient because the eigenvalues in question can be obtained explicitly. In fact, if the buffer size is denoted by N, there are exactly $N+1$ eigenvalues, and each can be found in constant time, which leads to a computational complexity of $O(N)$. On the other hand, using matrix analytic methods requires several matrix multiplications per iteration, and matrix multiplications have a complexity of $O(N^3)$. Since matrix analytic methods often require many iterations [4], or many matrix multiplications per iteration [9], this

*Received by the editors August 21, 2001; accepted for publication (in revised form) by D. O'Leary April 19, 2002; published electronically November 6, 2002. This research was supported by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/simax/24-2/39408.html
†Department of Computer Science, University of Saskatchewan, 57 Campus Drive, Saskatoon, Saskatchewan, S7N 5A9, Canada (grassman@cs.usask.ca).
‡Science Department, SIFC Regina Campus, Room 118, Regina, Saskatchewan, S4S 0A2, Canada (jtavakoli@sifc.edu).

means that matrix analytic methods are often literally more than 100 times slower than the method suggested here.

2. The model. The problem under consideration has two state variables, namely, the lengths of lines 1 and 2. These variables will be denoted by X_1 and X_2, respectively. The arrival rate to the first line is denoted by λ. The service rate of the first station is μ_1, and the rate of the second station is μ_2. The objective is to find $\pi_{i,j}$ for all i and j, where $\pi_{i,j}$ is the steady-state probability that $X_1 = i$ and $X_2 = j$. We assume $\mu_1 > \lambda$ to ensure that a steady-state solution exists. Because of loss, we need not put any restriction on μ_2; that is, μ_2 can be either less than or greater than λ. The transition matrix Q is block-structured with the blocks $Q_{i,j}$ containing all transitions where X_1 changes from i to j. Except for $Q_{0,0}$, the $Q_{i,j}$ depend only on the difference between i and j, and we therefore set

$$Q_{j-i} = Q_{j,i}.$$

Since X_1 can only change by 1, the only nonzero Q_j are Q_1, Q_0, and Q_{-1}. They are defined by the $(N+1) \times (N+1)$ matrices

$$Q_1 = \lambda I,$$

$$Q_0 = \begin{bmatrix} -(\lambda+\mu_1) & 0 & \dots & \dots & 0 \\ \mu_2 & -(\lambda+\mu_1+\mu_2) & 0 & \ddots & \vdots \\ 0 & \mu_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu_2 & -(\lambda+\mu_1+\mu_2) \end{bmatrix},$$

$$Q_{-1} = \begin{bmatrix} 0 & \mu_1 & 0 & \dots & 0 \\ 0 & 0 & \mu_1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mu_1 \\ 0 & 0 & \dots & \dots & \mu_1 \end{bmatrix}.$$

Furthermore, the matrix $Q_{0,0}$ reflects the fact that, when the first server is idle, it helps the second server, and this means that the rate of the second server is μ_3 at this time. Hence

$$Q_{0,0} = \begin{bmatrix} -(\lambda) & 0 & \dots & \dots & 0 \\ \mu_3 & -(\lambda+\mu_3) & 0 & \ddots & \vdots \\ 0 & \mu_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu_3 & -(\lambda+\mu_3) \end{bmatrix}.$$

The transition matrix can now be written as

$$Q = \begin{bmatrix} Q_{0,0} & Q_1 & 0 & \dots \\ Q_{-1} & Q_0 & Q_1 & \ddots \\ 0 & Q_{-1} & Q_0 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

If the vector π_n contains the equilibrium probabilities for which $X_1 = n$, one can write the equilibrium equations in block form as follows:

$$0 = \pi_0 Q_{00} + \pi_1 Q_{-1}, \tag{2.1}$$

$$0 = \pi_{n-1} Q_1 + \pi_n Q_0 + \pi_{n+1} Q_{-1}, \quad n > 0. \tag{2.2}$$

According to Bertsimas [2], Mitrani and Chakka [10], and Morse [11], (2.2) has solutions of the form

$$\pi_n = g x^n, \tag{2.3}$$

where $g = [g_0, g_1, \ldots, g_N]$ must be different from 0. As will be shown, this problem has several solutions. Of these solutions, only the ones satisfying $|x| < 1$ are of interest, because any solution with $|x| \geq 1$ would not converge. Substituting (2.3) into (2.2) yields

$$0 = g x^{n-1} Q_1 + g x^n Q_0 + g x^{n+1} Q_{-1}. \tag{2.4}$$

If we define

$$Q(x) = Q_1 + Q_0 x + Q_{-1} x^2,$$

then (2.4) implies

$$0 = gQ(x).$$

The problem is to find the eigenvalues x and the corresponding eigenvectors $g \neq 0$ which satisfy $gQ(x) = 0$. It is known [5] that, in the context of recurrent queues, $Q(x)$ has exactly $N+1$ eigenvalues with $|x| < 1$. Moreover, if, in such a matrix, Q_0 is lower diagonal with all subdiagonal elements positive, then all eigenvalues inside the unit circle are real and between 0 and 1 (see [6, Corollary 1]). Hence there are exactly $N + 1$ solutions of the form given by (2.3) inside the unit circle, and these solutions can be combined to satisfy (2.1).

3. Difference equations for the eigenvectors. The approach used is similar to that in [7], and we repeat it here for reference. In view of the matrices Q_1, Q_0 and Q_{-1}, $gQ(x) = 0$ expands to give

$$0 = -g_0((\lambda + \mu_1)x - \lambda) + g_1 \mu_2 x, \tag{3.1}$$

$$0 = g_{i-1}\mu_1 x^2 - g_i((\lambda + \mu_1 + \mu_2)x - \lambda) + g_{i+1}\mu_2 x, \quad i = 1, 2, \ldots, N-1, \tag{3.2}$$

$$0 = g_{N-1}\mu_1 x^2 - g_N(-\mu_1 x^2 + (\mu_1 + \mu_2 + \lambda)x - \lambda). \tag{3.3}$$

The solution $x = 0$ can be excluded because, in this case, the vector g is zero, and only the trivial solution is possible. Similarly, if $g_0 = 0$, then the entire vector $g = 0$. Hence we can set $g_0 = 1$, and, since $x \neq 0$, (3.1) yields

$$g_1 = \frac{\lambda + \mu_1 - \lambda/x}{\mu_2}. \tag{3.4}$$

The remaining g_i can now be calculated by solving (3.2), which yields

$$g_{i+1} = g_i \frac{\lambda + \mu_1 + \mu_2 - \lambda/x}{\mu_2} - g_{i-1}\frac{\mu_1}{\mu_2}x, \quad i = 1, 2, \ldots, N-1. \tag{3.5}$$

Given these g_i, any x satisfying (3.3) is obviously an eigenvalue. It is convenient to introduce g_{N+1} as follows:

$$g_{N+1} = g_N \frac{(\mu_1 + \mu_2 + \lambda) - \mu_1 x - \lambda/x}{\mu_2} - g_{N-1} \frac{\mu_1}{\mu_2} x. \tag{3.6}$$

Here g_{N+1} is a function of x, and every eigenvalue x must satisfy $g_{N+1}(x) = 0$.

It is well known from the theory of difference equations that $g_i = y^i$ for an appropriate y is a solution of (3.2). If we substitute y^i for g_i in (3.2), we get the following quadratic equation in terms of y:

$$0 = y^2 - y \frac{\lambda + \mu_1 + \mu_2 - \lambda/x}{\mu_2} + \frac{\mu_1}{\mu_2} x. \tag{3.7}$$

Let

$$b(x) = \frac{\lambda + \mu_1 + \mu_2 - \lambda/x}{\mu_2}, \tag{3.8}$$

$$d(x) = b(x)^2 - 4 \frac{\mu_1}{\mu_2} x. \tag{3.9}$$

The solutions of (3.7) can then be written as

$$y_1 = \frac{b(x) - \sqrt{d(x)}}{2}, \qquad y_2 = \frac{b(x) + \sqrt{d(x)}}{2}. \tag{3.10}$$

We will be able to find all $N+1$ eigenvalues without having to resort to any eigenvalue x such that $d(x) = 0$, and we therefore assume $d(x) \neq 0$. Considering also the fact that $g_0 = 1$ and $g_1 = b(x) - 1$, we find, after some calculation (see [7] for details),

$$g_i = \frac{1}{\sqrt{d(x)}} \left((y_2 - 1) y_2^i - (y_1 - 1) y_1^i \right). \tag{3.11}$$

From (3.10), $y_1 + y_2 = b(x)$, and $y_1 y_2 = \frac{\mu_1}{\mu_2} x$, we obtain

$$\begin{aligned} g_{N+1} &= g_N \left(b(x) - \frac{\mu_1}{\mu_2} x \right) - g_{N-1} \frac{\mu_1}{\mu_2} x \\ &= g_N \left(y_1 + y_2 - y_1 y_2 \right) - g_{N-1} y_1 y_2. \end{aligned}$$

Substituting (3.11) into the above equation yields

$$g_{N+1} = \frac{1}{\sqrt{d(x)}} \left[(y_1 - 1)(y_2 - 1)(y_1^{N+1} - y_2^{N+1}) \right]. \tag{3.12}$$

If $y_1^{N+1} - y_2^{N+1} \neq 0$, then $g_{N+1}(x) = 0$ implies that $y = 1$ is a root of (3.7). If this is the case, it follows that $x_0 = \lambda/\mu_1$ is the only eigenvalue inside the unit circle. To find the remaining N eigenvalues, we consider the case where y is complex. Using arguments similar to those given in [7], one can show that these eigenvalues must either be outside the unit interval or must satisfy

$$\frac{3}{4} \frac{\lambda}{\lambda + \mu_1 + \mu_2} < x < 3 \frac{\lambda}{\lambda + \mu_1 + \mu_2}. \tag{3.13}$$

To deal with the complex case, we express y_1 and y_2 in polar coordinates as follows:

$$y_1 = (\cos\phi - i\sin\phi)r, \qquad y_2 = (\cos\phi + i\sin\phi)r.$$

Here r and $\cos\phi$ are given by

$$r = \sqrt{\frac{\mu_1}{\mu_2}x}, \qquad \cos\phi = \frac{b(x)}{2r}. \tag{3.14}$$

Because of (3.12), $g_{N+1} = 0$ if $y_1^{N+1} - y_2^{N+1} = 0$. Since y_1^{N+1} and y_2^{N+1} are conjugate complex, one has

$$y_1^{N+1} - y_2^{N+1} = -2ir^{N+1}\sin((N+1)\phi).$$

Consequently, $g_{N+1} = 0$ if $\sin((N+1)\phi) = 0$. Clearly, this is the case if $\phi = \phi_\nu = \pi\,\nu/(N+1)$, and, since we consider only complex zeros, $(d(x) < 0)$, ν runs from 1 to N. This yields N eigenvalues, say, $x_1, x_2, \ldots, x_N$, which can be determined according to the equation

$$\cos(\nu\pi/(N+1)) = b(x_\nu)/(2r) = \frac{\lambda + \mu_1 + \mu_2 - \lambda/x_\nu}{2\sqrt{x_\nu\mu_1\mu_2}}, \quad \nu = 1, 2, \ldots, N. \tag{3.15}$$

We note that this result is substantially simpler than that obtained in [7], where customers are blocked rather than lost. Equation (3.15) can be converted into an equation of the third degree, and, for these equations, closed form solutions are available (see, e.g., [12]). These solutions simplify considerably if the quadratic term vanishes. This can be achieved by introducing $z = 1/\sqrt{x}$. If $\cos(\nu\pi/(N+1))$ is denoted by a, one finds, after some calculation,

$$z^3 - \frac{\lambda + \mu_1 + \mu_2}{\lambda}z + 2a\frac{\sqrt{\mu_1\mu_2}}{\lambda} = 0. \tag{3.16}$$

According to [12], solving equations of the third degree involves the calculation of intermediate results Q, R, and D as follows:

$$Q = -\frac{\lambda + \mu_1 + \mu_2}{3\lambda},$$

$$R = a\frac{\sqrt{\mu_1\mu_2}}{\lambda},$$

$$D = Q^3 + R^2 = -\frac{(\lambda + \mu_1 + \mu_2)^3}{(3\lambda)^3} + a^2\frac{\mu_1\mu_2}{\lambda^2}.$$

In our case, D is less than zero because

$$\frac{\lambda + \mu_1 + \mu_2}{3} > \sqrt[3]{\mu_1\mu_2\lambda} > a^{2/3}\sqrt[3]{\mu_1\mu_2\lambda}.$$

If $D < 0$, one must calculate

$$\theta = \arccos\left(\frac{-R}{\sqrt{-Q^3}}\right),$$

where

$$\frac{-R}{\sqrt{-Q^3}} = -a\sqrt{\frac{27\mu_1\mu_2\lambda}{(\mu_1+\mu_2+\lambda)^3}}.$$

Finally, if we let $\alpha = \sqrt{\frac{\lambda+\mu_1+\mu_2}{3\lambda}}$, then the solutions for (3.16) will be given by

$$\begin{aligned} z_1 &= 2\alpha\cos(\theta/3), \\ z_2 &= 2\alpha\cos(\theta/3+2\pi/3), \\ z_3 &= 2\alpha\cos(\theta/3+4\pi/3). \end{aligned}$$

Of these three roots, only z_1 satisfies (3.13). To see this, note that the cubic function on the left side of (3.16) has one maximum at $-\alpha$ and one minimum at α. Consequently, there is a zero between $-\infty$ and $-\alpha$, a zero between $-\alpha$ and α, and a zero above α. When comparing α to (3.13) and using $z = 1/\sqrt{x}$, one immediately concludes that the only acceptable root is the one above α. Since $z_1 > 2\alpha\cos(\pi/3) = \alpha$, it follows that we need only consider z_1. Hence there are N eigenvalues corresponding to complex values of y, and these eigenvalues can be obtained by calculating θ using $a = \cos(\nu\pi/(N+1))$ for $\nu = 1, 2, \ldots, N$ and calculating

$$x = 1/z_1^2 = \frac{3}{4}\frac{\lambda}{\lambda+\mu_1+\mu_2}\cdot\frac{1}{\cos^2(\theta/3)}.$$

Note that the expressions $b_1 = 3/(\lambda+\mu_1+\mu_2)$ and $b_2 = -\sqrt{b_1^3\mu_1\mu_2\lambda}$ are the same for all ν and need not be recalculated. Once b_1 and b_2 are found, the following three expressions must be evaluated for $\nu = 1, 2, \ldots, N$:

$$a = \cos(\nu\pi/(N+1)),\ \theta = \arccos(ab_2),\ x_\nu = \lambda b_1(4\cos^2(\theta/3))^{-1}.$$

Essentially, one therefore has to evaluate $2N$ cosines and N arc cosines, and this is a trivial task even when N has a high value. In fact, even for N as high as 500, a spreadsheet is sufficient.

Including the eigenvalue $x_0 = \lambda/\mu_1$, we have thus found all $N+1$ eigenvalues known to exist inside the unit circle, and we need not consider any other zeros of $y_1^{N+1} - y_2^{N+1}$.

4. The initial conditions. For each eigenvalue x_ν, $\nu = 0, 1, \ldots, N$, the eigenvector $g^{(\nu)}$ is given by (3.11). Any solution $\pi_n = g^{(\nu)}x_\nu^n$ solves (2.2), and so does any linear combination of these solutions. In other words, all possible solutions have the form

$$\pi_n = \sum_{\nu=0}^{N} c_\nu g^{(\nu)} x_\nu^n. \tag{4.1}$$

Let $\Lambda = \mathrm{diag}(x_\nu)$ and G be two $(N+1)\times(N+1)$ matrices, where G contains the row vector $g^{(\nu)}$ as its νth row. Then (4.1) can be written as

$$\pi_n = c\Lambda^n G. \tag{4.2}$$

We need to determine $c = [c_0, c_1, \ldots, c_N]$ in such a way that (2.1) is satisfied. Clearly,

$$\pi_0 = cG, \qquad \pi_1 = c\Lambda G. \tag{4.3}$$

Hence (2.1) leads to

$$cGQ_{0\,0} + c\Lambda GQ_{-1} = 0. \tag{4.4}$$

Also, the sum of all probabilities must be 1; that is,

$$1 = \sum_{n=0}^{\infty} \pi_n e = \sum_{n=0}^{\infty} c\Lambda^n Ge = c\ \mathrm{diag}(1/(1-x_\nu))Ge. \tag{4.5}$$

Here e is a column vector with all entries equal to 1. We will show that this equation reduces to

$$1 = c_0 \frac{1-(\lambda/\mu_2)^{N+1}}{1-\lambda/\mu_2} \frac{1}{1-\lambda/\mu_1}. \tag{4.6}$$

This equation yields c_0, and the other c_ν can then be found from (4.4).

To prove (4.6), note that the νth element of Ge is equal to $\sum_{i=0}^{N} g_i^{(\nu)}$. For the eigenvector corresponding to x_0, one has

$$\sum_{i=0}^{N} g_i^{(0)} = \sum_{i=0}^{N} (\lambda/\mu_2)^i = \frac{1-(\lambda/\mu_2)^{N+1}}{1-\lambda/\mu_2}.$$

This yields the first element of the column vector Ge. All other elements turn out to be zero. This is the case because, for eigenvectors corresponding to $x \neq x_0$, (3.11) implies

$$\sum_{i=0}^{N} g_i^{(\nu)} \sqrt{d(x)} = y_2^{N+1} - y_1^{N+1}.$$

This is equal to zero because of (3.12). Equation (4.6) can now be derived readily.

The proof also shows that the following equation holds:

$$Ge = \left[\frac{1-(\lambda/\mu_2)^{N+1}}{1-\lambda/\mu_2}, 0, 0, \ldots, 0 \right]^t. \tag{4.7}$$

Also note that

$$c_0 = \frac{(1-\lambda/\mu_1)(1-\lambda/\mu_2)}{1-(\lambda/\mu_2)^{N+1}} \tag{4.8}$$

does not depend on μ_3, which means that c_0 does not change if the servers become movable.

5. Marginal distributions and throughput. The marginal distributions can now be derived. For the distribution of X_1, one has, using (4.2), (4.7), and (4.8),

$$P\{X_1 = n\} = c\Lambda^n Ge = (1-\lambda/\mu_1)(\lambda/\mu_1)^n.$$

Hence making the first server movable has no effect on the distribution of X_1, which is to be expected in view of the fact that the server moves only when idle. For the distribution of X_2, we use (4.1) to obtain

$$P\{X_2 = j\} = \sum_{\nu=0}^{N} c_\nu g_j^{(\nu)}/(1-x_\nu).$$

Since $g_0^{(\nu)} = 1$ for all ν, one has

$$P\{X_2 = 0\} = \sum_{\nu=0}^{N} c_\nu/(1 - x_\nu).$$

We will also need the probability that both servers are idle:

$$P\{X_1 = X_2 = 0\} = (cG)_0 = \sum_{\nu=0}^{N} c_\nu.$$

The main motivation for having movable servers is to increase the throughput T. One finds

$$\begin{aligned} T &= \mu_2 P\{X_1 > 0,\, X_2 > 0\} + \mu_3 P\{X_1 = 0,\, X_2 > 0\} \\ &= \mu_2(1 - P\{X_1 = 0\} - P\{X_2 = 0\} + P\{X_1 = X_2 = 0\}) \\ &\quad + \mu_3(P\{X_1 = 0\} - P\{X_1 = X_2 = 0\}) \\ &= \mu_2 \left(\frac{\lambda}{\mu_1} - \sum_{\nu=0}^{N} c_\nu/(1 - x_\nu) + \sum_{\nu=0}^{N} c_\nu \right) + \mu_3 \left(1 - \frac{\lambda}{\mu_1} - \sum_{\nu=0}^{N} c_\nu \right) \\ &= (\mu_2 - \mu_3) \left(\frac{\lambda}{\mu_1} + \sum_{\nu=0}^{N} c_\nu \right) + \mu_3 - \mu_2 \sum_{\nu=0}^{N} c_\nu/(1 - x_\nu). \end{aligned}$$

In the case where $\mu_2 = \mu_3$, one can apply the formula of the $M/M/1/N$ queue,

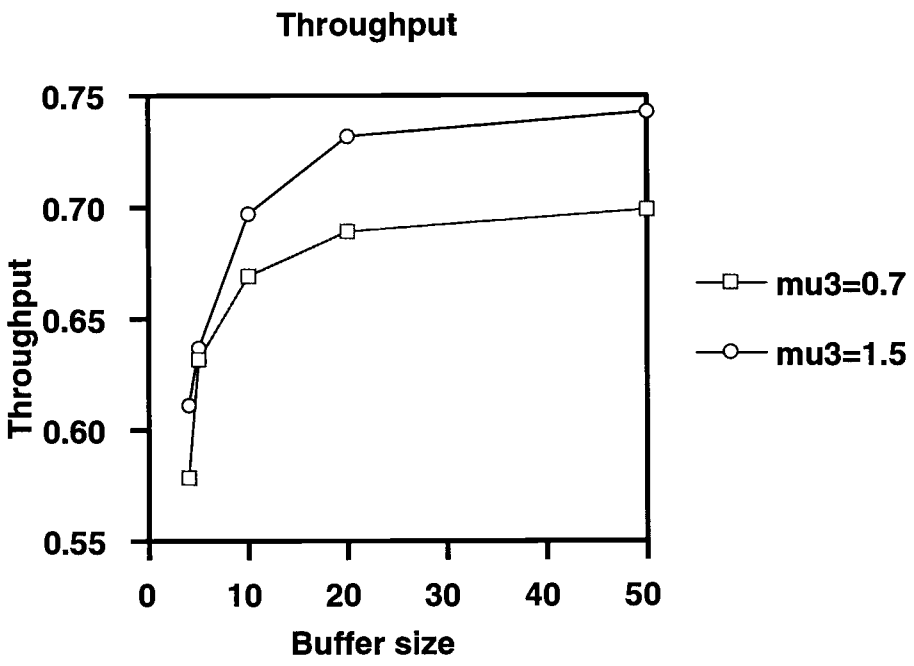


FIG. 5.1.

which is

$$P\{X_2 = \nu\} = (\lambda/\mu_2)^\nu \frac{1 - \lambda/\mu_2}{1 - (\lambda/\mu_2)^{N+1}}.$$

In this case, it follows that the throughput is given by

$$T = \mu_2(1 - P\{X_2 = 0\}) = \mu_2 - \frac{\mu_2 - \lambda}{1 - (\lambda/\mu_2)^{N+1}}.$$

Sometimes upper bounds are useful. If the first server is stationary, the upper bound of the throughput is given by

$$\min\{\lambda, \mu_2\}.$$

By making the server movable, one obtains the following bound:

$$\min\{\lambda, \mu_3(1 - \lambda/\mu_1) + \mu_2\lambda/\mu_1\}.$$

If $\lambda > \mu_2$, the second upper bound can be substantially higher than the first one. In this case, making the server movable increases the throughput considerably, as indicated by Figure 5.1. There $\lambda = 0.75$, $\mu_1 = 1$, and $\mu_2 = 0.7$. The values of μ_3 are 0.7 and 1.5, and the values of N are 4, 5, 10, 20, and 50.

6. Conclusions. As was shown here and elsewhere (see, e.g., [1] or [3]), movable servers can increase throughput, which means that tandem queues with movable servers need not be rebalanced as frequently as tandem queues with stationary servers. Hence, where possible, one should use movable servers. However, as shown in this paper, the analysis of systems with movable servers is inherently more difficult than its stationary counterpart.

The method used here is based on eigenvalues. The interesting point is that product form solutions are really eigenvalue solutions, except that only the dominant eigenvalue enters the final solution. However, other eigenvalues exist, but they affect the solution only if the initial conditions are different from the ones encountered in models having product form solutions. This leads to a completely new perspective on product form solutions and their limitations.

For the problem under investigation, eigenvalue solutions turned out to be extremely efficient. One reason was that the eigenvalues were available in closed form. It is an open question if there are other nontrivial models for which such closed form solutions exist.

There is a close relation between our method and matrix geometric methods. As is well known, in matrix geometric methods, one has to find the rate matrix R, which is the minimal nonnegative solution of

$$0 = Q_1 + RQ_0 + R^2Q_{-1}.$$

The vector π_n is then given as $\pi_0 R^n$. It is not difficult to prove that R must have the same eigenvalues and left eigenvectors as $Q(x)$. However, all eigenvalues of $Q(x)$ are distinct [6]. Therefore, all eigenvalues of R are distinct. This means that G is nonsingular, and $R = G^{-1}\Lambda G$. Hence the eigenvalue method can be used to find R. Of course, there are many other methods to find R (see, for example, [8] and [9]), but, in these methods, each iteration requires several matrix multiplications, which implies that one has $O(N^3)$ operations per iteration. Finding an eigenvalue

and the corresponding eigenvector, on the other hand, requires $O(N)$ operations per iteration, which means that finding all $N+1$ eigensolutions requires $O(N^2)$ operations. Of course, finding R from the eigenvalues still requires $O(N^3)$ operations, and so does finding c_i from (4.4). Hence both algorithms are $O(N^3)$ in the end even though they differ in time complexity by a large constant factor. We also note that finding π_n for a specific value of n takes $O(N^2)$ operations when using eigenvalues, whereas it takes $O(nN^2)$ operations when using matrix analytic methods. Hence using eigenvalues is advantageous in the case considered.

Acknowledgment. We would like to thank the referee for an extremely careful and pertinent review, which greatly improved the paper.

REFERENCES

[1] S. Andradottir, H. Ayhan, and D. Down, *Server assignment policies for maximizing the steady-state throughput of finite state queueing system*, Management Science, 47 (2001), pp. 1421–1439.

[2] D. Bertsimas, *An analytic approach to a general class of G/G/s queueing systems*, Oper. Res., 38 (1990), pp. 139–155.

[3] D. P. Bischak, *Performance of a manufacturing module with moving workers*, IIE Transactions, 28 (1996), pp. 723–733.

[4] J. N. Daigle and D. M. Lucantoni, *Queueing systems having phase-dependent arrival and service rates*, in Proceedings of the First International Conference on the Numerical Solution of Markov Chains, W. J. Stewart, ed., Marcel Dekker, New York, 1991, pp. 161–202.

[5] H. R. Gail, S. L. Hantler, and B. A. Taylor, *Use of characteristic roots for solving infinite state Markov chains*, in Computational Probability, W. K. Grassmann, ed., International Series in Operations Research and Management Science, Kluwer, Boston, 2000, pp. 205–254.

[6] W. K. Grassmann, *Real eigenvalues of certain tridiagonal matrix polynomials, with queueing applications*, Linear Algebra Appl., 342 (2002), pp. 93–106.

[7] W. K. Grassmann and S. Drekic, *An analytical solution for a tandem queue with blocking*, Queueing Systems Theory Appl., 36 (2000), pp. 221–235.

[8] G. Latouche, *Algorithms for Infinite Markov Chains with Repeating Columns*, Vol. 48, Springer-Verlag, Heidelberg, 1992, pp. 231–265.

[9] G. Latouche and V. Ramaswami, *A logarithmic reduction algorithm for quasi-birth-death processes*, J. Appl. Probab., 30 (1993), pp. 650–674.

[10] I. Mitrani and R. Chakka, *Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method*, Performance Evaluation, 23 (1995), pp. 241–260.

[11] P. M. Morse, *Queues, Inventories and Maintenance*, John Wiley, New York, 1958.

[12] M. R. Spiegel and J. Liu, *Mathematical Handbook of Formulas and Tables*, 2nd ed., Schaum's Outline Series, McGraw–Hill, New York, 1999.

[13] N. M. van Dijk and M. Rumsewicz, *Networks of queues with service anticipating routing*, Comm. Statist. Stochastic Models, 7 (1991), pp. 295–310.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 475–483

ALGEBRAIC MULTIPLICITY OF THE EIGENVALUES OF A BIPARTITE TOURNAMENT MATRIX*

YI-ZHENG FAN† AND JIONG-SHENG LI‡

Abstract. Let $\mathcal{B}$ be a bipartite tournament, and let $A(\mathcal{B})$ be its adjacency matrix. $A(\mathcal{B})$ is called a bipartite tournament matrix. In this paper, we mainly discuss the spectral properties of a bipartite tournament matrix, especially for the algebraic multiplicity of the eigenvalue 0 and the number of distinct eigenvalues.

Key words. bipartite tournament matrix, nonnegative matrix, algebraic multiplicity, index of imprimitivity

AMS subject classifications. 05C50, 15A48

PII. S0895479801386547

1. Introduction.

Let $G = (V, E)$ be a simple undirected graph of order n with vertex set $V = \{1, 2, \ldots, n\}$ and edge set $E \subset V \times V$. By assigning to each edge of G one of the two possible orientations, we turn G into a directed graph, denoted by $\vec{G}$. The adjacency matrix of $\vec{G}$ is given by $A(\vec{G}) = [a_{ij}]$ of order n, where $a_{ij} = 1$ if (i, j) is an arc of $\vec{G}$ and $a_{ij} = 0$ otherwise. If G is a complete graph, or generally, a k-partite ($k \geq 1$) complete graph, then $\vec{G}$ is called a tournament, or a k-partite tournament. The adjacency matrix of a k-partite tournament is called a k-partite tournament matrix.

In the literature, there are many results on the spectral properties of a tournament matrix; see, for instance, those for the bounds for the real parts and the imaginary parts of the eigenvalues [2], for the algebraic multiplicity of the eigenvalue 0, and for the number of distinct eigenvalues [8], [3]. Little is known for those of a bipartite tournament matrix, except for an upper bound for its spectral radius given by Li [5]. In this paper, we mainly discuss the spectral properties of a bipartite tournament matrix, especially for the algebraic multiplicity of the eigenvalue 0 and the number of distinct eigenvalues.

Let $\mathcal{B}$ be a bipartite tournament. Then, by a certain label of its vertices, the bipartite tournament matrix $A(\mathcal{B})$ has the form

$$\begin{bmatrix} O_{n_1} & B \\ C & O_{n_2} \end{bmatrix}, \tag{1.1}$$

where O_{n_1}, O_{n_2} $(n_1 \geq n_2)$ are, respectively, the square zero matrices of order n_1, n_2. Denote by $J_{s,t}$ the $s \times t$ all-ones matrix and by A^t the transpose of A. It is easily seen that $B + C^t = J_{n_1,n_2}$ and $A(\mathcal{B})$ is irreducible (i.e., $\mathcal{B}$ is strongly connected) only if $n_1 \geq n_2 \geq 2$. Since a reducible bipartite tournament matrix is permutation similar to a triangular block matrix with irreducible bipartite tournament matrices on the

*Received by the editors March 20, 2001; accepted for publication (in revised form) by R. Brualdi April 19, 2002; published electronically November 6, 2002. This work was supported by the National Natural Science Foundation of China (grant 19971086).
http://www.siam.org/journals/simax/24-2/38654.html
†Department of Mathematics, Anhui University, Hefei, Anhui 230039, The People's Republic of China (fanyz@mars.ahu.edu.cn).
‡Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, The People's Republic of China (lijs@ustc.edu.cn).

main diagonal blocks except the zero matrices of order one, we deal only with the irreducible bipartite tournament matrices in this paper.

Denote by T_{n_1,n_2} the set of irreducible matrices of the form of (1.1), and denote by E_{n_1,n_2} the set of integers t for which there is a matrix $A \in T_{n_1,n_2}$ with algebraic multiplicity of the eigenvalue 0 equal to t. We adopt the convention that the expression of every $A \in T_{n_1,n_2}$ is exactly the matrix in (1.1).

By the famous Perron–Frobenius theorem, in section 2 of this paper, we show that the index of imprimitivity of $A \in T_{n_1,n_2}$ is 2 or 4 so that the nonzero eigenvalues of A can be grouped in pairs or quadruples. For the above matrix A with index of imprimitivity equal to 4, we completely characterize the structures of its spectrum and the bipartite tournament corresponding to A. We establish the set E_{n_1,n_2} in section 3. In section 4, we discuss the matrix $A \in T_{n_1,n_2}$ with the least or largest number of distinct eigenvalues. For the former, an equivalent condition is obtained, and, for the latter, the singular and nonsingular matrices $A \in T_{n_1,n_2}$ are, respectively, given.

2. Preliminaries. Denote by $A \geq 0$ the (entrywise) nonnegative matrix A, and denote by $\rho(A)$ the spectral radius of A, $S(A)$ the spectrum of A. We first introduce the following theorem about the nonnegative matrix which is a part of the famous Perron–Frobenius theorem [1].

THEOREM 2.1. *Let $A \geq 0$ be irreducible of order n. Then the following hold.*

(1) *$\rho(A)$ is a simple eigenvalue, and any eigenvalue of A of the same modulus is also simple.*
(2) *If A has h eigenvalues $\lambda_0 = re^{i\theta_0}, \lambda_1 = re^{i\theta_1}, \ldots, \lambda_{h-1} = re^{i\theta_{h-1}}$ of modulus $\rho(A) = r$, with $0 = \theta_0 < \theta_1 < \cdots < \theta_{h-1} < 2\pi$, then these numbers are the distinct roots of $\lambda^h - r^h = 0$.*
(3) *More generally, the spectrum $S(A) = \{\lambda_0, \lambda_1, \ldots, \lambda_{n-1}\}$ goes over into itself under a rotation of the complex plane by $2\pi/h$.*
(4) *If $h > 1$, there exists a permutation matrix P such that*

$$PAP^t = \begin{bmatrix} O & A_{12} & O & \cdots & O \\ O & O & A_{23} & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & A_{h-1,h} \\ A_{h1} & O & O & \cdots & O \end{bmatrix}, \tag{2.1}$$

where the zero blocks along the main diagonal are square.

Let $A \geq 0$ be irreducible of order n; the number h of eigenvalues of A of modulus $\rho(A)$ is called the index of imprimitivity (or index of cyclicity) of A, denoted by $h(A)$. $h(A)$ can be obtained from the associated directed graph $D(A)$ of A by the following theorem. Note that $D(A)$ has n vertices and an arc (i, j) if and only if $a_{ij} > 0$. A circuit of length l of $D(A)$ is a sequence of arcs $(i_1, i_2), (i_2, i_3), \ldots, (i_{l-1}, i_l), (i_l, i_1)$ of $D(A)$.

THEOREM 2.2 (see [1]). *Let $A \geq 0$ be irreducible of order n. Let S_i be the set of all of the lengths m_i of circuits in $D(A)$ through the vertex i, and $h_i = \text{g.c.d.}_{m_i \in S_i}\{m_i\}$. Then $h_1 = h_2 = \cdots = h_n = h(A)$.*

LEMMA 2.3. *Let $A \in T_{n_1,n_2}$. Then $h(A) = 2$ or $h(A) = 4$.*

Proof. Let $\mathcal{B}$ be the bipartite tournament corresponding to A. Then $\mathcal{B}$ is strongly connected as A is irreducible, and the length of each circuit is an even integer greater than 2. By Theorem 2.2, $h(A)$ is also even. If $h(A) > 4$, then, by Theorem 2.1 (4),

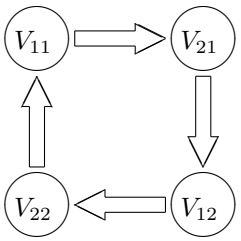


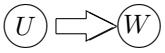
FIG. 1.

the vertex set of $\mathcal{B}$ can be partitioned into $h(A)$ disjoint subsets $V_1, \ldots, V_{h(A)}$ such that each arc of $\mathcal{B}$ is from V_i to V_{i+1} for some $1 \leq i \leq h(A)-1$, or $V_{h(A)}$ to V_1. Suppose $v \in V_1$. Then the set $U = V_3 \cup \cdots \cup V_{h(A)-1}$ has no vertices incident to v. So each pair of vertices of U are not incident by the definition of bipartite tournaments, which is impossible. The result follows. □

COROLLARY 2.4. *Let $A \in T_{n_1,n_2}$. Then the numbers of nonzero eigenvalues and distinct nonzero eigenvalues are both even.*

Proof. Let $\lambda \neq 0$ be an eigenvalue of A. Then, by Theorem 2.1 (3), $\lambda e^{i\frac{2k\pi}{h(A)}}$, $k = 0, \ldots, h(A)-1$, are $h(A)$ eigenvalues of A, which are also distinct. By Lemma 2.3, $h(A)$ is even, so the result follows. □

Before characterizing the matrix $A \in T_{n_1,n_2}$ with $h(A) = 4$, we introduce the graph in Figure 1, where



means that there exists an arc from each vertex of U to every vertex of W.

Denote by $\lambda^{(m)} \in S(A)$ the number λ as an eigenvalue of A with algebraic multiplicity equal to m.

THEOREM 2.5. *Let $A \in T_{n_1,n_2}$ with the corresponding bipartite tournament $\mathcal{B}$. Then the following are equivalent.*

(1) $h(A) = 4$.

(2) *$\mathcal{B}$ has the structure of Figure* 1.

(3) *The spectrum* $S(A) = \{\, \rho(A), -\rho(A), i\rho(A), -i\rho(A), 0^{(n_1+n_2-4)} \,\}$.

(4) *The algebraic multiplicity of eigenvalue* 0 *of A is* $n_1 + n_2 - 4$.

Proof. If (1) holds, by Theorem 2.1 (4), there exists a permutation matrix P such that

$$PAP^t = \begin{bmatrix} 0_{n_{11}} & A_{12} & 0 & 0 \\ 0 & 0_{n_{21}} & A_{23} & 0 \\ 0 & 0 & 0_{n_{12}} & A_{34} \\ A_{41} & 0 & 0 & 0_{n_{22}} \end{bmatrix}. \tag{2.2}$$

By the definition of bipartite tournaments, we know that $A_{i,i+1}, i = 1, 2, 3$, and A_{41} are all all-ones matrices. So (2) holds. Also, if (1) holds, then $\rho(A), -\rho(A), i\rho(A), -i\rho(A)$ are the eigenvalues of A. So (3) and (4) both hold as the rank of A is exactly 4. By Theorem 2.2, (2) implies (1); and by Theorem 2.1 (2), (3) also implies (1). Finally, we prove that (4) implies (1). If $h(A) \neq 4$, then $h(A) = 2$ by Lemma 2.3, and $S(A) = \{\, \rho(A), -\rho(A), re^{i\theta}, re^{i\theta+\pi}, 0^{(n_1+n_2-4)} \,\} (0 < r < \rho(A))$. Since the trace of A^2 is zero, $2\rho(A)^2 + 2r^2e^{i(2\theta)} = 0$, which is a contradiction. The result follows. □

Remark. (i) Let $A \in T_{n_1,n_2}$ with $h(A) = 2$. Then $re^{i\theta}, re^{i(\theta+\pi)}, re^{-i\theta}, re^{-i(\theta+\pi)}$ $(0 < r < \rho(A),\ \theta \neq l\pi,\ \theta \neq l\pi + \frac{\pi}{2},\ l \in \mathbb{Z})$, may be the eigenvalues of A. We show this by the following example.

Let $A \in T_{4,4}$ with

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

Then, by a calculation on Matlab, BC has eigenvalues approximating to 3.3423, -0.2494, $-1.5465+0.0903i$, $-1.5465-0.0903i$. So $S(A)$ is of the form $\{\ \rho, -\rho, ir_1, -ir_1, r_2e^{i\theta}, r_2e^{i(\theta+\pi)}, r_2e^{-i\theta}, r_2e^{-i(\theta+\pi)}\ \}$, where ρ, r_1, r_2 are positive real numbers.

(ii) By Corollary 2.4 and the proof of Theorem 2.5, we know that the largest algebraic multiplicity of the eigenvalue 0 of all matrices $A \in T_{n_1,n_2}$ is exactly $n_1 + n_2 - 4$. So we can add the following condition to Theorem 2.5.

(5) The algebraic multiplicity of eigenvalue 0 of A is the largest for all matrices in T_{n_1,n_2}.

3. The algebraic multiplicity of the eigenvalue 0. In the last part of section 2, we show that $n_1 + n_2 - 4$ is the largest number in the set E_{n_1,n_2} and completely characterize the bipartite tournament matrices $A \in T_{n_1,n_2}$ with algebraic multiplicity equal to that number. In this section, we shall establish the set E_{n_1,n_2}. First, we introduce the following lemma, where $O_{s,t}$ denotes the $s \times t$ zero matrix.

LEMMA 3.1. *Let $A \in T_{n_1,n_2}$ with the form of* (3.1)*, where $l_1 + l_2 + \cdots + l_k = n_1, m_1 + m_2 + \cdots + m_k = n_2, 2 \le k \le n_2$. Then A has exactly $2k$ nonzero eigenvalues and $n_1 + n_2 - 2k$ zero eigenvalues, and, for each of these eigenvalues, the algebraic multiplicity is same as the geometric multiplicty.*

$$A = \begin{bmatrix} O_{l_1} & & & & J_{l_1,m_1} & & & \\ & O_{l_2} & & & & J_{l_2,m_2} & & \\ & & \ddots & & & & \ddots & \\ & & & O_{l_k} & & & & J_{l_k,m_k} \\ O_{m_1,l_1} & J_{m_1,l_2} & \cdots & J_{m_1,l_k} & O_{m_1} & & & \\ J_{m_2,l_1} & O_{m_2,l_2} & \ddots & \vdots & & O_{m_2} & & \\ \vdots & \ddots & \ddots & J_{m_{k-1},l_k} & & & \ddots & \\ J_{m_k,l_1} & \cdots & J_{m_k,l_{k-1}} & O_{m_k,l_k} & & & & O_{m_k} \end{bmatrix}. \tag{3.1}$$

Proof. Let $\lambda \neq 0$ be an eigenvalue of A with the corresponding eigenvector $x = ((y^1)^t, (y^2)^t, \ldots, (y^k)^t, (z^1)^t, (z^2)^t, \ldots, (z^k)^t)^t$, where $y^i \in \mathbb{C}^{l_i}$ and $z^i \in \mathbb{C}^{m_i}$, $i = 1, 2, \ldots, k$. By $Ax = \lambda x$, one may get that, for each $i = 1, 2, \ldots, k$, $y^i = a_i J_{l_i,1}$ and $z^i = b_i J_{m_i,1}$ for some a_i and b_i. Then

$$\begin{aligned} J_{l_i,m_i} b_i J_{m_i,1} &= \lambda a_i J_{l_i,1}, \quad i = 1, 2, \ldots, k, \\ \sum_{j \neq i} J_{m_i,l_j} a_j J_{l_j,1} &= \lambda b_i J_{m_i,1}, \quad i = 1, 2, \ldots, k. \end{aligned}$$

So

$$\begin{aligned} m_i b_i &= \lambda a_i, \quad i = 1, 2, \ldots, k, \\ \sum_{j \neq i} l_j a_j &= \lambda b_i, \quad i = 1, 2, \ldots, k. \end{aligned}$$

The above equalities imply that λ is an eigenvalue of the matrix of order $2k$

$$D = \begin{bmatrix} 0 & & & & m_1 & & & \\ & 0 & & & & m_2 & & \\ & & \ddots & & & & \ddots & \\ & & & 0 & & & & m_k \\ 0 & l_2 & \cdots & l_k & 0 & & & \\ l_1 & 0 & \cdots & l_k & & 0 & & \\ \vdots & \vdots & \ddots & \vdots & & & \ddots & \\ l_1 & l_2 & \cdots & 0 & & & & 0 \end{bmatrix}, \tag{3.2}$$

and $u = (a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_k)^t$ is the corresponding eigenvector.

On the other hand, let $Du = \lambda u$ with $u = (a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_k)^t \neq 0$, and let $x = ((y^1)^t, (y^2)^t, \ldots, (y^k)^t, (z^1)^t, (z^2)^t, \ldots, (z^k)^t)^t$ with $y^i = a_i J_{l_i,1}$ and $z^i = b_i J_{m_i,1}$. Since D is nonsingular, $\lambda \neq 0$. One can get $Ax = \lambda x$, and hence λ is also a nonzero eigenvalue of A.

In addition, $D^2 = E(J_k - I_k)F \oplus (J_k - I_k)FE$, where $E = \operatorname{diag}\{m_1, m_2, \ldots, m_k\}$ and $F = \operatorname{diag}\{l_1, l_2, \ldots, l_k\}$. Since D^2 is similar to the symmetric matrix $(EF)^{\frac{1}{2}}(J_k - I_k)(EF)^{\frac{1}{2}} \oplus (EF)^{\frac{1}{2}}(J_k - I_k)(EF)^{\frac{1}{2}}$, D^2, and hence D, is similar to a diagonal matrix as D is nonsingular. D has $2k$ linear independent eigenvectors (corresponding to nonzero eigenvalues), and so does A. Since the rank of A is $2k$, A has exactly $2k$ nonzero eigenvalues and $n_1 + n_2 - 2k$ zero eigenvalues, and the algebraic and geometric multiplicities of the eigenvalue 0 are equal. By the above discussion, A is similar to a diagonal matrix, and the result follows. ∎

THEOREM 3.2.

$$E_{n_1,n_2} = \{n_1 + n_2 - 2k : k = 2, 3, \ldots, n_2\}.$$

Further, for each $t \in E_{n_1,n_2}$, there exists some matrix in T_{n_1,n_2} whose eigenvalue 0 has the same algebraic and geometric multiplicity equal to t.

Proof. Let $A \in T_{n_1,n_2}$. By Corollary 2.4, the algebraic multiplicity of the eigenvalue 0 of A is $n_1 + n_2 - 2k$ for some integer k. Since A is irreducible and nonnegative with even $h(A)$, $\rho(A), -\rho(A)$ are its two eigenvalues. A has at least two conjugate imaginary eigenvalues as the trace of A^2 is 0. So $k \geq 2$. Since $A^2 = BC \oplus CB$, BC, then A^2 has the eigenvalue 0 with algebraic multiplicity at least $n_1 - n_2$. So $k \leq n_2$. The result follows by Lemma 3.1. ∎

4. The number of distinct eigenvalues. Let $A \in T_{n_1,n_2}$. By Corollary 2.4, the number of the distinct eigenvalues of H is $2k + \delta$ for some k, where $\delta = 1$ if A is singular and $\delta = 0$ otherwise. Also, we can see that the number of the distinct nonzero eigenvalues of A is double that of CB (or BC). Since CB is nonnegative with trace equal to 0, CB has at least 2 distinct nonzero eigenvalues. CB has at most n_2 distinct nonzero eigenvalues. So $2 \leq k \leq n_2$. In this section, we will discuss, respectively, the bipartite tournament matrices with the least and largest numbers of distinct eigenvalues.

THEOREM 4.1. *Let $A \in T_{n_1,n_2}$. Then A has 4 distinct eigenvalues if and only if*

$$n_1 = n_2,\ CB = (r_1^2, r_2^2, r_3^2, \ldots, r_{n_2}^2)^t \left(1, \frac{r_1^2}{r_2^2}, \frac{r_1^2}{r_3^2}, \ldots, \frac{r_1^2}{r_{n_2}^2}\right) - r_1^2 I_{n_2}, \tag{4.1}$$

where $r_i^2, \frac{r_1^4}{r_i^2}, \frac{r_1^2 r_i^2}{r_j^2}$ $(i \neq j)$ are positive integers for $i, j = 2, 3, \ldots, n_2$.

Proof. Let $A \in T_{n_1,n_2}$ with 4 distinct eigenvalues. If $h(A) = 4$, then $n_1 + n_2 = 4$ by Theorem 2.5, and the conditions in (4.1) are easily verified. Suppose $h(A) = 2$ (or, equivalently, $n_1 + n_2 > 4$) in the following. Then CB is nonsingular and has exactly two distinct eigenvalues. So the spectrum $S(CB) = \{\, \rho(A)^2, (-r_1^2)^{(n_2-1)} \,\}$, and $r_1^2 = \frac{\rho(A)^2}{n_2-1}$, as its trace is 0. Then $CB + r_1^2 I_{n_2}$ is of rank 1, and $CB + r_1^2 I_{n_1}$ can be written as $(r_1^2, r_2^2, r_3^2, \ldots, r_{n_2}^2)^t (1, \frac{r_1^2}{r_2^2}, \frac{r_1^2}{r_3^2}, \ldots, \frac{r_1^2}{r_{n_2}^2})$. Also, CB is an integer matrix, so the necessity holds. To prove the converse, it is easily seen that the spectrum $S(CB)$ (and $S(BC)$) is $\{\, (n_2-1)r_1^2, (-r_1^2)^{(n_2-1)} \,\}$. So $S(A) = \{\, r_1\sqrt{n_2-1},\ -r_1\sqrt{n_2-1},\ ir_1^{(n_2-1)}, -ir_1^{(n_2-1)} \,\}$. The result follows. □

By Theorem 2.5 and the proof of Theorem 4.1, we get the following corollary.

COROLLARY 4.2. *Let $A \in T_{n_1,n_2}(n_1 + n_2 \geq 5)$ with the corresponding bipartite tournament $\mathcal{B}$. Then A has 5 distinct eigenvalues if one of the following holds:*

(1) $h(A) = 4$.

(2) *$\mathcal{B}$ has the structure of Figure* 1.

(3)

$$n_1 > n_2,\ CB = (r_1^2, r_2^2, r_3^2, \ldots, r_{n_2}^2)^t \left(1, \frac{r_1^2}{r_2^2}, \frac{r_1^2}{r_3^2}, \ldots, \frac{r_1^2}{r_{n_2}^2}\right) - r_1^2 I_{n_2},$$

where $r_i^2, \frac{r_1^4}{r_i^2}, \frac{r_1^2 r_i^2}{r_j^2}$ $(i \neq j)$ are positive integers for $i, j = 2, 3, \ldots, n_2$.

Remark. If the matrix CB in Theorem 4.1 is symmetric (i.e., the row sums of C are constant), then $r_i^2 = r_j^2$ for any $i \neq j$. So $CB = r_1^2(J_{n_2} - I_{n_2})$. Assume the column sums of C are also constant. We get the following two equations:

$$CC^t = (r - \lambda)I_{n_2} + \lambda J_{n_2},\ J_{n_2,1}C = kJ_{n_2,1}, \tag{4.2}$$

where $r = r_1^2 + \lambda$ and k are, respectively, the row sum and column sum of C. The equations (4.2) imply that the solution C is an incidence matrix of some balanced incomplete block design [4, p. 127].

Now we shall construct, respectively, the singular and nonsingular bipartite tournament matrices $A \in T_{n_1,n_2}$ with the largest number of distinct eigenvalues. One can easily verify the following lemma.

LEMMA 4.3. *Let $A \in T_{n_1,n_2}$. Then A has $n_1 + n_2$ distinct eigenvalues if and only if one of the following holds:*

(1) *$n_1 = n_2$, and CB has n_2 distinct nonzero eigenvalues.*

(2) *$n_1 = n_2 + 1$, and CB has n_2 distinct nonzero eigenvalues.*

LEMMA 4.4. *Let*

$$X = \begin{bmatrix} \operatorname{diag}\{\lambda_1, \ldots, \lambda_{n-1}\} & \alpha \\ \beta^t & a \end{bmatrix}$$

be a real matrix of order n, where $\alpha = (b_1, b_2, \ldots, b_{n-1})^t$ and $\beta = (c_1, c_2, \ldots, c_{n-1})^t$ such that $b_j c_j > 0$ for $j = 1, 2, \ldots, n-1$, and $\lambda_1 < \lambda_2 < \cdots < \lambda_{n-1}$. Then X has n distinct real eigenvalues.

Proof. By observing the characteristic polynomial of X, we get the following:

$$(a - \lambda) \prod_{i=1}^{n-1} (\lambda_i - \lambda) - \sum_{j=1}^{n-1} b_j c_j \prod_{i \neq j} (\lambda_i - \lambda) = 0.$$

Since $\lambda_i\,(\,i = 1, 2, \ldots, n-1\,)$ is not the root of above equation, using $\prod_{i=1}^{n-1}(\lambda_i - \lambda)$ to divide the equation, we get

$$0 = (a - \lambda) - \sum_{j=1}^{n-1} b_j c_j (\lambda_j - \lambda)^{-1} \equiv a - f(\lambda).$$

The derivation of $f(\lambda)$ is $f'(\lambda) = 1 + \sum_{j=1}^{n-1} \frac{b_j c_j}{(\lambda_j - \lambda)^2} > 0$, so $f(\lambda)$ is a strictly increasing function on intervals $(-\infty, \lambda_1), (\lambda_1, \lambda_2), \ldots, (\lambda_{n-2}, \lambda_{n-1}), (\lambda_{n-1}, +\infty)$, and $\lambda = \lambda_1, \lambda = \lambda_2, \ldots, \lambda = \lambda_{n-1}$ are its asymptotes. Therefore, $f(\lambda) = a$ has n distinct eigenvalues, and the result follows. □

Let $A \in T_{n_1,n_2}$ $(n_1 = n_2 = m \geq 3)$ with $C = [c_{ij}]$, whose nonzero entries are $c_{ii} = c_{i,i+1} = 1$ for $i = 1, \ldots, m-1$ and $c_{mm} = 1$. By a straightforward calculation, we get

$$CB = \begin{bmatrix} E_{m-1} & \alpha \\ \beta^t & 0 \end{bmatrix} = \left[\begin{array}{ccccc|c} 0 & 1 & 2 & \cdots & 2 & 2 \\ 1 & 0 & 1 & \ddots & \vdots & 2 \\ 2 & 1 & \ddots & \ddots & 2 & \vdots \\ \vdots & \ddots & \ddots & 0 & 1 & 2 \\ 2 & \cdots & 2 & 1 & 0 & 1 \\ \hline 1 & 1 & \cdots & 1 & 1 & 0 & 0 \end{array}\right], \tag{4.3}$$

where E_{m-1} is the $(m-1) \times (m-1)$ upper-left block of the last matrix in (4.3).

For the matrix CB in (4.3), one can get $\det(CB) = \det(J - C^t) = (-1)^{m-1}(\lceil \frac{m}{2} \rceil - 1)$, where $\lceil \frac{m}{2} \rceil$ denotes the least integer which is greater than or equal to $\frac{m}{2}$. By the Cauchy–Binet theorem [6, p. 14],

$$\begin{aligned} \det E_{m-1} &= \det C[\langle m-1\rangle | \langle m\rangle](J - C^t)[\langle m\rangle | \langle m-1\rangle] \\ &= \sum_{\substack{\gamma \subset \langle m\rangle \\ |\gamma| = m-1}} \det C[\langle m-1\rangle | \gamma] \det(J - C[\langle m-1\rangle | \gamma]), \end{aligned}$$

where $A[\alpha|\beta]$ denotes the submatrix of A with rows indexed by α and columns indexed by β, $\langle n\rangle$ denotes the set $\{1, 2, \ldots, n\}$, and $|\gamma|$ denotes the cardinality of the set γ. For any set $\gamma \subset \langle m\rangle$ with $|\gamma| = m-1$, $\det A[\langle m-1\rangle | \gamma] = 1$, and the sign of $\det(J - C[\langle m-1\rangle | \gamma])$ is $(-1)^m$, so the sign of $\det E_{m-1}$ is also $(-1)^m$. Also, we find that, if we permute simultaneously the rows and columns of $1, 2, \ldots, m$ of E_m to those of $m, m-1, \ldots, 2, 1$, then the resulting matrix is the same. Then, if $(x_1, x_2, \ldots, x_m)^t$ is an eigenvector of E_m, so is $(x_m, x_{m-1}, \ldots, x_2, x_1)^t$.

LEMMA 4.5. *Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ be eigenvalues of $E_m(\, m \geq 2)$. Then*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_{m-1} < 0 < \lambda_m.$$

Proof. For $m = 2$, one can easily verify the above lemma. Assume the result holds for the case when $m = k-1(\, k \geq 3)$. Let $u^1, u^2, \ldots, u^{k-1}$ be the orthonormal eigenvectors of E_{k-1} corresponding to its eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_{k-1}$, respectively. Then

$$J_{1,k-1} E_{k-1} u^i = \lambda_i J_{1,k-1} u^i,$$

and hence

$$(\lambda_i - (2k-6)) J_{1,k-1} u^i = (u_1^i + u_{k-1}^i), \tag{4.4}$$

where u_l^i denotes the lth entry of u^i.

By the discussion prior to this lemma, if $u_1^i = 0$, then $u_{k-1}^i = 0$, as λ_i is simple. Since E_{k-1} is irreducible and has distinct row sums, $2k-6 < \lambda_{k-1} < 2k-5$ [7, chapter 2]. Then $\lambda_i - (2k-6) \neq 0$ for each $i = 1, 2, \ldots, k-1$ by assumption. Hence $J_{1,k-1}u^i = 0$, and, therefore, $u^i = 0$ by the characteristic equation $E_{k-1}u^i = \lambda_i u^i$. So $u_1^i \neq 0$, and $u_1^i = u_{k-1}^i$ or $u_1^i = -u_{k-1}^i$.

Let $U_1 = (u^1, u^2, \ldots, u^{k-1})$, and let $U = U_1 \oplus 1$. Then

$$U^t E_k U = U^t \begin{bmatrix} E_{k-1} & \alpha \\ \alpha^t & 0 \end{bmatrix} U = \begin{bmatrix} \operatorname{diag}\{\lambda_1, \ldots, \lambda_{k-1}\} & U_1^t\alpha \\ \alpha^t U_1 & 0 \end{bmatrix}.$$

If $u_1^i = -u_{k-1}^i$, then $(u^i)^t J_{k-1,1} = 0$ by (4.4); otherwise, $(u^i)^t J_{k-1,1} = \frac{2\mu_i^1}{\lambda_i - (2k-6)}$. Since $U_1^t\alpha = ((u^1)^t, (u^2)^t, \ldots, (u^{k-1})^t)(2J_{k-1,1} - (0, \ldots, 0, 1)^t)$, then $(u^i)^t\alpha = u_1^i \neq 0$ or $(u^i)^t\alpha = 2(u^i)^t J_{k-1,1} - u_1^i = (\frac{4}{\lambda_i - (2k-6)} - 1)u_1^i \neq 0$. So each entry of $\alpha^t u^i$ is nonzero. By Lemma 4.4, the eigenvalues of E_{k-1} strictly interlace those of E_k. Since the sign of $\det E_k$ is $(-1)^{k-1}$, E_k has exactly one positive eigenvalue. The result follows. □

THEOREM 4.6. *Let $A \in T_{n_1,n_2}$ $(n_1 = n_2+1)$ with $C = [c_{ij}]$, whose nonzero entries are $c_{ii} = c_{i,i+1} = 1$ for $i = 1, 2, \ldots, n_2$. Then A has $n_1 + n_2$ distinct eigenvalues.*

Proof. Since $CB = E_{n_2}$, the result follows immediately from Lemmas 4.5 and 4.3. □

Note that, if $A \in T_{2,2}$, then A has 4 distinct eigenvalues.

THEOREM 4.7. *Let $A \in T_{n_1,n_2}$ $(n_1 = n_2 \geq 3)$ with $C = [c_{ij}]$, whose nonzero entries are $c_{ii} = c_{i,i+1} = 1$ for $i = 1, 2, \ldots, n_2 - 1$ and $c_{n_2,n_2} = 1$. Then A has $n_1 + n_2$ distinct eigenvalues.*

Proof. By (4.3),

$$CB = \begin{bmatrix} E_{n_2-1} & \alpha \\ \beta^t & 0 \end{bmatrix}.$$

There exists an orthogonal matrix $U_1 = (u^1, u^2, \ldots, u^{n_2-1})$ such that $U_1^t E_{n_2-1} U_1 = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_{n_2-1}\}$. Let $U = U_1 \oplus 1$. Then

$$U^t(CB)U = \begin{bmatrix} \operatorname{diag}\{\lambda_1, \ldots, \lambda_{n_2-1}\} & U_1^t\alpha \\ \beta^t U_1 & 0 \end{bmatrix}.$$

By Lemma 4.5, $\lambda_1 < \lambda_2 < \cdots < \lambda_{n_2-2} < 0 < \lambda_{n_2-1}$; also, by the proof of Lemma 4.5, $2n_2 - 6 < \lambda_{n_2-1} < 2n_2 - 5$, and, for each $i = 1, \ldots, n_2 - 1$, $u_1^i \neq 0$, $u_1^i = u_{n_2-1}^i$, or $u_1^i = -u_{n_2-1}^i$. Since $U_1^t\alpha = ((u^1)^t, (u^2)^t, \ldots, (u^{n_2-1})^t)\alpha$ and $\beta^t U_1 = (\alpha - J_{n_2-1,1})^t(u^1, u^2, \ldots, u^{n_2-1})$, then $(u^i)^t\alpha = \beta^t u^i = u_1^i$ (for the case of $u_1^i = -u_{n_2-1}^i$) or $(u^i)^t\alpha = (\frac{4}{\lambda_i - (2n_2-6))} - 1)u_1^i$, $\beta^t u^i = \alpha^t u^i - J_{n_2-1,1}^t u^i = (\frac{2}{\lambda_i - (2n_2-6))} - 1)u_1^i$ (for the case of $u_1^i = u_{n_2-1}^i$). So, for each $i = 1, 2, \ldots, n_2 - 1$, $((u^i)^t\alpha)(\beta^t u^i) > 0$. The result follows by Lemmas 4.4 and 4.3. □

Remark. By Lemma 4.5, the matrix $CB = E_{n_2}$ in Theorem 4.6 has exactly one positive eigenvalue and $n_2 - 1$ negative eigenvalues. The above property also holds for the matrix CB in Theorem 4.7 since the eigenvalues of CB strictly interlace those of E_{n_2-1} and the sign of $\det CB$ is $(-1)^{n_2-1}$. Therefore, the nonzero distinct eigenvalues of the matrix A in Theorems 4.6 and 4.7 are both distributed as follows: one positive, one negative, and other conjugate pairs of purely imaginary numbers. The following example will show that there exists a bipartite tournament matrix with the algebraic multiplicity of some eigenvalue greater than one.

Let $A \in T_{5,5}$ with

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

Then -1 is an eigenvalue of BC with algebraic multiplicity equal to 2. Therefore, i or $-i$ is an eigenvalue of A with algebraic multiplicity equal to 2.

REFERENCES

[1] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Chapter 2, Academic Press, New York, 1979.

[2] A. Brauer and I. Gentry, *On the characteristic roots of tournament matrices*, Bull. Amer. Math. Soc., 74 (1968), pp. 1133–1135.

[3] D. de Caen, D. A. Gregory, S. J. Kirkland, N. J. Pullman, and J. S. Maybee, *Algebraic multiplicity of the eigenvalues of a tournament matrix*, Linear Algebra Appl., 169 (1992), pp. 179–193.

[4] Jr. M. Hall, *Combinatorial Theory*, 2nd ed., Wiley, New York, 1986.

[5] J.-S. Li, *Eigenvalues of oriented-graph matrices*, Linear Algebra Appl., 221 (1995), pp. 103–110.

[6] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[7] H. Minc, *Nonnegative Matrices*, Chapter 2, Wiley, New York, 1988.

[8] N. Zagaglia Salvi, *Alcune proprietà delle matrici tornco regolari*, Rend. Sem. Mat. Brescia, 7 (1984), pp. 635–643.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 484–489

PIECEWISE CONTINUOUS TOEPLITZ MATRICES AND OPERATORS: SLOW APPROACH TO INFINITY*

ALBRECHT BÖTTCHER†, MARK EMBREE‡, AND LLOYD N. TREFETHEN‡

Abstract. The pseudospectra of banded finite dimensional Toeplitz matrices rapidly converge to the pseudospectra of the corresponding infinite dimensional operator. This exponential convergence makes a compelling case for analyzing pseudospectra of such Toeplitz matrices—not just eigenvalues. What if the matrix is dense and its symbol has a jump discontinuity? The pseudospectra of the finite matrices still converge, but it is shown here that the rate is no longer exponential in the matrix dimension—only algebraic.

Key words. Toeplitz matrix, piecewise continuous symbol, pseudospectra

AMS subject classifications. 47B35, 15A60

PII. S0895479800376971

Let T be a Toeplitz operator (singly infinite matrix) on $\ell^2(\mathbf{N})$ with symbol $a \in L^\infty(\mathbf{T})$, where $\mathbf{T}$ is the unit circle. Thus $T = (a_{j-k})_{j,k=0}^\infty$, where $\{a_n\}_{n=-\infty}^\infty$ is the sequence of Fourier coefficients for a, a complex-valued function on $\mathbf{T}$. If a is continuous, then the spectrum $\operatorname{sp} T$ is the curve $a(\mathbf{T})$ together with all of the points this curve encloses with nonzero winding number [6]. This result generalizes to piecewise continuous a: If $a^\#(\mathbf{T})$ is the curve consisting of the components of $a(\mathbf{T})$ connected by straight segments at points of discontinuity, then $\operatorname{sp} T$ is $a^\#(\mathbf{T})$ together with all of the points it encloses with nonzero winding number; see [5, section 1.8].

A long-recognized anomaly is that the spectra of Toeplitz matrices T_N of finite dimension N look very different, typically consisting of points distributed along curves rather than across regions, even as $N \to \infty$ [1, 5, 11, 12, 17]. Some kind of resolution of this anomaly was obtained with the discovery that, although the spectra of the matrix and the operator do not agree, the ε-pseudospectra may agree very closely [9, 10]. (The ε-pseudospectrum $\operatorname{sp}_\varepsilon A$ of a matrix or operator A is the set of points $z \in \mathbf{C}$ satisfying $\|(zI - A)^{-1}\| \geq \varepsilon^{-1}$, where we write $\|(zI - A)^{-1}\| = \infty$ when $z \in \operatorname{sp} A$; see, e.g., [13, 14].) In particular, if T_N is banded, then for each point z enclosed by $a(\mathbf{T})$ with nonzero winding number, $\|(zI - T_N)^{-1}\|$ grows exponentially as $N \to \infty$ [3, 10]; the condition number $\|V_N\|\|V_N^{-1}\|$ of any matrix V_N of eigenvectors of T_N is likewise exponentially large. As illustrated by numerical examples in [10], the result is that for small ε, the ε-pseudospectra of T_N typically look much like the spectrum of T for values of N on the order of hundreds.

A more general convergence result for $\operatorname{sp}_\varepsilon T_N$ has been proved in [2]. If $a \in L^\infty(\mathbf{T})$ is piecewise continuous, then, for each $\varepsilon > 0$, $\operatorname{sp}_\varepsilon T_N$ converges to $\operatorname{sp}_\varepsilon T$ as $N \to \infty$. The following question arises: If a is discontinuous, is the convergence still fast enough to be compelling for modest values of N?

*Received by the editors August 18, 2000; accepted for publication by P. Van Dooren May 3, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/simax/24-2/37697.html

†Fakultät für Mathematik, TU Chemnitz, 09107 Chemnitz, Germany (albrecht.boettcher@mathematik.tu-chemnitz.de).

‡Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK (LNT@comlab.ox.ac.uk). Current address of the second author: Department of Computational and Applied Mathematics, Rice University, 6100 Main Street—MS 134, Houston, TX 77005-1892 (embree@rice.edu). The research of this author was supported by UK Engineering and Physical Sciences Research Council grant GR/M12414.

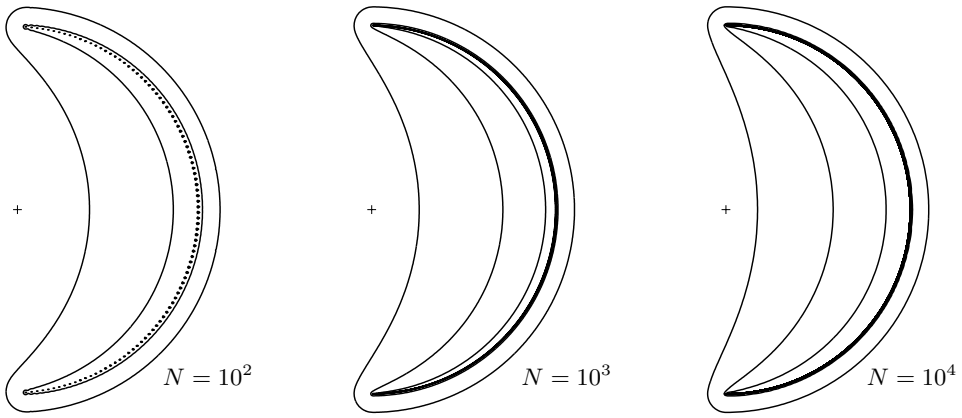


FIG. 1. *Eigenvalues and ε-pseudospectra for the Toeplitz matrices T_N given by* (1) *for three values of N with $\varepsilon = 10^{-1}$, 10^{-2}, and 10^{-3} (from the outside in). The cross (+) marks the origin. Except in the first image, the eigenvalues are so numerous that they appear fused into a curve. The thickness of this curve is actually due to the boundaries of the 10^{-2}- and 10^{-3}-pseudospectra; the boundary of the 10^{-3}-pseudospectrum also affects the thickness of the middle eigenvalues in the first plot. We believe these images are correct to plotting accuracy.*

We have found that the answer is no. If the symbol is discontinuous, the rate at which $\|(zI-T_N)^{-1}\|$ and $\|V_N\|\|V_N^{-1}\|$ increase as $N \to \infty$ may drop from exponential to algebraic, changing the qualitative nature of the pseudospectra strikingly.

We consider the following simple example. Take a such that $a(\mathbf{T})$ is the right half of the unit circle, specifically, $a(e^{i\theta}) = ie^{-i\theta/2}$ for $\theta \in [0, 2\pi)$. Then $\operatorname{sp} T$ is the closed right half of the unit disk, and T_N is a dense Toeplitz matrix whose entries are given by the Fourier coefficients of the symbol

$$(T_N)_{jk} := \frac{1}{\pi(j-k+\frac{1}{2})}, \qquad j,k = 1,\dots,N. \tag{1}$$

Figure 1 shows numerically computed ε-pseudospectra of T_N for $N = 100$, 1000, and 10 000, with $\varepsilon = 10^{-1}$, 10^{-2}, and 10^{-3}. Note how far they are from $\operatorname{sp} T$ for the smaller values of ε and how the interior arcs approximate circles passing through $\pm i$. Figure 2 shows resolvent norms as a function of N for points on the real axis. For $z = \frac{1}{2}$, the bound $\|(zI - T_N)^{-1}\|$ grows roughly like $3.8N^{0.30}$. At this rate, the resolvent norm will not exceed 10^5 until $N \approx 10^{15}$. For $z = 0$, $\|(zI - T_N)^{-1}\|$ grows roughly like $0.4 \log N + 1.5$; it will not exceed 10^5 until $N \approx 10^{108572}$. This behavior is related to the "Moler phenomenon," the observation that the norm of the matrix (1) approaches 1 spectacularly fast as $N \to \infty$, while the smallest singular value decays to 0 very slowly [5, section 4.5], [16].

Here is a mathematical foundation for these observations. Let a be a piecewise C^2 function with at most one jump discontinuity, say, at $e^{i\theta_0} \in \mathbf{T}$. For z outside $a(\mathbf{T})$, let $\arg(a - z)$ be any continuous argument of $a - z$ on $\mathbf{T} \setminus \{e^{i\theta_0}\}$. Define α_z, the Cauchy index of a with respect to z, by

$$\alpha_z = \frac{1}{2\pi}(\arg(a(e^{i(\theta_0+2\pi-0)}) - z) - \arg(a(e^{i(\theta_0+0)}) - z)),$$

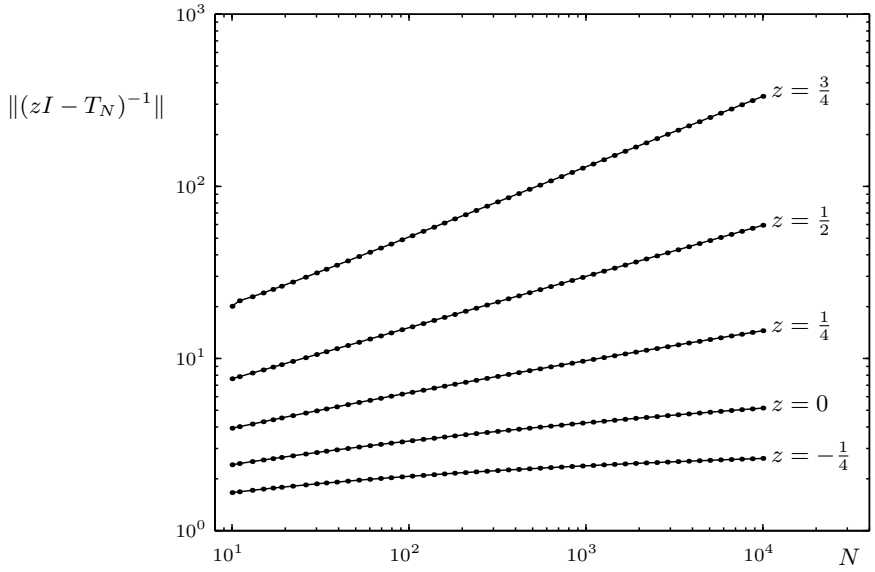


FIG. 2. *The resolvent norm as a function of N for the class of matrices* (1). *The growth is algebraic for $z = \frac{1}{4}, \frac{1}{2}$, and $\frac{3}{4}$ and logarithmic for $z = 0$. For $z = -\frac{1}{4}$, $\|(zI - T_N)^{-1}\|$ is bounded by* 4 (*see Theorem* 3.19 *of* [5]).

and put $\beta_z = |\alpha_z|$. If $\beta_z < \frac{1}{2}$, then $zI - T_N$ is invertible for all sufficiently large N, and it is well known that $\|(zI - T_N)^{-1}\| = O(1)$ in this case [7]. If $\beta_z \geq 1$, then $\|(zI - T_N)^{-1}\|$ may grow exponentially, as trigonometric polynomials (i.e., banded matrices) with nonzero winding number about z show. The following result tells us that, for $\frac{1}{2} \leq \beta_z < 1$, we have just algebraic growth at a known rate.

THEOREM. *If $\frac{1}{2} \leq \beta_z < 1$, then, for every $\delta > 0$, there exist positive constants C_z and $D_{z,\delta}$ such that*

$$C_z N^{2\beta_z - 1} \leq \|(zI - T_N)^{-1}\| \leq D_{z,\delta} N^{2\beta_z - 1 + \delta} \tag{2}$$

for all sufficiently large N.

In the example (1), we have $\beta_z < \frac{1}{2}$ for all z outside $\operatorname{sp} T$ and $\beta_z = \frac{1}{2}$ for $z \in (-i, i)$. For z in the interior of $\operatorname{sp} T$, we have

$$\beta_z = 1 - \frac{1}{\pi} \arctan \frac{1}{x}, \tag{3}$$

where $x \in (0, 1)$ is the point at which the circular arc through $-i, z, i$ intersects the real line. In particular, $\frac{1}{2} < \beta_z < \frac{3}{4}$, and hence, by our theorem, the resolvent norm increases at most like $O(N^{1/2})$ for z in the interior of $\operatorname{sp} T$, explaining the slow convergence seen in Figure 1. Moreover, formula (3) also reveals why the interior arcs of Figure 1 are close to circles passing through $-i$ and i. Finally, our theorem explains Figure 2. For $z = \frac{1}{2}$, for example, we have $2\beta_z - 1 = 0.295\ldots$, in good agreement with the growth $3.8N^{0.30}$ estimated numerically.

Sketch of the proof of the theorem. The proof of the upper bound in (2) can be based on the argument used to prove Theorem 6.1(c) of [4]: A theorem by Verbitsky and Krupnik (see, e.g., Theorem 7.20 of [5]) states that the resolvent norm is uniformly bounded on certain weighted ℓ^p spaces, and appropriate choice of these spaces,

together with Hölder's inequality, gives the ℓ^2 estimate $O(N^{2\beta_z-1+\delta})$. To prove the lower bound in (2), assume that $\frac{1}{2} \le \alpha_z < 1$. (The case $-1 < \alpha_z \le -\frac{1}{2}$ can be reduced to this case by passing to adjoints.) We can write $a - z = c_z \varphi_{\gamma_z}$, where c_z is a continuous and piecewise C^2 function with no zeros on $\mathbf{T}$ and with zero winding number and where φ_{γ_z} is a certain canonical piecewise continuous function with a single jump (see, e.g., pp. 170–171 and 182 of [5]). Here γ_z is a complex number whose real part equals α_z. By Cramer's rule, the $(N,1)$ entry of $(zI - T_N)^{-1}$ is $(-1)^{N+1}$ times the quotient of two Toeplitz determinants,

$$\left[(zI - T_N)^{-1}\right]_{N,1} = (-1)^{N+1} \frac{D_{N-1}(c_z \varphi_{\gamma_z - 1})}{D_N(c_z \varphi_{\gamma_z})},$$

and, since $|\operatorname{Re} \gamma_z| < 1$ and $|\operatorname{Re} \gamma_z - 1| < 1$, we can invoke Refinement 5.46 of [5] (which proves an important special case of the Fisher–Hartwig conjecture) to conclude that the absolute value of $[(zI - T_N)^{-1}]_{N,1}$ is asymptotically equal to a nonzero constant times

$$\left| \frac{N^{-(\gamma_z - 1)^2}}{N^{-\gamma_z^2}} \right| = \left| N^{2\gamma_z - 1} \right| = N^{2 \operatorname{Re} \gamma_z - 1} = N^{2\beta_z - 1}.$$

As the norm of $(zI - T_N)^{-1}$ is greater than the modulus of its $(N,1)$ entry, we arrive at the lower bound of (2). □

For the matrix (1) at $z = 0$, the estimate (2) asserts that $C \le \|T_N^{-1}\| \le D_\delta N^\delta$ for arbitrary $\delta > 0$. Using the Cauchy–Toeplitz structure of (1), Tyrtyshnikov [16] showed that we actually have

$$C \log N \le \|T_N^{-1}\| \le D \log N.$$

We may summarize our observations as follows. Since the pseudospectra, or resolvent norms, converge, T_N must "behave" as if $\operatorname{sp} T_N = \operatorname{sp} T$ for sufficiently large N. However, it is worth bearing in mind that a typical macroscopic physical system has on the order of 10^8 or 10^{10} atoms or molecules in each direction (on the order of the cube root of Avogadro's number or somewhat more). Thus, for T_N to behave like T, the dimension N will have to be larger than the numbers that usually pass for infinity in the physics of gases, liquids, and solids. Said another way, if one found a physical application governed by a matrix of the form (1), even if the dimension were very large, it is unlikely to be large enough to make approximation by the operator limit $N = \infty$ physically appropriate for spectral analysis of the system.

As a further example, Figure 3 presents the Toeplitz matrices associated with the symbol $a(e^{i\theta}) = \theta e^{i\theta}$. The eigenvalues of these finite Toeplitz matrices have been studied by Basor and Morrison [1]. Our theorem provides us with the growth rate of the resolvent norm as $N \to \infty$ in the regions where $\beta_z < 1$. Computational evidence suggests that the same rate is valid throughout the interior of the spectrum, although the values of β_z range up to $\frac{3}{2}$.

One could attempt to generalize our theorem and to raise conjectures suggested by our computations, but we will not pursue this here as our purpose is to point out the slow convergence phenomenon as briefly as possible.

Note added in proof. We wish to point out another class of problems where there exists a gap between algebraically and exponentially growing resolvent norms: certain nonsymmetric matrices related to the nonsymmetric Anderson models developed by Hatano and Nelson in the field sometimes known as nonhermitian quantum

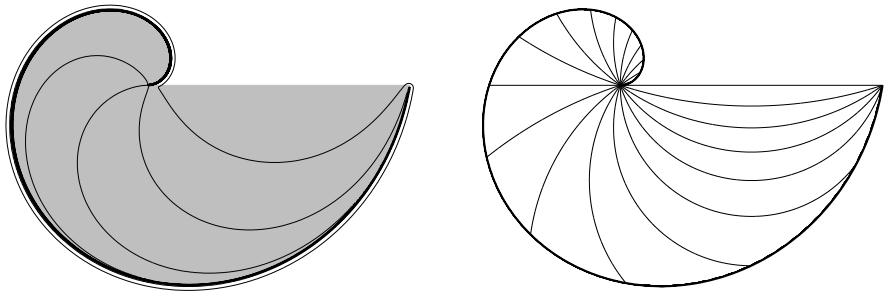


FIG. 3. *Slow growth for the symbol* $a(e^{i\theta}) = \theta e^{i\theta}$. *On the left are computed eigenvalues and* ε*-pseudospectra for the Toeplitz matrix of dimension* $N = 1000$ *for* $\varepsilon = 10^{-1}$, 10^{-2}, 10^{-3}, 10^{-4}. (*The eigenvalues appear fused into a curve near the essential range of* a.) *The shaded region shows the spectrum of the corresponding infinite dimensional operator. On the right are contour lines of constant* β_z *for* $\beta_z = 0.5, 0.55, \ldots, 1.45$ (*clockwise from right*).

mechanics [8]. In [15], it is shown that, for such matrices, the resolvent norm may grow algebraically in one part of the complex plane and exponentially in another. (An example is shown for a matrix of dimension one million, where the discrepancy in norms is between a few thousand and 10^{99698}.) In these problems, as for Toeplitz matrices, it is likely that regions of exponentially large resolvent norm would "act like spectrum" in a physical application whereas other regions would not.

Acknowledgments. Some of our calculations were performed on the SGI Cray Origin2000 at the Oxford Supercomputing Center; the third plot of Figure 1 involved computation of the minimum singular value of 10^4 dense square matrices each of dimension 10^4. We thank Richard Brent and Walter Gander for suggestions concerning fast Toeplitz algorithms and Nick Birkett, Jeremy Martin, and Tom Wright for advice concerning implementation and execution.

REFERENCES

[1] E. L. BASOR AND K. E. MORRISON, *The Fisher-Hartwig conjecture and Toeplitz eigenvalues*, Linear Algebra Appl., 202 (1994), pp. 129–142.

[2] A. BÖTTCHER, *Pseudospectra and singular values of large convolution operators*, J. Integral Equations Appl., 6 (1994), pp. 267–301.

[3] A. BÖTTCHER AND S. GRUDSKY, *Toeplitz band matrices with exponentially growing condition numbers*, Electron. J. Linear Algebra, 5 (1999), pp. 104–125.

[4] A. BÖTTCHER AND B. SILBERMANN, *Toeplitz operators and determinants generated by symbols with one Fisher–Hartwig singularity*, Math. Nachr., 127 (1986), pp. 95–124.

[5] A. BÖTTCHER AND B. SILBERMANN, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, 1999.

[6] I. GOHBERG, *On the application of the theory of normed rings to singular integral equations*, Uspekhi Mat. Nauk, 7 (1952), pp. 149–156 (in Russian).

[7] I. GOHBERG, *Toeplitz matrices composed of the Fourier coefficients of piecewise continuous functions*, Funktsional. Anal. i Prilozhen., 1 (1967), pp. 91–92 (in Russian).

[8] N. HATANO AND D. R. NELSON, *Localization transitions in non-Hermitian quantum mechanics*, Phys. Rev. Lett., 77 (1996), pp. 570–573.

[9] H. J. LANDAU, *On Szegő's eigenvalue distribution theory and non-Hermitian kernels*, J. Anal. Math., 28 (1975), pp. 335–357.

[10] L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162–164 (1992), pp. 153–185.
[11] P. Schmidt and F. Spitzer, *The Toeplitz matrices of an arbitrary Laurent polynomial*, Math. Scand., 8 (1960), pp. 15–38.
[12] P. Tilli, *Some results on complex Toeplitz eigenvalues*, J. Math. Anal. Appl., 239 (1999), pp. 390–401.
[13] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1992, pp. 234–266.
[14] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.
[15] L. N. Trefethen, M. Contedini, and M. Embree, *Spectra, pseudospectra, and localization for random bidiagonal matrices*, Comm. Pure Appl. Math., 54 (2001), pp. 595–623.
[16] E. E. Tyrtyshnikov, *Singular values of Cauchy–Toeplitz matrices*, Linear Algebra Appl., 161 (1992), pp. 99–116.
[17] H. Widom, *Eigenvalue distribution for nonselfadjoint Toeplitz matrices*, in Toeplitz Operators and Related Topics, Oper. Theory Adv. Appl. 71, Birkhäuser, Basel, 1994, pp. 1–8.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 490–508

ARRAY ALGEBRA EXPANSION OF MATRIX AND TENSOR CALCULUS: PART 1*

URHO A. RAUHALA†

Abstract. Array algebra expands the foundations of linear and nonlinear estimation theories, differential and integral calculus, numerical analysis, and fast transform techniques. It originates from an extension of the two-dimensional Kronecker or tensor products and related operators of the traditional vector, matrix, and tensor calculus using the general theory of matrix inverses called "loop inverses." A summary of the foundations of multilinear array algebra and loop inverse estimation is presented in part 1 of this paper. It is then expanded to include the latest developments in nonlinear estimation and applied mathematics using some unified matrix and tensor operators. The new operators are used in part 2 to derive the general theory of direct solution (one "hyper" iteration) techniques of rank-deficient nonlinear systems as an expansion of the loop inverse estimators and Q-surface tensor solution.

Key words. multilinear array algebra, tensor product, loop inverse estimation, fast transforms, nonlinear array polynomials, Q-surface, nonlinear tensor methods, decentered normals, multigrid normals, array relaxation

AMS subject classifications. 15A09, 15A69, 15A72, 65F20, 65F30, 65F50, 65H10, 65K10, 65M10

PII. S089547980240683X

1. Introduction.

Matrix notations serve to express consistent and overdetermined linear systems of equations

$$\underset{M,N}{A}\ \underset{N,1}{x} = \underset{M,1}{f+v},$$

$$\sum_{j=1}^{N} a(i,j)x(j) = f(i)+v(i), \quad i=1,2,\ldots,M,\ M\ge N, \tag{1.1}$$

where the vector f contains the observed values with the unknown residual error vector v and matrix A expresses each element of column vector $f+v$ as a linear function of the N elements in vector x. Certain types of problems can be formulated so that matrix A becomes a tensor or Kronecker product $\otimes$ of two matrices with dimensions m_1, n_1 and m_2, n_2 such that $M=m_1m_2$, $N=n_1n_2$. The starting point for achieving this problem formulation rearranges the $N=n_1n_2$ parameters of vector x into a two-dimensional (2D) n_1n_2 array X. An example is the separable polynomial parametric model $f(u_1,u_2)$ of the observed values in two variables u_1, u_2:

$$f(u_1,u_2) = \underset{1,n_1}{\begin{bmatrix} 1 & u_1 & u_1^2 & \cdots \end{bmatrix}}\ \underset{n_1n_2}{X}\ \underset{n_2,1}{\begin{bmatrix} 1 & u_2 & u_2^2 & \cdots \end{bmatrix}^T} = \underset{1,n_1}{a_1(u_1)}\ \underset{n_1n_2}{X}\ \underset{n_2,1}{a_2(u_2)^T}$$

$$= \sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2} u_1^{j_1-1}\,u_2^{j_2-1}\,x(j_1,j_2). \tag{1.2}$$

This equation can be interpreted as expressing a function value at a point (u_1,u_2), where u_1 and u_2 can be regarded as 2D coordinates. The tensor product often arises

*Received by the editors December 5, 2001; accepted for publication (in revised form) by A. H. Sayed April 9, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/simax/24-2/40683.html
†BAE SYSTEMS, P.O. Box 509008, San Diego, CA 92150-9008 (urho.rauhala@baesystems.com).

when the observed values of vector f (say, gray values of a digital image) form a regular $m_1 m_2$ grid F in two variables. The elements of the column vectors $f = vec(F)$ and $v = vec(V)$ of (1.1) are then declared as $m_1 m_2$ arrays F and V in the same fashion as a computer stores a 2D array. (The column-by-column stacking operator *vec* is assumed here.) The "design" matrix A now achieves the special block matrix structure of the tensor product of two small one-dimensional (1D) design matrices in each variable; namely,

$$\underset{M,N}{A} = \underset{m_1 n_1}{A_1} \otimes \underset{m_2 n_2}{A_2} = \{a_2(i_2, j_2)\}\, A_1,\ i_2 = 1, 2, \ldots, m_2,\ j_2 = 1, 2, \ldots, n_2. \tag{1.3}$$

Matrices A_1, A_2 are found by evaluating the polynomial basis functions in vectors $a_1(u_1)$ and $a_2(u_2)$ of (1.2) at the m_1 locations of u_1 and m_2 locations of u_2 of the observed $m_1 m_2$ grid F. The system of equations (1.1) with the special tensor product matrix A of (1.3) is expressed more efficiently by the pre- and postmultiplication of array X in the matrix equation

$$\begin{aligned} &\underset{m_1 n_1}{A_1}\ \underset{n_1 n_2}{X}\ \underset{n_2 m_2}{A_2^T} = \underset{m_1 m_2}{F} + \underset{m_1 m_2}{V}, \\ &\sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} a_1(i_1, j_1)\, a_2(i_2, j_2)\, x(j_1, j_2) = f(i_1, i_2) + v(i_1, i_2), \\ &\qquad\qquad i_1 = 1, 2, \ldots, m_1,\ i_2 = 1, 2, \ldots, m_2, \\ &a_1(i_1, j_1) = u_1(i_1)^{j_1 - 1},\ a_2(i_2, j_2) = u_2(i_2)^{j_2 - 1}. \end{aligned} \tag{1.4}$$

The basic idea of multilinear array algebra [18], [19], [20], [21], [22] expands the problem to three variables, which, in analogy to the problem of two variables, makes matrix A a tensor product of three matrices. The three-dimensional (3D) arrays of X, F, and V can no longer be expressed by the traditional vector, matrix, and tensor notations. The notation of a "matrix" is extended to a 3D array X simply by adding the third index to its elements $x(j_1, j_2, j_3)$. The matrix multiplication is extended beyond the pre- and postmultiplication of an array X into a new "backside"-multiplication by adding the third sum over index j_3 into (1.4). In terms of an extended matrix and tensor calculus, this can be interpreted as a matrix postmultiplication of each "depth" slice or n_2 2D $n_1 n_3$ subarrays of X by the $n_3 m_3$ matrix A_3^T of the third variable. The three matrix multiplications of a 3D array are expressed as

$$\begin{aligned} &\underset{m_1 n_1}{A_1}\ \underset{n_1 n_2 n_3}{X}\ \underset{n_3 m_3}{A_3^T}\ \underset{n_2 m_2}{A_2^T} = \underset{m_1 m_2 m_3}{F} + \underset{m_1 m_2 m_3}{V}, \quad A = A_1 \otimes A_2 \otimes A_3 = \{a_3(i_3, j_3) A_1 \otimes A_2\}, \\ &\sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \sum_{j_3=1}^{n_3} a_1(i_1, j_1)\, a_2(i_2, j_2) a_3(i_3, j_3)\, x(j_1, j_2, j_3) = f(i_1, i_2, i_3) + v(i_1, i_2, i_3). \end{aligned} \tag{1.5}$$

The tremendous savings in computing time and memory of the resulting array multiplications are more appreciated in the inverse process of finding some optimal estimators of the unknown parameter array X using the known observed array F. For instance, the entire set of parameters minimizing the norm of the residual array V is

found in the general multilinear least squares solution of (1.5) by [19]

$$\underset{n_1n_2n_3}{X} = \underset{n_1m_1}{G_1}\ \overset{\overset{G_3^T}{m_3n_3}}{\underset{m_1m_2m_3}{F}}\ \underset{m_2n_2}{G_2^T} + \underset{n_1n_2n_3}{U} - \underset{n_1n_1}{G_1A_1}\ \overbrace{\underset{n_1n_2n_3}{U}}^{\underset{n_3n_3}{(G_3A_3)^T}}\ \underset{n_2n_2}{(G_2A_2)^T}. \tag{1.6}$$

Array U is arbitrary in the general case where an infinite set of X can reflect the optimal and unique estimators of V and adjusted F in the observable domain. The general least squares inverses G_1, G_2, G_3 of matrices A_1, A_2, A_3 satisfy the single condition $A_j^TA_jG_j = A_j^T$ without any restrictions on the rank of matrices A_j, $j = 1, 2, 3$. The unique least squares minimum norm solution of X is found by having $U = 0$ (sufficient condition) and $G_j = A_j^+$, the pseudoinverse of A_j satisfying the four Moore–Penrose conditions [15]. The full-rank special case of $r(A) = N$ is called herein the L-inverse of A. It consists of the tensor product of the full-rank L-inverses $G_j = A_j^L = (A_j^TA_j)^{-1}A_j^T$ of all three matrices A_j in the traditional solution, where the regular inverse of a square matrix A^TA exists and all elements of X are unbiasedly estimable. They also have the minimum ("best") variance and are called the best linear unbiased estimator (BLUE) of estimable X. This explains the wide range of applications of the least squares estimation.

The remarkable operation count of the unique BLUE array estimator $\hat{X}$ is found from the algebraic expression of the general matrix and shorthand tensor notations of (1.6):

$$\hat{x}(j_1, j_2, j_3) = \sum_{i_3=1}^{m_3} g_3(j_3, i_3) \sum_{i_2=1}^{m_2} g_2(j_2, i_2) \sum_{i_1=1}^{m_1} g_1(j_1, i_1)\ f(i_1, i_2, i_3). \tag{1.7}$$

Although the end result of array $\hat{X}$ in (1.7) is invariant of the order of the summations or array multiplications, the operation count depends on this order and the values of n, m. The summation of the two inner loops over indices i_1, i_2 (pre- and postmultiplications of array F) in (1.7) yields

$$op1, 2 = m_3(n_1m_1m_2 + n_1m_2n_2),$$

resulting in the 2D subarrays $Y(i_3) = G_1F(i_3)G_2^T$, $i_3 = 1, 2, \ldots, m_3$. The summation over index i_3 involves a postmultiplication of n_2 2D n_1m_3 slices of array Y by the m_3n_3 matrix G_3^T with

$$op3 = n_2(n_1m_3n_3).$$

The total $op1, 2 + op3$ is orders of magnitude less than a typical count $N^3/6 = (n_1n_2n_3)^3/6$ in solving for N parameters. The operation count of explicitly forming the small inverse matrices G is only about $n_1^3 + n_2^3 + n_3^3$, which is negligible to the count of the array multiplications. The traditional rules of linear solution and matrix factorization techniques (such as SVD, Q-R, L-R) to avoid an explicit inversion of a square or rectangular matrix are often reversed in array algebra. In the example of a full-rank case with $N = M = n^3 = 100^3$, the solution of N or one million parameters would take on the order of $N^3/6 = n^9/6 = 0.167 \times 10^{18}$ operations in the traditional case if attempted. The array solution requires only $3n^4 = 3nN$ or 3×10^8 operations, taking just few seconds in modern computers. This fast operation count has no restrictions on the matrices G_1, G_2, and G_3.

Some practical applications of array algebra have exploited a special structure of matrices G, such as the sparseness or shift invariance of Toeplitz matrices. The tensor product properties of 1D design matrices A and their inverses G, such as the traditional "fast transform" matrices of Fourier, Karhunen–Loeve, cosine, Hadamard, and other transforms, are discussed in next section. Use of these special matrix operators results in a typical operation count of less than $100N$ for 2D and 3D array solutions [20], [22], [27].

Note the parallel structure of the three passes of array multiplications such that, for instance, the premultiplication of the m_2m_3 column vectors of F by G_1 could be done in parallel. The second pass would perform the resulting n_1m_3 row vector by matrix G_2^T multiplications in parallel, etc. Thus the example of $N = 100^3$ parameters would require three passes of a 100x100 matrix by 100x1 vector multiplications with 10,000 parallel processors. For full 100x100 matrices, each processor would perform the total of 30,000 operations to accomplish the solution of one million parameters. A sparse or other special structure of all matrices G could further reduce these operations such that the solution of one million parameters in 0.001 seconds is achieved with the 10,000 parallel processors, each with the performance of only one million operations per second [20].

The traditional solution has to invert or factorize a matrix of order $N = 10^6$, requiring $N^2 = 10^{12}$ storage elements (if attempted). The array solution inverts three matrices of order $n = 100$ with 10,000 elements. Needless to say, these tremendous savings in computing time and memory can be converted into huge savings in the computer capacity. Often, the most modest computers can beat the performance of the most powerful computers by a reformulation of the software solution of a given problem. Many new problem solutions and technologies become feasible, for the very first time, that otherwise would be unimaginable, as discussed in part 2 of this paper.

The analyst is challenged to rethink the problem in terms of the array formulation, such as in (1.4) and (1.5). This rethinking to exploit some special "fast" solution rules is related to the fast transform techniques in signal processing, as discussed in section 2. Array algebra expands this field to the general separable math models of linear algebra in the traditional parametric domain. Sections 3 and 4 introduce the more general tools of linear loop inverses and the general theory of linear estimators. They make the array algebra applicable to virtually any problem in linear algebra by performing the math modeling and adjustment in the directly observable and always unbiasedly estimable space domain. The inverse projection of the space domain estimators onto the original parametric domain recovers and expands the traditional operators of general matrix inverses and linear estimators. For more details, see [18], [19], [20], [21], [22], [23], [24] [25], [26], [27], [4], [8], [9], [13], [28], [31], [33].

2. General fast transforms of signal processing. The multilinear array multiplications and their inverse operators expand the special case of square (usually 2x2) matrices of tensor products used in the 1D fast Fourier transform (FFT) [7], [10]. The special tensor product structure of the symmetric and square NxN Fourier matrix A is achieved by a proper reordering of the columns/rows such that it can be split into a tensor product where A_2 is a 2x2 matrix of complex conjugates. The postmultiplication of the two-column array $F = [F_1, F_2]$ is found by multiplying F_2 of $N/2$ elements with the complex factor. The result is added and subtracted to/from F_1. The process is repeated by considering the remaining A_1 as new Fourier matrices A of the two branches of a tree structure, reducing to 2x2 matrices in the last pass. The 3D FFTs are special cases of an array solution (1.6), where $G = A^{-1} = A^T$, as detailed in [20],

[22], [24].

Similar tensor product factorizations are applicable for cosine transforms and the related finite element SVD regularization or Karhunen–Loeve transforms [27]. The starting idea transforms them into the complex numbers to identify their similarity with the complex FFT. The inverse discrete cosine transform (IDCT) of N elements is needed in the time-critical task of data decompression, such as in JPEG and MPEG industry standards. The traditional signal processing required two FFTs of $2N$ elements for its implementation, while the "fast" IDCT of array algebra uses two parallel FFT pipelines of $N/2$ elements. The design removes the bottleneck of interface memories between the passes by running both pipelines in parallel. The output of each pass feeds the input of the next pass without other than a few delay registers (vs. refresh memories) between the passes. The new process saves two passes such that, e.g., the brute-force $N = 8$ transform uses as many (four) passes as the new $N = 32$ IDCT.

The special 1D fast transform exploitation of tensor products is not generally applicable in real-world problems confronting an analyst. The fast transforms of signal processing have often served electronic industries solving some specific problems by a special hardware. Today, the general-purpose computers are getting so powerful that we, as the problem analysts and software designers, would prefer some fast but more generic modeling and solution techniques. Array algebra is centered on such a technique by expanding operators G of (1.6) and the underlying estimation theories. Its 1D version is outlined next to serve its expansion to multilinear modeling through (1.6). The polynomial model is used as an example expanding the bilinear Hadamard transform to more general matrices A and G from the orthogonal 2D case of (1.4) with $n_1 = n_2 = m_1 = m_2 = 2$ and

$$A_1 = A_2 = G_1^T = G_2^T = 1/2 \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

3. Full-rank basis transforms of loop inverse estimation. The idea of array algebra and loop inverses in [18], [19] started from separable modeling in the observable space domain using grid locations for the variables of the parametric interpolation model. The basis transforms among the model and space domain parameters can be identified using the 1D linear model $f(u)$ in variable u with n unknown model domain coefficients in column vector X with n basis functions, say, polynomials of u in the row vector $a(u)$ of n elements

$$f(u) = \underset{1,n}{a(u)}\ \underset{n,1}{X}, \quad a(u) = \left[1, u, u^2, \ldots, u^{n-1}\right]. \tag{3.1}$$

This parametric model of a single set of parameters X is equivalent to an infinite set of space domain interpolation functions. The interpolation takes place from an observable set of n unknown true values in column vector F_0 at any properly chosen (but otherwise freely selectable) variable u locations, such as a regular grid or profile, in

$$f(u) = \underset{1,n}{k(u)}\ \underset{n,1}{F_0}, \quad \underset{1,n}{k(u)} = \underset{1,n}{a(u)}\ \underset{n,n}{A_0^{-1}}. \tag{3.2}$$

The row vector $k(u)$ contains Lagrange's basis functions of variable u for the polynomial basis functions $a(u)$.

The two unknown sets of model domain X and space domain F_0 parameters are connected by the independent and consistent n equations at the chosen grid of (fictitious) variable locations

$$\underset{n,n}{A_0}\,\underset{n,1}{X} = \underset{n,1}{F_0} \Leftrightarrow X = A_0^{-1}F_0. \tag{3.3}$$

The full-rank n, n matrix A_0 of the parameter transform is found by evaluating the row vector $a(u)$ at the variable locations $u = u_1, u_2, \ldots, u_n$ chosen by the problem analyst.

The preferred computational model of fitting $m \geq n$ observed values F at arbitrary locations of variable u is the interpolation function (3.2) with the space domain parameters F_0, even in this starting case of one variable u with full-rank "problem matrix" A_0. The linear equations of m observed function values F and unknown residuals V are written for both models of $f(u)$ by

$$\underset{m,n}{K}\,\underset{n,1}{F_0} = \underset{m,1}{F} + \underset{m,1}{V} \;\Leftrightarrow\; \underset{m,n}{A}\,\underset{n,1}{X} = \underset{m,1}{F} + \underset{m,1}{V}\,, \tag{3.4}$$

where the m, n design matrices K and A are found by evaluating the row vectors $k(u)$ and $a(u)$ of n elements at the m locations of variable u of the observed values $f(u)$, some of which may or may not coincide with the n chosen grid locations of the unknown values F_0.

The standard BLUE full-rank least squares estimators of both sets of parameters are

$$\underset{n,1}{\hat{F}_0} = \underset{n,m}{H}\,\underset{m,1}{F} \Leftrightarrow \underset{n,1}{\hat{X}} = \underset{n,m}{G}\,\underset{m,1}{F}\,, \tag{3.5}$$

where H and G are the full-rank L-inverses of K and A. The substitution of the "interpolation matrix" $K = AA_0^{-1}$ yields the expected result for the "filter matrix" H,

$$\underset{n,m}{H} = \underset{n,n}{(K^TK)^{-1}}\,\underset{n,m}{K^T} = \underset{n,n}{A_0}\,\underset{n,m}{G} = A_0(A^TA)^{-1}A^T. \tag{3.6}$$

The same least squares solution $\hat{F}_0$ is achieved by evaluating the model domain estimator of $\hat{X} = G\,F$ at the chosen variable locations of F_0 by the linear transform $A_0\hat{X}$. However, the direct estimator $\hat{F}_0 = H\,F$ using the covariance matrix H has several advantages for a problem analyst:

- The condition number of matrix K is superior to that of A. Without loss of generality, the problem analyst can choose A_0 as a horizontal partition of A by coinciding the n "problem observations" of F_0 with a subset of F. The corresponding partition of K consists of a unit matrix of order n, and the remaining partition is often diagonally sparse. Thus this parameter transform acts as a good preconditioning tool for the adjustment of the observed values F. The sequential solution of this partitioned system of equations reveals the connection to the classical direct least squares solution of the residuals by the so-called condition adjustment. This was used for surveying net adjustments with hundreds of parameters by manual calculations before the birth of computers and numerical analysis in times when a matrix inversion beyond 5x5 was considered hopeless [16]. This finding serves the nonlinear estimation theory of loop inverse and Q-surface techniques to be introduced in part 2 of this paper.

- A regular distribution of uniform and uncorrelated observed values results in a Toeplitz filter matrix H with the exception of few first and last rows. A one-time simulation of some central rows of matrix H provides the general equivalent of the linearly shift invariant "impulse response" of signal processing, and the filtering solution $H\,F$ reduces into the convolution of F_0 from F [20], [21], [22], [27].
- The direct solution technique of $\hat{F}_0 = H\,F$ is especially applicable and useful in the rank-deficient case, $r(A) < n$, when the model domain parameters X are not uniquely estimable from F. The direct estimator of one set of parameters F_0, spanning the estimable space, can be projected to the model domain. The resulting estimators of X as a linear function of F reveal new inverse operators G of A in $X = GF + U - GAU$. They expand the theory of general matrix inverses and the foundations of linear estimation, as discussed next.

4. General theory of matrix inverses, linear estimation, and grid modeling. In many practical problems, the formulation of model $f(u)$ with parameters X and the associated basis functions of row vector $a(u)$ is the main task of the analyst before the measurements and adjustment of the observed values F are started. The BLUE property of the least squares solution implies that $E\{V\} = 0$. The model $f(u)$ should capture the systematically behaving "signal" component of the observables such that the remaining random residual errors are normally distributed. It is usually simple for a "problem expert" to find the first order parameters of "physical explanation" which account for the main component of model $f(u)$. As we keep refining the model with some added parameters, the plateau of diminishing returns is achieved. The additional parameters get highly correlated with the existing parameters. Their effect on the observable domain is still beneficial to satisfy the condition $E\{V\} = 0$ but at the expense of the estimability of modeling parameters X.

The estimability problem is related to the field of general matrix inverses and linear estimation theory. The Gauss–Markov model $E\{F\} = AX$, implying $E\{V\} = 0$, can be overly parameterized in X by a problem analyst who is not interested in the parameters themselves but in their linear functions A_0X, such as an entire set of parameters spanning the observable space. It is usually impossible to analytically derive the explicit mathematical expressions of the elements of the row vector $k(u)$ in (3.2) to interpolate the observed values F from any complete set of p parameters F_0 spanning the estimable space. Often, the rank of matrix A is not known before the adjustment in ill-posed problems of $p = r(A) < n$. In some new cases, which cannot be handled by the traditional theory of general inverses and estimation theory, the problem analyst may wish to focus on a subspace of the entire estimable space. This subspace is spanned by $p < r(A)$ independent observables $F_0 = A_0X$ in the consistent system of parameter transform (3.3). The pxn transform matrix A_0 has the rank of p, or its full-rank inverse called "m-inverse" exists in replacing the parameter transform $X = A_0^{-1}F_0$ of (3.3) by the nonunique transform

$$\underset{n,1}{X} = \underset{n,p}{A_0^m}\,\underset{p,1}{F_0} + \underset{n,1}{U} - \underset{n,p}{A_0^m}\,\underset{p,n}{A_0}\,\underset{n,1}{U}, \quad \underset{n,p}{A_0^m} = \underset{n,p}{A_0^T}\,(\underset{p,p}{A_0A_0^T})^{-1}. \tag{4.1}$$

For the minimum norm solution of X, the last two terms vanish (e.g., by having $U = 0$). The mxp "interpolator" $K = AA_0^m$ of (3.4) still has the full column rank of p, so its pxm L-inverse $K^L = (K^TK)^{-1}K^T$ exists. The solution of parameters X is found by the substitution of the direct estimator $\hat{F}_0 = K^LF = H\,F$ into the

minimum norm transform of X in (4.1). This results in the starting case of left sided loop inverses, the Lm-inverse, by

$$\begin{aligned}\underset{n,1}{\hat{X}} &= \underset{n,m}{A^{Lm}}\ \underset{m,1}{F} = \underset{n,p}{A_0^m}\ \underset{p,1}{\hat{F}_0} = \underset{n,p}{A_0^m}\ \underset{p,m}{H}\ \underset{m,1}{F} = \underset{n,p}{A_0^m}(\ \underset{m,n}{A}\ \underset{n,p}{A_0^m})^L\ \underset{m,1}{F} \\ &= \underset{n,p}{A_0^T}\ \underset{p,m}{D^L}\ \underset{m,1}{F} = \underset{n,p}{A_0^T}\ \underset{p,p}{(D^TD)^{-1}}\ \underset{p,m}{D^T}\ \underset{m,1}{F} = \underset{n,n}{N'^{-1}}\ \underset{n,m}{A^T}\ \underset{m,1}{F}\,,\end{aligned} \tag{4.2}$$

where

$$\underset{m,p}{D} = \underset{m,n}{A}\ \underset{n,p}{A_0^T} \qquad \text{and} \qquad \underset{n,n}{N'^{-1}} = \underset{n,p}{A_0^T}(\underset{p,m}{D^T}\ \underset{m,p}{D})^{-1}\ \underset{p,n}{A_0}\,.$$

The Lm-inverse of A becomes invariant of the chosen transform matrix A_0 when $p = r(A)$, producing the unique pseudoinverse A^+ as a special case. In the additional special case of $p = n$, the traditional full-rank L-inverse $A^L = (A^TA)^{-1}A^T$ is recovered. The general case of $p < r(A)$ still provides the BLUE values of the chosen subset F_0 in the estimable space by A_0GF, although the Lm-operator G does not satisfy the general least squares condition $A^TAG = A^T$ nor the starting g-inverse definition $AGA = A$ of the theory of general matrix inverses. The condition of linear functions A_0X (one element at a time), to be unbiasedly estimable as $H\,F$, becomes, under the model $E\{F\} = AX$,

$$A_0 = H\,A = A_0GA. \tag{4.3}$$

The condition (4.3) has no restrictions on G, such as $AGA = A$. For details, see [3], [11], [17], [18], [19], [20], [21], [22], [23]. Some properties of the Lm-inverse are exploited in [14] for $p = r(A)$ and in [12] for $p = n$. A simple example consists of three ($m = 3$) measured differences 1,2,4 in vector F among three ($n = 3$) unknown quantities in vector X. Select a 2x3 matrix A_0 to consist of any combination of two ($p = r(A) = 2$) rows among the 3x3 matrix A of

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

Before proceeding to the nonlinear tensor expansion of loop inverses, the reader is recommended to derive the linear Lm-inverse solution (4.2) and verify it with the pseudoinverse solution. Project the estimator $\hat{X}$ to the space domain to get the adjusted values of F and V. Then select A_0 to consist of any single row of matrix A such that $p = 1 < r(A)$, and repeat the computations. Verify the conditions of estimability $A_0 = A_0GA$, the least squares $A^TAG = A^T$, and the g-inverse $AGA = A$ in each choice of matrix A_0. Then find the direct estimators $\hat{F}_0$ without first computing estimators $\hat{X}$, and compare their values with the projections from $\hat{X}$ of the Lm-inverse computations. These solutions of $\hat{X}$ or $\hat{F}_0$ alone have limited industrial uses without the estimators of their quality. Find the estimable functions F_0 by the test $A_0GA = A_0$, and compute their full-rank covariance matrix

$$(K^TK)^{-1} = A_0GG^TA_0^T = A_0N'^{-1}A_0^T$$

of minimum trace to provide the standard errors for the BLUEs.

The theory of loop inverses expands the Lm-operator to additional parameter transforms and problem matrices A_0. For instance, the mLm-operator provides an

Lm-inverse solution of X satisfying some equality constraints $A_{00}X = C$. Most known cases of the adjustment calculus in surveying and other engineering sciences are recovered as a special case and then expanded to singular systems of equations [19]. This involves both left- and right-sided inverses and their weighted expansions without any practical use of the g-inverse or pseudoinverse computations. The judicious parameter transforms with the full-rank L- and m-inverses provide computational rules, where only regular full-rank inverses are required for square, symmetric, and positive definite matrices.

The problem analyst has to define the system of linear equations by a proper choice of parameters X, F_0 and observed grid F to get the inverse operators G of the fast separable solution (1.6). A general theory of multilinear functions $f(u_1, u_2, u_3, \dots)$ using the space domain grid F_0 is found by extending the functional model of each variable beyond the polynomials (1.2) by more general basis functions in vectors $a(u)$ of (3.1), such as covariance functions, fractals, wavelets, etc. Modeling functions of four or more variables cannot be expressed in the expanded matrix or shorthand tensor notations of (1.5) and (1.6), but the analogy of their algebraic summations (tensor contractions) is preserved. The matrix multiplications (and their inverses) of a k-dimensional array X with k matrices $A, B, C, \dots, K$ form the foundations of a multilinear estimation theory as an expansion of the fast transform techniques [7], [10]. Array algebra in [19] extended the shorthand summation convention of the 3D tensor and matrix equation (1.5) through the superscripts of the matrix and vector operators by

(4.4)
$$\underset{m_1n_1}{A^1}\ \underset{m_2n_2}{B^2}\ \underset{m_3n_3}{C^3} \dots \underset{m_kn_k}{K^k}\ \underset{n_1n_2n_3\dots n_k}{X} = \underset{m_1m_2m_3\dots m_k}{Y},$$
$$\sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2}\cdots\sum_{j_k=1}^{n_k} a(i_1,j_1)b(i_2,j_2)\dots k(i_k,j_k)x(j_1,j_2,j_3\dots j_k) = y(i_1,i_2,i_3\dots i_k).$$

The matrix superscript identifies the index of the array to be contracted with the second matrix index. This implies that the order of this array index equals the order of the matrix columns. The contracted index of the output array is replaced by the first matrix index. The inverse array multiplication of multilinear systems is found by applying the identifiers, such as L, m, Lm, mLm, etc., of the loop and other general matrix inverses as the superscripts of the matrices. For example, the 3D least squares minimum norm estimator of the array equation (1.6) can be denoted by

$$\underset{n_1n_2n_3}{\hat{X}} = \underset{n_1m_1}{A^{Lm1}}\ \underset{n_2m_2}{B^{Lm2}}\ \underset{n_3m_3}{C^{Lm3}}\ \underset{m_1m_2m_3}{F} = \underset{n_1m_1}{A^{Lm}}\ \underset{m_1m_2m_3}{\overset{m_3n_3}{\overset{C^{LmT}}{F}}}\ \underset{m_2n_2}{B^{LmT}}, \tag{4.5}$$

where the subscripts of matrix identifiers are avoided by denoting $A = A_1$, $B = A_2$, $C = A_3$. The order of writing the matrices $A, B, C \dots$ is immaterial in the same fashion as the order of summations in (4.4). These expanded matrix and shorthand tensor notations in the fashion of [8] require some standardization, especially when the work is expanded to the nonlinear systems, as shown next.

5. Transition to nonlinear estimation by global array solution. The introduced array algebra operators are applicable to many problems in linear algebra and related fields. The latest developments in array algebra have centered on the array algebra expansion of nonlinear estimation. Practical applications include the

design of digital image mapping systems in photogrammetric engineering, mission planning, and image understanding. This work was started from the "global" solution of the nonlinear problem in image registration. The fast linear array algebra was applied to combine a huge number of local nonlinear least squares match equations into a simultaneous linear solution of a finite element (globally constrained) network as follows.

The digital elevation model (DEM), digital data compression, image matching, and rectification with the related modeling problems of industrial photogrammetry, inertial network adjustment, and physical geodesy were some of the early applications of the multilinear array algebra [22], [24]. They lead to the theory of "fast" solutions of nonlinear estimation. The technique of regularization and finite element modeling was modified to be applicable for a "fast" solution of linear array relaxation, involving literally millions of parameters in one batch [6], [27]. This technique was applied to the global solution of nonlinear least squares matching (LSM) [20].

The local nonlinear LSM "observation equations" register one finite element window of reference gray values $f(x_i)$ with observed slave gray values $g(x_i)$ of $i = 1, 2, \ldots, m$ local window locations by

$$f(x_i + dx) = g(x_i) + v(x_i). \tag{5.1}$$

A constant unknown shift dx is sought such that the gray values match within the random noise of residuals v. An estimate of parameter dx is sought by minimizing the sum of θ powers of absolute values of residual errors v in

$$\sum_{i=1}^{m} \{abs[f(x_i + dx) - g(x_i)]\}^{\theta} = \text{minimum}. \tag{5.2}$$

This condition of minimum residuals results in the nonlinear normal equation $n(dx)$ of the unknown parameter dx by setting the partial derivative of (5.2) to zero in

$$n(dx) = \sum_{i=1}^{m} f'(x_i + dx)\{abs[f(x_i + dx) - g(x_i)]\}^{\theta-1} = 0. \tag{5.3}$$

The nonlinear normal equation (5.3) is solved iteratively at an initial value dx^0:

$$n'(dx^0)dx = -n(dx^0) \Rightarrow dx = -n'(dx^0)^{-1}n(dx^0). \tag{5.4}$$

Equation (5.4) reduces into the Newton–Raphson (N–R) solution of least squares, $\theta = 2$, at $dx^0 = 0$ by

$$dx = -\left\{\sum_{i=1}^{m}[f'(x_i)^2 + f''(x_i)l(x_i)]\right\}^{-1} \sum_{i=1}^{m}[f'(x_i)l(x_i)], \tag{5.5}$$

where f', f'' are the first and second order partials of function f with respect to the parameter dx at dx^0. The constant terms $l(x_i)$ contain the observed differences $f(x_i + dx^0) - g(x_i)$ by resampling or evaluating the nonlinear function at the shifted locations of the latest estimate dx^0.

Matrix calculus fails to express the 3D array F'' in the general case of more than one parameter in vector dX (vs. scalar dx of our transitional example in this section). The second order partial f'' is often ignored using the truncated Taylor expansion of

$$f(x_i + dx) = f(x_i) + f'(x_i)dx.$$

This makes the observation equations (5.1) linear. An estimator $d\hat{x}$ is found in each iteration from the column vector L of m differences between the predicted $f(x_i + dx^0)$ and observed $g(x_i)$ values by using the regular least squares inverse of the first order partial matrix F' of m rows and one column in

$$d\hat{x} = -F'^L L, \quad \text{where } F'^L = (F'^T F')^{-1} F'^T. \tag{5.6}$$

This solution technique of a nonlinear problem (5.1) is called Newton–Gauss (N–G) in this paper.

The above special cases of least squares solutions with $\theta = 2$ provide the most accurate (minimum variance) unbiased estimator $d\hat{x}$ when dx is estimable. By contrast, the less optimal cross correlation maximizes

$$\sum_{i=1}^{m} f(x_i + dx) g(x_i). \tag{5.7}$$

This object function results in a nonlinear normal equation of the shift parameter dx in analogy to (5.3). The interesting result (the derivation is left to the reader) is that the resulting normal equation of (5.7) is the same as for the minimum residual technique (5.2) when $\theta = 0$. The third interesting special case is $\theta = 1$ of Laplace minimum absolute residuals. It forms the "center of symmetry" for the linear polynomial term of the nonlinear normals.

The practical problem of LSM and many other nonlinear systems is that they cannot be solved locally by the N–G technique because of two main reasons. (1) The N–G technique fails to converge unless the initial values of parameter dx are known near the global minimum within the pull-in range of the first order partials f', and (2) the solution may converge to a local minimum of a biased estimator. The global finite element solution considers the weighted N–G normals (5.4) as the weighted observed values of the unknown 2D grid F_0 of the true dx values in the image space. The redundant observations consist of continuity constraints of finite elements in the fashion of the regularization techniques. The combined network normals of the continuity constraints and the diagonal LSM sample equations are solved very quickly by the linear array algebra expansion of SVD to 2D and 3D array equations involving grid parameters F_0, as detailed in [27] and [8]. The network solution fills optimal values at nodes where the local normal equation (5.4) is ill-posed or close to a homogeneous system, thereby preventing breakdowns in the automated process of image registration.

The global solution technique improved the quality and robustness of image matching in comparison to the traditional N–G method of no global constraints. Since the local nonlinear match equation has only one parameter dx, the Taylor series of $f(x + dx)$ and $f'(x + dx)$ can be expressed using the scalar polynomials of elementary calculus as nonlinear functions of the second and higher order partials. The high order terms in this special case improved the convergence rate and range, especially when $\theta < 2$. The question was how to expand these derivations to the general case, where the matrix and multilinear array notations were only partially applicable. This question and similar work in the general tensor methods prompted the inventions of several new shorthand tensor operators to solve the general loop inverse problem of nonlinear estimation, as discussed in next section.

6. Nonlinear array algebra. New tools of a unified matrix and tensor calculus of nonlinear array algebra will now be introduced for expressing the Taylor polynomials and their solutions (inverse Taylor series) in the general case of loop inverse estimation. The nonlinear transforms and normal equations in parameters dX and dF_0 cannot be expressed by the matrix calculus of only 2D arrays. The "long-hand" indical tensor notations are not compatible with the established matrix notations of nonlinear estimation, and the matrix-like notations of linear array algebra in section 1 are only partially applicable for the task. This prompted an expansion of the unified matrix and tensor notations in analogy to the scalar nonlinear image matching solution of section 5, as discussed next.

6.1. Array algebra expansion of Taylor polynomials. The general polynomial $N(X + dX) = 0$ of minimum residual normal equations is found for a set of $j = 1, 2, \ldots, n$ parameters in column vector dX by applying the Taylor series of $F(X + dX)$ and $F'(X + dX)$ to (5.3). The polynomial for the entire set $F(X + dX)$ of $i = 1, 2, \ldots, m$ observed function values $f(i)$ is found by expanding the array multiplication into *exponential* vector multiplications (repeat contractions). This results in a nonlinear *array polynomial* of the unified matrix and tensor notations in analogy to the scalar polynomials of elementary calculus

$$
\begin{aligned}
\underset{m,1}{F(X+dX)} &= \underset{m,1}{F(X)} + \underset{m,n}{F'}\,\underset{n,1}{dX} + 1/2\,\underbrace{\underset{m,n,n}{F''}\,\underset{n,1}{dX}\,{}^{**}2}_{m,1} + 1/6\,\underbrace{\underset{m,n,n,n}{F'''}\,\underset{n,1}{dX}\,{}^{**}3}_{m,1} + \cdots \\
&= f(i) + \sum_j f'(i,j)dx(j) + 1/2 \sum_j \sum_k f''(i,j,k)dx(j)dx(k) \\
&\quad + 1/6 \sum_j \sum_k \sum_l f'''(i,j,k,l)dx(j)dx(k)dx(l) + \cdots ,
\end{aligned}
\tag{6.1}
$$

where all indices j, k, l vary from 1 to n. The constant $F(X)$ and linear terms $dF = F'dX$ are well known in the traditional N–G solution. The second correction term due to the 3D second order partial array F'' is found as follows. The second order partial at a given point $f(i)$ of $F(X)$ is found by taking a partial derivative of the $j = 1, 2, \ldots, n$ nonlinear first order partials $f'(i, j)$ of matrix F' with respect to $k = 1, 2, \ldots, n$ parameters of X. This results in a symmetric matrix. It is premultiplied by the row vector dX^T, and the result is postmultiplied by the column vector dX, thereby contracting the last two indices j, k of elements $f(i, j, k)$. This is denoted, in analogy to the scalar $f''dx^2$, as $F''dX^{**}2$ [26]. The indical tensor notations of (6.1) are used in [5]. Other notations for the high order Taylor term operators appear in separate fields [1], [2], [29], [30], [32].

An array polynomial $\tilde{F}dX^{**}k$ has a vector dX of as many elements as there are implied by each of the k last indices of array $\tilde{F}$. Summations or contractions are performed over these indices. In (6.1), the sum of the product of all partial derivative arrays and the *vector powers* (repeated array contractions) of dX reduce into a scalar at each point $f(i)$. These scalars are collected into a vector of $i = 1, 2, \ldots, m$ elements representing the array polynomial approximation of the nonlinear parametric model $F(X + dX)$ for the observed values $g(i)$.

The array polynomial of matrix $F'(X + dX)$, needed in the generalized nonlinear

normals of (5.3), is found in analogy to its scalar counter part and (6.1) by

$$F'(X+dX) = \underset{m,n}{F'} + \underbrace{\underset{m,n,n}{F''}\,\underset{n,1}{dX}\,{}^{**}1}_{m,n} + 1/2\,\underbrace{\underset{m,n,n,n}{F'''}\,\underset{n,1}{dX}\,{}^{**}2}_{m,n} + \cdots$$

$$\begin{aligned} f'(X+dX)(i,j) &= f'(X)(i,j) + \sum_{k=1}^{n} f''(i,j,k)dx(k) \\ &\quad + 1/2 \sum_{k=1}^{n}\sum_{l=1}^{n} f'''(i,j,k,l)dx(k)dx(l) + \cdots , \end{aligned} \tag{6.2}$$

where the last index of F'' is contracted to get an m,n correction matrix dF'. The first power of dX has to be explicitly written as ${}^{**}1$. Otherwise, according to (1.5) of the 3D postmultiplication of array in $F''dX$, the second index of the 3D array F'' would be contracted. Due to the symmetry in indices j, k, the result would be correct in the m, 1, n array, but it requires an array transpose or exchange among the second and third index [19], [26].

6.2. Array polynomials of nonlinear least squares normal equations. In the special case of least squares, $\theta = 2$, the array polynomials of normals $N(X+dX) = 0$ are found in the analogy of $F(X+dX) = G + V(X+dX)$ to (5.3) from (6.1) and (6.2) by

$$\begin{aligned} \underset{n,1}{N(X+dX)} &= \underbrace{\underset{n,m}{F'(X+dX)^T}}\,\underset{m,1}{[F(X+dX)-G]} \\ &= \underbrace{\underset{n,m}{V'(X+dX)^T}}\,\underset{m,1}{V(X+dX)} = \underset{n,1}{0}\,, \\ \underbrace{\underset{n,1}{N(X+dX)}} &= \cdots + 1/12\,\underbrace{\underset{n,m}{F'''dX^{**}2^T}}\,\underbrace{\underset{m,1}{F'''dX^{**}3}} \\ &\quad + 1/6\,\underbrace{\underset{n,m}{F''dX^{**}1^T}}\,\underbrace{\underset{m,1}{F'''dX^{**}3}} + 1/4\,\underbrace{\underset{n,m}{F'''dX^{**}2^T}}\,\underbrace{\underset{m,1}{F''dX^{**}2}} \\ &\quad + 1/6\,\underset{n,m}{F'^T}\,\underbrace{\underset{m,1}{F'''dX^{**}3}} + 1/2\,\underbrace{\underset{n,m}{F''dX^{**}1^T}}\,\underbrace{\underset{m,1}{F''dX^{**}2}} \\ &\quad + 1/2\,\underbrace{\underset{n,m}{F'''dX^{**}2^T}}\,\underbrace{\underset{m,1}{F'dX}} + 1/2\,\underset{n,m}{F'^T}\,\underbrace{\underset{m,1}{F''dX^{**}2}} \\ &\quad + \underbrace{\underset{n,m}{F''dX^{**}1^T}}\,\underbrace{\underset{m,1}{F'dX}} + 1/2\,\underbrace{\underset{n,m}{F'''dX^{**}2^T}}\,\underset{m,1}{L} \\ &\quad + \underset{n,m}{F'^T}\,\underbrace{\underset{m,1}{F'dX}} + \underbrace{\underset{n,m}{F''dX^{**}1^T}}\,\underset{m,1}{L} + \underset{n,m}{F'^T}\,\underset{m,1}{L} = \underset{n,1}{0}\,. \end{aligned} \tag{6.3}$$

The column of constant terms for the difference of the predicted and measured values is denoted by

$$\underset{m,1}{L} = \underset{m,1}{F(X)} - \underset{m,1}{G}, \quad l(i) = f(i) - g(i).$$

Some extended shorthand notation conventions of matrix and tensor calculus are now introduced. The upper transpose ^{T} as a superscript of an array (including a vector and a matrix) exchanges the first and second indices in analogy to matrix calculus. The

upper transpose operator and the subsequent matrix or array multiplication can be replaced by a shorthand contraction (summation) operator, say, *, over the first array index i of the observables in (6.1). We will also introduce the lower or subtranspose operator $_T$ as a shorthand notation to [19, p. 92] for exchanging the first and last index of an array. These two new operators are now applied to the last line of (6.3):

(6.4)

$$\left[\underset{n,n}{\underset{n,m}{F'^T}\ \underset{m,n}{F'}} + \underbrace{(\underset{1,m}{L^T}\ \underset{m,n,n}{F''})_T}_{n,n}\right]\underset{n,1}{dX} = -\underset{n,m}{F'^T}\ \underset{m,1}{L} \Leftrightarrow \left(\underbrace{\underset{n,n}{F'^*F'}}_{n,n} + \underbrace{\underset{1,n,n}{F''^*L_T}}_{n,n}\right)\underset{n,1}{dX} = -\underbrace{\underset{n,1}{F'^*L}}_{n,1},$$

$$\sum_{k=1}^{n}\left\{\sum_{i=1}^{m} f'(i,j)f'(i,k) + \sum_{i=1}^{m} l(i)f''(i,j,k)\right\}dx(k) = -\sum_{i=1}^{m} f'(i,j)l(i), \quad j = 1,2,\ldots,n.$$

The lower transpose operator of (6.4) converts the 1xnxn array $L^TF'' = L^*F'' = F''^*L$ into the symmetric nxn matrix. The order of operators is analogous to the scalar calculus, powers (**), and upper transpose first, followed by the (matrix, *, or scalar) multiplications, lower transpose, and additions. Note the rules of regular matrix calculus in developing the matrix by vector product (6.3) from (6.1) and (6.2) using the * operator in place of the upper transpose. For instance, the matrix by vector product $F''dX^{**}1^*F'dX$ is different from $F'^*F''dX^{**}2$. In the special scalar case of one unknown, the terms with equal powers of dx can be combined in the fashion of [26].

7. General solution of nonlinear systems. The N–R normals (6.4) of least squares estimation are recovered by truncating the normals polynomial (6.3) after the constant and linear terms in vector dX. We are going to further expand these normals and introduce new ways for their solution using

1. the analytical multigrid technique of nonlinear array algebra,
2. the nonlinear direct solution (one hyper iteration) techniques of loop inverses and Q-surface,
3. the arbitrary power θ of the minimized absolute values of residuals,
4. the global expansion of LSM and other local nonlinear systems to exploit the fast multilinear array algebra.

The analytical multigrid technique has evolved from the global LSM expansion and has been also called the multiple initial value constrained (MIVC) solution of nonlinear array algebra [26]. The detailed derivations are too lengthy for this introductory paper, even with the new compact array notations. A general outline of the analytical multigrid method is given, and an example of the unweighted least squares solution is shown in the next section. It expands the N–G and N–R techniques such that the products among all terms of (6.1) are required already in the linearized normals.

The explicit use of the high order tensor partials of the nonlinear array polynomials in the multigrid solution improves the pull-in range and convergence rate such that a *direct* (one-iteration) solution of nonlinear systems of equations is becoming feasible. Similar concepts for utilizing the high order tensor partials have been explored in terms of mathematical geodesy by Blaha in [5]. They are related to the earlier mentioned (partitioned) loop inverse technique of parameter exchanges among the model domain of X and the observable space domain of F_0 through the direct estimators of F_0 and V such that the array polynomials represent either the traditional model domain functions $F(X)$ or their space domain interpolation functions

$K(F_0)$. The new array operators and principles used in the multigrid derivation of the combined solution of multiple array polynomials will be applied to the nonlinear Lm-inverse solution in part 2 of this paper.

7.1. Multigrid derivation of nonlinear array polynomials. The normal polynomial of (6.3) can be interpreted as "decentered," where dX is a known offset from some central initial values $X = X^0$. One practical application of the decentered polynomials avoids the re-evaluation of the vector $F(X)$, matrix F', 3D array F'', 4D array F''', etc. and their products in the search of the minimum residual solution in the neighborhood of X^0. The normals N of (6.3) and their partials N', N'', N''', ... are analytically evaluated or updated at a known offset dX by using the continuity of the local polynomials. Some interesting new solution techniques of nonlinear systems are thereby uncovered, as outlined next.

The decentered normals of (6.3) are evaluated at a discrete grid of $(2q+1)^n$ initial values within the expected "uncertainty basket" covering the range of $X^0 - dX, X^0 + dX$. Each parameter element is perturbed within its search limits into $2q+1$ (evenly distributed) locations of spacing dX/q. They form an n-dimensional grid of $(2q+1)^n$ elements of initial values.

The normal polynomials $N(X^0 - e \cdot dX/q), N(X^0 + e \cdot dX/q)$ are formed in symmetric pairs for $e = 1, 2, \ldots, q$ by replacing the "central" ($e = 0$) values F, F', F'', F''', ... of (6.3) by the decentered values at these locations. The weight of each pair can be decreased toward the edges of the uncertainty basket. The resulting two decentered normals are required to have the same solution at the unknown point $X^0 + ddX$. Thus, in the example of $e = q$, $N(dX) = 0$ with the shift $-dX + ddX$ from $X^0 + dX$, and $N(-dX) = 0$ with the shift $dX + ddX$ from $X^0 - dX$. The sum of the resulting two Taylor series of normal equations has to equal zero in

$$\begin{aligned} &N(-dX) + N'(-dX)(dX + ddX) + 1/2N''(-dX)(dX + ddX)^{**}2 \\ &\quad + 1/6N'''(-dX)()^{**}3 + \cdots + N(dX) + N'(dX)(-dX + ddX) \\ &\quad + 1/2N''(dX)(-dX + ddX)^{**}2 + 1/6N'''(dX)()^{**}3 + \cdots = 0. \end{aligned} \tag{7.1}$$

The derivation of this sum is straightforward using the extended matrix and tensor notations of (6.3) but requires a careful use of the known rules of matrix calculus. A useful corollary of these rules is the analogy of the array powers of vector sums $dX + ddX$ and differences $-dX + ddX$ with the scalar polynomials such that, for example,

$$\begin{aligned} {}_{m,n,n}F'' \underbrace{(-dX + ddX)}_{n,1}{}^{**}2 &= \underbrace{F''dX^{**}2}_{m,1} - 2\underbrace{F''dX^{**}1}_{m,n}\,\underset{n,1}{ddX} + \underbrace{F''ddX^{**}2}_{m,1}, \\ {}_{m,n,n,n}F''' \underbrace{(ddX - dX)}_{n,1}{}^{**}3 &= \underbrace{F'''ddX^{**}3}_{m,1} - 3\underbrace{F'''ddX^{**}2}_{m,n}\underset{n,1}{dX} \\ &\quad + 3\underbrace{F'''ddX^{**}1}_{m,n,n}\underset{n,1}{dX}{}^{**}2 - \underbrace{F'''dX^{**}3}_{m,1}. \end{aligned} \tag{7.2}$$

The constant terms of the sum in (7.1) are collected into

$$
\begin{aligned}
&[N(dX)+N(-dX)] - [N'(dX)-N'(-dX)]dX \\
&\qquad + 1/2[N''(dX)+N''(-dX)]dX{**}2 \\
&\qquad - 1/6[N'''(dX)-N'''(-dX)]dX{**}3 \\
&\qquad + 1/24[N^{IV}(dX)+N^{IV}(-dX)]dX{**}4 + 0dX{**}5.
\end{aligned} \tag{7.3}
$$

Because of the symmetry of the pair of the initial values, many terms of the central values F', F'', F''', ... in the sums and differences inside the brackets cancel out, while the others are doubled. A similar expression is found for the symmetric normal matrix of the linear term ddX by applying the rules of (7.2) in (7.1):

$$
\begin{aligned}
N'(dX) &+N'(-dX) - [N''(dX)-N''(-dX)]dX{**}1 \\
&\quad + 1/2[N'''(dX)+N'''(dX)]dX{**}2 \\
&\quad - 1/6[N^{IV}(dX)-N^{IV}(-dX)]dX{**}3 \\
&\quad + 1/24[N^{V}(dX)+N^{V}(-dX)]dX{**}4.
\end{aligned} \tag{7.4}
$$

An example is shown to express the 4th term in (7.4) as a function of the partials of $F(X)$ versus $N(X)$:

$$
\begin{aligned}
&- 1/6[N^{IV}(dX)-N^{IV}(-dX)]dX{**}3 \\
&\quad = -1/6\{4[(F''+F'''dX{**}1)dX{**}1{*}F'''dX{**}2 \\
&\qquad\qquad - (F''-F'''dX{**}1)dX{**}1{*}F'''dX{**}2] \\
&\qquad\quad + 6[F'''dX{**}2{*}(F''+F'''dX{**}1)dX{**}1 \\
&\qquad\qquad - F'''dX{**}2{*}(F''-F'''dX{**}1)dX{**}1]\} \\
&\quad = -2\underbrace{F'''dX{**}2}_{m,n} \,{*}\, \underbrace{F'''dX{**}2}_{m,n} = -2\underbrace{ddF'^{T}ddF'}_{n,n}.
\end{aligned}
$$

Similar expressions to the high order terms of parameters ddX are found in (7.1) at each pair of initial values. The contributions of all pairs are combined, and all terms are scaled with the scalar multiplier of the N–R term at the central initial value of $e = 0$. The resulting multigrid normal equations (7.1) recover (6.3) with the difference that dX is replaced by vector ddX. Thus exactly the same (combined) system of nonlinear normals is achieved from the entire basket of initial values as from using only the single central set X^0 of initial values. This cancellation of terms containing the known vector dX in (7.1) takes place only if *all* partials (up to the 5th order N^V, when $F(X+dX)$ is truncated after F''') are included. New types of the analytical multigrid normal equations are found by ignoring the effect of the high order terms. One of the first applications truncated the combined normals after N' in (7.3)–(7.4) in the fashion of the N–G and N–R techniques. This results in the following "superiteration" normals of [26], where vector dX is given the latest

confidence limits of the parametric model

$$
\begin{aligned}
&\{a[1/6(\underbrace{F'''dX^{**}3}_{m,1} * \underbrace{F'''dX^{**}1}_{m,n,n})_T + 1/4\,\underbrace{F'''dX^{**}2}_{m,n} * \underbrace{F'''dX^{**}2}_{m,n}] \\
&\quad + b[1/2\,\underbrace{F'}_{m,n} * \underbrace{F'''dX^{**}2}_{m,n} + 1/2\,\underbrace{F'''dX^{**}2}_{m,n} * \underbrace{F'}_{m,n} + 1/2(\underbrace{F''dX^{**}2}_{m,1} * \underbrace{F''}_{m,n,n})_T \\
&\qquad + \underbrace{(\underbrace{F'dX}_{m,1} * \underbrace{F'''dX^{**}1}_{m,n,n})_T}_{n,n} + \underbrace{\underbrace{F''dX^{**}1}_{m,n} * \underbrace{F''dX^{**}1}_{m,n}}_{n,n}] \\
&\qquad + \underbrace{L^T F''_T}_{n,n} + \underbrace{F'^T F'}_{n,n}\} \underset{n,1}{ddX} \\
&= -\underbrace{F'^T L}_{n,1} + 3a[1/6\,\underbrace{F''dX^{**}1}_{m,n} * \underbrace{F'''dX^{**}3}_{m,1} + 1/4\,\underbrace{F'''dX^{**}2}_{m,n} * \underbrace{F''dX^{**}2}_{m,1}] \\
&\quad + b[1/2F'^* F''dX^{**}2 + 1/2F'''dX^{**}2^* L + F''dX^{**}1^* F'dX] \\
&\quad - [\{..\}ddX^{**}2 + \{..\}ddX^{**}3 + \{..\}ddX^{**}4 + \{..\}ddX^{**}5 + \cdots].
\end{aligned}
\tag{7.5}
$$

Recall from (6.4) that the contraction operator * of the first array index could be replaced by the upper array transpose in the fashion of line 4 in (7.5). The scalars a, b depend on the value of q and the weighting of the pairwise normal contributions. As a special case of $dX = 0$ (a single initial value of X^0 very close to the true solution), the linear N–R normal terms of line 4 are recovered. The fourth order partials of $F(X)$ are needed in highly nonlinear systems, increasing the number of the shown constant and linear terms of parameters ddX.

Part 2 of this paper will outline a nonlinear expansion of the linear array relaxation [27], where the nonlinear high order array polynomials are evaluated using the solution ddX^0 of the shown linear system. This sum is shifted to the right-hand side as shown in the last line of (7.5) and multiplied with the linear inverse matrix, resulting in a refined correction vector $dddX$. The shown constant and linear terms already improved the pull-in range and speed of convergence in the industrial applications that have been feeding the evolution of array algebra theories.

8. Summary and continuation to part 2. The fast multilinear matrix, tensor, and array operators of sections 1 and 4 are applicable in the nonlinear basis transforms of sections 3 and 4. The solution of the nonlinear array polynomials of sections 5–7 can then be formulated in the estimable space domain and analytically transformed into the general solution of the original modeling parameters. Practical ways are found to avoid an explicit computation of the high order partials. This leads to a fast closed form direct solution of rank-deficient and ill-conditioned nonlinear systems using one hyper iteration of few internal solution steps. The initial linear N–G step requires 2–3 nonlinear correction steps to provide the rigorous "direct" nonlinear solutions both in the parametric and space domains.

REFERENCES

[1] T. M. Apostol, *Mathematical Analysis: A Modern Approach to Advanced Calculus*, Addison–Wesley, Reading, MA, 1957.

[2] W. Baarda, *S-Transformations and Criterion Matrices*, Netherlands Geodetic Commission 5, Delft University of Technology, Delft, The Netherlands, 1973.

[3] A. Bjerhammar, *Rectangular reciprocal matrices with special reference to geodetic calculations*, Bull. Géodésique, 1951, pp. 188–220.
[4] G. Blaha, *A few basic principles and techniques of array algebra*, Bull. Géodésique, 51 (1977), pp. 177–202.
[5] G. Blaha, *Non-iterative approach to nonlinear least-squares adjustment*, Manuscripta Geodaetica, 19 (1994), pp. 199–212.
[6] H. Ebner and P. Reiss, *Height interpolation by the method of finite elements*, in Proceedings of the DTM Symposium, (St. Louis, MO, 1978), ASPRS, VA, 1978, pp. 241–254.
[7] J. Cooley and J. Tukey, *An algorithm for machine computation of complex Fourier series*, Math. Comp., 19 (1965), pp. 297–301.
[8] L. De Lathauwer, B. De Moor, and J. Vandewille, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.
[9] D. W. Fausett and C. T. Fulton, *Large least squares problems involving Kronecker products*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 219–227.
[10] I. J. Good, *The interaction algorithm and practical Fourier series*, J. Roy. Statist. Soc. Ser. B, 20 (1958), pp. 361–372.
[11] E. Grafarend and B. Schaffrin, *Equivalence of Estimable Quantities and Invariants in Geodetic Networks*, Z. Vermessungwesen, 11 (1976).
[12] M. Gu, *New fast algorithms for structured linear least squares problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 244–269.
[13] V. Kratky, *Grid-modified polynomial transformation of satellite imagery*, Remote Sensing Environ, 5 (1976), pp. 67–74.
[14] B. Noble, *Methods for computing the Moore-Penrose generalized inverse and related matters*, Generalized Inverses and Applications, M. Z. Nashed, ed., Academic Press, New York, 1976, pp. 245–301.
[15] R. Penrose, *On best approximation solutions of linear matrix equations*, Proc. Cambridge Philos. Soc., 52 (1956), pp. 17–19.
[16] A. Pope, *Modern trends in adjustment calculus*, in Proceedings of the International Symposium of NA Geodetic Networks, New Brunswick, NJ, 1974.
[17] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Wiley, New York, 1971.
[18] U. A. Rauhala, *Calculus of Matrix Arrays and General Polynomial and Harmonic Interpolation by Least Squares with New Solutions in Photogrammetry and Geodesy*, Fot. Medd.VI:4, Department of Photogrammetry, Royal Institute of Technology, Stockholm, 1972.
[19] U. A. Rauhala, *Array Algebra with Applications in Photogrammetry and Geodesy*, Fot. Medd. VI:6, Department of Photogrammetry, Royal Institute of Technology, Stockholm, 1974.
[20] U. A. Rauhala, *Array algebra as general base of fast transforms*, in Proceedings of the Symposium on Image Processing—Interaction with Photogrammetry and Remote Sensing, Mitteilungen Der Geodaetischen Inst. Der T U Graz, Folge 29, 1977, pp. 175–188.
[21] U. A. Rauhala, *Intuitive derivation of loop inverses and array algebra*, Bull. Géodésique, 53 (1979), pp. 317–342.
[22] U. A. Rauhala, *Introduction to array algebra*, Photogrammetric Engrg. Remote Sensing, 46 (1980), pp. 177–192.
[23] U. A. Rauhala, *Note on general linear estimators and matrix inverses*, Manuscripta Geodaetica, 6 (1981), pp. 375–386; technical report 80-015, Geodetic Services, Inc., Melbourne, FL.
[24] U. A. Rauhala, *Compiler Positioning of Array Algebra Technology*, International Society of Photogrammetry and Remote Sensing, Vol. 26-3/3, Comm. III Symposium, Rovaniemi, 1986, pp. 173–198.
[25] U. A. Rauhala, *General theory of array algebra in nonlinear least squares and robust estimation*, ASPRS Spring Convention, Denver, 1990.
[26] U. A. Rauhala, *Nonlinear Array Algebra in Digital Photogrammetry*, International Society of Photogrammetry and Remote Sensing, Vol. 29 B2 II (1992), pp. 95–102.
[27] U. A. Rauhala, D. Davis, and K. Baker, *Automated DTM validation and progressive sampling algorithm of finite element array relaxation*, Photogrammetric Engrg. Remote Sensing, 4 (1989), pp. 449–465.
[28] P. A. Regalia and S. K. Mitra, *Kronecker products, unitary matrices and signal processing applications*, SIAM Rev., 31 (1989), pp. 586–613.
[29] P. Rozsa and I. Toth, *Eine direkte Methode zur Losung der Poissonschen Differentialgleichung mit Hilfe des 9- Punkte-Verfahrens*, ZAMM Z. Angew. Math. Mech., 50 (1970), pp. 713–720.
[30] R. B. Schnabel and P. D. Frank, *Tensor methods for nonlinear equations*, SIAM J. Numer. Anal., 21 (1984), pp. 815–843.

[31] R. A. Snay, *Applicability of array algebra*, Reviews of Geophysics and Space Physics, 16 (1978), pp. 459–464.
[32] M. Suzuki and K. Shimizu, *Analysis of distributed systems by array algebra*, International J. Systems Sci., 21 (1990), pp. 129–155.
[33] C. Van Loan and N. 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 Publishers, Dordrecht, The Netherlands, 1993, pp. 293–314.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 509–528

ARRAY ALGEBRA EXPANSION OF MATRIX AND TENSOR CALCULUS: PART 2*

URHO A. RAUHALA†

Abstract. Part 1 of this paper summarized some extended matrix and tensor operators of the multilinear array algebra and loop inverse estimation. It also introduced the compact shorthand tensor notations of array polynomials for expressing the Taylor expansion of a nonlinear function and its least squares normal equation. These matrix-like operators will be further expanded using the nonlinear tensor transforms of Blaha's Q-surface technique. They enable an analytical derivation of the direct (one hyper iteration) solution of nonlinear systems. New operators are found such as the nonlinear counterpart of the Lm- or general pseudoinverse and neural net operators. The increased pull-in range and rate of convergence of the new technique are applied in some standard examples of nonlinear equations. Some applications of the nonlinear array algebra involving literally billions of parameters are discussed, including the automated stereo mensuration of three-dimensional digital imaging and mapping systems.

Key words. array algebra, nonlinear Lm-inverse, tensor methods, Q-surface, hyper iteration, array polynomials

AMS subject classifications. 65F20, 65F50, 65H10, 65K05, 65K10

PII. S0895479802406841

1. Derivation of nonlinear loop inverse and Q-surface techniques. An algebraic equality of the linear condition adjustment with a singular parametric adjustment of the Lm-inverse will recover some starting equations of the geometric Q-surface approach of Blaha [1], [2]. The nonlinear Lm-inverse will be introduced by converting the indical tensor notations of the Q-surface approach into the extended matrix and tensor operators of [26]. The derivation is analogous to the linear Lm-inverse solution, (4.2) of [26], and consists of the following five steps.

Step 1. A subset of $p = r(F')$ independently observable (true values) L_1 of nonlinear functions $F_1(X)$ of n parameters X are chosen as the basis of the observable space domain. The nonlinear observables L_1 serve the same role as parameters F_0 in the linear Lm-inverse derivation of [26]. This parameter exchange more or less "solves the problem" by converting the two sets of nonlinear equations for the p and $m-p$ observed values L_{1obs} and L_{2obs},

$$\begin{aligned} F_1(X) &= \underset{p,1}{L_{1obs}} + V_1, \\ F_2(X) &= \underset{m-p,1}{L_{2obs}} + V_2, \end{aligned} \tag{1.1}$$

into one set of linear and another set of "less nonlinear" equations

$$\begin{aligned} I \;\; L_1 &= L_{1obs} + V_1, \;\; \underset{p,p}{I} = \text{unit matrix}, \\ K_2(L_1) &= F_2(F_1^{-1}(L_1)) = L_{2obs} + V_2. \end{aligned} \tag{1.2}$$

Here $F_1^{-1}(L_1)$ denotes X through the (known or approximate) inverse nonlinear function. Functions $L_2 = K_2(L_1)$ are nonlinear in parameters L_1. They can be interpreted

*Received by the editors December 5, 2001; accepted for publication (in revised form) by A. H. Sayed April 19, 2002; published electronically November 6, 2002.
http://www.siam.org/journals/simax/24-2/40684.html
†BAE SYSTEMS, P.O. Box 509008, San Diego, CA 92150-9008 (urho.rauhala@baesystems.com).

as the Monge form of a surface or as nonlinear Lagrange interpolation. Their partials K_2', K_2'', K_2''' are nonlinear functions of the partials F_2', F_2'', F_2''' of the $m-p$ redundant observables $L_2 = F_2(X)$ and partials F_1', F_1'', F_1''' of the p chosen basis functions in $L_1 = F_1(X)$, as shown by Blaha [1], [2]. In our example of a triangle with all $m = 3$ angles measured, parameters X represent the $n = 6$ coordinates of the three triangle points, and L_1 is a column vector of $p = 2$ parameters of the unknown true values of two angles [16], [20]. Vector L_2 contains the unknown $m - p = 1$ true value of the third angle, and column vectors V contain the total of $m = 3$ unknown residuals in V_1 and V_2 to be minimized. Ideally, the initial values of the parametric adjustment are computed as follows using the known forward and inverse parameter transforms.

Initial values of the n modeling parameters X are made consistent with the initial values of the chosen set L_1 of p space domain parameters. The process starts by arbitrarily fixing two of the triangle points, say, at coordinates 0,0 and 0,1. The observed L_{1obs} are used as the initial values of parameters L_1, and the initial (two adjustable) coordinates of the third point in X are intersected using the two angles in L_{1obs}. The nonlinear inverse mapping is not often known, so it has to be approximated using the Taylor series of the forward mapping. Since the forward Taylor series requires some crude initial values of parameters X, we are going to address the general problem, where the initial values X^0 are selected first and projected to the observed space by $L_1^0 = F_1(X^0)$. The problem consists of the parametric adjustment of vector L_1 and its (consistent) inverse transform to the estimators of parameter vector X. It is often useful to apply the consistent nonlinear solution of X from the first subset of (1.1) to get the initial values to the "pull-in range" of the observed values L_{1obs}. The resulting initial values X^0 are then used in computations for the partial derivatives of both sets of observables. We are following the course of [1], [2] and [22], [23] by imposing no restrictions on the high order partials in the derivation stage of the nonlinear loop inverse solution. The "fast" rules in the fashion of the (restricted) tensor methods [28], [3], and [7] will be recovered in the combined steps 1–5 after the derivation of the rigorous solution.

Step 2. In the fashion of the linear Lm-inverse, the least squares adjustment is performed in the space domain. The nonlinear parameter transform from X to L_1 cures many problems such as ill-conditioning (singularity) or range and rate of convergence in the sequential (horizontal or Kalman-type of partitioned) adjustment. The simple starting case selects the $p = r(F')$ basis functions L_1 among a suitable subset of the m observed functions $L = F(X)$. The parameter exchange among the unknown corrections dL_1 and dX using the notations of *array polynomials* in [26],

$$
\begin{aligned}
\underset{p,1}{dL_1} &= \underset{p,n}{F_1'}\,\underset{n,1}{dX} + 1/2\,\underbrace{\underset{p,n,n}{F_1''}\,\underset{n,1}{dX}^{**}2}_{p,1} + 1/6\,\underbrace{\underset{p,n,n,n}{F_1'''}\,\underset{n,1}{dX}^{**}3}_{p,1} + \cdots \\
&= \sum_{j=1}^{n} f_1'(i,j)dx(j) + 1/2 \sum_{j=1}^{n}\sum_{k=1}^{n} f_1''(i,j,k)dx(j)dx(k) \\
&\quad + 1/6 \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} f_1'''(i,j,k,l)dx(j)dx(k)dx(l) + \cdots ,
\end{aligned}
\tag{1.3}
$$

results in the following linear and nonlinear partitioned systems of equations for the

Taylor series of (1.2):

$$
\begin{aligned}
&\underset{p,p}{I}\quad \underset{p,1}{dL_1} = -\underset{p,1}{dL_{1obs}} + \underset{p,1}{V_1},\\
&\underset{m-p,p}{K_2{}'}\ \underset{p,1}{dL_1} + 1/2\ \underset{m-p,p,p}{K_2{}''}\ dL_1{}^{**}2 + 1/6\ \underset{m-p,p,p,p}{K_2{}'''}\ dL_1{}^{**}3 + \cdots\\
&\qquad = -\underset{m-p,1}{dL_{2obs}} + \underset{m-p,1}{V_2}\ .
\end{aligned}
\tag{1.4}
$$

Corrections dL_1 are the unknowns to be estimated using the discrepancies dL_{1obs} and dL_{2obs} (vectors L_{1obs} and L_{2obs} subtracted from the predicted $L_1^0 = F_1(X^0)$ and $L_2^0 = F_2(X^0)$, respectively). The partials of the indical tensor notations in [1], [2] are converted here into the extended matrix and shorthand tensor notations of [26]:

$$
\underset{m-p,p}{K_2{}'} = \underset{m-p,n}{F_2{}'}\ \underset{n,p}{F_1{}'^m},\ \underset{n,p}{F_1{}'^m} = \underset{n,p}{F_1{}'^T}\left(\underset{p,p}{F_1{}'F_1{}'^T}\right)^{-1},
\tag{1.5}
$$

$$
\begin{aligned}
&\underset{m-p,p,p}{K_2{}''} = \left(\underset{m-p,n,n}{F_2{}''} - \underset{m-p,p}{K_2{}'}\ \underset{p,n,n}{F_1{}''}\right)\underset{n,p}{F_1{}'^m}{}^{**}2 = \overbrace{\left(\underset{m-p,n,n}{F_2{}''} - \underset{m-p,n,n}{K_2{}'F_1{}''}\right)}^{F_1{}'^m}\underset{n,p}{F_1{}'^m},\\
&\underset{p,p,p}{K_1{}''} = \underset{p,n,n}{F_1{}''}\ \overset{F_1{}'^m}{\underset{n,p}{F_1{}'^m}}{}^{**}2 = \underset{p,n,n}{F_1{}''}\ \underset{n,p}{F_1{}'^m},\\
&K_2{}''dL_1{}^{**}1 = \underset{m-p,p}{dK_2{}'} = (\underbrace{F_2{}'' - K_2{}'F_1{}''}_{m-p,n,n})(\underbrace{F_1{}'^m dL_1}_{n,1})^{**}1\ \underset{n,p}{F_1{}'^m},
\end{aligned}
\tag{1.6}
$$

$$
\begin{aligned}
\underbrace{K_2{}'''dL_1{}^{**}2}_{m-p,p} &= (F_2{}''' - K_2{}'F_1{}''')F_1{}'^m{}^{**}3\, dL_1{}^{**}2\\
&\quad - 2K_2{}''dL_1{}^{**}1\, K_1{}''dL_1{}^{**}1 - K_2{}''(K_1{}''dL_1{}^{**}2)^{**}1\\
&= \left\{\left[\underset{m-p,n,n,n}{F_2{}'''} - \underset{m-p,p}{K_2{}'}\ \underset{p,n,n,n}{F_1{}'''}\right](\underbrace{F_1'^m dL_1}_{n,1})^{**}2 - \underset{m-p,n}{\Delta_2{}'}\right\}\underset{n,p}{F_1{}'^m}.
\end{aligned}
\tag{1.7}
$$

A shorthand notation $F_1{}'^{m**}k$ is used to k repeat matrix multiplications of an array using the same matrix $F_1{}'^m$, in analogy to the *exponential vector multiplications* $dX^{**}k$ or repeat contractions of arrays F'', F''' in (1.3). Starting from the last index, the operator $F_1{}'^{m**}k$ contracts the k last array indices by the first matrix index and replaces them by the second matrix index. The matrix by array premultiplications $K_2{}'F_1{}''$ and $K_2{}'F_1{}'''$ interpolate or predict the three-dimensional (3D) and four-dimensional (4D) partials $F_1{}''$, $F_1{}'''$ into the observed locations of L_2. The predicted partials are then subtracted from the actual partials $F_2{}''$ and $F_2{}'''$ of L_2. If the fourth order model domain partials are zero or constant, then the third order partials $F_1{}'''$ and $F_2{}'''$ cancel out in (1.7) such that the space domain partials $K_2{}'''$ of dL_{2obs} with respect to parameters dL_1 are only functions of the second and first order partials $F_2{}''$, $F_1{}''$ and $F_2{}'$, $F_1{}'$. A matrix above a 3D array in (1.6) denotes the "backside" array multiplication; see (1.5) and (4.5) of [26].

The third order partials K_2''' consist of four tensors as shown in [1], [2]. They are rarely used as such but are contracted with the vector powers of dL_1 as in (1.7). The product $K_2'''dL_1{}^{**}2 = ddK_2'$ in the first line of (1.7) combines the three permutation terms of K_2''' involving K_2'' into two terms as follows. The last index of K_2'' is contracted by vector dL_1 to produce the matrix dK_2' for a multiplication of matrix $dK_1' = K_1''dL_1{}^{**}1$. The second term is found by contracting the last index of the 3D array K_2'' with vector $K_1''dL_1{}^{**}2$. These terms involving K_2'', K_1'' are combined into the matrix product $\Delta_2'F_1'^m$ in the third line of (1.7).

Step 3. The least squares solution of the hybrid linear and nonlinear systems in (1.4) is started in the Q-surface approach by defining the nonlinear "gap" function at the point of expansion L_1^0 as follows. The normal equation gap is inverse multiplied by the Newton–Gauss (N–G) gradient matrix, resulting in the nonlinear gap function of the Q-surface

$$dL_1 = -(I + K_2'^T K_2')^{-1}(dL_{1obs} + K_2'^T dL_{2obs}), K_2' = F_2'F_1'^m = -[1 \quad 1]. \tag{1.8}$$

This nonlinear gap function $dL_1(dL_1)$ is equivalent to the Q-surface gap of Blaha's geometric interpretation in [1], [2]. Blaha's gap function is equivalent to updating an initial estimate $dL_1^0 = -dL_{1obs}$ of the first (linear) set of equations in (1.4) with the linear part of the redundant observations by [16, pp. 71–72] in the fashion of Kalman [13]:

$$dL_1 = -dL_{1obs} - K_2'^T(I + K_2'K_2'^T)^{-1}w, \quad w = dL_{2obs} - K_2'dL_{1obs}. \tag{1.9}$$

Note that the nonlinear functions in parameters dL_1 of (1.8) and (1.9) are identical (as the notations suggest). The direct estimation of residual vectors V_1 and V_2 of the linear condition adjustment (at the point of expansion L_1^0) becomes apparent by inserting the equivalent gap estimators $d\hat{L}_1$ of (1.8)–(1.9) into (1.4). Ignoring the second and higher order partials of (1.4), we have

$$\begin{aligned} \underset{m,1}{\hat{V}} &= \underset{m,m-p}{B^m}\ \underset{m-p,1}{w} = B^T(BB^T)^{-1}w, \\ \underset{m-p,m}{B} &= \left[-\underset{m-p,p}{K_2'}, \underset{m-p,m-p}{I} \right] = [1 \quad 1 \quad 1]. \end{aligned} \tag{1.10}$$

Thus each observed angle is corrected by $(BB^T)^{-1} = 1/3$ of the misclosure w (the deviation of the sum of the observed angles from 180 degrees). The corrected angles are projected into the adjusted parameters $\hat{X}$ (two coordinates of point 3) by the known nonlinear inverse model of intersection from the known datum points 1 and 2 using the adjusted angles. In the problems of few parameters with the known nonlinear forward and inverse functions among X and L_1, the initial values of the gap in (1.8)–(1.9) at the point of expansion $L_1^0 = L_{1obs}$ (with $dL_{1obs} = 0$) may already converge in one iteration. This N–G iteration in terms of the space domain parameters dL_1 merely initiates the rigorous solution of the nonlinear Q-surface and loop inverse techniques.

Step 4. Blaha in [1], [2] maps the Q-surface gap of (1.8)–(1.9) onto the final space domain parameters dL_1 using the Taylor series of inverse mapping. This was achieved by explicit tensor derivations of the first, second, and higher order partials of the known Q-surface gap of (1.9) with respect to the unknown space domain parameters dL_1. The resulting Taylor expansion was then reversed (using closed form inverse

tensor operators) to express the unknowns dL_1 as the inverse Taylor series of the observed gap. The nonlinear Lm-inverse solution will replace the forward Taylor series by the decentered array polynomials, (6.3) of [26], with respect to parameter vector dL_1 of (1.4). No explicit partials of parameters dL_1 with respect to the nonlinear gap function of normal equations are required as they already have been used (up to the fifth order) in the array polynomial of the normals.

The initial Q-surface gap dL_1^0 of (1.8) at the point of expansion L_1^0 ignores the high order terms in (1.4), so it usually does not provide the final estimate of the desired parameters dL_1. The task is to bridge the known gap dL_1^0 of (1.8), or, rather, the corresponding normals gap ("normal residuals" evaluated at the point L_1^0), by the linear and high order terms of the unknown parameters dL_1, ideally in one hyper iteration. The inverse mapping or estimation of dL_1 from the measured gap is facilitated by the fact that the parameter transformation from X to L_1 made the design matrix of the linear (1.4) partition into unity. In the fashion of the linear Lm-inverse estimation, the parametric adjustment from (1.4) is well conditioned. This is true even in the following four generalization steps:

- Singular or rank-deficient cases can be handled by transforming the n model domain parameters X into $p = r(F')$ space domain parameters L_1 as in the presented triangle example. It is also possible to choose the number p of the basis functions L_1 such that $p = r(F_1{}') < r(F')$, expanding the estimation theory of nonlinear functions and the "constrained nonlinear pseudoinverse" in analogy to the linear loop inverse estimators of (4.2)–(4.3) of [26]. For instance, if also one of the sides of the triangle is measured, the rank of the 4x6 matrix F' is three. The 2x6 (angular problem) matrix $F_1{}'$ of our example still produces the unbiased estimators of the basis functions L_1 in the "angular subspace" of the entire estimable space, as explained in a more detail in [18], [19], [20].
- It is not necessary to coincide the selected space domain parameters with the first p independent measured functions in L. Any p independent basis functions $L_0 = F_0(X)$ suffice, often forming a regular grid or profile of observables. Their estimators provide direct local interpolations $K(L_0) = F(X)$ by replacing the original modeling parameters X also in the user stage of the problem. As a matter of interest, some "pathological" problems cannot even be expressed in terms of "physical" modeling parameters X. The partials of a generic or empirical space domain model $K(L_0) = F(X)$ of undefined parameters X and undefined function $F(X)$ have to be approximated from the noisy measured data in the space domain with blunders and other violations of the Gauss–Markov model $E(V) = 0$. The decentered normals, (6.3) of [26], then already represent some empirical space domain parameters dL_1 of a nonlinear adjustment model (1.4) with some empirical approximations of the partials.
- The estimation theory of nonlinear array algebra expands the least squares to nonlinear "robust" estimation using an arbitrary power θ of the minimized residuals [22], [23].
- A solution dL_1 of the inverse mapping from the known normals gap is found as follows. It is then combined with the parameter transform into the nonlinear estimators of X from L_{obs} in analogy to the linear Lm-inverse. The

nonlinear Lm-inverse becomes applicable in the direct adjustment of the original modeling parameters using new operators to serve the role of "nonlinear matrix and tensor inverses."

The decentered normals of (1.4)–(1.7) vs. (6.1)–(6.2) of the model domain parameters dX of [26], in terms of the space domain parameters dL_1, form the following five-degree array polynomial $N(L_1^0 + dL_1) = 0$, which must be satisfied to minimize the nonlinear least squares object function $V_1^T V_1 + V_2^T V_2$ of the parameter vector dL_1:

(1.11)
$$\begin{aligned} N(L_1^0 + dL_1) &= dL_{1obs} + IdL_1 + (K_2' + K_2''dL_1\,{}^{**}1 + 1/2K_2'''dL_1\,{}^{**}2)^T \\ &\quad \cdot (dL_{2obs} + K_2'dL_1 + 1/2K_2''dL_1\,{}^{**}2 + 1/6K_2'''dL_1\,{}^{**}3) \\ &= dL_{1obs} + K_2{}'^T dL_{2obs} + (I + K_2{}'^T K_2' + dL_{2obs}^T K_2{}''_T)dL_1 \\ &\quad + 1/2K_2{}'^T K_2''dL_1\,{}^{**}2 + K_2''dL_1\,{}^{**}1^T K_2'dL_1 \\ &\quad + 1/2K_2'''dL_1\,{}^{**}2^T dL_{2obs} + 1/6K_2{}'^T K_2'''dL_1\,{}^{**}3 \\ &\quad + 1/2K_2''dL_1\,{}^{**}1^T K_2''dL_1\,{}^{**}2 + 1/2K_2'''dL_1\,{}^{**}2^T K_1'dL_1 \\ &\quad + 1/6K_2''dL_1\,{}^{**}1^T K_2'''dL_1\,{}^{**}3 + 1/4K_2'''dL_1\,{}^{**}2^T K_2''dL_1\,{}^{**}2 \\ &\quad + 1/12K_2'''dL_1\,{}^{**}2^T K_2'''dL_1\,{}^{**}3 = 0. \end{aligned}$$

The solution of (1.11) by "first order array relaxation" (FOAR) recursively inverts the decentered Taylor series in the fashion of the multigrid Newton–Raphson (N–R) normals; see (7.5) of [26]. The traditional iterative techniques are based on the fact that the Taylor series of $K_2{}'^T dL_{2obs}$ at the point L_1^0 with respect to the vector sum $dL_1 = dL_1^0 + ddL_1$ equals

$$K_2'(L_1^0 + dL_1^0)^T dL_{2obs}(L_1^0 + dL_1^0)$$

at point $L_1^0 + dL_1^0$ with respect to the vector ddL_1. We are going to reuse the decentered Taylor series at the original point of expansion, including the inverse matrix N'^{-1} of its first order partials. Several internal refinement steps of both space and model domain parameters are combined into one "hyper iteration," starting with an estimate of dL_1^0.

The FOAR solution of (1.11) shifts all except the linear term to the right-hand side. Both sides are multiplied by the inverse matrix of the linear term. This results in the recursive "inverse Taylor expansion" of (1.11),

(1.12)
$$\begin{aligned} d\hat{L}_1 &= -(I + K_2{}'^T K_2' + dL_{2obs}^T K_2{}''_T)^{-1}\{dL_{1obs} + K_2{}'^T dL_{2obs} \\ &\qquad + K_2{}'^T dF_2 + dK_2{}'^T(K_2'dL_1 + dF_2) \\ &\qquad + 1/2ddK_2{}'^T(dL_{2obs} + K_2'dL_1 + dF_2)\} \\ &= -(I + K_2{}'^T K_2' + dL_{2obs}^T K_2{}''_T)^{-1} F_1{}'^{mT}\{F_1{}'^T dL_{1obs} + F_2{}'^T dL_{2obs} \\ &\qquad + F_2{}'^T dF_2 + dF_2{}'^T(F_2'dX_1 + dF_2) \\ &\qquad + ddF_2{}'^T(dL_{2obs} + F_2'dX_1 + dF_2)\}, \end{aligned}$$

where the contractions with vector dL_1 are converted into contractions with its linear

transform $dX_1 = {F_1}'^m dL_1$ by

(1.13)
$$\begin{aligned}
d{K_2}' &= {K_2}'' dL_1 \,{}^{**}1 = ({F_2}'' - {K_2}'{F_1}'')dX_1 \,{}^{**}1{F_1}'^m = d{F_2}'{F_1}'^m,\\
dd{K_2}' &= {K_2}''' dL_1 \,{}^{**}2 = \{({F_2}''' - {K_2}'{F_1}''')dX_1 \,{}^{**}2 - {\Delta_2}'\}{F_1}'^m = 2dd{F_2}'{F_1}'^m,\\
dF_2 &= 1/2{K_2}'' dL_1 \,{}^{**}2 + 1/6{K_2}''' dL_1 \,{}^{**}3 + \cdots\\
&= 1/2({F_2}'' - {K_2}'{F_1}'')dX_1 \,{}^{**}2 + 1/6({F_2}''' - {K_2}'{F_1}''')dX_1 \,{}^{**}3\\
&\quad - 1/2d{K_2}'{F_1}'' dX_1 \,{}^{**}2 \ldots\\
&= K_2(F_1(X^0) + dL_1) - F_2(X^0) - {F_2}' dX_1.
\end{aligned}$$

The first line of (1.12) provides the initial estimator dL_1^0 used to evaluate the remaining "normals gap" by summing up the effect of the high order terms. This sum is multiplied by the same (already computed) inverse matrix at the point of expansion L_1^0, resulting in a correction term $dd\hat{L}_1$ to the latest estimate of dL_1^0. Like the Q-surface solution, (1.12), including the effect of the fifth order partials of the normals, may converge quickly. This is due to the stabilizing (unit matrix) effect of the well-conditioned linear partition in (1.4) and the reduction of nonlinearity in the differences (actual minus predicted) of the high order partials in the nonlinear partition.

Note that (1.11) can be premultiplied by a constant gradient matrix $(I + {K_2}'^T {K_2}')^{-1}$ such that the first line of the scaled (1.11) becomes the Q-surface gap of (1.8). The inverse expansion (1.12) is still recovered as the constant scaling matrix cancels out in satisfying the constraint ${V'}^T V = 0$ of (1.11). It differs from the Q-surface approach of [1], [2], which is based on the geometric interpretation of (1.9) and intuitive tensor derivation of the forward and reverse Taylor series of dL_1 with respect to the Q-surface gap vs. the scaled normals gap of (1.11). The explicit space domain solution (1.12) of the FOAR and the inverse Taylor expansion of (1.11) is bypassed next in the fashion of the linear Lm-inverse of (4.2) in [26] by combining the linear and nonlinear basis transforms and their inverses into the adjustment. This results in new solutions of nonlinear equations as follows.

Step 5. The adjustment part of the nonlinear estimation is complete when the final estimate,

$$\hat{L}_1 = L_1^0 + d\hat{L}_1, \;\; L_1^0 = F_1(X^0),$$

is found. It needs to be transformed into the original model domain using the nonlinear mapping among the chosen parameter sets X and L_1. A closed form nonlinear Lm-inverse replaces the m-inverse of the design matrix F_1' of linear estimation with the nonlinear inverse function $F_1^{-1}(L_1)$. By denoting the filtering operator of the solution $\hat{L}_1$ from the observed values $F = L_{obs}$ by the superscript "L," the nonlinear counterpart of the linear Lm-inverse solution is

$$\hat{X} = F_1^{-1}(\hat{L}_1) = F_1^{-1}[F^L(F_1^{-1}(L_1))]. \tag{1.14}$$

The nonlinear loop inverse and Q-surface techniques are causing some rethinking of many problems. For example, an idea for neural nets would interpret the "learning process" to establish a nonlinear mapping from one set of basis functions L_1 in the observable space to the model parameters X [23]. The solution of a "new case" with actual measured values L_{obs} starts with a filtering solution of the basis functions $\hat{L}_1$ followed by the "learned mapping" into $\hat{X}$. The nonlinear inverse function $X = F_1^{-1}(L_1)$ is usually not known, so it has to be approximated by the inversion of the

forward Taylor expansion of known $L_1 = F_1(X)$. This makes it possible to bypass an explicit solution of L_1. An equivalent but simplified solution is found directly from the measured values L_{obs} in terms of the original model domain parameters X in analogy to the linear Lm-inverse. Most inverse matrices ${F_1}'^m$ of the space domain partials K', K'', $K'''\ldots$ in (1.4)–(1.12) will cancel out in the fashion of (1.13) such that the model domain partials F', F'', $F'''\ldots$ suffice. This simplifies the practical applications.

Combined steps 1–5. The nonlinear transform from the adjusted $\hat{L}_1 = L_1^0 + d\hat{L}_1$ to the adjusted $\hat{X} = X^0 + d\hat{X} + dd\hat{X}$ is made from the Taylor polynomial (1.3) of forward mapping. It starts by using $d\hat{L}_1 = dL_1^0 + dd\hat{L}_1$ of (1.12) in the linear solution

$$d\hat{X} = {F_1}'^m d\hat{L}_1 = dX_1 + ddX_1 = -N'(X^0)^{-1}(N(X^0) + dN). \tag{1.15}$$

The insertion of (1.12) and (1.13) into (1.15) and the linear Lm-inverse of (4.2) in [26] yield

(1.16)

$$\begin{aligned}
\underbrace{N'\left(X^0\right)}_{n,n}^{-1} &= {F_1}'^m_{n,p}\left(I + {K_2}'^T K_2' + dL^T_{2obs} K_2''_T\right)^{-1}_{p,p} {F_1}'^{mT}_{p,n} \\
&= {F_1}'^T_{n,p}\left[{F_1}'_{p,n}\left({F'}^T F' + dL^T_{2obs}\Delta {F_2}''_T\right)_{n,n} {F_1'}^T_{n,p}\right]^{-1}_{p,n} {F_1}'_{p,n}, \\
\underbrace{N\left(X^0\right)}_{n,1} &= {F'}^T_{n,m}\, dL_{obs\,m,1} = {F_1}'^T_{n,p}\, dL_{1obs\,p,1} + {F_2}'^T_{n,m-p}\, dL_{2obs\,m-p,1}, \quad dX_{1\,n,1} = \underbrace{-N'\left(X^0\right)}_{n,n}^{-1}\underbrace{N\left(X^0\right)}_{n,1}, \\
dN_{n,1} &= {F_2}'^T_{n,m-p}\, dF_{2\,m-p,1} + dF_2'^T_{n,m-p}\left({F_2}'_{m-p,n}\, dX_{1\,n,1} + dF_{2\,m-p,1}\right) \\
&\quad + ddF_2'^T_{n,m-p}\left(dL_{2obs\,m-p,1} + {F_2}'_{m-p,n}\, dX_{1\,n,1} + dF_{2\,m-p,1}\right).
\end{aligned}$$

The high order partials are converted into the matrix and vector corrections of (1.16) by the ** contraction operator in the fashion of (1.3)–(1.13) by inserting the vector $dX_1 = {F_1}'^m dL_1^0$ into one FOAR iteration of (1.12) such that

(1.17)

$$\begin{aligned}
\Delta {F_2}''_{m-p,n,n} &= {F_2}''_{m-p,n,n} - {K_2}'_{m-p,p}\, {F_1}''_{p,n,n}, \quad \Delta {F_2}'''_{m-p,n,n,n} = {F_2}'''_{m-p,n,n,n} - {K_2}'_{m-p,p}\, {F_1}'''_{p,n,n,n}, \\
{dF_2}'_{m-p,n} &= \Delta {F_2}'_{m-p,n,n}\, dX_{1\,n,1}\, {}^{**}1, \quad {dF_1}'_{p,n} = {F_1}''_{p,n,n}\, dX_{1\,n,1}\, {}^{**}1, \\
{ddF_2}'_{m-p,n} &= 1/2\, \Delta {F_2}'''_{m-p,n,n,n}\, dX_{1\,n,1}\, {}^{**}2 - {dF_2}'_{m-p,n}\, {F_1}'^m_{n,p}\, {dF_1}'_{p,n} \\
&\quad -1/2\, \Delta {F_2}''_{m-p,n,n}\left({F_1}'^m_{n,p}\underbrace{{F_1}''dX_1\, {}^{**}2}_{p,1}\right) {}^{**}1, \\
dF_{2\,m-p,1} &= 1/2\underbrace{\Delta {F_2}''dX_1\, {}^{**}2}_{m-p,1} + 1/6\underbrace{\Delta {F_2}'''dX_1\, {}^{**}3}_{m-p,1} - 1/2\, {dF_2}'_{m-p,n}\, {F_1}'^m_{n,p}\underbrace{{F_1}''dX_1\, {}^{**}2}_{p,1}.
\end{aligned}$$

Vector dX_1 serves the same role as the initial linear estimator dL_1^0 in updating the high order terms of the decentered normals (1.12) in the space domain. It can exploit

the linear Lm-inverse in

$$dX_1 = -{F_1}'^m(F'{F_1}'^m)^L dL_{obs}, \quad \underset{m,n}{F'} = \begin{bmatrix} {F_1}' \\ {F_2}' \end{bmatrix}, \ \underset{m,1}{dL_{obs}} = \begin{bmatrix} dL_{1obs} \\ dL_{2obs} \end{bmatrix},$$

by moving the linear N–R term $dL_{2obs}^T \Delta {F_2}_T''$ from matrix N' to vector dN by $d{F_2}'^T dL_{2obs}$. Terms $d{F_2}'$ and $dd{F_2}'$ are then combined to update the normals in (1.15), and the change in their derivative matrix is included by

$$\begin{aligned} &dN = {F_2}'^T d{F_2} + (\Delta {F_2}')^T (dL_{2obs} + {F_2}' dX_1 + d{F_2}) + \Delta N' dX_1, \\ &\Delta {F_2}' = d{F_2}' + dd{F_2}', \ \ \Delta N' \simeq ({F_2}' + \Delta {F_2}')^T ({F_2}' + \Delta {F_2}') - {F_2}'^T {F_2}'. \end{aligned} \tag{1.18}$$

The nonlinear correction in $\hat{X} = X^0 + d\hat{X} + dd\hat{X}$ is after the first FOAR iteration of the parameter transform (1.3)

$$\begin{aligned} dd\hat{X} &= -{F_1}'^m[1/2 {F_1}'' d\hat{X}\,{}^{**}2 + 1/6 {F_1}''' d\hat{X}\,{}^{**}3 + \cdots] \\ &= -{F_1}'^m[F_1(X^0 + d\hat{X}) - F_1(X^0) - {F_1}' d\hat{X}] \\ &= d\hat{X} - {F_1}'^m[F_1(X^0 + d\hat{X}) - F_1(X^0)]. \end{aligned} \tag{1.19}$$

This "remainder rule" of the Taylor series avoids an explicit use of the second and higher order partials to get their combined effect without any restrictions unlike the tensor method of [28], [3], and [7]. This rule is now applied to simplify (1.18). The Taylor approximations involving the significant high order terms $\Delta {F_2}'', \Delta {F_2}''', \ldots$ are removed altogether by the evaluation of the following nonlinear functions and first order derivatives:

(1.20)

$$\begin{aligned} \Delta {F_2}' &= d{F_2}' + dd{F_2}' = \Delta \tilde{F}_2' - 1/2 \Delta_2', \\ d{F_2} &= F_2(X^0 + dX_1) - F_2(X^0) - {K_2}'[F_1(X^0 + dX_1) - F_1(X^0)] - 1/6 \Delta_2' dX_1 \\ &= F_2(X^0 + dX_1) - F_2(X^0) - (F_2' + \Delta \tilde{F}_2'){F_1}'^m[F_1(X^0 + dX_1) - F_1(X^0)] \\ &\quad + \Delta \tilde{F}_2' dX_1, \\ \Delta \tilde{F}_2' &= \Delta {F_2}'' dX_1\,{}^{**}1 + 1/2 \Delta {F_2}''' dX_1\,{}^{**}2 + \cdots \\ &= {F_2}'(X^0 + dX_1) - {K_2}' {F_1}'(X^0 + dX_1). \end{aligned}$$

The remaining effect of the second order partials in ${K_2}''' dL_1\,{}^{**}2$ of (1.7) and (1.13) is collected in (1.20) into

$$\underset{m-p,n}{\Delta_2'} = 2 \underset{m-p,n}{d{F_2}'} \underset{n,p}{{F_1}'^m} \underset{p,n}{d{F_1}'} + \underset{m-p,n,n}{\Delta {F_2}''} (\underbrace{{F_1}'^m {F_1}'' dX_1\,{}^{**}2}_{n,1})\,{}^{**}1. \tag{1.21}$$

The correction $d\hat{X}$ of (1.15) and adjusted $d\hat{L}_1 = {F_1}' d\hat{X}$ now use only the low order partials of function $F(X)$ in

(1.22)

$$d\hat{X} = dX_1 - {F_1}'^T (D^T D)^{-1} {F_1}' [{F_2}'^T d{F_2} + \Delta {F_2}'^T (dL_{2obs} + {F_2}' dX_1 + d{F_2}) + \Delta N' dX_1],$$

where $D = F' {F_1}'^T$; see (4.2) of [26]. The full-rank special case when $p = n$ recovers the N–G matrix

$$N'(X^0)^{-1} = {F_1}'^T (D^T D)^{-1} {F_1}' = (F'^T F')^{-1} = ({F_1}'^T {F_1}' + {F_2}'^T {F_2}')^{-1}. \tag{1.23}$$

In the general case, $L_1 = F_1(X)$ serves only the role of parameter transforms without any observed values L_{1obs} such that $L_{obs} = L_{2obs}$ and $F' = F_2{}'$.

The inverse Taylor expansion of (1.3) and its derivative refines (1.19) beyond the first pass terms of [1], [2]. The second pass terms are included without detailing the advanced tensor derivations (the subject of future proofs and refinements). These inverse expansions were extended to the fifth order partials to find a pattern in exploiting the fast remainder rule (1.19) of unlimited Taylor terms, resulting in the nonlinear direct solution of a "superiteration"

(1.24)
$$\begin{aligned}
X_3 &= X_2 + dd\hat{X} - F_1{}'^m[F_1(X_2) - \hat{L}_1 + F_1{}'(X_2)dd\hat{X}],\\
X_2 &= X_1 + dd\hat{X} = X^0 + d\hat{X} + dd\hat{X},\\
\hat{X} &= \overbrace{X_3 - F_1{}'^{mr}[F_1(X_3) - \hat{L}_1}^{X_4} + F_1(X_4) - \hat{L}_1 + F_1(X_5) - \hat{L}_1 + \cdots (\sim 0)],\\
X_5 &= X_4 - F_1{}'^{mr}[F_1(X_4) - \hat{L}_1],\\
F_1{}'^{mr} &= F_1{}'^m\{F_1{}'(X_2) - [F_1{}'(X_2) - F_1{}']F_1{}'^m[F_1{}'(X_1) - F_1{}']\}F_1{}'^m \simeq F_1{}'(X_3)^m,\\
\hat{L}_1 &= F_1(X^0) + F_1'd\hat{X}.
\end{aligned}$$

The inverse derivative of the Taylor expansion at point X_3 is denoted as the nonlinear mr-inverse of $F_1{}'$. The cancellation of the inverse matrices gets pronounced in the full-rank case of $p = r(F') = r(F_1{}') = n$. The direct solution (1.24) makes the $p = n$ basis function estimates of $F_1(X^0)$ agree with the measured values by

(1.25)
$$\begin{aligned}
X_3 &= X_0 - \widehat{F}_1{}'^{-1}[F_1(X_0) - L_{1obs} + F_1(X_1) - L_{1obs} + F_1(X_2) - L_{1obs}\\
&\qquad + F_1{}'(X_2)(X_2 - X_1)] + X_2 - X_1,\\
X^0 &= X_3 - \widehat{F}_1{}'^{-1r}[F_1(X_3) - L_{1obs} + F_1(X_4) - L_{1obs} + F_1(X_5) - L_{1obs} + \cdots (\sim 0)],\\
\widehat{F}_1{}'^{-1r} &= \widehat{F}_1{}'^{-1}\{F_1{}'(X_2) - [F_1{}'(X_2) - \widehat{F}_1{}']\widehat{F}_1{}'^{-1}[F_1{}'(X_1) - \widehat{F}_1{}']\}\widehat{F}_1{}'^{-1} \simeq F_1{}'(X_3)^{-1}.
\end{aligned}$$

The inverse matrix of $\widehat{F}_1{}'$ at crude initial values X_0 and its nonlinear r-inverse are reused in (1.25) in the inverse expansion of (1.3) to get $F_1(X^0)$ close to the observed values L_{1obs}.

The property of $dL_{1obs} = 0$ (or close to zero) in the full-rank special case of (1.22)–(1.24) at $L_1^0 = L_{1obs}$ illustrates the Q-surface, tensor, and N–G techniques by

(1.26)
$$\begin{aligned}
dX_1 &= F_1{}'^{-1}dL_1^0 = -F_1{}'^{-1}K_2{}'^T(I + K_2{}'K_2{}'^T)^{-1}dL_{2obs}\\
&= -F_1{}'^{-1}(I + K_2{}'^TK_2{}')^{-1}K_2{}'^TdL_{2obs}\\
&= -(F_1{}'^TF_1{}')^{-1}F_2{}'^T[I + F_2{}'(F_1{}'^TF_1{}')^{-1}F_2{}'^T]^{-1}dL_{2obs}\\
&= -(F_1{}'^TF_1{}' + F_2{}'^TF_2{}')^{-1}F_2{}'^TdL_{2obs},\\
d\hat{X} &= -N'^{-1}\{(F_2{}' + \Delta F_2{}')^T[dL_{2obs} + F_2{}'dX_1 + dF_2 + (F_2{}' + \Delta F_2{}')dX_1)]\\
&\qquad - 2F_2{}'^TF_2{}'dX_1\},\\
X_3 &= X_2 + dd\hat{X} - F_1{}'^{-1}[F_1(X_2) - \hat{L}_1 + F_1{}'(X_2)dd\hat{X}],\ X_2 = X^0 + d\hat{X} + dd\hat{X},\\
\hat{X} &= X_3 - F_1{}'(X_3)^{-1}[F_1(X_3) - \hat{L}_1 + F_1(X_4) - \hat{L}_1 + \cdots]\,\text{until}\,F_1(\hat{X}) \Rightarrow \hat{L}_1.
\end{aligned}$$

The first and second lines express dX_1 via the Q-surface gap dL_1^0 (1.9) in the special case of $dL_{1obs} = 0$ using the sequential and regular least squares solutions. Lines 3 and 4 provide the linear transform dX_1 of the gap (1.8) recovering the N–G solution of the modeling parameters with improved (but still biased) linear projections $F_1' dX_1 \neq 0$ of the basis functions in the space domain. The best linear unbiased estimator (BLUE) values $d\hat{L}_1$, as the linear projections $F_1{}' d\hat{X}$, are recovered after the nonlinear solution of the normals on lines 5 and 6. The unbiasedly estimable adjusted values at other locations than the chosen basis functions are nonlinear (Monge form) projections from $\hat{L}_1 = F_1(X^0) + F_1{}' d\hat{X}$. The adjusted basis functions $\hat{L}_1$ are converted into model domain estimators $\hat{X}$ on lines 7 and 8 of (1.26) using one superiteration (1.24) near the solution when $dL_{1obs} \simeq 0$. The tensor method of [28] uses the N–G correction dX_1 in a quadratic interpolation from the past point X_0 and present point X^0 by a rank restriction of $F_1{}''$.

The nonlinear Lm-inverse solutions (1.14)–(1.26) are now shown in the general case of infinite Taylor terms with poor initial values and no observed values dL_{1obs}. Thus $F_2{}' = F'$, $dL_{2obs} = dL_{obs}$, and the set $L_1 = F_1(X)$ serves only the role of parameter transforms. The derivation (not repeated in detail) is otherwise the same as in (1.1)–(1.24) but without the partition of dL_{1obs}. The row dimensions of all vectors, matrices, and arrays with subscript "2" change from $m - p$ into m. An expansion of (1.4) beyond the third order partials is required to find the general tensor structure of terms $\Delta F_2{}'$ and dF_2 in (1.20). The general solution must include the effect of an infinite number of Taylor terms in (1.4) and (1.11) and in the inverse Taylor expansion of (1.3) in the fashion of (1.24).

As mentioned before, the general pattern of the high order terms starts emerging when the derivations (using indical tensor notations) are extended to the fifth order partials such that the normals (1.11) and their inverse Taylor expansion consist of the ninth degree array polynomials. Their translation (which cannot be detailed in this paper) into matrix notations results in the following "hyper iteration" of direct solution, where an infinite number of Taylor terms is used in (1.4) and (1.3) to represent (1.2). This direct solution has the property that the inverse expansion of the space domain normals and the inverse transformation to the model domain require no explicit use of the second and higher order partials. They are replaced by the evaluations of the nonlinear functions and first order partials in a sequence of internal steps, such as

$$
\begin{aligned}
\underset{n,1}{dX_1} &= \underset{n,p}{F_1{}'^m}\, \underset{p,1}{dL_1^0} = -\underset{n,m}{F_2{}'^{Lm}}\, \underset{m,1}{dL_{2obs}} = -\underset{n,n}{N'^{-1}}\, \underbrace{\underset{n,m}{F_2{}'^T}\, \underset{m,1}{dL_{2obs}}}_{n,1},\\
\underset{n,n}{N'^{-1}} &= \underset{n,p}{F_1{}'^T} \left(\underset{p,n}{F_1{}'}\, \underset{n,m}{F_2{}'^T}\, \underset{m,n}{F_2{}'}\, \underset{n,p}{F_1{}'^T} \right)^{-1} \underset{p,n}{F_1{}'},\\
X_3 &= X_2 + ddX - F_1{}'^m[F_1(X_2) - \tilde{L}_1 + F_1{}'(X_2)ddX],\\
ddX &= -F_1{}'^m[F_1(X^0 + dX_1) - \tilde{L}_1],\ \ X_2 = X^0 + dX_1 + ddX,\\
\tilde{L}_1 &= F_1(X^0) + F_1{}' dX_1\\
ddX_1 &= F_1{}'^m d\hat{L}_1 = dX_1 - N'^{-1} F_2{}'(X_3)^T[F_2(X_3) - L_{2obs} + F_2{}'(X_3)dX_1],\\
\hat{X} &= \overbrace{X_4 - F_1{}'(X_3)^m[F_1(X_4) - \hat{L}_1}^{X_5} + F_1(X_5) - \hat{L}_1 + \cdots],\ \ X_4 = X_3 + ddX_1,\\
\hat{L}_1 &= \tilde{L}_1 + F_1{}' ddX_1.
\end{aligned}
\tag{1.27}
$$

The initial N–G corrections dX_1 of the linear functions $dX = F_1{}'^m dL_1$ indirectly

use the linear estimator dL_1^0 on line 1 of (1.27). The explicit computation of the space domain gap1 dL_1^0 is avoided by its linear transform into the "raw gap2" $dX^0 = dX_1$ of the nonlinear inverse transform from the space domain to the original model domain. The nonlinear normals of dL_1 in (1.11) are expressed in terms of the parameters $dX = {F_1}'^m dL_1$ and then expanded to unlimited Taylor terms using three steps of the "N–G super iteration" (1.24) on the third line such that

$$F_1(X_3) \Rightarrow \tilde{L}_1 = L_1^0 + dL_1^0 = F_1(X^0) + {F_1}' dX_1, \;\; F_2(X_3) \Rightarrow K_2(L_1^0 + dL_1^0). \tag{1.28}$$

The gap correction ddX_1 of line 6 is found without new matrix inversions or an explicit use of the high order tensors of the inverse Taylor expansion of the normals. It refines the consistent inverse transform from the space domain to the model domain on line 7. The nonlinear projections $F_1(\hat{X})$ should recover the adjusted space domain basis functions $\hat{L}_1 = \tilde{L}_1 + {F_1}' ddX_1$ within a sufficient accuracy. An evaluation of $F(\hat{X})$ at any location should produce the best (minimum variance) nonlinear unbiased estimator of an observable. The solution has converged when the least squares object function ${F_2}'(\hat{X})^T(F_2(\hat{X}) - L_{2obs}) = 0$ is satisfied. The hyper iteration of (1.27) is usually repeated in the process of "system pull-in" to refine the initial values of system parameters where the nonlinear parametric model is dynamically changed in each hyper iteration. A refinement of the system model often increases the number of local parameters. Their initial values are predicted from the global system solution of the previous hyper iteration.

Detailed computational algorithms of the nonlinear Lm-solutions and the related Q-surface technique are beyond the scope of this paper. Note that the terms involving the second order partials in $\Delta {F_2}'$ and dF_2 vanish or are negligible when either the direct solution technique of (1.27) is applied or $\Delta {F_2}''$, ${F_1}''$, or dX_1 get small. This is usually the case when the locations of the space domain basis functions L_1 are chosen to properly cover the observed space L_{obs} and/or when the initial linear N–G correction term dX_1 is made small. As stated before and demonstrated next, a careful search of good initial values (making dX_1 small) can cure many problems in nonlinear estimation.

2. Discussion and some standard examples. Solutions (1.15)–(1.28) resemble a special class of nonlinear problems, where the direct solution becomes linear [14]. Without any Taylor expansion, the (closed) nonlinear normals can be derived as the product of the nonlinear functions $v'(V) = f'(X)$ and $v(X)$ at any "generic" point of observables and then integrated over the entire space [23]. A smart problem analyst can make the linear solutions with ${F_1}'^{-1}$ become "fast" using the recipes of linear array algebra. Their predecessors consist of "Bolz arrays" and other pre-derived solutions used in the era before computers were available or were being pioneered in engineering [6], [8], [10], [11]. The analytical multigrid N–R technique in (7.5) of [26] and its emerging expansion by the integral calculus of nonlinear array algebra is returning to these "smart" and very fast computing methods.

The presented array algebra (and related Q-surface) technique expands the nonlinear estimation theory and its notational tools of matrix and tensor calculus in a fundamental way using the "fast" operators of array algebra and related tensor transforms. As an example, the second order partials of the space domain parameters dL_1,

$$\underset{m-p,p,p}{{K_2}''} = \underset{m-p,n,n}{\Delta {F_2}''} \; \underset{n,p}{{F_1}'^m} \; **2,$$

represent $m-p$ linear solutions of p^2 basis parameters of $K_2{''}$ in about $2(m-p)p^3/6$ versus $[(m-p)n^2]^3/6$ operations of a brute-force solution. For $m-p=p=n=10$ (a full-rank problem with 10 parameters and 20 observations), the "fast" count takes 50,000 times fewer operations. All operators in the solution of dX_1, dX, and ddX are reduced to matrices and vectors by the array contraction operator **. This helps in the derivation and use of these new operators, making them compatible with the shorthand contraction rules of matrix calculus. Note the connection of these "fast" array operators to the "long hand" operators of tensor products. The vectored (column-by-column stacked) solution of $vec(K_2{''})$ could be found by a premultiplication of $vec(\Delta F_2{''})$ with the $(m-p)p^2$ x $(m-p)n^2$ inverse matrix $I \otimes F_1{'}^{mT} \otimes F_1{'}^{mT}$. The above mental derivations of the nonlinear Lm-inverse technique eliminated the explicit nonlinear basis transforms among the space and model domain partials based on the commutative array multiplication rule such that $K_2'' dL_1 \,{}^{**}2 = \Delta F_2'' dX \,{}^{**}2$, where $dX = F_1{'}^{m} dL_1$.

The operation count of one hyper iteration (1.27) is not far from that of three to five N–G iterations, although many more correction steps, each similar to one N–G iteration and restricted tensor correction, are taken by reusing the same inverse matrices. The estimator chain X^0, dX_1, $d\hat{X}$, $\hat{X}$, and $\hat{L}_1 = F_1(\hat{X})$ can converge in one hyper iteration within a wide pull-in range in problems where the traditional techniques require more iterations or may not converge at all in problems of ill-conditioning and poor initial values [1], [2]. Explicit inversions of the derivative matrices between the point of expansion and the final solution can be avoided in (1.27) using the (computationally more expensive) rule of the nonlinear mr-inverse (1.24) to bridge the local minimum or maximum points. The "pathological cases" often require the use of the multigrid N–R technique and the nonlinear robust estimation. The number of hyper iterations does not matter in these problems as long as the best estimators are found for the estimable system parameters and some reasonable estimators are filled for the rest of the parameters at any cost of "automated computations." The cost of analyst-dependent "repair computations" get otherwise prohibitive in automation problems of imaging and information systems involving millions (even billions) of parameters.

Use of the rectangular p x n versus square n x n transform matrix $F_1{'}$ formally expands the nonlinear estimation into ill-conditioned, singular, or rank-deficient systems. In analogy to the linear Lm-inverse, the new nonlinear chain operator expands some foundations of mathematical statistics and estimation theories, requiring new research to serve the wide field of math, engineering, and computer sciences. This field has been application-driven such that the theory has evolved as a byproduct of the industrial array algebra applications. Some special cases of these theories have been used in least squares image matching as they can be locally handled by the scalar polynomials, including the elimination of linear modeling parameters from the nonlinear equations [18], [19], [20], [21], [22], [23]. The general hybrid linear and nonlinear solution is related to the method of "structured nonlinear total least norm" [27]. The hybrid problem is related to the techniques of "self-calibration" in photogrammetry using the function theory of tensor products and loop inverses [16, p. 124]. Before a further expansion of this theory by the global techniques in industrial applications of the imaging and information systems, some examples are discussed to solve the standard nonlinear problems.

Some standard problems of More, Garbow, and Hillstrom [15] are selected from the comparisons of Schnabel and Frank [28], where both the standard technique and the (restricted) tensor method had the most difficulties. The presented nonlinear

loop inverse ideas are now applied. Many of the problems can be solved using the initial value technique of (1.25) before any iterations of the nonlinear process are even started—with an error free result of "mental computing."

Problems of Wood gradient, Powell singular, Rosenbrock, Biggs Exp6, and Freudenstein and Roth (F–R) functions are simplified by the rank partition rule (1.1) of the loop inverse technique. The search of refined initial values (using the given crude initial values) is a major task of the problem analyst before the nonlinear solution algorithms are allowed to start. Most nonlinear solution techniques can converge only to the closest local solution when the initial values are within the "pull-in range" of the partials. The search of good initial values using the linear partition of the system of equations solves some of the selected problems without any nonlinear adjustment as follows.

The full-rank problem of the Wood function has $n = 4$, $m = 6$. Functions v_2, v_4, v_5, and v_6 are linear in all four parameters with $r(F_1{}') = n = 4$. The linear solution X^0 from this subset, to get $dL_{1obs} = 0$ in (1.25), needs no initial values, and the resulting solution also satisfies the two remaining nonlinear equations of v_1 and v_3. No adjustment is required.

The Powell singular function of rank 2 has $n = 4$, $m = 4$. The first two equations provide the linear p x n partition of $p = r(F')$ with $L_{1obs} = 0$ such that $X^0 = 0$. The solution also satisfies the two remaining nonlinear equations such that no iterations are needed.

The Rosenbrock function of $n = m = 2$ has $v_2 = 1 - x_1$, or a subset of X^0 with r elements can be derived from the measured values by a local linear solution. The resulting r initial values overrule the original crude initial values such that x_1^0 =1 (versus −1 of the original crude initial value). The remaining $n - r$ initial values are found from $r(F') - r$ nonlinear equations by substituting the known subset of r parameters into these equations and shifting their effect to the right-hand side. The principle of array relaxation is then applied to reduce the nonlinear effect of other parameters with the crude initial values such that the remaining effect becomes linear. The Rosenbrock function has only the linear part in $v_1 = 10(x_2 - x_1^2)$ for the remaining parameter x_2, so $x_2^0 = (x_1^0)^2 = 1$. An error-free solution is again achieved by the refined initial values X^0 without any iteration.

The Biggs Exp6 problem is discussed here because of a poor performance of the standard and restricted tensor method. It is analogous to a hybrid adjustment of linear and nonlinear systems in the early applications of array algebra leading to the discovery of the global least squares matching (LSM) technique or "global F–R function." In the fashion of the above exploitation of the linear partitions of the system, we continue converting the nonlinear systems into linear ones by proper parameter transforms. These transforms have to be analyzed at the very beginning of the system design by the problem analyst (expert) before the mensuration process and the associated data reduction algorithms are started. The elimination of the linear parameters reduces the size of the nonlinear problem and, in the fashion of the loop inverse technique, converts a large ill-posed hybrid linear and nonlinear problem into a small and "fully or almost linear" full-rank problem in a suitable parameter space.

The Biggs Exp6 function of the structured total least norm problem of [27] for Vandermonde matrices is related to the hybrid linear and nonlinear "global" solution of the LSM technique to be discussed in the section on image matching applications. The detailed solution of hybrid linear and nonlinear systems is beyond the scope of this paper, as it requires some modifications of the "purely nonlinear" loop inverse

solutions. Some findings in the ill-conditioning of the system will be discussed in connection with the F–R function. The early research of the loop inverse technique provides a connection among ill-posed Vandermonde matrices and linear least squares prediction using covariance functions. This connection allows a linear approximation of the Biggs function by a quadratic polynomial as follows.

The linear interpolation or prediction model of [9] can be written for three node locations u_1, u_2, u_3 by specifying the basis functions of vector $a(u)$ in (3.1) of [26] as

$$a(u) = [q_1^{(u-u_1)^2},\ q_2^{(u-u_2)^2},\ q_3^{(u-u_3)^2}],\ \underset{1,1}{f(u)} = \underset{1,n}{a(u)}\ \underset{n,1}{X}\ .$$

The covariance functions q_j, $j = 1, 2, 3$, of each node get the value of 1 when the square of distance (exponent of q_j) of the variable u from any node location is zero. A mental analytical derivation can show that as, $q_j > e^{-1/10} > 0.9$ approach the value of 1, the basis functions of the row vector $k(u) = a(u)A_0^{-1}$ in (3.2) of [26] approach the Lagrange quadratic polynomials [16, p. 118]. The 3×3 transform matrix A_0 is evaluated at the chosen node locations, say, at the first $u_1 = -1$, middle $u_2 = 0$, and last $u_3 = +1$ location of the observed values. The nonlinear "structural" parameters q_j (functions of frequency and damping factor) then have no effect on the interpolation function, so they are not estimable. The linear model domain parameters are not estimable because all elements of matrices A, A_0 consist of the value 1. Matrix $K = AA_0^{-1}$ is still well conditioned, representing Lagrange's interpolations from the true values F_0 at the chosen node locations. Its 3×3 partition at the node locations consists of the unit matrix.

Simulations with $n = 31$ and $m = 37$ in [16, p. 116] solved the Runge problem of polynomials by the Lm-inverse. Node locations of F_0 were coincided with 31 evenly distributed observed values. Having the redundant observations near both ends of the observed space reduced the interpolation error from 10^6 sigma to the order of 1 sigma value. The simulation used the exact limiting case of $q_j = 1$ by augmenting the 31 x 31 leading partition $K_1{}' = I$ by the 6 x 31 matrix $K_2{}'$ of Lagrange's interpolation coefficients. How does this relate to the Biggs or Vandermonde problem? The m x n elements of the Vandermonde matrix A of the linear amplitude and phase parameters approach 1 as q_j approach 1. The Biggs function can now be interpreted as the linear prediction model having the same effect on the observables y_i as a quadratic polynomial. Since the Biggs function has $q_j = 0.90 < 1$ and $m > n$, it is possible to improve the structural values of q_i in the fashion of [27], [16, p. 124], and the "self-calibration" of Brown [5].

We start approaching the main industrial application of the linear and nonlinear array algebra related to the hybrid system of linear and nonlinear equations of automated stereo image mensuration involving literally billions of unknowns. The local nonlinear model of LSM in (5.1) of [26] is expanded with a constant linear term db called "bias" to account for the systematic radiometric differences among the gray values of two match windows. Modeling the local gray values $f(x_i)$ with a quadratic polynomial within the window results in two nonlinear normal equations in parameters db and dx, similar to those of the F–R function. This allows the elimination of the linear parameter from one of the nonlinear equations. Its substitution into the second nonlinear equation produces the "reduced" (purely) nonlinear normal equation.

The elimination of the linear parameter x_1 in the F–R function requires no initial values. The resulting reduced nonlinear equation reads $x_2^3 - 2x_2^2 - 6x_2 - 8 = 0$ with a real root at $x_2 = 4$, which, substituted back into one of the equations, gives the linear solution of $x_1 = 5$. The reader is now urged to solve the modified F–R problem of

LSM with the linear bias term *db*. The reduced nonlinear normal equation becomes linear in the fashion of [14] but only if the high order partials are carried out in the derivations of the polynomial of the reduced nonlinear normal of $x_2 = dx$ [23]. Matrix calculus cannot handle these derivations in the general case of multiple parameters in vectors X_1 and X_2 versus scalars x_1, x_2 of the F–R function. The outlined loop inverse technique, including the effect of the high order partials, expands the nonlinear estimation theory, numerical analysis, and applied mathematics from the standard techniques in this hybrid system.

As mentioned before, a detailed solution of the hybrid systems is beyond the scope of this paper. The presented work on the "pure" nonlinear estimation is extended next to the general norm of the residual vector. The principles of the fast network solution using local nonlinear and global linear array algebra are outlined, and some completed industrial applications are discussed.

3. Nonlinear robust estimation and large-scale global applications.

3.1. Arbitrary residual power. The details of the above outlined derivations can be verified by starting from the general scalar case of (5.1) of [26]. An arbitrary power θ for the absolute values of residuals is introduced for one single parameter dx and then extended to the general case, where the parameters dX form a column vector. The rules of matrix calculus and scalar polynomials can be exploited in these derivations by the application of the array contraction operators * and ** to the high order partials of array polynomials. The following full-rank Newton–Raphson (N–R) solutions with the integer values $\theta = 0$ and $\theta = 1$ are found for a single set of initial values

$$ddX = -[-F'^T PF' + L^T PF''_T]^{-1} F'^T PL, \qquad \theta = 0, \;\; p(i,i) = l(i)^{-2}, \tag{3.1}$$

$$ddX = -[0 + sign(L)^T F_T]^{-1} F'^T sign(L), \qquad \theta = 1. \tag{3.2}$$

As mentioned in section 5 of [26], (3.1) represents the iterative N–R solution of nonlinear cross correlation using a single initial value of zero uncertainty basket. It resembles the counterpart of least squares $\theta = 2$ with the exception of the reverse sign of the significant linear term $F'^T PF'$. Its diagonal "structural weight matrix" P has the values of $1/l(i)^2$. A unit weight matrix W is assumed for the observables G (or L_{obs}) in all above derivations of this paper. Unlike (3.1), a properly converged solution $\hat{X}$ of the full-rank case of least squares with $\theta = 2$ is unbiased under the nonlinear model $E(L) = F(X)$ and with the minimum trace of the covariance matrix $[F'(\hat{X})^T F'(\hat{X})]^{-1}$. The nonlinear Laplace iterations are shown in (3.2) for $\theta = 1$ where the "structural weights"

$$p(i,i) = abs[l(i)]^{\theta - 2}$$

convert the elements of L into integers ± 1. Notice the cancellation of N–G term $F'^T PF'$ as it has been multiplied by $\theta - 1 = 0$ of this "central" power $\theta = 1$ of Laplace estimation.

3.2. Weighting of observables and global expansion of LSM. Let us return to the role of a problem analyst and software designer in image matching by applying the above discussed new modeling and solution techniques. A shortcoming of the traditional cross correlation and a single point LSM is the simplistic math model of a constant shift dx for the entire (large) window. The starting point of LSM in

the problem of entity, or ELSM video registration, considered an entire image as one window while expanding the model $dx(x, y)$ of a single dx shift into the "global" or image variant parameters X of two-dimensional (2D) similarity, affine, and separable polynomial transformations [17]. The global parameters also model the image-variant $dy(x, y)$ shifts in the 2D image matching problem of two variables. The unknown set X of global parameters is shared by all points of local LSM samples of a regular grid in the reference image. How are the local LSM solutions (automated mensuration results) folded into the normals of the global parameters such that their highly varying quality is taken into account?

The solution of the local LSM normals of a small sample window would produce estimators of shifts dx, dy and a constant difference among gray values. The 3 x 3 weight matrix W of the global model observations is found from the local gradient matrix N' by scaling its elements with the local variance of minimized residuals. The folding of the local LSM normals is now accomplished by a weighted linear or nonlinear least squares technique of the global model with a 3 x n design matrix A at each point. The contribution of the LSM on the left-hand side of the global normals is therefore $A^T W A$. The globally weighted right-hand side reduces to $A^T N$, where N is scaled by the local variance estimate. Thus the global solution is required to obey more LSM points of good weight matrix and low local registration errors than the weak local solutions. Note that the local LSM solution $N'^{-1}N$ is not explicitly required. (Scaled N', N suffice.)

The global solution of registration over an entire image using the robust estimation and multigrid N–R techniques is more robust than any local correlation or LSM-based technique. It allows the use of small LSM windows, ultimately 1x1 pixels. The sample speed of using $(2q + 1)^2$, say, 100x100, local initial values of 0.01 pixel spacing is not far from that of one initial value. It is $(2q + 1)^2$ times better than the brute-force search of cross correlation with equally large windows and equally many initial search values. Cross correlation typically needs about 10x10=100 times larger windows when no global model constraints are used. Thus the LSM sample technique provides on the order of 100 $(2q + 1)^2$ or *one million times* the computational savings for a smart problem analyst. This implies that we need efficient techniques for the linear solution of the global model to feed the pull-in or initial value computation of the LSM sample locations in the next hyper iteration. The uses of linear and nonlinear array algebra merge in a large field of applications related to the global finite element solution of automated image registration, which is outlined next.

3.3. Global minimum residual matching (GMRM). The $dx(x, y)$ registration model of epipolar images consists of an empirical finite element elevation model for its automated validation, smoothing, and detection of the break-lines [24]. The elevation model consists of a dense regular grid of x-shifts in the reference image, where the local MRM normal equations are formed. The resulting local estimators and weights are folded into the global normals of the 2D finite element grid or array. The weighted continuity (regularization) constraints are added on the normals, and the grid parameters are solved simultaneously using the fast array algebra with a solution speed of millions of points per sec per iteration in modern computers. Weights of the local nonlinear MRM solutions make the linear global solution obey more closely the good match points while automatically filling in or filtering out the poor areas.

The node density is so high, typically 1x1–2x2 pixels, that the capture of sharp elevation changes (including the ground canopy) is possible in areas of valid stereo coverage. Optimally smoothed topographic contour lines are achieved when the x-

shift grid is mapped into the object space. In tie point mensuration among unoriented images, $dy(x, y)$ consists of the relative orientation model. A challenging problem expands the 2-ray imaging geometry of human stereo vision into multiple (3-12) rays by applying the principles of photogrammetric triangulation as the automatic editing and on-line validation tool of GMRM results—requiring fast solutions of literally billions of parameters. New technologies of differential data reduction, self-calibration, and photogeodesy are causing a similar rethinking of photogrammetry as the synthetic aperture radar and interferometric synthetic aperture radar techniques are of radargrammetry. The multi-image GMRM registrations enable high-resolution image mapping and visualization in support of automated extraction and fusion of high-resolution terrain and geographic information systems [5], [6], [21], [25].

Note that the fast 2D SVD finite elements array solution of GMRM is closely related to solving partial differential equations and related problems in various fields of the applied mathematics [24].

4. Summary and ideas for future work. Parts 1–2 of this introductory paper expanded the shorthand contraction rules and notations of matrix calculus to the multilinear and nonlinear systems of array algebra and tensor calculus. A general theory of matrix inverses called loop inverses in linear and nonlinear estimation was shown using the new matrix and tensor operators in the basis transforms among the parametric model and space domains. New notations for repeat vector and matrix contractions of the high order tensors of the Taylor expansion produced the general nonlinear normals and their solutions without the traditional restrictions (truncation and rank) on the high order tensor terms. The number of Taylor terms was expanded to infinity in one superiteration of the inverse Taylor expansion of consistent systems of nonlinear equations. Two superiterations combined with a similar direct solution of the space domain normals then produced the nonlinear Lm-inverse solution in one hyper iteration. The new theory was applied in some standard problems and nonlinear industrial applications with billions of parameters.

The solution and elimination techniques of partitioned, sparse, and other hybrid linear and nonlinear systems can be derived as an expansion or special cases of the presented theories. The outlined nonlinear direct solutions (1.24)–(1.27) involve new work in the tensor calculus to include the effect of all high order partials in the nonlinear parameter transforms (1.3)–(1.7) and their inverses. These derivations in the indical tensor notations and their detailed translations into array algebra will further refine the presented solutions.

The general tensor notations of [1], [2] based on the mathematical geodesy of Hotine [12] were valuable in checking the shown array algebra derivations. The detailed derivations of the Q-surface, loop inverse, and other tensor methods require more work. One topic is a detailed comparison of the Q-surface solution with the presented results of this paper, with a translation of the Q-surface tensor operators into the extended matrix or array notations when applicable. The nonlinear loop inverse technique can be expanded to a solution of the normals of normals resulting in a hyper set of the Lm-inverse (general nonlinear pseudoinverse) and Q-surface solutions. They refine the nonlinear correction term involving $\Delta N'$ in (1.18), (1.26), and (1.27) due to the total derivative change of the space domain normals, and they will expand the nonlinear condition adjustment and Kalman updating. The presented derivations should be expanded to correlated observations in the fashion of [4], [16], and the Q-surface technique.

A translation of the new array operators into the indical tensor calculus would

show what shorthand operators and notations are still required to make tensors more compatible with the matrix notations. The new work should test and improve the standards of the new notations and concepts for their adoption to the wide fields where the traditional vector, matrix, and tensor calculus have been used in the estimation theory, numerical analysis, and signal processing. Many such fields, including mathematical physics and physical geodesy, have applied the notations and concepts of classical mathematics before the adoption of some applicable shorthand matrix notations [9].

The new unified matrix and tensor operators enable an expansion of the Lm-inverse technique to the nonlinear optimization and estimation with continuous or gridded (versus randomly located discrete) observables. The hybrid linear and nonlinear solution techniques of discrete and continuous signals will expand the differential and integral calculus with many applications. In the fashion of the array polynomials of Taylor expansion, many classical nonlinear scalar functions can be expanded to the vector variables such as Newton, Lagrange, and other interpolation or extrapolation functions and their new nonlinear inverse functions. The array algebra expansion of binomials is valuable in these derivations in the fashion of the multigrid N–R solution in [26]. The new solutions of nonlinear scalar and vector functions will be valuable for future education and "fast" library functions of computers.

Acknowledgments. This review spans the work performed at the following institutions: Helsinki Technical University (1963–68), Royal Institute of Technology in Stockholm (1968–75), DBA Systems and Geodetic Services, Inc., in Melbourne, Florida (1975–1983), General Dynamics Electronics, GDE Systems, Marconi Integrated Systems, and BAE SYSTEMS in San Diego since 1984. The initial writing of parts 1–2 of this paper under the title "Latest Developments in Array Algebra" was prompted by the invitation of Dr. Fulton for the SIAM Minisymposium on Kronecker products, held in Charlotte, NC, October 25, 1995 and in Atlanta, 1999. Dr. Blaha reviewed a 1997 draft of this paper using his tensor notations of the nonlinear Q-surface estimation. The author is indebted to the referenced and other pioneers of the estimation theory of the matrix and tensor calculus used in photogrammetry and geodesy to provide the mathematical foundations of digital imaging and information systems. An anonymous reviewer and the editor helped to organize the paper into these two more digestible parts.

REFERENCES

[1] G. Blaha, *Non-Iterative Geometric Approach to Nonlinear Parametric Least-Squares Adjustment with or without Constraints*, PL-TR-91-2136, Phillips Laboratory, Hanscom AFB, Bedford, MA, 1991.

[2] G. Blaha, *Non-iterative approach to nonlinear least-squares adjustment*, Manuscripta Geodaetica, 19 (1994), pp. 199–212.

[3] A. Bouaricha, *Tensor methods for large, sparse unconstrained optimization*, SIAM J. Optim., 7 (1997), pp. 732–756.

[4] D. C. Brown, *A Matrix Treatment of General Problems of Least Squares Considering Correlated Observations*, RCA Data Reduction Technical Report 43, Aberdeen, MD, 1955, pp. 1–25.

[5] D. C. Brown, *Evolution, application and potential of the bundle method of photogrammetric triangulation*, in Proceedings of the ISP Comm. III Symposium, Stuttgart, 1974, pp. 1–95.

[6] D. C. Brown, *New developments in photogeodesy*, Photogrammetric Engrg. Remote Sensing, 60 (1994), pp. 877–894.

[7] D. Feng and T. H. Pulliam, *Tensor-GMRES method for large systems of nonlinear equations*, SIAM J. Optim., 7 (1997), pp. 757–779.

[8] B. Hallert, *Fotogrammetri*, Norstedt and Soner, Stockholm, 1964.
[9] W. A. Heiskanen and H. Moritz, *Physical Geodesy*, W. H. Freeman, San Francisco, London, 1967.
[10] U. V. Helava, *Mathematical Methods in the Design of Photogrammetric Plotters*, Photogrammetria 2, International Society of Photogrammetry and Remote Sensing, 1959–1960.
[11] R. A. Hirvonen, *Tasoituslasku*, Academy of Technical Sciences, Helsinki, 1965.
[12] M. Hotine, *Mathematical Geodesy*, ESSA Monograph 2, U.S. Department of Commerce, Washington, D.C., 1969.
[13] R. E. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic Engineering (1960), pp. 35–44.
[14] P. Meissl, *Direct solution of overdetermined algebraic problems with examples from geometric geodesy*, Manuscripta Geodaetica, 4 (1979), pp. 309–358.
[15] J. J. More, B. S. Garbow, and E. Hillstrom, *Testing unconstrained optimization*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[16] U. A. Rauhala, *Array Algebra with Applications in Photogrammetry and Geodesy*, Fot. Medd. VI:6, Department of Photogrammetry, Royal Institute of Technology, Stockholm, 1974.
[17] U. A. Rauhala, *Array algebra as general base of fast transforms*, in Proceedings of the Symposium on Image Processing—Interaction with Photogrammetry and Remote Sensing, Mitteilungen Der Geodaetischen Inst. Der T U Graz, Folge 29, 1977, pp. 175–188.
[18] U. A. Rauhala, *Intuitive derivation of loop inverses and array algebra*, Bull. Géodésique, 53 (1979), pp. 317–342.
[19] U. A. Rauhala, *Introduction to array algebra*, Photogrammetric Engrg. Remote Sensing, 46 (1980), pp. 177–192.
[20] U. A. Rauhala, *Note on general linear estimators and matrix inverses*, Manuscripta Geodaetica, 6 (1981), pp. 375–386.
[21] U. A. Rauhala, *Compiler Positioning of Array Algebra Technology*, International Society of Photogrammetry and Remote Sensing Vol. 26-3/3, Comm. III Symposium, Rovaniemi, 1986, pp. 173–198.
[22] U. A. Rauhala, *General theory of array algebra in nonlinear least squares and robust estimation*, ASPRS Spring Convention, Denver, 1990.
[23] U. A. Rauhala, *Nonlinear Array Algebra in Digital Photogrammetry*, International Society of Photogrammetry and Remote Sensing Vol. 29 B2 II, 1992, pp. 95–102.
[24] U. A. Rauhala, D. Davis, and K. Baker, *Automated DTM validation and progressive sampling algorithm of finite element array relaxation*, Photogrammetric Engrg. Remote Sensing, 4 (1989), pp. 449–465.
[25] U. A. Rauhala, G. Kunkel, V. Kluth, and S. Wong, *Global least squares matching automates the image registration process*, Proceedings of ASPRS/ACSM/RT92, 1 (1992), pp. 223–232.
[26] U. A. Rauhala, *Array algebra expansion of matrix and tensor calculus: Part* 1, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 490–508.
[27] J. B. Rosen, H. Park, and J. Glick, *Structured total least norm for nonlinear problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 14–30.
[28] R. B. Schnabel and P. D. Frank, *Tensor methods for nonlinear equations*, SIAM J. Numer. Anal., 21 (1984), pp. 815–843.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 529–552

IMPROVED SYMBOLIC AND NUMERICAL FACTORIZATION ALGORITHMS FOR UNSYMMETRIC SPARSE MATRICES*

ANSHUL GUPTA†

Abstract. We present algorithms for the symbolic and numerical factorization phases in the direct solution of sparse unsymmetric systems of linear equations. We have modified a classical symbolic factorization algorithm for unsymmetric matrices to inexpensively compute minimal elimination structures. We give an efficient algorithm to compute a near-minimal data-dependency graph for unsymmetric multifrontal factorization that is valid irrespective of the amount of dynamic pivoting performed during factorization. Finally, we describe an unsymmetric-pattern multifrontal algorithm for Gaussian elimination with partial pivoting that uses the task- and data-dependency graphs computed during the symbolic phase. These algorithms have been implemented in WSMP—an industrial strength sparse solver package—and have enabled WSMP to significantly outperform other similar solvers. We present experimental results to demonstrate the merits of the new algorithms.

Key words. sparse matrix factorization, parallel sparse solvers, multifrontal methods

AMS subject classifications. 05C50, 65F05, 65F50, 65Y05

PII. S089547980139604X

1. Introduction.

Typical direct solvers for general sparse systems of linear equations of the form $Ax = b$ have four distinct phases: *analysis,* comprising ordering for fill-in reduction and symbolic factorization; *numerical factorization* of the sparse coefficient matrix A into triangular factors L and U using Gaussian elimination with partial pivoting; *forward and backward elimination* to solve for x using the triangular factors L and U and the right-hand-side vector b; and *iterative refinement* of the computed solution. In this paper, we describe some of the algorithms that are used in the unsymmetric symbolic and numerical factorization phases of the Watson Sparse Matrix Package (WSMP)—a high-performance and robust software for solving general sparse linear systems. These algorithms are crucial to WSMP's performance, which has been shown to be significantly better than that of other similar solvers [18]. An important contribution of this paper is to show that, contrary to conventional wisdom, it is possible to symbolically determine a static communication pattern for parallel unsymmetric sparse LU factorization even in the presence of partial pivoting.

The process of factoring a sparse matrix can be expressed by a directed acyclic task-dependency graph (task-DAG). The vertices of this DAG correspond to the tasks of factoring rows or columns, or groups of rows and columns, of the sparse matrix, and the edges correspond to the dependencies between the tasks. A task is ready for execution if and only if all tasks with incoming edges to it have completed. In addition to a task-DAG, there is a data-dependency graph (data-DAG) associated with sparse matrix factorization. The vertex set of the data-DAG is the same as that of the task-DAG for a given sparse matrix. An edge from a vertex i to a vertex j in the data-DAG denotes that at least some of the output data of task i is required as input by task j. In this paper, we define task i as the task of computing column i of L and row i of U. Once the tasks are defined, the task-DAG is unique to a sparse

*Received by the editors October 3, 2001; accepted for publication (in revised form) by E. G. Ng October 18, 2002; published electronically December 3, 2002.
http://www.siam.org/journals/simax/24-2/39604.html
†IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598 (anshul@watson.ibm.com).

matrix for a given permutation of rows and columns; however, the data-DAG is a function of the sparse factorization algorithm. Multifrontal algorithms [9, 14, 23] for sparse factorization can work with a minimal data-DAG (i.e., a data-DAG with the smallest possible number of edges) for a given matrix.

In the case of symmetric sparse matrices, the minimal task- and data-DAGs for the factorization process are a tree called the elimination tree [22]. However, for unsymmetric sparse matrices, the task- and data-DAGs are general DAGs. Moreover, the edge-set of the minimal data-DAG for unsymmetric sparse factorization can be a superset of the edge-set of a task-DAG. Gilbert and Liu [16] describe elimination structures for unsymmetric sparse LU factors and give an algorithm for sparse unsymmetric symbolic factorization. These elimination structures are two DAGs that are transitive reductions of the graphs of the factor matrices L and U, respectively, and can be used to derive a task-DAG for sparse LU factorization. Some researchers have argued that computing an exact transitive reduction can be too expensive [9, 15] and have proposed using subminimal DAGs with more edges than necessary. However, traversing unnecessary DAG edges during numerical factorization can be a source of overhead. Moreover, in a parallel implementation, extra DAG edges can be potential sources of unnecessary synchronization or communication.

In this paper, we show how a relatively straightforward modification to Gilbert and Liu's symbolic factorization algorithm enables an efficient computation of the minimal elimination DAGs. We also define a set of edges that must be added to the task-DAG in order to generate a minimal data-DAG that is valid as long as partial pivoting with dynamic row and column exchanges is not performed during factorization. Finally, we describe how supplementing this data-DAG further with a small set of extra edges can yield a near-minimal data-DAG that is sufficient to handle an arbitrary number of pivot failures and the resulting row and column exchanges during numerical factorization. A *pivot failure* occurs when the pivot order predicted by the analysis phase must be altered during numerical factorization because the numerical value of the pivot is too small. By means of experiments on a suite of unsymmetric sparse matrices from real applications, we show that computing the final data-DAG is extremely fast. Furthermore, for the matrices in our test suite, this data-DAG has only a slightly higher number of edges than the task-DAG constructed using complete transitive reduction.

The multifrontal method [9, 14, 23] for sparse matrix factorization usually offers a significant performance advantage over conventional factorization schemes by permitting efficient utilization of parallelism and memory hierarchy. Duff and Reid [14] described a symmetric-pattern multifrontal algorithm for unsymmetric matrices that generates an elimination tree based on the symmetric structure of the union of the structures of A and the transpose of A to guide the numerical factorization. This algorithm works on square frontal matrices (see section 4.1) and can incur a substantial overhead for very unsymmetric matrices due to unnecessary data dependencies in the elimination tree and due to extra zeros in the artificially symmetrized frontal matrices. Davis and Duff [9] and Hadfield [20] introduced an unsymmetric-pattern multifrontal algorithm that overcomes the deficiencies of a symmetric-pattern algorithm. Our powerful symbolic phase enables us to use a much more simplified and efficient version of the unsymmetric-pattern multifrontal algorithm with partial pivoting. We describe the unsymmetric-pattern multifrontal algorithm that is used in WSMP and experimentally compare it with other state-of-the-art sparse unsymmetric factorization codes.

TABLE 1.1
Test matrices with their order (N), number of nonzeros (NNZ), and the application area of origin.

Number	Matrix	*N*	*NNZ*	Application
1	af23560	23560	484256	Fluid dynamics
2	av41092	41092	1683902	Finite element analysis
3	bayer01	57735	277774	Chemistry
4	bbmat	38744	1771722	Fluid dynamics
5	comp2c	16783	578665	Linear programming
6	e40r0000	17281	553956	Fluid dynamics
7	e40r5000	17281	553956	Fluid dynamics
8	ecl32	51993	380415	Circuit simulation
9	epb3	84617	463625	Thermodynamics
10	fidap011	16614	1091362	Fluid dynamics
11	fidapm11	22294	623554	Fluid dynamics
12	invextr1	30412	1793881	Fluid dynamics
13	mil053	530238	3715330	Structural engineering
14	mixtank	29957	1995041	Fluid dynamics
15	nasasrb	54870	2677324	Structural engineering
16	onetone1	36057	341088	Circuit simulation
17	onetone2	36057	227628	Circuit simulation
18	pre2	659033	5959282	Circuit simulation
19	raefsky3	21200	1488768	Fluid dynamics
20	raefsky4	19779	1316789	Fluid dynamics
21	rma10	46835	2374001	Fluid dynamics
22	tib	18510	145149	Circuit simulation
23	twotone	120750	1224224	Circuit simulation
24	wang3old	26064	177168	Circuit simulation
25	wang4	26068	177196	Circuit simulation

In Table 1.1, we introduce the suite of randomly chosen test matrices that we will use in experiments throughout this paper. The table shows the order of each matrix, the number of nonzeros in it, and the application area of the origin of the matrix. All matrices in our test suite arise in real-life problems and are in the public domain. The experiments reported in this paper were conducted on an IBM RS6000 WH-2 with a 375 MHz Power3 CPU, 2 Gbytes of RAM, 8 Mbytes of level-2 cache, and 64 Kbytes of level-1 cache.

The organization of this paper is as follows. Section 2 introduces the terms, conventions, and notations used in the paper. A symbolic factorization algorithm that computes the structure of the triangular factors and minimal elimination structures is described in section 3. In section 4, we describe how to compute near-minimal data-DAGs for unsymmetric multifrontal factorization. The numerical factorization algorithm is discussed in detail in section 5. We finish with concluding remarks in section 6. The last subsection of each major section contains experimental results pertaining to the algorithms in that section.

2. Terminology and conventions. We assume that the original $n \times n$ sparse unsymmetric coefficient matrix is irreducible and cannot be permuted into a block-triangular form. This is not a serious restriction, because a general matrix can first be reduced to a block-triangular form and then only the irreducible diagonal blocks need to be factored [12]. We assume that the coefficient matrix A is factored into a lower triangular matrix L and an upper triangular matrix U. Multiple row and column permutations may be applied to A during various stages of the solution process. However, for the sake of clarity, we will always denote the coefficient matrix by A and the factors by L and U. The state of permutation of A, L, and U will usually be clear

from the context.

We denote the directed graph corresponding to an $n \times n$ matrix M by $G_M(V_M, E_M)$, where $V_M = \{1, 2, \ldots, n\}$. A graph may not always be associated with an explicitly defined matrix. However, when it is, then an edge $i \rightarrow j \in E_M$ if and only if m_{ij} is a structural nonzero entry in the sparse matrix M. The transpose of a matrix M is represented by M'. If $i \rightarrow j \in E_M$, then $j \rightarrow i \in E_{M'}$, and vice-versa.

Struct$(M_{i,*})$ is the set of indices of the columns in M that have a structural nonzero entry in row i. This is also the set of all vertices to which i has an outbound edge in G_M. Similarly, Struct$(M_{*,i})$ is the set of indices of the rows in M that have a structural nonzero entry in column i and is also the set of all vertices from which i has an inbound edge in G_M. A directed path from node i to node j in the directed graph G_M is denoted by $i \rightsquigarrow j$. The transitive reduction $G_{M^O}(V_M, E_{M^O})$ of a graph $G_M(V_M, E_M)$ is the graph with the smallest number of edges that has a directed path $i \rightsquigarrow j$ if and only if G_M has a directed path $i \rightsquigarrow j$. Since we are primarily dealing with the nonzero structure of matrices rather than the actual values, we may also loosely refer to M^O as the transitive reduction of M if G_{M^O} is a transitive reduction of G_M. The leading $i \times i$ submatrix of M is denoted by M_i and the corresponding graph and its transitive reduction by G_{M_i} and $G_{M_i^O}$, respectively.

The edges and paths in some of the graphs used in this paper are labeled. An edge in a labeled graph can have one of the three labels—L, U, or LU. Depending on its label, an edge can be an L-edge, a U-edge, or an LU-edge. L-, U-, and LU-edges from vertex i to j are denoted by $i \xrightarrow{\text{L}} j$, $i \xrightarrow{\text{U}} j$, and $i \xrightarrow{\text{LU}} j$, respectively. An L-path from i to j, denoted by $i \overset{\text{L}}{\rightsquigarrow} j$, is a directed path containing only L- and LU-edges. Similarly, a U-path from i to j, denoted by $i \overset{\text{U}}{\rightsquigarrow} j$, is a directed path containing only U- and LU-edges. If an L-edge $i \xrightarrow{\text{L}} j$ exists in the graph, then $j = \text{L-parent}(i)$. Similarly, if $i \xrightarrow{\text{U}} j$ exists, then $j = \text{U-parent}(i)$, and if $i \xrightarrow{\text{LU}} j$ exists, then $j = \text{LU-parent}(i)$.

We define[1] a supernode $[q : r]$ as a maximal set of consecutive indices $\{q, q+1, \ldots, r\}$ such that for all $i \in [q{:}r]$, Struct$(L_{*,i}) =$ Struct$(L_{*,q}) - \{q, q+1, \ldots, i-1\}$ and Struct$(U_{i,*}) =$ Struct$(U_{q,*}) - \{q, q+1, \ldots, i-1\}$. For $n \times n$ matrices L and U, we define $m \times m$ supernodal matrices $\mathcal{L}$ and $\mathcal{U}$ such that each supernode $[q{:}r]$ in L and U is represented by a single row and column $g = \sigma([q : r])$ in $\mathcal{L}$ and $\mathcal{U}$. Here $m \leq n$ is the total number of supernodes. Furthermore, if $g = \sigma([q{:}r])$, $h = \sigma([s{:}t])$, and $r < s$, then $g < h$; that is, the column and row indices in $\mathcal{L}$ and $\mathcal{U}$ maintain the relative order of supernodes in L and U.

3. Computing a task-DAG and the structures of L and U. Gilbert and Liu [16] present an unsymmetric symbolic factorization algorithm to compute the structures of the factors L and U and their transitive reductions L^O and U^O. Figure 3.1 summarizes Gilbert and Liu's algorithm. The algorithm computes the structure of L, U, and L^O row by row and computes the structure of U^O by columns.

The total time that the algorithm shown in Figure 3.1 spends in step 1 is bounded by flops(LU^O) [16], which is the number of operations required to multiply the sparse matrices L and U^O. Similarly, the time spent in step 3 is bounded by flops(UL^O). The total computational cost of steps 2 and 4 is $O(n(|E_{L^O}| + |E_{U^O}|))$. This is because transitive reduction is performed on n rows of U and columns of L, and the ith step could potentially traverse all edges in $G_{L_i^O}$ and $G_{U_i^O}$. Steps 2 and 4 of Gilbert and Liu's algorithm are much more costly than steps 1 and 3. The cost of these steps

[1] Other definitions of supernodes in the context of unsymmetric sparse factorization have been used in the literature [11].

for $i = 1$ to n **do**

1. Compute Struct$(L_{i,*})$ from Struct$(A_{i,*})$ by traversing $G_{U^O_{i-1}}$ and using the fact that $\forall\, j < i$, $j \in$ Struct$(L_{i,*})$ if and only if $\exists\, k \leq j$ such that $k \in$ Struct$(A_{i,*})$ and there is a path $k \rightsquigarrow j$ in U^O_{i-1}.
2. Transitively reduce Struct$(L_{i,*})$ using $G_{L^O_{i-1}}$ and extend it to $G_{L^O_i}$.
3. Compute Struct$(U_{i,*})$ = $((\cup_{j:j\rightarrow i\in E_{L'^O_i}}$ Struct$(U_{j,*}))$ $\cup$ Struct$(A_{i,*}))$ $-$ $\{1, 2, \ldots, i-1\}$.
4. Transitively reduce Struct$(U_{*,i})$ using $G_{U'^O_{i-1}}$ and extend it to $G_{U'^O_i}$.

end for

FIG. 3.1. *Gilbert and Liu's unsymmetric symbolic factorization algorithm* [16].

for $i = 1$ to n **do**

1. Transitively reduce Struct$(L_{i,*})$ using $G_{L^O_{i-1}}$ and extend it to $G_{L^O_i}$.
2. Compute Struct$(U_{i,*})$ = $((\cup_{j:j\rightarrow i\in E_{L'^O_i}}$ Struct$(U_{j,*}))$ $\cup$ Struct$(A_{i,*}))$ $-$ $\{1, 2, \ldots, i-1\}$.
3. Transitively reduce Struct$(U_{*,i})$ using $G_{U'^O_{i-1}}$ and extend it to $G_{U'^O_i}$.
4. Compute Struct$(L_{*,i})$ = $((\cup_{j:j\rightarrow i\in E_{U^O_i}}$ Struct$(L_{*,j}))$ $\cup$ Struct$(A_{*,i}))$ $-$ $\{1, 2, \ldots, i-1\}$.

end for

FIG. 3.2. *A modified symbolic factorization algorithm.*

has prompted researchers to seek alternatives, such as computing fast but incomplete transitive reduction [9, 15]. The use of such alternatives to G_{L^O} and G_{U^O} with more edges than G_{L^O} and G_{U^O}, respectively, can increase the cost of steps 1 and 3, as well as that of numerical factorization.

3.1. A modification to Gilbert and Liu's algorithm. We now describe a relatively simple modification to the algorithm shown in Figure 3.1. We start by splitting the original coefficient matrix into a lower triangular part stored by columns and an upper triangular part stored by rows. In our modified symbolic factorization algorithm, we compute the structure of L by the columns (i.e., L' by rows) and that of U by the rows. This is achieved by simply reformulating the algorithm shown in Figure 3.1 to perform only steps 2 and 3, but twice for each i on two sets of identical data structures—one corresponding to L' and the other corresponding to U. The modified algorithm is shown in Figure 3.2.

Note that in the algorithm of Figure 3.2, steps 3 and 4 are identical to steps 1 and 2, respectively. The first two steps compute the ith rows of L^O and U and the last two steps compute the ith columns of U^O and L. An actual code of this algorithm can use the same pair of routines with different arguments to implement all four steps. The reduction in the size of the code by half, however, is a secondary benefit of the modified algorithm. The primary advantage of this scheme is that it allows immediate detection of supernodes during symbolic factorization. This, as we shall explain in section 3.2, allows us to avoid computing and storing G_{L^O} and $G_{U'^O}$ explicitly. Instead, we can work only with their supernodal counterparts $G_{\mathcal{L}^O}$ and $G_{\mathcal{U}'^O}$.

3.2. Use of supernodes to speed up transitive reduction. Most modern sparse factorization codes rely heavily on supernodes to efficiently utilize memory hierarchies and parallelism in the hardware. Supernodes are so crucial to high performance in sparse matrix factorization that the criterion for the inclusion of rows and columns in the same supernode is often relaxed [7] to increase the size of the supernodes. Consecutive rows and columns with nearly the same but not identical structures are often included in the same supernode, and artificial nonzero entries with a numerical value of 0 are added to maintain identical row and column structures for all members of a supernode. The rationale is that the slight increase in the number of nonzeros and floating-point operations involved in the factorization is more than compensated for by a higher factorization speed.

WSMP's LU factorization algorithm also works on the relaxed supernodes generated by its symbolic factorization. In the symbolic factorization algorithm, as soon as Struct($L_{*,i}$) and Struct($U_{i,*}$) are computed in the ith iteration of the outer loop, they can be compared with Struct($L_{*,i-1}$) and Struct($U_{i-1,*}$) to determine if they belong to the current supernode. A new row-column pair is added to the current supernode if its structure is either identical or nearly identical to the previous row-column pair. If the ith row-column pair fails to meet the criterion for membership into the current supernode, then a new supernode is started at i.

The use of supernodes allows us to significantly reduce the cost of computing the transitive reductions. In step 1 of the algorithm shown in Figure 3.2, instead of transitively reducing the entire Struct($L_{i,*}$), we reduce only the set $\{h : h = \sigma([q\!:\!r]),$ where $[q\!:\!r] \subseteq$ Struct$(L_{i,*})\}$. Step 3 is treated similarly. As a result of working only with supernodes, the upper bound on the cost of computing the transitive reduction decreases from $O(n(|E_{L^O}| + |E_{U^O}|))$ to $O(n(|E_{\mathcal{L}^O}| + |E_{\mathcal{U}^O}|))$. This is because only the supernodal DAGs $G_{\mathcal{L}^O}$ and $G_{\mathcal{U}^O}$ are searched during each of the n transitive reduction steps. Strict supernodal graphs $G_{\mathcal{L}^O}$ and $G_{\mathcal{U}^O}$ would have at least $n - m$ fewer edges than G_{L^O} and G_{U^O}, where m is the number of supernodes. The reason is that $\mathcal{U}^O$ and $\mathcal{L}'^O$ do not contain any edges $i \rightarrow j$, where $j = i + 1$, $q \leq i < r$, and $[q : r]$ is a supernode. The use of relaxed supernodes reduces the number of edges even further because some potential edges of the form $i \rightarrow j$, where $j > i + 1$, may be eliminated from the task-DAG when nodes i and $i + 1$ are artificially merged.

3.3. Task-DAGs for LU factorization. In this paper, we will refer to two types of task-DAGs: a conventional DAG denoted by T^C and a supernodal DAG denoted by T^S. Each vertex of the conventional task-DAG refers to the task of computing a single row of U and the corresponding column of L. On the other hand, a vertex of the supernodal task-DAG corresponds to a set of row-column pairs that constitute a supernode. Although, in a practical implementation, we always work with supernodal DAGs, we will often use conventional task- and data-DAGs in the remainder of the paper to keep the exposition simple. All results and descriptions presented in terms of the conventional DAGs map naturally to the supernodal case.

We first show how to compute T^C in terms of the conventional structures L'^O and U^O. The transpose matrix L' is used to indicate that for all $i \rightarrow j \in E_{T^C}$, $j > i$.

THEOREM 3.1. *T^C is a task-DAG for LU factorization if its vertex set $V_{T^C} = \{1, 2, \ldots, n\}$ and its edge-set $E_{T^C} = E_{U^O} \cup E_{L'^O}$.*

Proof. To prove that T^C is a task-DAG, we show that E_{T^C} is sufficient to represent a proper ordering of the n elimination tasks denoted by V_{T^C}. Struct($L_{*,i}$) can contribute to Struct($L_{*,j}$) only if $i \in$ Struct($U_{*,j}$), and if this is the case, then the symbolic factorization algorithm of Figure 3.2 ensures that U^O contains either $i \rightarrow j$

TABLE 3.1

Comparison of conventional symbolic factorization (due to Gilbert and Liu [16]) with supernodal symbolic factorization. $|V|$ is the size of the largest diagonal block in the matrix on which symbolic factorization is performed; N_{sup} is the number of supernodes; t^C and t^S are the times in seconds of the two symbolic factorization algorithms; and $|E_{T^C}|$ and $|E_{T^S}|$ are the number of edges in the task-DAGs produced by the two algorithms.

Matrix	$\|V\|$	N_{sup}	Conventional symbolic		Supernodal symbolic				
	(n)	(m)	t^C	$\|E_{T^C}\|$	t^S	$\|E_{T^S}\|$	$\frac{n}{m}$	$\frac{\|E_{T^C}\|}{\|E_{T^S}\|}$	$\frac{t^C}{t^S}$
af23560	23560	4744	2.6	23644	.47	4793	5.0	4.9	5.5
av41092	41086	19302	3.1	62197	.83	34708	2.1	1.8	3.7
bayer01	48803	33891	.45	113948	.28	87028	1.4	1.3	1.6
bbmat	38744	4877	10.	41260	1.7	6077	7.9	6.8	5.9
comp2c	6756	996	.66	8005	.22	1736	6.8	4.6	3.0
e40r0000	17281	3049	.47	19262	.14	3225	5.7	6.0	3.4
e40r5000	17281	2755	.60	19891	.16	3182	6.3	6.3	3.8
ecl32	42341	12087	7.9	48779	1.2	15239	3.5	3.2	6.6
epb3	84617	30009	1.2	106137	.50	38088	2.8	2.8	2.4
fidap011	16614	1262	2.3	16613	.42	1261	13.	13.	5.5
fidapm11	22294	2327	3.7	23144	.65	2651	9.6	8.7	5.7
invextr1	30412	5295	4.5	37685	.93	10108	5.7	3.8	4.8
mil053	530238	166155	15.	530237	4.5	166154	3.2	3.2	3.3
mixtank	29957	2984	7.8	30949	1.2	3203	10.	9.7	6.5
nasasrb	54870	3808	4.9	54869	.97	3807	14.	14.	5.1
onetone1	32211	14215	1.1	45227	.31	23585	2.3	1.9	3.5
onetone2	32211	14843	.44	45073	.18	23999	2.2	1.9	2.4
pre2	629628	243693	30.	765210	6.4	317216	2.6	2.4	4.7
raefsky3	21200	1282	2.1	21199	.41	1281	17.	17.	5.1
raefsky4	19779	1359	2.9	19778	.50	1358	15.	15.	5.8
rma10	46835	3855	2.1	47152	.56	3911	12.	12.	3.8
tib	17583	7823	.11	22904	.07	10060	2.2	2.3	1.6
twotone	105740	34304	2.6	126656	.91	44856	3.1	2.8	2.9
wang3old	26064	8451	3.1	26063	.54	8450	3.1	3.1	5.7
wang4	26068	8254	3.0	26067	.53	8253	3.2	3.2	5.7

or $i \rightsquigarrow j$. The same is true for Struct$(U_{i,*})$, Struct$(U_{j,*})$, and L'^O. Therefore, every row-column pair i that updates row and column j must be eliminated before j. □

Theorem 3.1 can be easily extended to the supernodal case. The supernodal task-DAG T^S is defined by a vertex set $V_{T^S} = \{1, 2, \ldots, m\}$ and an edge set $E_{T^S} = E_{\mathcal{U}^O} \cup E_{\mathcal{L}'^O}$, where m is the number of supernodes.

3.4. Experimental results. In Table 3.1, we compare Gilbert and Liu's symbolic factorization algorithm [16] with the supernodal symbolic factorization algorithm described in section 3.2. We report their CPU times t^C and t^S, respectively, and the number edges in task DAGs T^C and T^S generated by them.

The last column of Table 3.1 shows the factor by which the supernodal symbolic factorization is faster than the conventional algorithm. The table also shows average supernode size (n/m) and the ratio of edges in T^C and T^S for each matrix. These two ratios are closely related. The ratio of t^C and t^S bears some correlation to the ratio of edges in T^C and T^S, but the actual ratio is matrix dependent. Note that only the time of transitive reduction steps 1 and 3 of the algorithm in Figure 3.2 is reduced by the use of supernodes; the time of computing the structures of L and U in steps 2 and 4 remains mostly unchanged (other than some reduction in the number of structures merged due to supernode relaxation). Therefore, the actual reduction achieved in the symbolic factorization time depends on the relative amounts of time spent in

transitive reduction and computing L and U structures. Moreover, Table 3.1 reports the number of edges in the task-DAGs, not the number of edges in the actual lower and upper triangular transitively reduced graphs that are traversed during symbolic factorization. Recall that the edge-set of a task-DAG is the union of the edge-sets of the corresponding lower and upper triangular transitively reduced graphs. The amount of structural symmetry in the matrix affects the number of common edges between the upper and lower transitively reduced graphs, which in turn determines the actual number of edges in the task-DAG.

Eisenstat and Liu [15] present an alternative to complete transitive reduction to reduce the cost of this step in sparse unsymmetric symbolic factorization. They propose exploiting structural symmetry in the matrix to compute partial transitive reductions. Although they present experimental results on a different set of much smaller matrices, it appears that the use of supernodes as proposed in section 3.2 can achieve much higher speedups in symbolic factorization while computing exact transitive reductions than the partial transitive reduction scheme proposed in [15]. However, Eisenstat and Liu's algorithm too can be sped up by the use of supernodes. A supernodal version of this algorithm has been implemented in the SuperLU$_{dist}$ [21] sparse solver package. We compared our symbolic factorization time with that of SuperLU$_{dist}$ and found the latter to be slower by about 25% overall on our test suite. This could be partly due to implementation differences and partly due to the fact that while Eisenstat and Liu's algorithm saves time in the transitive reduction computation, it spends extra time in merging structures due to redundant edges in the DAG. It appears that the use of supernodes in Gilbert and Liu's algorithm can speed up its transitive reduction enough for it to match or outperform even a supernodal version of Eisenstat and Liu's algorithm in execution time.

4. Data-DAGs for unsymmetric multifrontal LU factorization. The original multifrontal algorithm [14, 23] was described in the context of a symmetric-pattern coefficient matrix but has been applied to matrices with unsymmetric patterns by introducing zero-valued entries at appropriate locations to convert the original matrix into one with the pattern of $A + A'$ [14, 2, 4]. This can cause a substantial overhead for very unsymmetric matrices due to the extra computation performed on the introduced entries and the resulting fill-in. Davis and Duff [9] and Hadfield [20] introduced an unsymmetric-pattern multifrontal algorithm to overcome this shortcoming. In this section, we develop near-minimal data-DAGs for the unsymmetric multifrontal algorithm—an aspect of unsymmetric multifrontal factorization that has not been well investigated in previous works. As we shall show in section 5, the availability of a near-minimal data-DAG aids in the efficient implementation of the numerical factorization phase. It would also help minimize the synchronization and communication overheads in a parallel implementation.

4.1. Outline of the symmetric multifrontal algorithm. The symmetric-pattern multifrontal algorithm is guided by an assembly or elimination tree [22, 23, 19], which serves as both the task- and data-dependency graphs for the factorization process. The data associated with each supernode of the elimination tree is a square frontal matrix. A *frontal matrix* F^g associated with a supernode $g = \sigma([q:r])$ is a dense matrix whose dimensions are equal to $|\text{Struct}(L_{*,q})|$ or $|\text{Struct}(U_{q,*})|$. The contiguous local row and column indices in the dense frontal matrix correspond to noncontiguous global indices of the matrix $L + U$. Each entry in a frontal matrix corresponds to a structural nonzero entry in the global matrix. After a frontal matrix F^g is fully assembled or populated, the leading $r - q + 1$ rows and columns corresponding to

the supernode (also known as the pivot block) are factored and become parts of the factors U and L, respectively. The remaining trailing part of the frontal matrix is now called the update or the *contribution matrix*, denoted by C^g. The contribution matrix corresponding to a supernode is assembled completely into the frontal matrix of its only parent supernode and is never accessed again. This is because if $h = \sigma([s:t])$ is the parent of supernode $g = \sigma([q:r])$ in the elimination tree, then $\text{Struct}(L_{*,r}) - \{r\} \subseteq \text{Struct}(L_{*,s})$. The same is true for columns of U due to symmetry.

In a recursive formulation of the symmetric-pattern multifrontal algorithm, the task corresponding to a supernode first completes identical subtasks for each of its children in the elimination tree, then assembles their contribution matrices into its frontal matrix, and finally performs the partial factorization on the frontal matrix. Calling a recursive procedure to perform the task described above on the root supernode of the elimination tree completes the factorization of a sparse matrix with a symmetric structure.

4.2. Outline of the unsymmetric multifrontal algorithm. The overall structure of an unsymmetric-pattern multifrontal algorithm is similar to its symmetric counterpart and can be expressed in the form of a recursive procedure starting at the root (the supernode with no outgoing edges) of the task-DAG. However, there are two major differences. The first difference is in the control-flow. In the unsymmetric multifrontal algorithm, before starting a subtask for a child, the task corresponding to the parent supernode must check to see if the child supernode has already been processed by another parent. Only the first parent to reach a child actually performs the recursive computation starting at that child. The second difference is in the data-flow, or the way contribution matrices are assembled into frontal matrices. This is explained below in greater detail.

Recall that the edge-set E_{T^C} of the task-DAG T^C is the union of the edge-sets $E_{L'^O}$ and E_{U^O} of the transitive reductions of L' and U, respectively. We now assign labels to the edges in T^C. The edges contributed to E_{T^C} solely by $E_{L'^O}$ are labeled as L-edges. Similarly, edges contributed to E_{T^C} solely by E_{U^O} are labeled as U-edges. The third type of label, the LU-label, is assigned to the edges that belong to the intersection $E_{L'^O}$ and E_{U^O}. Finally, an L-edge $i \xrightarrow{\text{L}} j$ is converted to an LU-edge $i \xrightarrow{\text{LU}} j$ if there is a U-path $i \overset{\text{U}}{\rightsquigarrow} j$ in T^C, and a U-edge $i \xrightarrow{\text{U}} j$ is converted to $i \xrightarrow{\text{LU}} j$ if there is an L-path $i \overset{\text{L}}{\rightsquigarrow} j$ in T^C. The edges of the supernodal task-DAG T^S are defined similarly.

Unlike the symmetric multifrontal algorithm, the frontal and contribution matrices in the unsymmetric multifrontal algorithm are, in general, rectangular rather than square. Furthermore, a contribution matrix in the unsymmetric multifrontal algorithm can potentially be assembled into more than one frontal matrix because a supernode in the data-DAG can have more than one parent. As described in [20], the assembly of contribution matrices into the parent frontal matrices in the unsymmetric multifrontal algorithm proceeds as follows.

Let $g \xrightarrow{\text{L}} h$ be an L-edge in the data-DAG, where $g = \sigma([q:r])$ and $h = \sigma([s:t])$. If $\text{Struct}(L_{*,q})$ and $\text{Struct}(L_{*,s})$ have an index i in common, then all elements of row i of U in C^g can potentially be assembled into F^h. Similarly, if $g \xrightarrow{\text{U}} h$ is a U-edge and $\text{Struct}(U_{q,*})$ and $\text{Struct}(L_{s,*})$ have an index i in common, then all elements of column i of L in C^g can potentially be assembled into F^h. Finally, if $g \xrightarrow{\text{LU}} h$ is an LU-edge, then the entire trailing submatrix of C^g with global row and column indices greater than or equal to s can be assembled into F^h.

Certain entries of C^g may have potential destinations in the frontal matrices of

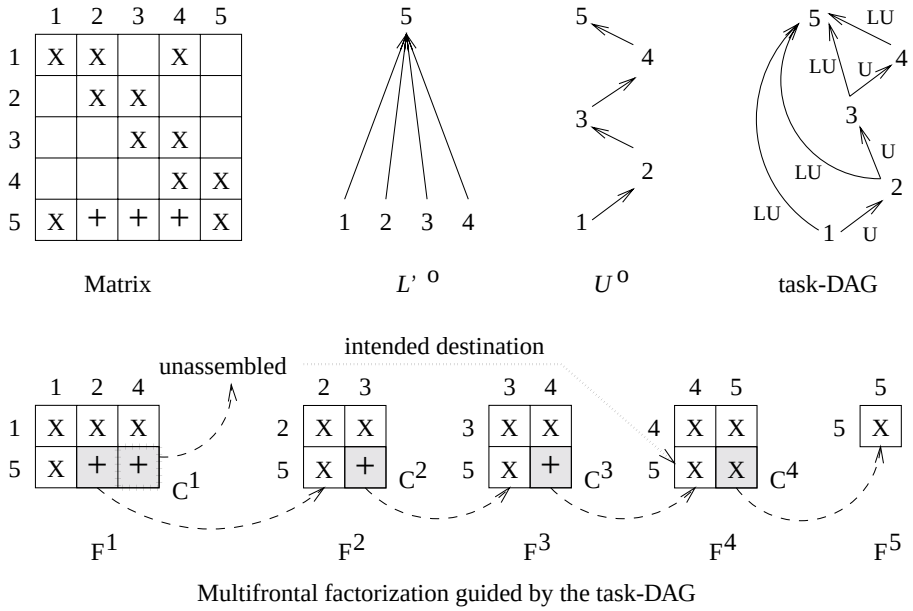


FIG. 4.1. *An example of the inability of a task-DAG to guide complete assembly of all contribution matrices in the unsymmetric multifrontal algorithm. An* 'X' *denotes a nonzero in the coefficient matrix and a '+' denotes a nonzero created due to fill-in.*

more than one parent of g even if the data-DAG contains no unnecessary edges. This is because C^g can have common rows (columns) with the frontal matrices of more than one among g's LU- and L-parents (U-parents). The unsymmetric multifrontal algorithm must ensure that any entry of a contribution matrix is not used to update more than one frontal matrix. Additionally, a correct data-DAG must have sufficient outgoing edges from all supernodes so that each entry of a contribution matrix has a potential destination in at least one frontal matrix.

4.3. Inadequacy of task-DAG for unsymmetric multifrontal algorithm. By means of a small example in Figure 4.1, we show that if the task-DAG defined in section 3.3 is used as a data-DAG, then all contribution matrices may not be fully absorbed into their parent frontal matrices. The figure shows a sparse matrix with factorization fill-in, the transitively reduced DAGs $\mathcal{L}'^O$ and $\mathcal{U}^O$, and the task-DAG with its edges labeled as described in section 4.2. For the sake of clarity, each supernode is chosen to be of size 1. The figure shows all frontal and contribution (shaded portions) matrices and the flow of data from the contribution to frontal matrices along the edges of the task-DAG. Note that all edges may not lead to a data transfer; e.g., $1 \xrightarrow{\text{LU}} 5$. It is easily seen that the U-edge $1 \xrightarrow{\text{U}} 4$, which is absent from the task-DAG (because it is removed while transitively reducing $\mathcal{U}$ to $\mathcal{U}^O$), is necessary for the complete assembly of C^1.

4.4. A data-DAG for a predefined pivot sequence. Having shown that the minimal task-DAG cannot serve as a data-DAG for unsymmetric multifrontal factorization, we now define a data-DAG that is sufficient for the proper assembly of all contribution matrices, as long as rows and columns are not exchanged among different supernodes for pivoting. We will use D^N to denote such a DAG, where the superscript N stands for "no pivoting." A data-DAG D^P that can accommodate

pivoting will be described in section 4.5.

THEOREM 4.1. *If a column index $j \in$ Struct($U_{i,*}$) satisfies all of the following conditions, then a U-edge $i \xrightarrow{U} j$ is necessary for C^i to be completely assembled into its parents' frontal matrices:*

1. *The LU-parent of i, if it exists, is greater than j.*
2. *None of i's U-parents are in Struct($U_{*,j}$).*
3. *There exists a $k \in$ Struct($L_{*,i}$) such that $k > j$.*

The transpose of this theorem can be stated similarly.

Proof. The contribution matrix C^i has a column that contributes to $L_{*,j}$, because, at the least, there is an element corresponding to $L_{k,j}$ in C^i. At the same time, none of i's U-parents' frontal matrices have column j, so they cannot absorb $L_{*,j}$ from C^i. Since the LU-parent of i is greater than j, it too cannot absorb $L_{*,j}$ from C^i. The addition of $i \xrightarrow{U} j$ makes it possible for C^i to contribute $L_{*,j}$ to F^j. The transpose case can be proven similarly. ∎

Theorem 4.1 captures the situation illustrated in Figure 4.1 for $i = 1$, $j = 4$, and $k = 5$ and prescribes the addition of $1 \xrightarrow{U} 4$ to ensure complete assembly of C^1.

THEOREM 4.2. *If D^N is a DAG formed by adding all possible edges to T^C according to Theorem* 4.1, *provided that these edges don't already exist, then D^N is a data-dependency DAG for the unsymmetric multifrontal algorithm without pivoting.*

Proof. To show that D^N is a data-DAG, we must show that its edge-set is sufficient for the complete absorption of all contribution matrices into their parent frontal matrices. We prove this by contradiction.

Without loss of generality, assume that an element corresponding to $L_{k,j}$ in C^i is not assembled. Note that $i < j < k$. If $L_{k,j}$ is in C^i, then $j \in$ Struct($U_{i,*}$) and $k \in$ Struct($L_{*,i}$). Since $j \in$ Struct($U_{i,*}$), either $i \xrightarrow{U} j \in E_{T^C}$ or there is a U-path $i \overset{U}{\rightsquigarrow} j$ in T^C. If $i \xrightarrow{U} j \in E_{T^C}$, then all entries with row indices greater than or equal to j in column j of C^i will be absorbed by F^j, and these entries include the one corresponding to $L_{k,j}$. If $i \xrightarrow{U} j \notin E_{T^C}$, then a U-path $i \overset{U}{\rightsquigarrow} j$ exists in T^C and there are two possibilities: either LU-parent(i) $\leq j$ or LU-parent(i) $> j$. Let $l =$ LU-parent(i). If $l \leq j$, then the entire trailing submatrix of C^i with row and column indices greater than l, including $L_{k,j}$, will be assembled into F^l. If $l > j$, then consider two further possibilities: either one of i's U-parents is in Struct($U_{*,j}$) or is not. If one is, then its frontal matrix will absorb column j from C^i. If none of i's U-parents is in Struct($U_{*,j}$), then all conditions for the applicability of Theorem 4.1 are satisfied. Therefore, $i \xrightarrow{U} j$ would have been added to D^N and would have caused the entry corresponding to $L_{k,j}$ in C^i to be absorbed into F^j. Thus, it is not possible for the entry corresponding to $L_{k,j}$ to be left unassembled in any C^i. Similarly, it can be shown that the entry corresponding to any $U_{j,k}$ cannot be left unassembled in any C^i. ∎

Having shown that the edge-set of D^N is sufficient for unsymmetric multifrontal factorization without pivoting, we now show that not all edges that D^N inherits from T^C may be necessary if pivoting is not performed during factorization.

THEOREM 4.3. *For LU factorization without pivoting, an edge $i \xrightarrow{U} j$ ($i \xrightarrow{L} j$) or $i \xrightarrow{LU} j$ in T^C is redundant if the maximum index in Struct($L_{*,i}$) (Struct($U_{i,*}$)) is smaller than j.*

Proof. Recall that Struct($L_{*,j}$) = ((∪$_{i:i\to j\in E_{U^O}}$ Struct($L_{*,i}$)) ∪ Struct($A_{*,j}$)) − $\{1, 2, \ldots, j-1\}$. If the maximum index in Struct($L_{*,i}$) is smaller than j, then Struct($L_{*,i}$) $\subseteq \{1, 2, \ldots, j-1\}$ and does not contribute to Struct($L_{*,j}$). The proof for L'^O and Struct($U_{i,*}$) is similar. ∎

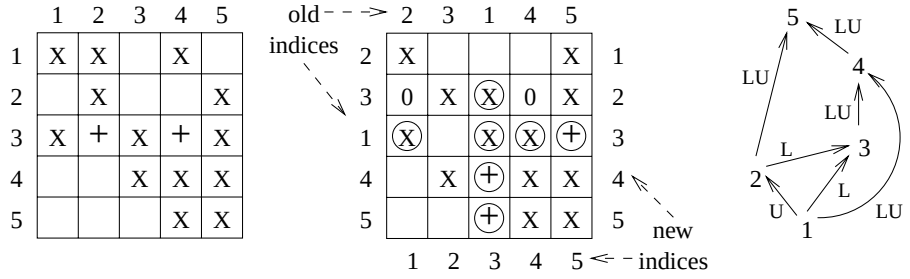
Note that Theorem 4.3 is valid only if row and column exchanges are not performed during LU factorization. Otherwise, additional fill-in caused by pivoting could create an index greater than or equal to j in Struct($L_{*,i}$) or Struct($U_{i,*}$), even if it is not predicted by the symbolic factorization on the original permutation of the matrix. Therefore, all edges in T^C could potentially be used.

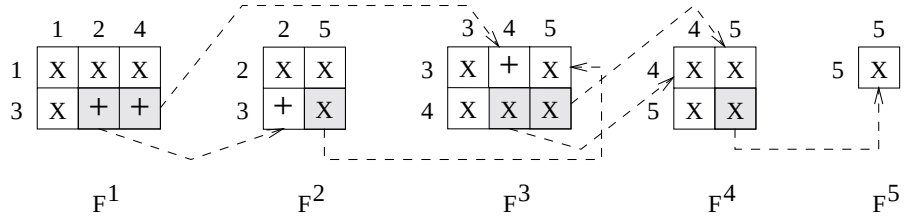
Supernodal versions of Theorems 4.1–4.3 for T^S can be proven similarly. To summarize the results of this subsection, we have shown how to construct a data-DAG for unsymmetric multifrontal factorization without pivoting from a task-DAG and we have shown that although the task-DAG is derived from the strict transitive reductions of L' and U (or $\mathcal{L}'$ and $\mathcal{U}$), it may still pass on edges to the data-DAG that are redundant if pivoting is not performed during factorization. Therefore, the data-DAG is not minimal. However, if pivoting is performed, then potentially all the edges could get used.

4.5. Supplementing the data-DAG for dynamic pivoting. We will now show that the edge-set of data-DAG D^N constructed in section 4.4 may not be sufficient if pivoting is performed during factorization. We also discuss how to supplement E_{D^N} to generate a data-DAG D^P whose edge-set is sufficient to handle any amount of pivoting. We start with an overview of the pivoting methodology in the unsymmetric multifrontal algorithm, which has been described in detail in [20].

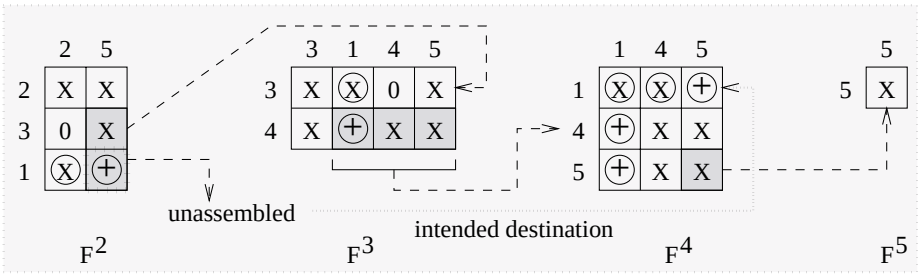
If a diagonal element $A_{i,i}$ $(q \leq i \leq r)$ in a supernode $[q:r]$ fails to meet the pivoting criterion, then first an attempt is made to exchange row and column i with a row j and a column k such that $i < j \leq r$, $i < k \leq r$ and $A_{j,k}$ satisfies the pivoting criterion. Such intrasupernode pivoting has no effect on the structure of the factors and factorization can continue as usual. However, it may not always be possible to find a suitable row-column pair within a supernode's pivot block to satisfy the pivoting criterion. In this situation, intersupernode pivoting is necessary. If $h = \sigma([s:t])$ is the LU-parent of $g = \sigma([q:r])$ in the data-DAG and a suitable ith pivot cannot be found within the pivot block of F^g, then all row-column pairs from i to r are symmetrically permuted to new locations from $s-(r-i+1)$ to $s-1$. Thus, effectively, supernode $[q:r]$ shrinks to $[q:i-1]$ and the supernode $[s:t]$ expands to $[s-(r-i+1):t]$. As a side effect of this pivoting, there is additional fill-in in all the ancestors of g in the data-DAG that are smaller than h. In particular, the columns of L of all of g's U-ancestors smaller than h get extra row indices $[i:r]$ and the rows of U of all of g's L-ancestors smaller than h get extra column indices $[i:r]$. A failure in supernode h is handled similarly in a recursive manner.

In D^N, whose construction is described in section 4.4, all supernodes may not have an LU-parent to support the symmetric pivoting method described above. Therefore, as the first step towards deriving D^P from D^N, we alter the edge-set of the latter as follows. For each g from 1 to m (where m is the total number of supernodes), the smallest supernode h to which both $g \overset{\text{L}}{\rightsquigarrow} h$ and $g \overset{\text{U}}{\rightsquigarrow} h$ exist is designated as the LU-parent of g; that is, if an edge $g \rightarrow h$ does not exist, then an LU-edge $g \overset{\text{LU}}{\rightarrow} h$ is added to the data-DAG, or if an L- or a U-edge $g \rightarrow h$ exists, then it is converted to an LU-edge. Then, all edges $g \rightarrow k$ such that $k > h$ are deleted. If the original matrix is not reducible to a block-triangular form, then after this modification, each supernode other than the root supernode has an LU-parent to accommodate row-column pairs that fail to satisfy the pivoting criterion in their original locations [20]. It is easily seen that this modification has no effect on Theorems 4.1–4.3 because $g \overset{\text{L}}{\rightarrow} h$ ($g \overset{\text{U}}{\rightarrow} h$) is in the modified D^N only if $g \overset{\text{L}}{\rightarrow} h$ ($g \overset{\text{U}}{\rightarrow} h$) is in the original D^N as defined in section 4.4.





Multifrontal factorization without any pivot failure



Factorization with failure of pivot 1

FIG. 4.2. *An example factorization to show how the failure of pivot* 1 *is handled by a symmetric permutation of row and column* 1 *to merge them with their LU-parent supernode,* 4. *An* 'X' *denotes a nonzero in the coefficient matrix and a '+' denotes a fill-in. The circled* 'X' *and '+' are created due to pivoting. A* '0' *denotes a fill-in predicted by the original symbolic factorization that has a value of zero due to pivoting-related movement of rows and columns. The figure also shows that the absence of* $2 \xrightarrow{L} 4$ *leaves the entry* $U_{1,5}$ *unassembled from* C^2.

Figure 4.2 shows how the failure of pivot row and column 1 is attempted in the unsymmetric multifrontal factorization of a small 5×5 example matrix. Row-column 1 is symmetrically permuted to a new location adjacent to 1's LU-parent 4 in the data-DAG. This results in an addition of row index 1 to 1's U-parent 2 and an addition of column index 1 to 1's L-parent 3. Additionally, after moving to their new locations, row 1 in U and column 1 in L get fill-in in column and row positions where row 4 in U and column 4 in L have nonzeros (i.e., $U_{1,5}$, $L_{4,1}$, and $L_{5,1}$). Figure 4.2 also shows that after pivoting, the new row 1 of C^2 cannot be fully assembled in the absence of an L-edge $2 \xrightarrow{L} 4$. Clearly, in addition to adding LU-edges as described earlier, D^N requires further modifications in order to serve as a data-DAG for unsymmetric multifrontal

algorithm with dynamic pivoting.

Figure 4.3 gives another example of a factorization where D^N is unable to guide a complete assembly in the event of a pivot failure. Note that this example satisfies the first two conditions of Theorem 4.1. However, since it does not satisfy condition 3, no edges are added and the absence of $2 \xrightarrow{U} 8$ precludes a complete assembly of C^2 into its parents' frontal matrices when pivot 1 fails. We now state and prove a theorem that prescribes a modification of D^N to prevent the situation illustrated in Figure 4.3.

THEOREM 4.4. *If a column index $j \in$ Struct$(U_{i,*})$ satisfies all of the following conditions, then a U-edge $i \xrightarrow{U} j$ is necessary for C^i to be completely assembled into its parents' frontal matrices in the event of failure of pivot k.*

1. *The LU-parent of i is greater than j.*
2. *None of i's U-parents are in Struct$(U_{*,j})$.*
3. *A k exists such that there is a U-path $k \overset{U}{\rightsquigarrow} i$ in T^C and LU-parent$(k) > j$.*

The transpose of this theorem can be stated similarly.

Proof. Note that Theorem 4.4 is very similar to Theorem 4.1. The only difference is condition 3. If pivot k fails, then it will add a row in Struct$(L_{*,i})$ that corresponds to LU-parent(k) -1, which is the new location of k and is greater than $j-1$, the new index for j. Thus, the failure of pivot k transforms condition 3 of Theorem 4.4 into condition 3 for the applicability of Theorem 4.1, which has already been proved. □

Theorem 4.4 states that even if Struct$(L_{*,i})$ does not have any index greater than j but all other conditions for the applicability of Theorem 4.1 are satisfied and $i \xrightarrow{U} j$ is not present in the DAG, then pivoting may result in incomplete assembly unless this edge is added. This is because pivoting can create a nonzero entry $L_{k,i}$ such that $k > j$. This is what happens in the example shown in Figure 4.3 for $k = 1$, $i = 2$, and $j = 8$ in the original indices. Pivoting changes i, j, and k to 1, 7, and 8, respectively. In light of Theorem 4.4, we introduce another modification to D^N. Instead of using Theorem 4.1 strictly to derive D^N from T^C, we omit checking for condition 3 and derive D^N by adding all those edges to T^C (T^S in practice) that satisfy conditions 1 and 2.

Now, by means of Theorem 4.5, we will show that the data-DAG D^N, even after the modifications described above, is not sufficient to ensure complete assembly of all contribution matrices in the event of intersupernode pivoting. The reader can verify that Figure 4.2 illustrates the transpose case of Theorem 4.5 for $j = 1$, $i = 2$, $h = 4$, and $k = 5$. Finally, Theorem 4.6 will show that supplementing the data-DAG with additional edges prescribed by Theorem 4.5 makes it sufficient to handle all contribution matrices in the face of intersupernode pivoting. As we did earlier in this paper, for the sake of clarity and simplicity, we will state and prove Theorems 4.5 and 4.6 in the context of conventional DAGs with single-node supernodes. The results naturally extend to supernodal DAGs.

THEOREM 4.5. *If h is the LU-parent of j and all of the following conditions hold, then a U-edge $i \xrightarrow{U} h$ is necessary for C^i to be completely assembled into its parents' frontal matrices in the event that j fails to meet the pivot criterion in its original location.*

1. *There exists an L-path $j \overset{L}{\rightsquigarrow} i$ such that $i < h$ and LU-parent$(i) > h$.*
2. *None of i's U-parents are in Struct$(L_{*,j})$.*
3. *Either $\exists\, k \in$ Struct$(L_{*,i})$ such that $k > h$, or there is a U-path $k \overset{U}{\rightsquigarrow} i$ and LU-parent$(k) > h$.*

The transpose case can be stated similarly.

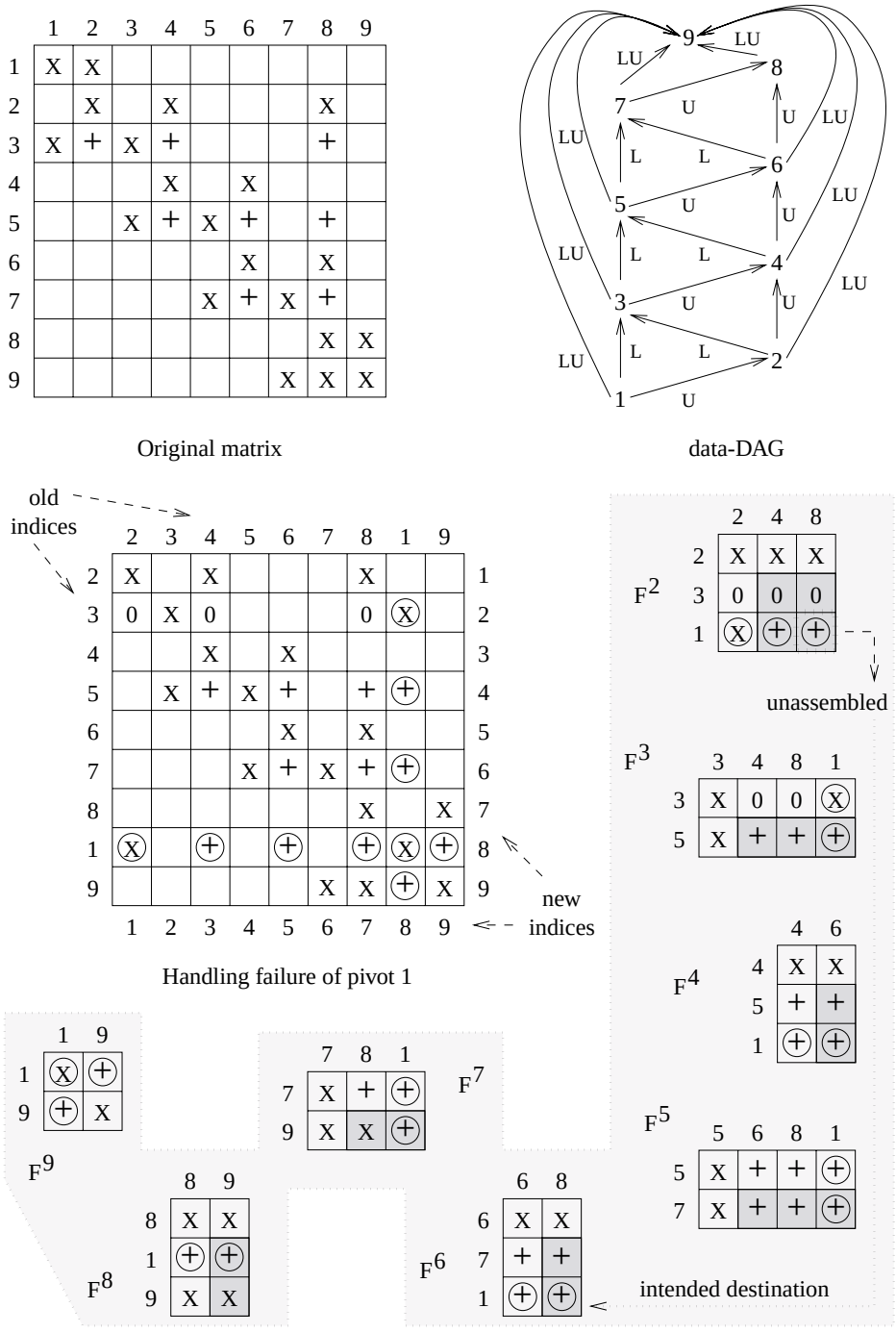


FIG. 4.3. *An example factorization to show that the edges in D^N are not sufficient to assemble C^2 into its parents' frontal matrices in the event of the failure of pivot* 1. *The convention for representing different types of structural nonzeros is the same as in Figure* 4.2.

Proof. If pivot j fails, then, along with the other failed LU-children of h, it occupies a new position just before h. Since there is an L-path $j \overset{\text{L}}{\rightsquigarrow} i$, column j is added to C^i after the failure of pivot j; that is, in the new matrix after pivoting, $j \in \text{Struct}(U_{i,*})$. We know that the LU-parent of i is greater than the new j, because LU-parent$(i) > h$. Since none of i's U-parents were in the old Struct$(L_{*,j})$, they are not in the new Struct$(U_{*,j})$ either. Thus the first two conditions for the applicability of Theorems 4.1 and 4.4 are satisfied. Condition 3 of Theorem 4.5 is equivalent to condition 3 of Theorems 4.1 and 4.4. Therefore, a U-edge $i \overset{\text{U}}{\rightarrow} j$ is needed for proper multifrontal factorization of the new matrix after permuting j to its new location. Since, in its new location, j is merged with h into a common supernode, a U-edge $i \overset{\text{U}}{\rightarrow} h$ in the original matrix would have sufficed. The proof of the transpose case is similar. ∎

THEOREM 4.6. *If D^P is a DAG formed by adding all possible edges according to Theorem* 4.5 *to D^N, then D^P is an adequate data-DAG for unsymmetric multifrontal factorization with potentially unlimited intersupernode pivoting.*

Proof. We prove this by showing that with D^P, it is not possible for any element of a contribution matrix C^i to remain unassembled. Without loss of generality, consider an element corresponding to $L_{k,j}$ in C^i. If $L_{k,j}$ is in C^i, then either $k \in \text{Struct}(L_{*,i})$ and $j \in \text{Struct}(U_{i,*})$ in the original L and U predicted by symbolic factorization, or row k or column j or both were added to C^i due to pivoting. If row k and column j are parts of the original structure of C^i, then Theorem 4.2 has already shown that the edge-set of D^N, which is a subset of the edge-set of D^P, is sufficient to assemble $L_{k,j}$. We now show that $L_{k,j}$ will be absorbed from C^i by one of i's parents in D^P when column j was added to C^i due to pivoting, irrespective of whether row k belonged to the original Struct$(L_{*,i})$ or if it too was added to C^i due to pivoting.

Let $g = \text{LU-parent}(i)$ and $h = \text{LU-parent}(j)$. We consider two cases: (1) $g \leq h$ and (2) $g > h$. If $g \leq h$, F^g will have both row k and column j and will absorb the element corresponding to $L_{k,j}$ from C^i. If $g > h$, then the first condition for the applicability of Theorem 4.5 has been satisfied. Now we consider two further scenarios: (2a) At least one of i's U-parents is in the original Struct$(L_{*,j})$, or (2b) none of i's U-parents is in the original Struct$(L_{*,j})$. In case of (2a), after pivoting, at least one of i's U-parents is in the new Struct$(U_{*,j})$ and the frontal matrix of this U-parent will absorb column j from C^i, including the entry corresponding to $L_{k,j}$. In case (2b), the second condition for the applicability of Theorem 4.5 has been satisfied. Finally, whether row k was in the original Struct$(L_{*,i})$ or was added to C^i due to the failure of a U-descendent k, in its final location, k must be greater than h. The reason is that if $j \leq k \leq h$ (i.e., k's new location is in the extended supernode h), then h must be an LU-ancestor of i because $j \leq k \leq h$ implies that there are both $i \overset{\text{L}}{\rightsquigarrow} h$ and $i \overset{\text{U}}{\rightsquigarrow} h$ in the data-DAG. But that is not possible because we are already working under the assumption that the LU-parent g of i is greater than h. Therefore, $k > h$ and the third condition of Theorem 4.5 has also been satisfied. As a result, Theorem 4.5 would have ensured that a U-edge $i \overset{\text{U}}{\rightarrow} h$ is present in D^P to assemble column j from C^i into F^h.

Similarly, we can prove that no entry corresponding to any $U_{j,k}$ will be left unassembled in C^i. ∎

4.6. Experimental results. In sections 4.4 and 4.5, we showed how to supplement the edge-set of the task-DAG to construct a data-DAG for the unsymmetric multifrontal algorithm. Table 4.1 shows experimental results of WSMP's

TABLE 4.1
Time required for constructing T^S, D^N, and D^P and the number of edges in each DAG.

Matrix	Symbolic		Supplement-1		Supplement-2		Total Time	
	t^S	$\lvert E_{T^S}\rvert$	t^1	$\lvert E_{D^N}\rvert$	t^2	$\lvert E_{D^P}\rvert$	$t^S+t^1+t^2$	$\frac{\lvert E_{D^P}\rvert}{\lvert E_{T^S}\rvert}$
af23560	.47	4793	.03	4794	.00	4794	0.50	1.00
av41092	.83	34708	.47	36346	.02	37092	1.32	1.07
bayer01	.28	87028	.13	95285	.05	96818	0.46	1.11
bbmat	1.7	6077	.06	6142	.00	6181	1.76	1.02
comp2c	.22	1736	.03	1929	.00	1978	0.25	1.14
e40r0000	.14	3225	.01	3264	.00	3264	0.15	1.01
e40r5000	.16	3182	.01	3235	.00	3237	0.17	1.02
ecl32	1.2	15239	.07	15244	.00	15260	1.27	1.00
epb3	.50	38088	.08	38173	.01	38300	0.59	1.01
fidap011	.42	1261	.02	1261	.00	1261	0.44	1.00
fidapm11	.65	2651	.03	2654	.00	2654	0.68	1.00
invextr1	.93	10108	.11	10813	.01	11244	1.05	1.11
mil053	4.5	166154	.51	166154	.07	166154	5.08	1.00
mixtank	1.2	3203	.05	3203	.00	3203	1.25	1.00
nasasrb	.97	3807	.05	3807	.00	3807	1.02	1.00
onetone1	.31	23585	.05	24523	.01	24691	0.37	1.05
onetone2	.18	23999	.04	24818	.01	24928	0.23	1.04
pre2	6.4	317216	.84	320063	.16	320942	7.40	1.01
raefsky3	.41	1281	.02	1281	.00	1281	0.43	1.00
raefsky4	.50	1358	.02	1358	.00	1358	0.52	1.00
rma10	.56	3911	.03	3911	.00	3911	0.59	1.00
tib	.07	10060	.01	10517	.00	10655	0.08	1.06
twotone	.91	44856	.12	45866	.01	45918	1.04	1.02
wang3old	.54	8450	.03	8450	.00	8450	0.57	1.00
wang4	.53	8253	.03	8253	.00	8253	0.56	1.00

implementation of the procedures to generate the various DAGs. Three DAGs are considered in Table 4.1: the supernodal task-DAG T^S, the supernodal data-DAG D^N for unsymmetric multifrontal factorization without pivoting, and the supernodal data-DAG D^P for unsymmetric multifrontal factorization with pivoting. The table shows the time to compute each of the DAGs and the number of edges in them for the 25 matrices in our test suite.

T^S is computed by the basic symbolic factorization algorithm described in section 3; therefore, t^S is the basic symbolic factorization time. We refer to the process of computing D^N from T^S as *Supplement*-1. Supplement-1 checks for the first two conditions of Theorem 4.1 to find the edges to be added to E_{T^S} and then adds outgoing LU-edges from supernodes without LU-parents to yield E_{D^N}. *Supplement*-2 is the process that adds edges based on the first two conditions of Theorem 4.5 to E_{D^N} to yield E_{D^P}. The execution time of Supplement-1 and Supplement-2 is denoted by t^1 and t^2, respectively.

Note that not all the edges in D^N and D^P are necessary. For the sake of computational speed, Supplement-1 and Supplement-2 in WSMP do not check for all the conditions of Theorems 4.1, 4.4, and 4.5 while adding edges. The last conditions of all three theorems are skipped. Even if all conditions of these theorems were checked, not all the edges in the resulting data-DAGs may be necessary. Therefore, D^N and D^P are not minimal data-DAGs for unsymmetric multifrontal factorization. However, as Table 4.1 shows, these DAGs do not have many more edges than T^S for most real-life matrices. The average for excess edges in supernodal D^P over T^S is only about 4% for our test suite. We have shown that the edges in the task-DAGs T^C or T^S are

insufficient to direct the data flow in unsymmetric multifrontal factorization. On the other hand, the edges in D^P are sufficient, even with pivoting. Therefore, the number of edges in a truly minimal supernodal data-DAG is somewhere between the number of edges in T^S and in the supernodal D^P. The experimental results in Table 4.1 show that these two numbers are usually fairly close. The table also shows that the time required to construct D^N and D^P is also small compared to the basic symbolic factorization time. Thus, the methodology described in this section for the construction of data-DAGs for unsymmetric multifrontal factorization is efficient in both time and the number of DAG edges. A comparison of the $t^S + t^1 + t^2$ column of Table 4.1 with WSMP's LU factorization time given in Table 5.1 shows that the total symbolic time is usually significantly less than the numerical factorization time.

5. Implementation details of unsymmetric factorization. A brief outline of the unsymmetric multifrontal algorithm based on the work of Hadfield [20] and Davis and Duff [9] is found in section 4.2. We now add more details to it and present a complete algorithm that is implemented in WSMP. WSMP is geared towards multiple factorizations of matrices with the same sparsity pattern but different nonzero values. Therefore, symbolic phase is performed only once and its output is used in all subsequent numerical factorizations, even with different pivot sequences resulting from different numerical values.

A fundamental data-structure in our unsymmetric multifrontal algorithm is the frontal matrix. A frontal matrix is associated with each supernode. Figure 5.1 shows the organization of a typical frontal matrix for a supernode $g = \sigma([q : r])$. The core of this frontal matrix is a | Struct$(L_{*,q})$ | $\times$ | Struct$(U_{q,*})$ | portion, where Struct$(L_{*,q})$ and Struct$(U_{q,*})$ are predicted by the symbolic factorization. In the absence of pivoting, the first $r - q + 1$ rows and columns of this matrix would be factored and would be saved as parts of U and L, respectively. The remaining trailing submatrix would constitute the contribution matrix whose contents would be absorbed into the frontal matrices of the parents of g in D^P.

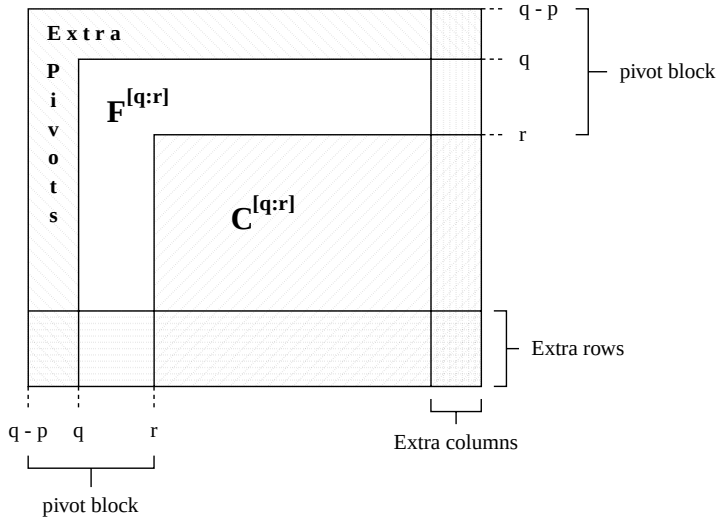


FIG. 5.1. *Organization of a typical frontal matrix for a supernode $g = \sigma([q : r])$. The p failed pivots from the LU-children of the supernode are appended at the beginning of the frontal matrix and the extra rows and columns inherited from U- and L-descendents, respectively, are appended at the end.*

In the presence of numerical pivoting, extra pivots as well as other rows and columns may be added to the frontal matrix depending on the labels and pivot failures of the children of g in D^P. Extra pivots (row-column pairs with the same indices) are added to F^g if some of the pivots of g's LU-children fail to satisfy the pivoting criterion. The LU-children of g themselves may have inherited some or all of these failed pivots from their own LU-children. Therefore, failed pivots from any of the LU-descendents of g can end up in its frontal matrix. If p such pivots are added, then the size of the pivot block increases from $r - q + 1$ to $r - q + p + 1$.

The frontal matrix F^g can similarly inherit extra rows corresponding to failed pivots in its U-descendents whose LU-parents are greater than g and extra columns corresponding to failed pivots in its L-descendents whose LU-parents are greater than g. Irrespective of their new indices, these extra rows and columns are always appended at the end of the original rows and columns of F^g and a sorted list of their indices is maintained at each supernode. Eventually, these are assembled into the extra pivots of the frontal matrices of the LU-parents of the supernodes where these pivots failed. The row and column structures predicted by symbolic factorization are kept intact for future factorizations of matrices with the same nonzero pattern. The additions to these structures due to pivoting, which depend on the nonzero values in the matrix being factored, are maintained separately and are discarded before each new factorization.

The availability of a static data-DAG D^P that is sufficient for handling an arbitrary amount of dynamic pivoting is critical to our implementation of the unsymmetric multifrontal algorithm. Figure 5.2 gives a high-level pseudocode of our factorization algorithm. The algorithm starts with the root supernode of task- and data-DAGs. At any supernode, first, it recursively factors all the unfactored children of that supernode. Then it looks at the failed pivots (if any) of its children to figure out the number and indices of the extra rows, columns, and pivots, if any, and accordingly allocates a frontal matrix of the appropriate size. In the next step, the contribution from the original coefficient matrix and the contribution matrices of the current supernode's children are accumulated in the appropriate locations inside the frontal matrix. Finally, the algorithm proceeds to factor the pivot block of the frontal matrix and updates the remainder of the frontal matrix. The leading successfully factored rows and columns are saved as portions of U and L for use during triangular solves. The remaining contribution matrix is eventually assembled into the frontal matrices of its parents and is released by the last parent to pick up its contribution.

The frontal matrix of the LU-parent of a supernode picks up all its failed pivot row-column pairs as well as the entire trailing submatrix of its contribution matrix with row and column indices greater than or equal to the first index of the parent supernode. The remaining rows and columns of a supernode's contribution matrix are assembled into the frontal matrices of its L- and U-parents in D^P. It is possible for more than one L- or U-parents' frontal matrices to have the same row or column indices in common with the child's contribution matrix. However, each element of a contribution matrix must be added into exactly one frontal matrix. Some simple bookkeeping to keep track of rows and columns that have been assembled suffices to ensure this condition for the relatively few rows and columns that have the potential to be copied into the frontal matrices of multiple L- and U-parents, respectively.

Figure 5.2 and the description in this section show that WSMP's unsymmetric multifrontal algorithm is fairly straightforward to implement. The static task- and data-DAGs computed during the symbolic phase and the use of recursion make the

```
function uns_mf (root) {
  /* 1. Recursive calls to root's children */
  for each child k of root in T^S do
    if not already processed k then
      Call uns_mf (k);
      Flag supernode k as already processed;
    end if
  end for
  /* 2. Collect pivoting info to determine size of F^root */
  for each child k of root in D^P do
    if k is an L-child then
      if k has failed pivots then
        Add them to the sorted list of F^root's extra columns;
      if C^k has extra columns then
        Add those whose LU-parent is greater than root to the
        sorted list of F^root's extra columns while checking for duplicates;
    else if k is a U-child then
      if k has failed pivots then
        Add them to the sorted list of F^root's extra rows;
      if C^k has extra rows then
        Add those whose LU-parent is greater than root to the
        sorted list of F^root's extra rows while checking for duplicates;
    else if k is an LU-child then
      if k has failed pivots then
        Add them to the sorted list of F^root's extra pivots;
      if C^k has extra columns then
        Add those whose LU-parent is greater than root to the
        sorted list of F^root's extra columns while checking for duplicates;
      if C^k has extra rows then
        Add those whose LU-parent is greater than root to the
        sorted list of F^root's extra rows while checking for duplicates;
    end if
  end for
  /* 3. Initialize root's frontal matrix */
  Allocate F^root of appropriate size and fill it with zeros;
  Populate F^root with entries from A corresponding to supernode root;
  /* 4. Assembly from children's contribution matrices into F^root */
  for each child k of root in D^P do
    Copy appropriate contribution from C^k into F^root;
    if root is the last parent of k to pick up C^k's contribution then
      Free the space occupied by C^k;
  end for
  /* 5. Numerical factorization */
  Factor the pivot block of F^root and update the trailing part to yield C^root;
end function uns_mf.
```

FIG. 5.2. *A simple and efficient unsymmetric multifrontal algorithm.*

numerical factorization algorithm much simpler to describe and implement than the earlier descriptions of the unsymmetric pattern multifrontal algorithm in [20] and [9]. Other than UMFPACK [8], WSMP is the only sparse solver available that is based on an unsymmetric pattern multifrontal algorithm. It is also the first such parallel solver available for general use. Although Hadfield [20] provided experimental results from a parallel implementation on the nCUBE, a practical parallel solver did not result from that effort.

The algorithm of Figure 5.2 is not only relatively simple in description but is also computationally lean because it minimizes the nonessential non–floating-point operations and can handle pivot failures fairly efficiently. It is also noteworthy that for structurally symmetric matrices, the algorithm in Figure 5.2 naturally reduces to a symmetric-pattern multifrontal algorithm guided by an elimination tree, which replaces both T^S and D^P. Other than a few "**if**" statements for each supernode, there is no overhead in using this algorithm for structurally symmetric matrices.

5.1. Experimental results. We now compare the unsymmetric LU factorization time of WSMP with that of three state-of-the-art multifrontal sparse solvers, namely, MUMPS version 4.1.6 [4, 5], MA41 [2, 3], and UMFPACK version 3.2 [8]. A detailed comparative study that includes more solvers can be found in [18]. The software compared in this section employ different variants of the multifrontal method. MUMPS contains a symmetric-pattern multifrontal factorization code based on the classical multifrontal algorithm [14]. MA41, in some sense, is a hybrid between symmetric and unsymmetric pattern multifrontal solvers. It uses an elimination tree to guide factorization, but the frontal matrices are pruned of all-zero rows and columns. UMFPACK 3.2 contains a variation of the unsymmetric-pattern multifrontal algorithm [9] that uses an elimination tree derived from the structure of $A'A$.

Apart from the factorization algorithm, there are other significant differences among the four software packages that affect their performance. First, they use different schemes for fill-reducing ordering. By default, WSMP uses a symmetric permutation based on a nested-dissection ordering [17] computed on the structure of $A+A'$. MUMPS and MA41 use a symmetric permutation based on the approximate minimum degree (AMD) algorithm [1] applied to the structure of $A+A'$. UMFPACK uses a column AMD algorithm [10] to prepermute only the columns of A and computes a row permutation based on numerical and sparsity criteria during factorization. The second difference is the use of a maximal matching algorithm [13] to permute the rows of the coefficient matrix to maximize the product of the magnitudes of its diagonal entries. As shown in [6, 18], this can affect factorization times because it changes the amount of structural symmetry and the amount of numerical pivoting during factorization. WSMP uses this preprocessing on all matrices, MUMPS and MA41 use it only if the structural symmetry in the original matrix is less than 50%, and UMFPACK does not use it at all. The third difference is that WSMP reduces the coefficient matrix into a block-triangular form, while MUMPS, MA41, and UMFPACK 3.2 do not.

Table 5.1 shows numerical factorization times and operation counts of MUMPS, MA41, UMFPACK, and WSMP run with the options in MUMPS, MA41, and WSMP changed to minimize the differences between the codes other than the factorization algorithm. We switched off the permutation to a heavy-diagonal form and the associated scaling in MUMPS, MA41, and WSMP. For WSMP, instead of its default nested-dissection ordering, we used an approximate minimum fill ordering, which is very similar to AMD. Even with these changes, differences remain between the four

TABLE 5.1
LU factorization times and operation counts of MUMPS, MA41, UMFPACK 3.2, and WSMP with similar permutation options. The best time is in boldface and the second best time is underlined.

Matrix	MUMPS		MA41		UMFPACK 3.2		WSMP	
	time (sec.)	ops $\times 10^9$	time (sec.)	ops $\times 10^9$	time (sec.)	ops $\times 10^9$	time (sec.)	ops $\times 10^9$
af23560	3.89	2.56	**3.58**	2.54	8.59	3.46	4.06	3.22
av41092	21.0	10.9	17.4	8.21	128.	30.1	**7.56**	3.38
bayer01	2.54	.697	1.51	.473	1.12	.024	**1.03**	.029
bbmat	54.3	41.6	56.3	41.1	78.7	39.1	**27.6**	21.5
comp2c	10.5	4.84	5.42	3.31	369.	113.	**3.79**	1.02
e40r0000	4.93	2.53	3.63	1.58	6.23	2.17	**0.80**	.419
e40r5000	14.5	5.43	29.9	1.74	6.76	2.09	**1.08**	.521
ecl32	**64.2**	64.6	67.1	64.4	191.	112.	139.	77.6
epb3	2.84	1.17	2.24	1.16	5.77	1.34	**1.70**	.547
fidap011	8.58	7.01	8.79	6.96	17.0	8.51	**6.51**	5.74
fidapm11	11.9	10.0	12.3	8.59	39.5	20.0	**7.29**	6.08
invextr1	**80.7**	71.5	82.1	34.2	178.	89.4	393.	92.9
mil053	43.5	31.8	40.0	31.8	107.	46.2	**28.2**	20.8
mixtank	151.	141.	152.	64.1	363.	243.	**76.3**	64.6
nasasrb	12.8	9.45	11.9	9.43	55.9	28.2	**10.4**	8.78
onetone1	17.1	8.19	12.6	4.86	5.85	2.33	**5.58**	3.57
onetone2	1.67	.605	1.17	.325	0.80	.080	**0.71**	.196
pre2	fail	fail	fail	fail	fail	fail	**346.**	301.
raefsky3	4.44	2.90	**3.88**	2.90	16.0	7.87	4.88	4.17
raefsky4	107.	74.4	92.9	44.7	**25.0**	12.9	43.4	22.5
rma10	4.00	1.39	2.89	1.38	8.83	3.44	**2.48**	1.31
tib	0.56	.122	0.37	.102	16.8	.203	**0.35**	.064
twotone	56.5	38.3	37.6	31.8	30.1	10.8	**2.87**	1.49
wang3old	72.9	57.8	57.7	51.0	**40.6**	24.2	45.8	32.3
wang4	11.8	10.5	12.2	10.5	53.4	30.7	**8.84**	7.94

codes. For instance, MUMPS, MA41, and WSMP first permute the matrix such that it has a diagonal of structural nonzeros. This initial permutation is the same for MUMPS and MA41 because both use the same code to compute it. However, it can be different for WSMP. The pivoting strategy of UMFPACK based on row interchanges is inherently different from the symmetric intersupernode pivoting strategy used in MUMPS, MA41, and WSMP. WSMP's algorithms work only with a permutation to the block-triangular form, which is not implemented in MUMPS, MA41, and UMFPACK. However, with the exception of *comp2c*, the effect of block-triangularization on the operation count for factorization is insignificant, if any. As a result of these differences and due to the fact that MUMPS may perform more operations than necessary on structurally unsymmetric matrices, the factorization operation counts for the four codes in Table 5.1 are different even with a similar ordering algorithm for fill-reduction.

In Table 5.1, the fastest factorization time for each matrix is in boldface and the second fastest time is underlined. Although differences other than the factorization algorithm itself affect the performance of these codes, it is easy to see the broad picture that emerges from Table 5.1. Most of the boldface entries are in the WSMP column and most of the underlined entries are in the MA41 column. For many matrices, the effect of the algorithmic choices of the software is evident in the factorization statistics in Table 5.1. MUMPS usually requires more floating-point operations for factorization than MA41 and WSMP because it uses artificially symmetrized frontal matrices padded with zeros. For the same reason, UMFPACK is faster than MUMPS

for very unsymmetric matrices (such as *bayer*01, *onetone*2, and *twotone*); however, it is slower for matrices with more structural symmetry (such as *fidap*011, *mil*053, and *wang*4) partly because it uses a fill-reducing permutation on the columns of the coefficient matrix before starting LU factorization. MA41 offers a significant improvement over MUMPS for matrices with a very unsymmetric structure, such as *comp*2*c*, *onetone*1, and *twotone*. It seems that MA41's mechanism for finding all-zero rows and columns incurs a slight overhead that it cannot offset for matrices with a nearly symmetric structure (such as *ecl*32, *fidap*011, and *wang*4), for which it is somewhat slower than MUMPS. It is clear from Table 5.1 that WSMP has the smallest overall factorization times even when its default options are modified to compare it with the other solvers. With its default options, WSMP's factorization times are usually much smaller [18] than those shown in Table 5.1.

6. Concluding remarks. This paper describes sparse unsymmetric symbolic and numerical factorization algorithms that improve previous similar algorithms. Our symbolic factorization phase, in particular, is more powerful than others described in the literature. It inexpensively computes minimal elimination structures that are transitive reductions of the upper and lower triangular factors of the original coefficient matrix. In addition, it computes near-minimal data-dependency DAGs for unsymmetric multifrontal factorization with and without pivoting. A data-DAG that has only a slightly higher number of edges than a minimal task-DAG and that is capable of expressing all possible data-dependencies in the face of dynamic pivoting is a key feature of our symbolic phase. We show how this data-DAG aids in a high-performance implementation of the unsymmetric multifrontal LU factorization algorithm. This factorization algorithm is not only faster than other sparse LU factorization algorithms but is also simpler than the unsymmetric multifrontal algorithms described previously in the literature.

Our algorithms do not introduce additional overheads while factoring matrices with a symmetric nonzero pattern. When presented with a sparse matrix with a symmetric structure, both the symbolic and the numerical factorization algorithms and the data-structures generated by them gracefully transform into their symmetric counterparts without requiring any significant amount of extra processing or storage.

In a distributed-memory parallel implementation of unsymmetric sparse LU factorization, the edges of the data-DAG connecting tasks mapped onto different processes determine the interprocess communication pattern. The static and near-minimal nature of the data-DAG used in our algorithms would be extremely useful for potential parallel implementations of unsymmetric multifrontal factorization, where changing the data-DAG dynamically could be cumbersome and inefficient and the unnecessary DAG edges could increase synchronization and communication overheads.

Acknowledgments. The author wishes to thank Andrew Conn, Fred Gustavson, Joseph Liu, Sivan Toledo, and the anonymous referees for extremely useful comments on earlier drafts of this paper.

REFERENCES

[1] P. R. Amestoy, T. A. Davis, and I. S. Duff, *An approximate minimum degree ordering algorithm,* SIAM J. Matrix Anal. Appl., 17 (1996), pp. 886–905.

[2] P. R. Amestoy and I. S. Duff, *Vectorization of a multiprocessor multifrontal code,* Internat. J. Supercomputer Appl., 3 (1989), pp. 41–59.

[3] P. R. Amestoy and I. S. Duff, *Memory management issues in sparse multifrontal methods on multiprocessors,* Internat. J. Supercomputer Appl., 7 (1993), pp. 64–82.

[4] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent, *A fully asynchronous multifrontal solver using distributed dynamic scheduling,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 15–41.

[5] P. R. Amestoy, I. S. Duff, and J. Y. L'Excellent, *Multifrontal parallel distributed symmetric and unsymmetric solvers,* Comput. Methods Appl. Mech. Engrg., 184 (2000), pp. 501–520.

[6] P. R. Amestoy, I. S. Duff, J. Y. L'Excellent, and X. S. Li, *Analysis and comparison of two general sparse solvers for distributed memory computers,* ACM Trans. Math. Software, 27 (2001), pp. 388–421.

[7] C. Ashcraft and R. G. Grimes, *The influence of relaxed supernode partitions on the multifrontal method,* ACM Trans. Math. Software, 15 (1989), pp. 291–309.

[8] T. A. Davis, *UMFPACK* V3.2: *An Unsymmetric-Pattern Multifrontal Method with a Column Pre-ordering Strategy,* Technical Report TR-02-2002, Computer and Information Sciences Department, University of Florida, Gainesville, FL, 2002.

[9] T. A. Davis and I. S. Duff, *An unsymmetric-pattern multifrontal method for sparse LU factorization,* SIAM J. Matrix Anal. Appl., 18 (1997), pp. 140–158.

[10] T. A. Davis, J. R. Gilbert, S. I. Larimore, and E. G.-Y. Ng, *A Column Approximate Minimum Degree Ordering Algorithm,* Technical Report TR-00-005, Computer and Information Sciences Department, University of Florida, Gainesville, FL, 2000.

[11] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu, *A supernodal approach to sparse partial pivoting,* SIAM J. Matrix Anal. Appl., 20 (1999), pp. 720–755.

[12] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Oxford University Press, Oxford, UK, 1990.

[13] I. S. Duff and J. Koster, *On algorithms for permuting large entries to the diagonal of a sparse matrix,* SIAM J. Matrix Anal. Appl., 22 (2001), pp. 973–996.

[14] I. S. Duff and J. K. Reid, *The multifrontal solution of unsymmetric sets of linear equations,* SIAM J. Sci. Statist. Comput., 5 (1984), pp. 633–641.

[15] S. C. Eisenstat and J. W. H. Liu, *Exploiting structural symmetry in unsymmetric sparse symbolic factorization,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 202–211.

[16] J. R. Gilbert and J. W. H. Liu, *Elimination structures for unsymmetric sparse LU factors,* SIAM J. Matrix Anal. Appl., 14 (1993), pp. 334–352.

[17] A. Gupta, *Fast and effective algorithms for graph partitioning and sparse matrix ordering,* IBM J. Res. Development, 41 (1997), pp. 171–183.

[18] A. Gupta, *Recent advances in direct methods for solving unsymmetric sparse systems of linear equations,* ACM Trans. Math. Software, 28 (2002), pp. 301–324.

[19] A. Gupta, G. Karypis, and V. Kumar, *Highly scalable parallel algorithms for sparse matrix factorization,* IEEE Trans. Parallel Distrib. Syst., 8 (1997), pp. 502–520.

[20] S. M. Hadfield, *On the LU Factorization of Sequences of Identically Structured Sparse Matrices within a Distributed Memory Environment*, Ph.D. thesis, University of Florida, Gainesville, FL, 1994.

[21] X. S. Li and J. W. Demmel, *A scalable sparse direct solver using static pivoting,* in Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing, SIAM, Philadelphia, 1999, CD-ROM.

[22] J. W. H. Liu, *The role of elimination trees in sparse factorization,* SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.

[23] J. W. H. Liu, *The multifrontal method for sparse matrix solution: Theory and practice,* SIAM Rev., 34 (1992), pp. 82–109.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 553–569

AN UNSYMMETRIZED MULTIFRONTAL LU FACTORIZATION*

PATRICK R. AMESTOY† AND CHIARA PUGLISI‡

Abstract. A well-known approach to computing the LU factorization of a general unsymmetric matrix $\mathbf{A}$ is to build the elimination tree associated with the pattern of the symmetric matrix $\mathbf{A}+\mathbf{A}^{\mathrm{T}}$ and use it as a computational graph to drive the numerical factorization. This approach, although very efficient on a large range of unsymmetric matrices, does not capture the unsymmetric structure of the matrices. We introduce a new algorithm which detects and exploits the structural asymmetry of the submatrices involved during the processing of the elimination tree. We show that with the new algorithm, significant gains, both in memory and in time, to perform the factorization can be obtained.

Key words. sparse linear equations, unsymmetric matrices, Gaussian elimination, multifrontal methods, elimination tree

AMS subject classifications. 65F05, 65F50

PII. S0895479800375370

1. Introduction.

We consider the direct solution of sparse linear equations based on a multifrontal approach. The systems are of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is an $n \times n$ unsymmetric sparse matrix. Duff and Reid [14, 15] developed the multifrontal method for computing the solution of indefinite sparse symmetric linear equations using Gaussian elimination and later extended it to solve more general unsymmetric matrices [16].

The multifrontal method belongs to a class of methods that separate the factorization into an analysis phase and a numerical factorization phase. The analysis phase involves a reordering step that reduces the fill-in during numerical factorization and a symbolic phase that builds the computational tree, the so-called *elimination tree* [10, 22, 24], whose structure gives the dependency graph of the multifrontal approach. The analysis phase is generally not concerned with numerical values and is based only on the sparsity pattern of the matrix.

As far as the analysis phase is concerned, the approaches introduced by Duff and Reid for both symmetric and unsymmetric matrices are almost identical. When the matrix is unsymmetric, the structurally symmetric matrix $\mathbf{M} = \mathbf{A} + \mathbf{A}^{\mathrm{T}}$, where the summation is performed symbolically, is used in place of the original matrix $\mathbf{A}$. The elimination tree of the unsymmetric LU factorization is thus identical to that of the Cholesky factorization of the symmetrized matrix $\mathbf{M}$.

To control the growth of the factors during the LU factorization, partial pivoting with a threshold criterion (see, for example, [11]) is used during the numerical factorization phase. The threshold value will define an interval in which pivots are acceptable. The pivot order, used during the analysis to build the elimination tree, might be modified during numerical factorization. Numerical pivoting can then result

*Received by the editors July 18, 2000; accepted for publication (in revised form) by S. A. Vavasis February 19, 2002; published electronically December 3, 2002. This research was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract DE-AC03-76SF00098.

http://www.siam.org/journals/simax/24-2/37537.html

†ENSEEIHT-IRIT, 2 rue Camichel, 31071 Toulouse, France, and NERSC, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720 (amestoy@enseeiht.fr).

‡NERSC, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720 (puglisi@enseeiht.fr).

in an increase in the estimated size of the factors and in the number of operations. To improve the numerical behavior of the multifrontal approach it is common to perform a step of preprocessing based on the numerical values. In fact if the matrix is not well scaled, which means that the entries in the original matrix do not have the same order of magnitude, a good prescaling of the matrix can have a significant impact on the accuracy and performance of the sparse solver. In some cases it is also very beneficial to precede the ordering by performing an unsymmetric permutation to place large entries on the diagonal. Duff and Koster [13] designed algorithms to permute large entries onto the diagonal and showed that it can very significantly improve the behavior of multifrontal solvers.

The multifrontal approach by Duff and Reid [16] is used in the Harwell Subroutine Library (HSL) code `ma41` [2, 3] and in the distributed memory code `MUMPS` developed in the context of the PARASOL project (EU ESPRIT IV LTR project 20160) [4, 5]. Another way to represent the symbolic LU factorization of a structurally unsymmetric matrix is to use directed acyclic graphs (see, for example, [17, 18]). These structures are more costly and complicated to handle than a tree, although they are better at capturing the asymmetry of the matrix. Davis and Duff [6] implicitly used this structure to drive their unsymmetric-pattern multifrontal approach.

We explain in this article how to use the simple elimination tree structure of the symmetric matrix $\mathbf{M}$ to detect, during the numerical factorization phase, structural asymmetry in the factors. We show that the new factorization phase has a very significant reduction in the computational time, the size of the LU factors, and the total memory requirements, as compared to the standard multifrontal approach [16]. In section 2, we first recall the main properties of the elimination tree and describe the standard multifrontal factorization algorithm. We then introduce the new algorithm and use a simple example to show the benefits that can be expected from the new approach. In section 3, our set of test matrices is introduced, and we analyze the performance gains (in terms of size of the factors, memory requirement, and factorization time) of the new approach with respect to the standard multifrontal code on these matrices. In section 4, we compare the performance of our approach with the unsymmetric-pattern multifrontal approach (`UMFPACK` [6, 7]) and with the supernodal partial pivoting code (`SuperLU` [9]). We add some concluding remarks in section 5.

2. Description of the multifrontal factorization algorithms. Let $\mathbf{A}$ be an unsymmetric matrix and let $\mathbf{M}$ denote its "symmetrized" form. The structure of $\mathbf{M}$ is $\mathrm{Struct}(\mathbf{A}) \cup \mathrm{Struct}(\mathbf{A}^{\mathrm{T}})$, where Struct() denotes the matrix pattern. Note that $\mathbf{M}$ has a symmetric structure and will contain several entries not present in $\mathbf{A}$. The elimination tree associated with the multifrontal factorization of $\mathbf{A}$ is computed performing a symbolic Cholesky factorization on $\mathbf{M}$. If the matrix $\mathbf{M}$ is reducible, then the tree will be a forest. Liu [22] defines the elimination tree as the transitive reduction of the directed graph of the Cholesky factors of $\mathbf{M}$. The characterization of the elimination tree and the description of its properties are beyond the scope of this article. In our context, we are interested in the elimination tree only as the computational graph for the multifrontal factorization. For a complete description of the elimination tree, the reader can consult [22, 23].

In the multifrontal approaches, we actually use an amalgamated elimination tree, referred to as the *assembly tree* [15], which can be obtained from the classical elimination tree. Each node of the assembly tree corresponds to Gaussian elimination operations on a full submatrix, called a *frontal matrix*. The frontal matrix can be partitioned as shown in Figure 1.

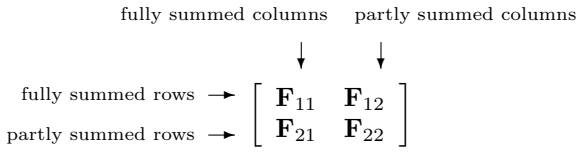


FIG. 1. *Partitioning of a frontal matrix.*

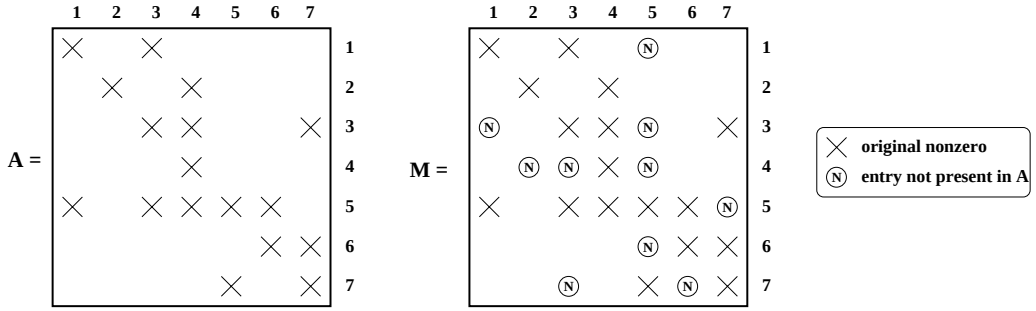


FIG. 2. *Patterns of an example matrix* $\mathbf{A}$ *and its "symmetrized" form* $\mathbf{M}$.

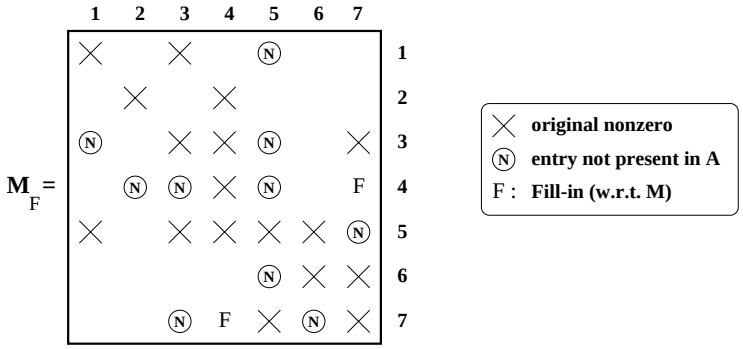


FIG. 3. *Structure of the Cholesky factors of the matrix* $\mathbf{M}$.

Each frontal matrix factorization involves the computation of a block of columns of $\mathbf{L}$, termed *fully summed columns* of the frontal matrix, a block of rows of $\mathbf{U}$, termed *fully summed rows*, and the computation of a Schur complement matrix $\mathbf{F}_{22} - \mathbf{F}_{21}\mathbf{F}_{11}^{-1}\mathbf{F}_{12}$, called a *contribution block*. The rows (columns) of the $\mathbf{F}_{22}$ block are referred to as *partly summed rows (columns)*.

The unsymmetric matrix $\mathbf{A}$, whose structure is shown on the left-hand side of Figure 2, will be used to illustrate the main properties of the assembly tree and to introduce the new algorithm. In Figures 2 and 3, an "$\times$" denotes a nonzero position from the original matrix $\mathbf{A}$ and a "$\textcircled{N}$" corresponds to an entry introduced by symmetrization. In Figure 3, we indicate the structure of the filled matrix $\mathbf{M}_{\mathrm{F}} = \mathbf{L} + \mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is the matrix of the Cholesky factor of $\mathbf{M}$. Entries with an "F" correspond to fill-in entries in the $\mathbf{L}$ factor.

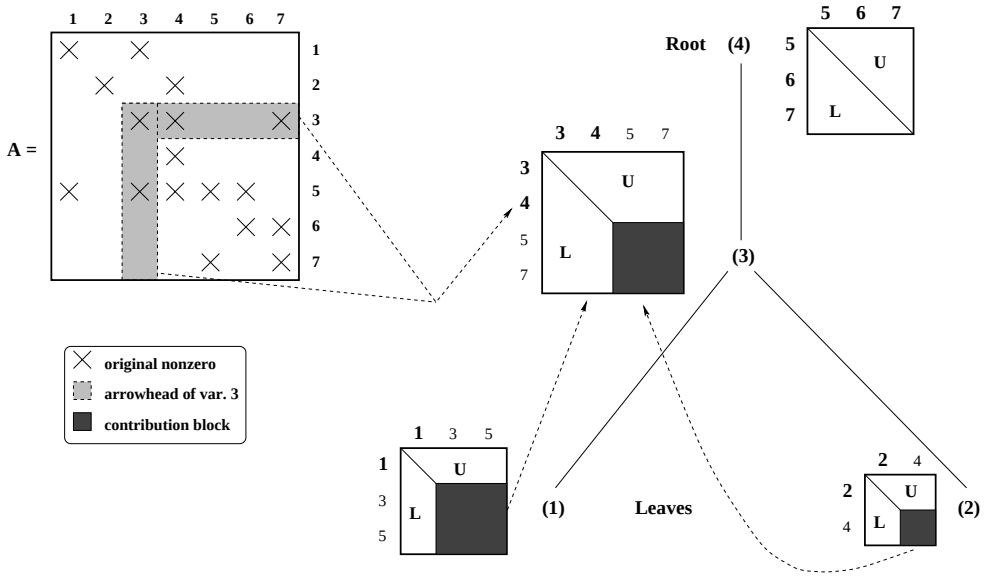


FIG. 4. *Assembly tree associated with our test matrix.*

The structure of the matrix $\mathbf{M}_{\mathrm{F}}$ is used to define the assembly tree (see Figure 4) associated with the multifrontal LU factorization of the matrix $\mathbf{A}$. From the fact that the factorization of $\mathbf{A}$ is based on the assembly tree associated with the symbolic Cholesky factorization of $\mathbf{M}$, we have

$$\mathrm{Struct}(\mathbf{M}_{\mathrm{F}}) = \mathrm{Struct}(\mathbf{L}) + \mathrm{Struct}(\mathbf{U}) \quad \text{and} \quad \mathrm{Struct}(\mathbf{L}^{\mathrm{T}}) = \mathrm{Struct}(\mathbf{U}).$$

Let us use the term *structural zero* to denote a numerical zero that does not result from numerical cancellation. Typically, due to the symmetrization process, the matrix $\mathbf{M}$ might contain many structural zeros that will propagate during the numerical factorization phase. In Figures 2 and 3, "(N)" corresponds to a structural zero in $\mathbf{M}$. What has motivated our work is the following question: Is it possible, during the processing of the assembly tree, to efficiently detect and remove structural zeros that appear in the matrix $\mathbf{M}_{\mathrm{F}}$ and that are a direct or indirect consequence of the symmetrization of the matrix $\mathbf{A}$? Although it is not so clear from the structure of the matrix $\mathbf{M}_{\mathrm{F}}$, we will show that blocks of structural zeros can be identified during the processing of the assembly tree.

In the following, we first describe how the assembly tree is exploited during the standard multifrontal algorithm. We then report and analyze the sparsity structure of the frontal matrices involved in the processing of the assembly tree associated with our example matrix. Based on these observations, we will introduce the new factorization algorithm.

The assembly tree is rooted (a node of the tree called the *root* is chosen to give an orientation to the tree) and is processed from the leaf nodes to the root node. If two nodes are adjacent in the tree, then the one nearer the root is the *parent node*, and the other is termed its *child*. Each edge of the assembly tree indicates a data dependency between parent and child. It involves sending a contribution block from the child to the parent. A parent node process will start when the processes associated with all of its children are completed.

For example, in Figure 4, node (3) must wait for the completion of nodes (1) and (2) before starting its computations. The subset of variables which can be used as pivots (boldface variables in Figure 4) are the *fully summed* variables of node (k). The contribution blocks of the children and the entries from the original matrix corresponding to the fully summed variables of node (k) are used to build the frontal matrix of the node. This will be referred to as the *assembly process*. During the assembly process of a frontal matrix, we need for each fully summed variable j to access the nonzero elements in the original matrix that are in rows/columns of indices greater than j. A way to efficiently access the original matrix is to store it in arrowheads according to the reordered matrix. For example, during the assembly process of node (3) the arrowheads of variables 3 and 4 from matrix $\mathbf{A}$ together with the contribution blocks of nodes (1) and (2) are used to assemble the frontal matrix of node (3). One should note that because the assembly tree is constructed by performing the symbolic Cholesky factorization of the symmetric matrix $\mathbf{M}$, the list of indices in the partly summed rows is identical to that of the partly summed columns (see row and column indices of block $\mathbf{F}_{22}$ in Figure 1). Therefore, during the assembly process, only the list of row indices of the partly summed rows is built. This list is obtained by merging all the row and column indices of the arrowheads of the matrix $\mathbf{A}$ with the row indices of the contribution blocks of all the children. Once the structure of the frontal matrix is built, the numerical values from both the arrowheads and the contribution blocks can be assembled at the right place in the frontal matrix. The floating-point operations involved during the assembly process will be referred to as *assembly operations* (only additions), whereas floating-point operations involved during the factorization of the frontal matrices will be referred to as *elimination operations*.

Partial pivoting with threshold is used to control the element growth in the factors. Note that pivots can be chosen only from within the block $\mathbf{F}_{11}$ of the frontal matrix. The LU factors corresponding to the fully summed variables are computed and a new contribution block is produced. When a fully summed variable of node (k) cannot be eliminated during the node process because of numerical considerations, then the corresponding arrowhead in the frontal matrix is added to the contribution block, and the fully summed variable will be included in the fully summed variables at the parent of node (k). This process creates additional fill-in in the LU factors.

In a multifrontal algorithm, we have to provide space for the frontal matrices and the contribution blocks and reserve space for storing the factors. We need working space to store both real and integer information. This will be referred to as the *maximum memory used* of the factorization phase. The same integer array can be used to describe a frontal matrix, its corresponding LU factors, and its contribution block. The management of the integer working array can thus be done in a simple and efficient way. In a uniprocessor environment, it is possible to determine the order in which the assembly tree will be processed. Furthermore, if we process the assembly tree with a depth first search order, we can use a stack to manage the storage of the factors and the contribution blocks. This mechanism is efficient both in terms of total memory requirement and amount of data movement (see [15]). A stack mechanism, starting from the beginning of the real working array, is used to store the LU factors. Another stack mechanism starting from the end of the real working array is used to store the contribution blocks. After the assembly phase of a node, the working space used by the contribution blocks of its children can be freed and, because the assembly tree is processed with a depth first search order, the contribution blocks will always be at the top of the stack. In the remainder of this paper, the maximum stack size of the contribution blocks will be referred to as the *maximum stack size*.

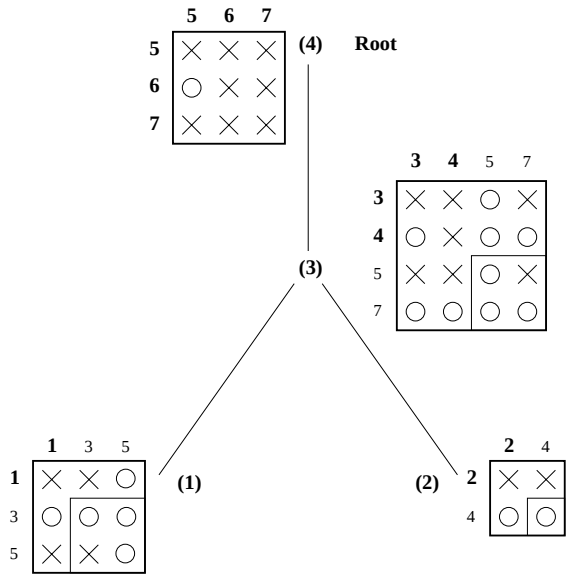


FIG. 5. *Processing the assembly tree associated with the matrix* **A**, *whose structure is showed in Figure* 2 *using the standard algorithm.*

2.1. The standard and new algorithms for multifrontal factorization. During a multifrontal factorization, each frontal matrix can be viewed as the minimum structure to perform the elimination of the fully summed variables and to carry the contribution blocks from all of its children. In Figure 5, we have a closer look at the frontal matrices involved in the processing of the assembly tree of Figure 4 to identify the structural zeros.

We report, beside each node, the structure of the factorized frontal matrix assuming that the pivots are chosen down the diagonal of the fully summed block in order (i.e., no numerical pivoting is required). An "×" corresponds to a nonzero entry and an "○" corresponds to a structural zero.

One can see, for our example, that the frontal matrices have many structural zeros. There are two kinds of structural zeros: those forming a complete zero column (or row), and the more isolated zero entries in a nonzero column or row (for example, entries (4,3) and (4,7) in the frontal matrix of node (3)). If one knows how to detect a partly summed row (or column) with only structural zeros, then the corresponding row (or column) can be suppressed from the frontal matrix because this row (or column) will not add any contribution to the father node.

Structural zero rows (or columns) can be detected during the assembly process of a frontal matrix because of the following property: If a row (or column) index does not appear in the row (or column) indices both of the arrowheads of the original matrix and of the contribution blocks of the children, then this index will correspond to a row (or column) with only structural zeros. This property is used to deduce the assembly process of the new algorithm. Note that if the matrix is not structurally deficient, then each fully summed row (or column) must have at least one nonzero entry. Therefore, we can restrict our search for zero rows (columns) to the partly summed rows (columns).

In the new assembly algorithm, the list of indices of the partly summed rows of

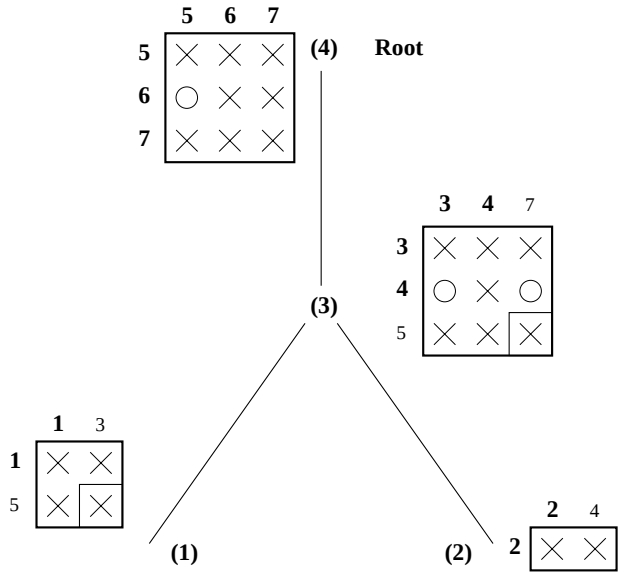


FIG. 6. *Processing the assembly tree associated with the matrix* **A**, *whose structure is showed in Figure* 2 *using the new algorithm.*

a frontal matrix is defined as the merge of the row indices in the arrowheads of the fully summed variables of the node with the row indices of the contribution blocks of its children. The column indices are defined similarly. As illustrated in Figure 6, the new assembly process can result in significant modifications in the processing of the assembly tree. For example, at node (1), row 3 and column 5 are suppressed from the frontal matrix; at node (2), all the partly summed rows are suppressed; at node (3), row 7 and column 5 are suppressed. Note that zero rows (columns) do not result only from propagation of zero structures detected at the leaf nodes (see, for example, row 7 at node (3)). As can be seen in Figure 6, the frontal matrices may become unsymmetric in structure. (Note, however, that we do not fully exploit the sparsity structure of the frontal matrices.)

We finally indicate in Figure 7 the structure of the LU factors obtained with the new algorithm. This should be compared to the matrix $\mathbf{M}_{\mathrm{F}}$ in Figure 3 showing the structure of the factors obtained with the standard algorithm.

It can be seen that nonzero entries corresponding to fill-in (for example, (7,4) in $\mathbf{M}_{\mathrm{F}}$) or introduced in $\mathbf{M}$ by the symmetrization process (for example, (4,5) in $\mathbf{M}_{\mathrm{F}}$) might be suppressed by the new algorithm. On the other hand, the new algorithm will never suppress structural zeros in a block of fully summed variables (for example, (4,3) in node (3) of Figure 5 and 6). On our small example, the total number of entries in the factors reduces from 29 to 22.

Comparing Figures 5 and 6, we see that the new algorithm might also lead to a significant reduction in both the number of operations involved during the assembly process and the maximum stack size. The latter, combined with a reduction in the size of the factors, will result in a reduction in the maximum memory used. In our example, the number of assembly operations drops from 26 (17 entries from **A** plus 9 from the contribution blocks) to 19 (2 entries from the contribution blocks). The

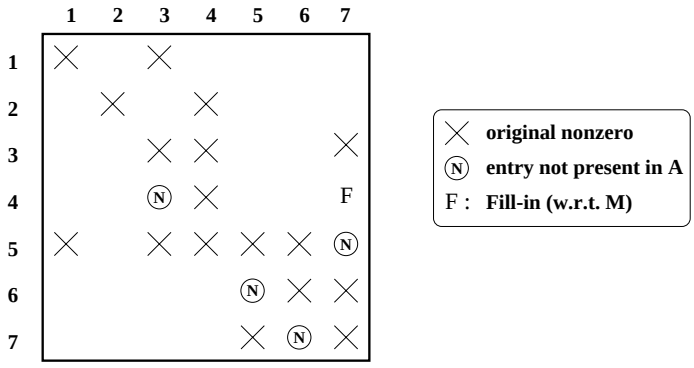


FIG. 7. *Structure of the LU factors obtained using the new algorithm.*

maximum stack size reduces from 5 to 1 (obtained in both cases after stacking the contribution blocks of nodes (1) and (2)).

3. Results and performance analysis. We describe in Table 1 the set of test matrices (order, number of nonzero entries, structural symmetry, and origin). We define the structural symmetry as the percentage of the number of nonzeros matched by nonzeros in symmetric locations over the total number of entries. A symmetric matrix has a value of 100. Although our performance analysis will focus on matrices with a relatively small structural symmetry, all classes of unsymmetric matrices are represented in this set. The selected matrices come from the forthcoming Rutherford–Boeing Sparse Matrix Collection [12],[1] the Tim Davis collection,[2] and SPARSKIT2.[3]

The HSL [20] code `ma41`[4] has been used to obtain the results for the standard multifrontal method. The factorization phase of `ma41` was then modified to use the new algorithm. The modified version of the code will be available in the next release of the HSL and is available under the same conditions as the original `ma41` code. The `ma41` code has a set of parameters to control its efficiency. We have used the default values for our target computer (SGI Cray Origin 2000). The approximate minimum degree ordering (`AMD` [1]) has been used to reorder the matrix. As we have mentioned in the introduction, it is often quite beneficial for very unsymmetric matrices to precede the ordering by performing an unsymmetric permutation to place large entries on the diagonal and then scaling the matrix so that the diagonal entries are all of modulus one and the off-diagonals have modulus less than or equal to one. We use the HSL code `mc64` [13] to perform this preordering and scaling on all matrices of structural symmetry smaller than 55. When `mc64` is not used, our matrices are always row and column scaled (each row/column is divided by its maximum value). All results presented in this section have been obtained on one processor (R10000 MIPS RISC 64-bit processor) of the SGI Cray Origin 2000 from Parallab (University of Bergen, Norway). The processor runs at a frequency of 195 Mhertz and has a peak performance of 390 Mflops per second.

In the following graphs, we show the performance ratio of the new factorization algorithm over the standard algorithm. Apart from Table 1, in all figures and tables of

[1] http://www.cse.clrc.ac.uk/Activity/SparseMatrices/
[2] http://www.cise.ufl.edu/research/sparse/matrices
[3] http://iftp.cs.umn.edu/pub/sparse/
[4] Available from the authors or at http://www.cse.clrc.ac.uk/Activity/HSL.

TABLE 1

Test matrices. Structural symmetry: 100 ≡ *symmetric. Original: structural symmetry of the original matrix. Modified: structural symmetry after permutation when* `mc64` *is used.*

Matrix name	Order	No. entries	Struct. symmetry Original	(modified)	Origin (discipline)
AV4408	4408	95752	0	(15)	Vavasis (Partial diff. eqn.) [25]
AV41092	41092	1683902	0	(8)	Vavasis (Partial diff. eqn.) [25]
BBMAT	38744	1771722	54	(50)	Rutherford–Boeing (CFD)
CAVITY15	2597	76367	94		SPARSKIT2 (CFD)
CAVITY26	4562	138187	95		SPARSKIT2 (CFD)
EX11	16614	1096948	100		SPARSKIT2 (CFD)
FIDAPM11	22294	623554	100		SPARSKIT2 (CFD)
GOODWIN	7320	324784	64		Davis (CFD)
LHR02	2954	37206	1	(17)	Davis (Chemical engineering)
LHR14C	14270	307858	1	(15)	Davis (Chemical engineering)
LHR17C	17576	381975	0	(20)	Davis (Chemical engineering)
LHR34C	35152	764014	0	(19)	Davis (Chemical engineering)
LHR71C	70304	1528092	0	(21)	Davis (Chemical engineering)
LNS_3937	3937	25407	87		Rutherford–Boeing (CFD)
OLAF1	16146	1015156	100		Davis (Structural engineering)
ONETONE1	36057	341088	10	(43)	Davis (Circuit simulation)
ONETONE2	36057	227628	15	(57)	Davis (Circuit simulation)
ORANI678	2529	90158	7	(9)	Rutherford–Boeing (Economics)
PSMIGR_1	3140	543162	48	(48)	Rutherford–Boeing (Demography)
RAEFSKY5	6316	168658	4	(4)	Davis (Structural engineering)
RAEFSKY6	3402	137845	2	(2)	Davis (Structural engineering)
RDIST1	4134	94408	6	(49)	Rutherford–Boeing (Chemical engineering)
RIM	22560	1014951	65		Davis (CFD)
SHERMAN5	3312	20793	78		Rutherford–Boeing (Oil reservoir simul.)
SHYY161	76480	329762	77		Davis (CFD)
SHYY41	4720	20042	77		Davis (CFD)
TWOTONE	120750	1224224	28	(43)	Davis (Circuit simulation)
UTM3060	3060	42211	56		SPARSKIT2
UTM5940	5940	83842	56		SPARSKIT2
WANG4	26068	177196	100		Rutherford–Boeing (Semiconductor)

this section, matrices will be sorted by increasing structural symmetry of the matrix to be factored, i.e., after application of the column permutation when `mc64` is used. We use the same matrix order in the graphs and in the complete set of results provided in Tables 2 and 3 in order to easily find, given a point in the graph, its corresponding entry in the tables.

On the complete set of test matrices, we first study in Figure 8 what is probably of main concern for the user of a sparse solver, i.e., the time needed to factor the matrix and the total memory used. We recall that the memory used includes the storage of both the integer (4 bytes per integer entry) and the reals (8 bytes per real). In Figure 8, we divide the matrices into three categories: matrices of structural symmetry smaller than 50 for which the time reduction is between 20% and 80%, matrices whose structural symmetry is between 50 and 80 for which the time reduction is between 3% and 20% (see horizontal lines), and nearly structurally symmetric matrices for which there is almost no difference between the standard and new version. It is interesting to notice that even on symmetric matrices, the added work to detect asymmetry does not have much of an effect on the performance of the factorization. In the remainder of this section, we will not consider further the results on almost structurally symmetric matrices (symmetry greater than 80).

TABLE 2
Comparison of the standard (Stnd) and the new algorithms on matrices of structural symmetry < 50. *Timings are in seconds.*

Matrix (Str.Sym.)	Version	LU (number of real entries)	Stack (number of real entries)	Mem. used (Kbytes)	Operations during Elimin.	Operations during Assemb.	Facto. Time
RAEFSKY6	Stnd	1509016	606458	16356	4.795E+08	4.347E+06	2.14
(2)	New	998064	145575	9173	2.313E+08	1.049E+06	0.99
RAEFSKY5	Stnd	1757680	378792	17168	3.746E+08	3.874E+06	1.75
(4)	New	1226376	172619	11488	2.081E+08	1.459E+06	0.98
AV41092	Stnd	13977898	3877302	130684	7.798E+09	4.697E+07	38.23
(8)	New	10629026	1880351	96707	4.380E+09	2.376E+07	19.62
ORANI678	Stnd	422713	5482454	46599	9.012E+07	8.122E+06	1.20
(9)	New	304199	457312	6473	5.179E+07	7.633E+05	0.31
AV4408	Stnd	552360	227787	5722	6.914E+07	1.453E+06	0.45
(15)	New	439648	105157	4396	4.713E+07	7.740E+05	0.32
LHR14C	Stnd	2166692	414756	20457	2.091E+08	7.942E+06	1.72
(15)	New	1747085	144216	16169	1.432E+08	3.114E+06	1.11
LHR02	Stnd	230116	135059	2930	1.255E+07	7.616E+05	0.15
(17)	New	174145	23073	1827	7.508E+06	2.543E+05	0.09
LHR34C	Stnd	5618356	753854	52335	6.284E+08	2.066E+07	4.99
(19)	New	4534033	279728	41760	4.362E+08	8.336E+06	3.14
LHR17C	Stnd	2833254	642175	26957	3.155E+08	1.093E+07	2.58
(20)	New	2296194	192785	21331	2.228E+08	4.425E+06	1.69
LHR71C	Stnd	11657690	731783	106686	1.417E+09	4.317E+07	10.66
(21)	New	9400102	258847	85395	9.711E+08	1.783E+07	6.82
TWOTONE	Stnd	22085646	15899616	283957	2.933E+10	2.171E+08	155.20
(43)	New	17004114	5344194	182298	1.838E+10	6.993E+07	80.27
ONETONE1	Stnd	4713485	3348215	52076	2.282E+09	2.675E+07	13.42
(43)	New	3918207	1434965	40719	1.660E+09	1.198E+07	7.78
PSMIGR_1	Stnd	6316254	12896617	148636	9.313E+09	8.326E+07	54.30
(48)	New	6075412	8587331	117372	8.889E+09	5.087E+07	46.08
RDIST1	Stnd	258096	53767	2509	8.150E+06	5.054E+05	0.13
(49)	New	227436	7865	2234	6.504E+06	3.507E+05	0.10

In Figure 9, we relate the gain in the factorization time to the reduction in the number of elimination operations and in the number of assembly operations. Although the number of operations due to the assembly is always much smaller than the number of operations involved during factorization (see Tables 2 and 3), the assembly process can still represent a significant part of the time spent in the factorization phase (see, for example, [2]). This is illustrated in Figure 9, where we see that the high reduction in the number of assembly operations (more than 50%) significantly contributes to reducing the factorization time. Note that, on a relatively large matrix (ONETONE1) of symmetry 57 (after permutation based on `mc64`), significant gains in time and in the number of assembly operations (more than 40%) can still be obtained. In Figure 10, we relate the memory reduction to the size of the factors and to the maximum stack size. Although a reduction in the maximum size of the stack might not always introduce a reduction in the total memory used, we see that in practice it is often the case. An extreme example of this reduction is matrix ORANI678 of symmetry 9 (see Table 2) for which the maximum stack size is reduced by more than one order of magnitude (5482454 to 457312). Finally, we notice that a large reduction in the maximum stack size (Figure 10) is generally correlated with a large reduction in the number of assembly operations (Figure 9).

4. Comparison with other unsymmetric solvers. In this section, we compare the `ma41` codes with three other sparse unsymmetric solvers. The aim of this

TABLE 3
Comparison of the standard (Stnd) and the new algorithms on matrices of structural symmetry ≥ 50. *Timings are in seconds.*

Matrix (Str.Sym.)	Version	LU (number of real entries)	Stack (number of real entries)	Mem. used (Kbytes)	Operations during Elimin.	Operations during Assemb.	Facto. Time
BBMAT	Stnd	44107862	8352717	386786	3.675E+10	2.283E+08	162.76
(50)	New	41078591	6786540	358434	3.246E+10	1.655E+08	137.73
UTM3060	Stnd	324640	78679	3287	2.683E+07	6.973E+05	0.20
(56)	New	309700	70390	3135	2.486E+07	6.198E+05	0.19
UTM5940	Stnd	701496	131839	6844	6.640E+07	1.529E+06	0.47
(56)	New	675079	124245	6590	6.264E+07	1.391E+06	0.45
ONETONE2	Stnd	2253553	898540	24776	5.085E+08	7.627E+06	3.20
(57)	New	1858514	495096	20206	3.624E+08	4.069E+06	1.98
GOODWIN	Stnd	1264140	307777	11560	1.612E+08	2.841E+06	0.95
(64)	New	1217726	283920	11089	1.505E+08	2.503E+06	0.90
RIM	Stnd	4127204	833290	36437	5.648E+08	9.193E+06	3.23
(65)	New	3973769	780826	35070	5.221E+08	8.209E+06	3.14
SHYY161	Stnd	7437816	377535	70254	9.945E+08	1.209E+07	6.13
(77)	New	7204304	355293	68220	9.583E+08	1.114E+07	6.03
SHYY41	Stnd	251336	28523	2642	1.036E+07	3.240E+05	0.13
(77)	New	239696	26015	2536	9.681E+06	2.843E+05	0.12
SHERMAN5	Stnd	167412	61227	1972	1.284E+07	4.413E+05	0.13
(78)	New	148176	44670	1735	1.029E+07	3.131E+05	0.10
LNS_3937	Stnd	285517	89578	2946	1.920E+07	5.483E+05	0.21
(87)	New	284874	89390	2947	1.914E+07	5.463E+05	0.19
CAVITY15	Stnd	202629	33004	2009	1.033E+07	3.452E+05	0.10
(94)	New	197553	31796	1959	9.896E+06	3.319E+05	0.10
CAVITY26	Stnd	394164	58589	3874	2.433E+07	6.877E+05	0.21
(95)	New	386700	57061	3800	2.358E+07	6.670E+05	0.21
EX11	Stnd	11981558	3960507	109987	6.678E+09	3.835E+07	27.78
(100)	New	11981558	3960507	109992	6.678E+09	3.835E+07	27.84
FIDAPM11	Stnd	15997220	4863371	154021	9.599E+09	4.705E+07	39.68
(100)	New	15997220	4863371	154041	9.599E+09	4.705E+07	39.74
OLAF1	Stnd	5880174	1506068	55402	1.965E+09	1.684E+07	8.83
(100)	New	5880174	1506068	55407	1.965E+09	1.684E+07	8.89
WANG4	Stnd	11561486	5063375	128968	1.048E+10	4.087E+07	42.70
(100)	New	11561486	5063375	129016	1.048E+10	4.087E+07	42.82

comparison is merely to show that the new version of the `ma41` code is competitive on a large class of matrices, including matrices very unsymmetric in structure. A comprehensive comparison of the codes is beyond the scope of this article. We compare the two versions of `ma41` with the unsymmetric pattern multifrontal code (`UMFPACK`2.2[5] [6, 7]) and with the supernodal partial pivoting code (`SuperLU`2.2[6] [9]). We also report results obtained with the new release of `UMFPACK` (`UMFPACK`3.0, April 30, 2001).

`UMFPACK`2 is an unsymmetric pattern multifrontal code in which unifrontal and multifrontal schemes are combined. It first tries to permute the matrix to a block triangular form [11] but does not use any other preordering of the matrix and does not perform a symbolic factorization of the matrix. Rather, pivots are chosen during numerical factorization to balance considerations of stability and sparsity by using approximate Markowitz counts with a pivot threshold. A directed acyclic graph is implicitly used to drive the numerical factorization. In both `SuperLU` and `UMFPACK`3, a preordering of the columns is computed. The column elimination tree [19] is used

[5]Available at http://www.cise.ufl.edu/research/sparse. A fully supported library version which is functionally equivalent is `ma38` in HSL.

[6]Available at http://www.nersc.gov/∼xiaoye/SuperLU.

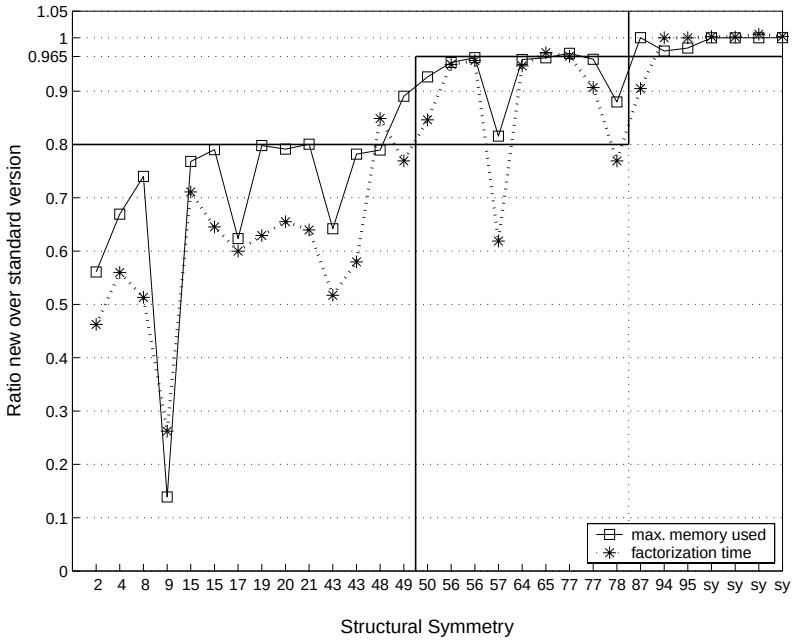


FIG. 8. *Study of the factorization time and the maximum memory used. sy, on the x-axis, corresponds to structurally symmetric matrices.*

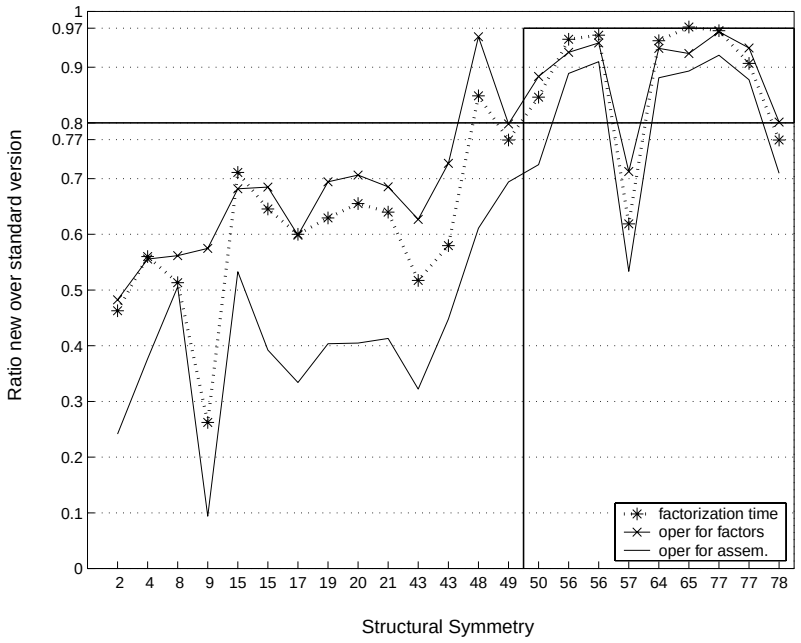


FIG. 9. *Impact of the reduction in the number of floating-point operations on the time.*

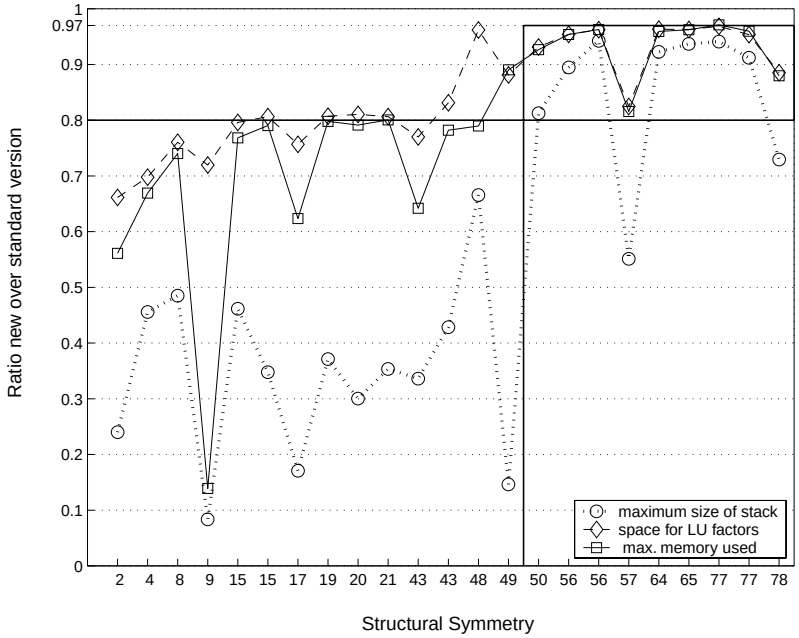


FIG. 10. *Correlation between the factor size, the maximum stack size, and the maximum memory used.*

during the analysis phase to get memory estimates and during the factorization phase to drive the computations. All solvers factorize general unsymmetric matrices and use dense matrix kernels. In each code, fill-in reduction has been set to use some variant of the minimum degree ordering algorithm (the approximate minimum degree (`AMD` [1]) for `ma41`, the unsymmetric approximate minimum degree [6] for `UMFPACK`2, and the column approximate minimum degree (`COLAMD`[7] [8]) on $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, without forming $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ explicitly, for both `SuperLU` and `UMFPACK`3). Finally, each code uses partial pivoting with a threshold value which has been set to 0.01.

We limit our study to matrices for which the factorization time using the standard `ma41` code is greater than two seconds. This leads to seven matrices of structural symmetry smaller than 50 and to seven matrices of structural symmetry higher than 50. The results, shown in Tables 4 and 5, include the number of entries in the LU factors, the maximum memory effectively used (in millions of bytes), the number of floating-point operations during the factorization phase (in millions), and the total time (in seconds) to solve a system of equations once (i.e., the analysis phase when necessary, the factorization phase, and the solve phase excluding iterative refinement).

One should first mention that with highly reducible matrices, permuting the matrix to a block triangular form can have a serious impact on the performance. (Compare results obtained with `UMFPACK`2 and with the other codes on the LHR and RAEFSKY6 matrices.) Note that the RAEFSKY6 matrix is triangular after permutation to block triangular form. As we might have expected, on matrices with structural symmetry higher than 50, the new `ma41` code is in general better than all other codes in all

[7] Available at http://www.cise.ufl.edu/research/sparse.

TABLE 4
Comparison with `SuperLU` *and* `UMFPACK` *on matrices with structural symmetry* < 50.

Matrix	Code	LU factors (number of entries)	Mem. used (Mbytes)	Oper. count (10^6)	Total time (seconds)
RAEFSKY6	Stnd	1509016	16	484	2.42
	New	998064	9	232	1.25
	SuperLU	373270	5	14	0.74
	UMFPACK2	137845	4	0	0.09
	UMFPACK3	410320	7	25	1.08
AV41092	Stnd	13977898	131	7845	79.38
	New	10629026	97	4404	48.19
	SuperLU	43656157	458	75531	858.57
	UMFPACK2	32999672	467	48444	405.11
	UMFPACK3	36595005	368	29651	441.19
LHR34C	Stnd	5618356	52	649	13.57
	New	4534033	42	445	11.48
	SuperLU	3589899	46	239	8.38
	UMFPACK2	3299425	43	224	5.76
	UMFPACK3	2956721	28	174	9.30
LHR17C	Stnd	2833254	27	326	5.70
	New	2296194	21	227	4.80
	SuperLU	1732026	22	111	3.98
	UMFPACK2	1516902	20	94	2.65
	UMFPACK3	1442412	14	82	4.43
LHR71C	Stnd	11657690	107	1460	32.70
	New	9400102	85	989	24.97
	SuperLU	7314370	94	506	17.48
	UMFPACK2	6516265	79	451	11.82
	UMFPACK3	5914747	56	357	19.44
TWOTONE	Stnd	22085646	284	29551	162.94
	New	17004114	182	18452	87.66
	SuperLU	21164814	261	8850	198.43
	UMFPACK2	9967842	265	9242	91.49
	UMFPACK3	15180013	164	10778	126.91
ONETONE1	Stnd	4713485	52	2308	15.29
	New	3918207	41	1672	9.59
	SuperLU	4857094	58	2859	29.64
	UMFPACK2	4741621	96	2352	20.37
	UMFPACK3	3749921	47	2080	19.81
PSMIGR_1	Stnd	6316254	149	9396	57.00
	New	6075412	117	8940	48.77
	SuperLU	8668085	88	16650	159.52
	UMFPACK2	5758871	329	8134	66.33
	UMFPACK3	6800169	75	10071	235.83

respects (even if, on this class of matrices, the total time and the memory used by `UMFPACK3` have been very significantly reduced with respect to `UMFPACK2`). On highly unsymmetric matrices (symmetry smaller than 50), the new `ma41` code is often comparable in terms of total time to the best unsymmetric solver (and this even if the number of operations remains sometimes higher). Finally, it is interesting to notice in Table 4 that the new `ma41` code is competitive in terms of the size of LU factors and of the maximum memory effectively used.

5. Concluding remarks. We have described a modification of the standard multifrontal LU factorization algorithm that can lead to a significant reduction in both the computational time and the maximum memory used. The standard multifrontal algorithm [16] for unsymmetric matrices is based on the assembly tree of a

TABLE 5
Comparison with `SuperLU` *and* `UMFPACK` *on matrices with structural symmetry* ≥ 50.

Matrix	Code	LU factors (number of entries)	Mem. used (Mbytes)	Oper. count (10^6)	Total time (seconds)
BBMAT	Stnd	44107862	387	36983	170.05
	New	41078591	358	32630	144.88
	SuperLU	48531243	525	42623	615.55
	UMFPACK2	140169268	1619	315717	2261.67
	UMFPACK3	43811565	394	37341	341.19
ONETONE2	Stnd	2253553	25	516	4.31
	New	1858514	20	366	2.91
	SuperLU	1153266	20	95	2.80
	UMFPACK2	1235606	30	140	4.22
	UMFPACK3	879719	12	76	3.08
RIM	Stnd	4127204	36	574	4.58
	New	3973769	35	530	4.87
	SuperLU	19149809	217	6857	93.06
	UMFPACK2	19644383	256	7336	59.03
	UMFPACK3	19367669	172	7798	56.82
SHYY161	Stnd	7437816	70	1007	7.61
	New	7204304	68	969	7.50
	SuperLU	6081407	78	1029	13.77
	UMFPACK2	9058119	159	4737	33.97
	UMFPACK3	5409691	55	1011	11.13
EX11	Stnd	11981558	110	6716	29.52
	New	11981558	110	6716	29.56
	SuperLU	14788186	157	7826	90.38
	UMFPACK2	38912590	493	56244	368.34
	UMFPACK3	13821878	128	6632	54.33
FIDAPM11	Stnd	15997220	154	9646	41.32
	New	15997220	154	9646	41.37
	SuperLU	25580332	275	20965	285.34
	UMFPACK2	72540520	1580	156717	1107.53
	UMFPACK3	24028406	243	19506	151.44
OLAF1	Stnd	5880174	55	1981	10.19
	New	5880174	55	1981	10.24
	SuperLU	7159053	77	2062	36.94
	UMFPACK2	7366098	122	2684	17.82
	UMFPACK3	6992964	62	2003	15.59
WANG4	Stnd	11561486	129	10525	43.91
	New	11561486	129	10525	44.04
	SuperLU	26220584	268	33726	323.18
	UMFPACK2	43489529	675	90296	602.18
	UMFPACK3	23450633	271	29518	188.47

symmetrized matrix and involves frontal matrices that are symmetric in structure. Therefore, it produces LU factors such that the matrix $\mathbf{F} = \mathbf{L} + \mathbf{U}$ is symmetric in structure. This approach is currently used in the context of two publicly available packages (`ma41` [2, 3] and `MUMPS`[8] [5, 4]) and has the advantage, with respect to other unsymmetric factorization algorithms [6, 7, 21], of having the LU factorization based on the processing of an assembly tree, while the other approaches use a graph structure and/or irregular sparsity patterns that are more complex to handle.

We have demonstrated that, based on the same assembly tree, one can derive a new multifrontal algorithm that will introduce asymmetry in the frontal matrices and in the matrix of the factors $\mathbf{F}$. The detection of the asymmetry is only based

[8] Available at http://www.enseeiht.fr/apo/MUMPS.

TABLE 6
Performance ratios of the new algorithm over the standard algorithm.

	Space for			Operations		
	LU	Stack	Total	Elimin.	Assemb.	Time
$0 \leq$ Structural symmetry < 50						
mean	0.79	0.35	0.70	0.67	0.41	0.60
median	0.80	0.35	0.78	0.68	0.40	0.61
$50 \leq$ Structural symmetry < 80						
mean	0.93	0.85	0.93	0.89	0.82	0.88
median	0.95	0.91	0.96	0.93	0.88	0.95

on structural information and is not costly to compute, as has been illustrated with structurally symmetric matrices, for which both algorithms behave similarly. On a set of unsymmetric matrices, we have shown that the new algorithm will reduce both the factor size and the number of operations by a significant factor. We have also observed that the reduction in the number of indirect memory access operations during the assembly process is generally much higher than the reduction in the number of elimination operations. We have noticed that the reduction in the maximum stack size is also relatively high and is comparable to the reduction in the number of assembly operations. Finally, we have shown that the new `ma41` code is very competitive with respect to `UMFPACK2`, `UMFPACK3`, and `SuperLU` codes, and this even on matrices with structural symmetry smaller than 50.

To conclude, we show in Table 6 a summary of the results (mean and median) obtained on the test matrices with structural symmetry smaller than 80. For very unsymmetric matrices (structural symmetry smaller that 50), we obtain an average reduction of 30% in the total maximum memory used and 40% in the factorization time. The maximum stack size and the number of assembly operations are reduced by 65% and 59%, respectively. Finally, it is interesting to observe that, even on fairly symmetric matrices ($50 \leq$ structural symmetry < 80), it can still be worth trying to identify and exploit asymmetry during the processing of the assembly tree.

Acknowledgments. We want to thank Horst Simon and Esmond Ng, who gave us the opportunity to work at NERSC (LBNL) for one year. We are grateful to Petter Bjørstad for providing access to the SGI Cray Origin 2000 from Parallab (Bergen, Norway). We also want to thank Sherry Li, Iain Duff, Jacko Koster, and Jean-Yves L'Excellent for helpful comments on an earlier version of this paper.

REFERENCES

[1] P. R. AMESTOY, T. A. DAVIS, AND I. S. DUFF, *An approximate minimum degree ordering algorithm,* SIAM J. Matrix Anal. Appl., 17 (1996), pp. 886–905.

[2] P. R. AMESTOY AND I. S. DUFF, *Vectorization of a multiprocessor multifrontal code,* Internat. J. Supercomput. Appl., 3 (1989), pp. 41–59.

[3] P. R. AMESTOY AND I. S. DUFF, *Memory management issues in sparse multifrontal methods on multiprocessors,* Internat. J. Supercomput. Appl., 7 (1993), pp. 64–82.

[4] P. R. AMESTOY, I. S. DUFF, J.-Y. L'EXCELLENT, AND J. KOSTER, *A fully asynchronous multifrontal solver using distributed dynamic scheduling,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 15–41.

[5] P. R. AMESTOY, J.-Y. L'EXCELLENT, AND I. S. DUFF, *Multifrontal parallel distributed symmetric and unsymmetric solvers,* Comput. Methods Appl. Mech. Engrg., 184 (2000), pp. 501–520.

[6] T. A. DAVIS AND I. S. DUFF, *An unsymmetric-pattern multifrontal method for sparse LU factorization,* SIAM J. Matrix Anal. Appl., 18 (1997), pp. 140–158.

[7] T. A. Davis and I. S. Duff, *A combined unifrontal/multifrontal method for unsymmetric sparse matrices,* ACM Trans. Math. Software, 25 (1999), pp. 1–19.

[8] T. A. Davis, J. R. Gilbert, S. I. Larimore, and E. G.-Y. Ng, *A Column Approximate Minimum Degree Ordering Algorithm,* Technical Report TR-00-005, Computer and Information Sciences Department, University of Florida, Gainesville, FL, 2000.

[9] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu, *A supernodal approach to sparse partial pivoting,* SIAM J. Matrix Anal. Appl., 20 (1999), pp. 720–755.

[10] I. S. Duff, *Full matrix techniques in sparse Gaussian elimination,* in Numerical Analysis (Dundee, 1981), Lecture Notes in Math. 912, G. A. Watson, ed., Springer-Verlag, Berlin, 1981, pp. 71–84.

[11] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices,* Oxford University Press, London, 1986.

[12] I. S. Duff, R. G. Grimes, and J. G. Lewis, *The Rutherford–Boeing Sparse Matrix Collection,* Technical Report RAL-TR-97-031, Rutherford Appleton Laboratory, Oxfordshire, UK, 1997; also Technical Report ISSTECH-97-017 from Boeing Information and Support Services and Report TR/PA/97/36 from CERFACS, Toulouse, France.

[13] I. S. Duff and J. Koster, *On algorithms for permuting large entries to the diagonal of a sparse matrix,* SIAM J. Matrix Anal. Appl., 22 (2001), pp. 973–996.

[14] I. S. Duff and J. K. Reid, *MA27—A Set of Fortran Subroutines for Solving Sparse Symmetric Sets of Linear Equations,* Technical Report R.10533, AERE, Harwell, UK, 1982.

[15] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear systems,* ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[16] I. S. Duff and J. K. Reid, *The multifrontal solution of unsymmetric sets of linear systems,* SIAM J. Sci. Statist. Comput., 5 (1984), pp. 633–641.

[17] S. C. Eisenstat and J. W. H. Liu, *Exploiting structural symmetry in unsymmetric sparse symbolic factorization,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 202–211.

[18] J. R. Gilbert and J. W. H. Liu, *Elimination structures for unsymmetric sparse LU factors,* SIAM J. Matrix Anal. Appl., 14 (1993), pp. 334–352.

[19] J. R. Gilbert and E. G. Ng, *Predicting structure in nonsymmetric sparse matrix factorizations,* in Graph Theory and Sparse Matrix Computations, J. R. Gilbert, A. George, and J. W. H. Liu, eds., Springer-Verlag, New York, 1993, pp. 107–140.

[20] Harwell Subroutine Library, 2000, A collection of Fortran codes for large scale scientific computation, available from http://www.cse.clrc.ac.uk/Activity/HSL.

[21] X. S. Li and J. W. Demmel, *A scalable sparse direct solver using static pivoting*, in Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing, San Antonio, TX, 1999, CD-ROM, SIAM, Philadelphia, 1999.

[22] J. W. H. Liu, *The role of elimination trees in sparse factorization,* SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.

[23] J. W. H. Liu, *The multifrontal method for sparse matrix solution: Theory and practice,* SIAM Rev., 34 (1992), pp. 82–109.

[24] R. Schreiber, *A new implementation of sparse Gaussian elimination,* ACM Trans. Math. Software, 8 (1982), pp. 256–276.

[25] S. A. Vavasis, *Stable finite elements for problems with wild constraints,* SIAM J. Numer. Anal., 33 (1996), pp. 909–916.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 570–580

ON MATRICES WITH SIGNED NULL-SPACES*

SI-JU KIM†, BRYAN L. SHADER‡, AND SUK-GEUN HWANG§

Abstract. We denote by $\mathcal{Q}(A)$ the set of all matrices with the same sign pattern as A. A matrix A has *signed null-space* provided there exists a set $\mathcal{S}$ of sign patterns such that the set of sign patterns of vectors in the null-space of $\widetilde{A}$ is $\mathcal{S}$ for each $\widetilde{A} \in \mathcal{Q}(A)$. Some properties of matrices with signed null-spaces are investigated.

Key words. totally L-matrices, signed compounds, signed null-spaces

AMS subject classification. 05C50

PII. S0895479801396221

1. Introduction.

The *sign* of a real number a is defined by

$$\operatorname{sign}(a) = \begin{cases} -1 & \text{if } a < 0, \\ 0 & \text{if } a = 0, \text{ and} \\ 1 & \text{if } a > 0. \end{cases}$$

A *sign pattern* is a $(0, 1, -1)$-matrix. The *sign pattern of a matrix* A is the matrix obtained from A by replacing each entry with its sign. We denote by $\mathcal{Q}(A)$ the set of all matrices with the same sign pattern as A.

Let A be an m by n matrix and b an m by 1 vector. The linear system $Ax = b$ has *signed solutions* provided there exists a collection $\mathcal{S}$ of n by 1 sign patterns such that the set of sign patterns of the solutions to $\widetilde{A}x = \tilde{b}$ is $\mathcal{S}$ for each $\widetilde{A} \in \mathcal{Q}(A)$ and $\tilde{b} \in \mathcal{Q}(b)$. This notion generalizes that of a sign-solvable linear system (see [1] and references therein). The linear system, $Ax = b$, is *sign-solvable* provided each linear system $\widetilde{A}x = \tilde{b}$ ($\widetilde{A} \in \mathcal{Q}(A)$, $\tilde{b} \in \mathcal{Q}(b)$) has a solution and all solutions have the same sign pattern. Thus $Ax = b$ is sign-solvable if and only if $Ax = b$ has signed solutions and the set $\mathcal{S}$ has cardinality 1.

The matrix A has *signed null-space* provided $Ax = 0$ has signed solutions. Thus A has signed null-space if and only if there exists a set $\mathcal{S}$ of sign patterns such that the set of sign patterns of vectors in the null-space of $\widetilde{A}$ is $\mathcal{S}$ for each $\widetilde{A} \in \mathcal{Q}(A)$. An *$L$-matrix* is a matrix A, with the property that each matrix in $\mathcal{Q}(A)$ has linearly independent rows. A square L-matrix is a *sign-nonsingular* (SNS)-matrix. A *totally L-matrix* is an m by n matrix such that each m by m submatrix is an SNS-matrix. It is known that totally L-matrices have signed null-spaces [3]. We also have the fact as a corollary of some results in this paper. Thus matrices with signed null-spaces generalize totally L-matrices.

A vector is *mixed* if it has a positive entry and a negative entry. A matrix is *row-mixed* if each of its rows is mixed. A *signing* is a nonzero diagonal $(0, 1, -1)$-matrix.

*Received by the editors October 8, 2001; accepted for publication (in revised form) by R. Wabben May 23, 2002; published electronically December 19, 2002.
http://www.siam.org/journals/simax/24-2/39622.html

†Department of Mathematics Education, Andong National University, Andong, 760-749 Republic of Korea, (sjkim@andong.ac.kr). The research of this author was supported by Andong National University.

‡Department of Mathematics, University of Wyoming, Laramie, Wyoming 82071 (bshader@uwyo.edu).

§Department of Mathematics Education, Kyungpook University, Taegu, 702-701, Republic of Korea (sghwang@knu.ac.kr). The research of this author was supported by Com^2MaC.

A signing is *strict* if each of its diagonal entries is nonzero. A matrix B is *strictly row-mixable* provided there exists a strict signing D such that BD is row-mixed.

In this paper, some properties of matrices with signed null-spaces are investigated, and we show that there exists an m by n matrix A with signed null-space such that A contains a totally L-matrix with m rows as its submatrix and the columns of A are distinct up to multiplication by -1 for any $n \in \{m, m+1, \ldots, 2m\}$.

We use the following standard notation throughout the paper. If k is a positive integer, then $\langle k \rangle$ denotes the set $\{1, 2, \ldots, k\}$. Let A be an m by n matrix. If α is a subset of $\{1, 2, \ldots, m\}$ and β is a subset of $\{1, 2, \ldots, n\}$, then $A[\alpha|\beta]$ denotes the submatrix of A determined by the rows whose indices are in α and the columns whose indices are in β. We sometimes use $A[*|\beta]$ instead of $A[\langle m \rangle|\beta]$. The submatrix complementary to $A[\alpha|\beta]$ is denoted by $A(\alpha|\beta)$. In particular, $A(-|\beta)$ denotes the submatrix obtained from A by deleting columns whose indices are in β. We write $\mathrm{diag}(d_1, d_2, \ldots, d_n)$ for the n by n diagonal matrix whose (i,i)-entry is d_i. Let $J_{m,n}$ denote the m by n matrix, all of whose entries are 1, and let e_i denote the column vector, all of whose entries are 0 except for the ith entry, which is 1.

2. Matrices with signed null-space. We say that an m by n matrix $A = [a_{ij}]$ *contains a mixed cycle* provided there exist distinct $i_1, i_2, \ldots, i_k$ and distinct $j_1, j_2, \ldots, j_k$ such that

$$a_{i_t,j_t} a_{i_t,j_{t+1}} < 0 \text{ for } t = 1, \ldots, k-1 \text{ and } a_{i_k,j_k} a_{i_k,j_1} < 0.$$

An m by n $(0, 1, -1)$-matrix has *signed mth compound* provided each of its m by m submatrices having term rank m is an SNS-matrix.

We make use of the following results of matrices with signed null-spaces.

THEOREM 2.1 (see [3]). *Let A be a strictly row-mixable m by n matrix. Then the following three conditions are equivalent.*

(a) *A has signed null-space.*

(b) *A has term rank m, and its mth compound is signed.*

(c) *AD has no mixed cycle for each strict signing such that AD is row-mixed.*

THEOREM 2.2 (see [2], [3]). *If a strictly row-mixable matrix A has signed null-space, then there exist matrices B and C (possibly with no rows) and nonzero vectors b and c such that B and C are strictly row-mixable matrices with signed null-spaces,*

$$\begin{bmatrix} B \\ b \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} c \\ C \end{bmatrix}$$

have signed null-spaces, and, up to permutation of rows and columns,

$$A = \begin{bmatrix} B & O \\ b & c \\ O & C \end{bmatrix}.$$

The converse also holds.

Let A be an m by n $(0, 1, -1)$-matrix. The matrix B is *conformally contractible* to A provided there exists an index k such that the rows and columns of B can be permuted so that B has the form

$$\left[\begin{array}{c|c|c} A[\langle m \rangle | \langle n \rangle \setminus \{k\}] & x & y \\ \hline 0 \quad \cdots \quad 0 & 1 & -1 \end{array}\right],$$

where $x = [x_1, \ldots, x_m]^T$ and $y = [y_1, \ldots, y_m]^T$ are $(0, 1, -1)$-vectors such that $x_i y_i \geq 0$ for $i = 1, 2, \ldots, m$, and the sign pattern of $x + y$ is the kth column of A.

Let B be conformally contractible to A. It is known that A has signed null-space if and only if B has signed null-space [3]. All matrices we consider from now on are assumed to be $(0, 1, -1)$-matrices.

THEOREM 2.3 (see [4]). *Let an m by n matrix A have a k by $k+1$ submatrix B whose complementary submatrix in A has term rank $m-k$. If there is a matrix B^* obtained from B by replacing some nonzero entries with 0's if necessary such that $J_{2,3}$ is the zero pattern of a matrix obtained from B^* by a sequence of conformal contractions, then A does not have signed null-space.*

Let M be an m by n strictly row-mixable matrix of the form

$$M = \begin{bmatrix} & & 0 \\ & * & \vdots \\ & & 0 \\ & 1 & 1 \end{bmatrix}. \tag{2.1}$$

PROPOSITION 2.4. *M has signed null-space if and only if*

$$A = \left[\begin{array}{ccccc|c} & & M & & & 0 \\ & & & & & \vdots \\ & & & & & 0 \\ \hline 0 & \cdots & 0 & 1 & -1 & 1 \end{array}\right]$$

has signed null-space.

Proof. Let M have signed null-space, and let C be any $m+1$ by $m+1$ submatrix of A. If C contains the last column of A, then $C(m+1|m+1)$ is an m by m submatrix of M. Hence $C(m+1|m+1)$ is an SNS-matrix, or $C(m+1|m+1)$ has identically zero determinant by Theorem 2.1. Thus C is an SNS-matrix, or C has identically zero determinant. Hence we may assume that C does not contain the last column of A. If C contains neither the $n-1$th column nor the nth column, then clearly C has identically zero determinant. If C contains only one of the $(n-1)$th column or the nth column, then $C(m+1|m+1)$ is an m by m submatrix of M. Hence $C(m+1|m+1)$ is an SNS-matrix, or $C(m+1|m+1)$ has identically zero determinant. Therefore, C is an SNS-matrix, or C has identically zero determinant. Let C contain both the $(n-1)$th column and the nth column of A. Then $C(m+1|m+1)$ is an SNS-matrix, or $C(m+1|m+1)$ has identically zero determinant. If $C(m+1|m+1)$ has identically zero determinant, then there exists an s by t zero submatrix of $C(m+1|m+1)$ such that $s+t = m+1$. From this, it is easy to show that C has a p by q zero submatrix such that $p+q = m+2$; i.e., C has identically zero determinant. Let $C(m+1|m+1)$ be an SNS-matrix. Since C is conformally contractible to $C(m+1|m+1)$, C is also an SNS-matrix. Thus the $(m+1)$th compound of A is signed. Since M has signed null-space, the term rank of M is m, and hence the term rank of A is $m+1$. Thus A has signed null-space by Theorem 2.1. The converse is trivial. □

We say that A is a *single extension* of M in Proposition 2.4. Proposition 2.4 means that a strictly row-mixable matrix has signed null-space if and only if its single extension has signed null-space.

Let

$$G = \begin{bmatrix} & & & & & 0 \\ & & * & & & \vdots \\ & & & & & 0 \\ 0 & \cdots & 0 & 1 & 1 & 1 \end{bmatrix}$$

be an m by n matrix, and let

$$H = \left[\begin{array}{cccccc|cc} & & & & & & & O & \\ \hline 0 & \cdots & 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & \cdots & 0 & 1 & 0 & -1 & 0 & 1 \end{array}\right].$$

PROPOSITION 2.5. *The m by n strictly row-mixable matrix G has signed null-space if and only if H has signed null-space.*

Proof. Let G have signed null-space, and let $C = [c_{ij}]$ be an $m+2$ by $m+2$ submatrix of H. That is, $C = H[*|\beta]$ for some $\beta \subset \langle n+2\rangle$. If $n+2 \in \beta$, then $H[\langle m+1\rangle|\beta \setminus \{n+2\}]$ is an SNS-matrix, or it has identically zero determinant since $H(m+2|n+2)$ is a single extension of G. Hence C is an SNS-matrix, or C has identically zero determinant. Similarly, we can show that C is an SNS-matrix or C has identically zero determinant if $n+1 \in \beta$. Hence we may assume that β contains neither $n+1$ nor $n+2$. Then it is easy to show that C has identically zero determinant if β contains at most two among $n-2,\ n-1$, and n. Let $\{n-2,\ n-1,\ n\} \subset \beta$. Then $H[\langle m\rangle|\beta \setminus \{n-1, n\}]$ is an SNS-matrix or it has identically zero determinant since G has signed null-space. If $H[\langle m\rangle|\beta \setminus \{n-1, n\}]$ has identically zero determinant, then clearly C has identically zero determinant. Let $H[\langle m\rangle|\beta \setminus \{n-1, n\}]$ be an SNS-matrix. Then $H[\langle m-1\rangle|\beta \setminus \{n-2, n-1, n\}]$ is an SNS-matrix since $c_{mm} = 1$. Since C is in the form of

$$\left[\begin{array}{c|ccc} H[\langle m-1\rangle|\beta \setminus \{n-2, n-1, n\}] & * & & \\ \hline & 1 & 1 & 1 \\ O & 0 & 1 & -1 \\ & 1 & 0 & -1 \end{array}\right]$$

and $C[m, m+1, m+2|m, m+1, m+2]$ is also an SNS-matrix, C is an SNS-matrix. The converse is trivial. □

We say that H is a *double extension* of G in Proposition 2.5. That G should have a row with exactly three ones is necessary in Proposition 2.5. For example, let

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

and

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

Then B is a double extension of A that has signed null-space. But $B[1, 2, 3, 4|1, 2, 3, 4]$ is a mixed submatrix of A, and hence B does not have signed null-space.

COROLLARY 2.6. *Every totally L-matrix has signed null-space.*

Proof. From Propositions 2.4 and 2.5, we have the result. □

PROPOSITION 2.7. *Let A be a strictly row-mixable m by n matrix with no duplicate columns up to multiplication by -1. If A has signed null-space and is not conformally contractible to a matrix, then it has at least two rows with exactly three nonzero entries.*

Proof. Without loss of generality, we may assume that each row of A has at least three nonzero entries and A has no zero column. Notice that $m \geq 2$ comes from the

condition. We prove the result by induction on m. Trivially, we have the result for $m = 2$. Let $m \geq 3$. By Theorem 2.2, A can be rearranged as

$$A = \begin{bmatrix} B & O \\ b & c \\ O & C \end{bmatrix},$$

where matrices B and C (possibly with no rows) are strictly row-mixable matrices which have signed null-spaces, and vectors b and c are nonzero.

$$\begin{bmatrix} B \\ b \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} c \\ C \end{bmatrix}$$

also have signed null-spaces. Let $A[\alpha|\beta] = \begin{bmatrix} B \\ b \end{bmatrix}$ and $A[\gamma|\delta] = \begin{bmatrix} c \\ C \end{bmatrix}$ such that $|\alpha| = k,\ |\beta| = s,\ |\gamma| = l$, and $|\delta| = t$. Then $k + l - 1 = m$ and $s + t = n$.

Let $k > 1$ and $l > 1$. If $A[\alpha|\beta]$ has one of the unit vectors $\pm e_k$ as a column, then we can assume that $A[\alpha|\beta]$ is of the form

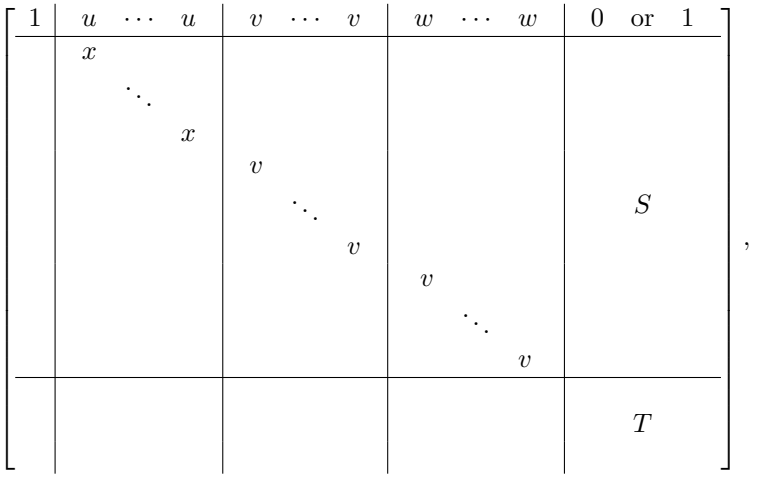
$$\begin{bmatrix} B' & O \\ b' & 1 \end{bmatrix}.$$

Let B' have no duplicate columns up to multiplication by -1. By induction, B' and hence A have at least two rows with exactly three nonzero entries. Thus we are done. Therefore, we assume that B' has duplicate columns up to multiplication by -1. Then $b' \neq 0$. If b' has at least two nonzero entries, then $A[\alpha|\beta]$ is a strictly row-mixable matrix with no duplicate columns up to multiplication by -1. Since $A[\alpha|\beta]$ is not conformally contractible to a matrix, B has at least one row with exactly three nonzero entries. Let b' have exactly one nonzero entry. Let the columns $1, 2$ of B' be a pair of duplicate columns up to multiplication by -1, and let p be the number of nonzero entries in the column 1 of B'. Let D be a strict signing such that $M = B'D$ is row-mixed. Since B' has signed null-space, M has no mixed cycle, and hence the columns 1 and 2 of M must be identical or $p = 1$. If $p \geq 2$, then the matrix M' obtained from M by multiplying the column 2 by -1 has a mixed cycle. Thus M' is a row-mixed matrix with signed null space, which is impossible by Theorem 2.1. Hence $p = 1$. Therefore, every duplicate column of B' is of the form e_i or $-e_i$ for some i. Hence B' has only one pair of duplicate columns, which are e_i or $-e_i$ for some $i(< k)$. The matrix obtained from B' by deleting one of the duplicate columns, which are e_i or $-e_i$, satisfies the conditions of the hypothesis if its ith row has at least three nonzero entries. This implies that B has at least one row with exactly three nonzero entries. Let $C' = A[\gamma|\{s\} \cup \delta]$. Similarly, C' has a row i with exactly three nonzero entries for some $i(\neq 1)$. Hence C has at least one row with exactly three nonzero entries. Therefore, A has at least two rows with exactly three nonzero entries. Similarly, in the case in which $A[\gamma|\delta]$ has one of the unit vectors $\pm e_1$ as a column, we have the result. Assume that $A[\alpha|\beta]$ and $A[\gamma|\delta]$ do not have the unit vectors $\pm e_k$ and $\pm e_1$, respectively, as columns. Since b is nonzero, the k by $s + 1$ matrix B^* obtained from $A[\alpha|\beta]$ by adding e_k as a column is a strictly row-mixable matrix with no duplicate columns up to multiplication by -1. Since B has signed null-space, B^* also has signed null-space. Applying the similar method above to B^*, we can derive that B has at least one row with exactly three nonzero entries. Similarly, C also has at least one row with exactly three nonzero entries. Hence we have the result when $k > 1$ and $l > 1$.

Let $k = 1$. Then $s = 1$ since the columns of A are distinct up to multiplication by -1. Hence we may assume that $A = [a_{ij}]$ is of the form

$$\begin{bmatrix} 1 & c \\ O & C \end{bmatrix}.$$

If C has no duplicate columns up to multiplication by -1, then we have the result for C by induction, and hence we have the result for A. Let C have duplicate columns up to multiplication by -1. Then the duplicate columns of C are of the form e_i or $-e_i$ for some i, as we have shown before. This implies that the number of identical columns of C up to multiplication by -1 is at most 3. Therefore, we may assume that the zero pattern of A is of the form



where $u = (1, 1, 0)$, $v = (1, 1)$, $w = (1, 0)$, and $x = (1, 1, 1)$, and the unspecified entries are zero. Let ϵ be the set of indices of columns in A corresponding to $\begin{bmatrix} S \\ T \end{bmatrix}$. Then we may also assume that $A[\gamma \setminus \{1\}|\epsilon]$ has no duplicate columns up to multiplication by -1, and the columns are also different from the ones of $A(1|\epsilon)$ up to multiplication by -1. If $\begin{bmatrix} S \\ T \end{bmatrix}$ is vacuous, we are done since $l \geq 3$ and every row but the first row of A has at least three nonzero entries. Let only T be vacuous. Notice that each column of S has at least two nonzero entries. Hence each row of S has at most one nonzero entry. For, suppose that a row of S has two nonzero entries. Since the columns of $A[\gamma \setminus \{1\}|\epsilon]$ are distinct up to multiplication by -1, we may assume that there exists a submatrix of A whose zero pattern is

$$\begin{bmatrix} 1 & 1 & * & * \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 1 & 1 & * & * \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & * \\ 0 & 0 & 1 & * & 1 \end{bmatrix},$$

where * is 0 or 1. By Theorem 2.3, A does not have signed null-space. This is a contradiction. Next, suppose that a row r of $A[\gamma \setminus \{1\}|\langle n \rangle]$ has four nonzero entries. Since each row of S has at most one nonzero entry and each column of S has at least two nonzero entries, we have a submatrx of A whose zero pattern is

$$\begin{bmatrix} 1 & 1 & 1 & * \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

which is also impossible by Theorem 2.3. Hence each row of $A[\gamma \setminus \{1\}|\langle n\rangle]$ has exactly three nonzero entries. Thus we have the result when T is vacuous. Let T be nonvacuous. Notice that the submatrix of A corresponding to T is a strictly row-mixable matrix with signed null-space. Let $T^{'}$ be the matrix obtained from T by deleting zero columns. Then we may assume that T is of the form $[O\ T^{'}]$. If the submatrix $A^{'}$ of A corresponding to $T^{'}$ has no duplicate columns up to multiplication by -1, then $A^{'}$ has at least two rows with exactly three nonzero entries by induction. Hence we have the result. Suppose that $A^{'}$ has duplicate columns up to multiplication by -1. It is easy to show that such columns of $A^{'}$ have exactly one nonzero entry as we have shown above. We want to show that the number of identical columns of $A^{'}$ is at most three. Suppose that there are four identical columns in $A^{'}$ up to multiplication by -1. We may assume that the zero pattern of the submatrix consisting of such duplicate columns of $A^{'}$ is of the form

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ & O & & \end{bmatrix}.$$

Since $A[\gamma\setminus\{1\}|\epsilon]$ has no duplicate columns up to multiplication by -1, we may assume that $A[\gamma \setminus \{1\}|\epsilon]$ has a submatrix whose zero pattern is

$$\begin{bmatrix} 1 & * & * \\ * & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & * & * \\ * & 1 & * \\ * & * & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

where * is 0 or 1. Hence we can have a submatrix N of A whose zero pattern is

$$\begin{bmatrix} 1 & 1 & * & * & * \\ 1 & 0 & 1 & * & * \\ 0 & 1 & * & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & 1 & 1 & * & * & * \\ 1 & 0 & 0 & 1 & * & * \\ 0 & 1 & 0 & * & 1 & * \\ 0 & 0 & 1 & * & * & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix},$$

where * is 0 or 1. By Theorem 2.3, A does not have signed null-space. This is a contradiction. Thus we can assume that $T^{'}$ is of the form

$$\begin{bmatrix} T_1^{'} & T_2^{'} \\ O & T_3^{'} \end{bmatrix},$$

where $T_1^{'}$ is a block diagonal matrix whose diagonal blocks are either [1 1] or [1 1 1], and the submatrix of A corresponding to $\begin{bmatrix} T_2^{'} \\ T_3^{'} \end{bmatrix}$ has no duplicate columns up to multiplication by -1. Continuing this process, we can assume that T is of the form

$$\begin{bmatrix} T_1 & & * \\ O & \ddots & \\ & & T_q \end{bmatrix},$$

where $T_i = [O\ T_i^{'}]$ for $i = 1, 2, \ldots, q$ and $T_i^{'}$ are block diagonal matrices whose diagonal blocks are either [1 1] or [1 1 1] for $i = 1, 2, \ldots, q-1$.

Let λ_i be the set of indices of rows in A corresponding to T_i. Let ϵ_i and δ_i be the set of indices of nonzero columns in A and zero columns in A corresponding

to T_i, respectively. It is easy to show that each row of $A[\lambda_i|\epsilon_i \cup \delta_{i+1}]$ has at most three nonzero entries for $i = 1, 2, \ldots, q-1$ by a method similar to that used in the case in which only T is vacuous. If the submatrix $A_q^{'}$ of A corresponding to $T_q^{'}$ has no duplicate columns up to multiplication by -1, then $A_q^{'}$ satisfies the hypothesis. Hence we have the result. If $A_q^{'}$ has duplicate columns up to multiplication by -1, then we may assume that $T_q^{'} = [T_q^{''}\ T_q^{'''}]$, where $T_q^{''}$ is a block diagonal matrix whose diagonal blocks are [1 1] or [1 1 1]. As we have shown above in the case in which T is vacuous, each row of $T_q^{'}$ has exactly three nonzero entries. If $T_q^{'}$ has at least two rows, then we are done.

Thus we may assume that $T_q^{'} = [1\ 1\ 1]$. Then $A[\langle m-1\rangle|n-2, n-1, n]$ cannot have a row whose zero pattern is equal to $(1, 1, 1)$ because, if so, then A has $J_{2,3}$ as a submatrix, and this is impossible by Theorem 2.3. If $A[m-1|n-2, n-1, n] = O$, then we are done. Hence we may assume that the zero pattern of $A[m-1|n-2, n-1, n]$ is either [1 1 0] or [1 0 0].

Let the zero pattern of $A[m-1|n-2, n-1, n]$ be [1 1 0]. If the rth row of $A[\langle m-2\rangle|n-2, n-1, n]$ has the zero pattern $(1, 1, 0)$ for some r, then there exist distinct $i_1, i_2, \ldots, i_k$ and distinct $j_1, j_2, \ldots, j_k$ such that $a_{i_1,j_1}, a_{i_2,j_1}, \ldots, a_{i_k,j_k}$ are nonzero, where $i_1 = 1, i_k = r$, and $j_k = n-2$. There also exist distinct $p_1, p_2, \ldots, p_t$ and distinct $q_1, q_2, \ldots, q_t$ such that $a_{p_1,q_1}, a_{p_2,q_1}, \ldots, a_{p_t,q_t}$ are nonzero, where $p_1 = 1$, $p_t = m-1$, and $q_t = n-2$. Choosing some entries from these entries, we have a matrix which is conformally contractible to a matrix whose zero pattern is $J_{2,3}$. This is impossible by Theorem 2.3. We can apply a method similar to that used above to show that $A[\langle m-2\rangle|n] = O$. Hence each row of $A[\langle m-2\rangle|n-2, n-1, n]$ has a zero pattern of the forms $(0, 0, 0)$, $(1, 0, 0)$, or $(0, 1, 0)$. Let $T_{q-1}^{'}$ have at least two rows. It is easy to show that, if each row of $A[\lambda_{q-1}|\epsilon_{q-1} \cup \delta_q \cup \epsilon_q]$ has at least four nonzero entries, we have a submatrix of A which is conformally contractible to a matrix whose zero pattern is $J_{2,3}$ by the method just used above. By Theorem 2.3, it is impossible. Hence some row of $A[\lambda_{q-1}|\epsilon_{q-1} \cup \delta_q \cup \epsilon_q]$ has exactly three nonzero entries. Thus we have the result when $T_{q-1}^{'}$ has at least two rows. Therefore, we may assume that $T_{q-1}^{'}$ is either [1 1] or [1 1 1]. Notice that $T_q = T_q^{'} = [1\ 1\ 1]$.

Let $T_{q-1}^{'} = [1\ 1\ 1]$. If $A[\langle m-2\rangle|n-2, n-1, n] \neq O$, then we can show that there exists a submatrix of A which is conformally contractible to a matrix whose zero pattern is $J_{2,3}$. This is impossible. Hence we may assume that $A[\langle m-2\rangle|n-2, n-1, n] = O$. Then $A[\langle m-1\rangle|\langle n-3\rangle]$ has at least two rows with exactly three nonzero entries by induction. Hence we are done. Let $T_{q-1}^{'} = [1\ 1]$. Notice that $A[\langle m-2\rangle|n-4, n-3]$ has no submatrix whose zero pattern is $J_{2,2}$ by Theorem 2.1. That is, all rows of $A[\langle m-2\rangle|n-4, n-3]$ except for one row have at least one zero entry. Since the conformal contraction of $A[\langle m-1\rangle|\langle m-3\rangle]$ on the last row has signed null-space, $A[\langle m-1\rangle|\langle n-3\rangle]$ has at least one row with exactly three nonzero entries. Thus we have the result if $A[\langle m-2\rangle|n-2, n-1, n] = O$. Let $A[\langle m-2\rangle|n-2, n-1, n] \neq O$. Since we are done if the $(m-2)$nd row of A has exactly three nonzero entries, we may assume that the $(m-2)$nd row of A has at least four nonzero entries. Deleting the cases in which a contradiction occurs, we may assume that the zero pattern of $A[m-2, m-1, m|n-6, n-5, n-4, n-3, n-2, n-1, n]$ is

$$\left[\begin{array}{cccc|ccc} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right].$$

It is easy to show that $A[\langle m-2\rangle|n-1,n]=O$ by using a method similar to that used above. If the columns of $A[m-2,m-1|n-4,n-2]$ are identical up to multiplication by -1, then it is easy to find a strict signing D such that AD is a row-mixed matrix and $A[m-2,m-1|n-4,n-2]D$ contains a mixed cycle. This is impossible by Theorem 2.1. Hence the columns of $A[\langle m-1\rangle|\langle n\rangle]$ are not identical up to multiplication by -1. Therefore, $A[\langle m-1\rangle|\langle n-2\rangle]$ satisfies the hypothesis. Thus we have the result when the zero pattern of $A[m-1|n-2,n-1,n]$ is [1 1 0].

In the case in which the zero pattern of $A[m-1|n-2,n-1,n]$ is [1 0 0], the last row of $A[\langle m-1\rangle|\langle n-3\rangle]$ must have two or three nonzero entries. If it has two nonzero entries, then we are done. Let it have three nonzero entries. Then we have a submatrix of A which is conformally contractible to a matrix whose zero pattern is $J_{2,3}$ if $A[\langle m-2\rangle|n-2,n-1,n]\neq O$ by a method similar to that used above. Hence we have $A[\langle m-2\rangle|n-2,n-1,n]=O$. Therefore, $A[\langle m-1\rangle|\langle n-3\rangle]$ has at least two rows with exactly three nonzero entries by induction. Thus we have the result for $k=1$. Similarly, we have the same result for $l=1$. □

3. Matrices containing totally *L*-matrices. Let A be a matrix with signed null-space. A is a *maximal matrix with signed null-space* if any matrix obtained from A by replacing a zero entry with a nonzero entry does not have signed null-space.

LEMMA 3.1. *An m by $m+2$ totally L-matrix is a maximal matrix with signed null-space.*

Proof. Let A be an m by $m+2$ totally L-matrix. Let A^* be an m by $m+2$ matrix obtained from A by replacing a zero entry with 1 or -1. Notice that every m by m submatrix of A^* has term rank m. Since A^* has a row with four nonzero entries, A^* is not a totally L-matrix. Therefore, there exists an m by m submatrix of A^* that is not an SNS-matrix. Hence A^* does not have signed null-space by Theorem 2.1. □

LEMMA 3.2. *Let A be an m by $m+2$ totally L-matrix, and let $\mathbf{x}$ be an m by 1 column vector which has at least two nonzero entries. Then $B=[A\ \mathbf{x}]$ does not have signed null-space.*

Proof. We will prove the result by induction on m. The statement is clear for $m=2$. We may assume that

$$B=[b_{ij}]=\left[\begin{array}{c|c|c} M' & \begin{matrix}O\\ I_2\end{matrix} & \mathbf{x}\end{array}\right],$$

where I_2 is the identity matrix of order 2. If $b_{m-1,m+3}=0$ or $b_{m,m+3}=0$, say, $b_{m,m+3}=0$, then $B(m|m+2)$ does not have signed null-space by induction. Hence we have the result by Theorem 2.3. Therefore, we may assume that the last two positions of $\mathbf{x}$ have nonzero entries. Since a totally L-matrix is a maximal matrix with signed null-space, $B(-|m+2)$ does not have signed null-space. Hence B does not have signed null-space. □

We say that an m by $m+2$ totally L-matrix contains k *double-extensions* (or $m-2k-2$ *single-extensions*) if A is obtained from

$$\begin{bmatrix} 1 & 1 & 1 & 0\\ 1 & -1 & 0 & 1\end{bmatrix}$$

by a sequence of $m-2k-2$ single-extensions and k double-extensions up to row and column permutations and multiplication of rows and columns by -1.

PROPOSITION 3.3. *Let A be an m by n matrix with signed null-space whose*

columns are nonzero and distinct up to multiplication by -1. *If* A *contains an* m *by* $m+2$ *totally* L*-matrix with* k *double-extensions, then* $n \leq 2m-2k$.

Proof. We will prove the result by induction on k. Let T_k be an m by $m+2$ totally L-matrix with k double-extensions contained in A. Notice that each column of A which does not correspond to T_k has exactly one zero entry by Lemma 3.2. If $k=0$, then it is known [1] that T_k has a signed rth compound for each $r=1,2,\ldots,m$. Hence we can have the identity matrix I_m as a submatrix of A. Since T_0 has exactly two columns with exactly one nonzero entry, $n \leq m+2+(m-2)=2m=2m-2k$. Let $k \neq 0$. By Proposition 2.4 and Lemma 3.2, we may assume that A is of the form

$$\left[\begin{array}{c|cc|ccc} A_1 & \multicolumn{2}{c|}{A_2} & \multicolumn{3}{c}{O} \\ \hline A_3 & 1 & -1 & 0 & 0 & 0 \\ & 1 & 1 & -1 & 0 & 0 \\ \hline O & 0 & 1 & 1 & 1 & 0 \\ & 1 & 0 & 1 & 0 & 1 \end{array}\right].$$

Then $A(m-1,m|n-1,n)$ has signed null-space, and it contains an $m-2$ by m totally L-matrix with $k-1$ double-extensions. The columns of $A(m-1,m|n-1,n)$ are distinct up to multiplication by -1 because, if not, then A_3 has a column of the forms $(0,1)^T$ or $(0,-1)^T$, say, $(0,1)^T$. Then A has a submatrix

$$B = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & 1 & 1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \tag{3.1}$$

which is not an SNS-matrix. Since A contains an m by $m+2$ totally L-matrix, the complementary submatrix to B in A has term rank $m-4$. Hence A does not have signed null-space by Theorem 2.1. This is a contradiction. Therefore, $n-2 \leq 2(m-2)-2(k-1)=2m-2k-2$ by induction. Thus we have $n \leq 2m-2k$. □

Let l be the number of single-extensions contained in A. Then we have $l=m-2k-2$. Hence we can restate the result of Proposition 3.3 in terms of l: $n \leq m+l+2$.

COROLLARY 3.4. *Let* T *be an* m *by* $m+2$ *totally* L*-matrix which contains no single-extensions. Then there is no* m *by* n *matrix* A *with signed null-space such that* A *contains* T *properly, and the columns of* A *are nonzero and distinct up to multiplication by* -1.

Proof. Let A be an m by n matrix with signed null-space, and let A contain T. Since T contains no single-extensions, $l=0$. Hence $n \leq m+l+2=m+2$. Hence $A=T$. □

Let M be an m by n matrix of the form in (2.1) with signed null-space, and let A be the $m+1$ by $n+2$ matrix such that

$$A = \left[\begin{array}{ccccc|cc} & & & & & 0 & 0 \\ & & M & & & \vdots & \vdots \\ & & & & & 0 & 0 \\ & & & & & 1 & 0 \\ \hline 0 & \cdots & 0 & 1 & 0 & -1 & 1 \end{array}\right].$$

Since $A(-|n+2)$ is conformally contractible to M, $A(-|n+2)$ has signed null-space. Since M has signed null-space, A has signed null-space by Theorem 2.1. Let T_k be an m by $m+2$ totally L-matrix with k double-extensions. Let $\{i_1,i_2,\ldots,i_l\}$ with

$i_1 < i_2 < \cdots < i_l$ be the set of indices of rows used when single-extensions are constructed in T_k. Notice that T_k does not contain any e_{i_j}, $j = 1, 2, \ldots, l$. The remark above and Proposition 2.5 imply that

$$T = [T_k \ e_{i_1} \ e_{i_2} \ \ldots \ e_{i_l}] \tag{3.2}$$

is an m by $2m - 2k$ matrix whose columns are distinct up to multiplication by -1, and it has signed null-space.

Let j be the index of a row of T_k used when a double-extension is done, and suppose that T_k does not have e_j as a column. $[T_k \ e_j]$ has a submatrix of the form in (3.1), and hence it does not have signed null-space, as we have shown in the proof of Proposition 3.3.

Let $\mathcal{T}_k$ be the set of all matrices of the form in (3.2). Notice that columns of $A \in \mathcal{T}_k$ are nonzero and distinct up to multiplication by -1. We can express the m by n matrices A with $n = 2m - 2k$ in Proposition 3.3 in terms of elements of $\mathcal{T}_k$.

PROPOSITION 3.5. *In Proposition* 3.3, $n = 2m - 2k$ *if and only if there exists a permutation matrix* Q *such that* A *is equal to* TQ *up to multiplication of rows and columns by* -1 *for some* $T \in \mathcal{T}_k$.

Proof. Let A be an m by n matrix such that $A = TQ$ for some permutation matrix Q and $T \in \mathcal{T}_k$. Then $m = 2k+l+2$, and hence $n = m+2+l = m+2+(m-2-2k) = 2m - 2k$. Conversely, let A be an m by $2m - 2k$ matrix satisfying the conditions in Proposition 3.3. Let T_k be an m by $m + 2$ totally L-matrix with k double-extensions contained in A. Then there exists a permutation matrix Q and strict signings D, E such that $DAQE$ is a submatrix of matrix T of the form in (3.2) by Lemma 3.2 and the remark above. Since T is an m by $2m - 2k$ matrix, $A = DTQ^{-1}E$. Since $T \in \mathcal{T}_k$, we have the result. □

COROLLARY 3.6. *Let* m *be a positive integer with* $m \geq 2$, *and let* n *be any integer in* $\{m, m + 1, \ldots, 2m\}$. *Then there exists an* m *by* n *matrix* A *with signed null-space such that* A *contains a totally* L*-matrix with* m *rows as its submatrix and the columns of* A *are nonzero and distinct up to multiplication by* -1.

Proof. Let n be any integer in $\{m, m + 1, \ldots, 2m\}$. If $n \leq m + 2$, then we can take an m by n totally L-matrix as such a matrix A. If $n > m + 2$, there exists an m by $m + 2$ totally L-matrix T_{n-m-2} with $n - m - 2$ single-extensions. Hence there exists an m by n matrix $A \in \mathcal{T}_{n-m-2}$ which contains T_{n-m-2} by the remark above. □

Acknowledgment. We would like to thank the referees for their comments.

REFERENCES

[1] R. A. BRUALDI AND B. L. SHADER, *The Matrices of Sign-Solvable Linear Systems*, Cambridge University Press, Cambridge, UK, 1995.

[2] K. G. FISHER, W. MORRIS, AND J. SHAPIRO, *Mixed dominating matrices*, Linear Algebra Appl., 270 (1998), pp. 191–214.

[3] S.-J. KIM AND B. L. SHADER, *Linear systems with signed solutions*, Linear Algebra Appl., 313 (2000), pp. 21–40.

[4] S.-J. KIM AND B. L. SHADER, *On matrices which have signed null-spaces*, Linear Algebra Appl., 353 (2002), pp. 245–255.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 581–598

FAST ITERATIVE METHODS FOR SINC SYSTEMS*

MICHAEL K. NG† AND DANIEL POTTS‡

Abstract. We consider linear systems of equations arising from the sinc method of boundary value problems which are typically nonsymmetric and dense. For the solutions of these systems we propose Krylov subspace methods with banded preconditioners. We prove that our preconditioners are invertible and discuss the convergence behavior of the conjugate gradient method for normal equations (CGNE). In particular, we show that the solution of an n-by-n discrete sinc system arising from the model problem can be obtained in $\mathcal{O}(n \log^2 n)$ operations by using the preconditioned CGNE method. Numerical results are given to illustrate the effectiveness of our fast iterative solvers.

Key words. sinc method, Toeplitz matrices, Krylov subspace methods, preconditioners, banded matrices

AMS subject classifications. 65F10, 65F15, 65T10

PII. S0895479800369773

1. Introduction.

In the sinc-Galerkin method, the basis functions are derived from the Whittaker cardinal (sinc) function

$$\operatorname{sinc}(x) := \begin{cases} \dfrac{\sin(\pi x)}{\pi x}, & x \in \mathbb{R} \setminus 0, \\ 1, & x = 0, \end{cases}$$

and its translates into

$$s(k,h)(x) := \operatorname{sinc}\left(\frac{x - kh}{h}\right) \qquad (x \in \mathbb{R},\ k \in \mathbb{Z},\ h > 0).$$

The globally supported basis functions can be transformed via a composition with a suitable conformal map to any connected subset of the real line. This basis has been proved useful in the numerical analysis of a number of problems [17, 23, 24].

We seek an approximate solution of the linear two-point boundary value problem

$$\begin{aligned} \mathcal{L}u = u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \quad a < x < b, \\ u(a) = u(b) = 0. \end{aligned} \tag{1.1}$$

We approximate u by

$$u_{M+N+1}(x) = \sum_{k=-M}^{N} u_k s(k,h) \circ \phi(x), \tag{1.2}$$

*Received by the editors March 16, 2001; accepted for publication (in revised form) by Z. Strakoš April 18, 2002; published electronically December 19, 2002. The work described in this paper was supported by a grant from the German Academic Exchange Services and the Research Grants Council of the Hong Kong Joint Research Scheme (project G-HK020/00).

http://www.siam.org/journals/simax/24-2/36977.html

†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 grant 7132/00P and HKU CRCG grants 10203501, 10203907, and 10203408.

‡Institute of Mathematics, Medical University of Lübeck, Wallstr. 40, D-23560 Lübeck, Germany (potts@math.mu-luebeck.de). The research of this author was supported in part by the Hong Kong–German Joint Research Collaboration Grant from the Deutscher Akademischer Austauschdienst and the Hong Kong Research Grants Council. Part of this work was done while he was visiting the Department of Mathematics of The Chinese University of Hong Kong.

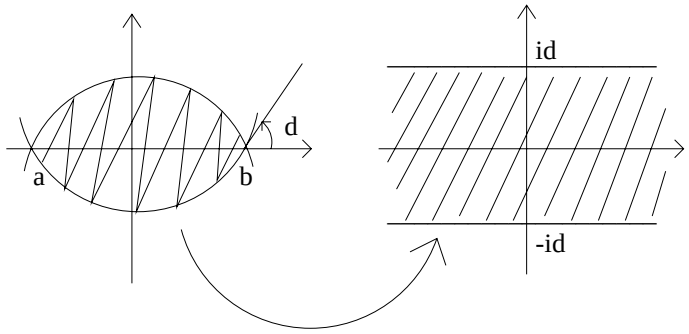


FIG. 1.1. *The conformal map* $\phi(z) = \log\left(\frac{z-a}{b-z}\right)$.

where ϕ is a conformal map of a simply connected domain $\mathcal{S}$ with boundary points $a \neq b$ onto

$$\mathcal{S}_d = \{z : z = x + \mathrm{i}y, \quad |y| < d, \quad d > 0\} \tag{1.3}$$

such that $\phi(a) = -\infty$ and $\phi(b) = \infty$. In Figure 1.1 (see, for instance, Lund and Bowers [17, p. 118], and Stenger [24, pp. 67–68]), we give an example of such a conformal map. The simply connected domain $\mathcal{S}$ is the eye-shaped region

$$\left\{ z : \left| \arg\ \left(\frac{z-a}{b-z}\right)\right| < d \right\},$$

and the conformal map is given by

$$\phi(z) = \log\left(\frac{z-a}{b-z}\right).$$

Other conformal maps can also be found in [17, 23, 24]. The general Galerkin method enables us to determine $\{u_k\}_{k=-M}^{N}$ by solving the linear system of equations

$$\langle \mathcal{L}u_{M+N+1} - f, s(k,h) \circ \phi \rangle = 0, \quad -M \leq k \leq N, \tag{1.4}$$

where the inner product is defined by

$$\langle f, g \rangle := \int_a^b f(x)g(x)w(x)\mathrm{d}x.$$

Here w plays the role of a weight function. For the case of second order problems, it is convenient to take $w(x) = \frac{1}{\phi'(x)}$; see [17, p. 116]. The most distinctive feature of the sinc basis is the resulting exponential convergence rate of the error. Moreover, the convergence rate is maintained when the solution of the boundary value problem has boundary singularities.

The approximate explicit expressions for the inner products in (1.4) have been thoroughly treated in [17, 23]. The resulting discrete sinc-Galerkin matrix coupling with collocation (see [24, p. 465]) is given by the dense matrix

$$\boldsymbol{A} = \boldsymbol{T}_2 + \boldsymbol{D}_1\boldsymbol{T}_1 + \boldsymbol{T}_1\boldsymbol{D}_1 + \boldsymbol{D}_2 \qquad (\boldsymbol{A} \in \mathbb{R}^{n\times n}), \tag{1.5}$$

where $\boldsymbol{T}_2$ is a symmetric Toeplitz matrix, $\boldsymbol{T}_1$ is a skew-symmetric Toeplitz matrix, and $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices. Here $n = M + N + 1$. A straightforward application of the Gaussian elimination method will result in an algorithm, which takes $\mathcal{O}(n^3)$ arithmetical operations.

For n-by-n Toeplitz systems, fast and superfast direct solvers requiring $\mathcal{O}(n^2)$ and $\mathcal{O}(n \log^2 n)$ arithmetical operations, respectively, have been developed; see, for instance, Levinson [15] and Ammar and Gragg [1]. However, there exist no fast direct solvers for solving the system in (1.5). This is mainly because the displacement rank of the coefficient matrix can take any value between 0 and n. Hence fast Toeplitz solvers that are based on low displacement rank of matrices cannot be applied. The details of displacement ranks can be found in [14].

However, we note that given any n-vector $\boldsymbol{q}$, the matrix–vector product $\boldsymbol{Aq}$ can be computed in $\mathcal{O}(n \log n)$ operations [26]. In fact, $\boldsymbol{T}_l\boldsymbol{q}$ ($l \in \{1, 2\}$) can be obtained by using fast trigonometric transforms; see, e.g., [11, 21]. Since $\boldsymbol{D}_l$ is a diagonal matrix, the product $\boldsymbol{D}_l\boldsymbol{q}$ ($l \in \{1, 2\}$) can be computed in $\mathcal{O}(n)$ operations. Thus Krylov subspace methods, which are based on matrix–vector products, can be employed for solving sinc systems. Since $\boldsymbol{A}$ is nonsymmetric, for solving the equations

$$\boldsymbol{Au} = \boldsymbol{f}, \tag{1.6}$$

we would suggest using conjugate-gradient-type methods like GMRES [22, p. 158], BiCGSTAB [22, p. 217], or the conjugate gradient method for normal equations (CGNE) [22, p. 238].

One way to speed up the convergence rate of CGNE is to precondition the coefficient matrix. Instead of solving the original system $\boldsymbol{Au} = \boldsymbol{f}$, we solve the preconditioned system

$$(\boldsymbol{M}^{-1}\boldsymbol{A})\boldsymbol{u} = \boldsymbol{M}^{-1}\boldsymbol{f}. \tag{1.7}$$

We note that the convergence rate of the CGNE method depends on the singular values of the preconditioned matrix [5, 28]. The matrix $\boldsymbol{M}$, called a preconditioner to the matrix $\boldsymbol{A}$, should be chosen with two criteria in mind: $\boldsymbol{Mr} = \boldsymbol{d}$ is easy to solve for any vector $\boldsymbol{d}$; the spectrum of $(\boldsymbol{M}^{-1}\boldsymbol{A})(\boldsymbol{M}^{-1}\boldsymbol{A})^{\mathrm{T}}$ is uniformly bounded and well separated from the origin compared to that of $\boldsymbol{AA}^{\mathrm{T}}$.

In [19], we have considered the symmetric sinc-Galerkin method [16] for discretization of the second order self-adjoint boundary value problem. In this case, the sinc-Galerkin matrix $\boldsymbol{A}$ is the sum of a symmetric Toeplitz matrix and a diagonal matrix. We have used banded matrices $\boldsymbol{R}$ with bandwidths independent of the size of the matrix as preconditioners. We have shown that they give rise to the fast convergence of the preconditioned conjugate gradient (PCG) method [10]. In particular, we proved that the spectra of $\boldsymbol{R}^{-1}\boldsymbol{A}$ are uniformly bounded from above and below by positive constants independent of the size of the matrix. The banded system $\boldsymbol{Rr} = \boldsymbol{d}$ can be solved in $\mathcal{O}(n)$ operations, where n is the size of the matrix. Therefore the cost of each PCG iteration is of $\mathcal{O}(n \log n)$ operations. It follows that the solution of $\boldsymbol{Au} = \boldsymbol{f}$ can be obtained in $\mathcal{O}(n \log n)$ operations. However, these preconditioners cannot be applied to nonsymmetric sinc systems.

The main aim of this paper is to propose other banded preconditioners $\boldsymbol{B}$ for $\boldsymbol{A}$, given by (1.5). We show that the singular values of the preconditioned sinc matrix arising from the model problem are uniformly bounded except for at most a finite number of outliers. Using this result, we show that the CGNE method applied to (1.7)

converges at most in $\mathcal{O}(\log n)$ iteration steps. Hence the method requires $\mathcal{O}(n \log^2 n)$ operations.

The outline of this paper is as follows: In section 2, we study some properties of the discrete sinc system. In section 3, we introduce our preconditioners. The convergence analysis of the CGNE method is given in section 4. Numerical results are presented in section 5 to illustrate the effectiveness of our method. Furthermore we compare the CGNE method with Krylov subspace methods like GMRES or BiCGSTAB, which do not require the translation of (1.7) to the normal equations. Finally, section 6 contains some concluding remarks.

2. Properties of discrete systems. Let $\mathcal{S}$ be a simply connected domain in the complex plane with boundary points $a \neq b$. Let ϕ be a conformal mapping of $\mathcal{S}$ onto the strip $\mathcal{S}_d$ defined by (1.3) such that $\phi(a) = -\infty$ and $\phi(b) = \infty$. For $1 \leq k \leq \infty$, let $\mathcal{H}^k(\mathcal{S})$ denote the family of all functions f that are analytic in $\mathcal{S}$ and fulfill

$$\begin{cases} \left(\int_{\partial \mathcal{S}} |f(z)|^k \mathrm{d}z \right)^{1/k} < \infty, & 1 \leq k < \infty, \\ \sup_{z \in \mathcal{S}} |f(z)| < \infty, & k = \infty. \end{cases}$$

Corresponding to the number α, let $\mathcal{L}_\alpha(\mathcal{S})$ denote the family of all analytic functions on $\mathcal{S}$ for which there exists a constant C such that

$$|f(z)| \leq C \, \frac{|\mathrm{e}^{\phi(z)}|^\alpha}{(1 + |\mathrm{e}^{\phi(z)}|)^{2\alpha}} \quad \forall z \in \mathcal{S} \, .$$

To study the convergence of the sinc-Galerkin method for differential problems, assumptions on the functions ϕ, p, and q are required.

Assumption (A1) (see [24, pp. 467, 469]). Assume for the differential equation (1.1) that p/ϕ', $(p/\phi')/\phi'$, $q/(\phi')^2$, $(1/\phi')'$, and $(1/\phi')''/\phi'$ are real valued and belong to $\mathcal{H}^\infty(\mathcal{S})$ and that problem (1.1) has a unique solution $u \in \mathcal{L}_\alpha(\mathcal{S})$.

Assumption (A2) (see [24, p. 478]). Assume for the differential equation (1.1) that

$$\mathrm{Re} \left(\frac{1}{\phi'(x)} \left(\frac{1}{\phi'(x)} \right)'' - \frac{1}{\phi'(x)} \left(\frac{p(x)}{\phi'(x)} \right)' + \frac{2q(x)}{(\phi'(x))^2} \right) \leq 0 \quad \text{for } a < x < b.$$

The following theorem about the approximate solution was given in [24].

THEOREM 2.1 (see [24, Theorem 7.2.6]). *Let Assumptions* (A1) *and* (A2) *be satisfied. Let*

$$\begin{aligned} \boldsymbol{A}_n^{(g)} := {} & \boldsymbol{T}_n[g_2] + h \boldsymbol{T}_n[g_1] \boldsymbol{D}_n \left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'} \right] \\ & + h^2 \boldsymbol{D}_n \left[\frac{1}{\phi'} \left(\frac{1}{\phi'} \right)'' - \frac{1}{\phi'} \left(\frac{p}{\phi'} \right)' + \frac{q}{(\phi')^2} \right], \end{aligned} \tag{2.1}$$

$$\boldsymbol{A}_n^{(c)} := \boldsymbol{T}_n[g_2] + h \boldsymbol{D}_n \left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'} \right] \boldsymbol{T}_n[g_1] + h^2 \boldsymbol{D}_n \left[\frac{q}{(\phi')^2} \right], \tag{2.2}$$

and

$$\boldsymbol{A}_n := \frac{1}{2} \left(\boldsymbol{A}_n^{(g)} + \boldsymbol{A}_n^{(c)} \right). \tag{2.3}$$

Here $\boldsymbol{T}_n[g_\ell]$ $(\ell \in \{1,2\})$ *denotes the* n*-by-*n *Toeplitz matrix with the* (j,k)*th entry given by the* $(j-k)$*th Fourier coefficient of the function*

$$g_\ell(\theta) = (\mathrm{i}\theta)^\ell \quad \forall \theta \in [-\pi, \pi]\,, \tag{2.4}$$

$\boldsymbol{D}_n[\psi]$ *is an* n*-by-*n *diagonal matrix given by*

$$\boldsymbol{D}_n[\psi] = \ \mathrm{diag}\ [\psi(x_{-M}), \ldots, \psi(x_0), \ldots, \psi(x_N)],$$

with $x_k = \phi^{-1}(kh)$ *for* $k = 0, \pm 1, \pm 2, \ldots$. *If the vector* $\boldsymbol{u} = (u_{-M}, \ldots, u_N)^T$ *denotes the exact solution of the system of equations*

$$\boldsymbol{A}_n \boldsymbol{u} = h^2 \boldsymbol{D}_n \left[\frac{1}{(\phi')^2}\right] \boldsymbol{f}, \tag{2.5}$$

where $\boldsymbol{f} = [f(x_{-M}), \ldots, f(x_N)]^T$, *then*

$$|u(x) - u_n(x)| \le C n^{1/2} \mathrm{e}^{-(\pi d \alpha n)^{1/2}} \quad \text{for } a < x < b. \tag{2.6}$$

THEOREM 2.2 (see [24, Lemma 7.2.5]). *Let Assumptions* (A1) *and* (A2) *be satisfied. Let* $\boldsymbol{A}_n^{(g)}$, $\boldsymbol{A}_n^{(c)}$, *and* $\boldsymbol{A}_n$ *be defined as in* (2.1), (2.2), *and* (2.3), *respectively. Then the following hold true:*

(i) *There exists a constant* c_1 *independent of* n *such that*

$$\|\boldsymbol{A}_n^{(g)}\|_2,\ \|\boldsymbol{A}_n^{(c)}\|_2,\ \|\boldsymbol{A}_n\|_2 \le \pi^2 \left(1 + \frac{c_1}{\sqrt{n}}\right).$$

(ii) *There exists a constant* c_2 *independent of* n *such that*

$$\|(\boldsymbol{A}_n^{(g)})^{-1}\|_2,\ \|(\boldsymbol{A}_n^{(c)})^{-1}\|_2,\ \|\boldsymbol{A}_n^{-1}\|_2 \le \frac{4n^2}{\pi^2}\left(1 + \frac{c_2}{n}\right).$$

In particular, the condition number $\kappa(\boldsymbol{A}_n \boldsymbol{A}_n^T)$ *of* $\boldsymbol{A}_n \boldsymbol{A}_n^T$ *satisfies*

$$\kappa(\boldsymbol{A}_n \boldsymbol{A}_n^T) \le 4n^2 \left(1 + \frac{c_1}{\sqrt{n}}\right)\left(1 + \frac{c_2}{n}\right).$$

Since $\kappa(\boldsymbol{A}_n \boldsymbol{A}_n^{\mathrm{T}}) = \mathcal{O}(n^2)$, the convergence of the CGNE method might be very slow with increasing n; see, for instance, Theorem 4.1 in section 4. In the next section, we introduce the banded preconditioner to precondition the sinc coefficient matrix in order to speed up the convergence rate of the CGNE method.

3. Banded preconditioners. Recall that the coefficient matrix $\boldsymbol{A}_n$ in (2.3) is the sum of Toeplitz-times-diagonal matrices and diagonal matrices. There are many "good" preconditioners for the individual parts. For instance, the diagonal matrix system can be solved easily. For Toeplitz systems, circulant preconditioners have been proved to be successful choices; see the recent survey paper by Chan and Ng [3]. However, we remark that circulant preconditioners do not work for Toeplitz-plus-banded systems. Even T. Chan's circulant preconditioner [6], which is well defined for non-Toeplitz matrices, will not—while defined for $\boldsymbol{A}_n$—work well when $\boldsymbol{D}_n[\cdot]$ are not identity matrices; see the numerical results in [4]. If we approximate $\boldsymbol{T}_n[g_\ell]$ in (2.3) by a circulant preconditioner $\boldsymbol{C}_n[g_\ell]$, then

$$\boldsymbol{C}_n[g_2] + \frac{h}{2}(\boldsymbol{D}_n^I \boldsymbol{C}_n[g_1] + \boldsymbol{C}_n[g_1]\boldsymbol{D}_n^I) + \frac{h^2}{2}\boldsymbol{D}_n^{II},$$

where

$$\boldsymbol{D}_n^I := \boldsymbol{D}_n \left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'} \right] \quad \text{and} \quad \boldsymbol{D}_n^{II} := \boldsymbol{D}_n \left[\frac{1}{\phi'} \left(\frac{1}{\phi'} \right)'' - \frac{1}{\phi'} \left(\frac{p}{\phi'} \right)' + \frac{2q}{(\phi')^2} \right]$$

can be expected to be a "good" approximation to $\boldsymbol{A}_n$. Unfortunately, the resulting circulant-type matrix system *cannot* be solved easily in general. Hence, this approach to constructing a preconditioner for $\boldsymbol{A}_n$ cannot work in most situations. In this paper, we consider a preconditioner which is easily invertible.

In [19], we have proposed to use banded matrices as preconditioners for symmetric sinc-Galerkin systems. Following this approach, we introduce our preconditioners $\boldsymbol{B}_n$ by

$$\boldsymbol{B}_n := \boldsymbol{P}_n^{II} + \frac{h}{2}(\boldsymbol{D}_n^I \boldsymbol{P}_n^I + \boldsymbol{P}_n^I \boldsymbol{D}_n^I) + \frac{h^2}{2} \boldsymbol{D}_n^{II}, \tag{3.1}$$

where $\boldsymbol{P}_n^{II}$ and $\boldsymbol{P}_n^I$ are the banded Toeplitz matrices

$$\boldsymbol{P}_n^{II} := \boldsymbol{T}_n(p_2) \;=\; \text{tridiag}\,[1, -2, 1] \quad \text{and} \quad \boldsymbol{P}_n^I := \boldsymbol{T}_n(p_1) \;=\; \text{tridiag}\,\left[-\frac{1}{2}, 0, \frac{1}{2} \right]$$

with generating functions of $\boldsymbol{P}_n^I$ and $\boldsymbol{P}_n^{II}$ given by

$$p_1(\theta) := \mathrm{i} \sin\theta \quad \text{and} \quad p_2(\theta) := -2 + 2\cos\theta \quad \forall \theta \in [-\pi, \pi], \tag{3.2}$$

respectively.

We note that the preconditioner $\boldsymbol{B}_n$ is just an n-by-n tridiagonal matrix. It follows that the system $\boldsymbol{B}_n \boldsymbol{r} = \boldsymbol{d}$ can be solved by using any efficient tridiagonal solver in $\mathcal{O}(n)$ operations.

The symmetric and skew-symmetric parts of $\boldsymbol{B}_n$ are given by

$$\boldsymbol{B}_n^{(h)} := \boldsymbol{P}_n^{II} + \frac{h^2}{2} \boldsymbol{D}_n^{II} \quad \text{and} \quad \boldsymbol{B}_n^{(s)} := \frac{h}{2}(\boldsymbol{D}_n^I \boldsymbol{P}_n^I + \boldsymbol{P}_n^I \boldsymbol{D}_n^I),$$

respectively. Moreover, we have by the theorem of Bendixson [25, p. 418] that

$$\lambda_{\min}(\boldsymbol{B}_n^{(h)}) \leq \; \mathrm{Re}[\lambda(\boldsymbol{B}_n)] \leq \lambda_{\max}(\boldsymbol{B}_n^{(h)})$$

and

$$\lambda_{\min}\left(\frac{1}{\mathrm{i}} \boldsymbol{B}_n^{(s)} \right) \leq \; \mathrm{Im}[\lambda(\boldsymbol{B}_n)] \leq \lambda_{\max}\left(\frac{1}{\mathrm{i}} \boldsymbol{B}_n^{(s)} \right),$$

where $\lambda(\boldsymbol{B})$ denotes the eigenvalues of the matrix $\boldsymbol{B}$.

LEMMA 3.1. *Let Assumption* (A2) *be satisfied. Further, let*

$$d_2 := \min_{x \in \phi^{-1}(\mathbb{R})} \left\{ \frac{1}{\phi'(x)} \left(\frac{1}{\phi'(x)} \right)'' - \frac{1}{\phi'(x)} \left(\frac{p(x)}{\phi'(x)} \right)' + \frac{2q(x)}{(\phi'(x))^2} \right\}$$

and

$$d_3 := \max_{x \in \phi^{-1}(\mathbb{R})} \left\{ \frac{1}{\phi'(x)} \left(\frac{1}{\phi'(x)} \right)'' - \frac{1}{\phi'(x)} \left(\frac{p(x)}{\phi'(x)} \right)' + \frac{2q(x)}{(\phi'(x))^2} \right\}.$$

Then we have

$$\boldsymbol{P}_n^{II} + \frac{d_2 h^2}{2}\,\boldsymbol{I}_n \le \boldsymbol{B}_n^{(h)} \le \boldsymbol{P}_n^{II} + \frac{d_3 h^2}{2}\,\boldsymbol{I}_n.$$

In particular, the preconditioners $\boldsymbol{B}_n$ are nonsingular for all n.

Proof. The assertion follows from (A2) and the fact that the matrices $\boldsymbol{B}_n^{(h)}$ are negative definite. ∎

Remark. In [24, p. 481], Stenger showed that the approximate solution for $u_{M+N+1}(x)$ in (1.2) can also be obtained by solving the linear systems involving the coefficient matrices $\boldsymbol{A}_n^{(g)}$ and $\boldsymbol{A}_n^{(c)}$ given in (2.1) and (2.2), respectively. We note that we can also develop similar banded preconditioners

$$\boldsymbol{B}_n^{(g)} := \boldsymbol{P}_n^{II} + h\,\boldsymbol{P}_n^{I}\boldsymbol{D}_n^{I} + h^2\boldsymbol{D}_n\left[\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' - \frac{1}{\phi'}\left(\frac{p}{\phi'}\right)' + \frac{q}{(\phi')^2}\right]$$

and

$$\boldsymbol{B}_n^{(c)} := \boldsymbol{P}_n^{II} + h\,\boldsymbol{D}_n^{I}\boldsymbol{P}_n^{I} + h^2\,\boldsymbol{D}_n\left[\frac{q}{(\phi')^2}\right]$$

for the matrices $\boldsymbol{A}_n^{(g)}$ and $\boldsymbol{A}_n^{(c)}$, respectively. Numerical tests show that these preconditioners work similarly well as the preconditioner $\boldsymbol{B}_n$ for $\boldsymbol{A}_n$. However, we remark that the convergence analysis for these preconditioned systems $(\boldsymbol{B}_n^{(g)})^{-1}\boldsymbol{A}_n^{(g)}$ and $(\boldsymbol{B}_n^{(c)})^{-1}\boldsymbol{A}_n^{(c)}$ is still an open problem. ∎

3.1. The model problem. In this subsection, we consider some model sinc-Galerkin matrices and analyze the spectra of these preconditioned matrices. By using the Bendixson theorem again, we obtain that symmetric and skew-symmetric parts of $\boldsymbol{A}_n$ are given by

$$\boldsymbol{A}_n^{(h)} := \boldsymbol{T}_n[g_2] + \frac{h^2}{2}\boldsymbol{D}_n^{II} \quad \text{and} \quad \boldsymbol{A}_n^{(s)} := \frac{h}{2}\left(\boldsymbol{D}_n^{I}\boldsymbol{T}_n[g_1] + \boldsymbol{T}_n[g_1]\boldsymbol{D}_n^{I}\right),$$

respectively, and that

$$\lambda_{\min}(\boldsymbol{A}_n^{(h)}) \le \ \mathrm{Re}[\lambda(\boldsymbol{A}_n)] \le \lambda_{\max}(\boldsymbol{A}_n^{(h)})$$

and

$$\lambda_{\min}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right) \le \ \mathrm{Im}[\lambda(\boldsymbol{A}_n)] \le \lambda_{\max}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right).$$

Let

$$d_1 := \max_{x\in\phi^{-1}(\mathbb{R})}\left\{\left|\frac{-\phi''(x)}{(\phi'(x))^2} - \frac{p(x)}{\phi'(x)}\right|\right\}. \tag{3.3}$$

Then we have

$$-\lambda_{\max}\left(\frac{d_1 h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right) \le \lambda_{\min}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right) \le \lambda_{\max}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right) \le \lambda_{\max}\left(\frac{d_1 h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right).$$

For the symmetric part of $\boldsymbol{A}_n$, we find

$$\boldsymbol{T}_n[g_2] + \frac{d_2 h^2}{2}\boldsymbol{I}_n \le \boldsymbol{A}_n^{(h)} \le \boldsymbol{T}_n[g_2] + \frac{d_3 h^2}{2}\boldsymbol{I}_n,$$

where d_2 and d_3 are defined as in Lemma 3.1. In particular, we have

$$\lambda_{\min}\left(\boldsymbol{T}_n[g_2]+\frac{d_2h^2}{2}\boldsymbol{I}_n\right)\le\lambda_{\min}\left(\boldsymbol{A}_n^{(h)}\right)$$
$$\le\lambda_{\max}\left(\boldsymbol{A}_n^{(h)}\right)\ \le\ \lambda_{\max}\left(\boldsymbol{T}_n[g_2]+\frac{d_2h^2}{2}\boldsymbol{I}_n\right).$$

The spectrum of the matrix $\boldsymbol{A}_n$ is contained in the box

$$\left[\lambda_{\min}\left(\boldsymbol{T}_n[g_2]+\frac{d_2h^2}{2}\boldsymbol{I}_n\right),\lambda_{\max}\left(\boldsymbol{T}_n[g_2]+\frac{d_3h^2}{2}\boldsymbol{I}_n\right)\right]$$
$$\times\ \left[-\lambda_{\max}\left(\frac{d_1h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right),\lambda_{\max}\left(\frac{d_1h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right)\right]$$

in the complex plane. This suggests that we analyze the banded preconditioners for the following model sinc-Galerkin matrices:

$$\boldsymbol{T}_n[g_2]+h\gamma_1\boldsymbol{T}_n[g_1]+h^2\,\gamma_2\boldsymbol{I}_n \text{ with } \gamma_1\in\{\pm d_1\} \text{ and } \gamma_2\in\{d_2/2,d_3/2\}. \tag{3.4}$$

If the corresponding banded matrices are good preconditioners of these model sinc-Galerkin matrices, then we expect that $\boldsymbol{B}_n$ will be a good preconditioner for $\boldsymbol{A}_n$. Numerical results in section 5 will show that our banded preconditioners give rise to fast convergence of the iterative method.

3.2. Spectra of the preconditioned matrices for the model problem. We note that the model problem matrices in (3.4) are Toeplitz matrices. Therefore, we analyze the spectra of their corresponding preconditioned matrices by using their generating functions. We first establish the following lemma.

LEMMA 3.2. *Let $c_1\in\mathbb{R}$, let c_2 be a negative number, and let h be a positive number. Let $g_1(\theta),g_2(\theta),p_1(\theta)$, and $p_2(\theta)$ be defined as in* (2.4) *and* (3.2). *If*

$$r(\theta)=\frac{hc_1g_1(\theta)+g_2(\theta)+h^2c_2}{hc_1p_1(\theta)+p_2(\theta)+h^2c_2}\quad\forall\theta\in[-\pi,\pi],$$

then

$$1\le\ \mathrm{Re}(r(\theta))<\frac{3\pi^3}{8}+\frac{\pi^2}{16}h^2c_1^2\quad\forall\theta\in[-\pi,\pi] \tag{3.5}$$

and

$$-\frac{h|c_1|\pi}{4-h^2c_2}\le\ \mathrm{Im}(r(\theta))\le\frac{h|c_1|\pi}{4-h^2c_2}\quad\forall\theta\in[-\pi,\pi],$$

where

$$r(\theta)\ =\ \mathrm{Re}(r(\theta))+i\,\mathrm{Im}(r(\theta)). \tag{3.6}$$

Proof. We have

$$\mathrm{Re}(r(\theta))=\frac{(-\theta^2+h^2c_2)(-2+2\cos\theta+h^2c_2)+h^2c_1^2\theta\sin\theta}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2}$$

and

$$\mathrm{Im}(r(\theta)) = \frac{hc_1\theta(-2+2\cos\theta+h^2c_2) - hc_1\sin\theta(-\theta^2+h^2c_2)}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2}\,.$$

Let us start with the real part. First we see that $\mathrm{Re}(r)-1$ is nonnegative because

$$\begin{aligned}&\frac{(-\theta^2+h^2c_2)(-2+2\cos\theta+h^2c_2)+(hc_1)^2\theta\sin\theta}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2} - 1\\ &= \frac{h^2c_1^2(\theta-\sin\theta)\sin\theta + (-2+2\cos\theta+h^2c_2)(-\theta^2+2-2\cos\theta)}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2} \qquad (3.7)\end{aligned}$$

and both functions $(\theta-\sin\theta)\sin\theta$ and $(-2+2\cos\theta+h^2c_2)(-\theta^2+2-2\cos\theta)$ are nonnegative on $[-\pi,\pi]$.

Since

$$\mathrm{Re}(r(\theta)) = \frac{(-\theta^2+h^2c_2)(-2+2\cos\theta+h^2c_2)+h^2c_1^2\theta\sin\theta}{(2-2\cos\theta-h^2c_2)^2+h^2c_1^2\sin^2\theta} \qquad (3.8)$$

we get with

$$\frac{2}{\pi}\theta^2 \le 2-2\cos\theta \le \theta^2 \quad \left(0\le\theta\le\frac{\pi}{2}\right) \qquad \text{and} \qquad \frac{2}{\pi}\theta \le \sin\theta \le \theta \quad \left(0\le\theta\le\frac{\pi}{2}\right)$$

that

$$\begin{aligned}\mathrm{Re}(r(\theta)) &\le \frac{(\theta^2-h^2c_2)(\theta^2-h^2c_2)+h^2\gamma_1^2\theta^2}{(\frac{2}{\pi}\theta^2-h^2c_2)^2+h^2\gamma_1^2\left(\frac{2}{\pi}\theta\right)^2}\\ &\le \max\left\{\left(\frac{\theta^2-h^2c_2}{\frac{2}{\pi}\theta^2-h^2c_2}\right)^2,\quad \frac{h^2\gamma_1^2\theta^2}{h^2\gamma_1^2\frac{4}{\pi^2}\theta^2}\right\}\\ &\le \max\left\{\left(\max\left\{\frac{\pi}{2},1\right\}\right)^2,\quad \frac{\pi^2}{4}\right\} \;=\; \frac{\pi^2}{4} \qquad \left(0\le\theta\le\frac{\pi}{2}\right).\end{aligned}$$

On the other hand, we have for $\frac{\pi}{2}\le\theta\le\pi$ that

$$\frac{4}{\pi}\theta \le 2-2\cos\theta \le \frac{3}{2}\theta \quad \text{and} \quad \sin\theta \le \theta - \frac{1}{\pi^2}\theta^3$$

and further by (3.8) we have

$$\begin{aligned}\mathrm{Re}(r(\theta)) &\le \frac{(\theta^2-h^2c_2)\left(\frac{3}{2}\theta-h^2c_2\right)+h^2\gamma_1^2\left(\theta^2-\frac{1}{\pi^2}\theta^4\right)}{\left(\frac{4}{\pi}\theta-h^2c_2\right)^2}\\ &\le \frac{(\theta^2-h^2c_2)\left(\frac{3}{2}\theta-h^2c_2\right)}{\left(\frac{4}{\pi}\theta-h^2c_2\right)^2} + h^2c_1^2\frac{\theta^2-\frac{1}{\pi^2}\theta^4}{\left(\frac{4}{\pi}\theta\right)^2}\\ &\le \frac{(\pi^2-h^2c_2)\left(\frac{3}{2}\pi-h^2c_2\right)}{(2-h^2c_2)^2} + \frac{\pi^2}{16}h^2\gamma_1^2\\ &\le \frac{\pi^2}{2}\cdot\frac{3}{4}\pi + \frac{\pi^2}{16}h^2c_1^2 \;=\; \frac{3\pi^3}{8}+\frac{\pi^2}{16}h^2c_1^2 \qquad \left(\frac{\pi}{2}\le\theta\le\pi\right).\end{aligned}$$

Since $\mathrm{Re}(r)$ is even, (3.5) follows. Furthermore, we have

(3.9)
$$\mathrm{Im}(r(\theta)) = \frac{hc_1(-2\theta + 2\theta\cos\theta + \theta h^2 c_2 + \theta^2 \sin\theta - h^2 c_2 \sin\theta)}{4 - 8\cos\theta - 4h^2 c_2 + 4\cos^2\theta + 4h^2 c_2 \cos\theta + h^4 c_2^2 + h^2 c_1^2 - h^2 c_1^2 \cos^2\theta}.$$

By using the Taylor series of $\cos\theta$ and $\sin\theta$ we get for $c_1 > 0$ that the numerator of the right-hand side of (3.9) is less than $\frac{h^3 c_1 c_2}{6}\theta$, but the denominator is bigger than $h^4 c_2^2 + h^2(c_1^2 - 2c_2)\theta^2$. Hence the maximum and minimum values of $\mathrm{Im}(r(\theta))$ are attained at $\theta = \pi$ and $\theta = -\pi$. The result follows by noting that

$$\mathrm{Im}(r(-\pi)) = -\mathrm{Im}(r(\pi)) = \frac{hc_1\pi}{4 - h^2 c_2}. \qquad \square$$

The next lemma follows immediately from the close relationship between the spectrum of a Toeplitz matrix and its generating function [9].

LEMMA 3.3. *Let γ_1 and γ_2 be defined as in* (3.4). *Then we have*

$$1 \le \lambda(\boldsymbol{T}_n[\mathrm{Re}(r)]) < \frac{3\pi^3}{8} + \frac{\pi^2}{16} h^2 \gamma_1^2$$

and

$$-\frac{h|\gamma_1|\pi}{4 - h^2\gamma_2} \le \lambda(\boldsymbol{T}_n[\mathrm{Im}(r)]) \le \frac{h|\gamma_1|\pi}{4 - h^2\gamma_2} \quad \forall \theta \in [-\pi, \pi].$$

Next we prove the following lemma.

LEMMA 3.4. *Let Assumptions* (A1) *and* (A2) *be satisfied. Then, for all n,*

$$\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n = (\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)\boldsymbol{T}_n[r] + \boldsymbol{L}_n, \tag{3.10}$$

where $\boldsymbol{L}_n$ has only nonzero entries in the first and last columns.

Proof. The result can be derived by noting that $\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n$ is a tridiagonal Toeplitz matrix. $\square$

With Lemma 3.4, we have that the spectra of the preconditioned matrices are also essentially bounded.

THEOREM 3.5. *Let Assumptions* (A1) *and* (A2) *be satisfied. Then at most* 8 *eigenvalues of*

$$(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n) \tag{3.11}$$

are outside the box

$$\left[1, \frac{3\pi^3}{8} + \frac{\pi^2}{16} h^2 \gamma_1^2\right] \times \left[-\frac{h|\gamma_1|\pi}{4 - h^2\gamma_2}, \frac{h|\gamma_1|\pi}{4 - h^2\gamma_2}\right]$$

in the complex plane.

Proof. Since the matrix $\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n$ is nonsingular, we obtain from (3.10) that

$$(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n) = \boldsymbol{T}_n[r] + \tilde{\boldsymbol{L}}_n, \tag{3.12}$$

where $\tilde{\boldsymbol{L}}_n = (\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}\boldsymbol{L}_n$ and the rank of $\tilde{\boldsymbol{L}}_n$ is at most 2. Let λ be an eigenvalue of the preconditioned matrix in (3.11). Then we get by Bendixson's theorem that

$$\lambda_{\min}\left(\boldsymbol{T}_n[\mathrm{Re}(r)] + \frac{\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{\mathrm{T}}}{2}\right) \le \mathrm{Re}(\lambda) \le \lambda_{\max}\left(\boldsymbol{T}_n[\mathrm{Re}(r)] + \frac{\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{\mathrm{T}}}{2}\right),$$

where $\mathrm{Re}(r)$ and $\mathrm{Im}(r)$ are defined as in (3.6). Since

$$\mathrm{rank}\left(\frac{\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{\mathrm{T}}}{2}\right) = 4,$$

by using Weyl's theorem [12, Theorem 4.3.1], at most 4 eigenvalues of $\boldsymbol{T}_n[\mathrm{Re}(r)] + (\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{*})/2$ are not contained in the interval $[\min \mathrm{Re}(r(\theta)), \max \mathrm{Re}(r(\theta))]$. Similarly, we prove that at most 4 eigenvalues of $\boldsymbol{T}_n[\mathrm{Im}(r)] + (\tilde{\boldsymbol{L}}_n - \tilde{\boldsymbol{L}}_n^{\mathrm{T}})/2$ are not contained in the interval $[\min \mathrm{Im}(r(\theta)), \max \mathrm{Im}(r(\theta))]$ of the imaginary axis. Now the assertion follows from Lemma 3.3. ∎

We remark that it is well known that the knowledge of the eigenvalues alone is not sufficient to estimate the convergence rate of GMRES; see, for instance, [8, 18]. As a matter of fact, it still remains an open problem to describe the convergence of GMRES in terms of some simple characteristic properties of the coefficient matrix. Even though we show in Theorem 3.5 that the eigenvalues of the preconditioned matrices are contained in a bounded region except for a finite number of outliers, we cannot provide a tight convergence bound of GMRES. However, we expect that GMRES may converge very fast when we apply GMRES to solve these preconditioned systems. Our numerical results in section 5 will show that GMRES indeed converges very fast.

Next we consider the singular values distribution of the preconditioned matrix. This will be useful in estimating the number of iterations required for convergence of the CGNE method.

With Lemmas 3.3 and 3.4, we have our main theorem, which states that the spectra of the preconditioned normal equations matrices are essentially bounded.

THEOREM 3.6. *Let Assumptions* (A1) *and* (A2) *be satisfied. Then there exist* $\beta \ge 1$ *independent of* n *such that at most* 6 *singular values of*

$$(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)$$

are outside the interval $[1, \beta]$.

Proof. By Lemma 3.4, we obtain

$$\begin{aligned}
&[(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)] \cdot \\
&[(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)]^{\mathrm{T}} \\
&= \boldsymbol{T}_n[r]\boldsymbol{T}_n[r]^{\mathrm{T}} + \hat{\boldsymbol{L}}_n,
\end{aligned}$$

where $\hat{\boldsymbol{L}}_n$ is Hermitian and $\mathrm{rank}(\hat{\boldsymbol{L}}_n) = 6$. By using the Courant–Fischer theorem about the inequalities between individual singular values of $\boldsymbol{T}_n[r]$ and eigenvalues of its Hermitian part [13, p. 151], we have

$$\sigma_{\min}(\boldsymbol{T}_n[r]) \ge \lambda_{\min}(\boldsymbol{T}_n[\mathrm{Re}(r)]) \ge 1\,.$$

Here $\sigma(\cdot)$ denotes the singular values of a matrix. By using Lemma 3.3, we get

$$\begin{aligned}\sigma_{\max}(\boldsymbol{T}_n[r]) \;\le\; \|\boldsymbol{T}_n[r]\|_2 \;\le\; 2\,\|r\|_\infty &\le 2\sqrt{\left(\frac{3\pi^3}{8}+\frac{\pi^2}{16}h^2\gamma_1^2\right)^2+\left(\frac{h|\gamma_1|\pi}{4-h^2\gamma_2}\right)^2}\\ &\le 2\sqrt{\left(\frac{3\pi^3}{8}+\frac{\pi^2}{16}\gamma_1^2\right)^2+\left(\frac{|\gamma_1|\pi}{4}\right)^2} := \beta.\end{aligned} \tag{3.13}$$

Hence the result follows. □

4. Convergence analysis of CGNE. An important practical aspect of solving boundary value problem (1.1) is the efficient solution of the resulting linear system

$$\boldsymbol{C}_n\boldsymbol{x} = \boldsymbol{B}_n^{-1}\boldsymbol{b} = \tilde{\boldsymbol{b}}, \tag{4.1}$$

with $\boldsymbol{C}_n = \boldsymbol{B}_n^{-1}\boldsymbol{A}_n$.

CGNE for solving the linear system (4.1) amounts to applying conjugate gradients to the system $\boldsymbol{C}_n\boldsymbol{C}_n^{\mathrm{T}}\boldsymbol{y} = \tilde{\boldsymbol{b}}$ under the change of variables $\boldsymbol{x} = \boldsymbol{C}_n^{\mathrm{T}}\boldsymbol{y}$; see [8, p. 105]. We note that the convergence rate of the CGNE method depends on the singular values of the preconditioned matrix. Since the singular values of the preconditioned sinc matrix arising from the model problem are uniformly bounded except for at most a finite number of outliers (cf. Theorem 3.6), we will show that the convergence rate of the PCG method for the normal equations will converge in at most $\mathcal{O}(\log n)$ steps. We begin by noting the following error estimate of the conjugate gradient method for the normal equations; see [28].

THEOREM 4.1. *Let $\boldsymbol{x}$ be the solution to $\boldsymbol{C}_n\boldsymbol{x} = \tilde{\boldsymbol{b}}$ and let $\boldsymbol{x}^{(j)}$ be the jth iterate of CGNE applied to the system $\boldsymbol{C}_n\boldsymbol{C}_n^T\boldsymbol{y} = \tilde{\boldsymbol{b}}$ under the change of variables $\boldsymbol{x} = \boldsymbol{C}_n^T\boldsymbol{y}$. If the eigenvalues $\{\delta_k\}$ of $\boldsymbol{C}_n\boldsymbol{C}_n^T$ are such that*

$$0 < \delta_1 \le \cdots \le \delta_p \le b_1 \le \delta_{p+1} \le \cdots \le \delta_{n-q} \le b_2 \le \delta_{n-q+1} \le \cdots \le \delta_n,$$

then

$$\frac{||\boldsymbol{x}-\boldsymbol{x}^{(j)}||_2}{||\boldsymbol{x}-\boldsymbol{x}^{(0)}||_2} \le 2\left(\frac{b-1}{b+1}\right)^{j-p-q} \cdot \max_{\delta\in[b_1,b_2]}\left\{\prod_{k=1}^{p}\left(\frac{\delta-\delta_k}{\delta_k}\right)\prod_{k=n-q+1}^{n}\left(\frac{\delta_k-\delta}{\delta_k}\right)\right\} \tag{4.2}$$

for $j \ge p+q$. Here $b \equiv (b_2/b_1)^{\frac{1}{2}} \ge 1$.

We can derive (4.2) by passing linear polynomials through the outlying eigenvalues δ_k for $1 \le k \le p$ and $n-q+1 \le k \le n$ and using a $(j-p-q)$th degree Chebyshev polynomial to minimize the error in the interval $[b_1, b_2]$. Since we always have

$$0 \le \frac{\delta_k-\delta}{\delta_k} \le 1, \quad n-q+1 \le k \le n$$

for $\delta \in [b_1, b_2]$, (4.2) can be simplified to

$$\frac{||\boldsymbol{x}-\boldsymbol{x}^{(j)}||_2}{||\boldsymbol{x}-\boldsymbol{x}^{(0)}||_2} \le 2\left(\frac{b-1}{b+1}\right)^{j-p-q} \cdot \max_{\delta\in[b_1,b_2]}\left\{\prod_{k=1}^{p}\left(\frac{\delta-\delta_k}{\delta_k}\right)\right\}. \tag{4.3}$$

For the preconditioned system, the iteration matrix $\boldsymbol{C}_n$ is given by

$$\boldsymbol{C}_n = (\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n).$$

Theorem 3.6 implies that we can choose $b_1 = 1$ and $b_2 = \beta$ in (3.13). Then, p and q are constants that are independent of n. In order to use (4.3), we need a lower bound for δ_k, $1 \le k \le p$. We note that

$$\begin{aligned} &\|(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)\|_2 \\ &\le \|\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n\|_2^{-1} \|\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n\|_2 \\ &\quad \cdot \kappa(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n), \end{aligned}$$

and there exists a constant $c_3 > 0$ independent of n such that

$$\|\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n\|_2 \le c_3 := 4 + \gamma_1\pi + \gamma_2.$$

Therefore, it remains to show that there exists $c_4 > 0$ independent of n such that

$$\left\|\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n\right\|_2 \ge c_4. \tag{4.4}$$

But this follows from the fact that

$$\|\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n\|_2 \ge \|\boldsymbol{T}_n[g_2] + h^2\gamma_2 \boldsymbol{I}_n\|_2 - \|h\gamma_1 \boldsymbol{T}_n[g_1]\|_2.$$

We remark that the singular values of $\boldsymbol{T}_n[g_2]$ and $\boldsymbol{T}_n[g_1]$ are distributed as $|g_2| = \theta^2$ and $|g_1| = |\theta|$, respectively (see [20, 27]). Therefore, for sufficiently small h, we have the inequality stated in (4.4). It follows by Theorem 2.2 that

$$\begin{aligned} \delta_k &\ge \min_{\ell} \delta_\ell \\ &= \left\|(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)\right\|_2^{-2} \\ &\ge \left(\frac{c_4}{c_3}\right)^2 16n^4 \left(1 + \frac{c_1}{\sqrt{n}}\right)^2 \left(1 + \frac{c_2}{n}\right)^2 \;=\; cn^{-4} \end{aligned}$$

for $1 \le k \le n$, where c is a positive constant. Thus, for $1 \le k \le p$ and $\delta \in [1, \beta]$, we have that

$$0 \le \frac{\delta - \delta_k}{\delta_k} \le cn^4.$$

Hence, (4.2) becomes

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(j)}||_2}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_2} < c^p n^{4p} \left(\frac{b-1}{b+1}\right)^{j-p-q}.$$

Given arbitrary tolerance $\epsilon > 0$, an upper bound for the number of iterations required to make

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(j_0)}||_2}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_2} < \epsilon$$

is therefore given by

$$j_0 \equiv p + q - \frac{p\log c + 4p\log n - \log\epsilon}{\log\left(\frac{b-1}{b+1}\right)} = \mathcal{O}(\log n).$$

Since each CGNE iteration requires $\mathcal{O}(n\log n)$ operations, the total cost of CGNE is at most $\mathcal{O}(n\log^2 n)$ arithmetical operations.

TABLE 5.1
Results for Example 5.1.

		CGNE		GMRES		BiCGSTAB	
n	E	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$
10	4.50e-03	24	12	11	8	21	10
20	8.48e-04	56	26	21	9	39	9
40	5.92e-05	154	28	40	8	67	9
80	1.05e-06	492	26	72	6	117	6
160	2.77e-09	1696	24	107	4	181	4
320	5.08e-13	*	24	153	3	261	3

TABLE 5.2
Results for Example 5.2.

		CGNE		GMRES		BiCGSTAB	
n	E	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$
8	3.14e-02	18	18	9	9	17	12
16	4.01e-03	40	28	17	12	35	14
32	3.55e-04	106	32	33	13	75	14
64	1.37e-05	312	32	64	12	138	12
128	1.18e-07	1020	30	125	10	250	10
256	1.15e-10	*	28	213	7	437	7
512	5.07e-14	*	26	373	5	921	5

5. Numerical results. In this section, we test our banded preconditioners on an SGI O2 workstation. All experiments were performed in MATLAB with a machine precision of 10^{-16}.

Our problems have homogeneous Dirichlet boundary conditions and known solutions. We apply GMRES, BiCGSTAB, and CGNE methods to

$$\boldsymbol{B}_n^{-1}\boldsymbol{A}_n\boldsymbol{x} = \boldsymbol{B}_n^{-1}\boldsymbol{b}.$$

Here $\boldsymbol{B}_n$ represents the banded preconditioner (3.1). The iterative method started with the zero vector and the vector $\boldsymbol{b}$ is given by (2.5).

Tables 5.1–5.4 list the number of matrix–vector products of $\boldsymbol{A}_n$ or $\boldsymbol{A}_n^{\mathrm{T}}$ required until the residual norms produced by the different iterative method satisfied $\|\boldsymbol{r}^{(j)}\|_2/\|\boldsymbol{r}^{(0)}\|_2 < 10^{-7}$. The symbol $*$ denotes that the method stopped without converging to the desired tolerance in 1000 iteration steps. We remark that GMRES uses one matrix–vector product per step, and BiCGSTAB and CGNE use two matrix–vector products per step. Note that the preconditioned systems need in addition the solution of $\boldsymbol{B}_n\boldsymbol{x} = \boldsymbol{y}$ or $\boldsymbol{B}_n^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{y}$. But since $\boldsymbol{B}_n$ is a tridiagonal matrix we compute the solution quickly by a permuted back-substitution algorithm as implemented in MATLAB. In the tables, the symbol $\boldsymbol{I}_n$ means that the system is solved without using a preconditioner.

In the tables, we also determine the error between the numerical approximation and the true solution at the sinc points defined as follows:

$$E := \sqrt{\sum_{k=-M}^{N} |u_k - u(x_k)|^2}.$$

TABLE 5.3
Results for Example 5.3 *with* $\kappa = 100$.

		CGNE		GMRES		BiCGSTAB	
n	E	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$
16	1.12e-01	48	34	17	13	61	19
32	2.07e-02	132	44	31	14	107	18
64	1.02e-03	420	44	52	13	206	18
128	9.77e-06	1408	38	95	12	347	16
256	1.06e-08	*	38	144	6	491	6
512	4.54e-13	*	30	206	4	890	4

TABLE 5.4
Results for Example 5.4 *for* $\kappa = 100$.

		CGNE		GMRES		BiCGSTAB	
n	E	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$	$\mathtt{I}_n$	$\mathtt{B}_n$
8	1.50e-01	18	20	9	9	29	17
16	1.06e-01	52	36	17	14	95	21
32	2.09e-02	160	56	33	17	517	29
64	1.04e-03	384	70	65	21	*	35
128	9.83e-06	*	92	129	52	*	42
256	1.02e-08	*	108	240	55	*	45
512	4.67e-13	*	102	430	6	*	12

Here we obtain this error by determining $\{u_k\}_{k=-M}^{N}$, where we solve the system (2.5) by a direct method.

In the numerical tests, we consider the following examples.

Example 5.1 (see [17, p. 119]). The discretization of

$$u''(x) + \frac{1}{6x}u'(x) - \frac{1}{x^2}u(x) = -\frac{19}{6}\sqrt{x} \quad (x \in (0,1)),$$

$$u(0) = u(1) = 0,$$

which has the solution $u(x) = x^{3/2}(1-x)$, is given by (2.3) with

$$\boldsymbol{D}_n^I \;=\; \boldsymbol{D}_n\left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'}\right] = \boldsymbol{D}_n\left[\frac{5-11x}{6}\right]$$

and

$$\boldsymbol{D}_n^{II} \;=\; \boldsymbol{D}_n\left[\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' - \frac{1}{\phi'}\left(\frac{p}{\phi'}\right)' + \frac{2q}{(\phi')^2}\right] = \boldsymbol{D}_n\left[\frac{(x-1)(12-x)}{6}\right].$$

We choose the conformal map $\phi(z) = \log(\frac{z}{1-z})$ and, as in [17, p. 119], $M = 2^l, N = \frac{3M}{2} - 1$, and $h = \frac{\pi}{\sqrt{3M}}$. This problem has a regular singular point at $x = 0$.

Example 5.2 (see [17, p. 126]). The discretization for the problem on $(0,\infty)$ given by

$$u''(x) - \frac{x}{x^2+1}u'(x) - \frac{1}{x^2+1}u(x) = \frac{2x(x^2-4)}{(x^2+1)^3} \quad (x \in (0,\infty)),$$

$$u(0) = \lim_{x\to\infty} u(x) = 0,$$

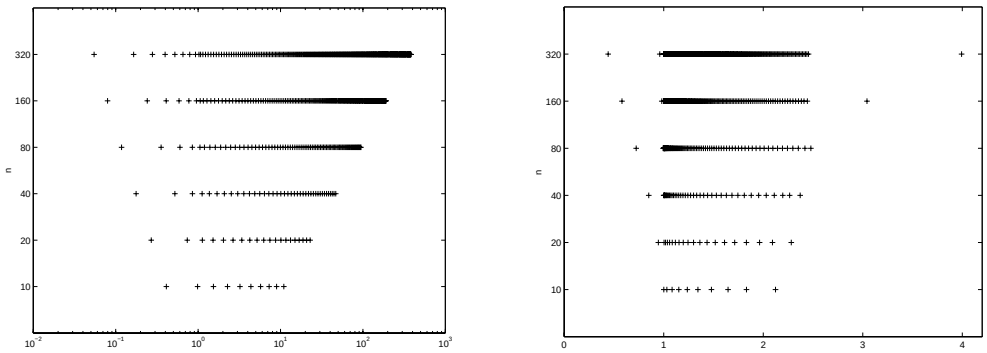


FIG. 5.1. *Singular values of* A_n *(left) and of* $\mathsf{B}_n^{-1}\mathsf{A}_n$ *(right) for* $n \in \{10, 20, 40, 80, 160, 320\}$ *given in Example* 5.1.

which has the solution $u(x) = \frac{x}{x^2+1}$, takes the form (2.3) with

$$\boldsymbol{D}_n^I = \boldsymbol{D}_n \left[\frac{2x^2+1}{x^2+1}\right] \quad \text{and} \quad \boldsymbol{D}_n^{II} = \boldsymbol{D}_n \left[\frac{-2x^4}{(x^2+1)^2}\right].$$

We choose the conformal map $\phi(z) = \log(z)$ and, as in [17, p. 126], $M = 2^l, N = M-1$, and $h = \frac{\pi}{\sqrt{2M}}$.

For Examples 5.1 and 5.2, Assumptions (A1) and (A2) are fulfilled. In Figure 5.1 we plot the singular values of $\boldsymbol{A}_n$ and of the preconditioned matrix $\boldsymbol{B}_n^{-1}\boldsymbol{A}_n$. We see that except for some outliers the singular values of $\boldsymbol{B}_n^{-1}\boldsymbol{A}_n$ lie in an fixed interval independent of n. For CGNE, our numerical results confirm our expected theoretical results, that the number of CGNE iterations is of order $\mathcal{O}(\log n)$.

We note that BiCGSTAB and GMRES use different Krylov subspaces [8, p. 90] than CGNE, and therefore we cannot compare their iteration results directly. However, we observe in all the tables that GMRES and BiCGSTAB converge very fast. These numerical results illustrate the effectiveness of our proposed preconditioners.

In the following examples we apply the banded preconditioner to precondition the sinc coefficient matrix when Assumption (A2) is not fulfilled.

Example 5.3 (see [2, 7]). We consider the convection problem

$$\begin{aligned} u''(x) - \kappa u'(x) &= f(x) \quad (x \in (0,1)), \\ u(0) &= u(1) = 0. \end{aligned} \tag{5.1}$$

The solution of (5.1) is difficult to compute for large values κ. We compute the solution for $f(x) = -\kappa$. The discretization is given by (2.3) with

$$\boldsymbol{D}_n^I = \boldsymbol{D}_n \left[1 - 2x + \kappa x(1-x)\right] \quad \text{and} \quad \boldsymbol{D}_n^{II} = \boldsymbol{D}_n \left[x(x-1)(2+\kappa(2x-1))\right].$$

We choose $\phi(z) = \log(\frac{z}{1-z})$, $h = \frac{\pi}{\sqrt{2M}}$ and $N = 2^l, M = N-1$. Note that Ernst [7] used a discretization based on the Galerkin finite element method and solved the resulting linear system by GMRES without a preconditioner.

Example 5.4 (see [2]). Consider the differential equation (for $\kappa > 0$) defined by

$$\begin{aligned} u'' - \tfrac{\kappa}{x} u'(x) &= -\kappa(\kappa+1)x^{\kappa-1} \quad (x \in (0,1)), \\ u(0) &= u(1) = 0. \end{aligned}$$

This problem has the difficulty represented by a regular singular point at $x = 0$ and a boundary layer at $x = 1$ when $\kappa \gg 0$. The linear system (1.7) takes the form (2.3) with

$$\boldsymbol{D}_n^I = \boldsymbol{D}_n \left[1 - 2x + \kappa(1 - x)\right] \quad \text{and} \quad \boldsymbol{D}_n^{II} = \boldsymbol{D}_n \left[x(x - 1)(2 + \kappa)\right].$$

We choose the conformal map $\phi(z) = \log(\frac{z}{1-z})$ and $h = \frac{\pi}{\sqrt{2M}}$ and $N = 2^l, M = N-1$.

6. Concluding remarks. We remark that the accuracy of the computed solution depends only on the Galerkin method used in the discretization of the boundary value problem. However, the convergence rate of the discrete system and the costs per iteration of the iterative method depend on how we discretize the boundary value problem. It is advantageous to use the sinc method to discretize the boundary value problem because the sinc-Galerkin method for boundary value problems is convergent exponentially (see (2.6) and Tables 5.1–5.4). However, we are required to solve n-by-n sinc systems where their coefficient matrices are dense. A straightforward application of the Gaussian elimination method will result in an algorithm, which takes $\mathcal{O}(n^3)$ arithmetical operations. The main contribution of this paper is to propose banded preconditioners to precondition sinc matrices and speed up the convergence rate of conjugate-gradient-type methods. The cost of our proposed method for sinc systems is significantly less than the $\mathcal{O}(n^3)$ cost required by the Gaussian elimination method for solving sinc systems.

Finally, we remark that we can employ the finite difference or the finite element method to discretize the boundary value problem, and therefore banded system solvers can be used to solve the corresponding linear system in $\mathcal{O}(n)$ operations. However, in order to obtain a reasonably accurate solution, a small step-size has to be used in the finite difference or the finite element method, and hence the dimension of the resulting matrix system will be very large compared to the size of the sinc system [19].

Acknowledgment. The authors would like to acknowledge R. Chan, B. Fischer, and G. Steidl for numerous fruitful and enlightening discussions. We would also like to thank the referees for their valuable suggestions.

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] T. S. Carlson, J. Lund, and K. L. Bowers, *A Sinc–Galerkin method for convection dominated transport*, in Computation and Control III, Progr. Systems Control Theory 15, Birkhäuser Boston, Boston, 1993, pp. 121–139.
[3] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
[4] R. H. Chan and K.-P. Ng, *Fast iterative solvers for Toeplitz-plus-band systems*, SIAM J. Sci. Comput., 14 (1993), pp. 1013–1019.
[5] R. H. Chan and M.-C. Yeung, *Circulant preconditioners for complex Toeplitz systems*, SIAM J. Numer. Anal., 30 (1993), pp. 1193–1207.
[6] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.
[7] O. G. Ernst, *Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1079–1101.
[8] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[9] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Los Angeles, 1958.

[10] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, 1989.
[11] G. Heinig and K. Rost, *Representations of Toeplitz–plus–Hankel matrices using trigonometric transforms with application to fast matrix-vector multiplication*, Linear Algebra Appl., 254 (1997), pp. 193–226.
[12] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[13] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[14] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.
[15] N. Levinson, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. and Phys., 25 (1946), pp. 261–278.
[16] J. Lund, *Symmetrization of the Sinc-Galerkin method for boundary value problems*, Math. Comp., 47 (1986), pp. 571–588.
[17] J. Lund and K. L. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, 1992.
[18] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795.
[19] M. Ng, *Fast iterative methods for symmetric Sinc-Galerkin systems*, IMA J. Numer. Anal., 19 (1999), pp. 357–373.
[20] S. V. Parter, *On the distribution of singular values of Toeplitz matrices*, Linear Algebra Appl., 80 (1986), pp. 115–130.
[21] D. Potts and G. Steidl, *Optimal trigonometric preconditioners for nonsymmetric Toeplitz systems*, Linear Algebra Appl., 281 (1998), pp. 265–292.
[22] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.
[23] F. Stenger, *Numerical methods based on Whittaker cardinal, or sinc functions*, SIAM Rev., 23 (1981), pp. 165–224.
[24] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer Ser. Comput. Math. 20, Springer-Verlag, New York, 1993.
[25] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1993.
[26] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.
[27] E. E. Tyrtyshnikov, *A unifying approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[28] H. Van der Vorst, *Preconditioning by Incomplete Decomposition*, Ph.D. thesis, Rijksuniversiteit Utrecht, The Netherlands, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 2, pp. 599–601

ADDENDUM TO "A KRYLOV–SCHUR ALGORITHM FOR LARGE EIGENPROBLEMS"*

G. W. STEWART†

Abstract. In this addendum to an earlier paper by the author, it is shown how to compute a Krylov decomposition corresponding to an arbitrary Rayleigh quotient. This decomposition can be used to restart an Arnoldi process, with a selection of the Ritz vectors corresponding to that Rayleigh quotient.

Key words. large eigenproblem, Krylov sequence, Arnoldi algorithm, Krylov decomposition, restarting, deflation

AMS subject classifications. 65F14, 65F50

PII. S0895479802403150

In [4] the author introduced a decomposition of the form

$$AU = UB + ub^{\mathrm{H}}, \tag{1}$$

where A is a matrix of order n and $(U\ u)$ has full column rank. It was shown that the column space of $(U\ u)$ (called the *subspace of the decomposition*) is a (possibly restarted) Krylov subspace of A and conversely that every Krylov subspace has such a representation, so that the *Krylov decomposition* (1) is a characterization of Krylov subspaces.[1] Arnoldi and Lanczos decompositions are special cases of Krylov decompositions.

The advantage of working with Krylov decompositions is that their subspaces remain invariant under two classes of transformations. The first, called a *similarity*, transforms the decomposition into

$$A(UW^{-1}) = (UW^{-1})(WBW^{-1}) + u(b^{\mathrm{H}}W^{-1}) \equiv A\tilde{U} = \tilde{U}\tilde{B} + u\tilde{b}^{\mathrm{H}},$$

where W is any nonsingular matrix. The second, called a *translation,* transforms the decomposition to the form

$$AU = U\tilde{B} + \tilde{u}\tilde{b}^{\mathrm{H}},$$

where

$$\tilde{B} = B + gb^{\mathrm{H}}, \quad \tilde{u} = \frac{u - Ug}{\gamma}, \quad \text{and} \quad \tilde{b}^{\mathrm{H}} = \gamma b^{\mathrm{H}}$$

for any vector g and any scalar $\gamma \neq 0$.

The computational algorithms in [4] were based on similarities. Translations were used primarily in the derivation of the properties of Krylov decompositions. The purpose of this note is to show that translations have a computational role to play in restarting an Arnoldi process with a selection of Rayleigh–Ritz approximations to a set of eigenvectors.

*Received by the editors February 25, 2002; accepted for publication (in revised form) by H. A. van der Vorst June 10, 2002; published electronically December 19, 2002.
http://www.siam.org/journals/simax/24-2/40315.html

†Department of Computer Science, University of Maryland, College Park MD 20742 (stewart@cs.umd.edu).

[1] A related characterization, cast in terms of subspaces, is given by Genseberger and Sleijpen [1].

The Rayleigh–Ritz method for extracting approximations to eigenvectors from a subspace does not depend on whether the subspace in question is a Krylov subspace. It can be presented in different ways. The one we give here leads most directly to the main result of this note. Let U be a basis for the subspace $\mathcal{U}$ in question, and let V be such that $V^{\mathrm{H}}U$ is nonsingular. (The space spanned by V is sometimes called the test subspace, and V itself the test matrix.) Then the matrix

$$\hat{B} = (V^{\mathrm{H}}U)^{-1}V^{\mathrm{H}}AU \tag{2}$$

has the property that if (μ, Uw) is an eigenpair of A, then (μ, w) is an eigenpair of $\hat{B}$. Specifically,

$$\hat{B}w = (V^{\mathrm{H}}U)^{-1}V^{\mathrm{H}}AUw = \mu(V^{\mathrm{H}}U)^{-1}V^{\mathrm{H}}Uw = \mu w.$$

By continuity one might expect that if $\mathcal{U}$ contains an approximate eigenvector of A, then it can be found by computing an appropriate eigenpair (μ, w) of $\hat{B}$ and forming Uw. This is the essence of the Rayleigh–Ritz method. (For an analysis of the method, see [2].) The matrix B is called a *Rayleigh quotient* (with respect to U and V) because (2) is a generalization of the ordinary Rayleigh quotient $v^{\mathrm{H}}Au/v^{\mathrm{H}}u$.

It was observed in [4] that the matrix B in the Krylov decomposition (1) is a Rayleigh quotient. Specifically, let $(V\ v)^{\mathrm{H}}$ be a left inverse of $(U\ u)$. Then $V^{\mathrm{H}}U = I$ and $V^{\mathrm{H}}u = 0$. It follows from (1) that $B = V^{\mathrm{H}}AU$ is a Rayleigh quotient, which can be used in the Rayleigh–Ritz procedure. In particular, we can discard undesirable Ritz vectors by a process known as Krylov–Schur restarting.

In some cases, however, we may not have the freedom to choose V. For example, in the harmonic Rayleigh–Ritz method, which has superior properties for approximating interior eigenvalues [3], [5, pp. 292–294], we must take $V = (A - \kappa I)U$, where κ is near the eigenvalues of interest. Now for a general test matrix V, there is no problem in computing the Rayleigh quotient. In fact, on multiplying (1) by $(V^{\mathrm{H}}U)^{-1}V^{\mathrm{H}}$, we find that

$$\hat{B} = (V^{\mathrm{H}}U)^{-1}V^{\mathrm{H}}AU = B + gb^{\mathrm{H}}, \tag{3}$$

where

$$g = (V^{\mathrm{H}}U)^{-1}V^{\mathrm{H}}u. \tag{4}$$

The problem is that $\hat{B}$ is seemingly not associated with a Krylov decomposition, so that the Krylov–Schur restarting procedure of [4] cannot be applied to remove undesirable Ritz vectors.

But in fact $\hat{B}$ *is* associated with a Krylov decomposition.

THEOREM 1. *Let $V^{\mathrm{H}}U$ be nonsingular, and let $\hat{B}$ and g be defined by* (3) *and* (4). *If*

$$\hat{u} = u - Ug,$$

then the Krylov decomposition

$$AU = U\hat{B} + \hat{u}b^{\mathrm{H}} \tag{5}$$

is a translation of the decomposition (1)*, whose Rayleigh quotient with respect to the test matrix V is $\hat{B}$.*

The proof consists of verifying that (5) is indeed a translation of (1) and that the matrix $(V^{\mathrm{H}}U)^{-1}VAU$ is indeed equal to $\hat{B}$.

To see how we can use (5) to restart the Krylov decomposition, suppose U is orthonormal (as it will be in practice). Let

$$\begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} = \begin{pmatrix} W_1^{\mathrm{H}} \\ W_2^{\mathrm{H}} \end{pmatrix} \hat{B}(W_1 \ W_2)$$

be a partition Schur decomposition of $\hat{B}$, where T_{11} contains the Ritz values corresponding to the Ritz vectors we wish to retain. Then by a similarity, we have

$$A(UW_1 \ UW_2) = (UW_1 \ UW_2) \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} + \hat{u}(b^{\mathrm{H}}W_1 \ b^{\mathrm{H}}W_2).$$

Hence

$$A(UW_1) = (UW_1)T_{11} + \hat{u}b^{\mathrm{H}}W_1$$

is a Krylov decomposition containing the desired Ritz vectors. The matrix UW_1 is orthonormal, but the vector $\hat{u}$ will not in general be orthogonal to the columns of UW_1. However, by a second translation we can orthogonalize it. The resulting decomposition is an orthogonal Krylov decomposition, which can be extended by the Arnoldi process in the usual way.

Acknowledgment. I am indebted to the Mathematical and Computational Sciences Division of the National Institute of Standards and Technology for the use of their research facilities.

REFERENCES

[1] M. Genseberger and G. L. G. Sleijpen, *Alternative correction equations in the Jacobi–Davidson method*, Numer. Linear Algebra Appl., 6 (1999), pp. 235–253.

[2] Z. Jia and G. W. Stewart, *An analysis of the Rayleigh–Ritz method for approximating eigenspaces*, Math. Comp., 70 (2001), pp. 637–647.

[3] R. B. Morgan, *Computing interior eigenvalues of large matrices*, Linear Algebra Appl., 154–156 (1991), pp. 289–309.

[4] G. W. Stewart, *A Krylov–Schur algorithm for large eigenproblems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 601–614.

[5] G. W. Stewart, *Matrix Algorithms Volume* II: *Eigensystems*, SIAM, Philadelphia, 2001.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 603–626

HERMITIAN AND SKEW-HERMITIAN SPLITTING METHODS FOR NON-HERMITIAN POSITIVE DEFINITE LINEAR SYSTEMS*

ZHONG-ZHI BAI†, GENE H. GOLUB‡, AND MICHAEL K. NG§

Abstract. We study efficient iterative methods for the large sparse non-Hermitian positive definite system of linear equations based on the Hermitian and skew-Hermitian splitting of the coefficient matrix. These methods include a Hermitian/skew-Hermitian splitting (HSS) iteration and its inexact variant, the inexact Hermitian/skew-Hermitian splitting (IHSS) iteration, which employs some Krylov subspace methods as its inner iteration processes at each step of the outer HSS iteration. Theoretical analyses show that the HSS method converges unconditionally to the unique solution of the system of linear equations. Moreover, we derive an upper bound of the contraction factor of the HSS iteration which is dependent solely on the spectrum of the Hermitian part and is independent of the eigenvectors of the matrices involved. Numerical examples are presented to illustrate the effectiveness of both HSS and IHSS iterations. In addition, a model problem of a three-dimensional convection-diffusion equation is used to illustrate the advantages of our methods.

Key words. non-Hermitian matrix, splitting, Hermitian matrix, skew-Hermitian matrix, iterative methods

AMS subject classifications. 65F10, 65F15, 65T10

PII. S0895479801395458

1. Introduction.

Many problems in scientific computing give rise to a system of linear equations

$$Ax = b, \quad A \in \mathbb{C}^{n\times n} \text{ nonsingular}, \quad \text{and } x, b \in \mathbb{C}^n, \tag{1.1}$$

with A a large sparse non-Hermitian and positive definite matrix.

Iterative methods for the system of linear equations (1.1) require efficient splittings of the coefficient matrix A. For example, the Jacobi and the Gauss–Seidel iterations [16] split the matrix A into its diagonal and off-diagonal (respectively, strictly lower and upper triangular) parts, and the generalized conjugate gradient (CG) method [7] and the generalized Lanczos method [27] split the matrix A into its Hermitian and skew-Hermitian parts; see also [11, 17, 26, 1] and [2], respectively. Because the matrix A naturally possesses a Hermitian/skew-Hermitian splitting (HSS) [7]

$$A = H + S, \tag{1.2}$$

*Received by the editors September 24, 2001; accepted for publication (in revised form) by D. P. O'Leary July 23, 2002; published electronically January 17, 2003.

http://www.siam.org/journals/simax/24-3/39545.html

†State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, The People's Republic of China (bzz@lsec.cc.ac.cn). The research of this author was subsidized by The Special Funds for Major State Basic Research Projects G1999032803.

‡Department of Computer Science, Stanford University, Stanford, CA 94305 (golub@sccm.stanford.edu). The work of this author was supported in part by DOE-FC02-01ER4177.

§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 Hong Kong Research Grants Council grants HKU 7132/00P and 7130/02P, and by HKU CRCG grants 10203501, 10203907, and 10203408.

where

$$H = \frac{1}{2}(A + A^*) \quad \text{and} \quad S = \frac{1}{2}(A - A^*), \tag{1.3}$$

we will study in this paper efficient iterative methods based on this particular matrix splitting for solving the system of linear equations (1.1).

Now $A = H(I + H^{-1}S)$, and thus $A^{-1} = (I + H^{-1}S)^{-1}H^{-1}$. Thus, if we replace $(I + H^{-1}S)^{-1}$ by its first order approximation $I - H^{-1}S$, then $(I - H^{-1}S)H^{-1}$ could be employed as a preconditioner to the matrix A. Of course, the preconditioning effect is completely determined by the spectral distribution of the matrix $H^{-1}S$, and it is satisfactory if the Hermitian part H is dominant [1]. On the other hand, if the skew-Hermitian part S is dominant, we can use an alternative preconditioning strategy recently presented by Golub and Vanderstraeten in [15]. Their basic idea is to invert the shifted skew-Hermitian matrix $\alpha I + S$ and then employ $(I - (S + \alpha I)^{-1}(H - \alpha I))(S + \alpha I)^{-1}$ as a preconditioner to the matrix A. In fact, the preconditioning effect for this preconditioner depends not only on the spectrum but also on the eigenvectors of the matrix $(S + \alpha I)^{-1}(H - \alpha I)$, which is, however, closely related to the shift α. For a nearly optimal α, numerical experiments in [15] on a variety of problems from real-world applications have shown that the reductions in terms of iteration count largely compensate for the additional work per iteration when compared to standard preconditioners. We remark that, for both preconditioners, exact inverses of the matrices H and $\alpha I + S$ are quite expensive, and, therefore, some further approximations, e.g., the incomplete Cholesky (IC) factorization [21, 20] and the incomplete orthogonal-triangular (IQR) factorization [3], to these two matrices may be respectively adopted in actual applications. However, theoretical analysis about existence, stability, and accuracy of the resulting iterative method are considerably difficult.

Moreover, based on the HSS (1.2)–(1.3), in this paper we present a different approach to solve the system of linear equations (1.1), called the HSS iteration, and it is as follows.

The HSS iteration method. *Given an initial guess $x^{(0)}$, for $k = 0, 1, 2, \ldots$, until $\{x^{(k)}\}$ converges, compute*

$$\begin{cases} (\alpha I + H)x^{(k+\frac{1}{2})} &= (\alpha I - S)x^{(k)} + b, \\ (\alpha I + S)x^{(k+1)} &= (\alpha I - H)x^{(k+\frac{1}{2})} + b, \end{cases} \tag{1.4}$$

where α is a given positive constant.

Evidently, each iterate of the HSS iteration alternates between the Hermitian part H and the skew-Hermitian part S of the matrix A, analogously to the classical alternating direction implicit (ADI) iteration for solving partial differential equations; see Peaceman and Rachford [23] and Douglas and Rachford [8]. Results associated to the stationary iterative method with alternation can be also found in Benzi and Szyld [4]. Theoretical analysis shows that the HSS iteration (1.4) converges unconditionally to the unique solution of the system of linear equations (1.1). The upper bound of the contraction factor of the HSS iteration is dependent on the spectrum of the Hermitian part H but is independent of the spectrum of the skew-Hermitian part S as well as the eigenvectors of the matrices H, S, and A. In addition, the optimal value of the parameter α for the upper bound of the contraction factor of the HSS iteration can be determined by the lower and the upper eigenvalue bounds of the matrix H.

Note that we can reverse the roles of the matrices H and S in the above HSS iteration method so that we may first solve the system of linear equations with coef-

ficient matrix $\alpha I + S$ and then solve the system of linear equations with coefficient matrix $\alpha I + H$.

The two half-steps at each HSS iterate require exact solutions with the n-by-n matrices $\alpha I + H$ and $\alpha I + S$. However, this is very costly and impractical in actual implementations. To further improve the computing efficiency of the HSS iteration, we can employ, for example, the CG method to solve the system of linear equations with coefficient matrix $\alpha I + H$ and some Krylov subspace method to solve the system of linear equations with coefficient matrix $\alpha I + S$ to some prescribed accuracy at each step of the HSS iteration. Other possible choices of inner iteration solvers are classical relaxation methods, multigrid methods or multilevel methods, etc. This results in an inexact Hermitian/skew-Hermitian splitting (IHSS) iteration. The tolerances (or numbers of inner iteration steps) for inner iterative methods may be different and may be changed according to the outer iteration scheme. Therefore, the IHSS iteration is actually a nonstationary iterative method for solving the system of linear equations (1.1).

Model problem analysis for a three-dimensional convection-diffusion equation and numerical implementations show that both HSS and IHSS iterations are feasible and efficient for solving the non-Hermitian positive definite system of linear equations (1.1).

The organization of this paper is as follows. In section 2, we study the convergence properties and analyze the convergence rate of the HSS iteration. In section 3, we establish the IHSS iteration and study its convergence property. The three-dimensional convection-diffusion equation is employed as a model problem to give intuitive illustration for the convergence theory for the HSS iteration in section 4. Numerical experiments are presented in section 5 to show the effectiveness of our methods. And, finally, in section 6, we draw a brief conclusion and include some remarks. Moreover, the basic lemma used in the model problem analysis in section 4 and some illustrative remarks can be found in the appendix.

2. Convergence analysis of the HSS iteration. In this section, we study the convergence rate of the HSS iteration. We first note that the HSS iteration method can be generalized to the two-step splitting iteration framework, and the following lemma describes a general convergence criterion for a two-step splitting iteration.

LEMMA 2.1. *Let $A \in \mathbb{C}^{n\times n}$, $A = M_i - N_i$ $(i = 1, 2)$ be two splittings*[1] *of the matrix A, and let $x^{(0)} \in \mathbb{C}^n$ be a given initial vector. If $\{x^{(k)}\}$ is a two-step iteration sequence defined by*

$$\begin{cases} M_1 x^{(k+\frac{1}{2})} &= N_1 x^{(k)} + b, \\ M_2 x^{(k+1)} &= N_2 x^{(k+\frac{1}{2})} + b, \end{cases}$$

$k = 0, 1, 2, \ldots$, then

$$x^{(k+1)} = M_2^{-1} N_2 M_1^{-1} N_1 x^{(k)} + M_2^{-1}(I + N_2 M_1^{-1})b, \quad k = 0, 1, 2, \ldots .$$

Moreover, if the spectral radius $\rho(M_2^{-1} N_2 M_1^{-1} N_1)$ of the iteration matrix $M_2^{-1} N_2 M_1^{-1} N_1$ is less than 1, then the iterative sequence $\{x^{(k)}\}$ converges to the unique solution $x^* \in \mathbb{C}^n$ of the system of linear equations (1.1) for all initial vectors $x^{(0)} \in \mathbb{C}^n$.

For the convergence property of the HSS iteration, we apply the above results to obtain the following main theorem.

[1] Here and in what follows, $A = M - N$ is called a splitting of the matrix A if M is a nonsingular matrix.

THEOREM 2.2. *Let $A \in \mathbb{C}^{n\times n}$ be a positive definite matrix, let $H = \frac{1}{2}(A + A^*)$ and $S = \frac{1}{2}(A - A^*)$ be its Hermitian and skew-Hermitian parts, and let α be a positive constant. Then the iteration matrix $M(\alpha)$ of the HSS iteration is given by*

$$M(\alpha) = (\alpha I + S)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S), \tag{2.1}$$

and its spectral radius $\rho(M(\alpha))$ is bounded by

$$\sigma(\alpha) \equiv \max_{\lambda_i \in \lambda(H)} \left| \frac{\alpha - \lambda_i}{\alpha + \lambda_i} \right|,$$

where $\lambda(H)$ is the spectral set of the matrix H. Therefore, it holds that

$$\rho(M(\alpha)) \le \sigma(\alpha) < 1 \quad \forall \alpha > 0;$$

i.e., the HSS iteration converges to the unique solution $x^* \in \mathbb{C}^n$ of the system of linear equations (1.1).

Proof. By putting

$$M_1 = \alpha I + H, \quad N_1 = \alpha I - S, \quad M_2 = \alpha I + S, \quad \text{and} \quad N_2 = \alpha I - H$$

in Lemma 2.1 and noting that $\alpha I + H$ and $\alpha I + S$ are nonsingular for any positive constant α, we obtain (2.1).

By the similarity invariance of the matrix spectrum, we have

$$\begin{aligned}\rho(M(\alpha)) &= \rho((\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S)(\alpha I + S)^{-1}) \\ &\le \|(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S)(\alpha I + S)^{-1}\|_2 \\ &\le \|(\alpha I - H)(\alpha I + H)^{-1}\|_2 \|(\alpha I - S)(\alpha I + S)^{-1}\|_2.\end{aligned}$$

Letting $Q(\alpha) = (\alpha I - S)(\alpha I + S)^{-1}$ and noting that $S^* = -S$, we see that

$$\begin{aligned}Q(\alpha)^* Q(\alpha) &= (\alpha I - S)^{-1}(\alpha I + S)(\alpha I - S)(\alpha I + S)^{-1} \\ &= (\alpha I - S)^{-1}(\alpha I - S)(\alpha I + S)(\alpha I + S)^{-1} \;=\; I.\end{aligned}$$

That is, $Q(\alpha)$ is a unitary matrix. ($Q(\alpha)$ is also called the Cayley transform of S.) Therefore, $\|Q(\alpha)\|_2 = 1$. It then follows that

$$\rho(M(\alpha)) \le \|(\alpha I - H)(\alpha I + H)^{-1}\|_2 = \max_{\lambda_i \in \lambda(H)} \left| \frac{\alpha - \lambda_i}{\alpha + \lambda_i} \right|.$$

Since $\lambda_i > 0 (i = 1, 2, \ldots, n)$ and α is a positive constant, it is easy to see that $\rho(M(\alpha)) \le \sigma(\alpha) < 1$. ∎

Theorem 2.2 shows that the convergence speed of the HSS iteration is bounded by $\sigma(\alpha)$, which depends only on the spectrum of the Hermitian part H but does not depend on the spectrum of the skew-Hermitian part S, on the spectrum of the coefficient matrix A, or on the eigenvectors of the matrices H, S, and A.

Now, if we introduce a vector norm $|||x||| = \|(\alpha I + S)x\|_2$ (for all $x \in \mathbb{C}^n$) and represent the induced matrix norm by $|||X||| = \|(\alpha I + S)X(\alpha I + S)^{-1}\|_2$ (for all $X \in \mathbb{C}^{n\times n}$), then, from the proof of Theorem 2.2, we see that

$$|||M(\alpha)||| = \|(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S)(\alpha I + S)^{-1}\|_2 \le \sigma(\alpha),$$

and it follows that

$$|||x^{(k+1)} - x^*||| \le \sigma(\alpha)|||x^{(k)} - x^*|||, \qquad k = 0, 1, 2, \ldots .$$

Therefore, $\sigma(\alpha)$ is also an upper bound of the contraction factor of the HSS iteration in the sense of the $|||\cdot|||$-norm.

We remark that if the minimum and the maximum eigenvalues of the Hermitian part H are known, then the optimal parameter α for $\sigma(\alpha)$ (or the upper bound of $\rho(M(\alpha))$ or $|||M(\alpha)|||$) can be obtained. This fact is precisely stated as the following corollary.

COROLLARY 2.3. *Let $A \in \mathbb{C}^{n\times n}$ be a positive definite matrix, let $H = \frac{1}{2}(A + A^*)$ and $S = \frac{1}{2}(A - A^*)$ be its Hermitian and skew-Hermitian parts, and let $\gamma_{\min}$ and $\gamma_{\max}$ be the minimum and the maximum eigenvalues of the matrix H, respectively, and let α be a positive constant. Then*

$$\alpha^* \equiv \arg\min_{\alpha} \left\{ \max_{\gamma_{\min} \le \lambda \le \gamma_{\max}} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| \right\} = \sqrt{\gamma_{\min}\gamma_{\max}},$$

and

$$\sigma(\alpha^*) = \frac{\sqrt{\gamma_{\max}} - \sqrt{\gamma_{\min}}}{\sqrt{\gamma_{\max}} + \sqrt{\gamma_{\min}}} = \frac{\sqrt{\kappa(H)} - 1}{\sqrt{\kappa(H)} + 1},$$

where $\kappa(H)$ is the spectral condition number of H.

Proof. Now,

$$\sigma(\alpha) = \max\left\{ \left| \frac{\alpha - \gamma_{\min}}{\alpha + \gamma_{\min}} \right|, \left| \frac{\alpha - \gamma_{\max}}{\alpha + \gamma_{\max}} \right| \right\}. \tag{2.2}$$

To compute an approximate optimal $\alpha > 0$ such that the convergence factor $\rho(M(\alpha))$ of the HSS iteration is minimized, we can minimize the upper bound $\sigma(\alpha)$ of $\rho(M(\alpha))$ instead. If α^* is such a minimum point, then it must satisfy $\alpha^* - \gamma_{\min} > 0$, $\alpha^* - \gamma_{\max} < 0$, and

$$\frac{\alpha^* - \gamma_{\min}}{\alpha^* + \gamma_{\min}} = \frac{\gamma_{\max} - \alpha^*}{\gamma_{\max} + \alpha^*}.$$

Therefore,

$$\alpha^* = \sqrt{\gamma_{\min}\gamma_{\max}},$$

and the result follows. ∎

We emphasize that, in Corollary 2.3, the optimal parameter α^* minimizes only the upper bound $\sigma(\alpha)$ of the spectral radius of the iteration matrix but does not minimize the spectral radius itself; for an illustration of this phenomenon, see, e.g., Table 2.

Corollary 2.3 shows that, when the so-called optimal parameter α^* is employed, the upper bound of the convergence rate of the HSS iteration is about the same as that of the CG method, and it does become the same when, in particular, the coefficient matrix A is Hermitian. It should be mentioned that, when the coefficient matrix A is normal, we have $HS = SH$, and, therefore, $\rho(M(\alpha)) = |||M(\alpha)||| = \sigma(\alpha)$. The optimal parameter α^* then minimizes all of these three quantities.

3. The IHSS iteration. The two half-steps at each step of the HSS iteration require finding solutions with the n-by-n matrices $\alpha I + H$ and $\alpha I + S$, which is, however, very costly and impractical in actual implementations. To overcome this disadvantage and further improve the efficiency of the HSS iteration, we can solve the two subproblems iteratively. More specifically, we may employ the CG method to solve the system of linear equations with coefficient matrix $\alpha I + H$, because $\alpha I + H$ is Hermitian positive definite, and some Krylov subspace method [7, 24, 18] to solve the system of linear equations with coefficient matrix $\alpha I + S$. This results in the following IHSS iteration for solving the system of linear equations (1.1).

The IHSS iteration method. *Given an initial guess $\bar{x}^{(0)}$, for $k = 0, 1, 2, \ldots$, until $\{\bar{x}^{(k)}\}$ converges, solve $\bar{x}^{(k+\frac{1}{2})}$ approximately from*

$$(\alpha I + H)\bar{x}^{(k+\frac{1}{2})} \approx (\alpha I - S)\bar{x}^{(k)} + b$$

by employing an inner iteration (e.g., the CG method) with $\bar{x}^{(k)}$ as the initial guess; then solve $\bar{x}^{(k+1)}$ approximately from

$$(\alpha I + S)\bar{x}^{(k+1)} \approx (\alpha I - H)\bar{x}^{(k+\frac{1}{2})} + b$$

by employing an inner iteration (e.g., some Krylov subspace method) with $\bar{x}^{(k+\frac{1}{2})}$ as the initial guess, where α is a given positive constant.

To simplify numerical implementation and convergence analysis, we may rewrite the above IHSS iteration method as the following equivalent scheme.

Given an initial guess $\bar{x}^{(0)}$, for $k = 0, 1, 2, \ldots$, until $\{\bar{x}^{(k)}\}$ converges,

1. *approximate the solution of $(\alpha I + H)\bar{z}^{(k)} = \bar{r}^{(k)}$ ($\bar{r}^{(k)} = b - A\bar{x}^{(k)}$) by iterating until $\bar{z}^{(k)}$ is such that the residual*

$$\bar{p}^{(k)} = \bar{r}^{(k)} - (\alpha I + H)\bar{z}^{(k)} \tag{3.1}$$

   satisfies

$$\|\bar{p}^{(k)}\| \le \varepsilon_k \|\bar{r}^{(k)}\|,$$

   and then compute $\bar{x}^{(k+\frac{1}{2})} = \bar{x}^{(k)} + \bar{z}^{(k)}$;

2. *approximate the solution of $(\alpha I + S)\bar{z}^{(k+\frac{1}{2})} = \bar{r}^{(k+\frac{1}{2})}$ ($\bar{r}^{(k+\frac{1}{2})} = b - A\bar{x}^{(k+\frac{1}{2})}$) by iterating until $\bar{z}^{(k+\frac{1}{2})}$ is such that the residual*

$$\bar{q}^{(k+\frac{1}{2})} = \bar{r}^{(k+\frac{1}{2})} - (\alpha I + S)\bar{z}^{(k+\frac{1}{2})} \tag{3.2}$$

   satisfies

$$\|\bar{q}^{(k+\frac{1}{2})}\| \le \eta_k \|\bar{r}^{(k+\frac{1}{2})}\|,$$

   and then compute $\bar{x}^{(k+1)} = \bar{x}^{(k+\frac{1}{2})} + \bar{z}^{(k+\frac{1}{2})}$. Here $\|\cdot\|$ is a norm of a vector.

In the following theorem, we analyze the above IHSS iteration method in slightly more general terms. In particular, we consider inexact iterations for the two-step splitting technique (cf. Lemma 2.1). To this end, we generalize the norm $|||\cdot|||$ to $|||\cdot|||_{M_2}$, which is defined by $|||x|||_{M_2} = \|M_2 x\|$ (for all $x \in \mathbb{C}^n$), which immediately induces the matrix norm $|||X|||_{M_2} = \|M_2 X M_2^{-1}\|$ (for all $X \in \mathbb{C}^{n\times n}$).

THEOREM 3.1. *Let $A \in \mathbb{C}^{n\times n}$ and $A = M_i - N_i$ $(i = 1, 2)$ be two splittings of the matrix A. If $\{\bar{x}^{(k)}\}$ is an iterative sequence defined as*

$$\bar{x}^{(k+\frac{1}{2})} = \bar{x}^{(k)} + \bar{z}^{(k)}, \quad \text{with} \quad M_1 \bar{z}^{(k)} = \bar{r}^{(k)} + \bar{p}^{(k)}, \tag{3.3}$$

satisfying $\frac{\|\bar{p}^{(k)}\|}{\|\bar{r}^{(k)}\|} \le \varepsilon_k$, *where* $\bar{r}^{(k)} = b - A\bar{x}^{(k)}$, *and*

$$\bar{x}^{(k+1)} = \bar{x}^{(k+\frac{1}{2})} + \bar{z}^{(k+\frac{1}{2})}, \quad \textit{with} \quad M_2\bar{z}^{(k+\frac{1}{2})} = \bar{r}^{(k+\frac{1}{2})} + \bar{q}^{(k+\frac{1}{2})}, \tag{3.4}$$

satisfying $\frac{\|\bar{q}^{(k+\frac{1}{2})}\|}{\|\bar{r}^{(k+\frac{1}{2})}\|} \le \eta_k$, *where* $\bar{r}^{(k+\frac{1}{2})} = b - A\bar{x}^{(k+\frac{1}{2})}$, *then* $\{\bar{x}^{(k)}\}$ *is of the form*

$$\begin{aligned} \bar{x}^{(k+1)} &= M_2^{-1}N_2M_1^{-1}N_1\bar{x}^{(k)} + M_2^{-1}(I + N_2M_1^{-1})b \\ &\quad + M_2^{-1}(N_2M_1^{-1}\bar{p}^{(k)} + \bar{q}^{(k+\frac{1}{2})}). \end{aligned} \tag{3.5}$$

Moreover, if $x^* \in \mathbb{C}^n$ *is the exact solution of the system of linear equations* (1.1), *then we have*

$$|||\bar{x}^{(k+1)} - x^*|||_{M_2} \le \left(\bar{\sigma} + \bar{\mu}\bar{\theta}\bar{\varepsilon}_k + \bar{\theta}(\bar{\rho} + \bar{\theta}\bar{\nu}\varepsilon_k)\eta_k\right) |||\bar{x}^{(k)} - x^*|||_{M_2}, \quad k = 0, 1, 2, \dots, \tag{3.6}$$

where

$$\begin{aligned} \bar{\sigma} &= \|N_2M_1^{-1}N_1M_2^{-1}\|, \quad \bar{\rho} = \|M_2M_1^{-1}N_1M_2^{-1}\|, \quad \bar{\mu} = \|N_2M_1^{-1}\|, \\ \bar{\theta} &= \|AM_2^{-1}\|, \qquad\qquad \bar{\nu} = \|M_2M_1^{-1}\|. \end{aligned}$$

In particular, if

$$\bar{\sigma} + \bar{\mu}\bar{\theta}\bar{\varepsilon}_{\max} + \bar{\theta}(\bar{\rho} + \bar{\theta}\bar{\nu}\varepsilon_{\max})\eta_{\max} < 1, \tag{3.7}$$

then the iterative sequence $\{\bar{x}^{(k)}\}$ *converges to* $x^* \in \mathbb{C}^n$, *where* $\varepsilon_{\max} = \max_k\{\varepsilon_k\}$ *and* $\eta_{\max} = \max_k\{\eta_k\}$.

Proof. From (3.3), we obtain

$$\begin{aligned} \bar{x}^{(k+\frac{1}{2})} &= \bar{x}^{(k)} + M_1^{-1}(\bar{r}^{(k)} + \bar{p}^{(k)}) \\ &= (I - M_1^{-1}A)\bar{x}^{(k)} + M_1^{-1}b + M_1^{-1}\bar{p}^{(k)} \\ &= M_1^{-1}N_1\bar{x}^{(k)} + M_1^{-1}b + M_1^{-1}\bar{p}^{(k)}. \end{aligned} \tag{3.8}$$

Similarly, from (3.4), we get

$$\begin{aligned} \bar{x}^{(k+1)} &= \bar{x}^{(k+\frac{1}{2})} + M_2^{-1}(\bar{r}^{(k+\frac{1}{2})} + \bar{q}^{(k+\frac{1}{2})}) \\ &= (I - M_2^{-1}A)\bar{x}^{(k+\frac{1}{2})} + M_2^{-1}b + M_2^{-1}\bar{q}^{(k+\frac{1}{2})} \\ &= M_2^{-1}N_2\bar{x}^{(k+\frac{1}{2})} + M_2^{-1}b + M_2^{-1}\bar{q}^{(k+\frac{1}{2})}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \bar{x}^{(k+1)} &= M_2^{-1}N_2(M_1^{-1}N_1\bar{x}^{(k)} + M_1^{-1}b + M_1^{-1}\bar{p}^{(k)}) \\ &\quad + M_2^{-1}b + M_2^{-1}\bar{q}^{(k+\frac{1}{2})} \\ &= M_2^{-1}N_2M_1^{-1}N_1\bar{x}^{(k)} + M_2^{-1}(I + N_2M_1^{-1})b \\ &\quad + M_2^{-1}(N_2M_1^{-1}\bar{p}^{(k)} + \bar{q}^{(k+\frac{1}{2})}), \end{aligned} \tag{3.9}$$

which is exactly (3.5).

Because $x^* \in \mathbb{C}^n$ is the exact solution of the system of linear equations (1.1), it must satisfy

$$x^* = M_1^{-1}N_1x^* + M_1^{-1}b \tag{3.10}$$

and

$$x^* = M_2^{-1}N_2M_1^{-1}N_1x^* + M_2^{-1}(I + N_2M_1^{-1})b. \tag{3.11}$$

By subtracting (3.10) from (3.8) and (3.11) from (3.9), respectively, we have

$$\bar{x}^{(k+\frac{1}{2})} - x^* = M_1^{-1}N_1(\bar{x}^{(k)} - x^*) + M_1^{-1}\bar{p}^{(k)} \tag{3.12}$$

and

$$\bar{x}^{(k+1)} - x^* = M_2^{-1}N_2M_1^{-1}N_1(\bar{x}^{(k)} - x^*) + M_2^{-1}(N_2M_1^{-1}\bar{p}^{(k)} + \bar{q}^{(k+\frac{1}{2})}). \tag{3.13}$$

Taking norms on both sides of the identities (3.12) and (3.13), we can obtain

$$\begin{array}{rcl} |||\bar{x}^{(k+\frac{1}{2})} - x^*|||_{M_2} & \leq & |||M_1^{-1}N_1(\bar{x}^{(k)} - x^*)|||_{M_2} + |||M_1^{-1}\bar{p}^{(k)}|||_{M_2} \\ & \leq & |||M_1^{-1}N_1|||_{M_2}|||\bar{x}^{(k)} - x^*|||_{M_2} + |||M_1^{-1}\bar{p}^{(k)}|||_{M_2} \\ & \leq & \|M_2M_1^{-1}N_1M_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +\|M_2M_1^{-1}\| \cdot \|\bar{p}^{(k)}\| \end{array} \tag{3.14}$$

and

$$\begin{array}{rcl} |||\bar{x}^{(k+1)} - x^*|||_{M_2} & \leq & |||M_2^{-1}N_2M_1^{-1}N_1|||_{M_2}|||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +|||M_2^{-1}(N_2M_1^{-1}\bar{p}^{(k)} + \bar{q}^{(k+\frac{1}{2})})|||_{M_2} \\ & = & \|N_2M_1^{-1}N_1M_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +\|N_2M_1^{-1}\bar{p}^{(k)} + \bar{q}^{(k+\frac{1}{2})}\| \\ & \leq & \|N_2M_1^{-1}N_1M_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +\|N_2M_1^{-1}\| \cdot \|\bar{p}^{(k)}\| + \|\bar{q}^{(k+\frac{1}{2})}\|. \end{array} \tag{3.15}$$

Noticing that

$$\|\bar{r}^{(k)}\| = \|b - A\bar{x}^{(k)}\| = \|A(x^* - \bar{x}^{(k)})\| \leq \|AM_2^{-1}\| \cdot |||x^* - \bar{x}^{(k)}|||_{M_2}$$

and

$$\|\bar{r}^{(k+\frac{1}{2})}\| = \|b - A\bar{x}^{(k+\frac{1}{2})}\| = \|A(x^* - \bar{x}^{(k+\frac{1}{2})})\| \leq \|AM_2^{-1}\| \cdot |||x^* - \bar{x}^{(k+\frac{1}{2})}|||_{M_2},$$

by (3.12), (3.14), and the definitions of the sequences $\{\bar{p}^{(k)}\}$ and $\{\bar{q}^{(k+\frac{1}{2})}\}$, we have

$$\|\bar{p}^{(k)}\| \leq \varepsilon_k\|\bar{r}^{(k)}\| \leq \varepsilon_k\|AM_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \tag{3.16}$$

and

$$\begin{array}{rcl} \|\bar{q}^{(k+\frac{1}{2})}\| & \leq & \eta_k\|\bar{r}^{(k+\frac{1}{2})}\| \\ & \leq & \eta_k\|AM_2^{-1}\|(\|M_2M_1^{-1}N_1M_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +\|M_2M_1^{-1}\| \cdot \|\bar{p}^{(k)}\|) \\ & \leq & \eta_k\|AM_2^{-1}\|(\|M_2M_1^{-1}N_1M_2^{-1}\| \\ & & +\varepsilon_k\|M_2M_1^{-1}\| \cdot \|AM_2^{-1}\|)|||\bar{x}^{(k)} - x^*|||_{M_2}. \end{array} \tag{3.17}$$

Through substituting (3.16) and (3.17) into (3.15), we finally obtain

$$\begin{array}{rcl} |||\bar{x}^{(k+1)} - x^*|||_{M_2} & \leq & \|N_2M_1^{-1}N_1M_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +\|N_2M_1^{-1}\| \cdot \varepsilon_k\|AM_2^{-1}\| \cdot |||\bar{x}^{(k)} - x^*|||_{M_2} \\ & & +\eta_k\|AM_2^{-1}\|(\|M_2M_1^{-1}N_1M_2^{-1}\| \\ & & +\varepsilon_k\|M_2M_1^{-1}\| \cdot \|AM_2^{-1}\|)|||\bar{x}^{(k)} - x^*|||_{M_2} \\ & \leq & \left(\bar{\sigma} + \bar{\mu}\bar{\theta}\varepsilon_k + \bar{\theta}(\bar{\rho} + \bar{\theta}\bar{\nu}\varepsilon_k)\eta_k\right)|||\bar{x}^{(k)} - x^*|||_{M_2}. \end{array}$$

□

We remark that, if the inner systems can be solved exactly in some applications, the corresponding quantities $\{\varepsilon_k\}$ and $\{\eta_k\}$, and hence $\varepsilon_{\max}$ and $\eta_{\max}$, are equal to zero. It then follows that the convergence rate of the IHSS iteration reduces to the same as that of the HSS iteration. In general, Theorem 3.1 shows that, in order to guarantee the convergence of the IHSS iteration, it is not necessary for $\{\varepsilon_k\}$ and $\{\eta_k\}$ to approach zero as k is increasing. All we need is that the condition (3.7) is satisfied.

By specializing Theorem 3.1 to the shifted Hermitian and skew-Hermitian splittings

$$\begin{aligned} A &= M_1 - N_1 \equiv (\alpha I + H) - (\alpha I - S) \\ &= M_2 - N_2 \equiv (\alpha I + S) - (\alpha I - H), \end{aligned}$$

we straightforwardly obtain the following convergence theorem about the IHSS iteration method.

THEOREM 3.2. *Let $A \in \mathbb{C}^{n\times n}$ be a positive definite matrix, let $H = \frac{1}{2}(A + A^*)$ and $S = \frac{1}{2}(A - A^*)$ be its Hermitian and skew-Hermitian parts, and let α be a positive constant. If $\{\bar{x}^{(k)}\}$ is an iterative sequence generated by the IHSS iteration method (cf.* (3.1) *and* (3.2)*) and if $x^* \in \mathbb{C}^n$ is the exact solution of the system of linear equations* (1.1)*, then it holds that*

$$|||\bar{x}^{(k+1)} - x^*||| \le (\sigma(\alpha) + \theta\rho\eta_k)(1 + \theta\varepsilon_k)|||\bar{x}^{(k)} - x^*|||, \quad k = 0, 1, 2, \dots,$$

where

$$\rho = \|(\alpha I + S)(\alpha I + H)^{-1}\|_2, \quad \theta = \|A(\alpha I + S)^{-1}\|_2. \tag{3.18}$$

In particular, if $(\sigma(\alpha) + \theta\rho\eta_{\max})(1 + \theta\varepsilon_{\max}) < 1$, then the iterative sequence $\{\bar{x}^{(k)}\}$ converges to $x^* \in \mathbb{C}^n$, where $\varepsilon_{\max} = \max_k\{\varepsilon_k\}$ and $\eta_{\max} = \max_k\{\eta_k\}$.

According to Theorem 3.1, we want to choose tolerances so that the computational work of the two-step splitting iteration method is minimized. In fact, as we have remarked previously, the tolerances $\{\varepsilon_k\}$ and $\{\eta_k\}$ are not required to approach zero as k increases in order to get the convergence of the IHSS iteration but are required to approach zero in order to asymptotically recover the original convergence rate (cf. Theorem 2.2) of the HSS iteration.

The following theorem presents one possible way of choosing the tolerances $\{\varepsilon_k\}$ and $\{\eta_k\}$ such that the original convergence rate (cf. Lemma 2.1) of the two-step splitting iterative scheme can be asymptotically recovered.

THEOREM 3.3. *Let the assumptions in Theorem* 3.1 *be satisfied. Suppose that both $\{\tau_1(k)\}$ and $\{\tau_2(k)\}$ are nondecreasing and positive sequences satisfying $\tau_1(k) \ge 1$, $\tau_2(k) \ge 1$, and $\lim_{k\to\infty} \sup \tau_1(k) = \lim_{k\to\infty} \sup \tau_2(k) = +\infty$, and that both δ_1 and δ_2 are real constants in the interval $(0, 1)$ satisfying*

$$\varepsilon_k \le c_1\delta_1^{\tau_1(k)} \quad \text{and} \quad \eta_k \le c_2\delta_2^{\tau_2(k)}, \qquad k = 0, 1, 2, \dots, \tag{3.19}$$

where c_1 and c_2 are nonnegative constants. Then we have

$$|||\bar{x}^{(k+1)} - x^*|||_{M_2} \le \left(\sqrt{\bar{\sigma}} + \bar{\omega}\bar{\theta}\delta^{\tau(k)}\right)^2 |||\bar{x}^{(k)} - x^*|||_{M_2}, \quad k = 0, 1, 2, \dots,$$

where

$$\tau(k) = \min\{\tau_1(k), \tau_2(k)\}, \quad \delta = \max\{\delta_1, \delta_2\}, \tag{3.20}$$

and

$$\bar{\omega} = \max\left\{\sqrt{c_1 c_2 \bar{\nu}}, \quad \frac{1}{2\sqrt{\bar{\sigma}}}(c_1\bar{\mu} + c_2\bar{\rho})\right\}.$$

In particular, we have

$$\lim_{k\to\infty} \sup \frac{|||\bar{x}^{(k+1)} - x^*|||_{M_2}}{|||\bar{x}^{(k)} - x^*|||_{M_2}} = \bar{\sigma};$$

i.e., the convergence rate of the inexact two-step splitting iterative scheme is asymptotically the same as that of the exact two-step splitting iterative scheme.

Proof. From (3.6) and (3.19), we obtain, for $k = 0, 1, 2, \dots$, that

$$\begin{aligned}
|||\bar{x}^{(k+1)} - x^*|||_{M_2} &\le \left(\bar{\sigma} + \bar{\mu}\bar{\theta}\varepsilon_k + \bar{\theta}(\bar{\rho} + \bar{\theta}\bar{\nu}\varepsilon_k)\eta_k\right)|||\bar{x}^{(k)} - x^*|||_{M_2} \\
&\le \left(\bar{\sigma} + \bar{\mu}\bar{\theta}c_1\delta_1^{\tau_1(k)} + \bar{\theta}(\bar{\rho} + \bar{\theta}\bar{\nu}c_1\delta_1^{\tau_1(k)})c_2\delta_2^{\tau_2(k)}\right)|||\bar{x}^{(k)} - x^*|||_{M_2} \\
&\le \left(\bar{\sigma} + \bar{\mu}\bar{\theta}c_1\delta^{\tau(k)} + \bar{\theta}(\bar{\rho} + \bar{\theta}\bar{\nu}c_1\delta^{\tau(k)})c_2\delta^{\tau(k)}\right)|||\bar{x}^{(k)} - x^*|||_{M_2} \\
&= \left(\bar{\sigma} + (c_1\bar{\mu} + c_2\bar{\rho})\bar{\theta}\delta^{\tau(k)} + c_1c_2\bar{\nu}\bar{\theta}^2\delta^{2\tau(k)}\right)|||\bar{x}^{(k)} - x^*|||_{M_2} \\
&\le \left(\bar{\sigma} + 2\bar{\omega}\sqrt{\bar{\sigma}}\bar{\theta}\delta^{\tau(k)} + \bar{\omega}^2\bar{\theta}^2\delta^{2\tau(k)}\right)|||\bar{x}^{(k)} - x^*|||_{M_2} \\
&= \left(\sqrt{\bar{\sigma}} + \bar{\omega}\bar{\theta}\delta^{\tau(k)}\right)^2|||\bar{x}^{(k)} - x^*|||_{M_2}.
\end{aligned}$$

The result follows straightforwardly. ∎

Theorems 3.2 and 3.3 immediately result in the following convergence result of the IHSS iteration method.

THEOREM 3.4. *Let the assumptions in Theorem* 3.2 *be satisfied. Suppose that both* $\{\tau_1(k)\}$ *and* $\{\tau_2(k)\}$ *are nondecreasing and positive sequences satisfying* $\tau_1(k) \ge 1$, $\tau_2(k) \ge 1$, *and* $\lim_{k\to\infty} \sup \tau_1(k) = \lim_{k\to\infty} \sup \tau_2(k) = +\infty$, *and that both* δ_1 *and* δ_2 *are real constants in the interval* $(0, 1)$ *satisfying* (3.19). *Then it holds that*

$$|||\bar{x}^{(k+1)} - x^*||| \le \left(\sqrt{\sigma(\alpha)} + \omega\theta\delta^{\tau(k)}\right)^2|||\bar{x}^{(k)} - x^*|||, \quad k = 0, 1, 2, \dots,$$

where ρ *and* θ *are defined by* (3.18), $\tau(k)$ *and* δ *are defined by* (3.20), *and*

$$\omega = \max\left\{\sqrt{c_1 c_2 \rho}, \quad \frac{1}{2\sqrt{\sigma(\alpha)}}(c_1\sigma(\alpha) + c_2\rho)\right\}.$$

In particular, we have

$$\lim_{k\to\infty} \sup \frac{|||\bar{x}^{(k+1)} - x^*|||}{|||\bar{x}^{(k)} - x^*|||} = \sigma(\alpha);$$

i.e., the convergence rate of the IHSS iteration method is asymptotically the same as that of the HSS iteration method.

According to Theorem 3.4, we show that, if the tolerances $\{\varepsilon_k\}$ and $\{\eta_k\}$ are chosen as in (3.19), then the IHSS iteration converges to the unique solution $x^* \in \mathbb{C}^n$ of the system of linear equations (1.1), and the upper bound of the asymptotic convergence factor of the IHSS iteration tends to $\sigma(\alpha)$ of that of the HSS iteration (cf. Theorem 2.2). Moreover, we remark that we may replace (3.19) by other rules for which $\{\varepsilon_k\}$ and $\{\eta_k\}$ approach zero. See [14].

TABLE 1
Work to compute a sweep of the IHSS method.

Operation	Work
$\bar{r}^{(k)} = b - A\bar{x}^{(k)}$	$n + a$
$(\alpha I + H)\bar{z}^{(k+\frac{1}{2})} = \bar{r}^{(k)}$	$\chi_k(H)$
$\bar{x}^{(k+\frac{1}{2})} = \bar{x}^{(k)} + \bar{z}^{(k+\frac{1}{2})}$	n
$\bar{r}^{(k+\frac{1}{2})} = b - A\bar{x}^{(k+\frac{1}{2})}$	$n + a$
$(\alpha I + S)\bar{z}^{(k+1)} = \bar{r}^{(k+\frac{1}{2})}$	$\chi_k(S)$
$\bar{x}^{(k+1)} = \bar{x}^{(k+\frac{1}{2})} + \bar{z}^{(k+1)}$	n

Computational complexity. To analyze the computational complexity of the HSS and the IHSS iterations, we need to estimate their computer times (via operation counts) and computer memories. Assume that a is the number of operations required to compute Ay for a given vector $y \in \mathbb{C}^n$ and $\chi_k(H)$ and $\chi_k(S)$ are the numbers of operations required to solve inner systems (3.1) and (3.2) inexactly with the tolerances $\{\varepsilon_k\}$ and $\{\eta_k\}$, respectively. Then the work to compute a sweep of the IHSS iteration is estimated using the results of Table 1. Straightforward calculations show that the total work to compute each step of the IHSS iteration is $\mathcal{O}(4n + 2a + \chi_k(H) + \chi_k(S))$.

In addition, a simple calculation shows that the memory is required to store $\bar{x}^{(k)}$, b, $\bar{r}^{(k)}$, $\bar{z}^{(k)}$. For the inexact solvers for inner systems (3.1) and (3.2), we require only some auxiliary vectors; for instance, CG-type methods need about five vectors [24]. Moreover, it is not necessary to store H and S explicitly as matrices, as all we need are two subroutines that perform the matrix-vector multiplications with respect to these two matrices. Therefore, the total amount of computer memory required is $\mathcal{O}(n)$, which has the same order of magnitude as the number of unknowns.

4. Application to the model convection-diffusion equation. We consider the three-dimensional convection-diffusion equation

$$-(u_{xx} + u_{yy} + u_{zz}) + q(u_x + u_y + u_z) = f(x, y, z) \tag{4.1}$$

on the unit cube $\Omega = [0, 1] \times [0, 1] \times [0, 1]$, with constant coefficient q and subject to Dirichlet-type boundary conditions. When the seven-point finite difference discretization, for example, the centered differences to the diffusive terms, and the centered differences or the first order upwind approximations to the convective terms are applied to the above model convection-diffusion equation, we get the system of linear equations (1.1) with the coefficient matrix

$$A = T_x \otimes I \otimes I + I \otimes T_y \otimes I + I \otimes I \otimes T_z, \tag{4.2}$$

where the equidistant step-size $h = \frac{1}{n+1}$ is used in the discretization on all of the three directions and the natural lexicographic ordering is employed to the unknowns. In addition, $\otimes$ denotes the Kronecker product, and T_x, T_y, and T_z are tridiagonal matrices given by

$$T_x = \text{ tridiag}(t_2, t_1, t_3), \quad T_y = \text{ tridiag}(t_2, 0, t_3), \quad \text{and} \quad T_z = \text{ tridiag}(t_2, 0, t_3),$$

with

$$t_1 = 6, \quad t_2 = -1 - r, \quad t_3 = -1 + r$$

if the first order derivatives are approximated by the centered difference scheme and

$$t_1 = 6 + 6r, \quad t_2 = -1 - 2r, \quad t_3 = -1$$

if the first order derivatives are approximated by the upwind difference scheme. Here

$$r = \frac{qh}{2}$$

is the mesh Reynolds number. For details, we refer to [9, 10] and [12, 13].

From (4.2), we know that the Hermitian part H and the skew-Hermitian part S of the matrix A are

$$H = H_x \otimes I \otimes I + I \otimes H_y \otimes I + I \otimes I \otimes H_z \tag{4.3}$$

and

$$S = S_x \otimes I \otimes I + I \otimes S_y \otimes I + I \otimes I \otimes S_z, \tag{4.4}$$

where

$$H_x = \text{ tridiag}\left(\frac{t_2+t_3}{2}, t_1, \frac{t_2+t_3}{2}\right), \quad H_y = H_z = \text{ tridiag}\left(\frac{t_2+t_3}{2}, 0, \frac{t_2+t_3}{2}\right),$$

$$S_\xi = \text{ tridiag}\left(\frac{t_2-t_3}{2}, 0, -\frac{t_2-t_3}{2}\right), \quad \xi \in \{x, y, z\}.$$

From Lemma A.1, we know, for the centered difference scheme, that

$$\begin{aligned} &\min_{1\le j,k,l\le n} \lambda_{j,k,l}(H) = 6(1-\cos(\pi h)), && \max_{1\le j,k,l\le n} \lambda_{j,k,l}(H) = 6(1+\cos(\pi h)), \\ &\min_{1\le j,k,l\le n} |\lambda_{j,k,l}(S)| = 0, && \max_{1\le j,k,l\le n} |\lambda_{j,k,l}(S)| = 6r\cos(\pi h). \end{aligned}$$

Therefore, the quantities in Theorem 2.2 can be obtained by concrete computations.

THEOREM 4.1. *For the system of linear equations* (1.1) *with the coefficient matrix* (4.2) *arising from the centered difference scheme for the three-dimensional model convection-diffusion equation* (4.1) *with the homogeneous Dirichlet boundary condition, the iteration sequence* $\{x^{(k)}\}$ *generated by the HSS iteration from an initial guess* $x^{(0)} \in \mathbb{C}^n$ *converges to its unique solution* $x^* \in \mathbb{C}^n$ *and satisfies*

$$|||x^{(k+1)} - x^*||| \le \left[1 - \pi h + \frac{1}{2}\pi^2 h^2 + \mathcal{O}(h^3)\right] \cdot |||x^{(k)} - x^*|||, \qquad k = 0, 1, 2, \ldots .$$

We note that this bound is independent of q and the mesh Reynolds number. The results for the upwind difference scheme can be obtained in an analogous fashion. Since H and S in (4.3) and (4.4) can be diagonalized by sine transforms, the number of operations required at each HSS iteration is about $O(n^3 \log n)$. It follows that the total complexity of the HSS iteration is about $O(n^4 \log n)$ operations. Here n is the number of grid points in all three directions. Here the model problem is used as an example to illustrate the convergence rate of the HSS iteration. We remark that there may be other efficient methods for solving the model convection-diffusion equation (see [12, 13, 6]).

For a three-dimensional convection-diffusion system of linear equations arising from performing one step of cyclic reduction on an equidistant mesh, discretized by the centered and the upwind difference schemes, Greif and Varah [9, 10] considered two ordering strategies, analyzed block splittings of the coefficient matrices, and showed that the associated block Jacobi iterations converge for both the one-dimensional and the two-dimensional splittings with their spectral radii bounded by

$$1-\left(\frac{10}{9}\pi^2+\frac{1}{6}q^2\right)h^2+O(h^3) \quad \text{and} \quad 1-\left(2\pi^2+\frac{9}{10}q^2\right)h^2+O(h^3),$$

respectively. It is clear that these two bounds are larger than those of the HSS and the IHSS methods. Moreover, for the three-dimensional convection-diffusion model equation, the number of operations required for each step of the block Jacobi iteration is about $O(n^3)$ operations, and hence its total complexity is about $O(n^5)$ operations. We remark that their methods can provide an ordering for block Jacobi which can be used for preconditioning.

5. Numerical examples. In this section, we perform some numerical examples to demonstrate the effectiveness of both HSS and IHSS iterations.

5.1. Spectral radius. In this subsection, we first show in Figures 1 and 2 the spectral radius $\rho(M(\alpha))$ of the iteration matrix $M(\alpha)$ and its upper bound $\sigma(\alpha)$ for different α. Here the coefficient matrices A arise from the discretization of the differential equation

$$-u''+qu'=0$$

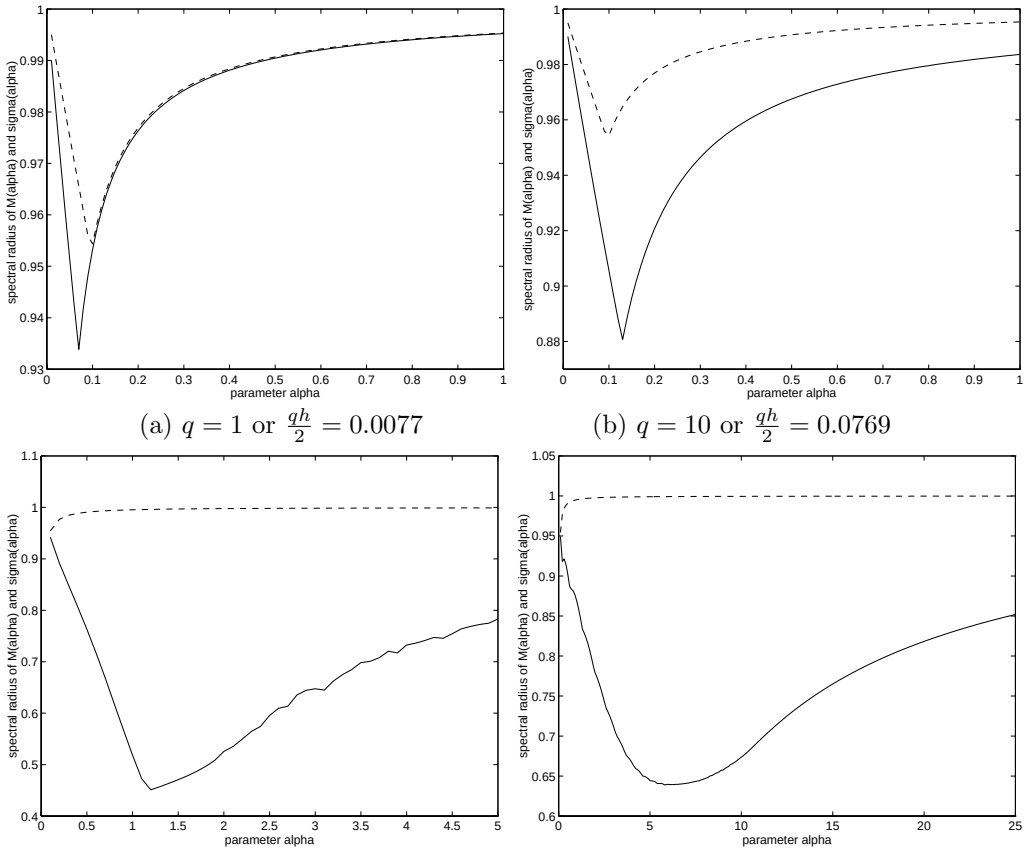
with the homogeneous boundary condition using the centered and the upwind difference schemes. In the tests, the size of the matrix A is 64-by-64. We see from the figures that both $\rho(M(\alpha))$ and $\sigma(\alpha)$ are always less than 1 for $\alpha>0$. These results show that the HSS iteration always converges. Moreover, when q (or $qh/2$) is small, $\sigma(\alpha)$ is close to $\rho(M(\alpha))$, i.e., $\sigma(\alpha)$ is a good approximation to $\rho(M(\alpha))$. However, when q (or $qh/2$) is large (the skew-Hermitian part is dominant), $\sigma(\alpha)$ deviates from $\rho(M(\alpha))$ very much. From Figures 1 and 2, we see that the optimal parameter α_t,

$$\alpha_t \equiv \arg\min_{\alpha}\{\rho(M(\alpha))\},$$

is roughly equal to $qh/2$. To further investigate $\sigma(\alpha)$, we examine the parameter α in the HSS iteration in Figure 3. In the figure, we depict the spectral radii of the iteration matrices for different q (or $qh/2$) by using α^* in Corollary 2.3, $\tilde{\alpha}=qh/2$, and the optimal parameter α_t. It is clear that, when q (or $qh/2$) is small (i.e., the skew-Hermitian part is not dominant), α^* is close to α_t, and $M(\alpha^*)$ is a good estimate of $M(\alpha_t)$. However, when q (or $qh/2$) is large, α^* is not very useful; see Table 2. In contrast to α^*, we observe that $\tilde{\alpha}$ is close to α_t when q (or $qh/2$) is large. In the appendix, we give a remark to further explain why the spectral radius of $M(\tilde{\alpha})$ is less than $\sigma(\alpha^*)$ by using a 2-by-2 matrix example.

In Figure 4, we depict the eigenvalue distributions of the iteration matrices using α_t when $q=1,10,100,1000$. We see that the spectral radius of the iteration matrix for large q is less than that of the iteration matrix for small q.

5.2. Results for the HSS iteration. In this subsection, we test the HSS iteration by numerical experiments. All tests are started from the zero vector, performed



(a) $q = 1$ or $\frac{qh}{2} = 0.0077$ (b) $q = 10$ or $\frac{qh}{2} = 0.0769$

(c) $q = 100$ or $\frac{qh}{2} = 0.7692$ (d) $q = 1000$ or $\frac{qh}{2} = 7.6923$

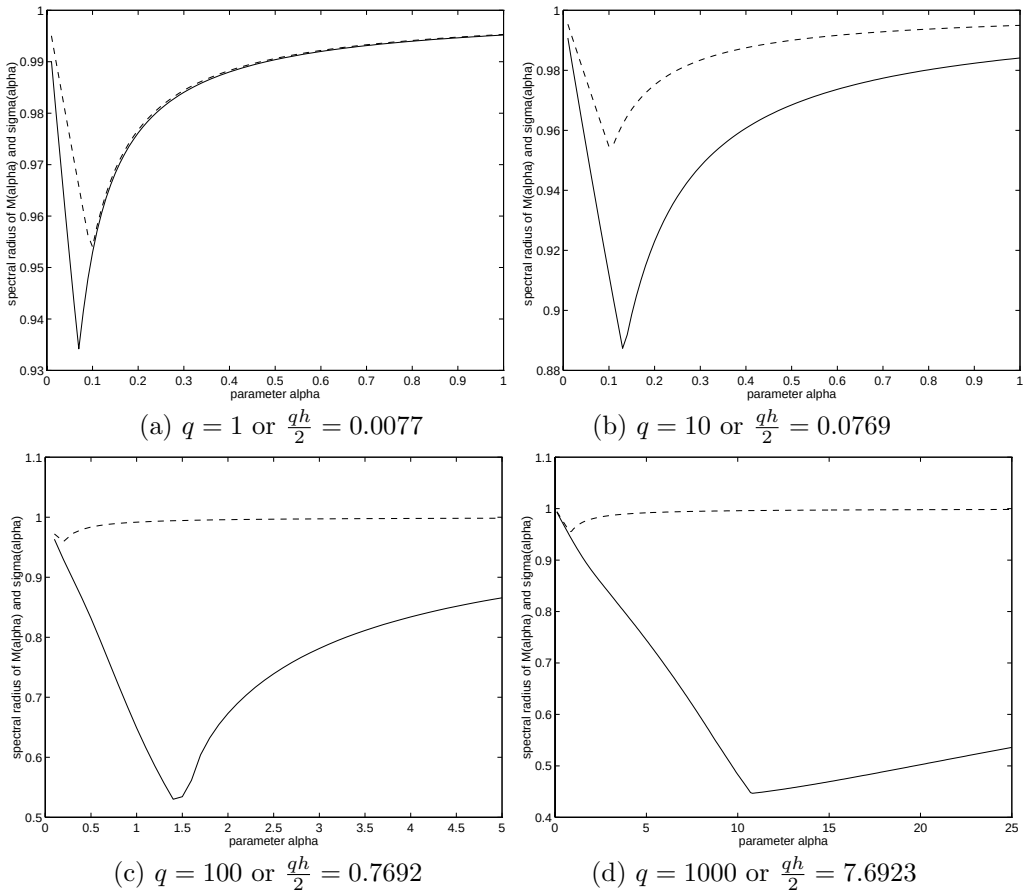
FIG. 1. *The spectral radius $\rho(M(\alpha))$ of the iteration matrices for different α: "——" and the upper bound $\sigma(\alpha)$ for different α: "- - - -" (centered difference scheme).*

TABLE 2
The spectral radii of the iteration matrices $M(\alpha^*)$, $M(\tilde{\alpha})$, and $M(\alpha_t)$ when $n = 64$.

Difference scheme	q	α^*	$\rho(M(\alpha^*))$	$\tilde{\alpha}$	$\rho(M(\tilde{\alpha}))$	α_t	$\rho(M(\alpha_t))$
centered	1	0.0966	0.9516	0.0077	0.9923	0.0700	0.9339
centered	10	0.0966	0.9086	0.0769	0.9264	0.1300	0.8807
centered	100	0.0966	0.9438	0.7692	0.6339	1.160	0.4487
centered	1000	0.0966	0.9511	7.6923	0.6445	5.800	0.6389
upwind	1	0.0974	0.9517	0.0077	0.9924	0.0700	0.9342
upwind	10	0.1041	0.9085	0.0769	0.9314	0.1300	0.8874
upwind	100	0.1710	0.9388	0.7692	0.7321	1.450	0.5237
upwind	1000	0.8399	0.9447	7.6923	0.6092	10.75	0.4466

in MATLAB with machine precision 10^{-16}, and terminated when the current iterate satisfies $\|r^{(k)}\|_2/\|r^{(0)}\|_2 < 10^{-6}$, where $r^{(k)}$ is the residual of the kth HSS iteration.

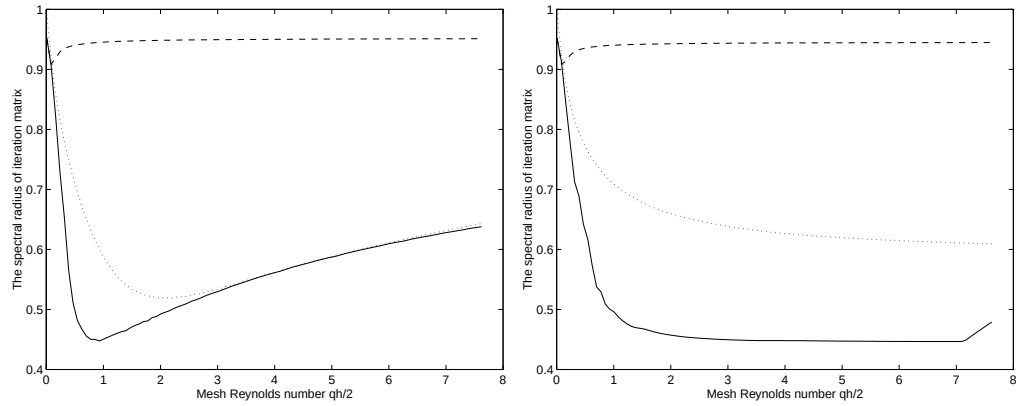
We solve the three-dimensional convection-diffusion equation (4.1) with the homogeneous Dirichlet boundary condition by the HSS iteration. The number n of grid points in all three directions is the same, and the n^3-by-n^3 linear systems with respect to the coefficient matrices $\alpha I + H$ and $\alpha I + S$ are solved efficiently by the sine and



(a) $q = 1$ or $\frac{qh}{2} = 0.0077$ (b) $q = 10$ or $\frac{qh}{2} = 0.0769$

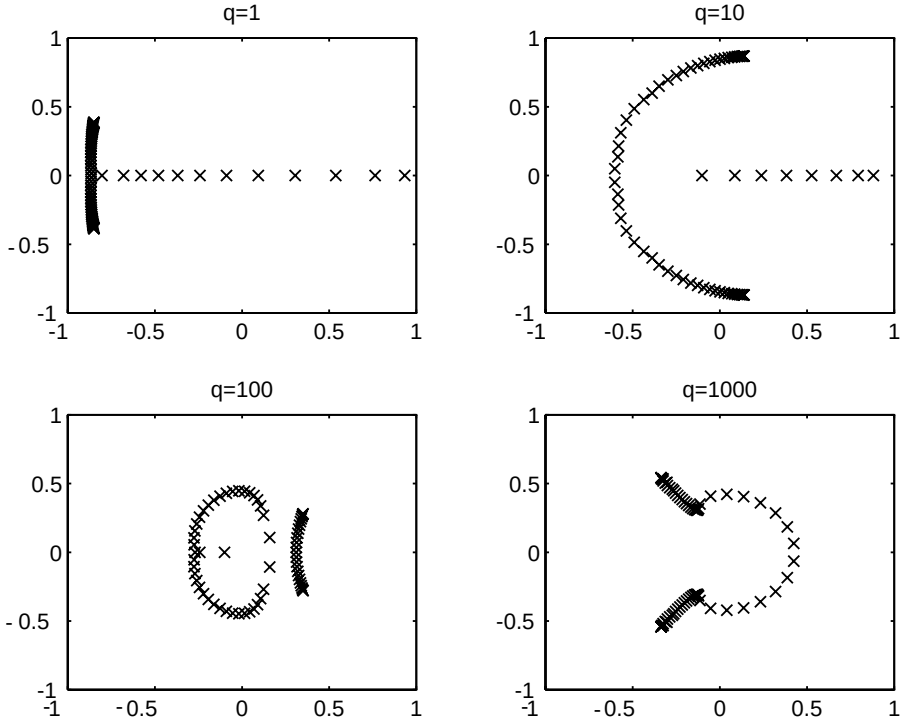
(c) $q = 100$ or $\frac{qh}{2} = 0.7692$ (d) $q = 1000$ or $\frac{qh}{2} = 7.6923$

FIG. 2. *The spectral radius $\rho(M(\alpha))$ of the iteration matrices for different α: "——" and the upper bound $\sigma(\alpha)$ for different α: "- - - -" (upwind difference scheme).*

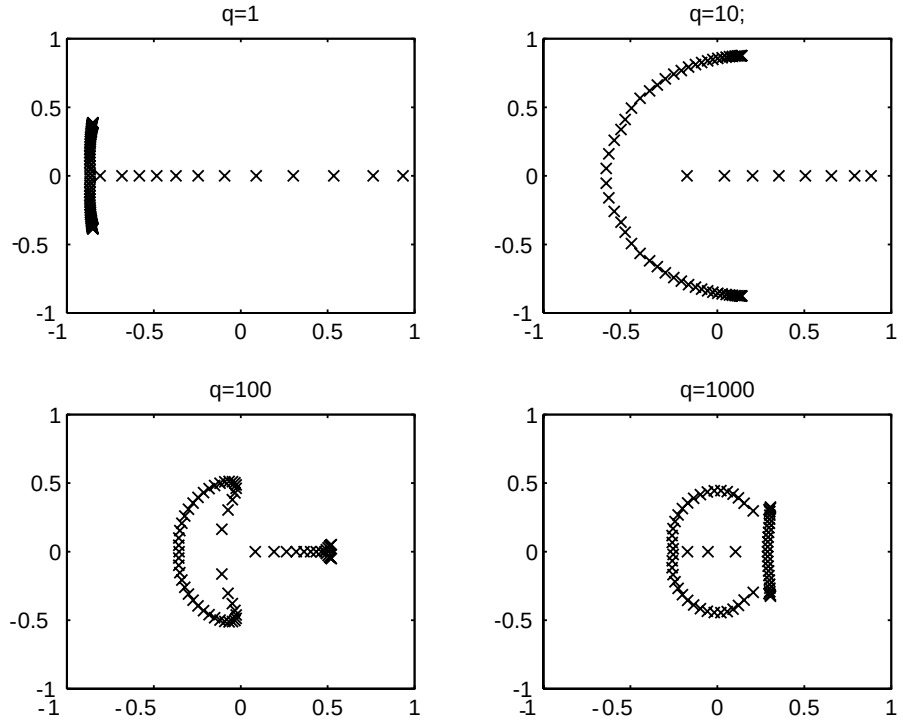


centered difference scheme upwind difference scheme

FIG. 3. *The spectral radius of the iteration matrices for different q: using α_t "——," α^* in Corollary* 2.3 *"- - - -," and $\tilde{\alpha} = qh/2$ "....."*



(a) centered difference scheme



(b) upwind difference scheme

FIG. 4. *The eigenvalue distributions of the iteration matrices when* $\alpha = \alpha_t$.

TABLE 3
Number of HSS iterations for the centered (left) and the upwind (right) difference schemes using α^* in Corollary 2.3.

	q			
n	1	10	100	1000
8	34	23	34	35
16	61	42	59	62
32	116	83	117	123
64	234	169	231	244

	q			
n	1	10	100	1000
8	33	22	27	28
16	59	42	52	53
32	114	82	102	109
64	226	158	205	228

TABLE 4
Number of HSS iterations for the centered (left) and the upwind (right) difference schemes using $\tilde{\alpha} = qh/2$.

	q			
n	1	10	100	1000
8	208	28	25	193
16	433	52	22	106
32	844	102	25	76
64	>1000	195	33	66

	q			
n	1	10	100	1000
8	220	40	22	20
16	446	63	26	22
32	852	115	33	25
64	>1000	208	48	33

TABLE 5
Number of HSS iterations for the centered (left) and the upwind (right) difference schemes using the optimal α_t.

	q			
n	1	10	100	1000
8	33	16	20	37
16	58	31	21	48
32	113	57	25	46
64	221	105	33	51

	q			
n	1	10	100	1000
8	33	22	15	15
16	59	35	18	18
32	114	63	26	23
64	204	109	40	33

the modified sine transforms, respectively (cf. Lemma A.1). In Table 3, we list the numerical results for the centered difference and the upwind difference schemes when $q = 1, 10, 100, 1000$. Evidently, when q is large, the cell Reynolds number is also large for each fixed n. Since the eigenvalues of H are known, the parameter α^* can be computed according to Corollary 2.3. We observe that the number of iterations is not only increasing linearly with n but also roughly independent of q as predicted from the convergence analysis in Corollary 2.3. We also test $\tilde{\alpha}$ and the optimal α given in Table 2. In Tables 4 and 5, we present their numbers of HSS iterations. We see from the tables that the number of iterations using the optimal α is less than that using α^* especially when q is large. Moreover, when q is large, the numbers of iterations using the optimal α and $\tilde{\alpha}$ are about the same.

5.3. Results for IHSS iterations. The second test is for the three-dimensional convection-diffusion equation

$$-(u_{xx} + u_{yy} + u_{zz}) + q \exp(x + y + z)(xu_x + yu_y + zu_z) = f(x, y, z)$$

on the unit cube $\Omega = [0, 1] \times [0, 1] \times [0, 1]$, with the homogeneous Dirichlet boundary conditions. For this problem, the n^3-by-n^3 linear systems with respect to the coefficient matrices $\alpha I + H$ and $\alpha I + S$ cannot be solved efficiently by the sine and the modified sine transforms. Therefore, we solve the linear systems with coefficient matrices $\alpha I + H$ iteratively by the preconditioned CG (PCG) method with the sine transform based preconditioner presented in [22], and we solve the linear systems

TABLE 6
Number of IHSS iterations for the centered difference scheme using α^* in Table 3.

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	38	37	36	25	21	21	28	28	28	35	39	35
16	72	65	60	45	45	38	55	55	54	59	59	59
32	171	160	142	91	86	84	105	104	103	114	114	114
64	462	339	298	249	210	172	205	202	202	237	233	233

TABLE 7
Number of IHSS iterations for the centered difference scheme using the optimal α_t in Table 5.

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	42	41	41	24	24	24	17	17	17	35	35	35
16	78	71	68	42	38	38	34	34	34	43	41	41
32	167	146	136	81	75	73	60	60	60	44	44	44
64	453	355	292	161	150	137	116	116	116	54	54	54

with the coefficient matrix $\alpha I + S$ iteratively by the preconditioned CG for normal equations (PCGNE) method with the modified sine transform based preconditioner given in [19]. This results in the IHSS iteration discussed in section 3. We choose CGNE as the inner solver because it is quite stable, convergent monotonically, and transpose-free. Therefore, as an inner iteration, it could produce an approximate solution satisfying a prescribed rough accuracy in a few iteration steps.

In our computations, the inner PCG and PCGNE iterates are terminated if the current residuals of the inner iterations satisfy

$$\frac{\|p^{(j)}\|_2}{\|r^{(k)}\|_2} \le \max\{0.1\delta^k, 1\times 10^{-7}\} \quad \text{and} \quad \frac{\|q^{(j)}\|_2}{\|r^{(k)}\|_2} \le \max\{0.1\delta^k, 1\times 10^{-6}\} \tag{5.1}$$

(cf. (3.19) and (3.20) in Theorem 3.3), where $p^{(j)}$ and $q^{(j)}$ are, respectively, the residuals of the jth inner PCG and iterates at the $(k+1)$st outer IHSS iterate, $r^{(k)}$ is the residual of the kth outer IHSS iterate, and δ is a control tolerance. In Tables 6–9, we list numerical results for the centered difference and the upwind difference schemes when $q = 1, 10, 100, 1000$. Since the eigenvalues of H cannot be explicitly obtained, the parameter α^* is not exactly known, and we employ the corresponding parameters used in HSS iterations in Tables 3 and 5 instead.

According to Tables 6–9, the number of IHSS iterations generally increases when δ increases. We see that these increases in IHSS iterations for small q are more significant than those for large q. We also observe that the number of IHSS iterations again increases linearly with n and roughly independent of q. In the tables, the number of iterations using the optimal α is again less than that using α^*, especially when q is large. Moreover, when the optimal α is used, the number of IHSS iterations is about the same for $q = 1, 10, 100, 1000$.

In Table 10, we list the average number of inner PCGNE iterations corresponding to the centered difference scheme. In this case, the Hermitian linear systems with the coefficient matrix $\alpha I + H$ can be solved efficiently by the sine transform. Therefore, we report only the average number of inner PCGNE iterations. In Tables 11 and 12, we report the average number of inner PCG and inner PCGNE iterations corresponding to the upwind difference scheme. It is obvious that, when the control parameter

TABLE 8
Number of IHSS iterations for the upwind difference scheme using α^* in Table 3.

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	33	33	32	24	23	23	27	26	26	28	28	28
16	78	68	68	46	42	42	63	61	60	70	70	69
32	171	155	129	103	87	82	131	127	127	166	164	164
64	460	348	306	263	180	164	248	248	246	370	367	366

TABLE 9
Number of IHSS iterations for the upwind difference scheme using the optimal α_t in Table 5.

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	42	35	35	24	24	22	30	30	30	29	29	29
16	80	70	70	44	44	44	48	48	48	59	59	59
32	165	144	131	83	82	80	85	85	85	95	95	95
64	316	258	239	179	143	141	137	137	137	143	143	143

TABLE 10
Average number of PCGNE iterations for the centered difference scheme using (a) *α^* in Table 3 and* (b) *the optimal α_t in Table 5.*

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.4	1.8	2.3	3.4	4.0	4.5	6.8	7.1	7.4	6.9	7.0	7.4
16	2.0	2.8	3.5	5.9	7.4	8.6	13.9	14.6	14.9	15.1	15.2	15.2
32	3.5	5.6	6.9	10.2	14.1	17.6	29.0	30.0	30.2	31.7	31.7	31.7
64	7.3	8.5	9.1	22.4	31.2	34.1	60.1	61.9	62.5	55.0	56.8	57.5

(a)

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.5	2.0	2.6	2.8	3.4	4.1	4.9	4.9	5.2	6.7	7.2	7.6
16	2.1	3.0	3.8	4.4	5.8	7.1	7.0	8.4	9.6	12.0	13.7	14.5
32	3.3	5.1	6.6	6.7	9.7	12.6	10.7	14.7	17.7	21.1	23.7	27.0
64	7.1	8.5	9.1	10.8	17.1	21.1	16.7	25.8	33.7	29.1	38.4	45.4

(b)

δ becomes small, the average number of inner PCG and inner PCGNE iterations becomes large. We observe from the tables that the average number of inner PCGNE iterations increases with q, but the average number of inner PCG iterations required is almost nonchanging. The reason is that the parameter q in the convection part does not affect the convergence rate of the Hermitian linear system but does affect the convergence rate of the shifted skew-Hermitian linear system. Moreover, the average number of inner PCGNE iterations using the optimal α_t is less than that of those using α^*, especially when q is large.

Moreover, we find that when δ decreases, the number of inner (PCG or PCGNE) iterations required increases in the numerical tests. In Figure 5, we show an example of this general phenomenon. This is mainly because the inner PCG and the inner PCGNE iterates are terminated if the current residuals of the inner iterations satisfy (5.1). When δ is small, more iterations are required to satisfy the stopping criterion.

Furthermore, instead of PCGNE, we solve the linear systems with the coefficient

TABLE 11
Average number of PCG iterations for the upwind difference scheme using (a) α^* *in Table* 3 *and* (b) *the optimal* α_t *in Table* 5.

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.5	2.3	3.1	1.3	1.9	2.5	1.4	2.0	2.7	1.4	2.1	2.8
16	2.8	4.3	5.3	2.1	3.1	4.3	2.5	4.0	5.1	2.6	4.4	5.3
32	5.4	6.6	7.0	3.8	5.6	6.4	4.6	6.3	7.0	5.3	6.7	7.2
64	7.9	8.3	8.5	7.1	7.7	8.1	7.0	8.1	8.4	7.6	8.4	8.6

(a)

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.3	1.6	2.1	1.0	1.5	1.6	1.0	1.6	2.0	1.0	1.6	1.9
16	1.6	2.3	2.6	1.3	1.9	2.3	1.4	2.0	2.4	1.5	2.2	2.5
32	3.0	3.5	3.6	2.2	3.1	3.4	2.2	3.1	3.4	2.3	3.2	3.5
64	4.3	4.6	4.7	3.8	4.3	4.6	3.4	4.3	4.5	3.5	4.3	4.6

(b)

TABLE 12
Average number of PCGNE iterations for the upwind difference scheme using (a) α^* *in Table* 3 *and* (b) *the optimal* α_t *in Table* 5.

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.4	1.7	2.1	2.7	3.2	3.8	4.6	4.9	5.5	4.7	5.2	6.0
16	2.0	2.8	3.8	4.9	6.3	7.7	10.4	12.5	13.9	11.6	13.8	14.6
32	3.4	5.4	6.5	9.8	13.0	16.0	24.6	28.6	29.8	28.4	29.9	30.5
64	7.3	8.6	9.2	21.9	27.6	32.3	55.8	59.9	61.2	62.5	63.2	63.5

(a)

	$q=1$			$q=10$			$q=100$			$q=1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.5	1.9	2.3	2.9	3.4	4.0	3.1	3.9	4.4	3.6	4.2	4.9
16	2.1	2.9	3.9	3.9	5.4	6.7	5.5	7.4	8.8	6.1	7.9	9.9
32	3.3	5.0	6.5	5.9	9.2	11.9	9.1	13.2	16.8	10.8	15.3	19.8
64	5.1	7.1	8.0	11.3	16.1	20.9	15.3	24.0	31.0	19.7	30.5	38.3

(b)

matrix $\alpha I + S$ iteratively by the preconditioned GMRES method (PGMRES [25, 24]) with the modified sine transform based preconditioner given in [19]. Using the same stopping criterion as for the PCGNE, we report the average number of inner PGMRES iterations in Table 13. We see from Tables 10 and 13 that, when q is small, the average number of inner PCGNE iterations is slightly less than that of inner PGMRES iterations. However, when q is large, the average number of inner PGMRES iterations is less than that of inner PCGNE iterations.

6. Conclusion and remarks. For the non-Hermitian positive definite system of linear equations, we present a class of (inexact) splitting iteration methods based on the HSS of the coefficient matrix and the Krylov subspace iterations such as CG and CGNE, and we demonstrate that these methods converge unconditionally to the unique solution of the linear system. In fact, this work presents a general framework of iteration methods for solving this class of system of linear equations. There are several combinations in the framework of iterations. We can solve the Hermitian

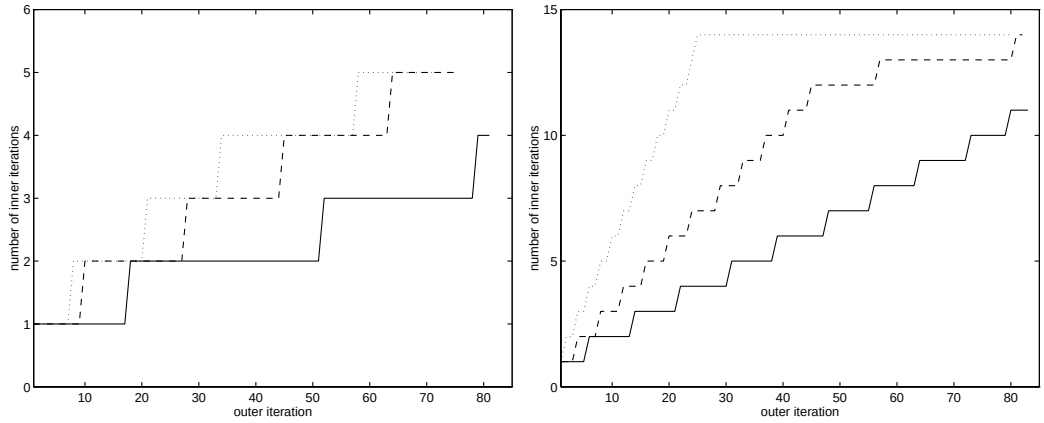


FIG. 5. *The number of inner iterations required for each outer iteration when* $n = 32$ *and* $q = 10$ *in the upwind difference scheme using the optimal* α_t*: (left) PCG inner iterations and (right) PCGNE inner iterations.* —– ($\delta = 0.9$), - - - - ($\delta = 0.8$), ... ($\delta = 0.7$).

TABLE 13

Average number of PGMRES iterations for the centered difference scheme using (a) α^* *in Table* 3 *and* (b) *the optimal* α_t *in Table* 5.

	$q = 1$			$q = 10$			$q = 100$			$q = 1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.6	1.9	2.3	4.0	4.3	4.9	6.9	7.1	7.4	6.9	7.1	7.5
16	2.2	2.9	3.6	6.4	7.5	8.6	14.5	14.6	14.9	15.2	15.3	15.5
32	3.7	5.7	7.0	10.9	14.5	17.6	27.8	28.1	30.2	28.5	29.7	30.7
64	7.5	9.1	9.2	23.6	31.3	34.1	48.1	51.3	54.5	48.5	51.8	52.5

(a)

	$q = 1$			$q = 10$			$q = 100$			$q = 1000$		
	δ			δ			δ			δ		
n	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7	0.9	0.8	0.7
8	1.8	2.1	2.5	2.8	3.5	4.1	4.9	5.0	5.2	6.8	7.4	7.7
16	2.3	3.0	3.9	4.4	5.8	7.2	7.0	8.5	9.5	12.1	13.9	14.5
32	3.3	5.2	6.7	6.8	9.7	12.7	9.1	12.5	15.1	16.5	18.3	21.0
64	7.3	8.5	9.2	10.8	17.2	21.2	13.3	19.4	24.9	19.9	23.8	33.5

(b)

part exactly or inexactly and the skew-Hermitian part exactly or inexactly. The best choice depends on the structures of the Hermitian and the skew-Hermitian matrices. Convergence theories for the correspondingly resulted exact HSS or IHSS iterations can be established following an analogous analysis to this paper with slight technical modifications.

Moreover, instead of CG and CGNE, we can employ other efficient iterative methods of types of Krylov subspace [24, 7], multigrid, multilevel, classical relaxation, etc. to solve the systems of linear equations with coefficient matrices $\alpha I + H$ and $\alpha I + S$ involved at each step of the HSS iteration. In particular, we mention that, when GMRES is applied to the linear system with coefficient matrix $\alpha I + S$, it automatically reduces to a two-term recurrence process, and its convergence property is dependent only on the eigenvalues, but independent of the eigenvectors, of the matrix $\alpha I + S$.

Appendix. The basic lemma used in the model problem analysis in section 4 is shown in this section.

LEMMA A.1 (see [5, 19]). *The matrix H in (4.3) can be diagonalized by the matrix $F^{(1)} \otimes F^{(1)} \otimes F^{(1)}$. Here $F^{(1)} = ([F^{(1)}]_{j,k})$ is the sine transform matrix defined by*

$$[F^{(1)}]_{j,k} = \sqrt{\frac{2}{n+1}} \sin\left(\frac{jk\pi}{n+1}\right), \quad j,k = 1,2,\ldots,n.$$

The corresponding eigenvalues of H are given by

$$\lambda_{j,k,l}(H) = t_1 + (t_2+t_3)\cdot\left[\cos\left(\frac{j\pi}{n+1}\right) + \cos\left(\frac{k\pi}{n+1}\right) + \cos\left(\frac{l\pi}{n+1}\right)\right], \quad j,k,l = 1,2,\ldots,n.$$

The matrix S in (4.4) can be diagonalized by the matrix $F^{(2)} \otimes F^{(2)} \otimes F^{(2)}$. Here $F^{(2)} = ([F^{(2)}]_{j,k})$ is the modified sine transform matrix defined by

$$[F^{(2)}]_{j,k} = \sqrt{\frac{2}{n+1}} i^{j+k+1} \sin\left(\frac{jk\pi}{n+1}\right), \quad j,k = 1,2,\ldots,n.$$

The corresponding eigenvalues of S are given by

$$\lambda_{j,k,l}(S) = i(t_2-t_3)\cdot\left[\cos\left(\frac{j\pi}{n+1}\right) + \cos\left(\frac{k\pi}{n+1}\right) + \cos\left(\frac{l\pi}{n+1}\right)\right], \quad j,k,l = 1,2,\ldots,n.$$

Here i is used to represent the imaginary unit.

Remark. We consider the 2-by-2 matrix

$$A = \begin{pmatrix} 2+2\cos(\pi h) & -qh/2 \\ qh/2 & 2-2\cos(\pi h) \end{pmatrix}$$

as an example to illustrate the use of the iteration parameter $\alpha = \tilde{\alpha} = qh/2$. It is clear that

$$H = \begin{pmatrix} 2+2\cos(\pi h) & 0 \\ 0 & 2-2\cos(\pi h) \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 0 & -qh/2 \\ qh/2 & 0 \end{pmatrix}.$$

We note that $2+2\cos(\pi h)$ and $2-2\cos(\pi h)$ are the largest and the smallest eigenvalues, respectively, of the Hermitian part of the discretization matrix of the differential equation $-u'' + qu' = 0$. In this case, the iteration matrix $M(\alpha)$ of the HSS iteration is similar to the matrix

$$\begin{aligned}\widetilde{M}(\alpha) = &\begin{pmatrix} \alpha-2-2\cos(\pi h) & 0 \\ 0 & \alpha-2+2\cos(\pi h) \end{pmatrix} \\ &\times \begin{pmatrix} \alpha+2+2\cos(\pi h) & 0 \\ 0 & \alpha+2-2\cos(\pi h) \end{pmatrix}^{-1} \times \begin{pmatrix} \alpha & qh/2 \\ -qh/2 & \alpha \end{pmatrix} \\ &\times \begin{pmatrix} \alpha & -qh/2 \\ qh/2 & \alpha \end{pmatrix}^{-1}.\end{aligned}$$

When $\alpha = \tilde{\alpha} = qh/2$, we have

$$\begin{pmatrix} \alpha & qh/2 \\ -qh/2 & \alpha \end{pmatrix} \times \begin{pmatrix} \alpha & -qh/2 \\ qh/2 & \alpha \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then we compute the eigenvalues λ of $\widetilde{M}(\tilde{\alpha})$, and they are given by

$$\pm\sqrt{\frac{(2+2\cos(\pi h)-\frac{qh}{2})(-2+2\cos(\pi h)+\frac{qh}{2})}{(2+2\cos(\pi h)+\frac{qh}{2})(2-2\cos(\pi h)+\frac{qh}{2})}}.$$

By using the series expansion of the above expression in terms of h, we obtain

$$\lambda = \pm\sqrt{\frac{-\pi+\frac{q}{2}}{\pi+\frac{q}{2}}}\cdot\left(1-\frac{qh}{4}+\mathcal{O}(h^2)\right).$$

However, if we use α^* as the iteration parameter, the upper bound $\sigma(\alpha^*)$ of the spectral radius $\rho(M(\alpha^*))$ of the iteration matrix $M(\alpha^*)$ is given by

$$\frac{\sqrt{2+2\cos(\pi h)}-\sqrt{2-2\cos(\pi h)}}{\sqrt{2+2\cos(\pi h)}+\sqrt{2-2\cos(\pi h)}} = 1-\pi h+\mathcal{O}(h^2);$$

see Corollary 2.3. Hence, when $q > 4\pi$, $\rho(M(\tilde{\alpha}))$ is less than $\sigma(\alpha^*)$. From this example, we see that $\tilde{\alpha}$ is a good iteration parameter when q is large. Figure 3 indeed shows that $\tilde{\alpha}$ is close to α_t.

REFERENCES

[1] O. Axelsson, Z.-Z. Bai, and S.-X. Qiu, *A class of nested iteration schemes for linear systems with a coefficient matrix with a dominant positive definite symmetric part*, Numer. Algorithms, to appear.

[2] Z.-Z. Bai, *Sharp error bounds of some Krylov subspace methods for non-Hermitian linear systems*, Appl. Math. Comput., 109 (2000), pp. 273–285.

[3] Z.-Z. Bai, I. Duff, and A. J. Wathen, *A class of incomplete orthogonal factorization methods* I: *Methods and theories*, BIT, 41 (2001), pp. 53–70.

[4] M. Benzi and D. Szyld, *Existence and uniqueness of splittings of stationary iterative methods with applications to alternating methods*, Numer. Math., 76 (1997), pp. 309–321.

[5] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.

[6] W. Cheung and M. K. Ng, *Block-circulant preconditioners for systems arising from discretization of the three-dimensional convection-diffusion equation*, J. Comput. Appl. Math., 140 (2002), pp. 143–158.

[7] P. Concus and G. H. Golub, *A generalized conjugate gradient method for non-symmetric systems of linear equations*, in Computing Methods in Applied Sciences and Engineering, Lecture Notes in Econom. and Math. Systems 134, R. Glowinski and J.R. Lions, eds., Springer-Verlag, Berlin, 1976, pp. 56–65; also available online from http://www-sccm.stanford.edu.

[8] J. Douglas, Jr. and H. H. Rachford, Jr., *Alternating direction methods for three space variables*, Numer. Math., 4 (1956), pp. 41–63.

[9] C. Greif and J. Varah, *Iterative solution of cyclically reduced systems arising from discretization of the three-dimensional convection-diffusion equation*, SIAM J. Sci. Comput., 19 (1998), pp. 1918–1940.

[10] C. Greif and J. Varah, *Block stationary methods for nonsymmetric cyclically reduced systems arising from three-dimensional elliptic equations*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1038–1059.

[11] M. Eiermann, W. Niethammer, and R. S. Varga, *Acceleration of relaxation methods for non-Hermitian linear systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 979–991.

[12] H. Elman and G. H. Golub, *Iterative methods for cyclically reduced non-self-adjoint linear systems*, Math. Comput., 54 (1990), pp. 671–700.

[13] H. Elman and G. H. Golub, *Iterative methods for cyclically reduced non-self-adjoint linear systems* II, Math. Comput., 56 (1991), pp. 215–242.

[14] E. Giladi, G. H. Golub, and J. B. Keller, *Inner and outer iterations for the Chebyshev algorithm*, SIAM J. Numer. Anal., 35 (1998), pp. 300–319.

[15] G. H. Golub and D. Vanderstraeten, *On the preconditioning of matrices with a dominant skew-symmetric component*, Numer. Algorithms, 25 (2000), pp. 223–239.
[16] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
[17] G. H. Golub and A. J. Wathen, *An iteration for indefinite systems and its application to the Navier–Stokes equations*, SIAM J. Sci. Comput., 19 (1998), pp. 530–539.
[18] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.
[19] L. Hemmingsson and K. Otto, *Analysis of semi-Toeplitz preconditioners for first-order PDEs*, SIAM J. Sci. Comput., 17 (1996), pp. 47–64.
[20] T. Manteuffel, *An incomplete factorization technique for positive definite linear systems*, Math. Comput., 34 (1980), pp. 473–497.
[21] J. Meijerink and H. Van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comput., 31 (1977), pp. 148–162.
[22] M. K. Ng, *Preconditioning of elliptic problems by approximation in the transform domain*, BIT, 37 (1997), pp. 885–900.
[23] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41.
[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.
[25] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[26] C.-L. Wang and Z.-Z. Bai, *Sufficient conditions for the convergent splittings of non-Hermitian positive definite matrices*, Linear Algebra Appl., 330 (2001), pp. 215–218.
[27] O. Widlund, *A Lanczos method for a class of nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 801–812.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 627–644

CONVERGENCE ANALYSIS OF INEXACT RAYLEIGH QUOTIENT ITERATION*

YVAN NOTAY†

Abstract. We consider the computation of the smallest eigenvalue and associated eigenvector of a Hermitian positive definite pencil. Rayleigh quotient iteration (RQI) is known to converge cubically, and we first analyze how this convergence is affected when the arising linear systems are solved only approximately. We introduce a special measure of the relative error made in the solution of these systems and derive a sharp bound on the convergence factor of the eigenpair in a function of this quantity. This analysis holds independently of the way the linear systems are solved and applies to any type of error. For instance, it applies to rounding errors as well.

We next consider the Jacobi–Davidson method. It acts as an inexact RQI method in which the use of iterative solvers is made easier because the arising linear systems involve a projected matrix that is better conditioned than the shifted matrix arising in classical RQI. We show that our general convergence result straightforwardly applies in this context and permits us to trace the convergence of the eigenpair in a function of the number of inner iterations performed at each step. On this basis, we also compare this method with some form of inexact inverse iteration, as recently analyzed by Neymeyr and Knyazev.

Key words. eigenvalue, Rayleigh quotient, Jacobi–Davidson, preconditioning

AMS subject classifications. 65F10, 65B99, 65N20

PII. S0895479801399596

1. Introduction.

We consider the computation of the smallest eigenvalue and associated eigenvector of a Hermitian positive definite pencil $A - \lambda B$.

In this context, the Rayleigh quotient iteration (RQI) method is known to converge very quickly, and cubically in the asymptotic phase [1, 15]. However, it requires solving at each step a system with the shifted matrix $A - \theta B$, with shift θ equal to the Rayleigh quotient, i.e., changing from step to step. For large sparse matrices, this makes the use of direct solvers impractical, and, therefore, several works focus on the use of iterative solvers either by a direct approach [2, 19, 24] or indirectly via the use of the Jacobi–Davidson (JD) method [3, 14, 20, 21, 22, 23]. However, how an inexact solution may affect the convergence seems up to now not very well understood, despite the various analyses developed in these papers. The answer is actually far from obvious because, on the one hand, the systems to solve are very ill conditioned, and hence reducing the error measured with respect to any standard norm may involve a lot of numerical effort. On the other hand, it has been known for a long time from the error analysis made in connection with direct solvers that large errors in the computed solution do not necessarily spoil the convergence [16, 26].

In this paper, we first bring some new light on the actual convergence of inexact RQI. We introduce a special measure of the relative error made in the solution of the linear systems and bound the convergence factor of the eigenpair in a function of this quantity. Moreover, we show that the bound is sharp, indicating that the analysis takes the errors into proper account. This is further demonstrated by showing that

*Received by the editors December 12, 2001; accepted for publication (in revised form) by H. van der Vorst July 16, 2002; published electronically January 17, 2003. This research was supported by the Fonds National de la Recherche Scientifique, Maître de recherches.

http://www.siam.org/journals/simax/24-3/39959.html

†Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165/84), 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium (ynotay@ulb.ac.be).

our bound allows a straightforward analysis of the rounding errors arising with a backward stable direct solver when θ is numerically equal to λ_1.

We next consider the JD method. Although it may be motivated in a different way (see [21]), it acts as an inexact RQI method and may even be seen as one of the easiest ways to implement robustly iterative solvers within RQI, since the ill-conditioned systems are not attacked directly (see the above references or section 4 for details). Here we show that our special measure of the error is equal to some standard relative error for the linear systems arising in the JD method. Hence our general convergence result straightforwardly applies, allowing us to trace the convergence of the eigenpair in a function of the number of *inner* iterations. This also allows some comparison with the predicted convergence of schemes based on inexact inverse iteration, as analyzed by Neymeyr [11, 12] and Knyazev and Neymeyr [9] (see [5, 10] for alternative analyses of inexact inverse iteration that, however, do not allow us to directly bound the convergence rate).

The remainder of the paper is organized as follows. In section 2, we recall some needed results on the convergence of RQI with exact solution of the arising linear systems. Our convergence analysis of inexact RQI is developed in section 3, and the JD method is discussed in section 4.

Notation. Throughout this paper, A and B are Hermitian $n \times n$ matrices. We further assume that B is positive definite and that the smallest eigenvalue of the pencil $A - \lambda B$ is simple. The eigenpairs are denoted $(\lambda_i, \mathbf{u}_i)$, $i = 1, \ldots, n$, with the eigenvalues ordered increasingly (i.e., $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$) and the eigenvectors orthonormal with respect to the $(\cdot, B\,\cdot)$ inner product (i.e., $(\mathbf{u}_i, B\,\mathbf{u}_j) = \delta_{ij}$).

For any symmetric and positive definite matrix C, we denote $\|\cdot\|_C$ as the C-norm, that is, the norm associated to the $(\cdot, C\,\cdot)$ inner product: $\|\mathbf{v}\|_C = \sqrt{(\mathbf{v}, C\,\mathbf{v})}$ for all $\mathbf{v}$.

2. Convergence of standard RQI. Let first recall the basic algorithm: if $\mathbf{u}$ is some approximate eigenvector and

$$\theta \;=\; \frac{(\mathbf{u},\, A\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})} \tag{2.1}$$

is the associated Rayleigh quotient, the RQI method computes the next approximate eigenpair $(\widehat{\mathbf{u}},\, \widehat{\theta}\,)$ as

$$\widehat{\mathbf{u}} = (A - \theta\, B)^{-1} B\,\mathbf{u}, \tag{2.2}$$

$$\widehat{\theta} = \frac{(\widehat{\mathbf{u}},\, A\,\widehat{\mathbf{u}})}{(\widehat{\mathbf{u}},\, B\,\widehat{\mathbf{u}})}. \tag{2.3}$$

(In practice, some form of normalization is performed on $\widehat{\mathbf{u}}$, but this does not matter for the discussion here.) Note that the RQI method favors the convergence toward the eigenvalue closest to θ. Here we analyze the convergence toward the smallest eigenvalue, and thus we assume that

$$\theta \;<\; \frac{\lambda_1 + \lambda_2}{2}, \tag{2.4}$$

which implies (since θ cannot be smaller than λ_1) that

$$\frac{\theta - \lambda_1}{\lambda_2 - \theta} \;\in\; [0, 1). \tag{2.5}$$

To assess the convergence, we introduce the decompositions

$$\mathbf{u} = \|\mathbf{u}\|_B \ (\cos\varphi \ \mathbf{u}_1 + \sin\varphi \ \mathbf{v}),$$
$$\widehat{\mathbf{u}} = \|\widehat{\mathbf{u}}\|_B \ (\cos\widehat{\varphi} \ \mathbf{u}_1 + \sin\widehat{\varphi} \ \widehat{\mathbf{v}}),$$

where $(\mathbf{v},\, B\,\mathbf{u}_1) = (\widehat{\mathbf{v}},\, B\,\mathbf{u}_1) = 0$ and $\|\mathbf{v}\|_B = \|\widehat{\mathbf{v}}\|_B = 1$. Then (see [15, p. 73])

$$\begin{aligned} \tan\widehat{\varphi} &= (\theta - \lambda_1) \ \|(A - \theta\, B)^{-1} B\,\mathbf{v}\|_B \ \tan\varphi \\ &\leq \frac{\theta - \lambda_1}{\lambda_2 - \theta} \ \tan\varphi, \end{aligned} \tag{2.6}$$

and the cubic convergence follows from $(\theta - \lambda_1) = \mathcal{O}(\sin^2\varphi)$.

Now, to prove our main theorem, we need a sharp bound on $\widehat{\theta}$. This is obtained with Knyazev's analysis as developed in [6, 7]. Indeed, particularizing [6, Theorem 2.3.1] to our context (see also [7, Theorem 2.5]), one gets

$$\frac{\widehat{\theta} - \lambda_1}{\lambda_2 - \widehat{\theta}} \ \leq \ \left(\frac{\theta - \lambda_1}{\lambda_2 - \theta}\right)^3, \tag{2.7}$$

which is simpler than (2.6) to work with.

Knyazev's proof is general and elegant. (It covers a family of methods and not only the RQI method; see [13, Theorem 4.4] for an English translation.) However, for our analysis, we need to know in which cases the above bound is sharp. This can be seen by deriving (2.7) directly from (2.6). To this purpose, let

$$\eta \ = \ (\mathbf{v},\, A\,\mathbf{v}), \quad \widehat{\eta} \ = \ (\widehat{\mathbf{v}},\, A\,\widehat{\mathbf{v}})$$

be the Rayleigh quotients associated to $\mathbf{v}$, $\widehat{\mathbf{v}}$, respectively (remember that $\|\mathbf{v}\|_B = \|\widehat{\mathbf{v}}\|_B = 1$). Note that $\widehat{\eta} \leq \eta$ because $\widehat{\mathbf{v}}$ is the vector resulting from one step of the shift and invert iteration applied to $\mathbf{v}$ with shift θ smaller than the smallest eigenvalue for which $\mathbf{v}$ has a nonzero component in the direction of the corresponding eigenvector. Since

$$\theta = \cos^2\varphi \ \ \lambda_1 \ + \ \sin^2\varphi \ \ \eta,$$

one has

$$\begin{aligned} \theta - \lambda_1 &= \sin^2\varphi \ \ (\eta - \lambda_1), \\ \eta - \theta &= \cos^2\varphi \ \ (\eta - \lambda_1), \end{aligned} \tag{2.8}$$

whence

$$\tan^2\varphi \ = \ \frac{\theta - \lambda_1}{\eta - \theta}, \tag{2.9}$$

and, similarly,

$$\tan^2\widehat{\varphi} \ = \ \frac{\widehat{\theta} - \lambda_1}{\widehat{\eta} - \widehat{\theta}} \ \geq \ \frac{\widehat{\theta} - \lambda_1}{\eta - \widehat{\theta}}.$$

Inequality (2.6) therefore implies that (squaring both sides)

$$\widehat{\theta} - \lambda_1 \ \leq \ \frac{(\theta - \lambda_1)^3}{(\lambda_2 - \theta)^2} \ \frac{\eta - \widehat{\theta}}{\eta - \theta},$$

whence (2.7) because the last term of the right-hand side is a decreasing function of $\eta \geq \lambda_2$.

From these developments, one sees that the bound (2.7) is sharp when $\mathbf{v} = \mathbf{u}_2$, i.e., when $\mathbf{u} \in \mathrm{span}\{\mathbf{u}_1, \mathbf{u}_2\}$; then, (2.6) becomes indeed an equality, whereas one has $\eta = \widehat{\eta} = \lambda_2$, entailing that equality is attained in (2.7). Note that, since asymptotically $\mathbf{v}$ converges toward $\mathbf{u}_2$, it also means that the bound (2.7) gives the correct value of the asymptotic convergence factor.

Finally, observe that it is relevant to characterize the convergence by the ratio $(\theta - \lambda_1)/(\lambda_2 - \theta)$ even when one is primarily interested in the accuracy of the eigenvector. Indeed, the B^{-1}-norm of the residual

$$\mathbf{r} \;=\; (A - \theta\, B)\, \mathbf{u} \tag{2.10}$$

satisfies

$$\frac{\|\mathbf{r}\|^2_{B^{-1}}}{\|\mathbf{u}\|^2_B} \;\geq\; (\theta - \lambda_1)(\lambda_2 - \theta) \tag{2.11}$$

[17, Lemma 3.2], whence, with (2.9),

$$\tan\varphi \;\leq\; \sqrt{\frac{\theta - \lambda_1}{\lambda_2 - \theta}} \;\leq\; \frac{1}{\lambda_2 - \theta}\, \frac{\|\mathbf{r}\|_{B^{-1}}}{\|\mathbf{u}\|_B}. \tag{2.12}$$

The convergence factor for the eigenvector is, however, only the square root of the one for the ratio $(\theta - \lambda_1)/(\lambda_2 - \theta)$. Note also that this ratio actually has to be made very small to satisfy a stopping criterion based on the residual norm.

3. Convergence of inexact RQI. Assume that some errors are introduced in the computation of $\widehat{\mathbf{u}} = (A - \theta\, B)^{-1} B\, \mathbf{u}$. Let $\widetilde{\mathbf{u}}$ be the resulting vector. To analyze the influence on the convergence factor, we need a proper measure of these errors. The error vector $\mathbf{x} = \widehat{\mathbf{u}} - \widetilde{\mathbf{u}}$ is by itself meaningless because the scaling of $\widetilde{\mathbf{u}}$ is unimportant. Among other possibilities, one may consider

$$\widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, \mathbf{v})}{(\widetilde{\mathbf{u}},\, \mathbf{v})}\, \widetilde{\mathbf{u}}$$

for some vector $\mathbf{v}$ not orthogonal to $\widehat{\mathbf{u}}$, $\widetilde{\mathbf{u}}$. Somewhat arbitrarily, we select $\mathbf{v} = B\, \mathbf{u}$. Note, nevertheless, that $(\widehat{\mathbf{u}},\, B\, \mathbf{u}) = 0$ is not possible because this would imply $\widehat{\theta} - \theta = \|\widehat{\mathbf{u}}\|_B^{-1} (\widehat{\mathbf{u}},\, (A - \theta\, B)\, \widehat{\mathbf{u}}) = 0$, which contradicts (2.7). Accordingly, since $\widetilde{\mathbf{u}}$ approximates $\widehat{\mathbf{u}}$ at least in direction, it is not very restrictive to assume that $(\widetilde{\mathbf{u}},\, B\, \mathbf{u}) \neq 0$.

This choice leads us to characterize the error with

$$\mathbf{y} \;=\; \widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\, \mathbf{u})}{(\widetilde{\mathbf{u}},\, B\, \mathbf{u})}\, \widetilde{\mathbf{u}} \tag{3.1}$$

for which we have still to choose an appropriate norm. Here we state the following lemma, which is a straightforward generalization of [14, Lemma 3.1].

LEMMA 3.1. *Let A, B be $n \times n$ Hermitian matrices. Assume that B is positive definite, and let $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of the pencil $A - \lambda\, B$. For any nonzero vector $\mathbf{u}$, one has*

$$\min_{\substack{\mathbf{z} \,\perp\, B\mathbf{u} \\ \mathbf{z} \neq 0}} \frac{(\mathbf{z},\, (A - \theta\, B)\, \mathbf{z})}{(\mathbf{z},\, B\, \mathbf{z})} \;\geq\; \lambda_1 + \lambda_2 - 2\, \theta,$$

where $\theta = \frac{(\mathbf{u},\, A\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}$.

Moreover, the bound is sharp: if $\mathbf{u} \in \operatorname{span}\{\mathbf{u}_1,\, \mathbf{u}_2\}$, *where* $\mathbf{u}_1$, $\mathbf{u}_2$ *are eigenvectors associated to* λ_1, λ_2, *then* (3.1) *becomes an equality, the lower bound being attained for the vectors* $\mathbf{z}$ *in the one-dimensional subspace* $\operatorname{span}\{\mathbf{u}_1,\, \mathbf{u}_2\} \cap B\mathbf{u}^{\perp}$.

Hence, when the condition (2.4) holds, $A - \theta\, B$ is positive definite on $B\mathbf{u}^{\perp}$, and

$$\|\cdot\|_{A-\theta\, B} \;=\; \sqrt{(\cdot,\, (A - \theta\, B)\,\cdot)} \tag{3.2}$$

defines a particular (energy) norm on that subspace. Since $\mathbf{y}$ belongs to that subspace, we may therefore use that norm, and we find that this makes the theoretical analysis easier.

Now, results are often better expressed in a function of relative errors. In this view, we compare the actual norm of $\mathbf{y}$ with the norm one would obtain with $\widetilde{\mathbf{u}} = \mathbf{u}$, that is, if no progress at all were made in the computation of the eigenpair. We thus propose to measure the errors introduced in the RQI process with the number

$$\gamma \;=\; \frac{\left\| \widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\; \widetilde{\mathbf{u}} \right\|_{A-\theta\, B}}{\left\| \widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}\; \mathbf{u} \right\|_{A-\theta\, B}}. \tag{3.3}$$

This looks somewhat unusual, but, as recalled in the introduction, standard measures of the error are often meaningless as far as the convergence of the eigenvector is concerned. Moreover, we shall see in the next section that this measure allows a straightforward analysis of the JD method.

We now state our main result.

THEOREM 3.2. *Let* A, B *be* $n \times n$ *Hermitian matrices. Assume that* B *is positive definite, and let* $\lambda_1 < \lambda_2 \le \cdots \le \lambda_n$ *be the eigenvalues of the pencil* $A - \lambda\, B$. *Let* $\mathbf{u}$ *be any nonzero vector such that*

$$\theta \;=\; \frac{(\mathbf{u},\, A\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}$$

satisfies

$$\theta \;<\; \frac{\lambda_1 + \lambda_2}{2}.$$

Let

$$\widehat{\mathbf{u}} \;=\; (A - \theta\, B)^{-1} B\,\mathbf{u},$$

and let $\widetilde{\mathbf{u}}$ *be a vector such that* $(\widetilde{\mathbf{u}},\, B\,\mathbf{u}) \neq 0$ *and*

$$\gamma \;=\; \frac{\left\| \widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\; \widetilde{\mathbf{u}} \right\|_{A-\theta\, B}}{\left\| \widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}\; \mathbf{u} \right\|_{A-\theta\, B}} \;\le\; 1.$$

Then

$$\widetilde{\theta} \;=\; \frac{(\widetilde{\mathbf{u}},\, A\,\widetilde{\mathbf{u}})}{(\widetilde{\mathbf{u}},\, B\,\widetilde{\mathbf{u}})}$$

satisfies

$$\frac{\widetilde{\theta}-\lambda_1}{\lambda_2-\widetilde{\theta}} \;\le\; \sigma^2 \;\; \frac{\theta-\lambda_1}{\lambda_2-\theta}, \tag{3.4}$$

where

$$\sigma \;=\; \frac{(\theta-\lambda_1)+\gamma\,(\lambda_2-\theta)}{(\lambda_2-\theta)+\gamma\,(\theta-\lambda_1)}. \tag{3.5}$$

Moreover, the bound is sharp: let $\mathbf{u}_1$, $\mathbf{u}_2$ *be eigenvectors associated to* λ_1, λ_2*; if* $\mathbf{u} \in \mathrm{span}\{\mathbf{u}_1, \mathbf{u}_2\}$*, then, for all* $\mathbf{x} \in \mathrm{span}\{\mathbf{u}_1, \mathbf{u}_2\}$ *such that* $|(\mathbf{x},\, B\,\mathbf{u})| < |(\widehat{\mathbf{u}},\, B\,\mathbf{u})|$*, one has that equality is attained in* (3.4) *for either* $\widetilde{\mathbf{u}} = \widehat{\mathbf{u}} + \mathbf{x}$ *or* $\widetilde{\mathbf{u}} = \widehat{\mathbf{u}} - \mathbf{x}$*.*

Proof. Assume (without loss of generality) that $\|\mathbf{u}\|_B = 1$, let $\widehat{\theta}$ be defined by (2.3) and $\mathbf{y}$ by (3.1), and let

$$\widehat{\delta} \;=\; \theta - \widehat{\theta}, \quad \widetilde{\delta} \;=\; \theta - \widetilde{\theta}.$$

First observe that $(\widehat{\mathbf{u}},\, (A-\theta\,B)\,\widehat{\mathbf{u}}) = (\widehat{\mathbf{u}},\, B\,\mathbf{u})$, that $(\widehat{\mathbf{u}},\, (A-\theta\,B)\,\mathbf{u}) = \|\mathbf{u}\|_B = 1$, and that $(\mathbf{u},\, (A-\theta\,B)\,\mathbf{u}) = 0$. Hence

$$\|\widehat{\mathbf{u}} - (\widehat{\mathbf{u}},\, B\,\mathbf{u})\,\mathbf{u}\|^2_{A-\theta\,B} \;=\; -(\widehat{\mathbf{u}},\, B\,\mathbf{u}). \tag{3.6}$$

Further,

$$\widehat{\delta} \;=\; -\frac{(\widehat{\mathbf{u}},\, (A-\theta\,B)\,\widehat{\mathbf{u}})}{(\widehat{\mathbf{u}},\, B\,\widehat{\mathbf{u}})} \;=\; -\frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{\|\widehat{\mathbf{u}}\|^2_B}, \tag{3.7}$$

and, since $(\mathbf{y},\, (A-\theta\,B)\,\widehat{\mathbf{u}}) = (\mathbf{y},\, B\,\mathbf{u}) = 0$,

$$\begin{aligned}
\widetilde{\delta} &= -\frac{(\widetilde{\mathbf{u}},\, (A-\theta\,B)\,\widetilde{\mathbf{u}})}{(\widetilde{\mathbf{u}},\, B\,\widetilde{\mathbf{u}})} \\
&= -\frac{((\widehat{\mathbf{u}}+\mathbf{y}),\, (A-\theta\,B)\,(\widehat{\mathbf{u}}+\mathbf{y}))}{\|\widehat{\mathbf{u}}+\mathbf{y}\|^2_B} \\
&= -\frac{(\widehat{\mathbf{u}},\, (A-\theta\,B)\,\widehat{\mathbf{u}}) + \|\mathbf{y}\|^2_{A-\theta\,B}}{\|\widehat{\mathbf{u}}+\mathbf{y}\|^2_B} \\
&= -(1-\gamma^2)\;\frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{\|\widehat{\mathbf{u}}+\mathbf{y}\|^2_B}.
\end{aligned} \tag{3.8}$$

On the other hand, consider the projector $P = I - \mathbf{u}\,(B\mathbf{u})^*$. Observe that $(\mathbf{v},\, B\,P\,\mathbf{w}) = (P\,\mathbf{v},\, B\,\mathbf{w})$ for all $\mathbf{v}$, $\mathbf{w}$, i.e., that P is orthogonal with respect to the $(\cdot,\, B\,\cdot)$ inner product. Since $P\,\mathbf{y} = \mathbf{y}$, one has

$$\begin{aligned}
\|\widehat{\mathbf{u}}+\mathbf{y}\|^2_B &= \|(I-P)\,\widehat{\mathbf{u}}\|^2_B \;+\; \|P\,\widehat{\mathbf{u}}+\mathbf{y}\|^2_B \\
&\le \|(I-P)\,\widehat{\mathbf{u}}\|^2_B \;+\; \|P\,\widehat{\mathbf{u}}\|^2_B \;+\; \|\mathbf{y}\|^2_B \;+\; 2\,\|P\,\widehat{\mathbf{u}}\|_B\,\|\mathbf{y}\|_B \\
&= \|\widehat{\mathbf{u}}\|^2_B \;+\; \|\mathbf{y}\|^2_B \;+\; 2\,\|P\,\widehat{\mathbf{u}}\|_B\,\|\mathbf{y}\|_B.
\end{aligned} \tag{3.9}$$

Hence, using Lemma 3.1,

$$\begin{aligned}
\|\widehat{\mathbf{u}}+\mathbf{y}\|^2_B &\le \|\widehat{\mathbf{u}}\|^2_B \;+\; \frac{1}{\lambda_1+\lambda_2-2\,\theta}\,\left(\|\mathbf{y}\|^2_{A-\theta\,B} \;+\; 2\,\|P\,\widehat{\mathbf{u}}\|_{A-\theta\,B}\,\|\mathbf{y}\|_{A-\theta\,B}\right) \\
&= \|\widehat{\mathbf{u}}\|^2_B \;-\; (\gamma^2+2\,\gamma)\,\frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{\lambda_1+\lambda_2-2\,\theta},
\end{aligned} \tag{3.10}$$

and, therefore, with (3.7), (3.8),

$$\widetilde{\delta} \;\geq\; \frac{1-\gamma^2}{\widehat{\delta}^{-1}+\frac{\gamma^2+2\,\gamma}{\lambda_1+\lambda_2-2\,\theta}}.$$

We now use (2.7) to bound $\widehat{\delta}$. With $\beta=(\theta-\lambda_1)/(\lambda_2-\theta)$, the latter inequality may be rewritten as

$$\frac{\theta-\lambda_1-\widehat{\delta}}{\beta^{-1}(\theta-\lambda_1)+\widehat{\delta}} \;\leq\; \beta^3,$$

i.e.,

$$\widehat{\delta} \;\geq\; (\theta-\lambda_1)\,\frac{1-\beta^2}{1+\beta^3} \;=\; (\theta-\lambda_1)\,\frac{1-\beta}{1-\beta+\beta^2}.$$

We thus have, since $\lambda_1+\lambda_2-2\,\theta=(\theta-\lambda_1)(\beta^{-1}-1)$,

$$\widetilde{\delta} \;\geq\; (\theta-\lambda_1)\,\frac{(1-\gamma^2)(1-\beta)}{1-\beta+\beta^2+\beta(\gamma^2+2\,\gamma)},$$

whence, letting $D=1-\beta+\beta^2+\beta(\gamma^2+2\,\gamma)$,

$$\begin{aligned}\widetilde{\theta}-\lambda_1 \;&=\; \theta-\lambda_1-\widetilde{\delta}\leq\frac{\theta-\lambda_1}{D}\;\big(D-(1-\gamma^2)(1-\beta)\big)\\ &=\frac{\theta-\lambda_1}{D}\;(\beta+\gamma)^2\end{aligned}$$

and

$$\begin{aligned}\lambda_2-\widetilde{\theta} \;&=\; \lambda_2-\theta+\widetilde{\delta}\geq\frac{\lambda_2-\theta}{D}\;\big(D+\beta(1-\gamma^2)(1-\beta)\big)\\ &=\frac{\lambda_2-\theta}{D}\;(1+\beta\,\gamma)^2.\end{aligned}$$

Therefore, (3.4) holds with

$$\sigma \;=\; \frac{\beta+\gamma}{1+\beta\,\gamma} \;=\; \frac{(\theta-\lambda_1)+\gamma\,(\lambda_2-\theta)}{(\lambda_2-\theta)+\gamma\,(\theta-\lambda_1)}.$$

To prove the sharpness, first observe that the only inequalities used in the proof are (2.7), (3.9), (3.10). Moreover, it has already been noted in section 2 that (2.7) becomes an equality when $\mathbf{u}\in\mathrm{span}\{\mathbf{u}_1,\,\mathbf{u}_2\}$. On the other hand, under the given assumptions, both $P\,\widehat{\mathbf{u}}$ and $\mathbf{y}$ belong to the one-dimensional subspace $\mathrm{span}\{\mathbf{u}_1,\,\mathbf{u}_2\}\cap B\mathbf{u}^{\perp}$. For this subspace, it is shown in Lemma 3.1 that the inequality used to obtain (3.10) from (3.9) is actually also an equality, whereas, since $P\,\widehat{\mathbf{u}}$ and $\mathbf{y}$ are aligned, one has necessarily

$$|(P\,\widehat{\mathbf{u}},\,B\,\mathbf{y})| \;=\; \|P\,\widehat{\mathbf{u}}\|_B\,\|\mathbf{y}\|_B;$$

i.e., (3.9) becomes an equality too if and only if $(P\,\widehat{\mathbf{u}},\,B\,\mathbf{y})$ is positive. Let then $\widetilde{\mathbf{u}}=\widehat{\mathbf{u}}+c\,\mathbf{x}$, where c equals either 1 or -1. One finds

$$\mathbf{y} \;=\; \frac{c}{1+c\,\frac{(\mathbf{x},\,B\,\mathbf{u})}{(\widehat{\mathbf{u}},\,B\,\mathbf{u})}}\left(\frac{(\mathbf{x},\,B\,\mathbf{u})}{(\widehat{\mathbf{u}},\,B\,\mathbf{u})}\;\widehat{\mathbf{u}}\;-\;\mathbf{x}\right),$$

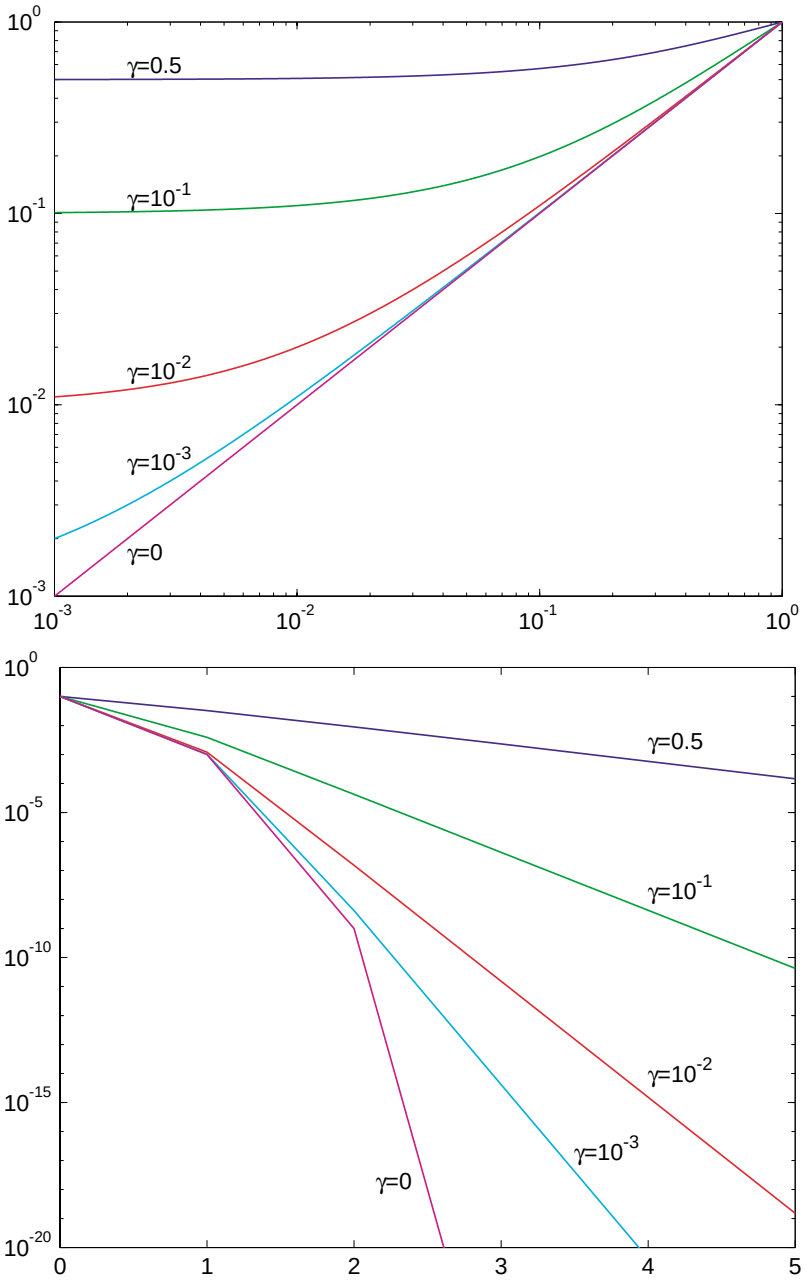


FIG. 1. *σ versus $(\theta-\lambda_1)/(\lambda_2-\theta)$ (top) and evolution of $(\theta-\lambda_1)/(\lambda_2-\theta)$ in a function of the number of steps (bottom).*

showing that, since $|(\mathbf{x},\, B\,\mathbf{u})| < |(\widehat{\mathbf{u}},\, B\,\mathbf{u})|$, one can always, by choosing appropriately the sign of c, select the direction of $\mathbf{y}$ in such a way that $(P\,\widehat{\mathbf{u}},\, B\,\mathbf{y}) > 0$ holds. ∎

Observe that the sharpness of the bound is not proved only for one special orientation of the error vector but that it holds for a two-dimensional subspace that includes both vectors aligned with $\mathbf{u}_1$ and vectors orthogonal to it.

To illustrate our result, we have plotted on Figure 1 (top) the convergence factor σ

against $(\theta - \lambda_1)/(\lambda_2 - \theta)$ for several values of γ. One sees that $\sigma \to \gamma$ when $\theta \to \lambda_1$; more precisely, one has

$$\sigma \;\approx\; \gamma \quad \text{when} \quad \frac{\theta - \lambda_1}{\lambda_2 - \theta} \;\ll\; \gamma. \tag{3.11}$$

On this figure (bottom), we also display the evolution of $(\theta-\lambda_1)/(\lambda_2-\theta)$ in a function of the number of RQI steps. (More precisely, we display it for the worst-case scenario, according to the bound (3.4).) One sees that it is not necessary to make the errors very small to essentially preserve the cubic convergence rate.

Practical estimation of γ. In practical situations, one generally does not have access to the exact value of γ. Nevertheless, some estimate can be obtained, based on the following reasoning. The situation is essentially similar to the one met in the context of the solution of Hermitian positive definite linear systems: theoretical results are expressed in a function of the energy norm of the error, which is not available in practical computations. However, one is generally satisfied with the computation of the residual norm, because it expresses the same error with respect to a different but equivalent norm, and in practice it most often happens that, on the whole, both measures of the error evolve similarly.

Here we want to follow the same approach, but we need to be careful because (3.2) defines a norm only on a particular subspace. Let then

$$P \;=\; I - \mathbf{u}\,(\mathbf{u},\, B\,\mathbf{u})^{-1}\,(B\mathbf{u})^* \tag{3.12}$$

be the projector with range $B\mathbf{u}^\perp$ and kernel $\operatorname{span}\{\mathbf{u}\}$, and note that $P^*(A - \theta\,B)P$ is Hermitian with range $B\mathbf{u}^\perp$. Hence, the pencil $P^*(A - \theta\,B)P - \lambda\,B$ possesses $n-1$ eigenvectors forming a B-orthonormal basis of $B\mathbf{u}^\perp$ and whose corresponding eigenvalues are, by Lemma 3.1, not smaller than $\lambda_1 + \lambda_2 - 2\,\theta$ and not larger than $\lambda_n - \theta$. Therefore, by expanding $\mathbf{v} \in B\mathbf{u}^\perp$ on this basis, one obtains, since $P\,\mathbf{v} = \mathbf{v}$ and thus $\|\mathbf{v}\|_{A-\theta\,B} = \|\mathbf{v}\|_{P^*(A-\theta\,B)P}$,

$$\alpha_1\;\|\mathbf{v}\|_{A-\theta\,B} \;\le\; \|P^*(A - \theta\,B)\,\mathbf{v}\|_{B^{-1}} \;\le\; \alpha_2\;\|\mathbf{v}\|_{A-\theta\,B}, \tag{3.13}$$

where $\alpha_1 = \sqrt{\lambda_1 + \lambda_2 - 2\,\theta}$ and $\alpha_2 = \sqrt{\lambda_n - \theta}$.

On the other hand,

$$\begin{aligned} P^*(A - \theta\,B)\;\left(\widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\;\widetilde{\mathbf{u}}\right) &= P^*\;\left(B\,\mathbf{u} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\,(A - \theta\,B)\,\widetilde{\mathbf{u}}\right) \\ &= -\frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\;P^*(A - \theta\,B)\,\widetilde{\mathbf{u}} \\ &= \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\;P^*\,\mathbf{g}, \end{aligned}$$

where

$$\mathbf{g} \;=\; B\,\mathbf{u} \;-\; (A - \theta\,B)\,\widetilde{\mathbf{u}} \tag{3.14}$$

is the residual of the linear system solved within the RQI process. Similarly, one finds

$$\begin{aligned} P^*(A - \theta\,B)\;\left(\widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}\;\mathbf{u}\right) &= -\frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}\;P^*(A - \theta\,B)\,\mathbf{u} \\ &= -\frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}\;\mathbf{r}, \end{aligned}$$

where $\mathbf{r}$ is the current residual of the eigenproblem (2.10).

Hence, with (3.13),

$$\alpha^{-1}\,\widetilde{\gamma}\;\le\;\gamma\;\le\;\alpha\,\widetilde{\gamma}, \tag{3.15}$$

where $\alpha=\sqrt{\frac{\lambda_n-\theta}{\lambda_1+\lambda_2-2\,\theta}}$ and where

$$\widetilde{\gamma}\;=\;\frac{(\mathbf{u},\,B\,\mathbf{u})}{|(\widetilde{\mathbf{u}},\,B\,\mathbf{u})|}\;\frac{\|P^*(A-\theta\,B)\,\widetilde{\mathbf{u}}\|_{B^{-1}}}{\|(A-\theta\,B)\,\mathbf{u}\|_{B^{-1}}}\;=\;\frac{(\mathbf{u},\,B\,\mathbf{u})}{|(\widetilde{\mathbf{u}},\,B\,\mathbf{u})|}\;\frac{\|P^*\,\mathbf{g}\|_{B^{-1}}}{\|\mathbf{r}\|_{B^{-1}}}. \tag{3.16}$$

Error analysis for $\boldsymbol{\theta\longrightarrow\lambda_1}$. For the sake of simplicity, here we confine ourselves to standard eigenproblems ($B=I$).

In the final phase of the process, one reaches $\theta=\lambda_1$ up to machine accuracy before the eigenvector has converged (see (2.12)). Some subtle reasoning is then needed to show that the ill-conditioning of $A-\theta\,I$ does not prevent further progress despite that the computed solution to $(A-\theta\,I)\,\mathbf{v}=\mathbf{u}$ cannot be accurate even with a backward stable direct solver [16, 26].

Here our results offer a straightforward way to prove that one more step is then enough to compute an accurate eigenvector. Indeed, with a backward stable direct solver, the computed solution $\widetilde{\mathbf{u}}$ is such that

$$\|\mathbf{g}\|\;=\;\|\mathbf{u}-(A-\theta\,I)\,\widetilde{\mathbf{u}}\|\;\le\;c\,\varepsilon_{\rm mach}\,\|A\|\,\|\widetilde{\mathbf{u}}\|.$$

Hence, since P is orthogonal (thus $\|P\,\mathbf{g}\|\le\|\mathbf{g}\|$) and using (2.11),

$$\begin{aligned}\widetilde{\gamma}&\le c\,\varepsilon_{\rm mach}\;\frac{\|\widetilde{\mathbf{u}}\|\,\|\mathbf{u}\|}{|(\widetilde{\mathbf{u}},\,\mathbf{u})|}\;\frac{\|A\|\,\|\mathbf{u}\|}{\|\mathbf{r}\|}\\&\le\;\frac{c\,\varepsilon_{\rm mach}}{\cos(\widetilde{\mathbf{u}},\,\mathbf{u})}\;\frac{\|A\|}{\sqrt{(\theta-\lambda_1)(\lambda_2-\theta)}}.\end{aligned}$$

Thus

$$\sigma\;\le\;\frac{\theta-\lambda_1}{\lambda_2-\theta}\;+\;\alpha\;\frac{c\,\varepsilon_{\rm mach}}{\cos(\widetilde{\mathbf{u}},\,\mathbf{u})}\;\frac{\|A\|}{\sqrt{(\theta-\lambda_1)(\lambda_2-\theta)}},$$

whence

$$\tan(\widetilde{\mathbf{u}},\,\mathbf{u}_1)\;\le\;\sqrt{\frac{\widetilde{\theta}-\lambda_1}{\lambda_2-\widetilde{\theta}}}\;\le\;\left(\frac{\theta-\lambda_1}{\lambda_2-\theta}\right)^{3/2}\;+\;\alpha\;\frac{c\,\varepsilon_{\rm mach}}{\cos(\widetilde{\mathbf{u}},\,\mathbf{u})}\;\frac{\|A\|}{\lambda_2-\theta};$$

that is, for $\theta\to\lambda_1$,

$$\tan(\widetilde{\mathbf{u}},\,\mathbf{u}_1)\;\le\;\alpha\;\frac{c\,\varepsilon_{\rm mach}}{\cos(\widetilde{\mathbf{u}},\,\mathbf{u})}\;\frac{\|A\|}{\lambda_2-\lambda_1}, \tag{3.17}$$

which is not far from the best attainable accuracy; see [26, pp. 69–70] (note that $\tan(\widetilde{\mathbf{u}},\,\mathbf{u})=\mathcal{O}(\tan(\mathbf{u},\,\mathbf{u}_1))=\mathcal{O}(\sqrt{\theta-\lambda_1}\,)$).

4. Convergence of the JD method. The JD method [3, 14, 20, 21, 22, 23] combines some form of inexact RQI with a Galerkin approach.

Let $\mathbf{u}$ be the current approximate eigenvector which we assume is normalized with respect to the B-norm. With this method, one first computes a correction $\mathbf{t}$ orthogonal to $B\mathbf{u}^{\perp}$ by solving (approximately) the so-called *correction equation*

$$P^*(A - \widetilde{\lambda}\, B)P\,\mathbf{t} \;=\; -\mathbf{r};\quad (\mathbf{t},\, B\,\mathbf{u}) = 0, \tag{4.1}$$

where $\widetilde{\lambda}$ is an approximation of the "target" eigenvalue, where $\mathbf{r}$ is the residual (2.10), and where

$$P \;=\; I - \mathbf{u}\,(B\mathbf{u})^*$$

is the projector (3.12). Next, one applies the Galerkin principle: the initial approximation and the successive corrections are gathered to form the basis of a subspace from which one extracts the best approximation of the searched eigenpair by the Rayleigh–Ritz procedure (see, e.g., [22] for algorithmic details).

The exact solution to (4.1) is

$$\widehat{\mathbf{t}} \;=\; \frac{1}{(B\mathbf{u},\, (A - \widetilde{\lambda}\, B)^{-1}B\mathbf{u})}\; (A - \widetilde{\lambda}\, B)^{-1}B\,\mathbf{u} \;-\; \mathbf{u}, \tag{4.2}$$

and hence the equivalence with RQI is recovered if one has used $\widetilde{\lambda} = \theta$ and if the next approximate eigenvector is $\mathbf{u} + \widehat{\mathbf{t}}$.

In practice, one does not always select $\widetilde{\lambda} = \theta$. One first sets $\widetilde{\lambda}$ equal to some fixed target, for instance, $\widetilde{\lambda} = 0$, if one searches for the smallest eigenvalue of a Hermitian positive definite eigenproblem; $\widetilde{\lambda} = \theta$ is then used when, according to some heuristic criterion, one detects that θ entered its final interval (see [14, 25] for examples of such criteria). Here we confine ourselves to this final phase.

On the other hand, the next approximate eigenvector resulting from the Rayleigh–Ritz procedure is generally not equal to $\mathbf{u}+\mathbf{t}$. However, in the context considered here, this procedure selects the vector from the subspace for which the associated Rayleigh quotient is minimal. Hence the convergence as measured through the evolution of the ratio $(\theta - \lambda_1)/(\lambda_2 - \theta)$ can only be improved by this approach. Note, however, that little improvement is expected in the final phase, at least if the correction equation is solved sufficiently accurately: RQI converges then so quickly that one hardly accelerates it. Actually, the Galerkin approach is mainly useful in the first phase, to bring θ into its final interval quickly and to avoid misconvergence if one has selected $\widetilde{\lambda} = \theta$ too early.

We thus continue assuming that one has selected $\widetilde{\lambda} = \theta < (\lambda_1 + \lambda_2)/2$, and we bound the convergence factor by analyzing the Rayleigh quotient associated to

$$\widetilde{\mathbf{u}} \;=\; \mathbf{u} + \widetilde{\mathbf{t}},$$

where $\widetilde{\mathbf{t}}$ is the computed approximate solution to (4.1). Note that $(\widetilde{\mathbf{t}},\, B\,\mathbf{u}) = 0$, whence $(\widetilde{\mathbf{u}},\, B\,\mathbf{u}) = (\mathbf{u},\, B\,\mathbf{u}) = 1$. Thus, since, by (4.2),

$$\widehat{\mathbf{u}} \;=\; (\widehat{\mathbf{u}},\, B\,\mathbf{u})\; (\widehat{\mathbf{t}} + \mathbf{u}),$$

one has

$$\widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\widetilde{\mathbf{u}},\, B\,\mathbf{u})}\; \widetilde{\mathbf{u}} \;=\; (\widehat{\mathbf{u}},\, B\,\mathbf{u})\; (\widehat{\mathbf{t}} - \widetilde{\mathbf{t}}),$$

whereas

$$\widehat{\mathbf{u}} \;-\; \frac{(\widehat{\mathbf{u}},\, B\,\mathbf{u})}{(\mathbf{u},\, B\,\mathbf{u})}\,\mathbf{u} \;=\; (\widehat{\mathbf{u}},\, B\,\mathbf{u})\,\widehat{\mathbf{t}}.$$

Hence, Theorem 3.2 applies with

$$\gamma \;=\; \frac{\|\widehat{\mathbf{t}}-\widetilde{\mathbf{t}}\|_{A-\theta\,B}}{\|\widehat{\mathbf{t}}\|_{A-\theta\,B}}; \tag{4.3}$$

i.e., γ is here the relative error in the correction equation (4.1) measured with respect to the standard energy norm for that equation (remember that, for all $\mathbf{v} \in B\mathbf{u}^{\perp}$, one has $P\,\mathbf{v} = \mathbf{v}$, and therefore $\|\mathbf{v}\|_{A-\theta\,B} = \|\mathbf{v}\|_{P^*(A-\theta\,B)P}$).

Now, since the correction equation is positive definite on $B\mathbf{u}^{\perp}$, one may solve it with the preconditioned conjugate gradient (PCG) method, as advised in [14]. Then γ may be directly bounded in a function of the number k of inner iterations [4, 18]: with the zero initial guess, one has

$$\gamma \;\leq\; 2\,\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k} + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{-k}\right)^{-1}, \tag{4.4}$$

where κ is the spectral condition number.

Considering (3.11) again, one will achieve $\sigma \approx \gamma$ if one is wise enough to stop inner iterations before γ becomes too small, so that further progress is useless. One then recovers our main conclusion from [14], where we analyze the evolution of the residual norm: with a proper stopping criterion, the convergence of the eigenvector goes along with that of the successive linear systems, and the main additional cost to compute the eigenpair compared with a mere linear system solution comes from the need to periodically restart the linear solver.

Now, one may wonder about the value of κ for such a projected system. To simplify the discussion, we assume here (and throughout the paper) that A is positive definite. (The general case is easily recovered with the shift transformation $A-\lambda\,B \rightarrow (A+\tau B)-(\lambda+\tau)\,B$.) We also recall that it is not advised to try to directly precondition the projected matrix (which is dense) nor even the shifted matrix $A-\theta\,B$ (which is indefinite). Instead, set up your favorite (positive definite) preconditioner K for A, and precondition $P^*(A-\theta\,B)P$ with

$$M \;=\; P^*\,K\,P.$$

Note that the singularity of M raises no practical difficulty; see [23]. Moreover, out of the four projection steps associated with the system matrix and the preconditioner, only one needs to be performed in practice; see [3, 14] for details.

Now, with such a preconditioner,

$$\kappa \;=\; \frac{\displaystyle\max_{\substack{\mathbf{z}\,\perp\,B\mathbf{u}\\ \mathbf{z}\neq 0}} \frac{(\mathbf{z},\,(A-\theta\,B)\,\mathbf{z})}{(\mathbf{z},\,K\,\mathbf{z})}}{\displaystyle\min_{\substack{\mathbf{z}\,\perp\,B\mathbf{u}\\ \mathbf{z}\neq 0}} \frac{(\mathbf{z},\,(A-\theta\,B)\,\mathbf{z})}{(\mathbf{z},\,K\,\mathbf{z})}}$$

may be bounded in a function of

$$\kappa(K^{-1}A) \;=\; \frac{\lambda_{\max}\left(K^{-1}A\right)}{\lambda_{\min}\left(K^{-1}A\right)}.$$

Indeed, one has (see also [14])

$$\begin{aligned}\max_{\substack{\mathbf{z}\perp B\mathbf{u}\\ \mathbf{z}\neq 0}} \frac{(\mathbf{z},\,(A-\theta\,B)\,\mathbf{z})}{(\mathbf{z},\,K\,\mathbf{z})} &\leq \max_{\mathbf{z}\neq 0} \frac{(\mathbf{z},\,A\,\mathbf{z})}{(\mathbf{z},\,K\,\mathbf{z})}\\ &= \lambda_{\max}\left(K^{-1}A\right),\end{aligned}$$

whereas, with Lemma 3.1,

$$\begin{aligned}\min_{\substack{\mathbf{z}\perp B\mathbf{u}\\ \mathbf{z}\neq 0}} \frac{(\mathbf{z},\,(A-\theta\,B)\,\mathbf{z})}{(\mathbf{z},\,K\,\mathbf{z})} &= \min_{\substack{\mathbf{z}\perp B\mathbf{u}\\ \mathbf{z}\neq 0}} \left(\frac{(\mathbf{z},\,A\,\mathbf{z})}{(\mathbf{z},\,K\,\mathbf{z})}\,\left(1+\frac{\theta\,(\mathbf{z},\,B\,\mathbf{z})}{(\mathbf{z},\,(A-\theta\,B)\,\mathbf{z})}\right)^{-1}\right)\\ &\geq \lambda_{\min}\left(K^{-1}A\right)\,\left(1+\frac{\theta}{\lambda_1+\lambda_2-2\,\theta}\right)^{-1}.\end{aligned}$$

Therefore,

$$\kappa \;\leq\; \kappa(K^{-1}A)\,\left(1+\frac{\theta}{\lambda_1+\lambda_2-2\,\theta}\right), \tag{4.5}$$

which, together with (4.4), allows us to bound γ and thus σ in a function of k, $\kappa(K^{-1}A)$, θ, λ_1, and λ_2.

Concerning the estimation of γ through $\widetilde{\gamma}$ (3.16), note that

$$\begin{aligned}P^*\,\mathbf{g} &= P^*\left(B\,\mathbf{u}\;-\;\big(A-\theta\,B\big)\big(\mathbf{u}+\widetilde{\mathbf{t}}\big)\right)\\ &= -\mathbf{r}\;-\;P^*\big(A-\theta\,B\big)\,\widetilde{\mathbf{t}}\\ &= \mathbf{g}_{\mathrm{ce}},\end{aligned}$$

where $\mathbf{g}_{\mathrm{ce}}$ is the residual in the correction equation (4.1). Hence, since $(\widetilde{\mathbf{u}},\,B\,\mathbf{u}) = (\mathbf{u},\,B\,\mathbf{u})$,

$$\widetilde{\gamma} \;=\; \frac{\|\mathbf{g}_{\mathrm{ce}}\|_{B^{-1}}}{\|\mathbf{r}\|_{B^{-1}}}, \tag{4.6}$$

which confirms that $\widetilde{\gamma}$ expresses the same error as γ but with respect to the residual norm instead of the energy norm.

Convergence of inexact inverse iteration. It is interesting to compare the above result with the convergence analysis of inexact inverse iteration as developed in [9, 11, 12]. Indeed, if schemes based on the RQI method are expected to converge faster in general, it does not mean that they are always more cost effective. On the other hand, the results in these papers also offer the best bounds to date for schemes based on nonlinear conjugate gradients as developed in, e.g., [8]. Thus the comparison may also give some insight into how the JD method compares with such methods.

Let first recall that inexact (or "preconditioned") inverse iteration also sets

$$\widetilde{\mathbf{u}} \;=\; \mathbf{u}+\widetilde{\mathbf{t}},$$

but here $\widetilde{\mathbf{t}}$ is obtained by solving approximately

$$A\,\mathbf{t} \;=\; -\mathbf{r}. \tag{4.7}$$

Stricto sensu, the analysis in [9, 11, 12] covers only the case in which $\widetilde{\mathbf{t}} = -K^{-1}\mathbf{r}$ for some positive definite preconditioner K. However, looking closely at Lemma 2.1

in [11] (which is the root of everything else), it clearly turns out that the main results also apply when several inner iterations are performed, with parameter γ equal to the relative error in (4.7) measured with respect to the energy norm, i.e.,

$$\gamma \;=\; \frac{\|\widetilde{\mathbf{t}} + A^{-1}\mathbf{r}\|_A}{\|A^{-1}\mathbf{r}\|_A}.$$

The main result is precisely a bound similar to (3.4) with convergence factor

$$\sigma \;\leq\; \overline{\sigma}(\gamma,\,\theta) \;\leq\; \overline{\sigma}(\gamma,\,\lambda_1) = 1-(1-\gamma)\left(1-\frac{\lambda_1}{\lambda_2}\right). \tag{4.8}$$

The first inequality is sharp for any θ and is due to Neymeyr [11, 12]; the resulting expression for $\overline{\sigma}(\gamma,\,\theta)$, however, is so complicated that it is not interesting to reproduce it here. The second inequality is due to Knyazev and Neymeyr [9] and actually gives a good approximation of the first one (see the figures).

Illustration and comparison. For both the JD method and inexact inverse iteration, we are able to bound the convergence factor σ in a function of the number of inner iterations k, given $\kappa(K^{-1}A)$, θ, λ_1, and λ_2. Let $\overline{\sigma}(k,\,\theta)$ be the resulting bound. The most interesting quantity is $\overline{\sigma}^{1/k}(k,\,\theta)$, which represents the convergence in the outer process *per inner iteration*. We have plotted it on Figure 2 against $(\theta-\lambda_1)/(\lambda_2-\theta)$. We consider different values of k for the JD method, but, to keep the figures readable, for inverse iteration we display $\overline{\sigma}^{1/k}(k,\,\theta)$ only for the value of k that minimizes $\overline{\sigma}^{1/k}(k,\,\lambda_1)$.

One sees that, for the JD method, the optimal value of k depends on θ, that is, on how far we are in the convergence process. On the other hand, with a proper choice of k, the JD method clearly outmatches inverse iteration, despite that we use for the latter the exact value of the condition number, whereas, for the JD method, the bound (4.4) is based on the worst-case estimate (4.5).

This is confirmed in Figure 3, where we have plotted the evolution of $(\theta-\lambda_1)/(\lambda_2-\theta)$ (worst-case scenario) against the cumulated number of inner iterations. One sees that the JD method converges faster when k is increased from step to step. In practice, this requires an adaptive stopping criterion such as the one proposed in [14].

Finally, we compared these results with the actual convergence on the following model example: $n = 10000$; $A = \mathrm{diag}(\lambda_i)$ with $\lambda_1 = 2$ and $\lambda_i = 3+i$, $i = 2,\ldots,n$; $K = \mathrm{diag}(\lambda_i(1+\eta_i))$, $i = 1,\ldots,n$, where the η_i are at random in $(0,1)$; the initial approximate eigenvector is given by $(\mathbf{u}_0)_i = \lambda_i^{-2}$.

Thus, we have $\lambda_2/\lambda_1 = 5/2$, $\kappa(K^{-1}A) \approx 2$, and $(\theta_0-\lambda_1)/(\lambda_2-\theta) = 0.953$; i.e., this situation is very similar to the one simulated in Figures 2 and 3 (top part). We therefore performed 2, 4, and 8 inner iterations (with the zero initial guess) in the successive JD steps and plotted on Figure 4 the corresponding evolution of $(\theta-\lambda_1)/(\lambda_2-\theta)$ against the cumulated number of inner iterations. Note that we do not consider the further improvement that could be obtained with the Galerkin approach mentioned at the beginning of this section and that the quantity given within inner steps corresponds to the quantity one would get if inner iterations were stopped at that moment. For illustration purposes, this actual convergence is compared with the theoretical bound and with the convergence of locally optimal block preconditioned conjugate gradient (LOBPCG) [8], which may be seen as an optimized version of preconditioned inverse iteration. The latter method is not of inner-outer type, and thus we here plot $(\theta-\lambda_1)/(\lambda_2-\theta)$ against the number of iterations.

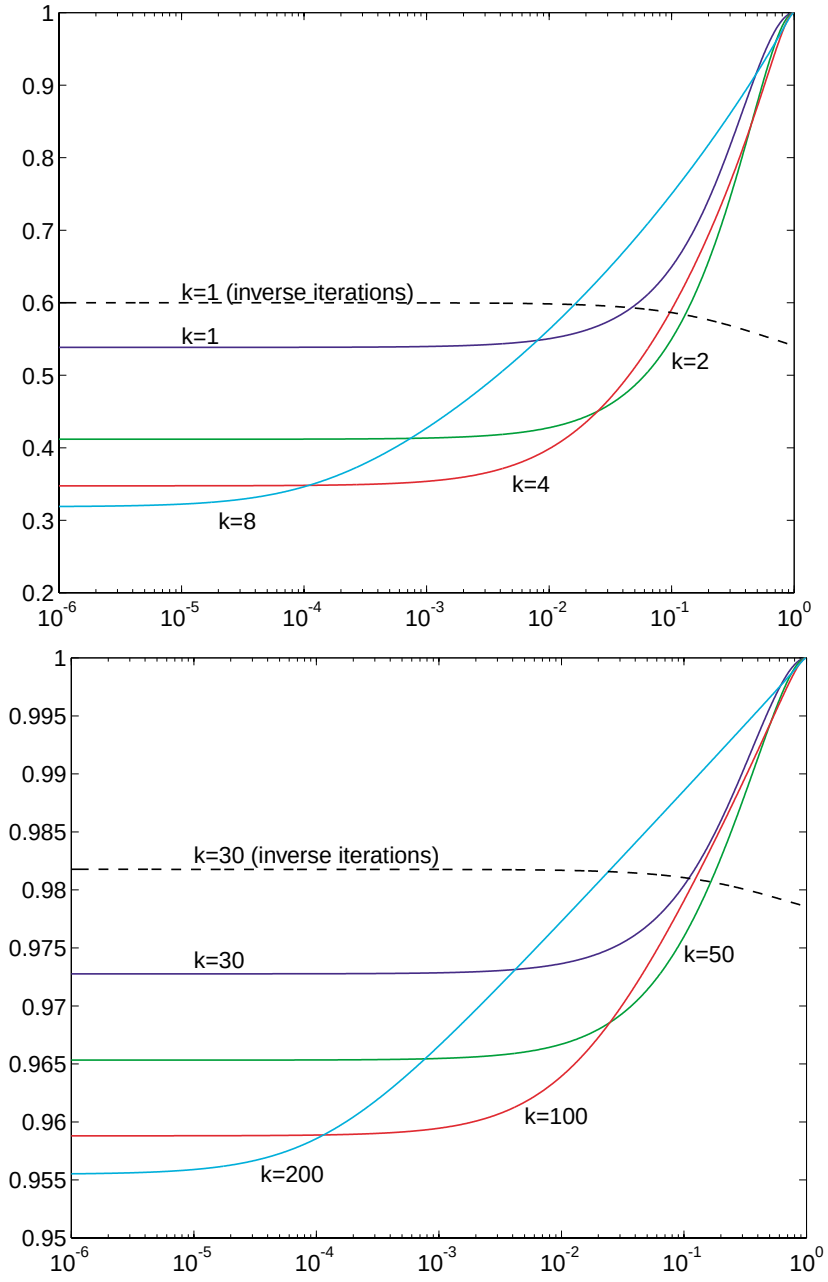


FIG. 2. $\bar{\sigma}^{1/k}(k,\theta)$ *versus* $(\theta-\lambda_1)/(\lambda_2-\theta)$ *for* $\kappa(K^{-1}A)=2$ *(top) and* $\kappa(K^{-1}A)=1000$ *(bottom);* $\lambda_2/\lambda_1=5/2$ *in both cases.*

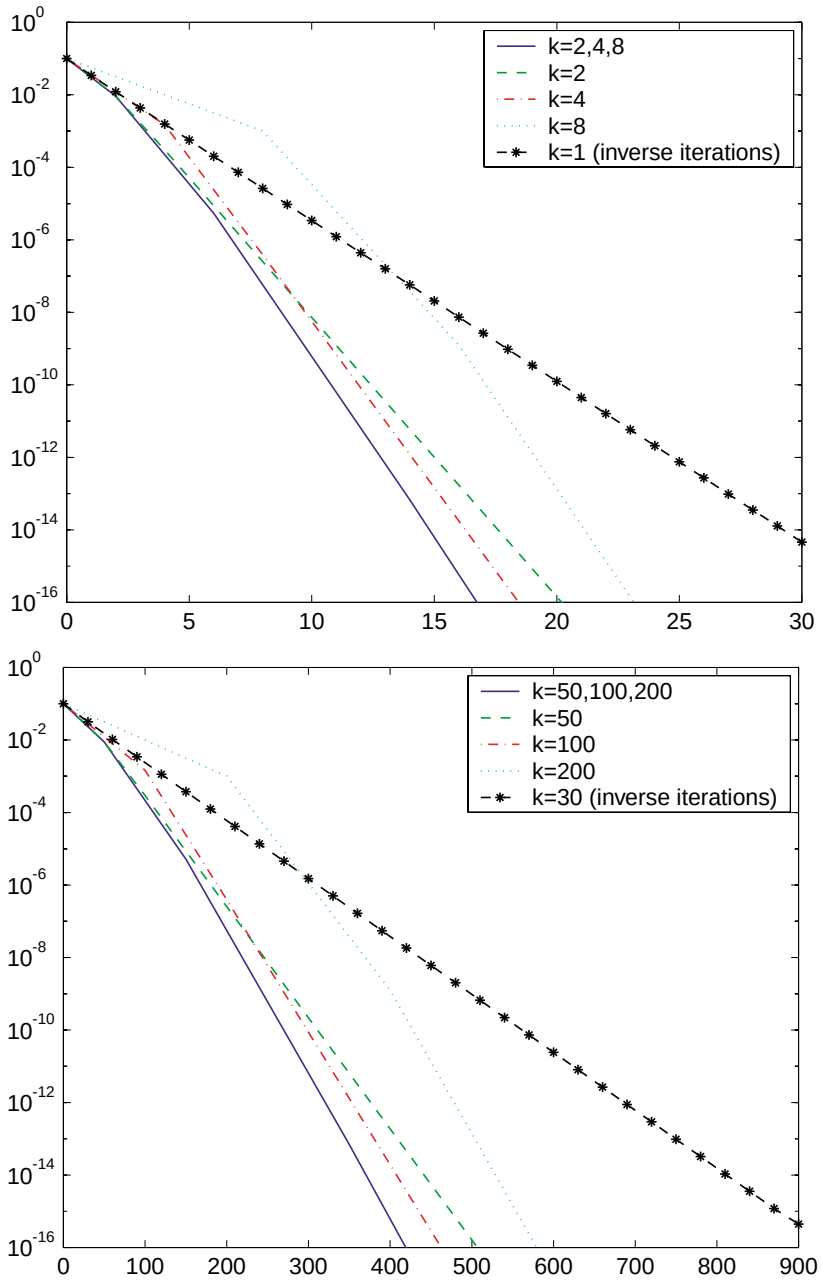


FIG. 3. *Evolution of* $(\theta-\lambda_1)/(\lambda_2-\theta)$ *in function of the cumulated number of inner iterations for* $\kappa(K^{-1}A)=2$ *(top) and* $\kappa(K^{-1}A)=1000$ *(bottom);* $\lambda_2/\lambda_1=5/2$ *in both cases.*

Acknowledgment. I thank Prof. A. Knyazev for having drawn my attention to (2.7) and for further stimulating discussions.

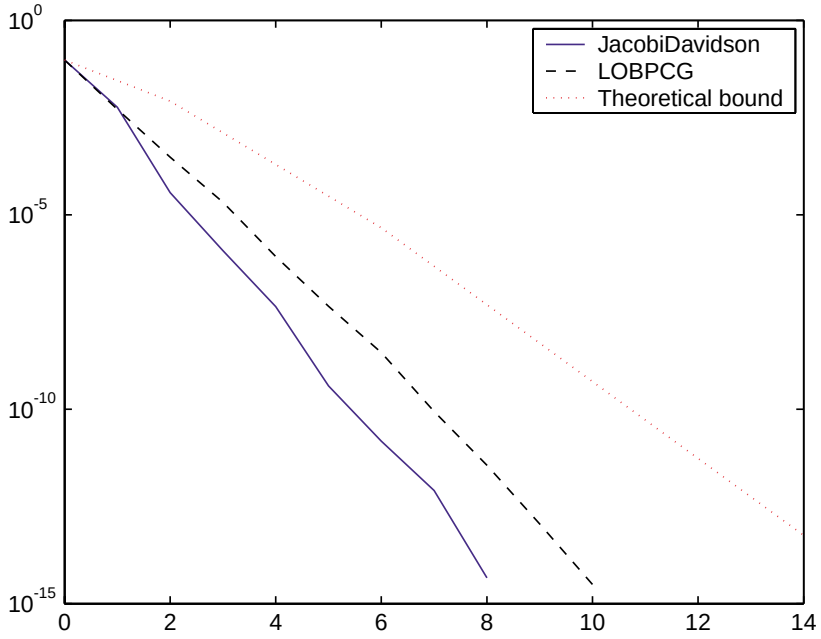


FIG. 4. *Evolution of* $(\theta - \lambda_1)/(\lambda_2 - \theta)$ *in function of the cumulated number of inner iterations (JD, theoretical bound) or in function of the number of iterations (LOBPCG) for the model example.*

REFERENCES

[1] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. A. VAN DER VORST, EDS., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000.

[2] F. A. DUL, *MINRES and MINERR are better than SYMMLQ in eigenpair computations*, SIAM J. Sci. Comput., 19 (1998), pp. 1767–1782.

[3] D. R. FOKKEMA, G. L. G. SLEIJPEN, AND H. A. VAN DER VORST, *Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1998), pp. 94–125.

[4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.

[5] G. H. GOLUB AND Q. YE, *Inexact inverse iterations for the generalized eigenvalue problems*, BIT, 40 (2000), pp. 672–684.

[6] A. V. KNYAZEV, *Computation of Eigenvalues and Eigenvectors for Mesh Problems: Algorithms and Error Estimates*, Dept. Numerical Math., USSR Academy of Sciences, Moscow, 1986 (in Russian).

[7] A. V. KNYAZEV, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[8] A. V. KNYAZEV, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[9] A. V. KNYAZEV AND K. NEYMEYR, *A geometric theory for preconditioned inverse iteration* III: *A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114; also available online from http://www-math.cudenver.edu/ccmreports/rep173.pdf, CU-Denver, 2001.

[10] Y.-L. LAI, K.-Y. LIN, AND W.-W. LIN, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 4 (1997), pp. 425–437.

[11] K. NEYMEYR, *A geometric theory for preconditioned inverse iteration* I: *Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001), pp. 61–85.

[12] K. NEYMEYR, *A geometric theory for preconditioned inverse iteration* II: *Convergence estimates*, Linear Algebra Appl., 322 (2001), pp. 87–104.

[13] K. Neymeyr, *A Hierarchy of Preconditioned Eigensolvers for Elliptic Differential Operators*, habilitationsschrift an der mathematischen fakultät, Universität Tübingen, Tübingen, Germany, 2001.
[14] Y. Notay, *Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.
[15] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[16] G. Peters and J. H. Wilkinson, *Inverse iteration, ill-conditioned equations and Newton's method*, SIAM Rev., 21 (1979), pp. 339–360.
[17] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.
[18] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, New York, 1996.
[19] V. Simoncini and L. Eldén, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.
[20] G. L. G. Sleijpen, A. Booten, D. R. Fokkema, and H. A. van der Vorst, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.
[21] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[22] G. L. G. Sleijpen and H. A. van der Vorst, *Hermitian eigenvalue problems*, *Generalized Hermitian eigenvalue problems*, in Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, Software Environ. Tools 11, SIAM, Philadelphia, 2000, Chapters 4.7, 5.6.
[23] G. L. G. Sleijpen, H. A. van der Vorst, and E. Meijerink, *Efficient expansion of subspaces in the Jacobi-Davidson method for standard and generalized eigenproblems*, Electron. Trans. Numer. Anal., 7 (1998), pp. 75–89.
[24] P. Smit and M. H. C. Paardekooper, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.
[25] D. B. Szyld, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM J. Numer. Anal., 25 (1988), pp. 1369–1375.
[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 645–659

REAL-VALUED, LOW RANK, CIRCULANT APPROXIMATION*

MOODY T. CHU† AND ROBERT J. PLEMMONS‡

Abstract. Partially due to the fact that the empirical data collected by devices with finite bandwidth often neither preserves the specified structure nor induces a certain desired rank, retrieving the nearest structured low rank approximation from a given data matrix becomes an imperative task in many applications. This paper investigates the case of approximating a given target matrix by a real-valued circulant matrix of a specified, fixed, and low rank. A fast Fourier transform (FFT)-based numerical procedure is proposed to speed up the computation. However, since a conjugate-even set of eigenvalues must be maintained to guarantee a real-valued matrix, it is shown by numerical examples that the nearest real-valued, low rank, and circulant approximation is sometimes surprisingly counterintuitive.

Key words. real-valued circulant matrix, lower rank, nearest approximation, conjugate-even, fast Fourier transform, truncated singular value decomposition

AMS subject classifications. 41A29, 41A50, 15A18, 65F35, 15A60

PII. S0895479801383166

1. Introduction.

Finding a low rank approximation of a general data matrix is a critical task in many aspects. The list of applications includes image compression, noise reduction, seismic inversion, latent semantic indexing, principal component analysis, regularization for ill-posed problems, and so on. Practical means to tackle this low rank approximation problem include the truncated singular value decomposition (TSVD) method [9], the Lanczos bidiagonalization process [14], and the Monte Carlo algorithm [11]. When the underlying matrix is also required to retain a certain structure, however, few techniques are available. Some preliminary discussion on structured low rank approximation regarding its mathematical properties, interesting applications, and an outline of some possible numerical procedures can be found in [5]. This paper concerns the special case of real-valued low rank approximation with circulant structure.

By an $n \times n$ circulant matrix, we mean a matrix C of the form

$$C = \begin{bmatrix} c_0 & c_1 & & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & \cdots & c_{n-3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ c_1 & c_2 & & c_{n-1} & c_0 \end{bmatrix},$$

where each of its rows is just the previous row cycled forward one step. A circulant matrix is uniquely determined by the entries of its first row. We shall denote

*Received by the editors January 3, 2001; accepted for publication (in revised form) by P. C. Hansen May 23, 2002; published electronically January 23, 2003.

http://www.siam.org/journals/simax/24-3/38316.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 grant DMS-9803759.

‡Department of Computer Science and Mathematics, Wake Forest University, Winston-Salem, NC 27109 (plemmons@mthcsc.wfu.edu). The research of this author was supported in part by the Air Force Office of Scientific Research under grant AFOSR-F49620-00-1-0155, the Army Research Office under grant DAAD-19-00-1-0540, and the National Science Foundation under grant CCR-9732070.

a circulant matrix by *Circul*($\mathbf{c}$) if its first row is $\mathbf{c}$. In this paper, we are mainly concerned with the case when $\mathbf{c} \in \mathbb{R}^n$.

Let Π ($= \Pi_n$) denote the specific permutation matrix of order n,

$$\Pi := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & & 1 \\ 1 & 0 & & \dots & 0 \end{bmatrix}. \tag{1.1}$$

It is easy to see that

$$C = \sum_{k=0}^{n-1} c_k \Pi^k \tag{1.2}$$

if and only if $C = \mathit{Circul}(\mathbf{c})$ with $\mathbf{c} := [c_0, \dots, c_{n-1}]$. It is convenient to represent this relationship as

$$\mathit{Circul}(\mathbf{c}) = P_{\mathbf{c}}(\Pi), \tag{1.3}$$

where

$$P_c(x) = \sum_{k=0}^{n-1} c_k x^k \tag{1.4}$$

is called the characteristic polynomial of *Circul*($\mathbf{c}$). Because of this representation, it follows that circulant matrices are closed under multiplication. It is also clear that circulant matrices commute under multiplication. Many important properties of circulant matrices can be traced back mainly to those of the matrix Π. The circulant structure often makes it possible to resolve many matrix-theoretic questions by "closed form" answers. The book by Davis [6] is generally considered the most complete reference on circulant matrices. It is also well known that circulant matrices are closely related to Fourier analysis [15]. That relationship will be used to develop a fast algorithm in this paper.

Circulant matrices have received much attention because the circulant form arises from areas such as acoustics, electrodynamics, image processing, mathematical statistics, number theory, numerical analysis, and stationary time series. To mention a few specific examples, circulant matrices often are used as preconditioners for ill-posed problems [2, 13]. In a recent book by Kailath and Sayed [10], circulant matrices are related to important applications of linear estimation theory. Circulant matrices even find applications to multiconjugate adaptive optics, as was discussed in [7, 8, 12].

Our goal in this paper is to retrieve as much information as possible from a given real-valued matrix A while enforcing a circulant structure and a rank condition; that is, we want to best approximate A with a real-valued circulant matrix C with a certain desired rank. Before moving on, we first point out the following three limitations imposed upon our approximation:

- We are emphasizing real-valued approximation. If there is no constraint requiring C to have real coefficients, then the nearest circulant approximation can easily be achieved via the notion of TSVD. (See Algorithm 3.1.)

- We are fixing the rank to a specific value and not to a certain range. In the latter, say, under the circumstances where singular values of the data matrix decay gradually to zero, the *precise* number of singular values included in a TSVD solution might not be very important. Such a flexibility on the rank condition, as was discussed in [5], is much easier to handle than the fixed rank condition.
- If the Frobenius matrix norm is used as the measurement of nearness, we may assume without loss of generality that the original matrix A is the Chan circulant matrix [3] to begin with. This can easily be seen from the fact that circulant matrices form a linear subspace, and thus the square of the distance from a target matrix to its nearest low rank circulant approximation is the sum of the squares of the distance from the target matrix to the linear subspace of circulant matrices (and hence its Chan circulant approximation) and the distance from the Chan circulant approximation to its nearest low rank circulant approximation.

Under these constraints, we follow the notion of the TSVD to propose a fast Fourier transform (FFT)-based fast algorithm. In order to keep the final low rank approximation a real-valued matrix, we recast the approximation as a data matching problem. As it turns out, we discover a situation where sometimes one may have to delete the largest eigenvalue in order to obtain a real-valued matrix. This surprising and somewhat counterintuitive case might not be significant in applications since, when the precise rank is not critically important, one may slightly relax the rank condition (say, from holding the given rank exactly to being no greater than the given rank), as we have indicated in the second bulleted item above. However, this discussion still might be worth noting in that it clearly demonstrates the disparity between fixed rank and variable rank and real-valued and complex-valued approximations.

2. Basic spectral properties. In this section, we briefly review some of the basic spectral properties relevant to our study. Most of the proofs can be found in [6, 15].

Let $i := \sqrt{-1}$. For a fixed integer $n \geq 1$, let ω $(= \omega_n)$ denote the primitive nth root of unity

$$\omega := \exp\left(\frac{2\pi i}{n}\right). \tag{2.1}$$

Let Ω $(= \Omega_n)$ denote the diagonal matrix

$$\Omega := \operatorname{diag}(1, \omega, \omega^2, \ldots, \omega^{n-1}), \tag{2.2}$$

and let F $(= F_n)$ denote the so-called discrete Fourier matrix whose Hermitian adjoint F^* is defined by

$$F^* := \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & \omega & \omega^2 & \ldots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \ldots & \omega^{2n-2} \\ \vdots & & & & \vdots \\ 1 & \omega^{n-1} & \omega^{n-2} & \ldots & \omega \end{bmatrix}. \tag{2.3}$$

Note that $\sqrt{n}F^*$ is the Vandermonde matrix generated by the row vector $[1, \omega, \omega^2, \ldots, \omega^{n-1}]$ and that F is a unitary matrix. The following spectral decomposition is a key to our discussion [6].

THEOREM 2.1. *The forward shift matrix* Π *is unitarily diagonalizable. Indeed,*

$$\Pi = F^* \Omega F. \tag{2.4}$$

The circulant matrix Circul$(\mathbf{c})$ *with any given row vector* $\mathbf{c}$ *has a spectral decomposition*

$$Circul(\mathbf{c}) = F^* P_{\mathbf{c}}(\Omega) F. \tag{2.5}$$

Observe that the vector of eigenvalues $\boldsymbol{\lambda} = [P_{\mathbf{c}}(1), \ldots, P_{\mathbf{c}}(\omega^{n-1})]$ of a circulant matrix *Circul*$(\mathbf{c})$ can quickly be calculated from

$$\boldsymbol{\lambda}^T = \sqrt{n} F^* \mathbf{c}^T. \tag{2.6}$$

From (2.6), the inverse eigenvalue problem of finding a circulant matrix with a prescribed spectrum can also be answered easily: Given any vector $\boldsymbol{\lambda} := [\lambda_0, \ldots, \lambda_{n-1}]$, the circulant matrix *Circul*$(\mathbf{c})$ with $\mathbf{c}$ defined by

$$\mathbf{c}^T = \frac{1}{\sqrt{n}} F \boldsymbol{\lambda}^T \tag{2.7}$$

will have eigenvalues $\{\lambda_0, \ldots, \lambda_{n-1}\}$. It is important to note that the matrix-vector multiplication involved in either (2.6) or (2.7) is precisely that involved in the FFT. Thus both the eigenvalue problem and the inverse problem for circulant matrices can be answered in $O(n \log_2 n)$ floating point operations [15]. Observe also that, if all of the eigenvalues are distinct, then there are precisely $n!$ many distinct circulant matrices with the prescribed spectrum.

For real circulant matrices, every complex-valued eigenvalue has the corresponding complex conjugate as another eigenvalue. Indeed, the spectrum of any real circulant matrix necessarily appears in a more special order, called *conjugate-even* in [15]. In order to obtain a *real-valued* circulant matrix by using the FFT in (2.7) for the inverse eigenvalue problem, the vector $\boldsymbol{\lambda}$ of the prescribed eigenvalues must also be arranged in a conjugate-even order. More precisely, the following arrangement of eigenvalues allows for efficient FFT calculation for real data [15].

THEOREM 2.2. *If the eigenvalues are arranged in the order that*

1. $\boldsymbol{\lambda} := [\lambda_0, \lambda_1, \ldots, \lambda_{m-1}, \lambda_m, \overline{\lambda_{m-1}}, \ldots, \overline{\lambda_1}]$, *where* $\lambda_0, \lambda_m \in \mathbb{R}$ *and* $n = 2m$, *or*
2. $\boldsymbol{\lambda} := [\lambda_0, \lambda_1, \ldots, \lambda_m, \overline{\lambda_m}, \ldots, \overline{\lambda_1}]$, *where* $\lambda_0 \in \mathbb{R}$ *and* $n = 2m + 1$,

then the circulant matrix Circul$(\mathbf{c})$ *with* $\mathbf{c}$ *obtained from* (2.7) *is real-valued and has entries in the prescribed vector* $\boldsymbol{\lambda}$ *as its spectrum.*

For later reference, we shall refer to λ_0 and λ_m, if $n = 2m$, and λ_0, if $n = 2m + 1$, in the above theorem as the *absolutely real* elements in $\boldsymbol{\lambda}$.

The singular value decomposition of *Circul*$(\mathbf{c})$ is also easy to establish. It follows from rewriting the expression (2.5) as

$$Circul(\mathbf{c}) = \left(F^* P_{\mathbf{c}}(\Omega) |P_{\mathbf{c}}(\Omega)|^{-1}\right) |P_{\mathbf{c}}(\Omega)| F, \tag{2.8}$$

where $|X|$ denotes the matrix of absolute values of the elements of X. The singular values of *Circul*$(\mathbf{c})$ are $|P_{\mathbf{c}}(\omega^k)|$, $k = 0, 1, \ldots, n-1$. Observe the following necessary characteristic for singular values of a real circulant matrix.

THEOREM 2.3. *Any* $n \times n$ *real-valued circulant matrix can have at most* $\lceil \frac{n+1}{2} \rceil$ *distinct singular values. More precisely, the singular values must appear in the following way:*

1. $\sigma_{n_0}, \sigma_{n_1}, \sigma_{n_1}, \ldots, \sigma_{n_{m-1}}, \sigma_{n_{m-1}}, \sigma_{n_m}$ *if* $n = 2m$ *or*
2. $\sigma_{n_0}, \sigma_{n_1}, \sigma_{n_1}, \ldots, \sigma_{n_m}, \sigma_{n_m}$ *if* $n = 2m + 1$.

3. Low rank approximation. Given a general matrix $A \in \mathbb{R}^{n\times n}$, its nearest circulant matrix approximation measured in the Frobenius norm is simply the Chan circulant matrix *Circul*($\mathbf{c}$) obtained by averaging over diagonals of A, as shown in [3]. Indeed, if $\mathbf{c} = [c_0, \ldots, c_{n-1}]$, then

$$c_k := \frac{1}{n}\langle A, \Pi^k\rangle, \quad k = 0, \ldots, n-1, \tag{3.1}$$

where

$$\langle X, Y\rangle = \operatorname{trace}(XY^T)$$

stands for the Frobenius inner product. This projection *Circul*($\mathbf{c}$) is generally of full rank even if A has lower rank to begin with. Recall that the TSVD gives rise to the nearest low rank approximation in the Frobenius norm. Observe further that the low rank approximation *Circul*($\hat{\mathbf{c}}$) of a circulant matrix *Circul*($\mathbf{c}$) by the TSVD is automatically circulant. We thus have the following algorithm for low rank circulant approximation.

ALGORITHM 3.1. *Given a general $n\times n$ matrix A, the matrix Circul($\hat{\mathbf{c}}$) computed below is a nearest circulant matrix to A with rank no higher than $\kappa \le n$.*

1. *Use the projection* (3.1) *to find the nearest circulant matrix approximation Circul($\mathbf{c}$) of A.*
2. *Use the inverse FFT* (2.6) *to calculate the spectrum $\boldsymbol{\lambda}$ of the matrix Circul($\mathbf{c}$).*
3. *Let $\hat{\boldsymbol{\lambda}}$ be the vector consisting of elements of $\boldsymbol{\lambda}$, but those corresponding to the $n-\kappa$ smallest (in modulus) singular values are set to zero.*
4. *Apply the FFT* (2.7) *to $\hat{\boldsymbol{\lambda}}$ to compute a nearest circulant matrix Circul($\hat{\mathbf{c}}$) of rank κ to A.*

The above algorithm is fast due to the employment of efficient FFT calculation. The resulting matrix *Circul*($\hat{\mathbf{c}}$), however, is complex-valued in general. To construct real-valued low rank approximation, the truncated singular values must be specifically selected so that the resulting vector $\hat{\boldsymbol{\lambda}}$ of *truncated* eigenvalues is conjugate-even. Recall from Theorem 2.3 that most of the singular values are paired. Thus, to preserve the conjugate-even property, the deletion of one complex eigenvalue necessitates the deletion of its complex conjugate as well. To achieve the desired rank, the criteria for truncation must be modified in a special way, as we shall now describe.

It is clear from Theorem 2.1 that all circulant matrices of the same size have the same set of unitary eigenvectors. The real-valued low rank circulant approximation problem, therefore, is equivalent to the following data matching problem (DMP):

(DMP) *Given a conjugate-even vector $\boldsymbol{\lambda} \in \mathbb{C}^n$, find its nearest conjugate-even approximation $\hat{\boldsymbol{\lambda}} \in \mathbb{C}^n$ in the 2-norm subject to the constraint that $\hat{\boldsymbol{\lambda}}$ has exactly $n-\kappa$ zeros.*

Note that finding the closest vector approximation $\hat{\boldsymbol{\lambda}}$ in the 2-norm produces the closest matrix approximation in the Frobenius norm. If there were no conjugate-even constraint, the DMP could easily be answered. See, for example, [1, 4]. With the conjugate-even constraint, we claim that the DMP could be solved according to the following sorting scheme.

THEOREM 3.1. *The optimal solution $\hat{\boldsymbol{\lambda}}$ to the DMP must be such that its nonzero entries match precisely with the first κ conjugate-even components of $\boldsymbol{\lambda}$ according to the descending order of their moduli.*

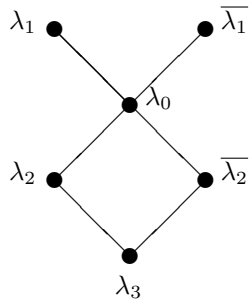


FIG. 1. *Tree graph of* $\lambda_1, \overline{\lambda_1}, \lambda_0, \lambda_2, \overline{\lambda_2}, \lambda_3$ *with* $|\lambda_1| \geq |\lambda_0| > |\lambda_2| \geq |\lambda_3|$.

Proof. Without loss of generality, we may write $\hat{\boldsymbol{\lambda}} = [\hat{\boldsymbol{\lambda}}_1, 0] \in \mathbb{C}^n$ with $\hat{\boldsymbol{\lambda}}_1 \in \mathbb{C}^\kappa$ to be determined and consider the problem of minimizing

$$F(P, \hat{\boldsymbol{\lambda}}) = \|P\hat{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T\|_2^2,$$

where the permutation matrix P is used to search for the match. Partition the permutation matrix into $P = [P_1, P_2]$ with $P_1 \in R^{n\times\kappa}$. The objective function in the least squares problem is reduced to

$$F(P, \hat{\boldsymbol{\lambda}}) = \|P_1\hat{\boldsymbol{\lambda}}_1^T - \boldsymbol{\lambda}^T\|_2^2,$$

which obviously has its optimal solution with

$$\hat{\boldsymbol{\lambda}}_1 = \boldsymbol{\lambda}P_1.$$

This proves the important fact that the entries of $\hat{\boldsymbol{\lambda}}_1$ must come from the rearrangement of κ components of $\boldsymbol{\lambda}$. Indeed, the objective function becomes

$$F(P, \hat{\boldsymbol{\lambda}}) = \|(P_1P_1^T - I)\boldsymbol{\lambda}\|_2^2,$$

where $P_1P_1^T - I$ is but a projection. To minimize $F(P, \boldsymbol{\lambda}P_1)$, the optimal permutation P should be such that $P_1P_1^T$ projects $\boldsymbol{\lambda}$ onto its first κ components with as large a modulus as possible while maintaining the conjugate-even condition. □

In other words, without the conjugate-even constraints, the answer to the DMP corresponds precisely to the usual selection criterion mentioned in Algorithm 3.1, i.e., $\hat{\boldsymbol{\lambda}}$ is obtained by setting the $n - \kappa$ elements of $\boldsymbol{\lambda}$ with smallest modulus to zeros. With the conjugate-even constraint, the above criterion remains effective, but the truncation also depends on the conjugate-even structure inside $\boldsymbol{\lambda}$, as we explain next.

Consider the case $n = 6$ as an example. We shall first assume that neither λ_1 nor λ_2 is a real number. There are six possible conjugate-even structures. For convenience, we shall denote each structure by a tree graph. Each node in the tree represents an element of $\boldsymbol{\lambda}$. Arrange the nodes from top to bottom according to the descending order of their moduli. In case of a tie, arrange the complex conjugate nodes at the same level, and place the real node below the complex nodes. Thus the conjugate-even structure $\lambda_1, \overline{\lambda_1}, \lambda_0, \lambda_2, \overline{\lambda_2}, \lambda_3$, arranged in the descending order of their moduli, will be denoted by the tree in Figure 1.

The nearest conjugate-even vectors to $\boldsymbol{\lambda}$ of rank 5, 3, and 2, respectively, are easy to determine. Their trees are given in Figure 2, where ∘ and • at each node denotes,

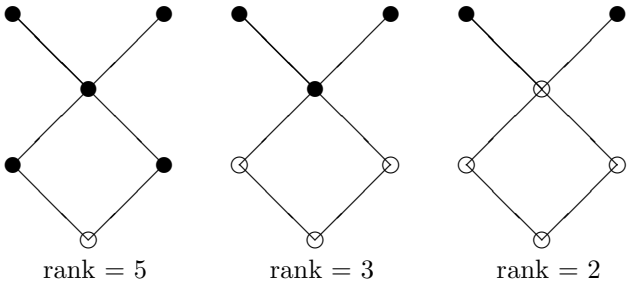


FIG. 2. *Tree graphs of* $\hat{\boldsymbol{\lambda}}$ *with rank* 5, 3, *and* 2.

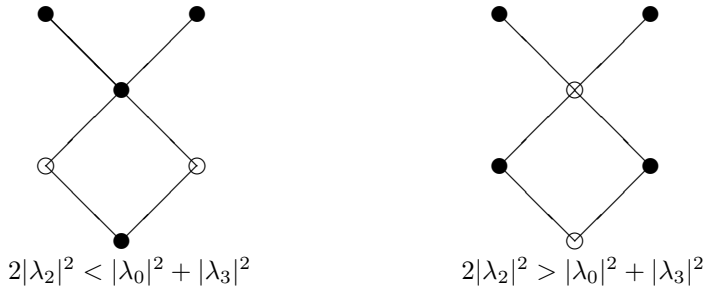


FIG. 3. *Tree graphs of* $\hat{\boldsymbol{\lambda}}$ *with rank* 4.

respectively, that the particular node is being replaced by zero or remains unchanged from the original tree. For these ranks and for this specific tree structure depicted in Figure 1, the conjugate-even requirement has no effect.

However, depending upon whether $2|\lambda_2|^2 > |\lambda_0|^2 + |\lambda_3|^2$, there are two choices for $\hat{\lambda}$ as the nearest conjugate-even approximation of rank 4. See Figure 3. Finally, the nearest rank-1 conjugate-even approximation for the tree of $\boldsymbol{\lambda}$ given by Figure 1 is depicted in Figure 4.

It should be noted that we have implicitly assumed that, if $n = 2m$, then the two absolutely real elements in a conjugate-even $\boldsymbol{\lambda}$ are λ_0 and λ_m and that $|\lambda_0| \geq |\lambda_m|$. We have also assumed that the remaining $2m - 2$ elements are "potentially" complex-valued (some of them could in fact turn out to be real-valued), that they are paired up (necessarily), and that they are arranged in descending order, i.e., $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{m-1}|$. A similar assumption can be made for the case in which $n = 2m + 1$. See the ordering stated in Theorem 2.2. Nevertheless, we will *never* assume any ordering relationship between the absolutely real element(s) and the potential complex elements. Indeed, it is precisely such an ordering relationship that will determine the truncation criteria as we have demonstrated above for the case in which $n = 6$. In other words, assuming that there are exactly $m + 1$ distinct absolute values of elements in $\boldsymbol{\lambda}$, there are exactly $m(m + 1)/2$ many possible conjugate-even structures for the case in which $n = 2m$, depending upon where the moduli of the absolutely real elements fit into the moduli of the potentially complex elements when a *total* ordering is taken.

Again, under the assumption that neither λ_1 or λ_2 is a real number, we further illustrate our point by considering other cases for $n = 6$ in Figure 5. The leftmost column in Figure 5 represents the six possible conjugate-even structures of $\boldsymbol{\lambda}$ when

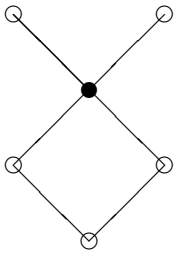


FIG. 4. *Tree graph of* $\hat{\boldsymbol{\lambda}}$ *with rank* 1.

elements are arranged in descending order of their moduli. For each fixed structure, moving from left to right, Figure 5 demonstrates the plan of how the nodes on the original tree should be "pruned" to solve the DMP for various lower rank conditions. There are four cases, A4, A2, B2, and F4, in which additional comparisons are needed to further discern which plan should be used. This situation happens when an even number of nodes from a "loop" are to be dropped. We have already discussed the case F4 in Figure 3. Other cases can easily be identified.

It is entirely possible that there are real-valued elements other than the two (when n is even) absolutely real elements in a conjugate-even λ. The eigenvalues of a symmetric circulant matrix, for instance, are conjugate-even and are all real. When this happens, these conjugate-even real-valued elements must appear in pairs, and the truncation criteria are further complicated. Using the example discussed in Figure 1 but assuming further that $\lambda_2 = \overline{\lambda_2}$, we illustrate our point below. First, we use a dashed link in Figure 6 and larger dots to indicate the occurrence of $\lambda_2 = \overline{\lambda_2}$. It is important to note that, in contrast to the two drawings in Figure 3, the tree graph of the nearest conjugate-even approximation $\hat{\lambda}$ with rank 4 changes its structure in this case. See Figure 7.

4. Algorithm. While the aforementioned graph-theoretic concept should be quite easy to follow, a general purpose code is not as straightforward. To facilitate the discussion, we now present an algorithm for computing the real-valued low rank circulant approximation. In order to highlight the notion on how the singular values of *Circul*($\mathbf{c}$) should be truncated, we simplify many computational operations by adopting a pseudo-MATLAB syntax. The commands for these abridged operations are denoted in boldface (whereas, to avoid distraction, the vectors $\mathbf{c}$ and $\boldsymbol{\lambda}$ are denoted as ordinary c and λ) in the steps of the following algorithm.

ALGORITHM 4.1. *Given* $\mathbf{c} \in \mathbb{R}^n$ *and a positive integer* $1 \le \kappa < n$, *let* $m = \lfloor \frac{n}{2} \rfloor$. *Define* $tol = n\epsilon\|\mathbf{c}\|$ *where* ϵ *is the machine accuracy as the threshold of system zero. The matrix Circul*($\hat{\mathbf{c}}$) *with* $\hat{\mathbf{c}}$ *computed at the end of the following steps has eigenvalues* $\hat{\boldsymbol{\lambda}}$ *containing exactly* $n - \kappa$ *zeros and is the nearest approximation to Circul*($\mathbf{c}$).

1. $\lambda = n * \mathbf{ifft}(c)$; *(Indices of* λ *start with* 1.*)*
   $\hat{\lambda} = \lambda(1 : m + 1)$;
2. *if* $n = 2m$
   $I_r = \mathbf{find}(\mathbf{abs}(\mathbf{imag}(\lambda(2 : m))) < tol) + 1$;
   $I_c = \mathbf{find}(\tilde{}\mathbf{ismember}(2 : m, I_r)) + 1$;
   else
   $I_r = \mathbf{find}(\mathbf{abs}(\mathbf{imag}(\lambda(2 : m + 1))) < tol) + 1$;
   $I_c = \mathbf{find}(\tilde{}\mathbf{ismember}(2 : m + 1, I_r)) + 1$;

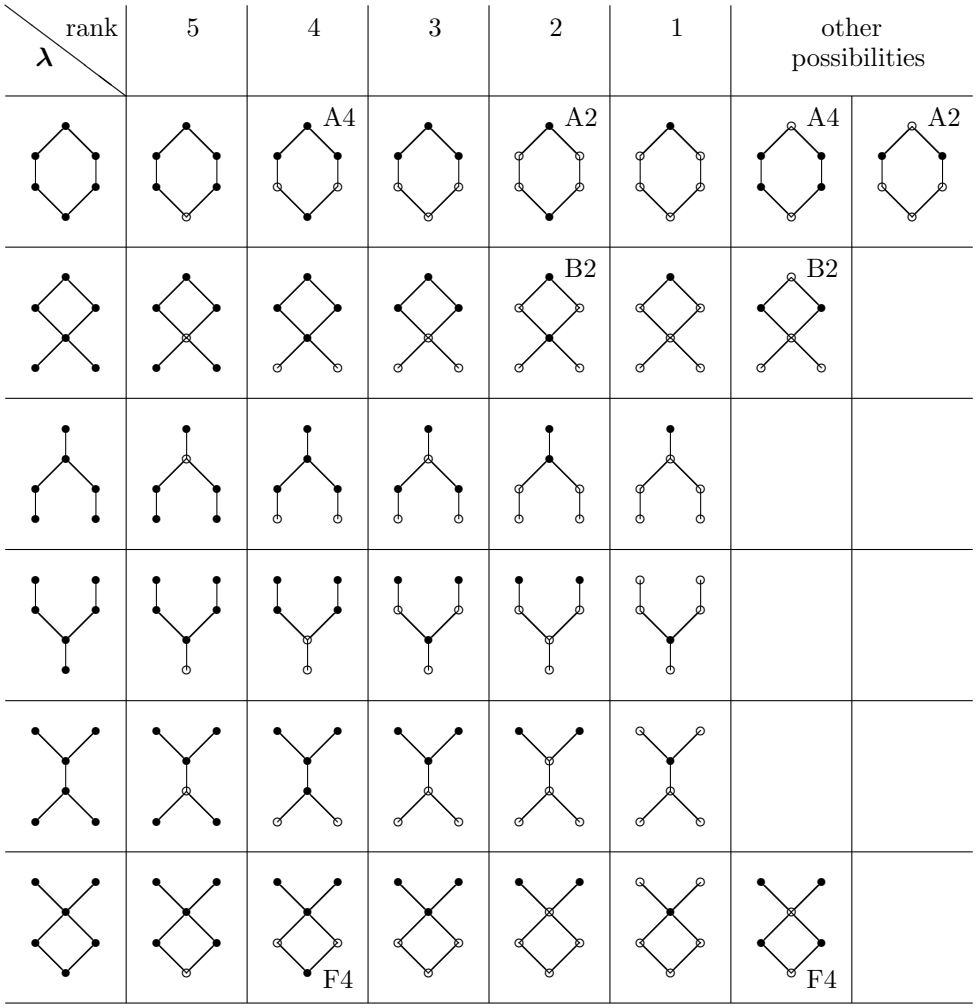


FIG. 5. *Possible solutions to the DMP when* $n = 6$.

end

3. $[t, J] = \mathbf{sort}(\mathbf{abs}(\lambda))$*;*
   $J = \mathbf{fliplr}(J)$*;* *(J is the index set sorting λ in descending order.)*
   for $i = 1 : m + 1$

$$I(:,:,i) = \begin{cases} [2,0]\,, & \textit{if } \mathbf{ismember}(J(i), I_c); \\ [2,1]\,, & \textit{if } \mathbf{ismember}(J(i), I_r); \\ [1,1]\,, & \textit{otherwise}; \end{cases}$$

   end

4. $\sigma = 0$*;*
   $s = m + 1$*;*
   while $\sigma < n - \kappa$
   $\sigma = \sigma + I(1, 1, s)$*;*
   $s = s - 1$*;*
   end
   $idx = s + 1$*;* *(idx indicates the place where λ is to be cut.)*

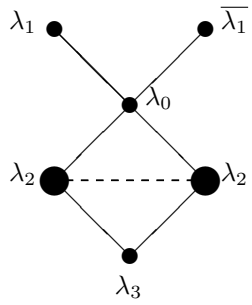


FIG. 6. *Tree graph of* $\lambda_1, \overline{\lambda_1}, \lambda_0, \lambda_2, \lambda_2, \lambda_3$ *with* $|\lambda_1| \geq |\lambda_0| > |\lambda_2| \geq |\lambda_3|$.

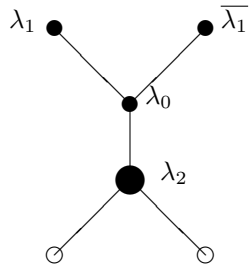


FIG. 7. *Tree graph of* $\hat{\boldsymbol{\lambda}}$ *with rank* 4 *when* $\lambda_2 = \overline{\lambda_2}$.

5. *if* $\sigma = n - \kappa$
   $\hat{\lambda}(J(idx : m+1)) = \mathbf{zeros}(1, m - idx + 2)$*;*
   go to 7
6. $k_\kappa = \mathbf{min}(\mathbf{find}(I(1,1,idx+1:m+1) == 1)) + idx$*;*
   $k_u = \mathbf{max}(\mathbf{find}(I(1,1,1:idx) == 1))$*;*
   if $I(:,:,idx) == [2,1]$
      if ~**isempty**(k_κ)
         $\hat{\lambda}(J(idx : m+1)) = \mathbf{zeros}(1, m - idx + 2)$*;*
         $\hat{\lambda}(J(k_\kappa)) = \lambda(J(idx))$*;*
      else
         $\hat{\lambda}(J(k_u)) = 0$*;*
         $\hat{\lambda}(J(idx+1 : m+1)) = \mathbf{zeros}(1, m - idx + 1)$*;*
      end
   else
      if ~**isempty**(k_κ)
         if **isempty**(k_u)
            $\hat{\lambda}(J(idx : m+1)) = \mathbf{zeros}(1, m - idx + 2)$*;*
            $\hat{\lambda}(J(k_\kappa)) = \lambda(J(k_\kappa))$*;*
         else
            $t_1 = 2 * \mathbf{abs}(\lambda(J(idx)))^2$*;*
            $t_2 = \mathbf{abs}(\lambda(J(k_u)))^2 + \mathbf{abs}(\lambda(J(k_\kappa)))^2$*;*
            if $t_1 \leq t_2$
               $\hat{\lambda}(J(idx : m+1)) = \mathbf{zeros}(1, m - idx + 2)$*;*
               $\hat{\lambda}(J(k_\kappa)) = \lambda(J(k_\kappa))$*;*

else
$\hat{\lambda}(J(idx+1:m+1)) = \mathbf{zeros}(1, m-idx+1)$*;*
$\hat{\lambda}(J(k_u)) = 0$*;*
end
end
else
$\hat{\lambda}(J(k_u)) = 0$*;*
$\hat{\lambda}(J(idx+1:m+1)) = \mathbf{zeros}(1, m-idx+1)$*;*
end
end

7. $\hat{\lambda} = \begin{cases} \left[\hat{\lambda}, \mathbf{fliplr}(\mathbf{conj}(\hat{\lambda}(2:m)))\right], & \text{if } n = 2m; \\ \left[\hat{\lambda}, \mathbf{fliplr}(\mathbf{conj}(\hat{\lambda}(2:m+1)))\right], & \text{if } n = 2m+1; \end{cases}$
$\hat{c} = \mathbf{real}(\mathbf{fft}(\hat{\lambda}))/n$*;*

5. Numerical examples. In this section, we illustrate our algorithm with some numerical examples. We report all numerics using only four significant digits, although entries of all matrices in consideration and the corresponding eigenvalues are originally the full length of the double precision. All calculations are done using MATLAB.

Example 1. Consider the 8×8 symmetric circulant matrix whose first row is given by a randomly generated vector

$$\mathbf{c} = [0.5404, 0.2794, 0.1801, -0.0253, -0.2178, -0.0253, 0.1801, 0.2794].$$

The corresponding eigenvalues, arranged in descending order of their moduli, are

$$\{1.1909, 1.1891, 1.1891, 0.3273, 0.3273, 0.1746, -0.0376, -0.0376\}.$$

The singular values clearly are given by the moduli of these eigenvalues. Observe the parity caused by the conjugate-evenness, whereas 1.1909 and 0.1746 are what we call absolutely real eigenvalues.

The nearest circulant approximation of rank 7 would be simply to set the last eigenvalue, i.e., -0.0376, to zero by using Algorithm 3.1, but such a TSVD approach would result in a complex matrix. To obtain the nearest real-valued circulant approximation of rank 7, we have to keep the pair of -0.0376 and zero out the value 0.1746. Using the conjugate-even eigenvalues

$$\hat{\boldsymbol{\lambda}} = [1.1909, 1.1891, -0.0376, 0.3273, 0, 0.3273, -0.0376, 1.1892],$$

we can construct the nearest real-valued rank-7 circulant approximation to $Circul(\mathbf{c})$ via the FFT and obtain the first row vector

$$\hat{\mathbf{c}} = [0.5186, 0.3012, 0.1583, -0.0035, -0.2396, -0.0035, 0.1583, 0.3012].$$

In yet another scenario, the first row of the nearest rank 4 circulant approximation is given by the row vector

$$\hat{\mathbf{c}} = [0.4871, 0.3182, 0.1898, -0.1023, -0.1075, -0.1023, 0.1898, 0.3182]$$

with eigenvalues $\hat{\boldsymbol{\lambda}}$

$$\hat{\boldsymbol{\lambda}} = [1.1909, 1.1892, 0, 0, 0.3273, 0, 0, 1.1892],$$

where we see that the last pair of eigenvalues in $\boldsymbol{\lambda}$ are set to zero while the value 0.1746 together with the value 0.3273 cause a topology change in the graph tree the same as in Figures 6 and 7.

Example 2. Consider the 9×9 circulant matrix whose first row is given by

$$\mathbf{c} = [1.6864, 1.7775, 1.9324, 2.9399, 1.9871, 1.7367, 4.0563, 1.2848, 2.5989].$$

The corresponding eigenvalues have conjugate-even structure given by

$$[20.0000, -2.8130 \pm 1.9106i, 3.0239 \pm 1.0554i, -1.3997 \pm 0.7715i, -1.2223 \pm 0.2185i].$$

Note the absolute real eigenvalue has modulus much larger than any other eigenvalues. To obtain a real-valued circulant approximation of rank 8, we have no choice but to select the vector (in its ordering)

$$\begin{aligned}\hat{\boldsymbol{\lambda}} = [&0, -2.8130 - 1.9106i, 3.0239 - 1.0554i,\\ &-1.3997 - 0.7715i, -1.2223 + 0.2185i, -1.2223 - 0.2185i,\\ &-1.3997 + 0.7715i, 3.0239 + 1.0554i, -2.8130 + 1.9106i]\end{aligned}$$

to produce

$$\hat{\mathbf{c}} = [-0.5358, -0.5872, -1.1736, -0.3212, 1.0198, 1.4013, -0.0761, -0.4115, 0.6844]$$

as the first row of its nearest real-valued circulant approximation. The fact that the *largest* eigenvalue (singular value) of *Circul*($\mathbf{c}$) must be set to zero to produce the nearest rank-8 approximation is quite counterintuitive to the usual sense of TSVD approximation.

On the other hand, it is worth noting that, if we slightly modify our approximation criteria by requesting only a nearest low rank approximation with rank *no greater than* 8, the answer could be completely different. In this particular example, such a nearest matrix turns out to be of rank 7 and is in agreement with the usual TSVD approximation by truncating the *pair* of eigenvalues with the smallest moduli.

Example 3. Let $C_\kappa \in \mathbb{R}^{n\times n}$ be a given circulant matrix of rank κ. With probability one, any random noise added to C_κ will destroy the circulant structure as well as the rank condition. To establish a comparison, we may assume, without loss of generality, that, after the projection step (3.1) mentioned in Algorithm 3.1, the added noise is a circulant matrix. Let $E \in \mathbb{R}^{n\times n}$ denote a random but fixed circulant matrix with unit Frobenius norm. Consider the perturbation of C_κ by an additive noise of magnitude (in Frobenius norm) 10^{-j}; i.e., consider the circulant matrices

$$W_j = C_\kappa + 10^{-j}E, \quad j = 1, \ldots, 12.$$

It is almost certain that, under such a random perturbation, the matrix W_j will be of full rank. Note that $\|W_j - C_\kappa\| = 10^{-j}$. It will be interesting to see if W_j has any closer circulant matrix approximation of rank κ than C_κ, especially when j is large.

Toward that end, we report a test case with $n = 100$, $\kappa = 73$, and a predetermined matrix C_{73}. In Figure 8, the (continuous) lines depict the distribution of singular values of the perturbed matrices W_j for $j = 1, \ldots, 12$, respectively, whereas the singular values of the original C_{73} are marked by $*$. Observe how the perturbation affects the last 27 (machine zero) singular values of C_{73} more significantly than the first 73 (larger) singular values according to the magnitude 10^{-j}.

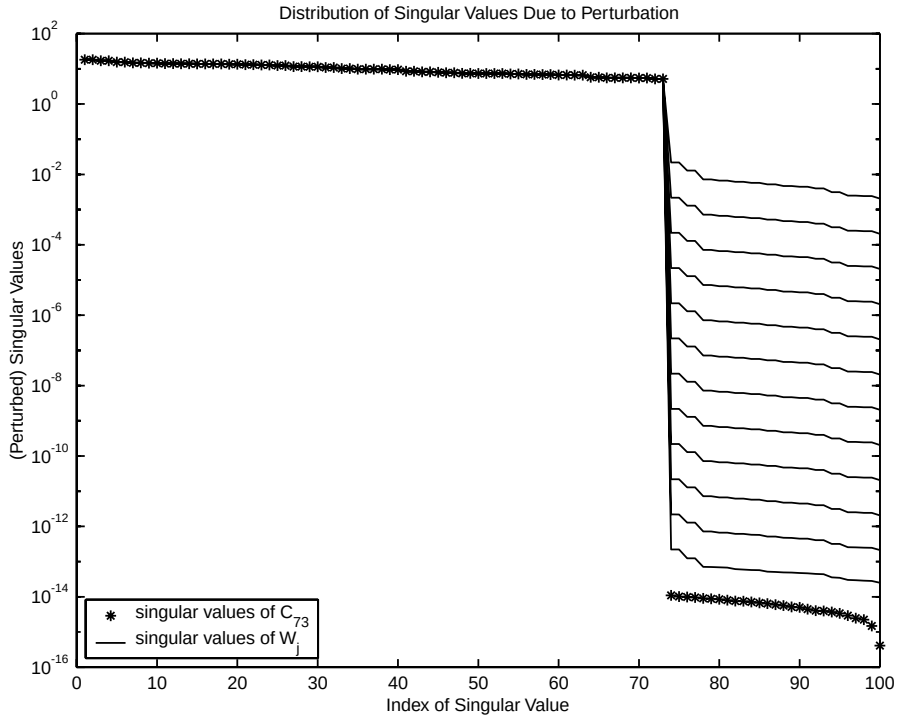


FIG. 8. *Distribution of singular values.*

Using our algorithm to find the best circulant approximation Z_j to W_j, we find that it is always the case that

$$\|W_j - Z_j\| < \|W_j - C_\kappa\|$$

for each j. This is indicated in Figure 9 by the fact that the circles $\circ$ are always below the diagonal line. Also marked in Figure 9 by $+$ signs is the difference between Z_j and C_κ.

6. Summary. Structured low rank approximation is an important and challenging task both theoretically and computationally. The special case of real-valued, low rank approximation with circulant structure is studied in this paper. For any given real data matrix, its nearest real circulant approximation can simply be determined from the average of its diagonal entries. Its nearest lower rank approximation can also be determined effectively from the TSVD and the FFT. However, such an approximation usually will be complex-valued. To simultaneously maintain the circulant structure, induce a specific lower rank, and keep the matrix real, the conjugate-even structure must be taken into account. These requirements, in turn, can substantially change the truncation criteria. Some counterintuitive examples illustrate the effect if the rank is fixed to a precise number and the approximation is required to be real-valued. We have proposed a fast algorithm to accomplish all of these approximation objectives.

Acknowledgments. The authors would like to thank Nikos Pitsianis and Xiaobai Sun for helpful comments that motivated much of this work.

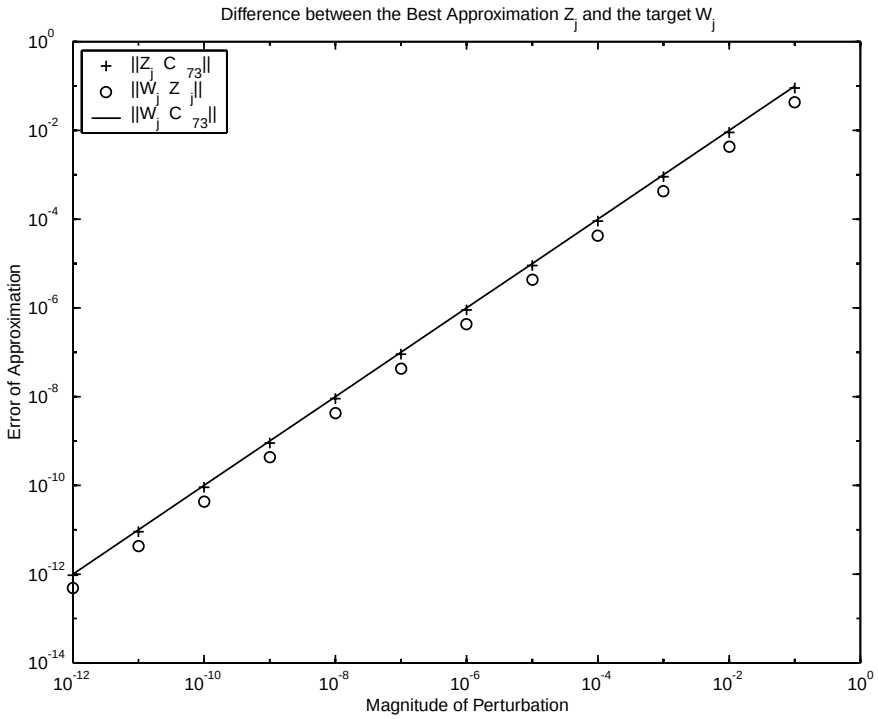


FIG. 9. *Errors in approximation.*

REFERENCES

[1] R. W. BROCKETT, *Least square matching problems*, Linear Algebra Appl., 122/123/124 (1989), pp. 761–777.

[2] R. H. CHAN, J. G. NAGY, AND R. J. PLEMMONS, *Circulant preconditioned Toeplitz least squares iterations*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 80–97.

[3] T. F. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

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

[5] M. T. CHU, R. E. FUNDERLIC, AND R. J. PLEMMONS, *Structured low rank approximation*, Linear Algebra Appl., to appear.

[6] P. J. DAVIS, *Circulant Matrices*, John Wiley and Sons, New York, 1979.

[7] B. L. ELLERBROEK AND F. J. RIGAUT, *Methods for correcting tilt anisoplanatism in laser-guide-star-based multi-conjugate adaptive optics*, J. Opt. Soc. Amer. A, 18 (2001), pp. 2539–2547.

[8] M. HANKE, J. G. NAGY, AND R. J. PLEMMONS, *Preconditioned iterative regularization for ill-posed problems*, in Numerical Linear Algebra and Scientific Computing, De Gruyter Press, Berlin, 1993, pp. 141–163.

[9] P. C. HANSEN, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM Monogr. Math. Model. Comput. 4, SIAM, Philadelphia, PA, 1997.

[10] T. KAILATH, A. H. SAYED, AND B. HASSIBI, *Linear Estimation*, Prentice Hall, Upper Saddle River, NJ, 2000.

[11] A. FRIEZE AND R. KANNAA, *Fast Monte-Carlo Algorithm for Finding Low Rank Approximations*, available online from http://www.cs.yale.edu/~kannan/ (1998).

[12] J. G. NAGY, V. P. PAUCA, R. J. PLEMMONS, AND T. C. TORGERSEN, *Space-varying restoration of optical images*, J. Opt. Soc. Amer. A, 14 (1997), pp. 3162–3174.

[13] M. Ng, *Circulant preconditioners for convolution-like integral equations with higher order quadrature rules*, Electron. Trans. Numer. Anal., 6 (1997), pp. 18–28.
[14] 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.
[15] C. F. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, Frontiers Appl. Math. 10, SIAM, Philadelphia, PA, 1992.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 660–677

INVERSION OF DISPLACEMENT OPERATORS*

VICTOR Y. PAN† AND XINMAO WANG‡

Abstract. We recall briefly the displacement rank approach to the computations with structured matrices, which we trace back to the seminal paper by Kailath, Kung, and Morf [*J. Math. Anal. Appl.*, 68 (1979), pp. 395–407]. The concluding stage of the computations is the recovery of the output from its compressed representation via the associated displacement operator L. The recovery amounts to the inversion of the operator. That is, one must express a structured matrix M via its image $L(M)$. We show a general method for obtaining such expressions that works for all displacement operators (under only the mildest nonsingularity assumptions) and thus provides the foundation for the displacement rank approach to practical computations with structured matrices. We also apply our techniques to specify the expressions for various important classes of matrices. Besides unified derivation of several known formulae, we obtain some new ones, in particular, for the matrices associated with the tangential Nevanlinna–Pick problems. This enables acceleration of the known solution algorithms. We show several new matrix representations of the problem in the important confluent case. Finally, we substantially improve the known estimates for the norms of the inverse displacement operators, which are critical numerical parameters for computations based on the displacement approach.

Key words. structured matrices, displacement rank, inverse displacement operators, tangential confluent Nevanlinna–Pick problem

AMS subject classifications. 47A57, 47B35, 47N40, 68Q25, 65F05

PII. S089547980238627X

1. Introduction.

1.1. Displacement rank approach to computations with structured matrices. Structured matrices are omnipresent in computations for communication, sciences, and engineering (see the extensive bibliographies in [KS95], [KS99], and [P01]). Figure 1.1 displays the four most popular classes of structured matrices. They are generalized to various other highly important matrix structures in the *displacement rank approach*, which we trace back to the seminal paper [KKM79]. We next follow [P01] to outline this approach, which treats various matrix structures in a unified way, based on their association with the *displacement operators*, and then focus on its most fundamental stage of the inversion of the displacement operators.

An $n \times n$ structured matrix M can be associated with an appropriate displacement operator L such that $r = \operatorname{rank}(L(M))$ is small, $r \ll n$. The image matrix $L(M)$ is called the *displacement* of M, and r is called the *displacement rank* of M. The n^2 entries of the displacement $L(M)$ can be represented via only $2rn$ parameters. Such a compressed representation of $L(M)$ can be extended to the matrix M by inverting the displacement operator.

Example 1.1 (Cauchy-like matrices; see [HR84], [GO94]). Let $D(\mathbf{s}) = \operatorname{diag}(s_i)_{i=1}^m$, $D(\mathbf{t}) = \operatorname{diag}(t_j)_{j=1}^n$, $\mathbf{s} = (s_i)_{i=1}^m$, and $\mathbf{t} = (t_j)_{j=1}^n$, where all s_i, t_j are distinct. Consider

*Received by the editors April 4, 2001; accepted for publication (in revised form) by S. Van Huffel June 28, 2002; published electronically January 23, 2003.
http://www.siam.org/journals/simax/24-3/38627.html
†Department of Mathematics and Computer Science, Lehman College of CUNY, Bronx, NY 10468 (vpan@lehman.cuny.edu).
‡Ph.D. Program in Mathematics, Graduate School of CUNY, New York, NY 10036 (xwang2@gc.cuny.edu).

Toeplitz matrices $(t_{i-j})_{i,j=1}^{m,n}$	Hankel matrices $(h_{i+j})_{i,j=1}^{m,n}$
Vandermonde matrices $\left(t_i^{j-1}\right)_{i,j=1}^{m,n}$	Cauchy matrices $\left(\frac{1}{s_i-t_j}\right)_{i,j=1}^{m,n}$

FIG. 1.1. *Four classes of structured matrices.*

the linear operator

$$L(M) = D(\mathbf{s})M - MD(\mathbf{t}).$$

Suppose that

$$L(M) = \sum_{k=1}^{l} \mathbf{g}_k \mathbf{h}_k^T = GH^T, \qquad G = (\mathbf{g}_1, \ldots, \mathbf{g}_l), \quad H = (\mathbf{h}_1, \ldots, \mathbf{h}_l). \tag{1.1}$$

It is immediately verified that (1.1) has a unique solution

$$M = \sum_{k=1}^{l} D(\mathbf{g}_k) C(\mathbf{s}, \mathbf{t}) D(\mathbf{h}_k), \tag{1.2}$$

where $C = C(\mathbf{s}, \mathbf{t}) = (\frac{1}{s_i - t_j})_{i,j=1}^{m,n}$ is the Cauchy matrix of Figure 1.1. Equation (1.2) reduces multiplication of a matrix M by a vector to multiplication of $C(\mathbf{s}, \mathbf{t})$ by l vectors. If $l \ll n$, M is called a *Cauchy-like* matrix. This class covers the important subclasses of *Loewner* and *Pick* matrices [P01, pp. 9, 96].

Similarly, the classes of Toeplitz, Hankel, and Vandermonde matrices are extended, and many other structured matrices are also compressed via (1.1) for appropriate operators L. This enables the performing of computations with the matrices M in terms of their *displacement generators* G, H by using much smaller amounts of computer memory and much less CPU time than with the general matrices as long as

(a) the ranks of the displacements are kept small and
(b) the desired output (e.g., the solution of a linear system) is easily recovered at the end.

Here is a flowchart of [P01] for this approach: COMPRESS, OPERATE, DECOMPRESS.

The COMPRESS stage consists in choosing a short displacement generator for the input matrix M (e.g., [P92], [P93], by computing the SVD of its displacement $L(M) = U\Sigma^2 V^T = GH^T$, $G = U\Sigma$, $H = V\Sigma$). Simple rules for operating with displacements at the OPERATE stage can be found in [P01, section 1.5]. They include expressions for displacements of the products, sums, linear combinations, Schur complements, and blocks of structured matrices. They also include algorithms for the recovery of a shorter generator from a longer one. These expressions and algorithms are stated for *symbolic displacement*, based on (1.1), where the operator L and matrix class M are not specified. Thus the rules and algorithms are *unified* over various classes of structured matrices. Application of these rules to some basic computations with structured matrices (such as the computation of short displacement generators for the inverses or for the bases of the null spaces) yields effective unified algorithms, which are superfast, that is, which run in $O(n \log^d n)$ time for $d \leq 3$, versus the orders of n^3 time in Gaussian elimination and n^2 time in some fast algorithms.

Furthermore, in [P90], the *displacement transformation* was proposed as a means of extending any successful algorithm available for one class of structured matrices to other classes, and sample transformations among the four classes of matrices of Hankel, Toeplitz, Vandermonde, and Cauchy types were displayed. This approach was pushed forward extensively, yielding effective practical algorithms [H95], [GKO95], [KO96], [G98], [G98a]. On the other hand, the DECOMPRESS stage never enjoyed the systematic treatment it deserves, and thus the entire approach hinged on a few ad hoc formulae scattered in [KKM79], [AG91], [GO94], and [BP94]. Particularly underdeveloped was this basic stage for the important applications using rectangular structured matrices (appearing, e.g., in structured least-squares computations) and singular displacement operators (appearing, e.g., in the study of the Nevanlinna–Pick celebrated problems). An important related issue is the estimation of $\|L^{-1}\|$, which is a critical numerical parameter. For example, whenever the solution of a linear system $M\mathbf{x} = \mathbf{b}$ is recovered from the displacement $L(M^{-1})$ computed numerically, the output errors are proportional to $\|L^{-1}\|$. In another example, a structured matrix is inverted numerically by means of Newton's iteration, and the COMPRESS stage is recursively applied in each iterative step [P92], [P01]. The convergence rate of the process and even the convergence itself critically depend on the residual norm $r_i = \|I - MX_i\|$, where X_i is an approximation to M^{-1} implicitly represented by its compressed displacement. The residual norm r_i is proportional to $\|L^{-1}\|$, and so the convergence is faster where $\|L^{-1}\|$ is smaller.

1.2. Our results and organization of the paper. In the present paper, we fill the cited void in a systematic regular way. We specify bilinear expressions of structured matrices via their displacements or, equivalently, invert the linear displacement operators, covering the most popular classes of structured matrices and *almost all known operators* L (see Remark 6.4). We treat the general case of rectangular matrices M and supply general inversion techniques for possibly singular operators L. We first invert them on the orthogonal complement of their null spaces and then extend the inversion to all matrices by using the first or the last row and/or column of M (see Examples 5.1(3), 5.4(2), and 5.6(2) and sections 6.2(ii) and 6.3(ii)). Because of the high importance of the approach, our work should inevitably have substantial practical impact on the computations with structured matrices supplying a solid foundation for them to replace the collection of random ad hoc recipes available so far. Within the limited space of this paper, we point only to some impact on the solution of the Nevanlinna–Pick and Nehari problems in Remarks 5.8 and 6.4, referring the reader to [BGR90], [OP98], [P01], and the bibliographies therein for further information on these problems and the impact. Sections 2–4 cover some simple and/or known auxiliary results. In sections 5 and 6, we derive the desired bilinear expressions. The derivation is elementary and rather straightforward in section 5 (apart from our novel treatment of the case of singular operators) but is more involved in section 6. There we invert the operators associated with a very general class of confluent matrices. The inversion of these operators is a basis for the design of effective algorithms for the confluent tangential Nevanlinna–Pick problem and was a highly important long-standing open issue. The superfast algorithms of [OP98], [OS00], and [P01] for the confluent Nevanlinna–Pick problem as well as their future amendments and improvements rely and must inevitably rely on the inversion of the associated displacement operators. In Remarks 5.9 and 6.4, we comment on the preceding works. In section 7, we briefly comment on the extension of our results to the products and inverses of structured matrices. Finally, in sections 8 and 9, we substantially advance the

known results in [P92], [P93], [PRW02], and [PKRC02] on estimating the norms of the inverse displacement operators.

2. Definitions and basic results. Let us begin with some definitions and simple basic results (cf. [P01] for a detailed and systematic exposition of structured matrix computations). We assume computations in an arbitrary field $\mathbb{F}$, which, in particular, covers computations in the fields of complex, real, or rational numbers ($\mathbb{C}$, $\mathbb{R}$, or $\mathbb{Q}$). $M \in \mathbb{F}^{m\times n}$ denotes an $m \times n$ matrix with the entries in the field $\mathbb{F}$.

- $\mathbf{t}^{-1} = (t_i^{-1})_{1\le i\le k}$, where $\mathbf{t} = (t_i)_{1\le i\le k} \in \mathbb{F}^{k\times 1}$. W^T, $\mathbf{v}^T$ are the *transposes* of a matrix W and a vector $\mathbf{v}$, respectively. W^{-T} is the transpose of W^{-1}, that is, $W^{-T} = (W^{-1})^T = (W^T)^{-1}$.
- $(W_1, \dots, W_n)$ is the $1 \times n$ block matrix with the blocks $W_1, \dots, W_n$. $D(\mathbf{v}) = \operatorname{diag}(\mathbf{v})$ is the $n \times n$ *diagonal* matrix, where $\mathbf{v} = (v_i)_{1\le i\le n}$.
- $\mathbf{e}_i$ is the ith coordinate vector, having its ith coordinate 1 and all other coordinates 0, and so $\mathbf{e}_1 = (1, 0, \dots, 0)^T$. $\mathbf{0} = (0, \dots, 0)^T$, $\mathbf{1} = (1, \dots, 1)^T$. $I = I_n = (\mathbf{e}_1, \dots, \mathbf{e}_n)$ is the $n \times n$ *identity matrix*. 0_n is the $n \times n$ *null matrix*. $J = J_n = (\mathbf{e}_n, \dots, \mathbf{e}_1)$ is the $n \times n$ *reflection matrix*.
- $Z_f = \begin{pmatrix} & f \\ I_{n-1} & \end{pmatrix} = (\mathbf{e}_2, \dots, \mathbf{e}_n, f\mathbf{e}_1)$ is the $n \times n$ *unit f-circulant* matrix. $Z = Z_0$ is the $n \times n$ unit lower triangular Toeplitz matrix. For a vector $\mathbf{v} = (v_1, \dots, v_m)^T$, write

$$Z_{f,m,n}(\mathbf{v}) = (z_{i,j})_{\substack{1\le i\le m\\ 1\le j\le n}},$$

$$z_{i,j} = \begin{cases} v_{i-j+1} & \text{if } i \ge j,\\ f^k v_{m-l} & \text{if } j-i-1 = km+l, 0 \le l \le m-1, k \ge 0.\end{cases}$$

  $Z_{f,m,n}(\mathbf{v})$ is the $m \times n$ f-circulant matrix with the first column $\mathbf{v}$. $Z_f(\mathbf{v}) = \sum_{i=1}^{m} v_i Z_f^{i-1} = Z_{f,m,m}(\mathbf{v})$. $Z(\mathbf{v}) = Z_0(\mathbf{v})$.
- $V_{m,n}(\mathbf{x}) = (x_i^{j-1})_{\substack{1\le i\le m\\ 1\le j\le n}}$ is the $m \times n$ Vandermonde matrix defined by its second column vector $\mathbf{x} = (x_i)_{1\le i\le m}$. $V(\mathbf{x}) = V_{m,m}(\mathbf{x})$.
- ω_n is a primitive nth root of 1 (that is, $\omega_n^n = 1$, $\omega_n^s \ne 1$, $s = 1, 2, \dots, n-1$); e.g., $\omega_n = e^{2\pi\sqrt{-1}/n}$ in the complex number field $\mathbb{C}$. $\mathbf{w}_n = (\omega_n^{i-1})_{1\le i\le n}$ is the vector of all nth roots of 1.
- $\Omega_n = (\omega_n^{(i-1)(j-1)})_{\substack{1\le i\le n\\ 1\le j\le n}}$ is the $n \times n$ matrix of the *discrete Fourier transform (DFT)*. The DFT of a vector $\mathbf{v}$ of dimension n is the vector $DFT(\mathbf{v}) = \Omega_n \mathbf{v}$.
- $\lceil x \rceil$ and $\lfloor x \rfloor$ denote two integers closest to a real x such that $\lfloor x \rfloor \le x \le \lceil x \rceil$.
- For any matrix A, let $\sigma_i(A)$ be its ith largest singular value if $i \le \operatorname{rank}(A)$, and let $\sigma_i(A) = 0$ if $i > \operatorname{rank}(A)$. For any $n \times n$ matrix A, let $\operatorname{spectrum}(A) = \{\lambda_1(A), \dots, \lambda_n(A)\}$ be the set of all of the eigenvalues of A. (We repeat m times any eigenvalue having algebraic multiplicity m.)

The following simple results can be easily verified.

THEOREM 2.1. *$J^2 = I$, $J\mathbf{v} = (v_{n+1-i})_{1\le i\le n}$, $JD(\mathbf{v})J = D(J\mathbf{v})$ for any vector $\mathbf{v} = (v_i)_{1\le i\le n}$.*

THEOREM 2.2. *For the $n \times n$ matrix Z_e and any scalar e, we have $Z_e^n = eI$, $Z_e^T = JZ_eJ$. For $e \ne 0$, we have $Z_e^{-1} = Z_{1/e}^T$.*

THEOREM 2.3 (see [CPW74]). *For the $n \times n$ matrix Z_e and scalar $e \ne 0$, we have $Z_e = V^{-1} \operatorname{diag}(\omega_n^i)_{i=0}^{n-1} V$, where $V = V(\mathbf{t}) = (\omega_n^{ij})_{i,j=0}^{n-1} \operatorname{diag}(t^i)_{i=0}^{n-1}$ and t is a primitive nth root of e.*

THEOREM 2.4. *$O(n\log n)$ flops are sufficient to multiply by a vector the matrices V and V^{-1} of Theorem 2.3 as well as the $n\times n$ Vandermonde matrix $V((\mathbf{t}+s\mathbf{1})^{-1})^T$ for any scalar s.*

Proof. Let $\mathbf{v}=(v_i)_i$, $\mathbf{u}=(u_k)_k$, $v(x)=\sum_{1\le i\le n} v_i x^{i-1}$. Then the vectors $V(\mathbf{u})\mathbf{v}=(v(u_k))_{k=1}^n$ for $\mathbf{u}=\mathbf{t}$, $\mathbf{u}=\mathbf{t}^{-1}$, and $\mathbf{u}=(\mathbf{t}+s\mathbf{1})^{-1}$ can be computed in $O(n\log n)$ flops [P01, p. 29], [PSD70]. ∎

The rest of this section is the basis for estimating the operator norms $\|L^{-1}\|$ in sections 8–9.

DEFINITION 2.5.

- *$\sigma_1(A)\ge\cdots\ge\sigma_r(A)>0$ are all of the singular values of a matrix A, $r=\operatorname{rank}(A)$.*
- *$\lambda_1(A),\dots,\lambda_n(A)$ are all of the eigenvalues of an $n\times n$ matrix A with $|\lambda_1(A)|\ge\cdots\ge|\lambda_n(A)|$.*
- *$m\mid n$ means that an integer m divides an integer n; $m\nmid n$ means the opposite.*
- *$\operatorname{lcm}(m,n)$ is the* least common multiple *of two positive integers m and n.*

DEFINITION 2.6 (norms of vectors, operators, and matrices).

- *For a vector $\mathbf{x}=(x_i)$, we define its (Euclidean) norm by $\|\mathbf{x}\|=(\sum_i|x_i|^2)^{1/2}$.*
- *For a linear operator L on a normed vector space V, we define the 2-norm by $\|L\|=\sup_{\mathbf{0}\ne\mathbf{x}\in V}\frac{\|L(\mathbf{x})\|}{\|\mathbf{x}\|}$.*
- *Viewing an $m\times n$ matrix $A=(a_{ij})=(\mathbf{a}_1,\dots,\mathbf{a}_n)$ as an mn-dimensional vector $\overrightarrow{A}=(\mathbf{a}_1^T,\dots,\mathbf{a}_n^T)^T$, we define its Frobenius norm $\|A\|_F=\|\overrightarrow{A}\|=(\sum_{i,j}|a_{i,j}|^2)^{1/2}$. Alternatively, we may view the matrix as a linear operator $L_A:\mathbf{x}\mapsto A\mathbf{x}$ (or $R_A:\mathbf{x}\mapsto\mathbf{x}^TA$) and define its operator norm (2-norm) $\|A\|=\|L_A\|$ (or $\|R_A\|$).*
- *Given a linear operator L on the $m\times n$ matrix space, we restrict L on the matrices A having rank of at most r and define $\|L\|_r=\sup_{\operatorname{rank}(A)\le r}\frac{\|L(A)\|}{\|A\|}$.*

Here are some simple results.

THEOREM 2.7. *For $r\ge1$ and a linear operator L on the matrix space, we have $\|L\|_{r-1}\le\|L\|_r\le r\|L\|_1$.*

Proof. (1) $\|L\|_{r-1}\le\|L\|_r$ is obvious. (2) For any matrix A, we know from its SVD that $A=A_1+\cdots+A_r$, where $r=\operatorname{rank}(A)$, each A_i is a rank-1 matrix, and $\|A_i\|=\sigma_i(A)$. Then $\|L(A)\|\le\|L(A_1)\|+\cdots+\|L(A_r)\|\le\|L\|_1(\|A_1\|+\cdots+\|A_r\|)\le r\|L\|_1\|A\|$. ∎

THEOREM 2.8. *For any matrix A, $\|A\|=\sigma_1(A)$, $\|A\|_F=(\sum_i\sigma_i(A)^2)^{1/2}$. Therefore, $\|A\|_F/\sqrt{\operatorname{rank}(A)}\le\|A\|\le\|A\|_F$. Furthermore, $\|A\|\ge|\lambda_1(A)|$ if A is a square matrix.*

Example 2.9. For $m\times m$ matrix Z_e,

$$\|Z_e^k\|=\begin{cases}|e|^{k/m} & \text{if } m\mid k,\\ |e|^{\lfloor k/m\rfloor}\max(1,|e|) & \text{if } m\nmid k.\end{cases}$$

3. Linear operators of Sylvester and Stein types. Let us associate structured matrices with *displacement linear operators of Sylvester type*, $L=\nabla_{A,B}$,

$$\nabla_{A,B}(M)=AM-MB, \tag{3.1}$$

and *Stein type*, $L=\Delta_{A,B}$,

$$\Delta_{A,B}(M)=M-AMB, \tag{3.2}$$

where A, B are two fixed *operator matrices*. The image $L(M)$ is called the *L-displacement* of a matrix M or just its *displacement*. We consider the general case of rectangular matrices A, B, and M.

Operators of both types are useful; the ∇ operators may be more effective at the OPERATE stage; the Δ operators may be more effective at the DECOMPRESS stage.

THEOREM 3.1. *$\nabla_{A,B} = A\Delta_{A^{-1},B}$ if the matrix A is nonsingular, and $\nabla_{A,B} = -\Delta_{A,B^{-1}}B$ if the matrix B is nonsingular.*

DEFINITION 3.2. *A linear operator L is* nonsingular *if the equation $L(M) = 0$ implies that $M = 0$.*

THEOREM 3.3 (cf. [P01, Theorem 4.3.2]). *$\nabla_{A,B}$ is nonsingular if and only if $\lambda_i(A) \neq \lambda_j(B)$ for all pairs of eigenvalues $(\lambda_i(A), \lambda_j(B))$; $\Delta_{A,B}$ is nonsingular if and only if $\lambda_i(A)\lambda_j(B) \neq 1$ for all pairs $(\lambda_i(A), \lambda_j(B))$.*

COROLLARY 3.4. *If the operator $\nabla_{A,B}$ is nonsingular, then A or B is nonsingular.*

Let us relate basic operations with matrices to operations with their displacements (cf. [P01]).

THEOREM 3.5. *For a nonsingular matrix M and a pair of operator matrices A and B, we have*

$$\nabla_{B,A}(M^{-1}) = -M^{-1}\nabla_{A,B}(M)M^{-1}.$$

Furthermore,

$$\Delta_{B,A}(M^{-1}) = BM^{-1}\Delta_{A,B}(M)B^{-1}M^{-1} \quad \text{if } B \text{ is nonsingular},$$
$$\Delta_{B,A}(M^{-1}) = M^{-1}A^{-1}\Delta_{A,B}(M)M^{-1}A \quad \text{if } A \text{ is nonsingular}.$$

THEOREM 3.6. *For any triple of matrices (A, B, M) of compatible sizes, we have*

$$\nabla_{A,B}(M^T) = -\nabla_{B^T,A^T}(M)^T, \qquad \Delta_{A,B}(M^T) = \Delta_{B^T,A^T}(M)^T.$$

THEOREM 3.7. *Let $\hat{A} = VAV^{-1}$, $\hat{B} = W^{-1}BW$ for some nonsingular matrices V and W. Then*

$$\nabla_{\hat{A},\hat{B}}(VMW) = V\nabla_{A,B}(M)W, \qquad \Delta_{\hat{A},\hat{B}}(VMW) = V\Delta_{A,B}(M)W.$$

4. Inversion of displacement operators. Our explicit expressions for a matrix M via its displacement rely on the next simple theorem.

THEOREM 4.1 (see [GO92], [W93], [PRW02]). *For any triple of matrices A, B, and M and for all natural numbers k, we have $M = A^kMB^k + \sum_{i=0}^{k-1} A^i\Delta_{A,B}(M)B^i$.*

By combining Theorems 3.1 and 4.1, we obtain the next result.

COROLLARY 4.2. *Given a triple of matrices A, B, and M and a natural number k, we have $M = A^{-k}MB^k + \sum_{i=0}^{k-1} A^{-i-1}\nabla_{A,B}(M)B^i$ if A is nonsingular and $M = A^kMB^{-k} - \sum_{i=0}^{k-1} A^i\nabla_{A,B}(M)B^{-i-1}$ if B is nonsingular.*

Theorem 4.1 and Corollary 4.2 enable simple expressions of a matrix M via its displacements $\Delta_{A,B}(M)$ and $\nabla_{A,B}(M)$, respectively, provided that $A^k = cI$ and/or $B^k = cI$ for a scalar c.

COROLLARY 4.3. *Under the assumptions of Theorem 4.1, we have $M(I - aB^k) = \sum_{i=0}^{k-1} A^i\Delta_{A,B}(M)B^i$ if $A^k = aI$ and $(I - bA^k)M = \sum_{i=0}^{k-1} A^i\Delta_{A,B}(M)B^i$ if $B^k = bI$.*

DEFINITION 4.4. *Hereafter,* $W = (\mathbf{w}_1, \dots, \mathbf{w}_s)$ *is a matrix with columns given by vectors* $\mathbf{w}_1, \dots, \mathbf{w}_s$. *Suppose, for an* $m \times n$ *matrix* M *and a linear operator* L, *that we have*

$$L(M) = GH^T = \sum_{k=1}^{l} \mathbf{g}_k \mathbf{h}_k^T, \tag{4.1}$$

$G = (\mathbf{g}_1, \dots, \mathbf{g}_l)$, $H = (\mathbf{h}_1, \dots, \mathbf{h}_l)$, *and* l *is "small"* ($l = O(1)$ *or* $l \ll \min(m, n)$). *Then* M *is called a* structured matrix with an L-generator (G, H) of length l.

DEFINITION 4.5. *For natural numbers* m *and* n, *an* $m \times m$ *matrix* P, *and an* m*-dimensional column vector* $\mathbf{v}$, *we define the* $m \times n$ Krylov matrix $K_{m,n}(P, \mathbf{v}) = (\mathbf{v}, P\mathbf{v}, \dots, P^{n-1}\mathbf{v})$.

Remark 4.6. The Krylov matrix $K_{m,n}(P, \mathbf{v})$ turns into

(a) the $m \times n$ f-circulant matrix $Z_{f,m,n}(\mathbf{v})$ when $P = Z_f$,
(b) $JZ_{f,m,n}(J\mathbf{v})$ when $P = Z_f^T$, and
(c) the product $D(\mathbf{v})V_{m,n}(P\mathbf{1})$ of the diagonal matrix $D(\mathbf{v})$ and the Vandermonde matrix $V_{m,n}(P\mathbf{1})$ when P is a diagonal matrix; $D(\mathbf{v}) = I_m$ when $\mathbf{v} = \mathbf{1}$.

Theorem 4.1 and Corollary 4.2 imply the next results.

THEOREM 4.7. *For an operator* $L = \Delta_{A,B}$, *an* $m \times n$ *matrix* M *satisfying* (4.1), *and all natural numbers* k, *we have* $M = A^k M B^k + \sum_{j=1}^{l} K_{m,k}(A, \mathbf{g}_j) K_{n,k}(B^T, \mathbf{h}_j)^T$.

THEOREM 4.8. *For an operator* $L = \nabla_{A,B}$, *an* $m \times n$ *matrix* M *satisfying* (4.1), *and all natural numbers* k, *we have* $M = A^{-k-1} M B^k + A^{-1} \sum_{j=1}^{l} K_{m,k}(A^{-1}, \mathbf{g}_j) \cdot K_{n,k}(B^T, \mathbf{h}_j)^T$ *if* A *is nonsingular and* $M = A^k M B^{-k-1} - \sum_{j=1}^{l} K_{m,k}(A, \mathbf{g}_j) \cdot K_{n,k}(B^{-T}, \mathbf{h}_j)^T B^{-1}$ *if* B *is nonsingular.*

5. Bilinear expressions via generators for fundamental matrix structures. In this section, we extend (1.2) to express a matrix $M \in \mathbb{F}^{m \times n}$ via its displacement $L(M)$, where $L = \Delta_{A,B}$ and $L = \nabla_{A,B}$ for some commonly used operator matrices $A \in \mathbb{F}^{m \times m}$, $B \in \mathbb{F}^{n \times n}$. Let $L(M) = GH^T = \sum_{1 \le j \le l} \mathbf{g}_j \mathbf{h}_j^T$, $G = (g_{i,j})_{1 \le i \le m, 1 \le j \le l}$, $H = (h_{i,j})_{1 \le i \le n, 1 \le j \le l}$.

Example 5.1. The Stein-type operators $L = \Delta_{Z_e, Z_f}$ are associated with Hankel-like matrices. Note that $Z_e^m = eI_m$, $Z_f^n = fI_n$. We begin with the special case in which $e = 0$ (the same tool applies to $f = 0$), then supply the expressions in the general nonsingular case, and finally cover all choices of e and f.

1. $e = 0$. Apply Theorem 4.7, take into account Remark 4.6, and obtain that

$$M = \sum_{j=1}^{l} K_{m,m}(Z, \mathbf{g}_j) K_{n,m}(Z_f^T, \mathbf{h}_j)^T = \sum_{j=1}^{l} Z(\mathbf{g}_j) Z_{f,n,m}(J\mathbf{h}_j)^T J.$$

2. Let the operator Δ_{Z_e, Z_f} be nonsingular. Then the matrix $I_n - eZ_f^m$ is nonsingular due to Definition 3.2. Apply Theorem 4.7 for $k = m$, then recall Remark 4.6 and obtain that

$$\begin{aligned} M &= \sum_{j=1}^{l} K_{m,m}(Z_e, \mathbf{g}_j) K_{n,m}(Z_f^T, \mathbf{h}_j)^T (I_n - eZ_f^m)^{-1} \\ &= \sum_{j=1}^{l} Z_e(\mathbf{g}_j) Z_{f,n,m}(J\mathbf{h}_j)^T J (I_n - eZ_f^m)^{-1}. \end{aligned}$$

3. If the operator Δ_{Z_e,Z_f} is singular, then we cannot recover the matrix M solely from its displacement. We need extra information about M. We begin with the matrix equation

$$\Delta_{Z_e,Z}(M) = \Delta_{Z_e,Z_f}(M) + Z_e M \begin{pmatrix} & f \\ 0_{n-1} & \end{pmatrix} = GH^T + fZ_e M \mathbf{e}_1 \mathbf{e}_n^T,$$

apply Theorem 4.7 to the operator $\Delta_{Z_e,Z}$, recall Remark 4.6, and deduce that

$$\begin{aligned} M &= \sum_{j=1}^{l} K_{m,n}(Z_e, \mathbf{g}_j) K_{n,n}(Z^T, \mathbf{h}_j)^T + fK_{m,n}(Z_e, Z_e M \mathbf{e}_1) K_{n,n}(Z^T, \mathbf{e}_n)^T \\ &= \sum_{j=1}^{l} Z_{e,m,n}(\mathbf{g}_j) Z(J\mathbf{h}_j)^T J + fZ_{e,m,n}(JZ_e M \mathbf{e}_1) J, \end{aligned}$$

where $(M_{i,1})_{1 \le i \le m} = M\mathbf{e}_1$ is the first column of M.

Remark 5.2. In case 2, we involve the matrix $(I_n - eZ_f^m)^{-1} = V_f^{-1} \operatorname{diag}(\frac{1}{1-et_i^m})_{1 \le i \le n} V_f$, where $V_f = (t_i^{j-1})_{1 \le i \le n, 1 \le j \le n}$, $t_1, \dots, t_n$ are all the nth roots of f.

Example 5.3. The case of the Stein-type operators $L = \Delta_{Z_e,Z_f^T}$, $L = \Delta_{Z_e^T,Z_f}$, and $L = \Delta_{Z_e^T,Z_f^T}$, associated with Toeplitz/Hankel-like matrices, can be reduced to Example 5.1 based on Theorem 2.2.

Example 5.4. The Sylvester-type operators $L = \nabla_{Z_e,Z_f}$ are associated with Toeplitz-like matrices.

1. If $e \neq 0$, then, by Theorems 3.1 and 2.2, we have

$$\Delta_{Z_{1/e}^T,Z_f}(M) = Z_{1/e}^T \nabla_{Z_e,Z_f}(M) = (Z_{1/e}^T G) H^T.$$

The latter equation immediately reduces the problem to the case of the Stein-type operators $\Delta_{Z_e^T,Z_f}$ of Example 5.3. The same tool applies to the case $f \neq 0$.

2. If $e = f = 0$, then we have (cf. [BP93], [BP94] for the proof)

$$M = J \sum_{j=1}^{l} Z(JZ^T \mathbf{g}_j) Z_{0,n,m}(J\mathbf{h}_j)^T J + JZ_{0,n,m}(JM^T \mathbf{e}_m)^T J.$$

Example 5.5. Similarly to Example 5.4, we express Hankel-like and Toeplitz-like matrices M associated with the Sylvester-type operators $L = \nabla_{Z_e,Z_f^T}$, $L = \nabla_{Z_e^T,Z_f}$, and $L = \nabla_{Z_e^T,Z_f^T}$.

Example 5.6. The Stein-type operators $L = \Delta_{D(\mathbf{v}),Z_f^T}$ are associated with the matrix structure of Vandermonde type.

1. If the operator $\Delta_{D(\mathbf{v}),Z_f^T}$ is nonsingular, then the matrix $I_m - fD(\mathbf{v})^n$ is nonsingular, and it follows from Theorem 4.7 and Remark 4.6 that

$$\begin{aligned} M &= (I_m - fD(\mathbf{v})^n)^{-1} \sum_{j=1}^{l} K_{m,n}(D(\mathbf{v}), \mathbf{g}_j) K_{n,n}(Z_f, \mathbf{h}_j)^T \\ &= \operatorname{diag}\left(\frac{1}{1 - fv_i^n}\right)_{1 \le i \le m} \sum_{j=1}^{l} D(\mathbf{g}_j) V_{m,n}(\mathbf{v}) Z_f(\mathbf{h}_j)^T. \end{aligned}$$

2. To relax the nonsingularity assumption, observe that

$$\Delta_{D(\mathbf{v}),Z^T}(M) = \Delta_{D(\mathbf{v}),Z_f^T}(M) + D(\mathbf{v})M \begin{pmatrix} 0_{n-1} \\ f \end{pmatrix} = GH^T + fD(\mathbf{v})M\mathbf{e}_n\mathbf{e}_1^T.$$

Therefore, we deduce from Theorem 4.7 and Remark 4.6 that

$$\begin{aligned} M &= \sum_{j=1}^{l} K_{m,n}(D(\mathbf{v}), \mathbf{g}_j)K_{n,n}(Z, \mathbf{h}_j)^T + fD(\mathbf{v})K_{m,n}(D(\mathbf{v}), M\mathbf{e}_n) \\ &= \sum_{j=1}^{l} D(\mathbf{g}_j)V_{m,n}(\mathbf{v})Z(\mathbf{h}_j)^T + fD(\mathbf{v})D(M\mathbf{e}_n)V_{m,n}(\mathbf{v}). \end{aligned}$$

Example 5.7. The Stein-type operators $L = \Delta_{D(\mathbf{v}),Z_f}$, $L = \Delta_{Z_e,D(\mathbf{v})}$, and $L = \Delta_{Z_e^T,D(\mathbf{v})}$ and the Sylvester-type operators $L = \nabla_{D(\mathbf{v}),Z_f}$, $L = \nabla_{D(\mathbf{v}),Z_f^T}$, $L = \nabla_{Z_e,D(\mathbf{v})}$, and $L = \nabla_{Z_e^T,D(\mathbf{v})}$ are also associated with structured matrices of Vandermonde type. Theorems 2.2, 3.1, 3.5, and 3.6 enable us to reduce the problems to the case of the operator $\Delta_{D(\mathbf{v}),Z_f^T}$ of Example 5.6.

Remark 5.8. In [P01], a variant of the displacement transformation techniques of [P90] reduced the solution of the tangential Nevanlinna–Pick and matrix Nehari problems to computations with *skew-Hankel-like* matrices associated with the operator $\nabla_{Z_e^T,-Z_f}$ for a pair of scalars e and f, $e \neq f$. Such matrices have no other natural derivation; their hidden structure appears to be hard to exploit, but the inversion formulae of Example 5.3 are immediately extended. More generally, *all presented expressions for the inversion of the operators* $\nabla_{A,B}$ *and* $\Delta_{A,B}$ *can be immediately extended to operators* $\nabla_{aA,bB}$ *and* $\Delta_{aA,bB}$ *for any pair of nonzero scalars* a *and* b. With this extension, the unified divide-and-conquer algorithms of [P99], [P00], and [P01] solve the cited important problems by using $O(N\log^2 N)$ flops (N being the input size) versus $O(N\log^3 N)$ announced in the proceedings paper [OP98] and the order of N^2 in the preceding works. Important superfast algorithms using $O(N\log^d N)$ flops are known for other rational interpolation problems [BVB97], [CJL96], [CJL96a], [KVB99], but different techniques are required for the Nevanlinna–Pick and Nehari problems.

Remark 5.9. The bilinear expressions of this section have been known for square matrices M and the following operators L: $\Delta_{Z^T,Z}$ and Δ_{Z,Z^T} [KKM79], $\nabla_{D(\mathbf{v}),Z^T}$ [GKKL87], $\nabla_{Z_1,Z_{-1}}$ [AG91], $\nabla_{Z,Z}, \nabla_{Z^T,Z^T}, \Delta_{Z,Z}$, and Δ_{Z^T,Z^T} [BP93], [BP94], and $\Delta_{Z_e,Z_{1/e}^T}$ and $\Delta_{D(\mathbf{v}),Z_{1/e}^T}$ for $e \neq 0$ [GO94]. Some other expressions of this section have been announced in [P01] (see the notes to section 4.4 therein) although only for square matrices M and nonsingular operators L.

6. Bilinear expressions via generators for confluent matrices.

Example 6.1. More general operator matrices $A = \lambda I_m + Z_e$ and $B = \mu I_n + Z_f$ (see Remark 6.3) cover the confluent matrices involved in the tangential confluent Nevanlinna–Pick problem [BGR90], [GO94a], [OS00], [P01]. Then $L = \nabla_{A,B} = \nabla_{(\lambda-\mu)I_m+Z_e,Z_f} = \nabla_{Z_e,(\mu-\lambda)I_n+Z_f}$, and we partition our study depending on the values e, f, and $\lambda - \mu$.

6.1. Case 1. $e = f = 0$. Assume that $\lambda \neq \mu$. Otherwise, see part 2 of Example 5.4. Combine Corollary 4.2, Remark 4.6, and the equation

$$\nabla_{A,B}(M) = ((\lambda - \mu)I_m + Z)M - MZ = GH^T$$

to obtain

$$M = \sum_{j=0}^{n-1}((\lambda-\mu)I_m + Z)^{-j-1}GH^TZ^j = \sum_{j=0}^{n-1}\sum_{i=0}^{m-1}\binom{-j-1}{i}(\lambda-\mu)^{-j-1-i}Z^iGH^TZ^j$$
$$= \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\frac{(-1)^i(i+j)!}{i!(\lambda-\mu)^{i+j+1}j!}Z^iGH^TZ^j = \sum_{k=1}^{l}K_{m,m}(Z,\mathbf{g}_k)\Theta_0(\lambda-\mu)K_{n,n}(Z^T,\mathbf{h}_k)^T$$
$$= \sum_{k=1}^{l}Z(\mathbf{g}_k)\Theta_0(\lambda-\mu)Z(J\mathbf{h}_k)^TJ,$$

$$\Theta_0(s) = \left(\frac{(-1)^{i-1}(i+j-2)!}{(i-1)!s^{i+j-1}(j-1)!}\right)_{i,j=1}^{m,n} = \operatorname{diag}\left(\frac{(-1)^{i-1}}{(i-1)!}\right)_{i=1}^{m} H \operatorname{diag}\left(\frac{1}{(j-1)!}\right)_{j=1}^{n},$$
$$H = \left(\frac{(i+j-2)!}{s^{i+j-1}}\right)_{\substack{1\le i\le m\\ 1\le j\le n}},$$

which is a Hankel matrix.

6.2. Case 2. $e \neq 0$, $f = 0$ (similarly if $e = 0$, $f \neq 0$).
(i) $(\mu-\lambda)^m \neq e$. Write $V = V(\mathbf{t})$, and combine the equation

$$\nabla_{A,B}(M) = ((\lambda-\mu)I_m + Z_e)M - MZ = GH^T$$

with Theorem 2.3, Corollary 4.2, and Remark 4.6 to obtain

$$M = \sum_{j=0}^{n-1}((\lambda-\mu)I_m + Z_e)^{-j-1}GH^TZ^j = \sum_{j=0}^{n-1}V^{-1}((\lambda-\mu)I_m + D)^{-j-1}VGH^TZ^j$$
$$= \sum_{k=1}^{l}V^{-1}((\lambda-\mu)I_m + D)^{-1}K_{m,n}(((\lambda-\mu)I_m + D)^{-1}, V\mathbf{g}_k)K_{n,n}(Z^T,\mathbf{h}_k)^T$$
$$= \sum_{k=1}^{l}V^{-1}\operatorname{diag}(V\mathbf{g}_k)\left((\tfrac{1}{\lambda-\mu+t_i})^j\right)_{\substack{1\le i\le m\\ 1\le j\le n}}K_{n,n}(Z^T,\mathbf{h}_k)^T$$
$$= \sum_{k=1}^{l}\left(V^{-1}\sum_{r=1}^{m}g_{r,k}D^{r-1}\right)\left((\tfrac{1}{\lambda-\mu+t_i})^j\right)_{\substack{1\le i\le m\\ 1\le j\le n}}K_{n,n}(Z^T,\mathbf{h}_k)^T$$
$$= \sum_{k=1}^{l}\left(\sum_{r=1}^{m}g_{r,k}Z_e^{r-1}V^{-1}\right)\left((\tfrac{1}{\lambda-\mu+t_i})^j\right)_{\substack{1\le i\le m\\ 1\le j\le n}}K_{n,n}(Z^T,\mathbf{h}_k)^T$$
$$= \sum_{k=1}^{l}K_{m,m}(Z_e,\mathbf{g}_k)V^{-1}\left((\tfrac{1}{\lambda-\mu+t_i})^j\right)_{\substack{1\le i\le m\\ 1\le j\le n}}K_{n,n}(Z^T,\mathbf{h}_k)^T$$
$$= \sum_{k=1}^{l}K_{m,m}(Z_e,\mathbf{g}_k)\Theta_1(\lambda-\mu)K_{n,n}(Z^T,\mathbf{h}_k)^T = \sum_{k=1}^{l}Z_e(\mathbf{g}_k)\Theta_1(\lambda-\mu)Z(J\mathbf{h}_k)^TJ.$$

Here $\Theta_1(s) = V^{-1}((\frac{1}{s+t_i})^j)_{\substack{1\le i\le m\\ 1\le j\le n}} = \frac{1}{m}V(\mathbf{t}^{-1})^TV_{m,n}((\mathbf{t}+s\mathbf{1})^{-1})D(\mathbf{t}+s\mathbf{1})^{-1}$, $\mathbf{t} = (t_i)_{1\le i\le m}$ is the vector of all the mth roots of e.

(ii) $(\mu-\lambda)^m = e$, so the operator L is singular. Note that

$$((\lambda-\mu)I_m + Z)M - MZ = \nabla_{A,B}(M) + (Z - Z_e)M = GH^T - e\mathbf{e}_1\mathbf{e}_m^T M.$$

Proceed similarly to Case 1 to obtain that

$$\begin{aligned} M &= \sum_{k=1}^{l} K_{m,m}(Z, \mathbf{g}_k)\Theta_0(\lambda-\mu)K_{n,n}(Z^T, \mathbf{h}_k)^T \ - \ e\Theta_0(\lambda-\mu)K_{n,n}(Z^T, M^T\mathbf{e}_m)^T \\ &= \sum_{k=1}^{l} Z(\mathbf{g}_k)\Theta_0(\lambda-\mu)Z(J\mathbf{h}_k)^T J \ - \ e\Theta_0(\lambda-\mu)Z(JM^T\mathbf{e}_m)^T J. \end{aligned}$$

6.3. Case 3. $ef \neq 0$.

(i) The operator L is nonsingular, so both matrices $I - f((\lambda-\mu)I_m + Z_e)^n$ and $I - e((\mu-\lambda)I_n + Z_f)^m$ are nonsingular. Apply Corollary 4.2 for $k = m$ and obtain that

$$M = Me((\mu-\lambda)I_n + Z_f)^m + \sum_{i=0}^{m-1} Z_e^{-i-1}GH^T((\mu-\lambda)I_n + Z_f)^i.$$

Therefore, we have

$$\begin{aligned} M &= \left(\sum_{i=0}^{m-1} Z_e^{-i-1}GH^T((\mu-\lambda)I_n + Z_f)^i\right)(I_n - e((\mu-\lambda)I_n + Z_f)^m)^{-1} \\ &= \left(\sum_{i=0}^{m-1} Z_e^{-i-1}GH^T \sum_{j=0}^{n-1}\binom{i}{j}(\mu-\lambda)^{i-j}Z_f^j\right)(I_n - e((\mu-\lambda)I_n + Z_f)^m)^{-1} \\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{n}\binom{i-1}{j-1}(\mu-\lambda)^{i-j}Z_e^{-i}GH^T Z_f^{j-1}\right)(I_n - e((\mu-\lambda)I_n + Z_f)^m)^{-1} \\ &= \left(\sum_{k=1}^{l} K_{m,m}(Z_e^{-1}, Z_e^{-1}\mathbf{g}_k)\Theta_2(\mu-\lambda)K_{n,m}(Z_f^T, \mathbf{h}_k)^T\right)(I_n - e((\mu-\lambda)I_n + Z_f)^m)^{-1}, \end{aligned}$$

where $\Theta_2(s) = \left(\frac{(i-1)!s^{i-j}}{(j-1)!(i-j)!}\right)_{\substack{1\le i\le m\\ 1\le j\le i}}$ is an $m \times m$ lower triangular matrix, $\Theta_2(s) = \operatorname{diag}((i-1)!)_{1\le i\le m}\left(\frac{s^{i-j}}{(i-j)!}\right)_{\substack{1\le i\le m\\ 1\le j\le i}}\operatorname{diag}\left(\frac{1}{(j-1)!}\right)_{1\le j\le m}$. Recall the equation $Z_e^{-1} = Z_{1/e}^T$ of Theorem 2.2, recall Remark 4.6, and rewrite this expression as follows:

$$M = \left(\sum_{k=1}^{l} JZ_{1/e}(JZ_e^{-1}\mathbf{g}_k)\Theta_3(\mu-\lambda)Z_{f,n,m}(J\mathbf{h}_k)^T J\right)(I_n - e((\mu-\lambda)I_n + Z_f)^m)^{-1}.$$

(ii) The operator L is singular. For any 4-tuple (λ, μ, e, f), apply the equation

$$(\lambda I_m + Z_e)M - M(\mu I_n + Z) = \nabla_{A,B}(M) + M(Z_f - Z) = GH^T + fM\mathbf{e}_1\mathbf{e}_n^T,$$

where $Z^n = 0$, and, as in Case 2, deduce from Theorem 4.7 and Remark 4.6 that

$$\begin{aligned} M &= \sum_{k=1}^{l} K_{m,m}(Z_e, \mathbf{g}_k)\Theta_1(\lambda-\mu)K_{n,n}(Z^T, \mathbf{h}_k)^T + fK_{m,m}(Z_e, M\mathbf{e}_1)\Theta_1(\lambda-\mu)J \\ &= \sum_{k=1}^{l} Z_e(\mathbf{g}_k)\Theta_1(\lambda-\mu)Z(J\mathbf{h}_k)^T J + fZ_e(M\mathbf{e}_1)\Theta_1(\lambda-\mu)J. \end{aligned}$$

6.4. Further remarks.

Remark 6.2. For $ef \neq 0$, we have (cf. Theorem 2.3)

$$(I_m - f((\lambda-\mu)I_m + Z_e)^{-n})^{-1} = V_e^{-1} \operatorname{diag}\left(\frac{(\lambda-\mu+s_i)^n}{(\lambda-\mu+s_i)^n - f}\right)_{1\le i\le m} V_e,$$
$$(I_n - e((\mu-\lambda)I_n + Z_f)^{-m})^{-1} = V_f^{-1} \operatorname{diag}\left(\frac{(\mu-\lambda+t_i)^m}{(\mu-\lambda+t_i)^m - e}\right)_{1\le i\le n} V_f,$$

where $V_e = (s_i^{j-1})_{\substack{1\le i\le m\\ 1\le j\le m}}$, $V_f = (t_i^{j-1})_{\substack{1\le i\le n\\ 1\le j\le n}}$, $s_1, \dots, s_m$, are all mth roots of e, and $t_1, \dots, t_n$ are all nth roots of f.

Remark 6.3. To invert the Sylvester-type operators $L = \nabla_{\lambda I_m + Z_e, \mu I_n + Z_f^T}$, $L = \nabla_{\lambda I_m + Z_e^T, \mu I_n + Z_f}$, and $L = \nabla_{\lambda I_m + Z_e^T, \mu I_n + Z_f^T}$, combine Theorem 2.2, Example 6.1, and the equations

$$\nabla_{\lambda I_m + Z_e, \mu I_n + Z_f}(MJ) = \nabla_{\lambda I_m + Z_e, \mu I_n + Z_f^T}(M)J = G(JH)^T,$$
$$\nabla_{\lambda I_m + Z_e, \mu I_n + Z_f}(JM) = J\nabla_{\lambda I_m + Z_e^T, \mu I_n + Z_f}(M) = (JG)H^T,$$
$$\nabla_{\lambda I_m + Z_e, \mu I_n + Z_f}(JMJ) = J\nabla_{\lambda I_m + Z_e^T, \mu I_n + Z_f^T}(M)J = (JG)(JH)^T.$$

Remark 6.4. For a Sylvester-type operator $L = \nabla_{A,B}$ for any pair of A and B, we have $PAP^{-1} = \operatorname{diag}(\lambda_i(A)I_{m_i} + Z)_{1\le i\le p}$, $QBQ^{-1} = \operatorname{diag}(\lambda_j(B)I_{n_j} + Z)_{1\le j\le q}$. Let us express a matrix M via its displacement $L(M) = GH^T$, the matrices P and Q, and the Jordan blocks $A_i = \lambda_i(A)I_{m_i} + Z$, $i = 1, \dots, p$; $B_j = \lambda_j(B)I_{n_j} + Z$, $j = 1, \dots, q$, of the operator matrices A and B. (Already for $P = I_m$, $Q = I_n$, this covers the general class of confluent matrices associated with the tangential confluent Nevanlinna–Pick problem [BGR90].) We recover the matrix M from its displacement $L(M) = GH^T$ by applying the following steps:

1. Represent the matrix PMQ^{-1} as a $p \times q$ block matrix with blocks $M_{i,j}$ of size $m_i \times n_j$; represent the matrix PG as a $p \times 1$ block matrix with blocks G_i of size $m_i \times l$; represent the matrix H^TQ^{-1} as a $1 \times q$ block matrix with blocks H_j^T of size $l \times n_j$.
2. Replace the matrix equation $\nabla_{A,B}(M) = GH^T$ by the block equations $\nabla_{A_i,B_j}(M_{i,j}) = G_iH_j^T$ for all pairs (i,j), $i = 1, \dots, p$; $j = 1, \dots, q$.
3. Express the blocks $M_{i,j}$ from their displacement generators (G_i, H_j) as in Example 6.1.
4. Express the matrix $M = P^{-1}(M_{i,j})_{\substack{1\le i\le p\\ 1\le j\le q}}Q$.

For $P = I_m$, $Q = I_n$, we arrive at the matrices M defining the tangential confluent Nevanlinna–Pick problem. In this case, extensively studied since [BGR90], stages 1 and 4 are trivialized. In Case 1 of Example 6.1, a distinct expression for M via $L(M)$ is stated in [OS00]. With omitted proofs and restricted to the case of square matrices M, some of our results were announced in [P01] with reference to the present paper (see the notes of section 4.4 therein).

7. Two implications.

(a) The basic structured matrices can be multiplied by vectors in nearly linear time (see [P01]). Our bilinear expressions of structured matrices via their generators enable immediate extension of these algorithms to more general classes of structured matrices. In particular, we multiply the $n \times n$ matrices of Examples 5.1, 5.3–5.5,

and 6.1 by a vector by using $O(ln \log n)$ flops. Similarly, we yield the cost bound of $O(ln \log^2 n)$ flops for the $n \times n$ matrices of Examples 5.6 and 5.7.

(b) Theorem 3.5 enables the extension of all of our expressions to the inverse matrix M^{-1} via the products of this matrix with the $2l$ vectors $\mathbf{g}_k$ and $\mathbf{h}_k$, $k = 1, \dots, l$.

8. Lower and upper bounds on the norms of the inverse displacement operators. In this section, we estimate the operator norm $\|L^{-1}\|$ for the operators $L = \Delta_{Z_e,Z_f}$, $L = \nabla_{Z_e,Z_f}$, $L = \Delta_{Z_e,D(\mathbf{v})}$, $L = \nabla_{Z_e,D(\mathbf{v})}$, and $L = \Delta_{D(\mathbf{u}),D(\mathbf{v})}$, $L = \nabla_{D(\mathbf{u}),D(\mathbf{v})}$. All of our proofs and estimates, however, are invariant to interchanging the operator matrices A and B and to transposing any of A and B, so the same estimates are immediately extended to the operators $\Delta_{Z_e^T,Z_f}$, $\nabla_{Z_e^T,Z_f}$, Δ_{Z_e,Z_f^T}, ∇_{Z_e,Z_f^T}, $\Delta_{Z_e^T,Z_f^T}$, $\nabla_{Z_e^T,Z_f^T}$, $\Delta_{Z_e^T,D(\mathbf{v})}$, $\nabla_{Z_e^T,D(\mathbf{v})}$, $\Delta_{D(\mathbf{v}),Z_e}$, $\nabla_{D(\mathbf{v}),Z_e}$, $\Delta_{D(\mathbf{v}),Z_e^T}$, $\nabla_{D(\mathbf{v}),Z_e^T}$, respectively. This covers the operators associated with the matrices of the most popular structures of Toeplitz, Hankel, Vandermonde, and Cauchy types.

THEOREM 8.1. *For any operator norm and any positive integer r, we have*

$$\max_{i,j} |1 - \lambda_i(A)\lambda_j(B)|^{-1} \le \|\Delta_{A,B}^{-1}\|_r \le \sqrt{r}\|(I - B^T \otimes A)^{-1}\|, \tag{8.1}$$

$$\max_{i,j} |\lambda_i(A) - \lambda_j(B)|^{-1} \le \|\nabla_{A,B}^{-1}\|_r \le \sqrt{r}\|(I \otimes A - B^T \otimes I)^{-1}\|, \tag{8.2}$$

where $\otimes$ is the Kronecker product and the lower bounds on $\|\Delta_{A,B}^{-1}\|_r$ and $\|\nabla_{A,B}^{-1}\|_r$ apply to any operator norm.

Proof. (1) Let $\mathbf{g}$ and $\mathbf{h}$ be two eigenvectors of A and B, respectively, such that $A\mathbf{g} = \lambda_i(A)\mathbf{g}$, $B^T\mathbf{h} = \lambda_j(B)\mathbf{h}$. Let $M = \mathbf{g}\mathbf{h}^T$. Then we have $\Delta_{A,B}(M) = (1 - \lambda_i(A)\lambda_j(B))M$, $\nabla_{A,B}(M) = (\lambda_i(A) - \lambda_j(B))M$; that is, M is an eigenvector of $\Delta_{A,B}$ and $\nabla_{A,B}$. This proves the lower bounds in (8.1) and (8.2).

(2) Recall that $\overrightarrow{\Delta_{A,B}(M)} = (I - B^T \otimes A)\overrightarrow{M}$, $\overrightarrow{\nabla_{A,B}(M)} = (I \otimes A - B^T \otimes I)\overrightarrow{M}$. By Theorem 2.8, $\|M\| \le \|M\|_F = \|\overrightarrow{M}\|$, $\|\Delta_{A,B}(M)\| \ge \|\Delta_{A,B}(M)\|_F/\sqrt{r} = \|\overrightarrow{\Delta_{A,B}(M)}\|/\sqrt{r}$, $\|\nabla_{A,B}(M)\| \ge \|\nabla_{A,B}(M)\|_F/\sqrt{r} = \|\overrightarrow{\nabla_{A,B}(M)}\|/\sqrt{r}$. This proves the upper bounds in (8.1) and (8.2). ∎

Our next upper bounds on $\|L^{-1}\|$ rely on the bilinear expressions for M implied by Example 2.9 and Corollary 4.3. Write $\hat{e} = \max(1, |e|)$, $\hat{f} = \max(1, |f|)$.

THEOREM 8.2. *Let $\ell = \mathrm{lcm}(m, n)$. We have*

$$\|\Delta_{Z_e,Z_f}^{-1}\| \le \frac{\hat{e}\hat{f}}{|1 - e^{\ell/m}f^{\ell/n}|} \sum_{k=0}^{\ell-1} |e|^{\lfloor k/m \rfloor} |f|^{\lfloor k/n \rfloor}, \tag{8.3}$$

$$\|\nabla_{Z_e,Z_f}^{-1}\| \le \frac{\hat{e}\hat{f}}{|e^{\ell/m} - f^{\ell/n}|} \sum_{k=0}^{\ell-1} |e|^{\lfloor (\ell-1-k)/m \rfloor} |f|^{\lfloor k/n \rfloor}. \tag{8.4}$$

Proof. (1) Let $\Delta = \Delta_{Z_e,Z_f}(M)$. Then $M = \frac{1}{1-e^{\ell/m}f^{\ell/n}} \sum_{k=0}^{\ell-1} Z_e^k \Delta Z_f^k$. So $\|M\| \le \frac{1}{|1-e^{\ell/m}f^{\ell/n}|} \sum_{k=0}^{\ell-1} \|Z_e^k\| \, \|\Delta\| \, \|Z_f^k\| \le \frac{\|\Delta\|\hat{e}\hat{f}}{|1-e^{\ell/m}f^{\ell/n}|} \sum_{k=0}^{\ell-1} |e|^{\lfloor k/m \rfloor} |f|^{\lfloor k/n \rfloor}$.

(2) Let $\nabla = \nabla_{Z_e,Z_f}(M)$. We may assume $e \ne 0$; otherwise, $\|\nabla_{Z_0,Z_f}^{-1}\| = \lim_{e \to 0} \|\nabla_{Z_e,Z_f}^{-1}\|$. Then $M = \frac{1}{1-f^{\ell/n}/e^{\ell/m}} \sum_{k=0}^{\ell-1} Z_e^{-1-k} \nabla Z_f^k$. So $\|M\| \le \frac{|e|^{\ell/m}}{|e^{\ell/m} - f^{\ell/n}|} \cdot \sum_{k=0}^{\ell-1} \|Z_e^{-1-k}\| \, \|\nabla\| \, \|Z_f^k\| \le \frac{\|\nabla\|\hat{e}\hat{f}}{|e^{\ell/m} - f^{\ell/n}|} \sum_{k=0}^{\ell-1} |e|^{\lfloor (\ell-1-k)/m \rfloor} |f|^{\lfloor k/n \rfloor}$. ∎

Theorems 8.1 (for $A = Z_e$, $B = Z_f$) and 8.2 together imply the following corollary.

COROLLARY 8.3. *Let $m = n$, so $\operatorname{lcm}(m,n) = n$. Then*

$$|1-|ef|^{1/n}\omega_{2n}|^{-1} \le \|\Delta_{Z_e,Z_f}^{-1}\| \le \frac{n\hat{e}\hat{f}}{|1-ef|},$$
$$||e|^{1/n}-|f|^{1/n}\omega_{2n}|^{-1} \le \|\nabla_{Z_e,Z_f}^{-1}\| \le \frac{n\hat{e}\hat{f}}{|e-f|}.$$

Remark 8.4. Comparing the latter lower and upper bounds as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \frac{|1-ef|}{n\hat{e}\hat{f}\left|1-|ef|^{1/n}\omega_{2n}\right|} = \begin{cases} \frac{|1-ef|}{\hat{e}\hat{f}\sqrt{\pi^2+\ln^2|ef|}} & \text{if } ef \neq 0, \\ 0 & \text{if } ef = 0, \end{cases}$$
$$\lim_{n\to\infty} \frac{|e-f|}{n\hat{e}\hat{f}\left||e|^{1/n}-|f|^{1/n}\omega_{2n}\right|} = \begin{cases} \frac{|e-f|}{\hat{e}\hat{f}\sqrt{\pi^2+\ln^2|e/f|}} & \text{if } ef \neq 0, \\ 0 & \text{if } ef = 0. \end{cases}$$

Here ln denotes the natural logarithm. Our estimates of Corollary 8.3 are asymptotically tight as $n \to \infty$; that is, the lower and upper bounds differ by a nonzero constant factor, provided $ef \neq 0$ and either $ef \neq 1$ (for Δ_{Z_e,Z_f}) or $e \neq f$ (for ∇_{Z_e,Z_f}).

Let us improve our lower bounds by sampling matrices in the case in which $f = 0$ for any e (similarly, where $e = 0$ for any f). Let $M = Z_e(\mathbf{1})JZ(\mathbf{1})$. Then $\Delta_{Z_e,Z}(M) = \mathbf{1}\,\mathbf{1}^T$. Write $e = x + y\sqrt{-1}$, where x, y are real numbers. We have $\|\Delta_{Z_e,Z}(M)\| = n$ and

$$\begin{aligned} \|M\|^2 &\ge \frac{1}{n}\|\mathbf{1}^T M\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left|in + \frac{i(i-1)}{2}(e-1)\right|^2 \\ &\ge \left(\frac{n^4}{20} - \frac{n^3}{8} + \frac{n^2}{12}\right)(x-1)^2 + \left(\frac{n^4}{4} - \frac{n^3}{3}\right)(x-1) + \frac{n^4}{3} \\ &\ge \frac{n^2}{48}\left(n^2 - \frac{2}{5}n + 25\right). \end{aligned}$$

Therefore, $\|\Delta_{Z_e,Z}^{-1}\|_1 \ge cn$ for some constant $c > 0$. Similarly, we have $\|\nabla_{Z_e,Z}^{-1}\|_1 \ge c'n$ for another constant $c' > 0$. This leads to much tighter bounds than the ones of Theorem 8.1 for $A = Z_e$, $B = Z_0$.

In all cases, we have $\|\nabla_{Z_e,Z_f}^{-1}\|_1 \ge cn$ for all $ef \neq 1$; $\|\Delta_{Z_e,Z_f}^{-1}\|_1 \ge cn$ for all $e \neq f$, where c is a positive constant independent of n.

THEOREM 8.5.

$$\max_j |1-|e|^{1/m}|v_j|\omega_{2m}|^{-1} \le \|\Delta_{Z_e,D(\mathbf{v})}^{-1}\| \le \hat{e}\sum_{k=0}^{m-1}\max_j\left|\frac{v_j^k}{1-ev_j^m}\right|, \tag{8.5}$$
$$\max_j ||e|^{1/m}-|v_j|\omega_{2m}|^{-1} \le \|\nabla_{Z_e,D(\mathbf{v})}^{-1}\| \le \hat{e}\sum_{k=0}^{m-1}\max_j\left|\frac{v_j^k}{e-v_j^m}\right|. \tag{8.6}$$

Proof. (1) The lower bounds in (8.5) and (8.6) follow from Theorem 8.1 for $A = Z_e$, $B = D(\mathbf{v})$.

(2) Let $\Delta = \Delta_{Z_e,D(\mathbf{v})}(M)$. Then $M = \sum_{k=0}^{m-1} Z_e^k \Delta D(\mathbf{v})^k (I_n - eD(\mathbf{v})^m)^{-1}$ (see Corollary 4.3). So $\|M\| \le \sum_{k=0}^{m-1} \|Z_e^k\| \ \|\Delta\| \ \|D(\mathbf{v})^k (I_n - eD(\mathbf{v})^m)^{-1}\| \le \|\Delta\| \hat{e} \sum_{k=0}^{m-1} \max_j |\frac{v_j^k}{e-v_j^m}|$.

(3) Assume $e \neq 0$; otherwise, $\|\nabla^{-1}_{Z_0,D(\mathbf{v})}\| = \lim_{e\to 0} \|\nabla^{-1}_{Z_e,D(\mathbf{v})}\|$. Let $\nabla = \nabla_{Z_e,D(\mathbf{v})}(M)$. Then $M = \sum_{k=0}^{m-1} Z_e^{-1-k} \nabla D(\mathbf{v})^k (I_n - eD(\mathbf{v})^m)^{-1}$ (see Corollary 4.3). So $\|M\| \le \sum_{k=0}^{m-1} \|Z_e^{-1-k}\| \ \|\nabla\| \ \|D(\mathbf{v})^k (I_n - eD(\mathbf{v})^m)^{-1}\| \le \|\nabla\| \hat{e} \sum_{k=0}^{m} \max_j |\frac{v_j^k}{e-v_j^m}|$. ∎

Remark 8.6. Suppose $|v_j| \notin (1-\epsilon, 1+\epsilon)$ for a constant $\epsilon > 0$ and for all j.

(a) If $e \neq 0$, then

$$\lim_{m\to\infty} \sum_{k=0}^{m-1} \max_j \left| \frac{v_j^k}{1 - ev_j^m} \right| < \frac{1}{\epsilon} \max\left(1, \frac{1}{|e|}\right),$$
$$\lim_{m\to\infty} \sum_{k=0}^{m-1} \max_j \left| \frac{v_j^k}{e - v_j^m} \right| < \frac{1}{\epsilon} \max\left(1, \frac{1}{|e|}\right).$$

(b) If $e = 0$, let us improve the lower bound by sampling $M = (v_j^{i-1})_{i=1}^n \mathbf{e}_j^T$. Then $\Delta_{Z,D(\mathbf{v})}(M) = \mathbf{e}_1 \mathbf{e}_j^T$. Write $v = \max_j |v_j|$. Since $\|\Delta_{Z,D(\mathbf{v})}(M)\| = 1$, we have $\sqrt{\frac{v^{2m}-1}{v^2-1}} \le \|\Delta^{-1}_{Z,D(\mathbf{v})}\|_1 \le \frac{v^m-1}{v-1}$. Compare the latter lower and upper bounds as $m \to \infty$ to obtain $\lim_{m\to\infty} \sqrt{\frac{v^{2m}-1}{v^2-1}} / \frac{v^m-1}{v-1} = |\frac{v+1}{v-1}|$. That is, our estimates (8.5), (8.6) are asymptotically tight as $m \to \infty$ for any e provided that $\{v_j\}$ are not clustered around 1.

Theorem 3.1 for $A = D(\mathbf{u})$, $B = D(\mathbf{v})$ implies the next corollary.

COROLLARY 8.7.

$$\frac{1}{\min_{i,j} |1 - u_i v_j|} \le \|\Delta^{-1}_{D(\mathbf{u}),D(\mathbf{v})}\|_r \le \frac{\sqrt{r}}{\min_{i,j} |1 - u_i v_j|}, \tag{8.7}$$

$$\frac{1}{\min_{i,j} |u_i - v_j|} \le \|\nabla^{-1}_{D(\mathbf{u}),D(\mathbf{v})}\|_r \le \frac{\sqrt{r}}{\min_{i,j} |u_i - v_j|}. \tag{8.8}$$

Remark 8.8. The lower and upper bounds of Corollary 8.7 are within the factor of $\sqrt{r}$ from each other, and r is small for structured matrices.

Next, we extend the upper estimates of Corollary 8.7 for $\|L^{-1}\|_r$ based on the displacement transformation techniques, with the goal of improving our estimates (8.3)–(8.6) when $|e|$ and $|f|$ are not too small or too large.

THEOREM 8.9. *Let $\hat{A} = VAV^{-1}$, $\hat{B} = W^{-1}BW$ for some nonsingular matrices V and W, $C = \|V\| \ \|V^{-1}\| \ \|W\| \ \|W^{-1}\|$. Then*

$$\|\Delta^{-1}_{A,B}\|_r \le C \|\Delta^{-1}_{\hat{A},\hat{B}}\|_r, \qquad \|\nabla^{-1}_{A,B}\|_r \le C \|\nabla^{-1}_{\hat{A},\hat{B}}\|_r.$$

Proof. $\Delta_{\hat{A},\hat{B}}(VMW) = V\Delta_{A,B}(M)W$, $\nabla_{\hat{A},\hat{B}}(VMW) = V\nabla_{A,B}(M)W$. ∎

By combining Theorems 8.9 and 2.3, we transform the operators $\Delta^{-1}_{Z_e,Z_f}$, $\nabla^{-1}_{Z_e,Z_f}$ and $\Delta^{-1}_{Z_e,D(\mathbf{v})}$, $\nabla^{-1}_{Z_e,D(\mathbf{v})}$ into the operators $\Delta^{-1}_{D(\mathbf{u}),D(\mathbf{v})}$, $\nabla^{-1}_{D(\mathbf{u}),D(\mathbf{v})}$ and then extend the bounds of Corollary 8.7 to the former operators. We arrive at a corollary showing the desired improvement of (8.3)–(8.6).

COROLLARY 8.10. *Suppose* $ef \neq 0$; $e^{\frac{1}{m}}$ *and* $f^{\frac{1}{n}}$ *are any* m*th and* n*th roots of* e *and* f*, respectively. Write* $\tilde{e} = \max(|e|, \frac{1}{|e|})$, $\tilde{f} = \max(|f|, \frac{1}{|f|})$, *and then*

$$\|\Delta^{-1}_{Z_e,Z_f}\|_r \le \sqrt{r}\tilde{e}^{\frac{m-1}{m}}\tilde{f}^{\frac{n-1}{n}} \max_{i,j} |1 - e^{\frac{1}{m}}\omega_m^i f^{\frac{1}{n}}\omega_n^j|^{-1},$$

$$\|\nabla^{-1}_{Z_e,Z_f}\|_r \le \sqrt{r}\tilde{e}^{\frac{m-1}{m}}\tilde{f}^{\frac{n-1}{n}} \max_{i,j} |e^{\frac{1}{m}}\omega_m^i - f^{\frac{1}{n}}\omega_n^j|^{-1},$$

$$\|\Delta_{Z_e,D(\mathbf{v})}\|_r \le \sqrt{r}\tilde{e}^{\frac{m-1}{m}} \max_{i,j} |1 - e^{\frac{1}{m}}\omega_m^i v_j|^{-1},$$

$$\|\nabla_{Z_e,D(\mathbf{v})}\|_r \le \sqrt{r}\tilde{e}^{\frac{m-1}{m}} \max_{i,j} |e^{\frac{1}{m}}\omega_m^i - v_j|^{-1}.$$

9. Decreasing the norm $\|L^{-1}\|$. Typically, in the DECOMPRESS stage of the displacement rank approach, the numerical problems diminish where $\|L^{-1}\|$ is smaller. In particular, this factor is critical for rapid convergence of Newton's iteration with recursive compression applied to invert a structured matrix [P92], [P01], [PRW02], [PKRC02], [CPVBW02]. It is, therefore, desired to decrease $\|L^{-1}\|$. Surprisingly, this is possible based on the displacement transformation approach proposed in [P90]. According to this approach, a successful algorithm or method of study for a specific class of structured matrices can be extended to other classes of structured matrices via appropriate transformation of the associated displacement operators. At the end of the preceding section, we applied this approach to extend our estimates of Corollary 8.7 from Cauchy-like to Toeplitz/Hankel-like and Vandermonde-like matrices. Let us demonstrate how this works by another example. Suppose we seek the solution of a nonsingular linear system $M\mathbf{x} = \mathbf{b}$, where the Cauchy-like input matrix M is associated with an operator $L_0 = \nabla_{D(\mathbf{s}),D(\mathbf{t})}$, and suppose the norm $\|L_0^{-1}\|$ is too large. Let us solve the problem by using the displacement transformation method. Choose a vector $\mathbf{v} = (a\omega_n^i)_{i=0}^{n-1}$ for a scalar a such that $\|L^{-1}\|$ is small for $L = \nabla_{D(\mathbf{s}),D(\mathbf{v})}$. According to Corollary 8.7, this is the case if the component sets of the vectors $\mathbf{s}$ and $\mathbf{v}$ are well isolated from each other. Solve the linear system $MC(\mathbf{t},\mathbf{v})\mathbf{y} = \mathbf{b}$ whose coefficient matrix is associated with the operator $L = \nabla_{D(\mathbf{s}),D(\mathbf{v})}$ and is typically not worse conditioned than M; finally, recover $\mathbf{x} = C(\mathbf{t},\mathbf{v})\mathbf{y}$. Due to the transition from L_0 to L, the critical stage of the solution can be dramatically simplified (for instance, if the solution is obtained by using Newton's iteration). The above recipe can be immediately extended to the case of Toeplitz-like, Hankel-like, Vandermonde-like, and other structured matrices M based on their well-known simple transformations into Cauchy-like matrices (see Theorems 2.3, 3.7, and 8.9 and [P90]).

Acknowledgments. We thank both referees for many helpful comments and the editor Prof. Dr. Ir. Sabine Van Huffel for ensuring fast processing of the paper and compression of its first draft.

REFERENCES

[AG91] G. AMMAR AND P. GADER, *A variant of the Gohberg–Semencul formula involving circulant matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 534–540.

[BGR90] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Oper. Theory Adv. Appl. 45, Birkhäuser, Basel, 1990.

[BP93] D. A. BINI AND V. Y. PAN, *Improved parallel computations with Toeplitz-like and Hankel-like matrices*, Linear Algebra Appl., 188/189 (1993), pp. 3–29.

[BP94] D. A. BINI AND V. Y. PAN, *Polynomial and Matrix Computations, Vol.* 1: *Fundamental Algorithms*, Birkhäuser Boston, Boston, 1994.

[BVB97] A. Bultheel and M. Van Barel, *Linear Algebra, Rational Approximation and Orthogonal Polynomials*, Stud. Comput. Math. 6, North–Holland, Amsterdam, 1997.

[CJL96] S. Cabay, A. R. Jones, and G. Labahn, *Computation of numerical Padé–Hermite and simultaneous Padé systems* I: *Near inversion of generalized Sylvester matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 248–267.

[CJL96a] S. Cabay, A. R. Jones, and G. Labahn, *Computation of numerical Padé–Hermite and simultaneous Padé systems* II: *A weakly stable aglorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 268–297.

[CPVBW02] G. Codevico, V. Y. Pan, M. van Barel, and X. Wang, *Iterative Inversion of Structured Matrices*, Tech. report 200214, Ph.D. Program in Computer Science, The Graduate Center of the City University of New York, New York, NY, 2002 (submitted).

[CPW74] R. E. Cline, R. J. Plemmons, and G. Worm, *Generalized inverses of certain Toeplitz matrices*, Linear Algebra Appl., 8 (1974), pp. 25–33.

[GKKL87] I. Gohberg, T. Kailath, I. Koltracht, and P. Lancaster, *Linear complexity parallel algorithms for linear systems of equations with recursive structure*, Linear Algebra Appl., 88/89 (1987), pp. 271–315.

[GKO95] 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.

[GO92] I. Gohberg and V. Olshevsky, *Circulants, displacements and decompositions of matrices*, Integral Equations Operator Theory, 15 (1992), pp. 730–743.

[GO94] I. Gohberg and V. Olshevsky, *Complexity of multiplication with vectors for structured matrices*, Linear Algebra Appl., 202 (1994), pp. 163–192.

[GO94a] I. Gohberg and V. Olshevsky, *Fast state-space algorithms for matrix Nehari and Nehari–Takagi interpolation problems*, Integral Equations Operator Theory, 20 (1994), pp. 44–83.

[G98] M. Gu, *Stable and efficient algorithms for structured systems of linear equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 279–306.

[G98a] M. Gu, *New fast algorithms for structured linear least squares problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 244–269.

[H95] G. Heinig, *Inversion of generalized Cauchy matrices and the other classes of structured matrices*, in Linear Algebra for Signal Processing, IMA Vol. Math. Appl. 69, Springer-Verlag, New York, 1995, pp. 95–114.

[HR84] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Oper. Theory Adv. Appl. 13, Birkhäuser, Basel, 1984.

[KKM79] T. Kailath, S. Y. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[KO96] T. Kailath and V. Olshevsky, *Displacement structure approach to discrete transform based preconditioners of G. Strang type and of T. Chan type*, Calcolo, 33 (1996), pp. 191–208.

[KS95] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[KS99] T. Kailath and A. H. Sayed, eds., *Fast Reliable Algorithms for Matrices with Structure*, SIAM, Philadelphia, 1999.

[KVB99] P. Kravanja and M. Van Barel, *Algorithms for solving rational interpolation problems related to fast and superfast solvers for Toeplitz systems*, in Proceedings of Advanced Signal Processing Algorithms, Architecture and Implementation IX (Denver, CO), SPIE Publications, Bellingham, WA, 1999, pp. 359–370.

[OP98] V. Olshevsky and V. Y. Pan, *A unified superfast algorithm for boundary rational tangential interpolation problem*, in Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Los Alamitos, CA, 1998, pp. 192–201.

[OS00] V. Olshevsky and M. A. Shokrollahi, *A unified superfast algorithm for confluent tangential interpolation problem and for structured matrices*, in Proceedings of the 14th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2000), University of Perpignan, Perpignan, France, 2000.

[P90] V. Y. Pan, *On computations with dense structured matrices*, Math. Comput., 55 (1990), pp. 179–190. Proceedings version in Proceedings of International Symposium on Symbolic and Algebraic Computation (ISSAC '89), ACM, New York, 1989, pp. 34–42.

[P92] V. Y. Pan, *Parallel solution of Toeplitz-like linear systems*, J. Complexity, 8 (1992), pp. 1–21.

[P93] V. Y. Pan, *Decreasing the displacement rank of a matrix*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 118–121.

[P99] V. Y. Pan, *A Unified Superfast Divide-and-Conquer Algorithm for Structured Matrices over Abstract Fields*, MSRI Preprint 1999-033, Mathematical Sciences Research Institute, Berkeley, CA, 1999.

[P00] V. Y. Pan, *Nearly optimal computations with structured matrices*, in Proceedings of the 11th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA 2000), ACM, New York, SIAM, Philadelphia, 2000, pp. 953–962.

[P01] V. Y. Pan, *Structured Matrices and Polynomials: Unified Superfast Algorithms*, Birkhäuser Boston, Boston, Springer-Verlag, New York, 2001.

[PKRC02] V. Y. Pan, M. Kunin, R. Rosholt, and H. Cebecioğlu, *Homotopic Residual Correction Processes*, Tech. report 200215, Ph.D. Program in Computer Science, The Graduate Center of the City University of New York, New York, NY, 2002 (submitted).

[PRW02] V. Y. Pan, Y. Rami, and X. Wang, *Structured matrices and Newton's iteration: Unified approach,* Linear Algebra Appl., 343/344 (2002), pp. 233–265.

[PSD70] P. Penfield, Jr., R. Spencer, and S. Duinker, *Tellegen's Theorem and Electrical Networks*, MIT Press, Cambridge, MA, 1970.

[W93] D. H. Wood, *Product rules for the displacement of nearly-Toeplitz matrices*, Linear Algebra Appl., 188/189 (1993), pp. 641–663.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 678–692

MODIFIED FINITE SECTIONS FOR TOEPLITZ OPERATORS AND THEIR SINGULAR VALUES*

BERND SILBERMANN†

Abstract. The topic of this paper is the study of modified finite sections of Toeplitz operators and their singular values. We prove the splitting property for the singular values and consider two important consequences. We show that the kernel dimension of a Fredholm Toeplitz operator with a piecewise continuous matrix-valued generating function can be extracted from the singular values behavior of the modified sections. Second, we generalize the results on asymptotic Moore–Penrose invertibility of Heinig and Hellinger [*Integral Equations Operator Theory*, 19 (1994), pp. 419–446] to piecewise continuous generating functions.

Key words. Toeplitz operators, singular values, finite sections

AMS subject classifications. 45E10, 65F20

PII. S089547980139515X

1. Introduction.

Let PC denote the C^*-algebra of all piecewise continuous functions defined on the unit circle $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$, and let $PC_{N\times N}$ be the C^*-algebra of all $N \times N$ matrices with entries from PC. We shall mainly deal with the question of how the singular values of matrices A_n approximating the Toeplitz operator $T(a)$ acting on the space l^2_N are distributed, where $a \in PC_{N\times N}$ and the operator $T(a)$ is supposed to be Fredholm. Of course one expects that the answer depends strongly on the kind of the matrices A_n. There are many possible approximations A_n; here we restrict ourselves to the so-called modified finite sections. If the approximations are the familiar finite sections $T_n(a)$ (which are square matrices), the complete answer was obtained by Roch and the author in [R/S 2]. It was shown that the set Λ_n of the singular values of the finite sections $T_n(a)$ of a Fredholm Toeplitz operator is subject to the splitting property. We say that the singular values (computed via $A_n^*A_n$) of a sequence (A_n) of $k(n) \times l(n)$ matrices A_n have the splitting property if there exist a sequence $c_n \longrightarrow 0\,(c_n \geq 0)$ and a number $d > 0$ such that

$$\Lambda_n \subset [0, c_n] \cup [d, \infty) \quad \text{for all} \quad n\,,$$

and the singular values of A_n are said to meet the k-splitting property if, in addition, for all sufficiently large n exactly k singular values of A_n lie in $[0, c_n]$.

The mentioned result now reads as follows: If $T(a)$ is Fredholm, $a \in PC_{N\times N}$, then the sequence $(T_n(a))$ has the k-splitting property with

$$k = \dim\ \ker T(a) + \dim\ \ker T(\tilde{a})\,,$$

where $\tilde{a}(t) := a(1/t)$.

Thus, if we would know the number dim ker $T(\tilde{a})$, then we would know the kernel dimension of $T(a)$, provided that we would be able to compute the set $\Lambda_n \cap [0, c_n]$. As a rule, we know the number dim ker $T(\tilde{a})$ only in very special cases. So the question arises whether the operator can be approximated by matrices A_n such that

*Received by the editors September 14, 2001; accepted for publication by L. Elden May 13, 2002; published electronically January 23, 2003.
http://www.siam.org/journals/simax/24-3/39515.html
†Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany, (bernd.silbermann@mathematik.tu-chemnitz.de).

the splitting property still holds with some operator $\tilde{A}$ instead of $T(\tilde{a})$ and such that the kernel dimension of $\tilde{A}$ is available. In other words, we try to design approximations A_n to $T(a)$ having prescribed properties. We shall show that the so-called modified finite sections are good candidates for our aim. We also show that our approach is intimately related to the approximation of the Moore–Penrose inverse of the Toeplitz operator $T(a)$. In the course of the paper we do not recover only the results of Heinig and Hellinger [H/H] for Toeplitz operators $T(a)$ with a from the Wiener class $W_{N\times N}$, but we extend them to operators $T(a)$ with $a \in PC_{N\times N}$. Notice that the methods of [H/H] do not work in this more general situation. Our main tool is a C^*-algebra approach mainly developed by Roch and the author in the last years (see, for instance, the book [H/R/S]). Let us mention two results proved in sections 3 and 4. Define block matrices $T_{n,0,r}(a)$ and $T_{n,r,0}(a)$ (whose entries are $N \times N$-matrices) by

$$T_{n,0,r}(a) = (a_{i-j})\,,\ \ 0 \le i \le n,\, 0 \le j \le n-r,$$

$$T_{n,r,0}(a) = (a_{i-j})\,,\ \ 0 \le i \le n-r,\, 0 \le j \le n,$$

respectively, where $a_k\,(k \in \mathbb{Z})$ are the Fourier coefficients of $a \in PC_{N\times N}$. The following theorems are consequences of the main results obtained in sections 3 and 4, respectively.

THEOREM 1.1. *Let the Toeplitz operator $T(a) : l^2_N \longrightarrow l^2_N$ be Fredholm, $a \in PC_{N\times\mathbb{N}}$. Then the singular values of the sequence $(T_{n,0,r}(a))$ enjoy the k-splitting property, where k depends on r. Moreover,*

$$k = \dim\ker T(a)$$

for r large enough.

Examples will be presented in the appendix.

In what follows, let A^+ denote the Moore–Penrose inverse of an operator A.

THEOREM 1.2. *Let $T(a)$ be Fredholm, $a \in PC_{N\times N}$.*

(a) *If $T(a)$ is left invertible, then there is an r_0 such that the Moore–Penrose inverses $(T^+_{n,0,r}(a))$ converge strongly to $T^+(a)$ for all $r \ge r_0$ as n goes to infinity.*

(b) *If $T(a)$ is right invertible, then there is an r_0 such that the Moore–Penrose inverses $(T^+_{n,r,0}(a))$ converge strongly to $T^+(a)$ for all $r \ge r_0$ as n goes to infinity.*

2. Toeplitz operators and the algebra generated by familiar finite sections. We shall see in section 3 that the sequences $(T_{n,0,r}(a))$ and $(T_{n,r,0}(a))$ can be identified with some sequences of square matrices which belong to the algebra $\mathcal{A}$ generated by all sequences of familiar finite sections of Toeplitz operators with generating functions from $PC_{N\times N}$. This observation already shows that it would certainly be of importance to have as much knowledge on $\mathcal{A}$ as possible. Fortunately, the algebra $\mathcal{A}$ was intensively studied in the past. Here we merely recall definitions and some nontrivial facts needed in what follows. Let l^2_N denote the Hilbert space of all sequences $(x_i)_{i\in\mathbb{Z}^+}, \mathbb{Z}^+ := \{k \in \mathbb{Z} : k \ge 0\}$, where $x_i \in \mathbb{C}^N$ and

$$\|\,(x_i)\,\| := \left(\sum_{i=0}^{\infty} \|\,x_i\,\|^2\right)^{\frac{1}{2}} < \infty.$$

($\|\,x_i\,\|$ refers to the familiar euclidean norm in $\mathbb{C}^N$.)

Given an $N \times N$ matrix-valued function $a \in L^\infty_{N\times N}$ (where L^∞ means the essentially bounded functions defined on $\mathbb{T}$) denote the sequence of its Fourier coefficients by $(a_n)_{n\in\mathbb{Z}}$. The Toeplitz operator $T(a) : l^2_N \longrightarrow l^2_N$ is defined by $(x_i) \mapsto (y_i)$, where

$$y_i = \sum_{j=0}^{\infty} a_{i-j}x_j \ (i \in \mathbb{Z}^+) \, .$$

The Toeplitz operator $T(a)$ with generating function $a \in L^\infty(\mathbb{T})_{N\times N}$ is bounded, that is, $T(a) \in \mathcal{L}(l^2_N)$, and moreover $\| T(a) \| = \| a \|_\infty$ (see, for instance, [B/S 2]). Here, for a Hilbert space $\mathcal{H}$, we denote by $\mathcal{L}(\mathcal{H})$ the C^*-algebra of all bounded linear operators acting on $\mathcal{H}$. Further, let $\mathcal{K}(\mathcal{H})$ stand for the closed two-sided ideal of all compact operators. Now introduce operators P_n and W_n on l^2_N by

$$\begin{aligned} P_n(x_0, x_1, \ldots, x_n, \ldots) &= (x_0, \ldots, x_n, 0, \ldots) \, , \\ W_n(x_0, x_1, \ldots, x_n, \ldots) &= (x_n, x_{n-1}, \ldots, x_0, 0, \ldots) \, . \end{aligned} \tag{2.1}$$

Obviously, $P_n, W_n \in \mathcal{L}(l^2_N)$ and

$$P_n^2 = P_n, \quad W_n^2 = P_n \, .$$

In what follows we will identify operators acting on im P_n or on l^2 $(N = 1)$ with their matrix representation with respect to the standard basis of im P_n or l^2, respectively. We proceed analogously in the case $N > 1$. For $n \in \mathbb{Z}^+$ the (familiar) finite section $T_n(a)$ of $T(a)$ is defined by

$$T_n(a) := P_n T(a) P_n \, .$$

The finite section $T_n(a)$ is related to the operator W_n by

$$W_n T_n(a) W_n = T_n(\tilde{a}) \, ,$$

where $\tilde{a}$ is defined by $\tilde{a}(t) := a(1/t)$.

The matrix representation of $T_n(a)$ is given by the finite block Toeplitz matrix

$$(a_{i-j})_{i,j=0}^{n} \, ,$$

whereas the underlying matrix representation of $T(a)$ is given by the infinite Toeplitz matrix

$$(a_{i-j})_{i,j=0}^{\infty} \, .$$

Let $\mathcal{F}$ denote the collection of all operator sequences $(A_n)_{n\in\mathbb{Z}^+}$ with $A_n \in \mathcal{L}(\text{im } P_n)$ and

$$\| (A_n) \| := \sup_n \| A_n \| < \infty \, . \tag{2.2}$$

With the operations $(A_n) + (B_n) := (A_n + B_n)$, $(A_n)(B_n) := (A_n B_n)$, $(A_n)^* := (A_n^*)$ and the norm (2.2), $\mathcal{F}$ actually becomes a C^*-algebra. First of all, recall that a sequence $(A_n) \in \mathcal{F}$ is called norm stable if the operators $A_n :$ im $P_n \longrightarrow$ im P_n are invertible for n large enough (say, for $n \geq n_0$) and

$$\sup_{n\geq n_0} \| A_n^{-1} \| < \infty \, .$$

If, in addition, A_n converges strongly to some invertible operator A, then the sequence $(A_k^{-1})_{k\geq n_0}$ converges strongly to A^{-1}. We shall write $s\text{-lim } A_n = A$ if the sequence A_n tends strongly to A. Let $\mathcal{G}$ denote the collection of all sequences $(A_n) \in \mathcal{F}$ with $||A_n|| \to 0$. Clearly, $\mathcal{G}$ actually forms a closed two-sided ideal in $\mathcal{F}$.

Note the following.

PROPOSITION 2.1 (see [B/S 2] or [B/S 3]). *$(A_n) \in \mathcal{F}$ is norm stable if and only if the coset $(A_n) + \mathcal{G}$ is invertible in the quotient algebra $\mathcal{F}/\mathcal{G}$.*

Now consider the smallest C^*-subalgebra $\mathcal{A} \subset \mathcal{F}$ containing all sequences $(T_n(a))$ with $a \in PC_{N\times N}$. The algebra $\mathcal{A}$ has a lot of remarkable properties which will be of decisive importance in studying the problems formulated in the introduction.

PROPOSITION 2.2 (see [B/S 2] or [B/S 3]). *Let $K_1, K_2 \in \mathcal{K}(l_N^2), (C_n) \in \mathcal{G}$. Then*

$$(B_n) := (P_nK_1P_n + W_nK_2W_n + C_n) \in \mathcal{A}\,. \tag{2.3}$$

Moreover, all sequences of the form (2.3) *form a closed two-sided ideal in $\mathcal{A}$.*

PROPOSITION 2.3 (see [B/S 2] or [B/S 3]). *For each sequence $(A_n) \in \mathcal{A}$ there exist the strong limits*

$$\begin{aligned} \mathcal{W}_1\,(A_n) &:= s - \lim A_n\,, \\ \mathcal{W}_2\,(A_n) &:= s - \lim W_nA_nW_n\,. \end{aligned}$$

Moreover, $\mathcal{W}_1 : \mathcal{A} \to \mathcal{L}(l_N^2)$ and $\mathcal{W}_2 : \mathcal{A} \to \mathcal{L}(l_N^2)$ are $*$-homomorphisms that act as follows:

$$\begin{aligned} &\mathcal{W}_1\,(T_n(a)) = T(a)\,, \quad \mathcal{W}\,(T_n(a)) = T(\tilde{a})\,, \\ &\mathcal{W}_1\,(B_n) = K_1\,, \qquad\quad \mathcal{W}_2\,(B_n) = K_2\,, \end{aligned}$$

where (B_n) is the sequence (2.3).

Now we formulate a theorem, which is completely proved in [B/S 1] and [S 1]. This theorem was, however, not explicitly stated there, but it is a direct consequence of Theorems 1 and 2 in [B/S 1]. The first explicit formulation was published in [S 2].

THEOREM 2.4. *Let $(A_n) \in \mathcal{A}$ be arbitrarily given.*

(a) *The sequence (A_n) is norm stable if and only if the operators $\mathcal{W}_1(A_n)$ and $\mathcal{W}_2(A_n)$ are invertible.*

(b) *The operator $\mathcal{W}_1(A_n)$ is a Fredholm operator if and only if the operator $\mathcal{W}_2(A_n)$ is a Fredholm operator.*

We call a sequence $(A_n) \in \mathcal{A}$ a Fredholm sequence if $\mathcal{W}_1(A_n)$ is a Fredholm operator.

This theorem is a far going extension of classic results (see, for instance, [G/F]). It is easy to see that the mapping

$$\text{smb} : (A_n) \mapsto (\mathcal{W}_1(A_n), \mathcal{W}_2(A_n))$$

is a $*$-homomorphism of the C^*-algebra $\mathcal{A}$ into the C^*-algebra $\mathcal{L}_2 := \mathcal{L}(l_N^2) \oplus \mathcal{L}(l_N^2)$, the direct sum of two copies of $\mathcal{L}(l_N^2)$, with norm $\| (B, C) \| = \max\{\| B \|, \| C \|\}$. The image of $\mathcal{A}$ under this homomorphism is denoted by smb $\mathcal{A}$. The element smb (A_n) is called the stability symbol of (A_n).

THEOREM 2.5 (see [H/R/S] or [S 2]). *The algebras $\mathcal{A}/\mathcal{G}$ and* smb $\mathcal{A}$ *are isometrically isomorphic. The isomorphism is given by*

$$(A_n) + \mathcal{G} \mapsto \text{ smb } (A_n)\,.$$

This theorem shows that $\mathcal{A}/\mathcal{G}$ can be represented in a very nice way. Moreover, it says that all properties of a sequence $(A_n) \in \mathcal{A}$ which do not depend on the first

members of (A_n) should be stored in the operators $\mathcal{W}_1(A_n)$ and $\mathcal{W}_2(A_n)$. In other words the asymptotic properties of (A_n) should be reflected in the mentioned operators. The following theorem makes this precise for the asymptotic behavior of the singular values.

THEOREM 2.6 (see [H/R/S] or [R/S 1]). *Let $(A_n) \in \mathcal{A}$ be a Fredholm sequence, and let Λ_n denote the set of all singular values of A_n. Then (A_n) is subject to the k-splitting property with*

$$k = \dim \ker \mathcal{W}_1(A_n) + \dim \ker \mathcal{W}_2(A_n)\,.$$

One can show that the k-splitting property is also necessary for $\mathcal{W}_1(A_n)$ being Fredholm (see [H/R/S]*). We will make use of these theorems in the next sections.*

3. Modified finite sections and the splitting property. For each multiindex $\alpha = (\alpha_1, \ldots, \alpha_N)$, $\alpha_i \in \mathbb{Z}^+ (i = 1, \ldots, N)$ and for each operator $A \in \mathcal{L}(l^2)$ we define an operator $A^\alpha \in \mathcal{L}(l^2_N)$ by

$$\text{diag } (A^{\alpha_1}, \ldots, A^{\alpha_N})\,.$$

Further, let e_1 and e_{-1} stand for the functions $e_1, e_{-1} : \mathbb{T} \longrightarrow \mathbb{T}$ defined by $t \mapsto t$ and $t \mapsto t^{-1}$, respectively. The Toeplitz operator $T(e_1) : l^2 \longrightarrow l^2$ will also be denoted by V. Then it follows that $V^* = T(e_{-1})$. Notice that for any multiindices α and β and any function $a \in L^\infty_{N\times N}(\mathbb{T})$ the property

$$V^{*\beta} T(a) V^\alpha = T\left(e^\beta_{-1} a e^\alpha_1\right) \tag{3.1}$$

is fulfilled. We shall also use the projections

$$P_\alpha := \text{ diag } (P_{\alpha_1}, \ldots, P_{\alpha_N})\,, \tag{3.2}$$

where P_{α_i} is defined by (2.1) for $N = 1$. The multiindex $(n, n, \ldots, n)$ will also be denoted by n. In each case the meaning will become clear from the context. Notice also the relations (α, β,-multiindices)

$$V^\alpha P_\beta = P_{\beta+\alpha} V^\alpha\,, \quad P_\beta V^{*\alpha} = V^{*\alpha} P_{\beta+\alpha} \tag{3.3}$$

and

$$P_\beta V^\alpha = P_\beta V^\alpha P_\beta, \quad V^{*\alpha} P_\beta = P_\beta V^{*\alpha} P_\beta\,. \tag{3.4}$$

It is sufficient to prove these assertions in the case $N = 1$. Recall that the projection $P_m\,(m \in \mathbb{Z}^+)$ can be written as $P_m = I - V^m V^{*m}$. Now it follows that

$$V^k P_m = V^k - V^{m+k} V^{*m+k} V^k = P_{m+k} V^k\,.$$

By taking the adjoint we get (3.3). The proof of (3.4) is also very simple.

In what follows we shall consider modified finite sections of the Toeplitz operator $T(a)$ of the form

$$T_{n,\alpha,\beta}(a) := P_{n-\alpha} T(a) P_{n-\beta}\,, \tag{3.5}$$

where $n = (n, \ldots, n)$ and $P_{n-\alpha}(P_{n-\beta})$ is the zero operator if $n - \alpha$ ($n - \beta$) is not a multiindex, that is, if it has negative components. With help of (3.3) and (3.4) the finite sections (3.5) can be rewritten ($n - \alpha \geq 0$, $n - \beta \geq 0$) as

$$P_{n-\alpha} T(a) P_{n-\beta} = P_n V^{*\alpha} P_n T(e^{-\alpha}_{-1} a e_1^{-\beta}) P_n V^\beta P_n\,. \tag{3.6}$$

This simple observation is crucial: it shows that the sequence of the finite sections (3.5) belongs to the algebra $\mathcal{A}$ for $a \in PC_{N\times N}$. In what follows we will identify (3.5) with (3.6), which are square matrices and can be assumed to be extensions of the matrices (3.5) by zeros. The following theorem is a direct consequence of Theorem 2.6.

THEOREM 3.1. *Let the Toeplitz operator $T(a)$ be Fredholm and $a \in PC_{N\times N}$. Then the singular values of $(T_{n,\alpha,\beta}(a))$ meet the k-splitting property with*

$$k = \dim\ker T(a) + \dim\ker \tilde{T}_{\alpha,\beta}(a)\,, \tag{3.7}$$

where

$$\tilde{T}_{\alpha,\beta}(a) := V^{\alpha}T(e^{\alpha}_{-1}\tilde{a}e^{\beta}_{1})V^{*\beta}\,.$$

Proof. It is easy to see that

$$\begin{aligned}\mathcal{W}_1(T_{n,\alpha,\beta}(a)) &= T(a)\,,\\ \mathcal{W}_2(T_{n,\alpha,\beta}(a)) &= \tilde{T}_{\alpha,\beta}(a)\,.\end{aligned}$$

Now it remains to apply Theorem 2.6. ∎

We would like to employ this theorem in order to compute the kernel dimension of a Fredholm Toeplitz operator $T(a)$ with $a \in PC_{N\times N}$. To this aim we introduce the notion of generalized factorization for $p = 2$ (see [L/S]): a right factorization in $L^2(\mathbb{T})$ of a matrix function $G : \mathbb{T} \longrightarrow \mathbb{C}_{N\times N}$ is by definition a representation of the form

$$G(t) = G_-(t)\Lambda(t)G_+(t)\,, \tag{3.8}$$

where $G_+^{\pm 1} \in H^2$ (H^2 is the Hardy space), $G_-^{\pm 1} \in \overline{H}^2$, $\Lambda(t) = \text{ diag }(t^{\kappa_1},\dots,t^{\kappa_N})$, and $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_N$ are integers. It is known that the numbers κ_i, $i = 1,2,\dots,N$, are uniquely determined if the representation (3.8) exists. They are called the right partial indices. Analogously one defines a left factorization:

$$G(t) = \hat{G}_+(t)\Omega(t)\hat{G}_-(t)\,,$$

where $\hat{G}_+, \hat{G}_-$ and Ω fulfill the same conditions as above. Even if for a given matrix function G a left and a right factorization exists, then the right and left partial indices do not necessarily coincide. A simple but useful example is provided by the matrix function

$$G(t) = \begin{pmatrix} t & 1\\ 0 & t^{-1}\end{pmatrix}.$$

In fact, the left and right factorizations are given by

$$\begin{aligned} G(t) &= \begin{pmatrix} 1 & 0\\ 0 & 1\end{pmatrix}\quad\begin{pmatrix} t & 0\\ 0 & t^{-1}\end{pmatrix}\quad\begin{pmatrix} 1 & t^{-1}\\ 0 & 1\end{pmatrix}\\ &= \begin{pmatrix} 1 & 0\\ t^{-1} & -1\end{pmatrix}\quad\begin{pmatrix} 1 & 0\\ 0 & 1\end{pmatrix}\quad\begin{pmatrix} t & 1\\ 1 & 0\end{pmatrix}.\end{aligned}$$

This circumstance causes difficulties in the theory of Toeplitz operators with matrix valued generating functions. From the last example it follows that $T(G)$ is invertible but $T(\tilde{G})$ is not (contrary to scalar valued generating functions). Clearly, if G

possesses a right (left) factorization, then G^{-1} possesses a left (right) factorization, too.

The following fact will be used in what follows (see [L/S] or [G/K]): if $T(a)$, $a \in L^\infty_{N\times N}(\mathbb{T})$, is Fredholm in l^2_N, then a possesses a right factorization and

$$\dim\ \ker T(a) = \sum_{i=1}^{N} \max\{-\kappa_i, 0\}\,.$$

Now we specify Theorem 3.1.

THEOREM 3.2. *Let $T(a)$ be Fredholm, $a \in PC_{N\times N}$. Then there is an $r_0 \in \mathbb{Z}_+$ such that for all $r := (r, \dots, r) \geq (r_0, \dots, r_0)$ the operator $T(\tilde{a}e_1^r)$ has trivial kernel and all statements of Theorem* 3.1 *with respect to the modified section $T_{n,0,r}(a)$ hold and the kernel dimension of $\tilde{T}_{0,r}(a)$ equals $N \cdot r$.*

Proof. Theorem 3.1 ensures that the sequence $(T_{n,0,r}(a))$ has the k-splitting property with

$$k = \dim\ker T(a) + \dim\ker \tilde{T}_{0,r}(a)\,.$$

Since $T(\tilde{a})$ is Fredholm too, the function $\tilde{a}$ possesses a right factorization $\tilde{a}(t) = F_-(t)\Lambda(t)F_+(t), \Lambda(t) = \ \mathrm{diag}\ \{t^{s_1}, \dots, t^{s_N}\}$ and $s_1 \geq s_2 \geq \dots \geq s_N$. Then there exists a number r_0 such that for all multiindices $r = (r, \dots, r) \geq (r_0, \dots, r_0)$ the operator $T(\tilde{a}e_1^r)$ has trivial kernel. Indeed, take $r_0 = \max\{-s_1, \dots, -s_N, 0\}$. Obviously, $\tilde{a}e_1^r$ is subject to the factorization

$$\tilde{a}e_1^r = F_-\Lambda F_+ e_1^r = F_-\Lambda e_1^r F_+$$

and

$$\dim\ker T(\tilde{a}e_1^r) = \sum_{i=1}^{N} \max\{-s_i - r, 0\} = 0, \quad r \geq r_0\,.$$

Thus, $\dim\ \ker\ \tilde{T}_{0,r}(a) = \dim\ker V^{*r} = N \cdot r$, and we are done. ∎

Remark 3.1. In order to compute the kernel dimension of $T(a)$ one has to determine the singular values for $T_{n,0,r}(a)$ lying in $[0, c_n]$ and to subtract $N \cdot r$ (n, r large enough). How large r must be chosen? The following observation is useful: if r is replaced by $r + 1$ and the number of singular values in the respective set $[0, c_n]$ increases exactly by N, then a correct r is found, that is, $r \geq r_0$. Indeed, if $r < r_0$, then the difference of the kernel dimensions

$$\dim\ker \tilde{T}_{0,r+1}(a) - \dim\ker \tilde{T}_{0,r}(a) \tag{3.9}$$

is less than N because $\dim\ \ker\ T(\tilde{a}e_1^r) - \dim\ker T(\tilde{a}e_1^{r+1}) > 0$.

Remark 3.2. If the kernel dimensions of the operators $T(ae_1^r)$ ($r = (r, \dots, r)$) can be computed, then the right partial indices κ_i of a can also be computed.

Remark 3.3. The described procedure offers a way to compute the kernel dimension of a Fredholm Toeplitz operator $T(a)$ with $a \in PC_{N\times N}$. This might seem strange because the kernel dimension of a Fredholm operator A is not stable under small perturbations (it is, however, upper semicontinuous). Nevertheless, the proposal method of kernel computation is stable under small perturbations. The reason is at least the following: We compute the number of singular values of the related

matrices lying in $[0, c_n]$ (n large enough), that is, something like the sum of kernel dimensions, where the related singular values are far from the remaining part of the singular values. More precisely, we have to show that for given $\varepsilon > 0$ the singular values of $T_{n,0,r}(b), b \in PC_{n\times N}$, lie in the set $[0, c_n+\varepsilon]\cup[d-\varepsilon, \infty)$ if only $||T(a)-T(b)||$ is small enough; moreover, we have to show that the number of the singular values of $T_{n,0,r}(b)$ lying in $[0, c_n+\varepsilon]$ equals $\dim\ker T(a) + \dim\ker \tilde{T}_{0,r}(a)$. Further, the computation of the singular values of $T_{n,0,r}(b)$ leads again to computational errors. If they are small enough, we will get the same statement as above. How can one see this?

First observe that the uniform limiting set of the sets Λ_n equals

$$\begin{aligned}\Lambda(a) &= \text{ sp } (\mathcal{W}_1^*(T_{n,0,r})\mathcal{W}_1(T_{n,0,r}))^{\frac{1}{2}} \cup \text{ sp } (\mathcal{W}_2^*(T_{n,0,r})\mathcal{W}_2(T_{n,0,r}))^{\frac{1}{2}} \\ &= \text{ sp } (\text{smb } (T_{n,0,r})^* \text{ smb } (T_{n,0,r}))^{\frac{1}{2}}\end{aligned} \tag{3.10}$$

(see [R/S 2], Theorem 4.14).

If $0 \notin \Lambda(a)$ there is nothing to prove. Indeed, the property $0 \notin \Lambda(a)$ implies for a Fredholm operator $T(a)$, $a \in PC_{N\times N}$, the stability of $(T_{n,0,r})$ by Theorem 2.4. However, stability is stable under small perturbations; this is a direct consequence of Proposition 2.1. Suppose $0 \in \Lambda(a)$. Then the point 0 is an isolated point in $\Lambda(a)$ (by Proposition 4.2); moreover, the multiplicity of 0 equals

$$\dim \text{ ker } (\text{smb } (T_{n,0,r})^* \text{ smb } T_{n,0,r})^{\frac{1}{2}} = \dim \text{ ker } T(a) + \dim \text{ ker } \tilde{T}_{0,r}(a)\,. \tag{3.11}$$

If we approximate the Toeplitz operator $T(a)$ (in the class of Toeplitz operators with $PC_{N\times N}$ generating functions), then if $T(b)$ is close enough to $T(a)$ the point 0 can split into a finite number of points which lie in $[0, \varepsilon) \cap \Lambda(b)$ ($\varepsilon > 0$ given and sufficiently small), and their number (counted with respect to their multiplicity) equals again (3.11) (see Theorem 6.27(d) in [H/R/S]). Now one has to use Theorem 7.12 in [H/R/S] (recall that $\mathcal{A}$ is a standard algebra in the sense of [H/R/S]). Hence, the number of singular values of $T_{n,0,r}(b)$ lying in $[0, c_n + \varepsilon]$ equals again (3.7) for n large enough. This shows that the described procedure is as stable as it can be.

Remark 3.4. Many programs such as MATLAB use immediately the rectangular form of the matrix $T_{n,0,r}(a)$ (that is, they drop down the r last columns consisting of zero matrices) for computing the singular values. In this case the singular values of $(T_{n,0,r}(a))$ have the k-splitting property with

$$k = \dim \text{ ker } T(a)$$

if r is large enough. The above mentioned criterion now reads as follows: a correct r is found if $(T_{n,0,r}(a))$ and $(T_{n,0,r+1}(a))$ have the same k-splitting property. This fact is reflected in Theorem 1.1.

Remark 3.5. If K is compact and $a \in PC_{N\times N}$, then the described methods can also be used to compute dim ker $(T(a) + K)$, where $T(a)$ is Fredholm. One has only additionally to take into account Proposition 2.2.

4. Asymptotic Moore–Penrose invertibility. The splitting property proved in the last section is closely related to the asymptotic Moore–Penrose invertibility. Let H be a Hilbert space, and let us recall that an operator $A \in \mathcal{L}(H)$ is called Moore–Penrose invertible if there is an operator $B \in \mathcal{L}(H)$ such that

$$ABA = A, \quad BAB = B, \quad (AB)^* = AB, \quad (BA)^* = BA. \tag{4.1}$$

It is well known that an operator is Moore–Penrose invertible if and only if its range is closed (such operators are also called normally solvable). Moreover, the

operator B is uniquely determined and will be called the Moore–Penrose inverse of A (also written as $B = A^+$). If $A \in \mathcal{L}(H)$ is Moore–Penrose invertible, then A^+y is the pseudosolution of the equation $Ax = y$, that is, the element with the smallest norm of all the elements x for which $\| Ax - y \|$ is minimal.

We will heavily use the following (and well-known) characterization.

PROPOSITION 4.1. *The following statements are equivalent:*

(i) *The operator $A \in \mathcal{L}(H)$ is Moore–Penrose invertible.*

(ii) *The operator $A^*A + P_{\ker A}$ is invertible.*

(iii) *The operator $AA^* + P_{\ker A^*}$ is invertible.*

Moreover, if this is fulfilled, then

$$A^+ = (A^*A + P_{\ker A}A)^{-1}A^* = A^* \left(AA^* + P_{\ker A^*}\right)^{-1},$$

where P_M denotes the orthogonal projection onto the closed subspace $M \subset H$.

Sketch of the proof. That (i) is equivalent to (ii) is proved, for instance, in [H/R/S]. The equivalence of (i) to (iii) can be proved analogously.

Via the axioms (4.1) one can define Moore–Penrose invertibility for elements in arbitrary C^*-algebras. Again, the Moore–Penrose inverse of a given element is unique, provided it exists, which can be easily seen by representing the C^*-algebra as an algebra of operators.

Notice the following result.

PROPOSITION 4.2 (see [R/S 1] or [H/R/S]).

(i) *An element a of a C^*-algebra with identity is Moore–Penrose invertible if and only if the element a^*a is invertible or if 0 is an isolated point of the spectrum of a^*a. If this condition is fulfilled, then $||a^+|| = \min(sp\ a^*a \setminus \{0\})$.*

(ii) *C^*-subalgebras of C^*-algebras with identity are inverse closed with respect to Moore–Penrose invertibility; that is, if an element of a C^*-subalgebra C of a C^*-algebra B has a Moore–Penrose inverse in B, then this Moore–Penrose inverse necessarily belongs to C.*

The C^*-algebras to which we will apply this proposition are $\mathcal{F}/\mathcal{G}$ and some C^*-subalgebras of it ($\mathcal{F}$ and $\mathcal{G}$ are introduced in section 2). A sequence $(A_n) \in \mathcal{F}$ is said to be Moore–Penrose stable if

$$\sup_{n \geq 1} \| A_n^+ \| < \infty$$

(recall that A_n^+ exists for all n because dim im $P_n < \infty$). We are mainly interested in sequences belonging to $\mathcal{A} \subset \mathcal{F}$ ($\mathcal{A}$ defined also in section 2). It is not hard to find examples of sequences $(T_n(a))$ for which $(T_n^+(a))$ is not bounded, but $T(a)$ is Fredholm. Moreover, for $a \in PC \setminus C \quad (N = 1)$ the sequence $(T_n^+(a))$ is not bounded if $T(a)$ is Fredholm but not invertible (see [B/S 3]). If one allows modified finite sections, the picture changes dramatically. We will use an approach which first occured in [S 2] and temporarily study a weaker problem.

THEOREM 4.3 (see [S 2] or [H/R/S]). *The following assertions are equivalent for a sequence $(A_n) \in \mathcal{A}$:*

(i) *The operators $\mathcal{W}_1(A_n)$ and $\mathcal{W}_2(A_n)$ are normally solvable (that is, they have closed range).*

(ii) *There is a sequence $(B_n) \in \mathcal{A}$ such that*

$$\| A_nB_nA_n - A_n \| \longrightarrow 0, \quad \| B_nA_nB_n - B_n \| \longrightarrow 0,$$
$$\| (A_nB_n)^* - A_nB_n \| \longrightarrow 0, \quad \| (B_nA_n)^* - B_nA_n \| \longrightarrow 0$$

as $n \longrightarrow \infty$.

If one of the conditions is fulfilled, then (B_n) *is unique up to sequences in the ideal* $\mathcal{G}$ *(even in* $\mathcal{F}$*) and* (B_n) *tends strongly to* $\mathcal{W}_1^+(A_n)$.

If $\mathcal{W}_1(A_n)$ *(and therefore also* $\mathcal{W}_2(A_n)$*) is Fredholm, then the assertion is a consequence of Theorem* 2.5.

This theorem can be accomplished by the following proposition.

PROPOSITION 4.4. *Let the situation be as in the preceding theorem, and let* $(A_n) \in \mathcal{A}$. *If the operator* $\mathcal{W}_1(A_n)$ *is Fredholm (therefore,* $\mathcal{W}_2(A_n)$ *is also Fredholm), then the sequences* $(D_n), (D_n')$,

$$D_n = A_n^* A_n + P_n P_{\ker \mathcal{W}_1(A_n)} P_n + W_n P_{\ker \mathcal{W}_2(A_n)} W_n \,,$$
$$D_n' = A_n A_n^* + P_n P_{\ker \mathcal{W}_1(A_n^*)} P_n + W_n P_{\ker \mathcal{W}_2(A_n^*)} W_n \,,$$

belong to $\mathcal{A}$ *and are stable, and the sequences* $(B_n), (B_n')$ *given by*

$$\begin{aligned} B_n &= D_n^+ A_n^* \,, \\ B_n' &= A_n^* D_n'^+ \end{aligned} \tag{4.2}$$

are subject to condition (ii) *of Theorem* 4.1 *(whence it follows that* $(B_n) - (B_n') \in \mathcal{G}$*).*

The proof can be carried out as the proof of Theorem 6.4 in [H/R/S].

Now one might think that the Moore–Penrose inverses A_n^+ for a Moore–Penrose stable sequence $(A_n) \in \mathcal{A}$ have something to do with the operators (4.2). Under some additionally given conditions this is indeed the case. These conditions are summarized in the next proposition which is a special case of a general statement (Proposition 6.5, Theorem 6.7 in [H/R/S]).

PROPOSITION 4.5. *Let* $(A_n) \in \mathcal{A}$, *and let* $\mathcal{W}_1(A_n)$ *be Fredholm.*
Set $B_n := P_n P_{\ker \mathcal{W}_1(A_n)} P_n \,, C_n := W_n P_{\ker \mathcal{W}_2(A_n)} W_n$.

(a) *If* $A_n B_n = A_n C_n = 0$ *for* n *large enough and*

(b) B_n *and* C_n *are projections and* $B_n C_n = 0$ *for* n *large enough,*

then the sequence (A_n) *is Moore–Penrose stable and*

$$P_{\ker A_n} = B_n + C_n$$

for n *sufficiently large.*

The connection of this result with the k*-splitting property is almost obvious: We have (*n *large enough)*

$$\dim \ker A_n = \dim \ker \mathcal{W}_1(A_n) + \dim \ker \mathcal{W}_2(A_n) \,.$$

This observation already implies the Moore–Penrose stability of (A_n).

THEOREM 4.6. *Let* $a \in PC_{N\times N}$ *and let the operator* $T(a)$ *be Fredholm. Consider* $(T_{n,\alpha,\beta}(a)) \in \mathcal{A}$ *with given multiindices* α, β. *If there is an* n_0 *such that*

$$\ker T(a) \subset \ \textit{im}\ P_{n_0} \ \textit{and}\ \ker \tilde{T}_{\alpha,\beta}(a) \subset \ \textit{im}\ P_{n_0} \tag{4.3}$$

or

$$\ker T^*(a) \subset \ \textit{im}\ P_{n_0} \ \textit{and}\ \ker \tilde{T}_{\alpha,\beta}(a)^* \subset \ \textit{im}\ P_{n_0} \,,$$

then the sequence $(T_{n,\alpha,\beta})$ *is Moore–Penrose stable and* $(T_{n,\alpha,\beta}^+(a))$ *converges strongly to* $T^+(a)$. *Moreover, for* $n \geq n_0 + \alpha$ *we have*

$$P_{\ker T_{n,\alpha,\beta}(a)} = P_n P_{\ker T(a)} P_n + W_n P_{\ker \tilde{T}_{\alpha,\beta}(a)} W_n \,,$$
$$P_{\ker T_{n,\alpha,\beta}^*(a)} = P_n P_{\ker T^*(a)} P_n + W_n P_{\ker \tilde{T}_{\alpha,\beta}^*(a)} W_n \,,$$

respectively.

Proof. We have to check conditions (a) and (b) of Proposition 4.5. First consider the case where the first condition in (4.3) is fulfilled.

(a) For $n \geq n_0 + \beta$ we get

$$T_{n,\alpha,\beta}(a)P_nP_{\ker T(a)}P_n = P_{n-\alpha}T(a)P_{\ker T(a)}P_n = 0$$

and

$$\begin{aligned}
&T_{n,\alpha,\beta}W_nP_{\ker V^\alpha T(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\beta}}W_n \\
&= W_n(W_nT_{n,\alpha,\beta}W_nP_{\ker V^\alpha T(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\beta}})W_n \\
&= W_n(V^\alpha P_nT(e_{-1}^\alpha \tilde{a} e_1^\beta)P_nV^{*\beta}P_{\ker V^\alpha T(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\beta}})W_n \\
&= W_n(V^\alpha P_nT(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\beta}P_{\ker V^\alpha T(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\beta}})W_n\,.
\end{aligned}$$

Because the first condition of (4.3) is valid, the operator inside the brackets is zero (notice that $P_nV^{*\beta}P_{n_0} = P_nP_{n_0}V^{*\beta}P_{n_0} = V^{*\beta}P_{n_0}$ for $n \geq n_0$). Thus (a) is fulfilled, (b) is obvious, and the sequence $(T_{n,\alpha,\beta}(a))$ is Moore–Penrose stable and the Moore–Penrose inverses converge strongly to $T^+(a)$. If the second condition is fulfilled in (4.3), then $(T^{*+}_{n,\alpha,\beta}(a))$ tends strongly to $T^{*+}(a)$. Taking adjoints we get the claim. ∎

Conjecture 4.1. Let $T(a)$ be Fredholm, $a \in PC_{N\times N}$, and the sequence $(T_{n,\alpha,\beta}(a))$ be Moore–Penrose stable. Then one of the conditions (4.3) is fulfilled.

Remark 4.1. For $N = 1$ and $\alpha = \beta = 0$ this was proved by Heinig and Hellinger in [H/H]. A more general conjecture is the following.

Conjecture 4.2. Let the first condition of Conjecture 4.1 be fulfilled. Then there is an n_0 such that for $n \geq n_0$

$$\dim\ker T_{n,\alpha,\beta}(a) = \max\{\gamma, \gamma^*\}\,,$$

where

$$\gamma = \dim\left(\text{im } P_{n_0} \cap \ker T(a)\right) + \dim\left(\text{im } P_{n_0} \cap \ker V^\alpha T(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\beta}\right)$$

and

$$\gamma^* = \dim\left(\text{im } P_n \cap \ker T^*(a)\right) + \dim\left(\text{im } P_{n_0} \cap \ker V^\beta T^*(e_{-1}^\alpha \tilde{a} e_1^\beta)V^{*\alpha}\right)\,.$$

Next we describe a sufficiently large class of Fredholm operators, $a \in PC_{N\times N}$, for which $\ker T(a) \subset \text{ im } P_{n_0}$ for some n_0. Of course each left invertible Toeplitz operator owns this property. If a is such that $(a^{-1})_m = 0$ for all sufficiently large m (here $(a^{-1})_j$ denotes the jth Fourier coefficient of a^{-1}), then $T(a)$ has the mentioned property, too. This can be easily seen by factorization.

By specifying Theorem 4.6 we get the following theorem.

Theorem 4.7. *Let $a \in PC_{N\times N}$ and $T(a)$ be Fredholm.*

(a) *If $T(a)$ is left invertible or $(a^{-1})_m = 0$ for m large enough, then there is r_0 such that $(T^+_{n,0,r}(a))$ converges strongly to $T^+(a)$ for all $r \geq r_0$.*

(b) *If $T(a)$ is right invertible or $(a^{-1})_{-m} = 0$ for m large enough, then there is a r_0 such that $(T^+_{n,r,0}(a))$ converges strongly to $T^+(a)$ for all $r \geq r_0$.*

Proof. (a) If r is large enough, then the kernel of $T(\tilde{a}e_1^r)V^{*r}$ is contained in im P_r. Now it follows that the conditions of Theorem 4.7 are fulfilled, whence the claim follows.

(b) can be reduced to (a) by taking adjoints. The theorem is completely proved. □

Remark 4.2. The same results are true if one replaces $PC_{N\times N}$ by $QC_{N\times N}$ or more generally by $PQC_{N\times N}$. QC stands here for the algebra of all quasi-continuous functions and PQC for the algebra of all piecewise quasicontinuous functions defined on $\mathbb{T}$. The reason is that all results of section 2 again hold.

Remark 4.3. One can expect that analogous results are also true for further operator classes and their approximations. This will be considered in a forthcoming paper.

5. Appendix. Here we present two examples which show that at least for smooth generating functions the kernel dimension of Fredholm Toeplitz operators can be computed effectively. These examples are given via randomly chosen factors of the Wiener–Hopf factorization

1^0

$$\begin{aligned} a(t) &= \begin{pmatrix} t^2+3t+1+\frac{7}{2}t^{-1} & t^3+t+\frac{1}{2}t^{-1}+2t^{-2} \\ t+4 & t^2+1+4t^{-1} \end{pmatrix} \\ &= \begin{pmatrix} t^{-2}+1 & \frac{1}{2}t^{-1} \\ t^{-1} & 1 \end{pmatrix} \begin{pmatrix} t & 0 \\ 0 & t^{-1} \end{pmatrix} \begin{pmatrix} t+3 & t^2 \\ t & t+4 \end{pmatrix}. \end{aligned}$$

Therefore the kernel dimension of the Toeplitz operator $T(a)$ equals 1.

2^0

$$\begin{aligned} a(t) &= \begin{pmatrix} 2t^2+7t+3+\frac{1}{2}t^{-1} & \frac{1}{2}t^{-2} \\ t+3+t^{-1} & t^{-2} \end{pmatrix} \\ &= \begin{pmatrix} t^{-1}+2 & \frac{1}{2} \\ t^{-1} & 1 \end{pmatrix} \begin{pmatrix} t & 0 \\ 0 & t^{-2} \end{pmatrix} \begin{pmatrix} t+3 & 0 \\ t & 1 \end{pmatrix}. \end{aligned}$$

Thus, $T(a)$ is Fredholm with dim ker $T(a) = 2$.

In Figures 1–4 we plotted the singular values $s_j(T_{n,0,r}(a))$ versus $1 \le n \le 70$ for the generating functions a given in Examples 1^0 and 2^0 and for $r = 0, 1$, respectively. The computations showed that in all cases d can be chosen about $\frac{1}{4}$. The number of the lower singular values which approach to zero cannot be seen because to the computer they are equal to zero. However, the computer allows us also to determine their number.

The computations show that the sequence $(T_{n,0,0}(a))$ is subject to the 1-splitting property.

The next figure is devoted to the case $r = 1$. In this case the sequence $(T_{n,0,1}(a))$ is subject to the 3-splitting property, and we observe already stabilization in the sense of the remark made after Theorem 3.2. Thus, the computations lead to dim ker $T(a) = 1$ (recall that $N = 2$).

The computations give that $(T_{n,0,0}(a))$ and $(T_{n,0,1}(a))$ have the 2- and 4-splitting property, respectively. Thus, the developed theory gives dim ker $T(a) = 2$.

The examples show that the values c_n can be taken converging very fast to zero if the generating functions are smooth. For $N = 1$ and the familiar finite sections this result is already proved in [B/S 3]. It would be of interest to have a proof in the general case.

Example 1^0.

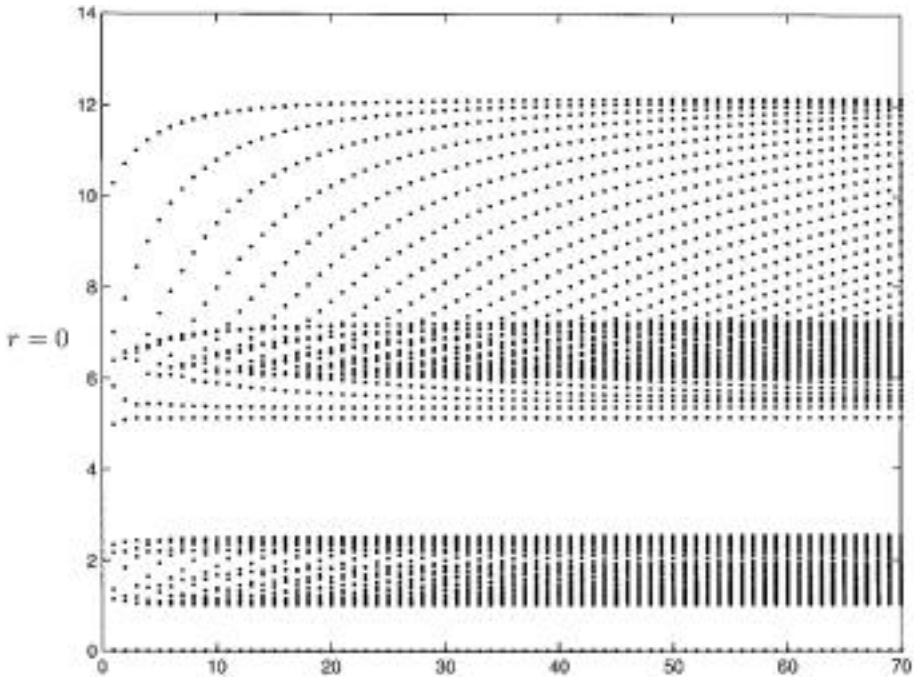


FIG. 1.

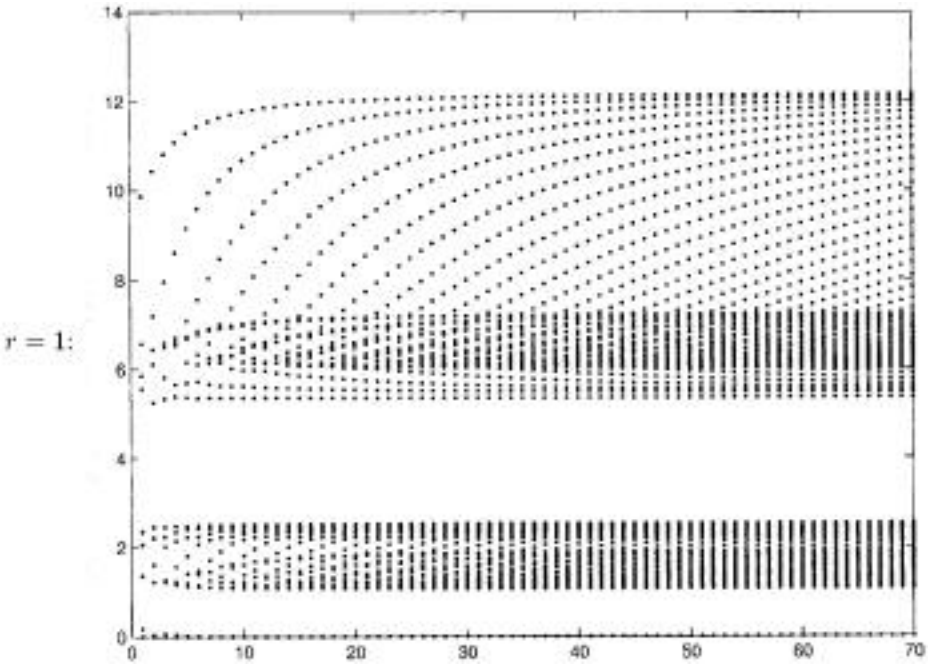


FIG. 2.

Example 2^0.

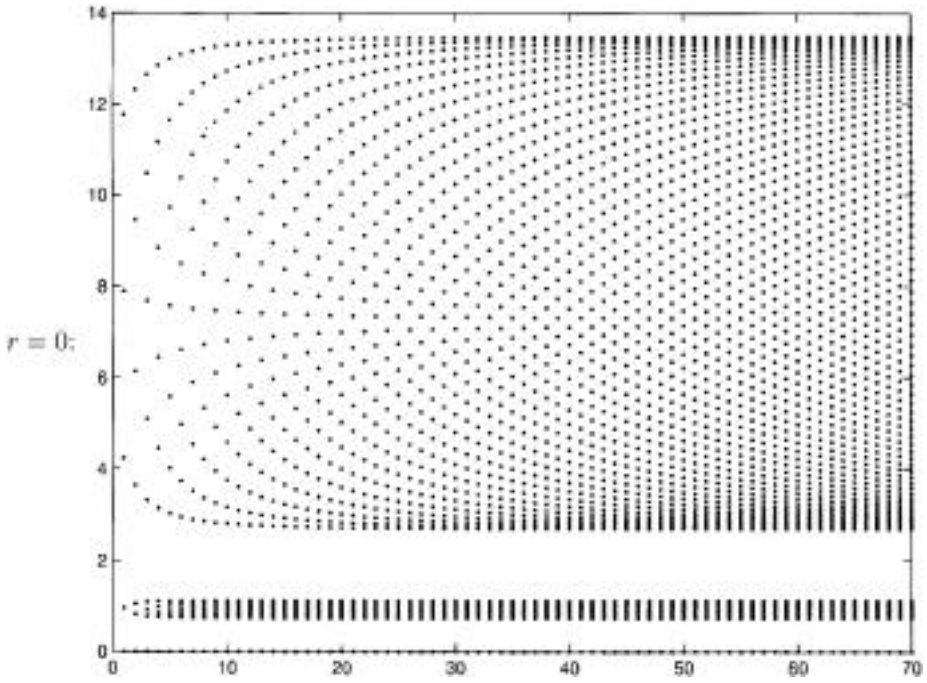


FIG. 3.

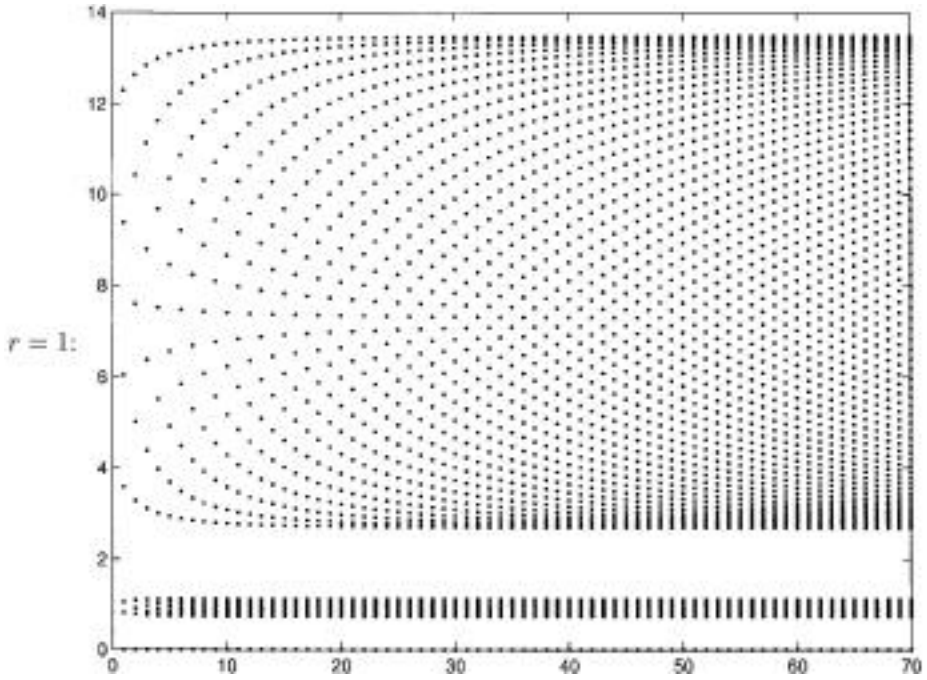


FIG. 4.

REFERENCES

[B/S 1] A. Böttcher and B. Silbermann, *The finite section method for Toeplitz operators on the quarter-plane with piecewise continuous symbols*, Math. Nachr., 110 (1983), pp. 297–291.

[B/S 2] A. Böttcher and B. Silbermann, *Analysis of Toeplitz operators*, Akademie-Verlag, Berlin, 1989, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[B/S 3] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, Berlin, Heidelberg, 1999.

[G/F] I. Gohberg and I. Feldmann, *Convolution Equations and Projection Methods for Their Solution*, Nauka, Moskva, 1971 (in Russian); Amer. Math. Soc., Translation of Math. Monogr., 41, Providence, RI 1974 (in English).

[G/K] I. Gohberg and N. Krupnik, *One-Dimensional Linear Singular Integral Operators* I. *Introduction*, Birkhäuser-Verlag, Basel, Boston, Stuttgart, 1992.

[H/H] G. Heinig and F. Hellinger, *The finite section method for Moore-Penrose inversion of Toeplitz operators*, Integral Equations Operator Theory, 19 (1994), pp. 419–446.

[H/R/S] R. Hagen, S. Roch, and B. Silbermann, *C^*-Algebras and Numerical Analysis*, Marcel Dekker, New York, Basel, 2001.

[L/S] S. Litvinchuk and I. Spitkovskii, *Factorization of Measurable Matrix Functions*, Birkhäuser-Verlag, Basel, Boston, 1987.

[R/S 1] S. Roch and B. Silbermann, *C^*-algebra techniques in numerical analysis*, J. Operator Theory, 35 (1996), pp. 241–280.

[R/S 2] S. Roch and B. Silbermann, *Index calculus for approximation methods and singular value decomposition*, J. Math. Anal. Appl., 225 (1998), pp. 401–426.

[S 1] B. Silbermann, *Local objects in the theory of Toeplitz operators*, Integral Equations Operator Theory, 9 (1986), pp. 706–738.

[S 2] B. Silbermann, *Asymptotic Moore–Penrose inversion of Toeplitz operators*, Linear Algebra Appl., 256 (1997), pp. 219–234.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 693–714

COMPUTING THE SMOOTHNESS EXPONENT OF A SYMMETRIC MULTIVARIATE REFINABLE FUNCTION*

BIN HAN†

Abstract. Smoothness and symmetry are two important properties of a refinable function. It is known that the Sobolev smoothness exponent of a refinable function can be estimated by computing the spectral radius of a certain finite matrix which is generated from a mask. However, the increase of dimension and the support of a mask tremendously increase the size of the matrix and therefore make the computation very expensive. In this paper, we shall present a simple and efficient algorithm for the numerical computation of the smoothness exponent of a symmetric refinable function with a general dilation matrix. By taking into account the symmetry of a refinable function, our algorithm greatly reduces the size of the matrix and enables us to numerically compute the Sobolev smoothness exponents of a large class of symmetric refinable functions. Step-by-step numerically stable algorithms are given. To illustrate our results by performing some numerical experiments, we construct a family of dyadic interpolatory masks in any dimension, and we compute the smoothness exponents of their refinable functions in dimension three. Several examples will also be presented for computing smoothness exponents of symmetric refinable functions on the quincunx lattice and on the hexagonal lattice.

Key words. eigenvalues of matrices, smoothness exponent, regularity, multivariate refinable functions, symmetry, interpolating functions, quincunx dilation matrix

AMS subject classifications. 42C40, 42C15, 46E35, 41A05, 41A63

PII. S0895479801390868

1. Introduction.

A $d \times d$ integer matrix M is called a *dilation matrix* if the condition $\lim_{k\to\infty} M^{-k} = 0$ holds. A dilation matrix M is *isotropic* if all of the eigenvalues of M have the same modulus. We say that a is a *mask* on $\mathbb{Z}^d$ if a is a finitely supported sequence on $\mathbb{Z}^d$ such that $\sum_{\beta\in\mathbb{Z}^d} a(\beta) = 1$. Wavelets are derived from refinable functions via a standard multiresolution technique. A *refinable function* ϕ is a solution to the refinement equation

$$\phi = |\det M| \sum_{\beta\in\mathbb{Z}^d} a(\beta)\phi(M\cdot -\beta), \tag{1.1}$$

where a is a mask and M is a dilation matrix. For a mask a on $\mathbb{Z}^d$ and a $d\times d$ dilation matrix M, it is known [2] that there exists a unique compactly supported distributional solution, denoted by ϕ_a^M throughout the paper, to the refinement equation (1.1) such that $\widehat{\phi}_a^M(0) = 1$, where the Fourier transform of $f \in L_1(\mathbb{R}^d)$ is defined to be

$$\widehat{f}(\xi) := \int_{\mathbb{R}^d} f(x)e^{-ix\cdot\xi}\,dx, \qquad \xi \in \mathbb{R}^d,$$

and can naturally be extended to tempered distributions. When the mask a and dilation matrix M are clear from the context, we write ϕ instead of ϕ_a^M for simplicity. Symmetric multivariate wavelets and refinable functions have proved to be very

*Received by the editors June 14, 2001; accepted for publication (in revised form) by M. Hanke May 15, 2002; published electronically January 23, 2003. This research was supported by NSERC Canada under grant G121210654 and by Alberta Innovation and Science REE under grant G227120136.

http://www.siam.org/journals/simax/24-3/39086.html

†Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1 (bhan@math.ualberta.ca, http://www.ualberta.ca/~bhan).

useful in many applications. For example, two-dimensional (2D) refinable functions and wavelets have been widely used in subdivision surfaces and image/mesh compression while three-dimensional (3D) refinable functions have been used in subdivision volumes, animation and video processing, etc.

For a compactly supported function ϕ in $\mathbb{R}^d$, we say that the shifts of ϕ are *stable* if, for every $\xi \in \mathbb{R}^d$, $\widehat{\phi}(\xi + 2\pi\beta) \neq 0$ for some $\beta \in \mathbb{Z}^d$. For a function $\phi \in L_2(\mathbb{R}^d)$, its *Sobolev smoothness exponent* is defined to be

$$\nu_2(\phi) := \sup\Big\{\nu \geqslant 0 \quad : \quad \int_{\mathbb{R}^d} |\widehat{\phi}(\xi)|^2 (1+|\xi|^2)^\nu \, d\xi < \infty\Big\}. \tag{1.2}$$

Smoothness is one of the most important properties of a wavelet system. Therefore, it is of great importance to have algorithms for the numerical computation of the smoothness exponent of a refinable function. Let a be a mask, and let M be a dilation matrix. We denote Π_{k-1} the set of all polynomials of total degree less than k. By convention, Π_{-1} is the empty set. We say that a satisfies the *sum rules* of order k with respect to the lattice $M\mathbb{Z}^d$ if

$$\sum_{\beta \in M\mathbb{Z}^d} a(\alpha+\beta) q(\alpha+\beta) = \sum_{\beta \in M\mathbb{Z}^d} a(\beta) q(\beta) \qquad \forall\, \alpha \in \mathbb{Z}^d,\ q \in \Pi_{k-1}.$$

Define a new sequence b from the mask a by

$$b(\alpha) := \sum_{\beta \in \mathbb{Z}^d} a(\alpha+\beta)\overline{a(\beta)}, \qquad \alpha \in \mathbb{Z}^d. \tag{1.3}$$

Let $\ell_0(\mathbb{Z}^d)$ denote the linear space of all finitely supported sequences on $\mathbb{Z}^d$. For a subset K of $\mathbb{Z}^d$, by $\ell(K)$ we denote the linear space of all finitely supported sequences on $\mathbb{Z}^d$ that vanish outside the set K.

The *transition operator* $T_{b,M}$ associated with the sequence b and the dilation matrix M is defined by

$$[T_{b,M} u](\alpha) = |\det M| \sum_{\beta \in \mathbb{Z}^d} b(M\alpha - \beta) u(\beta), \qquad \alpha \in \mathbb{Z}^d,\ u \in \ell_0(\mathbb{Z}^d). \tag{1.4}$$

Let $\phi \in L_2(\mathbb{R}^d)$ be a refinable function with a finitely supported mask a and a dilation matrix M such that the shifts of ϕ are stable and a satisfies the sum rules of order k but not $k+1$. Define the set $\Omega_{b,M}$ by

$$\begin{aligned} &\Omega_{b,M} := \Big[\sum_{j=1}^{\infty} M^{-j} K\Big] \cap \mathbb{Z}^d \quad \text{and} \\ &K := \{\alpha \in \mathbb{Z}^d \ :\ |\alpha| \leqslant k\} \cup \{\alpha \in \mathbb{Z}^d \ :\ b(\alpha) \neq 0\}, \end{aligned} \tag{1.5}$$

and define the slightly smaller subspace V_{2k-1} of $\ell(\Omega_{b,M})$ to be

$$V_j := \Big\{u \in \ell(\Omega_{b,M}) \ :\ \sum_{\beta \in \mathbb{Z}^d} u(\beta) q(\beta) = 0 \qquad \forall\, q \in \Pi_j\Big\}, \qquad j \in \mathbb{N}_0. \tag{1.6}$$

When M is isotropic, it was demonstrated in [4, 5, 6, 10, 21, 23, 24, 26, 27, 33, 35] in various forms under various conditions that

$$\nu_2(\phi) = -\frac{d}{2} \log_{|\det M|} \rho(T_{b,M}|_{V_{2k-1}}), \tag{1.7}$$

where $\rho(T_{b,M}|_{V_{2k-1}})$ is the spectral radius of the operator $T_{b,M}$ acting on the finite dimensional $T_{b,M}$-invariant subspace V_{2k-1} of $\ell(\Omega_{b,M})$.

However, from the point of view of numerical computation, there are some difficulties in obtaining the Sobolev smoothness exponent of a refinable function via (1.7) by computing the quantity $\rho(T_{b,M}|_{V_{2k-1}})$ due to the following considerations:

D1. It is not easy to find a simple basis for the space V_{2k-1} by a numerically stable procedure to obtain a representation matrix of $T_{b,M}$ under such a basis. Theoretically speaking, if some elements in a numerically found basis of V_{2k-1} cannot satisfy the equality in (1.6) exactly, then it will dramatically change the spectral radius since in general $T_{b,M}$ has significantly larger eigenvalues outside the subspace V_{2k-1}.

D2. When the dimension is greater than one and even when the mask has a relatively small support, in general, the dimensions of the spaces V_{2k-1} and $\ell(\Omega_{b,M})$ are very large. For example, for a 3D mask with support $[-7,7]^3$ and sum rules of order 4, we have $\dim(V_7) = 24269$ and $\dim(\ell(\Omega_{b,2I_3})) = 24389$. This makes the numerical computation using (1.7) very expensive or even impossible.

D3. In order to obtain the exact Sobolev smoothness exponent by (1.7), we have to check the assumption that the shifts of ϕ_a^M are stable, which is a far from trivial condition to be verified.

Fortunately, the difficulty in D1 was successfully overcome in Jia and Zhang [25], where they demonstrated that $\rho(T_{b,M}|_{V_{2k-1}})$ is the largest value in modulus in the set consisting of all of the eigenvalues of $T_{b,M}|_{\ell(\Omega_{b,M})}$, excluding some known special eigenvalues. Note that $\ell(\Omega_{b,M})$ has a simple basis $\{\delta_\alpha : \alpha \in \Omega_{b,M}\}$, where $\delta_\alpha(\alpha) = 1$ and $\delta_\alpha(\beta) = 0$ for all $\beta \in \mathbb{Z}^d \backslash \{\alpha\}$.

On the other hand, both symmetry and smoothness of a wavelet basis are very important and much desired properties in many applications. It is one of the purposes of this paper to try to overcome the difficulty in D2 for a symmetric refinable function. We shall demonstrate in Algorithm 2.1 that we can compute the Sobolev smoothness exponent of a symmetric refinable function by using a much smaller space than the space $\ell(\Omega_{b,M})$. In section 3, we shall see that, for many refinable functions, it is not necessary to directly verify the stability assumption since they are already implicitly implied by the computation. Therefore, the difficulty in D3 does not exist at all for many refinable functions. (Almost all interesting known examples fall into this class.)

To give the reader some idea of how symmetry can be of help in computing the Sobolev smoothness exponents of symmetric refinable functions, we give the following comparison result in Table 1. See section 2 for more detail and explanation of Table 1.

TABLE 1

The last two rows indicate the matrix sizes in computing the Sobolev smoothness exponents of symmetric refinable functions using both the method in [25] *and the method in Algorithm* 2.1 *in section* 2 *of this paper. This table demonstrates that Algorithm* 2.1 *can greatly reduce the size of the matrix in computing the Sobolev smoothness exponent of a symmetric refinable function.*

Mask	4D mask	3D mask	2D mask	2D mask	2D mask
Support	$[-5,5]^4$	$[-7,7]^3$	$[-27,27]^2$	$[-7,7]^2$	$[-12,12]^2$
Symmetry	full axes	full axes	hexagonal	full axes	hexagonal
Dilation matrix	$2I_4$	$2I_3$	$2I_2$	$\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -2\\ 2 & -1\end{bmatrix}$
Method in [25]	194481	24389	8911	5601	$\geqslant 3241$
Algorithm 2.1	715	560	756	707	294

Masks and refinable functions with extremely large supports rarely may be used in real world applications. For a given mask which is of interest in applications, very often there are some free parameters in the mask, and one needs to optimize the smoothness exponent of its refinable function [9, 12, 15, 17, 28, 30]. The efficient algorithms proposed in this paper will be of help for such a smoothness optimization problem. On the other hand, a refinable function vector satisfies the refinement equation (1.1) with a matrix mask of multiplicity r. A matrix mask of multiplicity r is a sequence of $r \times r$ matrices on $\mathbb{Z}^d$. (Masks discussed in this paper correspond to $r = 1$ and are called scalar masks.) Very recently, as demonstrated in [17], multivariate refinable function vectors with short support and symmetry are of interest in computer aided geometric design (CAGD) and in numerical solutions to partial differential equations. Let M be the quincunx dilation matrix (the fourth dilation matrix in Table 1), and let a be a matrix mask of multiplicity 3 with support $[-1, 1]^2$. (Hermite interpolatory masks of order 1, discussed in [17], are examples of such masks which often have many free parameters and are useful in CAGD.) In order to compute the Sobolev smoothness exponent of its refinable function vector with such a small mask, without using symmetry, we found that one has to deal with a 1161×1161 matrix (see also [23]). As a consequence, even in low dimensions and for masks with small supports, it is very important to take into account the symmetry of a refinable function (vector) in algorithms for the numerical computation of its smoothness exponent. Though we consider only scalar masks here for simplicity, results in this paper can be generalized to matrix masks and refinable function vectors which will be discussed elsewhere.

The structure of the paper is as follows. In section 2, we shall present step-by-step numerically stable and efficient algorithms for the numerical computation of the Sobolev smoothness exponent of a symmetric refinable function. In addition, an algorithm for computing the Hölder smoothness exponent of a symmetric refinable function will be given in section 2, provided that the symbol of its mask is nonnegative. In section 3, we shall study the relation of the spectral radius of a certain operator acting on different spaces. Such analysis enables us to overcome the difficulty in D3 for a large class of masks. In section 4, we shall apply the results in sections 2 and 3 to several examples, including refinable functions on quincunx lattice and hexagonal lattice. We shall also present a C^2 $\sqrt{3}$-interpolatory subdivision scheme in section 4. Next, we shall generalize the well-known univariate interpolatory masks in Deslauriers and Dubuc [8] and the bivariate interpolatory masks in [15] to any dimension. Finally, we shall use the results in sections 2 and 3 to compute Sobolev smoothness exponents of interpolating refinable functions associated with such interpolatory masks in dimension three.

Programs can be downloaded at http://www.ualberta.ca/~bhan for computing the Sobolev and Hölder smoothness exponents of symmetric refinable functions based on the Algorithms 2.1 and 2.5 in section 2. However, such programs come without warranty and are not yet optimized with respect to user interface.

2. Computing smoothness exponent using symmetry. In this section, taking into account the symmetry, we shall present an efficient algorithm for the numerical computation of the Sobolev smoothness exponent of a symmetric multivariate refinable function with a general dilation matrix. As the main result in this section, Algorithms 2.1 and 2.5 are quite simple and can be easily implemented, though their proofs and some notation are relatively technical.

Before proceeding further, let us introduce some notation and necessary background. Let $\mathbb{N}_0$ denote all of the nonnegative integers. For $\mu = (\mu_1, \dots, \mu_d) \in \mathbb{N}_0^d$,

$|\mu| := \mu_1 + \cdots + \mu_d$, $\mu! := \mu_1! \cdots \mu_d!$, and $\xi^\mu := \xi_1^{\mu_1} \cdots \xi_d^{\mu_d}$ for $\xi = (\xi_1, \ldots, \xi_d) \in \mathbb{R}^d$. For $\alpha \in \mathbb{Z}^d$ and $y \in \mathbb{R}^d$, we define

$$\nabla_\alpha u := u - u(\cdot - \alpha), \quad \nabla_y f := f - f(\cdot - y), \qquad u \in \ell_0(\mathbb{Z}^d),\ f \in L_p(\mathbb{R}^d).$$

For $\mu = (\mu_1, \ldots, \mu_d) \in \mathbb{N}_0^d$, $\nabla^\mu := \nabla_{e_1}^{\mu_1} \cdots \nabla_{e_d}^{\mu_d}$, where e_j is the jth coordinate unit vector in $\mathbb{R}^d$. Let $\delta = \delta_0$ denote the sequence such that $\delta(0) = 1$ and $\delta(\beta) = 0$ for all $\beta \in \mathbb{Z}^d \backslash \{0\}$. For $u \in \ell_0(\mathbb{Z}^d)$, its ℓ_p norm is defined to be $\|u\|_p := (\sum_{\beta \in \mathbb{Z}^d} |u(\beta)|^p)^{1/p}$. Let M be a $d \times d$ dilation matrix, and let a be a mask on $\mathbb{Z}^d$. Define the *subdivision operator* $S_{a,M} : \ell_0(\mathbb{Z}^d) \mapsto \ell_0(\mathbb{Z}^d)$ by

$$[S_{a,M} u](\alpha) := |\det M| \sum_{\beta \in \mathbb{Z}^d} a(\alpha - M\beta) u(\beta), \qquad \alpha \in \mathbb{Z}^d,\ u \in \ell_0(\mathbb{Z}^d).$$

For $1 \leqslant p \leqslant \infty$ and $k \in \mathbb{N}_0$, we define

$$\rho_k(a; M, p) := \max\Big\{ \lim_{n \to \infty} \|\nabla^\mu S_{a,M}^n \delta\|_p^{1/n} \ :\ |\mu| = k,\ \mu \in \mathbb{N}_0^d \Big\}. \tag{2.1}$$

Let M be a dilation matrix, and let λ_{max} be the spectral radius of M. (When M is isotropic, then $\lambda_{max} = |\det M|^{1/d}$.) When a mask a satisfies the sum rules of order k but not $k+1$, we define the following important quantity:

$$\nu_p(a; M) := -\log_{\lambda_{max}} \big[\, |\det M|^{-1/p} \rho_k(a; M, p) \,\big], \qquad 1 \leqslant p \leqslant \infty. \tag{2.2}$$

The above quantity $\nu_p(a; M)$ plays a very important role in characterizing the convergence of a subdivision scheme in a Sobolev space and in characterizing the L_p smoothness exponent of a refinable function.

The L_p smoothness of $f \in L_p(\mathbb{R}^d)$ is measured by its L_p *smoothness exponent*:

$$\begin{aligned} \nu_p(f) := \sup\{\nu \geqslant 0 \quad :\quad & \|\nabla_y^n f\|_p \leqslant C \|y\|^\nu \ \forall\ y \in \mathbb{R}^d, \quad \text{for some} \\ & \text{constant } C \ \text{ and for large enough positive integer } n\}. \end{aligned} \tag{2.3}$$

When $p = 2$, the above definition of $\nu_2(f)$ agrees with the definition in (1.2). By generalizing the results in [4, 5, 6, 10, 12, 21, 24, 26, 27, 32, 33, 35] and references therein, we have

$$\nu_p(\phi_a^M) \geqslant \nu_p(a; M), \qquad 1 \leqslant p \leqslant \infty,$$

and the equality holds when the shifts of ϕ_a^M are stable and M is an isotropic dilation matrix. When M is a general dilation matrix and the shifts of ϕ_a^M are stable, as demonstrated in [5] for the case when $p = 2$, one can have only the estimate

$$-\log_{\lambda_{max}} \big[\, |\det M|^{-1/p} \rho_k(a; M, p) \,\big] \leqslant \nu_p(\phi_a^M) \leqslant -\log_{\lambda_{min}} \big[\, |\det M|^{-1/p} \rho_k(a; M, p) \,\big],$$

where $\lambda_{min} := \min_{1 \leqslant j \leqslant d} |\lambda_j|$ and $\lambda_{max} := \max_{1 \leqslant j \leqslant d} |\lambda_j|$ with λ_j, $j = 1, \ldots, d$, being all of the eigenvalues of M. As pointed out in [5], the usual Sobolev smoothness defined in (1.2) and (2.3) is closely related to isotropic dilations, and anisotropic Sobolev spaces are needed in the case of an anisotropic dilation matrix. See [5] for more detail on this issue.

So, to compute the Sobolev smoothness exponent of a refinable function, we need to compute $\nu_2(a; M)$ and therefore to compute $\rho_k(a; M, 2)$. It is the purpose of this section to discuss how to efficiently compute $\rho_k(a; M, 2)$ when a is a symmetric mask.

Let Θ be a finite subset of integer matrices whose determinants are ± 1. We say that Θ is a *symmetry group* with respect to a dilation matrix M (see [13]) if Θ forms a group under matrix multiplication and $M\theta M^{-1} \in \Theta$ for all $\theta \in \Theta$. Obviously, each element in a symmetry group induces a linear isomorphism on $\mathbb{Z}^d$.

Let Θ_d^A denote the set of all linear transforms on $\mathbb{Z}^d$ which are given by

$$\theta_{\pi,\varepsilon}(\alpha_1, \dots, \alpha_d) := (\varepsilon_1 \alpha_{\pi(1)}, \dots, \varepsilon_d \alpha_{\pi(d)}), \qquad (\alpha_1, \dots, \alpha_d) \in \mathbb{Z}^d, \tag{2.4}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d) \in \{-1, 1\}^d$ and π is a permutation on $(1, \dots, d)$. Θ_d^A is called the *full axes symmetry group*. Obviously, Θ_d^A is a symmetry group with respect to the dilation matrix $2I_d$. It is also easy to check that Θ_2^A is a symmetry group with respect to the quincunx dilation matrices

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Another symmetry group with respect to $2I_2$ is the following group, which is called the *hexagonal symmetry group*:

$$\Theta^H = \left\{ \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \pm \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}, \pm \begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix}, \pm \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \pm \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix}, \pm \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix} \right\}. \tag{2.5}$$

Such a group Θ^H can be used to obtain wavelets on the hexagonal planar lattice (that is, the triangular mesh). For a symmetry group Θ and a sequence u on $\mathbb{Z}^d$, we define a new sequence $\Theta(u)$ as follows:

$$[\Theta(u)](\beta) := \frac{1}{\#\Theta} \sum_{\theta \in \Theta} u(\theta\beta), \qquad \beta \in \mathbb{Z}^d,\ u \in \ell_0(\mathbb{Z}^d), \tag{2.6}$$

where $\#\Theta$ denotes the cardinality of the set Θ. We say that a mask a is *invariant* under Θ if $\Theta(a) = a$. Obviously, for any sequence u, $\Theta(u)$ is invariant under Θ since $\Theta(\Theta(u)) = \Theta(u)$. When Θ is a symmetry group with respect to a dilation matrix M, then the fact that a is invariant under Θ implies that the refinable function ϕ_a^M is also invariant under Θ; that is, $\phi_a^M(\theta\cdot) = \phi_a^M$ for all $\theta \in \Theta$. See Han [13] for a detailed discussion on the symmetry property of multivariate refinable functions. We caution the reader that the condition $M\theta M^{-1} \in \Theta$ for all $\theta \in \Theta$ cannot be removed in the definition of a symmetry group with respect to a dilation matrix M. For example, as a subgroup of Θ_2^A,

$$\Theta = \left\{ \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \pm \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

is not a symmetry group with respect to the quincunx dilation matrices, though it is a symmetry group with respect to the dilation matrix $2I_2$. So, even when a mask a is invariant under such a group Θ, the refinable function ϕ_a^M with the quincunx dilation matrices may not be invariant under Θ.

Let $\mathbb{Z}_\Theta^d$ denote a subset of $\mathbb{Z}^d$ such that, for every $\alpha \in \mathbb{Z}^d$, there exists a unique $\beta \in \mathbb{Z}_\Theta^d$ satisfying $\theta\beta = \alpha$ for some $\theta \in \Theta$. In other words, $\mathbb{Z}_\Theta^d$ is a complete set of representatives of the distinct cosets of $\mathbb{Z}^d$ under the equivalence relation induced by Θ on $\mathbb{Z}^d$.

Taking into account the symmetry of a mask, now we have the following algorithm for the numerical computation of the important quantity $\nu_2(a; M)$.

ALGORITHM 2.1. *Let M be a $d \times d$ isotropic dilation matrix, and let Θ be a symmetry group with respect to the dilation matrix M. Let a be a mask on $\mathbb{Z}^d$ such that $\sum_{\beta\in\mathbb{Z}^d} a(\beta) = 1$. Define the sequence b as in (1.3). Suppose that b is invariant under the symmetry group Θ and that a satisfies the sum rules of order k but not $k+1$. The quantity $\nu_2(a; M)$, or, equivalently, $\rho_k(a; M, 2)$, is obtained via the following procedure:*

(a) *Find a finite subset K_Θ of $\mathbb{Z}^d_\Theta$ such that*

$$\{M^{-1}(\theta\alpha+\beta) \,:\, \theta\in\Theta,\, \alpha\in K_\Theta,\, \beta\in \operatorname{supp} b\}\cap\mathbb{Z}^d \subseteq \{\theta\beta \,:\, \beta\in K_\Theta,\, \theta\in\Theta\}$$

and

$$\dim(\Pi_{2k-1}|_{\{\theta\beta\,:\,\theta\in\Theta,\,\beta\in K_\Theta\}}) = \dim(\Pi_{2k-1}).$$

(b) *Obtain a $(\#K_\Theta)\times(\#K_\Theta)$ matrix T as follows:*

$$T[\alpha,\beta] := \frac{|\det M|}{\#\{\theta\in\Theta \,:\, \theta\beta=\beta\}} \sum_{\theta\in\Theta} b(M\beta-\theta\alpha), \qquad \alpha,\beta\in K_\Theta. \tag{2.7}$$

(c) *Let $\sigma(T)$ consist of the absolute values of all of the eigenvalues of the square matrix T counting the multiplicity of its eigenvalues. Then $\nu_2(a; M)$ is the smallest number in the following set:*

$$\left\{-\frac{d}{2}\log_{|\det M|}\rho \,:\, \rho\in\sigma(T)\right\}\Big\backslash \{j/2 \text{ with positive multiplicity } m_\Theta(j) \,:\, 0\leqslant j<2k\}, \tag{2.8}$$

where by default $\log_{|\det M|} 0 := -\infty$ and

$$m_\Theta(j) := \dim(\Theta(\Pi_j)) - \dim(\Theta(\Pi_{j-1})), \qquad j\in\mathbb{N}_0. \tag{2.9}$$

Before we give a proof of Algorithm 2.1, let us make some remarks and discuss how to compute the set K_Θ and the quantities $m_\Theta(j)$ in Algorithm 2.1. Since the matrix T in Algorithm 2.1 has a simple structure, it is not necessary to store the whole matrix T in order to compute its eigenvalues, and many techniques from numerical analysis (such as the subspace iteration method and Arnoldi's method as discussed in [34]) can be exploited to further improve the efficiency in computing the eigenvalues of T. We shall not discuss such an issue here. One satisfactory set K_Θ can easily be obtained as follows.

PROPOSITION 2.2. *Let $K_0 := \operatorname{supp} b \cup \{\theta\alpha\in\mathbb{Z}^d \,:\, |\alpha|\leqslant k, \theta\in\Theta\}$, where $\operatorname{supp} b := \{\beta\in\mathbb{Z}^d \,:\, b(\beta)\neq 0\}$. Recursively compute*

$$K_j := K_{j-1}\cup\big[\big(M^{-1}(K_{j-1}+\operatorname{supp} b)\big)\cap\mathbb{Z}^d\big], \qquad j\in\mathbb{N}.$$

Then $K_j = K_{j-1}$ for some $j\in\mathbb{N}$. Set $K_\Theta := K_j\cap\mathbb{Z}^d_\Theta$. Then K_Θ satisfies all the conditions in (a) *of Algorithm* 2.1.

Proof. Note that $K_j\subseteq(\sum_{j=1}^\infty M^{-j}K_0)\cap\mathbb{Z}^d\subseteq\{\alpha\in\mathbb{Z}^d \,:\, |\alpha|<r\}$ for some finite integer r. Therefore, there must exist $j\in\mathbb{N}$ such that $K_j = K_{j-1}$ by $K_{i-1}\subseteq K_i$ for all $i\in\mathbb{N}$. Consequently,

$$M^{-1}(K_j+\operatorname{supp} b)\cap\mathbb{Z}^d = M^{-1}(K_{j-1}+\operatorname{supp} b)\cap\mathbb{Z}^d \subseteq K_j.$$

Since $K_0 \subseteq K_j$, we have

$$\dim(\Pi_{2k-1}) = \dim(\Pi_{2k-1}|_{K_0}) \leqslant \dim(\Pi_{2k-1}|_{\{\theta\beta\,:\,\theta\in\Theta,\,\beta\in K_\Theta\}}) \leqslant \dim(\Pi_{2k-1}). \qquad \square$$

Let $O_j := \{\mu \in \mathbb{N}_0^d \ : \ |\mu| = j\}$. The set O_j can be ordered according to the lexicographic order. That is, $(\nu_1, \dots, \nu_d)$ is less than $(\mu_1, \dots, \mu_d)$ in lexicographic order if $\nu_j = \mu_j$ for $j = 1, \dots, i-1$ and $\nu_i < \mu_i$ for some i. For a $d \times d$ matrix A and any $j \in \mathbb{N}_0$, we define a $(\#O_j) \times (\#O_j)$ matrix $S(A, j)$, which is uniquely determined by

$$\frac{(Ax)^\mu}{\mu!} = \sum_{\nu\in O_j} [S(A,j)]_{\mu,\nu} \frac{x^\nu}{\nu!}, \qquad \mu \in O_j,\ j \in \mathbb{N}_0. \tag{2.10}$$

It is easy to verify that $S(AB, j) = S(A, j)S(B, j)$. When $\lambda_1, \dots, \lambda_d$ are all of the eigenvalues of A, then λ^μ, $\mu \in O_j$, are all of the eigenvalues of $S(A, j)$, where $\lambda = (\lambda_1, \dots, \lambda_d)$, since $S(A, j)$ is similar to $S(B, j)$ when A is similar to B. Moreover, $\mu! S(A^T, j)_{\mu,\nu} = \nu! S(A, j)_{\nu,\mu}$ for all $\mu, \nu \in O_j$ and $j \in \mathbb{N}_0$ by comparing the Taylor series of the same function $e^{x^T Ay}$ and $e^{y^T A^T x}$.

The quantities $m_\Theta(j)$, $j \in \mathbb{N}_0$, can be computed as follows.

PROPOSITION 2.3. *Let Θ be a symmetry group. Then*

$$m_\Theta(j) = \operatorname{rank}\bigg[\sum_{\theta\in\Theta} S(\theta, j)\bigg], \qquad j \in \mathbb{N}_0. \tag{2.11}$$

In particular, when $-I_d \in \Theta$, then $m_\Theta(2j-1) = 0$ for all $j \in \mathbb{N}$.

Proof. For $\mu \in \mathbb{N}_0^d$, let q_μ be the sequence given by $q_\mu(\alpha) = \alpha^\mu/\mu!$, $\alpha \in \mathbb{Z}^d$. Note that

$$\begin{aligned}(\#\Theta)[\Theta(q_\mu)](\alpha) &= \sum_{\theta\in\Theta} q_\mu(\theta\alpha) = \sum_{\theta\in\Theta} \frac{(\theta\alpha)^\mu}{\mu!} \\ &= \sum_{\theta\in\Theta} \sum_{\nu\in O_j} [S(\theta,j)]_{\mu,\nu} \frac{\alpha^\nu}{\nu!} \\ &= \sum_{\nu\in O_j} q_\nu(\alpha) \sum_{\theta\in\Theta} [S(\theta,j)]_{\mu,\nu}.\end{aligned}$$

Since $q_\mu, \mu \in O_j$ are linearly independent, we have

$$\begin{aligned}m_\Theta(j) &= \dim(\Theta(\Pi_j)) - \dim(\Theta(\Pi_{j-1})) \\ &= \dim(\operatorname{span}\{\Theta(q_\mu) \ : \ \mu \in O_j\}) \\ &= \operatorname{rank}\bigg[\sum_{\theta\in\Theta} S(\theta, j)\bigg].\end{aligned}$$

When $-I_d \in \Theta$, we observe that

$$\sum_{\theta\in\Theta} S(\theta, j) = \sum_{\theta\in\Theta} S(-\theta, j) = (-1)^j \sum_{\theta\in\Theta} S(\theta, j).$$

Therefore, $m_\Theta(2j-1) = 0$ for all $j \in \mathbb{N}$ since $\sum_{\theta\in\Theta} S(\theta, 2j-1) = 0$. $\square$

Note that $m_\Theta(j)$ depends only on the symmetry group Θ and is independent of the dilation matrix M. If Θ is a subgroup of the full axes symmetry group Θ_d^A,

then $m_\Theta(j)$ can easily be determined since the matrix $S(\theta, j)$ is very simple for every $\theta \in \Theta_d^A$. For example,

$$m_{\Theta_d^A}(2j) = \#\{(\mu_1, \dots, \mu_d) \in \mathbb{N}_0^d \,:\, 0 \leqslant \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_d,\ \mu_1 + \cdots + \mu_d = j\}.$$

For the convenience of the reader, we list the quantities $m_\Theta(j)$ in Algorithm 2.1 for some well-known symmetry groups in Table 2. In Table 2, the symmetry groups Θ_2^1 and Θ_2^2 are defined to be

$$\Theta_2^1 = \left\{ \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \pm \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \right\}, \qquad \Theta_2^2 = \left\{ \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \pm \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right\}.$$

TABLE 2

The quantities $m_\Theta(j)$, $j \in \mathbb{N}_0$, in Algorithm 2.1 for some known symmetry groups. Note that $m_\Theta(2j-1) = 0$, $j \in \mathbb{N}$, in this table.

	$m_\Theta(j)$, $j = 0, 2, 4, \dots, 32$																
j	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32
Θ_1^A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Θ_2^A	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
Θ^H	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6
Θ_2^1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Θ_2^2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Θ_3^A	1	1	2	3	4	5	7	8	10	12	14	16	19	21	24	27	30
Θ_4^A	1	1	2	3	5	6	9	11	15	18	23	27	34	39	47	54	64

For a sequence u on $\mathbb{Z}^d$, its *symbol* is given by

$$\widehat{u}(\xi) = \sum_{\beta \in \mathbb{Z}^d} u(\beta) e^{-i\beta \cdot \xi}, \qquad \xi \in \mathbb{R}^d. \tag{2.12}$$

For $j = 1, \dots, d$, let Δ_j denote the difference operator given by

$$\Delta_j u := -u(\cdot - e_j) + 2u - u(\cdot + e_j), \qquad u \in \ell_0(\mathbb{Z}^d),$$

and $\Delta^\mu := \Delta_1^{\mu_1} \cdots \Delta_d^{\mu_d}$ for $\mu = (\mu_1, \dots, \mu_d) \in \mathbb{N}_0^d$. Define

$$\langle u, v \rangle := \sum_{\beta \in \mathbb{Z}^d} u(\beta) \overline{v(\beta)}, \qquad u, v \in \ell_0(\mathbb{Z}^d).$$

To prove Algorithm 2.1, we need the following result.

THEOREM 2.4. *Let a be a finitely supported mask on $\mathbb{Z}^d$, and let b be the sequence defined in* (1.3). *Let Θ be a symmetry group with respect to a dilation matrix M. Suppose that b is invariant under Θ. Then $\Theta(T_{b,M} u) = T_{b,M} \Theta(u)$ for all $u \in \ell_0(\mathbb{Z}^d)$, and*

$$\rho_k(a; M, 2) = |\det M|^{1/2} \sqrt{\rho(T_{b,M}|_{W_k})}, \qquad k \in \mathbb{N}_0, \tag{2.13}$$

where $T_{b,M}$ is the transition operator defined in (1.4) *and W_k is the minimal $T_{b,M}$-invariant finite dimensional space which is generated by $\Theta(\Delta^\mu \delta)$, $\mu \in \mathbb{N}_0^d$, with $|\mu| = k$.*

Proof. Since Θ is a symmetry group with respect to the dilation matrix M and b is invariant under Θ, for $u \in \ell_0(\mathbb{Z}^d)$, we have

$$\begin{aligned}\Theta(T_{b,M}u)(\alpha) &= \frac{|\det M|}{\#\Theta}\sum_{\beta\in\mathbb{Z}^d}\sum_{\theta\in\Theta} b(M\theta M^{-1}M\alpha-\beta)u(\beta)\\ &= \frac{|\det M|}{\#\Theta}\sum_{\beta\in\mathbb{Z}^d}\sum_{\theta\in\Theta} b(M\alpha-\beta)u(\theta\beta)\\ &= |\det M|\sum_{\beta\in\mathbb{Z}^d} b(M\alpha-\beta)\Theta(u)(\beta).\end{aligned}$$

Therefore, $\Theta(T_{b,M}u) = T_{b,M}\Theta(u)$ for all $u \in \ell_0(\mathbb{Z}^d)$.

Note that $\widehat{b}(\xi) = |\widehat{a}(\xi)|^2 \geqslant 0$ for all $\xi \in \mathbb{R}^d$. Let $m := |\det M|$. By the Parseval identity, we have

$$\begin{aligned}\|\nabla^\mu S^n_{a,M}\delta\|_2^2 &= \frac{1}{(2\pi)^d}\int_{[0,2\pi)^d} |\widehat{\nabla^\mu S^n_{a,M}\delta}(\xi)|^2\,d\xi\\ &= \frac{m^n}{(2\pi)^d}\int_{[0,2\pi)^d} \widehat{\Delta^\mu S^n_{b,M}\delta}(\xi)\,d\xi\\ &= m^n\Delta^\mu S^n_{b,M}\delta(0).\end{aligned}$$

From the definition of the transition operator, it is easy to verify that

$$T^n_{b,M}\Delta^\mu\delta(0) = \langle T^n_{b,M}\Delta^\mu\delta,\delta\rangle = \langle\Delta^\mu\delta, S^n_{b,M}\delta\rangle = \langle\delta,\Delta^\mu S^n_{b,M}\delta\rangle = \Delta^\mu S^n_{b,M}\delta(0).$$

For a sequence u such that $\widehat{u}(\xi) \geqslant 0$ for all $\xi \in \mathbb{R}^d$, we observe that $\|u\|_\infty = u(0)$ (see [11]). From the facts that $T^n_{b,M}\widehat{\Theta(\Delta^\mu\delta)}(\xi) \geqslant 0$ and $\widehat{\Delta^\mu S^n_{b,M}\delta}(\xi) \geqslant 0$ for all $\xi \in \mathbb{R}^d$, it follows that

$$\begin{aligned}\|T^n_{b,M}\Theta(\Delta^\mu\delta)\|_\infty &= T^n_{b,M}\Theta(\Delta^\mu\delta)(0) = \Theta(T^n_{b,M}\Delta^\mu\delta)(0)\\ &= T^n_{b,M}\Delta^\mu\delta(0) = \Delta^\mu S^n_{b,M}\delta(0) = m^{-n}\|\nabla^\mu S^n_{a,M}\delta\|_2^2.\end{aligned}$$

Since W_k is the minimal $T_{b,M}$-invariant subspace generated by

$$\{\Theta(\Delta^\mu\delta)\ :\ \mu\in\mathbb{N}_0^d,\ |\mu| = k\},$$

we have

$$\begin{aligned}\rho(T_{b,M}|_{W_k}) &= \max\Big\{\lim_{n\to\infty}\|T^n_{b,M}\Theta(\Delta^\mu\delta)\|_\infty^{1/n}\ :\ |\mu| = k,\ \mu\in\mathbb{N}_0^d\Big\}\\ &= \max\Big\{\lim_{n\to\infty} m^{-1}\|\nabla^\mu S^n_{a,M}\delta\|_2^{2/n}\ :\ |\mu| = k,\ \mu\in\mathbb{N}_0^d\Big\}\\ &= m^{-1}\big(\rho_k(a;M,2)\big)^2,\end{aligned}$$

which completes the proof. □

Proof of Algorithm 2.1. Let $K := \{\theta\beta\ :\ \theta\in\Theta,\ \beta\in K_\Theta\}$. Then it is easy to check that both $\ell(K)$ and $\Theta(\ell(K))$ are invariant under $T_{b,M}$ (see [14, Lemma 2.3]). Since a satisfies the sum rules of order k, then the sequence b, which is defined in (1.3), satisfies the sum rules of order $2k$ and V_{2k-1} is invariant under $T_{b,M}$ (see [20, Theorem 5.2]), where

$$V_j := \Big\{u\in\ell_0(\mathbb{Z}^d)\ :\ \sum_{\beta\in\mathbb{Z}^d} u(\beta)q(\beta) = 0\quad\forall\ q\in\Pi_j\Big\}.$$

Define $U_j := \Theta(\ell(K) \cap V_j)$, $j \in \mathbb{N} \cup \{0, -1\}$. Let W_k denote the linear space in Theorem 2.4. Observe that $W_k \subseteq U_{2k-1} \subseteq V_{2k-1}$. By Theorem 2.4 and (1.7), we have $\rho_k(a; M, 2) = \sqrt{|\det M| \rho(T_{b,M}|_{U_{2k-1}})}$.

Since b satisfies the sum rules of order $2k$ and b is invariant under Θ, we have $T_{b,M} U_j \subseteq U_j$ for all $j = -1, 0, \dots, 2k-1$. Therefore, we have $\operatorname{spec}(T_{b,M}|_{\Theta(\ell(K))}) = \operatorname{spec}(T_{b,M}|_{U_{2k-1}}) \cup \operatorname{spec}(T_{b,M}|_{\Theta(\ell(K))/U_{2k-1}})$, where $\operatorname{spec}(T)$ denotes the set of all of the eigenvalues of T counting multiplicity, and the linear space $\Theta(\ell(K))/U_{2k-1}$ is a quotient group under addition. Note that $U_{-1} = \Theta(\ell(K))$. Since $T_{b,M} U_j \subseteq U_j$ for all $j = -1, 0, \dots, 2k-1$, the quotient group $\Theta(\ell(K))/U_{2k-1}$ is isomorphic to $U_{-1}/U_0 \oplus U_0/U_1 \oplus \cdots \oplus U_{2k-2}/U_{2k-1}$. Hence

$$\operatorname{spec}(T_{b,M}|_{\Theta(\ell(K))/U_{2k-1}}) = \bigcup_{j=0}^{2k-1} \operatorname{spec}(T_{b,M}|_{U_{j-1}/U_j}).$$

By [25, Theorem 3.2] or by the proof of Theorem 3.1 in section 3, we know that, for any $j = 0, \dots, 2k-1$, all of the eigenvalues of $T_{b,M}|_{V_{j-1}/V_j}$ have modulus $|\det M|^{-j/d}$, where we used the assumption that M is isotropic. Since U_{j-1}/U_j is a subgroup of V_{j-1}/V_j, we deduce that all of the eigenvalues of $T_{b,M}|_{U_{j-1}/U_j}$ have modulus $|\det M|^{-j/d}$. (In fact, by duality, we can prove that, for any $j = 0, \dots, 2k-1$,

$$\operatorname{spec}(T_{b,M}|_{U_{j-1}/U_j}) = \operatorname{spec}(\tau|_{\Theta(\Pi_j \setminus \Pi_{j-1})})$$

without assuming that M is isotropic, where $[\tau(q)](x) := q(M^{-1}x)$, $q \in \Pi_{2k-1}$.) By duality,

$$\dim(U_{j-1}/U_j) = \dim(U_{j-1}) - \dim(U_j) = \dim(\Theta(\Pi_j)) - \dim(\Theta(\Pi_{j-1})) = m_\Theta(j).$$

Note that $\{\Theta(\delta_\alpha) : \alpha \in K_\Theta\}$ is a basis of $\Theta(\ell(K))$, and the matrix T is the representation matrix of the linear operator $T_{b,M}$ acting on $\Theta(\ell(K))$ under the basis $\{\Theta(\delta_\alpha) : \alpha \in K_\Theta\}$. This completes the proof. □

From the above proof, without the assumption that M is isotropic, we observe that $\rho_k(a; M, 2)$ is the largest number in the set $\sigma(T) \setminus \{|\lambda| : \lambda \in \operatorname{spec}(\tau|_{\Theta(\Pi_{2k-1})})\}$, where $\sigma(T)$ is defined in Algorithm 2.1 and $[\tau(q)](x) := q(M^{-1}x)$, $x \in \mathbb{R}^d$, $q \in \Pi_{2k-1}$. Since Θ is a symmetry group with respect to the dilation matrix M, it is easy to see that $\tau\Theta = \Theta\tau$ and $\tau\Theta(\Pi_j) \subseteq \Theta(\Pi_j)$ for all $j \in \mathbb{N}$, where $\Theta(\Pi_j) := \{\frac{1}{\#\Theta} \sum_{\theta \in \Theta} q(\theta x) : q \in \Pi_j\}$. In passing, we mention that the calculation of the Sobolev smoothness for a bivariate mask which is invariant under Θ_2^A with the dilation matrix $2I_2$ was also discussed by Zhang in [36]. When a mask has a nonnegative symbol, then we can also compute $\rho_k(a; M, \infty)$ in a similar way (see [14, Theorem 4.1]). For completeness, we present the following algorithm, whose proof is almost identical to that of Algorithm 2.1.

ALGORITHM 2.5. *Let M be a $d \times d$ isotropic dilation matrix, and let Θ be a symmetry group with respect to the dilation matrix M. Let a be a mask on $\mathbb{Z}^d$ such that $\sum_{\beta \in \mathbb{Z}^d} a(\beta) = 1$. Suppose that a is invariant under the symmetry group Θ, the symbol of a is nonnegative (i.e., $\widehat{a}(\xi) \geqslant 0$ for all $\xi \in \mathbb{R}^d$), and a satisfies the sum rules of order k but not $k+1$. The quantity $\nu_\infty(a; M)$, or, equivalently, $\rho_k(a; M, \infty)$, is obtained via the following procedure:*

(a) *Find a finite subset K_Θ of $\mathbb{Z}^d_\Theta$ such that*

$$\{M^{-1}(\theta\alpha + \beta) : \theta \in \Theta, \alpha \in K_\Theta, \beta \in \operatorname{supp} a\} \cap \mathbb{Z}^d \subseteq \{\theta\beta : \beta \in K_\Theta, \theta \in \Theta\}$$

and

$$\dim(\Pi_{k-1}|_{\{\theta\beta\,:\,\theta\in\Theta,\,\beta\in K_\Theta\}}) = \dim(\Pi_{k-1}).$$

(b) *Obtain a* $(\#K_\Theta)\times(\#K_\Theta)$ *matrix* T *as follows:*

$$T[\alpha,\beta] := \frac{|\det M|}{\#\{\theta\in\Theta\ :\ \theta\beta=\beta\}}\sum_{\theta\in\Theta} a(M\beta-\theta\alpha), \qquad \alpha,\beta\in K_\Theta.$$

(c) *Let* $\sigma(T)$ *consist of the absolute values of all of the eigenvalues of the square matrix* T *counting the multiplicity of its eigenvalues. Then* $\nu_\infty(a;M)$ *is the smallest number in the following set:*

$$\{-d\log_{|\det M|}\rho\ :\ \rho\in\sigma(T)\}\backslash \{j \text{ with positive multiplicity } m_\Theta(j)\ :\ j=0,\dots,k-1\}.$$

Moreover, without the assumption that the symbol of the mask a *is nonnegative,* $\nu_\infty(a;M)$ *is equal to or less than the quantity obtained in* (c).

Cohen and Daubechies in [4] discussed how to estimate the smoothness exponent of a refinable function using the Fredholm determinant theory. Matlab routines for computing smoothness exponents using the method in [25] were developed and described in [28]. When a mask has a nonnegative symbol, Matlab routines for estimating the Hölder smoothness exponent were developed and described in [1], where symmetry is not taken into account and eigenvectors have to be explicitly computed and checked as to whether or not they belong to the subspace V_{k-1}.

3. Relations among $\rho_k(a;M,p)$, $k\in\mathbb{N}_0$. In this section, we shall study the relations among $\rho_k(a;M,p)$, $k\in\mathbb{N}_0$. Using such relations, we shall be able to overcome the difficulty in D3 in section 1 in order to check the stability condition for certain refinable functions.

The main results in this section are as follows.

THEOREM 3.1. *Let* M *be a dilation matrix. Let* a *be a finitely supported mask on* $\mathbb{Z}^d$ *such that* $\sum_{\beta\in\mathbb{Z}^d} a(\beta)=1$ *and* a *satisfies the sum rules of order* k *with respect to the lattice* $M\mathbb{Z}^d$. *Let* $\lambda_{min} := \min_{1\leqslant j\leqslant d}|\lambda_j|$ *and* $\lambda_{max} := \max_{1\leqslant j\leqslant d}|\lambda_j|$, *where* $\lambda_1,\dots,\lambda_d$ *are all of the eigenvalues of* M. *Then*

$$\rho_j(a;M,p) = \max\{\rho_k(a;M,p),\ \ |\det M|^{1/p}\lambda_{min}^{-j}\} \quad \forall\ 1\leqslant p\leqslant\infty,\ 0\leqslant j<k, \tag{3.1}$$

and

$$|\det M|^{1/q-1/p}\rho_j(a;M,p)\leqslant\rho_j(a;M,q)\leqslant\rho_j(a;M,p)$$

for all $j\in\mathbb{N}_0$ *and* $1\leqslant p\leqslant q\leqslant\infty$. *Consequently,*

$$\nu_p(a;M)\geqslant\nu_q(a;M)\geqslant\nu_p(a;M)+(1/q-1/p)\log_{\lambda_{max}}|\det M|$$

for all $1\leqslant p\leqslant q\leqslant\infty$.

We say that a mask a is an *interpolatory mask* with respect to the lattice $M\mathbb{Z}^d$ if $a(\beta)=0$ for all $\beta\in M\mathbb{Z}^d\backslash\{0\}$. Let a and b be two finitely supported masks on $\mathbb{Z}^d$. Define a sequence c by $\widehat{c}(\xi)=\widehat{a}(\xi)\overline{\widehat{b}(\xi)}$, $\xi\in\mathbb{R}^d$. If c is an interpolatory mask with respect to the lattice $M\mathbb{Z}^d$, then b is called a *dual mask* of a with respect to the lattice $M\mathbb{Z}^d$ and vice versa.

Let ϕ be a continuous function on $\mathbb{R}^d$. We say that ϕ is an *interpolating function* if $\phi(0) = 1$ and $\phi(\beta) = 0$ for all $\beta \in \mathbb{Z}^d\backslash\{0\}$. For discussion on interpolating refinable functions and interpolatory masks, the reader is referred to [7, 8, 9, 15, 16, 30, 31] and references therein. For a compactly supported function ϕ on $\mathbb{R}^d$, we say that the shifts of ϕ are *linearly independent* if, for every $\xi \in \mathbb{C}^d$, $\widehat{\phi}(\xi + 2\pi\beta) \neq 0$ for some $\beta \in \mathbb{Z}^d$. Clearly, if the shifts of ϕ are linearly independent, then the shifts of ϕ are stable. When ϕ is a compactly supported interpolating function, then the shifts of ϕ are linearly independent since $\sum_{\beta\in\mathbb{Z}^d} \widehat{\phi}(\xi + 2\pi\beta) = 1$. Let ϕ be the refinable function with a finitely supported mask and the dilation matrix $2I_d$. A method was proposed in Hogan and Jia [19] to check whether the shifts of ϕ are linearly independent or not. However, there are similar difficulties as mentioned in D1 and D2 in section 1 when applying such a method in [19]. In fact, the procedure in [19] is not numerically stable, and exact arithmetic is needed. Also see [29] on stability.

An iteration scheme can be employed to solve the refinement equation (1.1). Start with some initial function $\phi_0 \in L_p(\mathbb{R}^d)$ such that $\widehat{\phi_0}(0) = 1$ and $\widehat{\phi_0}(2\pi\beta) = 0$ for all $\beta \in \mathbb{Z}^d\backslash\{0\}$. We employ the iteration scheme $Q^n_{a,M}\phi_0$, $n \in \mathbb{N}_0$, where $Q_{a,M}$ is the linear operator on $L_p(\mathbb{R}^d)$ $(1 \leqslant p \leqslant \infty)$ given by

$$Q_{a,M} f := |\det M| \sum_{\beta\in\mathbb{Z}^d} a(\beta) f(M \cdot -\beta), \qquad f \in L_p(\mathbb{R}^d).$$

This iteration scheme is called a *subdivision scheme* or a *cascade algorithm* [2, 18]. When the sequence $Q^n_{a,M}\phi_0$ converges in the space $L_p(\mathbb{R}^d)$, then the limit function must be ϕ^M_a, and we say that the subdivision scheme associated with mask a and dilation matrix M converges in the L_p norm. It was proved in [14] that the subdivision scheme associated with the mask a and the dilation matrix M converges in the L_p norm if and only if $\rho_1(a; M, p) < |\det M|^{1/p}$. (By Theorem 3.1, we see that this is equivalent to $\nu_p(a; M) > 0$.) See [2, 9, 14, 18] and references therein on convergence of subdivision schemes.

Let ϕ be a refinable function with a finitely supported mask a and a dilation matrix M. It is known that ϕ is an interpolating refinable function if and only if the mask a is an interpolatory mask with respect to the lattice $M\mathbb{Z}^d$ and the subdivision scheme associated with mask a and dilation M converges in the L_∞ norm (equivalently, $\rho_1(a; M, \infty) < 1$; see [14]). However, in general, it is difficult to directly check the condition $\rho_1(a; M, \infty) < 1$. On the other hand, in order to check that ϕ is an interpolating refinable function with a finitely supported interpolatory mask a and a $d \times d$ dilation matrix M, it was known in the literature (for example, see [1, 29, 31, 34]) that one needs to check the following two alternative conditions: (1) ϕ is a continuous function. (Often, one computes the Sobolev smoothness exponent of ϕ to establish that $\nu_2(\phi) > d/2$, and, consequently, ϕ is a continuous function.) (2) Up to a scalar modification, δ is the unique eigenvector of the transition operator $T_{a,M}|_{\ell(\Omega_{a,M})}$ corresponding to a simple eigenvalue 1. In the following, we show that, if a is an interpolatory mask and $\nu_2(a; M) > d/2$, then (2) is automatically satisfied. In other words, for an interpolatory mask a with respect to the lattice $M\mathbb{Z}^d$, we show that $\nu_2(a; M) > d/2$ implies $\nu_\infty(a; M) > 0$. Consequently, $\rho_1(a; M, \infty) < 1$, and the corresponding subdivision scheme converges in the L_∞ norm, and its associated refinable function is indeed an interpolating refinable function.

COROLLARY 3.2. *Let a be a finitely supported mask on $\mathbb{Z}^d$, and let M be a dilation matrix. Suppose that b is a dual mask of a with respect to the lattice $M\mathbb{Z}^d$*

and

$$\nu_p(a;M)+\nu_q(b;M)>0 \quad \textit{for some} \quad 1\leqslant p,q\leqslant\infty \quad \textit{with} \quad 1/p+1/q=1. \tag{3.2}$$

Then the shifts of ϕ_a^M are linearly independent and consequently stable. If M is isotropic and (3.2) *holds, then $\nu_p(a;M)>0$ implies that $\phi_a^M\in L_p(\mathbb{R}^d)$ and $\nu_p(\phi_a^M)=\nu_p(a;M)$. In particular, if $\nu_2(a;M)>d/2$ (or, more generally, $\nu_p(a;M)>d/p$ for some $1\leqslant p\leqslant\infty$) and a is an interpolatory mask with respect to the lattice $M\mathbb{Z}^d$, then the subdivision scheme associated with mask a and dilation M converges in the L_∞ norm, and, consequently, ϕ_a^M is a continuous interpolating refinable function.*

Proof. Let $m:=|\det M|$. Define a sequence c by $\widehat{c}(\xi)=\widehat{a}(\xi)\overline{\widehat{b}(\xi)}$. By definition, c is an interpolatory mask with respect to the lattice $M\mathbb{Z}^d$. By [12, Theorem 5.2] and Young's inequality, when $1/p+1/q=1$, we have

$$\rho_{j+k}(c;M,\infty)\leqslant m^{-1}\rho_j(a;M,p)\rho_k(b;M,q)\qquad\forall\, j,k\in\mathbb{N}_0.$$

Note that

$$\nu_p(a;M)=-\log_{\lambda_{max}}[m^{-1/p}\rho_j(a;M,p)]$$

and

$$\nu_q(b;M)=-\log_{\lambda_{max}}[m^{-1/q}\rho_k(b;M,q)]$$

for some proper integers j and k. Therefore, $\rho_{j+k}(c;M,\infty)\leqslant\lambda_{max}^{-\nu_p(a;M)-\nu_q(b;M)}<1$. It follows from Theorem 3.1 that $\rho_1(c;M,\infty)<1$, and, therefore, the subdivision scheme associated with mask c and dilation M converges in the L_∞ norm. Consequently, ϕ_c^M is an interpolating refinable function, and so its shifts are linearly independent. Note that $\widehat{\phi}_c^M(\xi)=\widehat{\phi}_a^M(\xi)\overline{\widehat{\phi}_b^M(\xi)}$. Therefore, the shifts of ϕ_a^M must be linearly independent and consequently stable.

Note that δ is a dual mask of an interpolatory mask, and, for any $1\leqslant q\leqslant\infty$,

$$\nu_q(\delta;M)=(1/q-1)\log_{\lambda_{max}}m\geqslant d/q-d$$

since $\lambda_{max}\geqslant|\det M|^{1/d}$. The second part of Corollary 3.2 follows directly from the first part. The second part can also be proved directly. Since $\nu_p(a;M)>d/p$, by Theorem 3.1, we have

$$\rho_k(a;M,\infty)\leqslant\rho_k(a;M,p)=m^{1/p}\lambda_{max}^{-\nu_p(a;M)}<[m\lambda_{max}^{-d}]^{1/p}\leqslant 1$$

for some proper integer k. By Theorem 3.1, we have $\rho_1(a;M,\infty)<1$. So the subdivision scheme associated with the mask a and the dilation matrix M converges in the L_∞ norm, and, therefore, we conclude that ϕ_a^M is a continuous interpolating refinable function. ∎

Let k be a nonnegative integer. We mention that, if $\rho_j(a;M,p)<|\det M|^{1/p}\lambda_{max}^{-k}$ for some positive integer j, then one can prove that the mask a must satisfy the sum rules of order at least $k+1$ with respect to the lattice $M\mathbb{Z}^d$.

In order to prove Theorem 3.1, we need to introduce the concept of ℓ_p norm joint spectral radius. Let $\mathcal{A}$ be a finite collection of linear operators on a finite dimensional normed vector space V. We denote $\|A\|$ as the operator norm of A which is defined

to be $\|A\| := \sup\{\|Av\| \ : \ \|v\| = 1, v \in V\}$. For a positive integer n, $\mathcal{A}^n$ denotes the Cartesian power of $\mathcal{A}$,

$$\mathcal{A}^n = \{(A_1, \dots, A_n) \ : \ A_1, \dots, A_n \in \mathcal{A}\},$$

and, for $1 \leqslant p \leqslant \infty$, we define

$$\|\mathcal{A}^n\|_p := \Bigg(\sum_{(A_1,\dots,A_n)\in\mathcal{A}^n} \|A_1 \cdots A_n\|^p \Bigg)^{1/p} \qquad \text{when} \quad 1 \leqslant p < \infty,$$
$$\|\mathcal{A}^n\|_\infty := \max\{\|A_1 \cdots A_n\| \quad : \quad (A_1, \dots, A_n) \in \mathcal{A}^n\} \qquad \text{when} \quad p = \infty.$$

For any $1 \leqslant p \leqslant \infty$, the ℓ_p *norm joint spectral radius* (see [6, 15, 24] and references therein on the ℓ_p norm joint spectral radius) of $\mathcal{A}$ is defined to be

$$\rho_p(\mathcal{A}) := \lim_{n\to\infty} \|\mathcal{A}^n\|_p^{1/n} = \inf_{n\geqslant 1} \|\mathcal{A}^n\|_p^{1/n}.$$

Let $\mathcal{E}$ be a complete set of representatives of the distinct cosets of the quotient group $\mathbb{Z}^d/M\mathbb{Z}^d$. To relate the quantities $\rho_k(a; M, p)$ to the ℓ_p norm joint spectral radius, we introduce the linear operator $T_\varepsilon (\varepsilon \in \mathcal{E})$ on $\ell_0(\mathbb{Z}^d)$ as follows:

$$T_\varepsilon u(\alpha) := |\det M| \sum_{\beta\in\mathbb{Z}^d} a(M\alpha - \beta + \varepsilon) u(\beta), \qquad \alpha \in \mathbb{Z}^d,\ u \in \ell_0(\mathbb{Z}^d). \tag{3.3}$$

For $\nu = (\nu_1, \dots, \nu_d)$ and $\mu = (\mu_1, \dots, \mu_d)$, we say that $\nu \leqslant \mu$ if $\nu_j \leqslant \mu_j$ for all $j = 1, \dots, d$.

Proof of Theorem 3.1. Let $m := |\det M|$. Let $K_0 = \operatorname{supp} a \cup \{\beta \in \mathbb{Z}^d \ : \ |\beta| \leqslant k\}$ and $K = \mathbb{Z}^d \cap \sum_{j=1}^\infty M^{-j} K_0$. Define

$$V_j = \left\{ u \in \ell(K) \ : \ \langle u, q\rangle = \sum_{\beta\in\mathbb{Z}^d} u(\beta)\overline{q(\beta)} = 0 \qquad \forall\ q \in \Pi_j \right\}, \qquad j \in \mathbb{N}_0.$$

Since a satisfies the sum rules of order k, by [20, Theorem 5.2], $T_\varepsilon V_j \subseteq V_j$ for all $0 \leqslant j < k$ and for all $\varepsilon \in \mathcal{E}$. By [14, Theorem 2.5], we have

$$\rho_j(a; M, p) = \rho_p(\{T_\varepsilon|_{V_{j-1}} \ : \ \varepsilon \in \mathcal{E}\}), \qquad j = 0, \dots, k.$$

Note that $V_{j-1} = V_j \oplus W_j$, where $W_j := \operatorname{span}\{\nabla^\mu \delta \ : \ |\mu| = j,\ \mu \in \mathbb{N}_0^d\}$.

For any $\mu, \nu \in \mathbb{N}_0^d$ such that $|\nu| \leqslant |\mu| < k$, we have

$$\begin{aligned}\left\langle T_\varepsilon \nabla^\mu \delta, \frac{(M\cdot)^\nu}{\nu!} \right\rangle &= \sum_{\alpha\in\mathbb{Z}^d} [T_\varepsilon \nabla^\mu \delta](\alpha) \frac{(M\alpha)^\nu}{\nu!} \\ &= m \sum_{\beta\in\mathbb{Z}^d} \sum_{\alpha\in\mathbb{Z}^d} a(M\alpha - \beta + \varepsilon)[\nabla^\mu \delta](\beta) \frac{(M\alpha)^\nu}{\nu!}.\end{aligned}$$

Note that

$$\frac{(M\alpha)^\nu}{\nu!} = \frac{(M\alpha - \beta + \varepsilon + \beta - \varepsilon)^\nu}{\nu!} = \sum_{0\leqslant\eta\leqslant\nu} \frac{(M\alpha - \beta + \varepsilon)^{\nu-\eta}}{(\nu-\eta)!} \frac{(\beta - \varepsilon)^\eta}{\eta!}.$$

Since a satisfies the sum rules of order k, we have

$$\sum_{\alpha\in\mathbb{Z}^d} a(M\alpha-\beta+\varepsilon)\frac{(M\alpha)^\nu}{\nu!} = \sum_{0\leqslant\eta\leqslant\nu} \frac{(\beta-\varepsilon)^\eta}{\eta!} \sum_{\alpha\in\mathbb{Z}^d} a(M\alpha-\beta+\varepsilon)\frac{(M\alpha-\beta+\varepsilon)^{\nu-\eta}}{(\nu-\eta)!}$$
$$= \sum_{0\leqslant\eta\leqslant\nu} \frac{(\beta-\varepsilon)^\eta}{\eta!} \sum_{\alpha\in\mathbb{Z}^d} a(M\alpha)\frac{(M\alpha)^{\nu-\eta}}{(\nu-\eta)!}.$$

Thus, for $\nu,\mu\in\mathbb{N}_0^d$ such that $|\nu|\leqslant|\mu|<k$, we have

$$\left\langle T_\varepsilon\nabla^\mu\delta, \frac{(M\cdot)^\nu}{\nu!}\right\rangle = m\sum_{0\leqslant\eta\leqslant\nu}\sum_{\alpha\in\mathbb{Z}^d} a(M\alpha)\frac{(M\alpha)^{\nu-\eta}}{(\nu-\eta)!}\sum_{\beta\in\mathbb{Z}^d}[\nabla^\mu\delta](\beta)\frac{(\beta-\varepsilon)^\eta}{\eta!}.$$

It is evident that

$$\sum_{\beta\in\mathbb{Z}^d}[\nabla^\mu\delta](\beta)\frac{(\beta-\varepsilon)^\eta}{\eta!} = \left\langle\nabla^\mu\delta, \frac{(\cdot-\varepsilon)^\eta}{\eta!}\right\rangle = \delta(\mu-\eta) \qquad \forall\,|\eta|\leqslant|\mu|.$$

Therefore,

(3.4)
$$\left\langle T_\varepsilon\nabla^\mu\delta, \frac{(M\cdot)^\nu}{\nu!}\right\rangle = m\delta(\mu-\nu)\sum_{\alpha\in\mathbb{Z}^d}a(M\alpha) = \delta(\mu-\nu) \quad \forall\,\varepsilon\in\mathcal{E},\ |\nu|\leqslant|\mu|<k.$$

On the other hand, for any $|\nu|\leqslant|\mu|$,

$$\left\langle \sum_{|\eta|=|\mu|} S(M^{-1},|\mu|)_{\eta,\mu}\nabla^\eta\delta,\ \frac{(M\cdot)^\nu}{\nu!}\right\rangle = \sum_{\alpha\in\mathbb{Z}^d}\sum_{|\eta|=|\mu|} S(M^{-1},|\mu|)_{\eta,\mu}[\nabla^\eta\delta](\alpha)\frac{(M\alpha)^\nu}{\nu!}$$
$$= \sum_{\alpha\in\mathbb{Z}^d}\sum_{|\eta|=|\mu|} S(M^{-1},|\mu|)_{\eta,\mu}[\nabla^\eta\delta](\alpha)\sum_{|\lambda|=|\nu|}S(M,|\nu|)_{\nu,\lambda}\frac{\alpha^\lambda}{\lambda!}$$
$$= \sum_{|\eta|=|\mu|}\sum_{|\lambda|=|\nu|} S(M^{-1},|\mu|)_{\eta,\mu}S(M,|\nu|)_{\nu,\lambda}\sum_{\alpha\in\mathbb{Z}^d}[\nabla^\eta\delta](\alpha)\frac{\alpha^\lambda}{\lambda!}$$
$$= \sum_{|\eta|=|\mu|}\sum_{|\lambda|=|\nu|} S(M^{-1},|\mu|)_{\eta,\mu}S(M,|\nu|)_{\nu,\lambda}\delta(\eta-\lambda)$$
$$= \delta(|\mu|-|\nu|)\sum_{|\eta|=|\mu|} S(M,|\mu|)_{\nu,\eta}S(M^{-1},|\mu|)_{\eta,\mu}$$
$$= \delta(|\mu|-|\nu|)S(I_d,|\mu|)_{\nu,\mu}$$
$$= \delta(\mu-\nu),$$

where $S(M^{-1},|\mu|)$ is defined in (2.10). Therefore, we have

$$T_\varepsilon\nabla^\mu\delta - \sum_{|\eta|=|\mu|} S(M^{-1},j)^T_{\mu,\eta}\nabla^\eta\delta\in V_j \qquad \forall\,|\mu|=j<k,\ \varepsilon\in\mathcal{E}.$$

Since $V_{j-1}=V_j\oplus W_j$ and $\{\nabla^\mu\delta\,:\,|\mu|=j,\ \mu\in\mathbb{N}_0^d\}$ is a basis for W_j, we have

$$T_\varepsilon|_{V_{j-1}} = \begin{bmatrix} T_\varepsilon|_{V_j} & * \\ 0 & S(M^{-1},j)^T\end{bmatrix}, \qquad \varepsilon\in\mathcal{E},\ 0\leqslant j<k.$$

Note that the spectral radius of $S(M^{-1}, j)^T$ is λ_{min}^{-j} for all $j \in \mathbb{N}$. Therefore, we deduce that

$$\rho_p(\{T_\varepsilon|_{V_{j-1}} \,:\, \varepsilon \in \mathcal{E}\}) = \max\{m^{1/p}\lambda_{min}^{-j},\ \rho_p(\{T_\varepsilon|_{V_j} \,:\, \varepsilon \in \mathcal{E}\})\}.$$

So (3.1) holds.

By the definition of the ℓ_p norm joint spectral radius, using the Hölder inequality, we have

$$|\det M|^{1/q-1/p}\rho_p(\{T_\varepsilon|_{V_j} \,:\, \varepsilon \in \mathcal{E}\}) \leqslant \rho_q(\{T_\varepsilon|_{V_j} \,:\, \varepsilon \in \mathcal{E}\}) \leqslant \rho_p(\{T_\varepsilon|_{V_j} \,:\, \varepsilon \in \mathcal{E}\})$$

(see [14]) for all $1 \leqslant p \leqslant q \leqslant \infty$. This completes the proof. ∎

4. Some examples of symmetric refinable functions. In this section, we shall give several examples to demonstrate the advantages of the algorithms and results in sections 2 and 3 on computing smoothness exponents of symmetric refinable functions.

Example 4.1. Let $M = 2I_2$. The interpolatory mask a for the butterfly scheme in [9] is supported on $[-3, 3]^2$ and is given by

$$\frac{1}{64}\begin{bmatrix} 0 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & -1 & 0 & 2 & 0 & -1 \\ 0 & -1 & 2 & 8 & 8 & 2 & -1 \\ 0 & 0 & 8 & 16 & 8 & 0 & 0 \\ -1 & 2 & 8 & 8 & 2 & -1 & 0 \\ -1 & 0 & 2 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then a satisfies the sum rules of order 4, and a is invariant under the hexagonal symmetry group Θ^H. By Proposition 2.2, we have $\#K_{\Theta^H} = 11$, and, by computing the eigenvalues of the 11×11 matrix T in Algorithm 2.1, we have

$$\begin{aligned}&\{-\log_4 \rho \,:\, \rho \in \sigma(T)\}\\ &\qquad = \{0,\ 1,\ 2,\ 2.44077,\ 2.56925,\ 3,\ 3,\ 3.05923,\ 3.28397,\ 3.72404,\ 4\}.\end{aligned}$$

Let ϕ be the refinable function with mask a and the dilation matrix $2I_2$. So, by Algorithm 2.1, $\nu_2(a; 2I_2) \approx 2.44077 > 1$. Therefore, by Corollary 3.2, ϕ is an interpolating refinable function and $\nu_2(\phi) = \nu_2(a; 2I_2) \approx 2.44077$. Note that the matrix size using the method in [25] is $\#\Omega_{b,2I_2} = 109$, which is much larger than the matrix size $\#K_{\Theta^H} = 11$ used in Algorithm 2.1.

Example 4.2. Let $M = 2I_2$. A family of bivariate interpolatory masks RS_r ($r \in \mathbb{N}$) was given in Riemenschneider and Shen [31] (also see Jia [22]) such that RS_r is supported on $[1-2r, 2r-1]^2$, RS_r satisfies the sum rules of order $2r$ with respect to the lattice $2\mathbb{Z}^2$, and RS_r is invariant under the hexagonal symmetry group Θ^H. Using the fact that the symbol of RS_r has the factor $[(1 + e^{-i\xi_1})(1 + e^{-i\xi_2})(1 + e^{i(\xi_1+\xi_2)})]^r$, by taking out some of such factors, Jia and Zhang [25, Theorem 4.1] were able to compute the Sobolev smoothness exponents of ϕ_r for $r = 2, \ldots, 16$, where ϕ_r denotes the refinable function with the mask RS_r and the dilation matrix $2I_2$. Note that the mask RS_{16} is supported on $[-31, 31]^2$. In fact, in order to compute $\nu_2(\phi_{16})$,

the method in [25, Theorem 4.1] has to compute the eigenvalues of two matrices of size 4743. (Without factorization, the matrix size used in [25] is 11719.) Without using any factorization, for any mask a which is supported on $[-31,31]^2$ and is invariant under Θ^H, by Algorithm 2.1, we have $\#K_{\Theta^H} = 992$. So, to compute $\nu_2(\phi_{16})$, we need only to compute the eigenvalues of a matrix of size 992.

Example 4.3. Let $M = \left[\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right]$ be the quincunx dilation matrix. The interpolatory mask a is supported on $[-3,3]^2$ and is given by

$$\frac{1}{64}\begin{bmatrix} 0 & -1 & 0 & -1 & 0 \\ -1 & 0 & 10 & 0 & -1 \\ 0 & 10 & 32 & 10 & 0 \\ -1 & 0 & 10 & 0 & -1 \\ 0 & -1 & 0 & -1 & 0 \end{bmatrix}.$$

Note that a satisfies the sum rules of order 4 with respect to the quincunx lattice $M\mathbb{Z}^2$, and a is invariant under the full axes symmetry group Θ_2^A with respect to the dilation matrix M. This example was discussed in [25] and belongs to a family of quincunx interpolatory masks in [16]. Let ϕ be the refinable function with the mask a and dilation matrix M. By Algorithm 2.1, we have $\#K_{\Theta_2^A} = 46$ and $\nu_2(a;M) \approx 2.44792 > 1$. Therefore, $\nu_2(\phi) = \nu_2(a;M) \approx 2.44792$. Note that the matrix to compute $\nu_2(\phi)$ using the method in [25] has size 481 (see [25]), which is much larger than the size 46 when using Algorithm 2.1. Note that the symbol of a is nonnegative. By Algorithm 2.5, we have $\#K_{\Theta_2^A} = 13$ and $\nu_\infty(a;M) \approx 1.45934 > 0$. Therefore, by Corollary 3.2, $\nu_\infty(\phi) = \nu_\infty(a;M) \approx 1.45934$. However, using the method in [25], the matrix size is 129 (see [25]), which is much larger than the size 13 in Algorithm 2.5.

Example 4.4. Let $M = \left[\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right]$. A family of quincunx interpolatory masks g_r ($r \in \mathbb{N}$) was proposed in [16] such that g_r is supported on $[-r,r]^2$, satisfies the sum rules of order $2r$ with respect to $M\mathbb{Z}^2$, is an interpolatory mask with respect to $M\mathbb{Z}^2$, and is invariant under the full axes symmetry group Θ_2^A. Note that the mask in Example 4.3 corresponds to the mask g_2 in this family. Since the symbols of g_r are nonnegative, the L_∞ smoothness exponents $\nu_\infty(\phi_r)$ were computed in [16] for $r = 1,\dots,8$, where ϕ_r is the refinable function with mask g_r and the dilation matrix M. Using Algorithm 2.5, we are able to compute $\nu_\infty(\phi_r)$ for $r = 9,\dots,16$ in Table 3.

TABLE 3

The L_∞ (Hölder) smoothness exponent of the interpolating refinable function ϕ_r whose mask is g_r.

$\nu_\infty(\phi_9)$	$\nu_\infty(\phi_{10})$	$\nu_\infty(\phi_{11})$	$\nu_\infty(\phi_{12})$
5.71514	6.21534	6.70431	7.18321
$\nu_\infty(\phi_{13})$	$\nu_\infty(\phi_{14})$	$\nu_\infty(\phi_{15})$	$\nu_\infty(\phi_{16})$
7.65242	8.11171	8.56039	8.99752

A coset-by-coset (CBC) algorithm was proposed in [12, 16] to construct quincunx biorthogonal wavelets. Some examples of dual masks of g_r, denoted by $(g_r)_k^s$, were constructed in [16, Theorem 5.2], and some of their Sobolev smoothness exponents were given in Table 4 of [16]. Note that the dual mask $(g_r)_k^s$ is supported on $[-k-r, r+k]^2$, satisfies the sum rules of order $2k$, has nonnegative symbol, and is invariant under the full axes symmetry group Θ_2^A. However, in the paper [16], we are

unable to complete the computation in Table 4 in [16] due to the difficulty mentioned in D2 in section 1. In fact, to compute $\nu_2(a; M)$ for a mask supported on $[-k, k]^2$, the set $\Omega_{b,M}$ defined in (1.5) is given by

$$\{(i, j) \in \mathbb{Z}^2 \ : \ |i| \leqslant 6k,\ |j| \leqslant 6k,\ |i - j| \leqslant 8k,\ |i + j| \leqslant 8k\}.$$

For example, in order to compute $\nu_2((g_4)_8^s; M)$, the set $\Omega_{b,M}$ consists of 16321 points, which is beyond our ability to compute the eigenvalues of a 16321×16321 matrix. We now can complete the computation using Algorithm 2.1. Note that the quincunx dilation M here is denoted by Q in Table 4 of [16]. By computation, $\nu_2(\phi^M_{(g_4)_6^s}) \approx 2.47477$, and the rest of the computation is given in Table 4.

TABLE 4
Computing $\nu_2(\phi^M_{(g_r)_k^s})$ *by Algorithm* 2.1. *The result here completes Table* 4 *of* [16].

$\nu_2(\phi^M_{(g_1)_7^s})$	$\nu_2(\phi^M_{(g_2)_7^s})$	$\nu_2(\phi^M_{(g_3)_7^s})$	$\nu_2(\phi^M_{(g_4)_7^s})$
3.01166	2.92850	2.90251	2.91546
$\nu_2(\phi^M_{(g_1)_8^s})$	$\nu_2(\phi^M_{(g_2)_8^s})$	$\nu_2(\phi^M_{(g_3)_8^s})$	$\nu_2(\phi^M_{(g_4)_8^s})$
3.49499	3.38671	3.34268	3.32116

In passing, we mention that, if a finitely supported mask a on $\mathbb{Z}^2$ is invariant under the full axes symmetry group Θ_2^A, then it was proved in Han [13] that all of the refinable functions with the mask a and any of the quincunx dilation matrices

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

are the same function, which is also invariant under the full axes symmetry group Θ_2^A. Also see [3, 4] on quincunx wavelets. For any primal (matrix) mask and any dilation matrix, the CBC algorithm proposed in [12] can be used to construct dual (matrix) masks with any preassigned order of sum rules.

Example 4.5. Let $M = \left[\begin{smallmatrix} 1 & -2 \\ 2 & -1 \end{smallmatrix}\right]$ be the dilation matrix in a $\sqrt{3}$-subdivision scheme [28]. The interpolatory mask a is supported on $[-4, 4]^2$ and is given by

$$\frac{1}{2187} \begin{bmatrix} 0 & 0 & 0 & 0 & 7 & 4 & 0 & 4 & 7 \\ 0 & 0 & 0 & 4 & 0 & -32 & -32 & 0 & 4 \\ 0 & 0 & 0 & -32 & -20 & 0 & -20 & -32 & 0 \\ 0 & 4 & -32 & 0 & 312 & 312 & 0 & -32 & 4 \\ 7 & 0 & -20 & 312 & 729 & 312 & -20 & 0 & 7 \\ 4 & -32 & 0 & 312 & 312 & 0 & -32 & 4 & 0 \\ 0 & -32 & -20 & 0 & -20 & -32 & 0 & 0 & 0 \\ 4 & 0 & -32 & -32 & 0 & 4 & 0 & 0 & 0 \\ 7 & 4 & 0 & 4 & 7 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that a satisfies the sum rules of order 6 with respect to the lattice $M\mathbb{Z}^2$, and a is invariant under the hexagonal symmetry group Θ^H with respect to the dilation matrix M. By Algorithm 2.1, we have $\#K_{\Theta^H} = 38$ and $\nu_2(a; M) \approx 3.28036 > 1$. Let

ϕ be the refinable function with the mask a and the dilation matrix M. Therefore, by Corollary 3.2, ϕ is a C^2 interpolating refinable function, and $\nu_2(\phi) = \nu_2(a; M) \approx 3.28036$. By estimate, the matrix size $\#\Omega_{b,M}$, using the method in [25], is greater than 361, which is much larger than the size 38 when using Algorithm 2.1. Note that the symbol of a is nonnegative. By Algorithm 2.5, we have $\#K_{\Theta^H} = 11$ and $\nu_\infty(a; M) \approx 2.34654 > 0$. Therefore, by Corollary 3.2, $\nu_\infty(\phi) = \nu_\infty(a; M) \approx 2.34654$. Using the method in [25], the matrix size $\#\Omega_{a,M}$ is greater than 85, which is much larger than the size 11 in Algorithm 2.5. Since $\phi \in C^2$, this example gives us a C^2 $\sqrt{3}$-interpolatory subdivision scheme.

In the rest of this section, let us present some examples in dimension three. By generalizing the proof of [15, Theorem 4.3], we have the following result.

THEOREM 4.6. *Let $M = 2I_d$ be the dilation matrix. For each positive integer r, there exists a unique dyadic interpolatory mask g_r^d in $\mathbb{R}^d$ with the following properties:*

(a) *g_r^d is supported on the set $\{2\alpha + \varepsilon : \varepsilon \in \{-1, 0, 1\}^d, \alpha \in \mathbb{Z}^d, |\alpha| < r\}$;*

(b) *g_r^d is symmetric about all of the coordinate axes;*

(c) *g_r^d satisfies the sum rules of order $2r$ with respect to the lattice $2\mathbb{Z}^d$.*

By the uniqueness, we see that each g_r^d in Theorem 4.6 is invariant under the full axes symmetry group Θ_d^A. By the uniqueness of g_r^d in Theorem 4.6 again, we see that g_r^1 ($r \in \mathbb{N}$) were the masks given in [8] and g_r^2 ($r \in \mathbb{N}$) were the masks proposed in [15]. Moreover, the masks g_r^d can be obtained via a recursive formula without solving any equations.

In the following, let us give some examples of the above interpolatory masks in dimension three. Let

$$\mathbb{Z}^3_{\Theta_3^A} := \{(\beta_1, \beta_2, \beta_3) \in \mathbb{Z}^3 : 0 \leqslant \beta_1 \leqslant \beta_2 \leqslant \beta_3\}.$$

Clearly, if a is a mask invariant under the group Θ_3^A, then it is totally determined by all of the coefficients $a(\beta)$, $\beta \in \mathbb{Z}^3_{\Theta_3^A}$.

Example 4.7. The coefficients of the interpolatory mask g_2^3 on the set $\mathbb{Z}^3_{\Theta_3^A}$ are given by

$$\begin{aligned}
&g_2^3(0,0,0) = 1/8, && g_2^3(0,0,1) = 9/128, && g_2^3(0,1,1) = 5/128,\\
&g_2^3(1,1,1) = 11/512, && g_2^3(0,0,3) = -1/128, && g_2^3(0,1,3) = -1/256,\\
&g_2^3(1,1,3) = -1/512, && g_2^3(\alpha) = 0 && \text{for any other } \alpha \in \mathbb{Z}^3_{\Theta_3^A}.
\end{aligned}$$

Then g_2^3 satisfies the sum rules of order 4, and there are only 81 nonzero coefficients in the mask g_2^3. Let ϕ be the refinable function with the mask g_2^3 and the dilation matrix $2I_3$. By Algorithm 2.1, we have $\#K_{\Theta_3^A} = 36$ and $\nu_2(g_2^3; 2I_3) \approx 2.44077 > 1.5$. Therefore, by Corollary 3.2, ϕ is an interpolating refinable function, and $\nu_2(\phi) \approx 2.44077$. Note that $\#\Omega_{b,2I_3} = 965$ and $\#K_{\Theta_3^A} = 36$. Hence Algorithm 2.1 can greatly reduce the size of the matrix to compute $\nu_2(g_2^3; 2I_3)$.

Example 4.8. The coefficients of the interpolatory mask g_3^3 on the set $\mathbb{Z}^3_{\Theta_3^A}$ are given by

$$\begin{aligned}
&g_3^3(0,0,0) = 1/8, && g_3^3(0,0,1) = 75/1024, && g_3^3(0,1,1) = 87/2048,\\
&g_3^3(1,1,1) = 25/1024, && g_3^3(0,0,3) = -25/2048, && g_3^3(0,1,3) = -29/8192,\\
&g_3^3(1,1,3) = -29/8192, && g_3^3(0,3,3) = 1/2048, && g_3^3(1,3,3) = 1/4096,\\
&g_3^3(0,0,5) = 3/2048, && g_3^3(0,1,5) = 3/4096, && g_3^3(1,1,5) = 3/8192,\\
&g_2^3(\alpha) = 0 \quad \text{for other } \alpha \in \mathbb{Z}^3_{\Theta_3^A}.
\end{aligned}$$

Then g_3^3 satisfies the sum rules of order 6, and it has 171 nonzero coefficients. Let ϕ be the refinable function with the mask g_3^3 and the dilation matrix $2I_3$. By Algorithm 2.1, we have $\#K_{\Theta_3^A} = 101$ and $\nu_2(g_3^3; 2I_3) \approx 3.17513 > 1.5$. Therefore, by Corollary 3.2, ϕ is an interpolating refinable function, and $\nu_2(\phi) \approx 3.17513$. Note that $\#\Omega_{b,2I_3} = 3021$ and $\#K_{\Theta_3^A} = 101$. Hence Algorithm 2.1 can greatly reduce the size of the matrix to compute $\nu_2(g_3^3; 2I_3)$.

Let ϕ_r be the refinable function with the mask g_r^3 ($r \in \mathbb{N}$) and the dilation matrix $2I_3$. The Sobolev smoothness exponents of ϕ_r ($r = 2, \dots, 11$) are presented in Table 5. By [15, Theorem 3.3] and [12, Theorem 5.1], we see that g_r^3 ($r = 1, \dots, 11$) achieves the optimal Sobolev smoothness and optimal order of sum rules with respect to the support of their masks. In general, Algorithms 2.1 and 2.5 roughly reduce the size of the matrix to be $1/(\#\Theta)$ of the number of points in $\Omega_{b,M}$ in (1.5). Note that $\#\Theta_d^A = 2^d d!$ and $\#\Theta_3^A = 48$. So Algorithms 2.1 and 2.5 are very useful in computing the smoothness exponents of symmetric multivariate refinable functions.

TABLE 5
The Sobolev smoothness exponent of the refinable function ϕ_r whose mask is g_r^3 for $r = 2, \dots, 11$.

$\nu_2(\phi_2)$	$\nu_2(\phi_3)$	$\nu_2(\phi_4)$	$\nu_2(\phi_5)$	$\nu_2(\phi_6)$
2.44077	3.17513	3.79313	4.34408	4.86202
$\nu_2(\phi_7)$	$\nu_2(\phi_8)$	$\nu_2(\phi_9)$	$\nu_2(\phi_{10})$	$\nu_2(\phi_{11})$
5.36283	5.85293	6.33522	6.81143	7.28260

Acknowledgments. The author is indebted to Rong-Qing Jia for discussion on computing the smoothness of multivariate refinable functions. The author thanks IMA at the University of Minnesota for their hospitality during his visit to IMA in 2001. The author also thanks the referees for their helpful comments on improving the presentation of this paper and for suggesting the references [1, 4, 7, 18].

REFERENCES

[1] A. Barinka, S. Dahlke, and N. Mulders, *The IGPM Villemoes Machine*, IGPM-Report 184, Institut für Geometrie und Praktische Mathematik, Aachen, Germany, 2000.

[2] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991).

[3] A. Cohen and I. Daubechies, *Nonseparable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[4] A. Cohen and I. Daubechies, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591.

[5] A. Cohen, K. Gröchenig, and L. Villemoes, *Regularity of multivariate refinable functions*, Constr. Approx., 15 (1999), pp. 241–255.

[6] I. Daubechies and J. C. Lagarias, *Two-scale difference equations.* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[7] J. Derado, *Nonseparable, compactly supported interpolating refinable functions with arbitrary smoothness*, Appl. Comput. Harmon. Anal., 10 (2001), pp. 113–138.

[8] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.

[9] N. Dyn, J. A. Gregory, and D. Levin, *A butterfly subdivision scheme for surface interpolation with tension control*, ACM Trans. Graphics, 9 (1990), pp. 160–169.

[10] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[11] T. N. T. Goodman, C. A. Micchelli, and J. D. Ward, *Spectral radius formulas for subdivision operators*, in Recent Advances in Wavelet Analysis, Wavelet Anal. Appl. 3, Academic Press, Boston, 1994, pp. 335–360.

[12] B. Han, *Analysis and construction of optimal multivariate biorthogonal wavelets with compact support*, SIAM J. Math. Anal., 31 (1999), pp. 274–304.
[13] B. Han, *Symmetry property and construction of wavelets with a general dilation matrix*, Linear Algebra Appl., 353 (2002), pp. 207–225.
[14] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.
[15] B. Han and R.-Q. Jia, *Optimal interpolatory subdivision schemes in multidimensional spaces*, SIAM J. Numer. Anal., 36 (1998), pp. 105–124.
[16] B. Han and R.-Q. Jia, *Quincunx fundamental refinable functions and quincunx biorthogonal wavelets*, Math. Comp., 71 (2002), pp. 165–196.
[17] B. Han, T. P.-Y. Yu, and B. Piper, *Multivariate Refinable Hermite Interpolants*, preprint, University of Alberta, Edmonton, Alberta, Canada, 2002.
[18] L. Hervé, *Comportement asymptotique dans l'algorithme de transformée en ondelettes. Lien avec la régularité de l'ondelette*, Rev. Mat. Iberoamericana, 11 (1995), pp. 431–451.
[19] T. A. Hogan and R.-Q. Jia, *Dependence relations among the shifts of a multivariate refinable distribution*, Constr. Approx., 17 (2001), pp. 19–37.
[20] R.-Q. Jia, *Approximation properties of multivariate wavelets*, Math. Comp., 67 (1998), pp. 647–665.
[21] R.-Q. Jia, *Characterization of smoothness of multivariate refinable functions in Sobolev spaces*, Trans. Amer. Math. Soc., 351 (1999), pp. 4089–4112.
[22] R.-Q. Jia, *Interpolatory subdivision schemes induced by box splines*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 286–292.
[23] R.-Q. Jia and Q. T. Jiang, *Spectral Analysis of the Transition Operators and Its Applications to Smoothness Analysis of Wavelets*, preprint, University of Alberta, Edmonton, Alberta, Canada, 2001.
[24] R.-Q. Jia, S. D. Riemenschneider, and D.-X. Zhou, *Smoothness of multiple refinable functions and multiple wavelets*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 1–28.
[25] R.-Q. Jia and S. R. Zhang, *Spectral properties of the transition operator associated to a multivariate refinement equation*, Linear Algebra Appl., 292 (1999), pp. 155–178.
[26] Q. T. Jiang, *On the regularity of matrix refinable functions*, SIAM J. Math. Anal., 29 (1998), pp. 1157–1176.
[27] Q. T. Jiang, *Multivariate matrix refinable functions with arbitrary matrix dilation*, Trans. Amer. Math. Soc., 351 (1999), pp. 2407–2438.
[28] Q. T. Jiang and P. Oswald, *On the Analysis of $\sqrt{3}$-Subdivision Schemes*, preprint, University of Missouri at St. Louis, St. Louis, MO, 2001.
[29] W. Lawton, S. L. Lee, and Z. Shen, *Stability and orthonormality of multivariate refinable functions*, SIAM J. Math. Anal., 28 (1997), pp. 999–1014.
[30] C. A. Micchelli, *Interpolating subdivision schemes and wavelets*, J. Approx. Theory, 86 (1996), pp. 41–71.
[31] S. D. Riemenschneider and Z. Shen, *Multidimensional interpolatory subdivision schemes*, SIAM J. Numer. Anal., 34 (1997), pp. 2357–2381.
[32] O. Rioul, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.
[33] A. Ron and Z. Shen, *The Sobolev regularity of refinable functions*, J. Approx. Theory, 106 (2000), pp. 185–225.
[34] A. Ron, Z. Shen, and K.-C. Toh, *Computing the Sobolev regularity of refinable functions by the Arnoldi method*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 57–76.
[35] L. F. Villemoes, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 433–1460.
[36] S. R. Zhang, *Properties of Refinable Functions and Subdivision Schemes*, Ph.D. thesis, University of Alberta, Edmonton, Alberta, Canada, 1998.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 715–727

VALUES OF MINORS OF AN INFINITE FAMILY OF *D*-OPTIMAL DESIGNS AND THEIR APPLICATION TO THE GROWTH PROBLEM: II*

C. KOUKOUVINOS†, M. MITROULI‡, AND JENNIFER SEBERRY§

Abstract. We obtain explicit formulae for the values of the $2v-j$ minors, $j=0,1,2$, of D-optimal designs of order $2v=x^2+y^2$, v odd, where the design is constructed using two circulant or type 1 incidence matrices of $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ supplementary difference sets (SDS). This allows us to obtain information on the growth problem for families of matrices which have moderately large growth. Some of our theoretical formulae suggest that growth greater than $2v$ may occur, but experimentation has not yet supported this result. An open problem remains to establish whether the $(1,-1)$ completely pivoted (CP) incidence matrices of $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ SDS, which yield D-optimal designs, can have growth greater than $2v$.

Key words. D-optimal designs, supplementary difference sets, symmetric designs, Gaussian elimination, growth, complete pivoting

AMS subject classifications. 05B20, 15A15, 65F05, 65G05

PII. S0895479801386845

1. Introduction.

In this paper, we use several concepts from orthogonal design theory (e.g., see [5]), but here we will formulate those concepts in matrix notation.

Let $A=[a_{ij}]\in\mathcal{R}^{n\times n}$. We reduce A to upper triangular form by using Gaussian elimination with complete pivoting (GECP) [15]. Let $A^{(k)}=[a_{ij}^{(k)}]$, $k=1,2,\ldots,n$, denote the matrix obtained after the first k pivoting operations, so $A^{(n)}$ is the final upper triangular matrix. A diagonal entry of that final matrix will be called a pivot. Matrices with the property that no exchanges are actually needed during GECP are called completely pivoted (CP). Let

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

denote the growth associated with GECP on A and

$$g(n)=\sup\{\,g(n,A)/A\in\mathcal{R}^{n\times n}\,\}.$$

The problem of determining $g(n)$ for various values of n is called the growth problem [6].

The values of $g(n)$ are usually less than n. One of the curious frustrations of the growth problem is that it is difficult but possible to construct any examples of $n\times n$ matrices A for which $g(n,A)$ is greater than or equal to n [6].

*Received by the editors March 23, 2001; accepted for publication (in revised form) by N. J. Higham August 12, 2002; published electronically January 31, 2003.
http://www.siam.org/journals/simax/24-3/38684.html

†Department of Mathematics, National Technical University of Athens, Zografou 15773, Athens, Greece (ckoukouv@math.ntua.gr).

‡Department of Mathematics, University of Athens, Panepistemiopolis 15784, Athens, Greece (mmitroul@cc.uoa.gr). The work of this author was supported by a grant from the University of Athens.

§School of Information Technology and Computer Science, University of Wollongong, Wollongong, NSW, 2522, Australia (jennie@uow.edu.au).

A *Hadamard matrix* of order n is an $n \times n$ matrix of 1's and -1's with $HH^T = H^T H = nI_n$. Hadamard matrices were first studied by Sylvester in 1867. In 1893 Hadamard discovered that if $X = (x_{ij})$ is a matrix of order n, then

$$|\det X|^2 \leq \prod_{i=1}^{n} \sum_{j=1}^{n} |x_{ij}|^2.$$

Hadamard showed that matrices satisfying the equality and with entries in the unit disc (i.e., $|x_{ij}| \leq 1$) have order $1, 2$, or $\equiv 0 \pmod 4$ and entries $\{1, -1\}$. He produced examples for orders up to 20. Subsequently, matrices which satisfy the equality of Hadamard's inequality came to be known as Hadamard matrices. We refer the interested reader to [5] for more details. Two Hadamard matrices H_1 and H_2 are called equivalent (or Hadamard equivalent, or H-equivalent) if one can be obtained from the other by a sequence of row negations, row permutations, column negations, and column permutations. Equivalent Hadamard matrices give different pivot structures when GECP is performed on them. When GECP is done on an $n \times n$ Hadamard matrix H, the last pivot has magnitude n. This was proved by Cryer in [2] because it is the reciprocal of an entry from H^{-1} and that equals $(\frac{1}{n})A^T$. Thus $g(n, H) \geq n$. Cryer [2] also evaluated the two pivots preceding the last which take the value of $\frac{n}{2}$, and he remarked that it is unlikely any earlier pivot under GECP could exceed n. In [3] it was proved that the last six pivots cannot exceed n when GECP is done on a Hadamard matrix. The equality $g(n, H) = n$ has been proved for the equivalence class of $n \times n$ Hadamard matrices containing the Sylvester–Hadamard matrix [3]. This evidence supports Cryer's hunch that $g(n, H) = n$ for any Hadamard matrix H.

A matrix W with entries $\{0, \pm 1\}$ satisfying $WW^T = kI_n$, $k \in \{1, 2, \ldots, n\}$, is called a *weighing matrix* of order n and weight k. For more details and construction methods concerning Hadamard and weighing matrices, see [5]. It has also been observed that weighing matrices of order n can give $g(n, W) = n - 1$ [11].

Following Kharaghani [7] a matrix $\mathcal{B}$ of order n is a *D-optimal matrix* or *D-optimal design* if the determinant of $\mathcal{B}$ is the maximal determinant among all matrices with entries ± 1 (a ± 1 matrix) of order n. Let d_n denote the maximum absolute value of determinant of all $n \times n$ matrices with elements ± 1. It follows from Hadamard's inequality that $d_n \leq n^{\frac{n}{2}}$, and it is easily shown that equality can only hold if $n = 1$ or 2 or if $n \equiv 0 \pmod 4$, as described above. If $n \equiv 0 \pmod 4$ and a Hadamard matrix H of order n exists, then H has absolute value of determinant $n^{\frac{n}{2}}$, and thus it is a D-optimal matrix. It still remains open if a Hadamard matrix of order n exists for every $n \equiv 0 \pmod 4$. The smallest value of n is 428 for which a Hadamard matrix of order n and consequently a D-optimal design of the same order is not yet known.

A D-optimal design A of order n is said to be *constructible from two circulant matrices* if it can be written in the form $A = \left[\begin{smallmatrix} A_1 & A_2 \\ A_2^T & -A_1^T \end{smallmatrix}\right]$, where A_1, A_2 are circulant matrices of order $\frac{n}{2}$.

Let X be an $n \times n$ matrix of the form $aI + bJ$, where J is an $n \times n$ matrix, every entry of which is 1. The eigenvalues of this matrix are a with multiplicity $(n-1)$ and $a + bn$, thus $\det(X) = (a + bn)a^{n-1}$. This paper studies $(+1, -1)$ matrices C of size $(2v) \times (2v)$, where v is odd, and they satisfy $CC^T = \left[\begin{smallmatrix} X & 0 \\ 0 & X \end{smallmatrix}\right]$, where $a = 2v - 2$ and $b = 2$.

Thus $\det(C) = \det(X) = (4v-2)(2v-2)^{v-1}$. We shall here be concerned with the case $n \equiv 2 \pmod 4$, $n \neq 2$, and this will be implicitly assumed in what follows. Ehlich [4] has proved the following theorem.

THEOREM 1. *We have*

$$d_n \leq (2n-2)(n-2)^{\frac{n}{2}-1}$$

and equality can hold only if $2n-2 = x^2+y^2$, *where* x *and* y *are integers.*

Thus, the above matrices A are D-optimal by Ehlich's theorem.

Since Hadamard matrices of order n are D-optimal designs for $n \equiv 0 \pmod 4$, when they exist, and they have large growth, it is natural to inquire how big growth could be for other D-optimal matrices. This is examined in the present paper for an infinite family of D-optimal matrices.

Notation 1. Write A for a matrix of order n whose initial pivots p_i, $i = 1, 2, \ldots,$ are derived from matrices with CP structure. Write $A(j)$ for the absolute value of the determinant of the $j \times j$ principal submatrix in the upper left-hand corner of the matrix A . Throughout this paper, we find all possible values of the $n-j$ minors, $j = 1, 2$. Hence, if any minor is CP, it must have one of these values. It can be proved [2] that

$$g(n, A) = \max\left\{1, \max_{1 \leq k \leq n-1}\left|\frac{A(k+1)}{A(k)}\right|\right\}.$$

Thus, the magnitude of the pivots appearing after the application of Gaussian elimination operations on a CP matrix A is given by

$$p_j = \frac{A(j)}{A(j-1)}, \quad j = 1, 2, \ldots, n, \quad A(0) = 1. \tag{1}$$

2. D-optimal designs of order $2v \equiv 2 \pmod 4$ from symmetric balanced incomplete block designs. For the purpose of this paper we will define a symmetric balanced incomplete block design (SBIBD) (v, k, λ) to be a $v \times v$ matrix, B, with entries 0 or 1, which has exactly k entries $+1$ and $v-k$ entries 0 in each row and column and for which the inner product of any distinct pairs of rows and columns is λ. The $(1, -1)$ incidence matrix of B is obtained by letting $A = 2B - J$, where J is the $v \times v$ matrix with entries all $+1$. We write I for the identity matrix of order v. Then we have

$$BB^T = (k-\lambda)I + \lambda J \tag{2}$$

and

$$AA^T = 4(k-\lambda)I + (v - 4(k-\lambda))J. \tag{3}$$

It can be easily shown that

$$\det B = (k-\lambda)^{\frac{v-1}{2}}\sqrt{k + (v-1)\lambda},$$

and since $\lambda(v-1) = k^2 - k$,

$$\det A = 2^{v-1}(k-\lambda)^{\frac{v-1}{2}}|v-2k|. \tag{4}$$

In this paper we evaluate the $2v-j$, $j=0,1,2$, minors for $(1,-1)$ incidence matrices of certain SBIBDs which yield D-optimal designs.

For the purpose of this paper we will define two supplementary difference sets $2-\{v;k_1,k_2;\lambda\}$, abbreviated as SDS, to be two circulant (or type 1) $v\times v$ matrices B_1 and B_2, with entries 0 or 1, which have exactly k_i entries $+1$ and $v-k_i$ entries 0, $i=1,2$, respectively, in each row and column and for which the inner product of any pair of rows of $[\ B_1 \quad B_2\ ]$ is λ, where $\lambda=(k_1+k_2)-\frac{(v-1)}{2}$. We note that circulant matrices commute, and that the transpose of a circulant matrix is also a circulant matrix. The $(1,-1)$ incidence matrices of B_i are obtained by letting $A_i=2B_i-J,\ \ i=1,2$.

Then it is true that

$$A_1A_1^T+A_2A_2^T=(2v-2)I+2J \tag{5}$$

when the matrix $A=\left[\begin{smallmatrix}A_1 & A_2\\ A_2^T & -A_1^T\end{smallmatrix}\right]$ is constructible from two $(1,-1)$ circulants as just described. Thus when such A exists, it has determinant $(4v-2)(2v-2)^{v-1}$; hence by Ehlich's theorem, it is a D-optimal design.

Only two infinite families of D-optimal designs are known:

1. The first, which uses $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ SDS, is based on the family of $SBIBD(s^2+s+1,s+1,1)$ for s a prime power, found by Singer [12] and used extensively by Spence [13]. Koukouvinos, Kounias, and Seberry [8] showed how to use these SDS to form an infinite family of D-optimal designs, constructible from two circulant matrices, for $n=2(s^2+s+1)$, where $s=2,4,6,8$ or s is an odd prime power. This family is called the *Koukouvinos–Kounias–Seberry–Singer–Spence (KKSSS)* family. If the D-optimal design A is constructed from the above SDS, then

$$\det A=(4v-2)(2v-2)^{v-1}. \tag{6}$$

2. The second family is based on Brouwer's [1] family of $2-\{2s^2+2s+1;s^2,s^2;s(s-1)\}$ SDS, where the two SDS are in fact identical. Whiteman [14] showed how to form these SDS into an infinite family of D-optimal designs, constructible from two circulant matrices, for $n=2(2s^2+2s+1)$, where s is an odd prime power.

In [9] the pivot structure of $(1,-1)$ incidence matrices of $SBIBD(v,k,\lambda)$ was studied. In [10] values for the pivots of $2-\{2s^2+2s+1;s^2,s^2;s(s-1)\}$ SDS were evaluated (as previously noted, the two SDS are in fact identical for this case). In the present paper we obtain values for the pivots of $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ SDS and D-optimal designs made from them. Our calculations here and in [10] have given moderately large values of growth for the D-optimal matrices of both KKSSS and Brouwer types, but it is not known yet whether there exist any $(+1,-1)$ $n\times n$ D-optimal matrices with growth greater than n.

2.1. Minors of size $(2v-1)$. We denote by $A=\Delta(h,i,j,k,m)$ the following matrix of order $2v$:

$$A = \Delta(h,i,j,k,m) =$$

$$\left[\begin{array}{cccc|cccc|cccc|cccc}
\overbrace{m}^{h} & 1 & \cdots & 1 & \overbrace{3}^{i} & 3 & \cdots & 3 & \overbrace{-}^{j} & - & \cdots & - & \overbrace{1}^{k} & 1 & \cdots & 1 \\
1 & m & \cdots & 1 & 3 & 3 & \cdots & 3 & - & - & \cdots & - & 1 & 1 & \cdots & 1 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
1 & 1 & \cdots & m & 3 & 3 & \cdots & 3 & - & - & \cdots & - & 1 & 1 & \cdots & 1 \\
\hline
3 & 3 & \cdots & 3 & m & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 & - & - & \cdots & - \\
3 & 3 & \cdots & 3 & 1 & m & \cdots & 1 & 1 & 1 & \cdots & 1 & - & - & \cdots & - \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
3 & 3 & \cdots & 3 & 1 & 1 & \cdots & m & 1 & 1 & \cdots & 1 & - & - & \cdots & - \\
\hline
- & - & \cdots & - & 1 & 1 & \cdots & 1 & m & 1 & \cdots & 1 & 3 & 3 & \cdots & 3 \\
- & - & \cdots & - & 1 & 1 & \cdots & 1 & 1 & m & \cdots & 1 & 3 & 3 & \cdots & 3 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
- & - & \cdots & - & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & m & 3 & 3 & \cdots & 3 \\
\hline
1 & 1 & \cdots & 1 & - & - & \cdots & - & 3 & 3 & \cdots & 3 & m & 1 & \cdots & 1 \\
1 & 1 & \cdots & 1 & - & - & \cdots & - & 3 & 3 & \cdots & 3 & 1 & m & \cdots & 1 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
1 & 1 & \cdots & 1 & - & - & \cdots & - & 3 & 3 & \cdots & 3 & 1 & 1 & \cdots & m
\end{array}\right],$$

where $m = 2v = h + i + j + k$. Then by the determinant simplification theorem [10],

$$\det\Delta(h,i,j,k,m) = (m-1)^{m-4}\begin{vmatrix} m-1+h & 3h & -h & h \\ 3i & m-1+i & i & -i \\ -j & j & m-1+j & 3j \\ k & -k & 3k & m-1+k \end{vmatrix}$$

and

$$\begin{aligned}\det\Delta(h,i,j,k,m) = (m-1)^{(m-4)}[&(m-1)^4 + (m-1)^3(i+j+h+k) \\ &- 8(m-1)^2(jk+ih) - 16(m-1)(jk(i+h)+ih(j+k))].\end{aligned}$$

The $(2v-1)\times(2v-1)$ minors are obtained by removing a row and column from A to get D. We note that this means the number of rows becomes $m-1$ instead of m. The number of columns being reduced by one means we have one of $h-1,i,j,k$, or $h,i-1,j,k$, or $h,i,j-1,k$, or $h,i,j,k-1$, as the number of columns of each type. Thus $\det DD^T$ is $\det\Delta(h-1,i,j,k,m-1)$ or $\det\Delta(h,i-1,j,k,m-1)$ or $\det\Delta(h,i,j-1,k,m-1)$ or $\det\Delta(h,i,j,k-1,m-1)$.

Notation 2. We use the notation M_j to denote a $j\times j$ minor of A. We use "$-$" to denote "-1" throughout this paper.

Notation 3. In the work that follows we simplify the typesetting by defining two expressions $\mathcal{T}$ and $\mathcal{P}$:

$$\begin{aligned}\mathcal{T} &= 2^{s^2+s+1}s^{s^2+s}(s+1)^{s^2+s} = 2(2v-2)^{v-1},\\ \mathcal{P} &= 2s^2+2s+1 = 2v-1.\end{aligned}$$

LEMMA 1. *The $(2v-1)\times(2v-1)$ minors of the D-optimal designs of the KKSSS series are*

$$\frac{s}{s+1}\mathcal{T},\quad \frac{s+1}{s}\mathcal{T},\quad \frac{s^2+s+1}{s(s+1)}\mathcal{T},\quad \mathcal{T},$$

where $\mathcal{T} = 2^{s^2+s+1}s^{s^2+s}(s+1)^{s^2+s}$.

Proof. Here we use the $(1,-1)$ incidence matrices of the $2-\{s^2+s+1; \frac{s(s-1)}{2}, \frac{s(s+1)}{2}; \frac{s(s-1)}{2}\}$ SDS. By the reasoning above, with $v = s^2+s+1$, $h = \frac{(s+1)(s+2)}{2}$, $i = \frac{s(s-1)}{2}$, $j = \frac{s(s+1)}{2}$, $k = \frac{s^2+s+2}{2}$, $m = 2s^2+2s+2$ substituted into $\det \Delta(h-1,i,j,k,m-1)$, $\det \Delta(h,i,j-1,k,m-1)$, $\det \Delta(h,i-1,j,k,m-1)$, and $\det \Delta(h,i,j,k-1,m-1)$ we obtain the result.

Specifically the $(2v-1)\times(2v-1)$ minor is the square root of the determinant and is given by one of the following:

(1) $\det \Delta(h-1,i,j,k,m-1) = 2^{s^2+s+1}\ s^{s^2+s+1}\ (s+1)^{s^2+s-1}$.
(2) $\det \Delta(h,i-1,j,k,m-1) = 2^{s^2+s+1}\ s^{s^2+s-1}\ (s+1)^{s^2+s+1}$.
(3) $\det \Delta(h,i,j-1,k,m-1) = 2^{s^2+s+1}\ s^{s^2+s-1}\ (s+1)^{s^2+s-1}\ (s^2+s+1)$.
(4) $\det \Delta(h,i,j,k-1,m-1) = 2^{s^2+s+1}\ s^{s^2+s}\ (s+1)^{s^2+s}$. □

2.2. Minors of size $(2v-2)$. Now remove two rows and two columns of A. We have not included the generic matrix in expanded form, except for two cases, but moved straight to the determinant after it has been simplified using the determinant simplification theorem [10]. Thus the determinant of a submatrix of A obtained by removing two rows and two columns is $(2v-2)^{v-5}\sqrt{\det D}$, where

$$D = \begin{bmatrix} 2v-2 & 2u_2 & 2u_3 & 4u_4 & -2u_5 & 0 & 0 & 2u_8 \\ 2u_1 & 2v-2 & 4u_3 & 2u_4 & 0 & -2u_6 & 2u_7 & 0 \\ 2u_1 & 4u_2 & 2v-2 & 2u_4 & 0 & 2u_6 & -2u_7 & 0 \\ 4u_1 & 2u_2 & 2u_3 & 2v-2 & 2u_5 & 0 & 0 & -2u_8 \\ -2u_1 & 0 & 0 & 2u_4 & 2v-2 & 2u_6 & 2u_7 & 4u_8 \\ 0 & -2u_2 & 2u_3 & 0 & 2u_5 & 2v-2 & 4u_7 & 2u_8 \\ 0 & 2u_2 & -2u_3 & 0 & 2u_5 & 4u_6 & 2v-2 & 2u_8 \\ 2u_1 & 0 & 0 & -2u_4 & 4u_5 & 2u_6 & 2u_7 & 2v-2 \end{bmatrix}.$$

Diagrammatically, we have used the matrix form

$$\begin{bmatrix} A_1 & A_2 \\ A_2^T & -A_1^T \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}.$$

For Case I both rows and columns are removed from A_1; for Case II one row is from A_1 and one from A_3, but both columns are from A_1; for Case III one row is from A_1 and one from A_3, and one column is from A_1 and one column is from A_2.

To calculate the minors of size $(2v-2)$ we distinguish three major cases. This leads to the following seven subcases:

Case Ia. $\begin{bmatrix} x & y \\ x & \bar{y} \end{bmatrix}$, where the (1,1) and the (2,1) elements have the same sign, the (1,2) element and the (2,2) element have opposite signs, and the inner product of row one and two with each other is 2.

Case Ib. $\begin{bmatrix} x & y \\ x & y \end{bmatrix}$, where the (1,1) and the (2,1) elements have the same sign, the (1,2) element and the (2,2) element have the same signs, and the inner product of rows one and two with each other is +2.

Case Ic. $\begin{bmatrix} x & \bar{y} \\ \bar{x} & y \end{bmatrix}$, where the (1,1) and the (2,1) elements have opposite sign, the (1,2) element and the (2,2) element have opposite signs, and the inner product of row one and two with each other is 2.

Case IIa. $\begin{bmatrix} x & y \\ x & \bar{y} \end{bmatrix}$, where the (1,1) element and the (2,1) element have the same signs, the (1,2) element and the (2,2) element have different signs, and the inner product of rows one and two with each other is zero.

TABLE 1

2×2 submatrix	Number of Rows of Each Type Ia							
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$\begin{matrix}1 & 1\\ 1 & -\end{matrix}$	$\lambda_1 - 1$	$k_1 - \lambda_1 - 1$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & 1\\ - & -\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1 - 1$	$v_1 + \lambda_1 - 1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}1 & -\\ - & 1\end{matrix}$	λ_1	$k_1 - \lambda_1 - 1$	$k_1 - \lambda_1 - 1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$

TABLE 2

2×2 subsquare	Number of Rows of Each Type Ib							
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$\begin{matrix}1 & 1\\ 1 & 1\end{matrix}$	$\lambda_1 - 2$	$k_1 - \lambda_1$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & 1\\ - & 1\end{matrix}$	λ_1	$k_1 - \lambda_1 - 2$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}1 & -\\ 1 & -\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1 - 2$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & -\\ - & -\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1$	$v_1 + \lambda_1 - 2$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$

TABLE 3

2×2 subsquare	Number of Rows of Each Type IIa							
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$\begin{matrix}1 & 1\\ 1 & -\end{matrix}$	$\lambda_1 - 1$	$k_1 - \lambda_1$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1 - 1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & 1\\ - & -\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1 - 1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1 - 1$
$\begin{matrix}1 & -\\ 1 & 1\end{matrix}$	λ_1	$k_1 - \lambda_1 - 1$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1 - 1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & -\\ - & 1\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1$	$v_1 + \lambda_1 - 1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1 - 1$	$v_2 + \lambda - \lambda_1$

TABLE 4

2×2 subsquare	Number of Rows of Each Type IIb							
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$\begin{matrix}1 & 1\\ 1 & 1\end{matrix}$	$\lambda_1 - 1$	$k_1 - \lambda_1$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1 - 1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & 1\\ - & 1\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1 - 1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1 - 1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}1 & -\\ 1 & -\end{matrix}$	λ_1	$k_1 - \lambda_1 - 1$	$k_1 - \lambda_1$	$v_1 + \lambda_1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1 - 1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1$
$\begin{matrix}- & -\\ - & -\end{matrix}$	λ_1	$k_1 - \lambda_1$	$k_1 - \lambda_1$	$v_1 + \lambda_1 - 1$	$\lambda - \lambda_1$	$k_2 - \lambda + \lambda_1$	$k_2 - \lambda + \lambda_1$	$v_2 + \lambda - \lambda_1 - 1$

Case IIb. $\left[\begin{smallmatrix} x & y \\ x & y \end{smallmatrix}\right]$, where the (1,1) element and the (2,1) element have the same signs, the (1,2) element and the (2,2) element also have the same sign, and the inner product of row one and two with each other is zero.

Case IIIa. $\left[\begin{smallmatrix} x & y \\ x & \bar{y} \end{smallmatrix}\right]$, where one of the columns in the submatrix has two identical elements and the other has two different elements.

Case IIIb. $\left[\begin{smallmatrix} x & y \\ x & y \end{smallmatrix}\right]$, where both columns in the submatrix have identical elements.

In [10] we analyzed which 2×2 submatrices gave independent values for the distribution of rows in the minors of order $2v - 2$. These are summarized for the KKSSS family in Tables 1, 2, 3, 4. Case III is covered by Tables 5 and 6. Set $v_1 = v - 2k_1$, $v_2 = v - 2k_2$ in Tables 1–4.

TABLE 5

2×2 subsquare	Number of Rows of Each Type IIIa							
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$\begin{matrix}1 & 1\\ 1 & -\end{matrix}$	$\rho-1$	$k_1-\rho$	$k_2-\rho$	$v-k_1-k_2+\rho$	$k_2-\rho$	$\rho-1$	$v-k_1-k_2+\rho$	$k_1-\rho$
$\begin{matrix}1 & -\\ 1 & 1\end{matrix}$	ρ	$k_1-\rho-1$	$k_2-\rho$	$v-k_1-k_2+\rho$	$k_2-\rho-1$	ρ	$v-k_1-k_2+\rho$	$k_1-\rho$
$\begin{matrix}- & -\\ - & 1\end{matrix}$	ρ	$k_1-\rho$	$k_2-\rho$	$v-k_1-k_2+\rho-1$	$k_2-\rho$	ρ	$v-k_1-k_1+\rho-1$	$k_1-\rho$
$\begin{matrix}- & 1\\ 1 & 1\end{matrix}$	ρ	$k_1-\rho$	$k_2-\rho-1$	$v-k_1-k_2+\rho$	$k_2-\rho-1$	ρ	$v-k_1-k_2+\rho$	$k_1-\rho$
$\begin{matrix}- & -\\ 1 & -\end{matrix}$	ρ	$k_1-\rho$	$k_2-\rho$	$v-k_1-k_2+\rho-1$	$k_2-\rho$	$\rho-1$	$v-k_1-k_2+\rho$	$k_1-\rho$
$\begin{matrix}1 & -\\ - & -\end{matrix}$	ρ	$k_1-\rho-1$	$k_2-\rho$	$v-k_1-k_2+\rho$	$k_2-\rho$	ρ	$v-k_1-k_2+\rho$	$k_1-\rho-1$

TABLE 6

2×2 subsquare	Number of Rows of Each Type IIIb							
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$\begin{matrix}1 & 1\\ 1 & 1\end{matrix}$	$\rho-1$	$k_1-\rho$	$k_2-\rho$	$v-k_1-k_2+\rho$	$k_2-\rho-1$	ρ	$v-k_1-k_2+\rho$	$k_1-\rho$
$\begin{matrix}1 & -\\ 1 & -\end{matrix}$	ρ	$k_1-\rho-1$	$k_2-\rho$	$v-k_1-k_2+\rho$	$k_2-\rho$	$\rho-1$	$v-k_1-k_2+\rho$	$k_1-\rho$
$\begin{matrix}- & 1\\ - & 1\end{matrix}$	ρ	$k_1-\rho$	$k_2-\rho-1$	$v-k_1-k_2+\rho$	$k_2-\rho$	ρ	$v-k_1-k_1+\rho-1$	$k_1-\rho$
$\begin{matrix}- & -\\ - & -\end{matrix}$	ρ	$k_1-\rho$	$k_2-\rho$	$v-k_1-k_2+\rho-1$	$k_2-\rho$	ρ	$v-k_1-k_2+\rho$	$k_1-\rho-1$

Case Ia. To illustrate the derivation of the tables such as Table 1 we give Case Ia as an example.

$\begin{matrix}1 & y\\ 1 & \bar{y}\end{matrix}$		$\begin{matrix}- & y\\ - & \bar{y}\end{matrix}$		Inner Product of rows is 2
$\begin{matrix}1 & 1\\ \vdots & \vdots\\ 1 & 1\end{matrix}$	λ_1-1	$\begin{matrix}1 & 1\\ \vdots & \vdots\\ 1 & 1\end{matrix}$	λ_1	
$\begin{matrix}1 & -\\ \vdots & \vdots\\ - & 1\end{matrix}$	$k_1-\lambda_1-1$	$\begin{matrix}1 & -\\ \vdots & \vdots\\ - & 1\end{matrix}$	$k_1-\lambda_1$	$v-2$ rows which have inner product 2
$\begin{matrix}\vdots & \vdots\\ - & 1\\ - & -\end{matrix}$	$k_1-\lambda_1$	$\begin{matrix}\vdots & \vdots\\ - & 1\\ - & -\end{matrix}$	$k_1-\lambda_1-1$	with rows one and two
$\begin{matrix}\vdots & \vdots\\ - & -\end{matrix}$	$v-2k_1+\lambda_1$	$\begin{matrix}\vdots & \vdots\\ - & -\end{matrix}$	$v-2k_1+\lambda_1-1$	
$\begin{matrix}1 & 1\\ \vdots & \vdots\\ 1 & 1\end{matrix}$	λ_2	$\begin{matrix}1 & 1\\ \vdots & \vdots\\ 1 & 1\end{matrix}$	λ_2	$\lambda=\lambda_1+\lambda_2$
$\begin{matrix}1 & -\\ \vdots & \vdots\\ - & 1\end{matrix}$	$k_2-\lambda_2$	$\begin{matrix}1 & -\\ \vdots & \vdots\\ - & 1\end{matrix}$	$k_2-\lambda_2$	v rows which have inner product 0
$\begin{matrix}\vdots & \vdots\\ - & 1\\ - & -\end{matrix}$	$k_2-\lambda_2$	$\begin{matrix}\vdots & \vdots\\ - & 1\\ - & -\end{matrix}$	$k_2-\lambda_2$	with rows one and two
$\begin{matrix}\vdots & \vdots\\ - & -\end{matrix}$	$v-2k_2+\lambda_2$	$\begin{matrix}\vdots & \vdots\\ - & -\end{matrix}$	$v-2k_2+\lambda_2$	

Case III. To help understand Case III we recall that in this case one column removed comes from the columns with $k_1 + k_2$ ones per column and the other from the columns with $v - k_2 + k_1$ ones per column in the original design. This means the generic form of these two columns is

1		1	
		$\vdots$	
1		$\vdots$	ρ
1		1	
1	k_1	$-$	
$\vdots$		$\vdots$	
$\vdots$		$\vdots$	$k_1 - \rho$
1		$-$	
$-$		1	
$\vdots$		$\vdots$	
$\vdots$		$\vdots$	$k_2 - \rho$
$-$	$v - k_1$	1	
$-$		$-$	
$\vdots$		$\vdots$	
$\vdots$		$\vdots$	$v - k_1 - k_2 + \rho$
$-$		$-$	
1		1	
		$\vdots$	
1		$\vdots$	$k_2 - \rho$
1		1	
1	k_2	$-$	
$\vdots$		$\vdots$	
$\vdots$		$\vdots$	ρ
1		$-$	
$-$		1	
$\vdots$		$\vdots$	
$\vdots$		$\vdots$	$v - k_1 - k_2 + \rho$
$-$	$v - k_2$	1	
$-$		$-$	
$\vdots$		$\vdots$	
$\vdots$		$\vdots$	$k_1 - \rho$
$-$		$-$	

Note that they have inner product zero.

The results given are quite general for the minors of size $2v-2$ constructed from any $2-\{v; k_1, k_2; \lambda\}$ SDS. We now apply these results to the special case of the $2-\{s^2+s+1; \frac{s(s-1)}{2}, \frac{s(s+1)}{2}; \frac{s(s-1)}{2}\}$ SDS.

LEMMA 2. *The $(2v-2)\times(2v-2)$ minors of the D-optimal design of the KKSSS series are*

$$0,\quad \frac{1}{s^2}\mathcal{T},\quad \frac{1}{(s+1)^2}\mathcal{T},\quad \frac{1}{s(s+1)}\mathcal{T},\quad \frac{1}{s^2(s+1)^2}\mathcal{T},\quad \frac{1}{s(s+1)^2}\mathcal{T},$$

$$\frac{1}{s^2(s+1)}\mathcal{T},\quad \frac{2s+1}{s^2(s+1)^2}\mathcal{T},\quad \frac{s^2+1}{s^2(s+1)^2}\mathcal{T},\quad \frac{s^2+s+1}{s^2(s+1)^2}\mathcal{T},\quad \frac{s^2+2s+2}{s^2(s+1)^2}\mathcal{T},$$

where $\mathcal{T} = 2^{s^2+s+1}s^{s^2+s}(s+1)^{s^2+s}$.

Proof. Here $\lambda = \frac{1}{2}s(s-1)$, $k_1 = \frac{1}{2}s(s-1)$, $k_2 = \frac{1}{2}s(s+1)$, and $v = s^2+s+1$. The expressions for u_i, $i = 1, \ldots, 8$, were calculated in each case. Maple was then used to evaluate the determinant for D giving the required result. Case Ia gives the values $2^{10}s^4(s+1)^8$, $2^{10}s^8(s+1)^4$, and $2^{10}s^4(2s+1)^2(s+1)^4$. Case IIa gives the values $2^{10}s^4(s+1)^8$, $2^{10}s^6(s+1)^6$, $2^{10}s^4(s^2+s+1)^2(s+1)^4$, and $2^{10}s^8(s+1)^4$.

Case Ib gives the value zero for the determinant. Case IIb gives the value $2^{10}s^6(s+1)^4$ and the value zero for the determinant.

Case IIIa gives the values $2^{10}s^4(s^2+2s+2)^2(s+1)^4$, $2^{10}s^6(s+1)^6$, $2^{10}s^8(s+1)^4$, $2^{10}s^4(s^2+1)^2(s+1)^4$, and $2^{10}s^4(s+1)^8$, whereas Case IIIb gives the values $2^{10}s^4(s+1)^4$, $2^{10}s^4(s+1)^6$, and $2^{10}s^6(s+1)^4$.

Taking the square root and multiplying by $(2s^2+2s)^{s^2+s-4}$ gives the required result. □

Remark 1. The values $\frac{1}{s^2(s+1)^2}\mathcal{T}$, $\frac{1}{s(s+1)^2}\mathcal{T}$, $\frac{1}{s^2(s+1)}\mathcal{T}$ all arise from a 2×2 corner block

$$\begin{matrix} x & y \\ x & y \end{matrix},$$

which cannot occur as the leading 2×2 block when GECP is done here. Also the value $\frac{2s+1}{s^2(s+1)^2}\mathcal{T}$ arises from a 2×2 corner block

$$\begin{matrix} x & \bar{y} \\ \bar{x} & y \end{matrix},$$

which cannot occur as the leading 2×2 block when GECP is done here. □

3. Pivot structure for the *KKSSS* family of *D*-optimal designs.

Conjecture (growth conjecture for the KKSSS family). Let A be a $2v\times 2v$ CP D-optimal design of the KKSSS family which is constructed from $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ SDS. Reduce A by GECP and recall that $\mathcal{P}=2s^2+2s+1$. Then we conjecture the following:

(i) $g(v,A)=\frac{s+1}{s}\mathcal{P}$, or $\frac{s}{s+1}\mathcal{P}$, or $\frac{s(s+1)}{s^2+s+1}\mathcal{P}$, or $\mathcal{P}$;
(ii) the last pivot is equal to $\frac{s+1}{s}\mathcal{P}$, or $\frac{s}{s+1}\mathcal{P}$, or $\frac{s(s+1)}{s^2+s+1}\mathcal{P}$, or $\mathcal{P}$;
(iii) the second-to-last pivot can take the values given in Table 8;
(iv) every pivot before the last has magnitude at most $2v$;
(v) the first four pivots are equal to $1,2,2,4$;
(vi) the fifth pivot may be 2 or 3.

We prove (ii) and (iii) in this paper. (v) and (vi) were proved for Brouwer's *SBIBD*$(2s^2+2s+1,s^2,\frac{1}{2}s(s-1))$ in [9] and we also show they hold for the KKSSS family.

We recall that for any CP matrix A of *SBIBD*(v,k,λ), the two last pivots p_v and p_{v-1} are given from the formulae

$$p_v=\frac{A(v)}{A(v-1)},\quad p_{v-1}=\frac{A(v-1)}{A(v-2)}. \tag{7}$$

THEOREM 2. *Let A be the $2v\times 2v$ D-optimal design of the KKSSS family. Reduce A by GECP. Then the last pivot, p_{2v}, is $\frac{s+1}{s}\mathcal{P}$, or $\frac{s}{s+1}\mathcal{P}$, or $\frac{s(s+1)}{s^2+s+1}\mathcal{P}$, or $\mathcal{P}$. The only possible values of the second-to-last pivot, p_{2v-1}, are those given in Table* 8.

Proof. From (4), (6), and Lemma 1 we have for the D-optimal design made using $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ SDS the results given in Table 7, where the first row gives the values of M_{2v}, the first column gives the values of M_{2v-1}, and the entries are $p_{2v}=\frac{M_{2v}}{M_{2v-1}}$.

From (4) and Lemmas 1 and 2 we have for the D-optimal design made using $2-\{s^2+s+1;\frac{s(s-1)}{2},\frac{s(s+1)}{2};\frac{s(s-1)}{2}\}$ SDS the results given in Table 8, where the first row gives the values of M_{2v-1}, the first column gives the values of M_{2v-2}, and the other entries are the only possible values of $p_{2v-1}=\frac{M_{2v-1}}{M_{2v-2}}$.

Remark 2. The entries marked $*$ in Tables 7 and 8 are those obtained in experiments. It is not known whether all the values shown in Tables 7 and 8 can actually

TABLE 7
The only possible values of p_{2v}.

M_{2v}	$(2s^2+2s+1)\mathcal{T}$
M_{2v-1}	
$\frac{s}{s+1}\mathcal{T}$	$\frac{s+1}{s}\mathcal{P}$
$\frac{s+1}{s}\mathcal{T}$	$\frac{s}{s+1}\mathcal{P}$ *
$\frac{s^2+s+1}{s(s+1)}\mathcal{T}$	$\frac{s(s+1)}{s^2+s+1}\mathcal{P}$ *
$\mathcal{T}$	$\mathcal{P}$

TABLE 8
The only possible values of p_{2v-1}.

M_{2v-1}	$\frac{s}{s+1}\mathcal{T}$	$\frac{s+1}{s}\mathcal{T}$	$\frac{s^2+s+1}{s(s+1)}\mathcal{T}$	$\mathcal{T}$
M_{2v-2}				
$\frac{1}{s^2}\mathcal{T}$	$\frac{s^3}{s+1}$	$\frac{s}{s+1}$	$\frac{s(s^2+s+1)}{s+1}$	s^2
$\frac{1}{(s+1)^2}\mathcal{T}$	$s(s+1)$ *	$\frac{(s+1)^3}{s}$	$\frac{(s^2+s+1)(s+1)}{s}$	$(s+1)^2$
$\frac{1}{s(s+1)}\mathcal{T}$	s^2	$(s+1)^2$ *	s^2+s+1 *	$s(s+1)$
$\frac{s^2+1}{s^2(s+1)^2}\mathcal{T}$	$\frac{s^3(s+1)}{s^2+1}$	$\frac{s(s+1)}{s^2+1}$	$\frac{s(s+1)(s^2+s+1)}{s^2+1}$	$\frac{s^2(s+1)^2}{s^2+1}$
$\frac{s^2+s+1}{s^2(s+1)^2}\mathcal{T}$	$\frac{s^3(s+1)}{s^2+s+1}$	$\frac{s(s+1)^3}{s^2+s+1}$ *	$s(s+1)$	$\frac{s^2(s+1)^2}{s^2+s+1}$
$\frac{s^2+2s+2}{s^2(s+1)^2}\mathcal{T}$	$\frac{s^3(s+1)}{s^2+2s+1}$	$\frac{s(s+1)^3}{s^2+2s+2}$ *	$\frac{s(s+1)(s^2+s+1)}{s^2+2s+2}$	$\frac{s^2(s+1)^2}{s^2+2s+2}$

TABLE 9
Numerical values of p_{2v}.

$2v$	s	p_{2v}			
		$\frac{s+1}{s}\mathcal{P}$	$\frac{s}{s+1}\mathcal{P}$	$\frac{s(s+1)}{s^2+s+1}\mathcal{P}$	$\mathcal{P}$
14	2	19.5	$\frac{26}{3}$	$\frac{78}{7}$	13
26	3	$\frac{100}{3}$	$\frac{75}{4}$	$\frac{12\cdot25}{13}$	25
42	4	$\frac{5\cdot41}{4}$	$\frac{4\cdot41}{5}$	$\frac{20\cdot41}{21}$	41

occur as p_{2v} and p_{2v-1} when GECP is done to a matrix of KKSSS type. In particular, notice that the first value listed for p_{2v}, $\frac{s+1}{s}\mathcal{P}$, is greater than $2v$, but in experiments using GECP on such matrices we never saw it arise. □

In Tables 9 and 10 we give some values for the two last pivots, which we obtained in experiments, for the family KKSSS.

Remark 3. We experimented with $2v = 14$ by testing 100,000 equivalent transformations. The theoretical values for M_{2v-1} are $2^{14}\cdot 3^5$, $2^{12}\cdot 3^7$, $2^{12}\cdot 3^5\cdot 7$, and $2^{13}\cdot 3^6$. In our calculations we always found $p_{2v} = \frac{26}{3}$ and $\frac{78}{7}$. This leaves as an open problem the existence of a 14×14 matrix having growth equal to 19.5.

The next result is easy to prove using a counting argument and noting that the inner product of every pair of rows is $+1$ to see that the design always contains a 4×4 Hadamard matrix.

TABLE 10
Numerical values of p_{2v-1}.

$2v$	s	p_{2v-1}				
		$s(s+1)$	$(s+1)^2$	$\frac{s(s+1)^3}{s^2+s+1}$	$\frac{s(s+1)^3}{s^2+2s+2}$	s^2+s+1
14	2	6	9	$\frac{54}{7}$	5.4	7
26	3	12	16	$\frac{3\cdot 4^3}{13}$	$\frac{3\cdot 4^3}{17}$	13
42	4	20	25	$\frac{4\cdot 5^3}{21}$	$\frac{4\cdot 5^3}{26}$	26

TABLE 11
Growth factors and pivots patterns for small CP KKSSS designs.

s	$2v$	Growth	Pivot pattern
2	14	$\frac{78}{7}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 6, \frac{78}{7})$
2	14	$\frac{26}{3}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 5.4, \frac{26}{3})$
2	14	$\frac{26}{3}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 6, \frac{26}{3})$
2	14	$\frac{26}{3}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{16}{5}, \ldots, 6, \frac{26}{3})$
2	14	$\frac{26}{3}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 9, \frac{26}{3})$
2	14	$\frac{26}{3}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, \frac{54}{7}, \frac{26}{3})$
3	26	$\frac{75}{4}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 12, \frac{75}{4})$
3	26	$\frac{75}{4}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, \frac{3\cdot 4^3}{17}, \frac{75}{4})$
3	26	$\frac{12\cdot 25}{13}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 12, \frac{12\cdot 25}{13})$
3	26	$\frac{75}{4}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, \frac{3\cdot 4^3}{13}, \frac{75}{4})$
3	26	$\frac{75}{4}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 16, \frac{75}{4})$
3	26	$\frac{12\cdot 25}{13}$	$(1, 2, 2, 4, 3, \frac{10}{3}, \frac{18}{5}, \ldots, 13, \frac{12\cdot 25}{13})$

PROPOSITION 1 (see [10]). *Let A be the $2v \times 2v$ $(1, -1)$ incidence matrix of an SBIBD of the KKSSS family. Reduce A by GECP. Then the magnitudes of the first four pivots are $1, 2, 2,$ and 4; the magnitude of $|a_{55}^{(4)}|$ is 2 or 3.*

The values presented in Table 11 are those we saw in experiments when we used GECP on some small KKSSS matrices. The first seven pivots and the last two are presented. All the other intermediate pivots take a variety of values. At least 54 different pivot structures were detected for $2v = 14$ and over 20,000 for $2v = 26$.

Remark 4. We note that experimentally, for $s = 1$, we always found the unique pivot structure $(1, 2, 2, 4, 3, \frac{10}{3})$.

Acknowledgments. We would like to thank Professor Nick Higham and two anonymous referees for their valuable comments and suggestions, which led to a significant improvement in the presentation of the paper.

REFERENCES

[1] A. E. BROUWER, *An Infinite Series of Symmetric Designs*, Report ZW 202/83, Mathematisch Centrum, Amsterdam, 1983.
[2] C. W. CRYER, *Pivot size in Gaussian elimination*, Numer. Math., 12 (1968), pp. 335–345.
[3] J. DAY AND B. PETERSON, *Growth in Gaussian elimination*, Amer. Math. Monthly, 95 (1988), pp. 489–513.

[4] H. Ehlich, *Determinantenabschätzungen für binäre matrizen*, Math. Z., 83 (1964), pp. 123–132.
[5] A. V. Geramita and J. Seberry, *Orthogonal Designs: Quadratic forms and Hadamard matrices*, Marcel Dekker, New York, Basel, 1979.
[6] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[7] H. Kharaghani, *D-optimal matrices*, in The CRC Handbook of Combinatorial Designs, C. J. Colbourn and J. H. Dinitz, eds., CRC Press, Boca Raton, FL, 1996, pp. 321–323.
[8] C. Koukouvinos, S. Kounias, and J. Seberry, *Supplementary difference sets and optimal designs*, Discrete Math., 88 (1991), pp. 49–58.
[9] C. Koukouvinos, M. Mitrouli, and J. Seberry, *Values of minors of* $(1,-1)$ *incidence matrices of SBIBDs and their application to the growth problem*, Des. Codes Cryptogr., 23 (2001), pp. 267–281.
[10] C. Koukouvinos, M. Mitrouli, and J. Seberry, *Values of minors of an infinite family of D-optimal designs and their application to the growth problem*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 1–14.
[11] C. Koukouvinos, M. Mitrouli, and J. Seberry, *Growth in Gaussian elimination for weighing matrices* $W(n,n-1)$, Linear Algebra Appl., 306 (2000), pp. 189–202.
[12] J. Singer, *A theorem in finite projective geometry and some applications to number theory*, Trans. Amer. Math. Soc., 43 (1938), pp. 377–385.
[13] E. Spence, *Skew-Hadamard matrices of the Goethals-Seidel type*, Canad. J. Math., 27 (1975), pp. 555–560.
[14] A. L. Whiteman, *A family of D-optimal designs*, Ars Combin., 30 (1990), pp. 23–26.
[15] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 728–746

COMMUTATION RELATIONS FOR TOEPLITZ AND HANKEL MATRICES*

CAIXING GU† AND LINDA PATTON†

Abstract. Let A, B, C, and D be Toeplitz matrices. The main theorem of this paper determines if $AB - CD$ is Toeplitz. This theorem is used to prove a variety of new and previously known algebraic results about Hankel and Toeplitz matrices. For instance, the set of all Hankel matrices which commute with a given Hankel matrix is parametrized. This and similar classification results are straightforward to prove using our approach since it naturally produces the set of matrices to solve each problem.

Key words. Toeplitz, Hankel, structured matrices, commutator

AMS subject classifications. 15A27, 47B35

PII. S0895479800377320

1. Introduction. The main result of this paper provides a procedure for discovering algebraic properties of Toeplitz-like matrices. We apply this procedure to obtain simple proofs of some known results such as classifying commuting Toeplitz matrices (by Gel'fgat [4]), normal Toeplitz matrices (by Farenick et al. [2], Gel'fgat [3], Ikramov [8], [9], Ikramov and Chugunov [10], and Ito [12]), and invertible Toeplitz matrices with Toeplitz inverses (by Huang and Cline [7], Greville [5], and Shalom [14]). Our approach leads naturally to these results because the necessary and sufficient conditions to solve each problem are automatically derived. We observe that reversing the order of either the columns or the rows transforms a Toeplitz matrix into a Hankel matrix and vice versa. Therefore, our method also yields new algebraic results about Hankel matrices. For example, we show which pairs of Hankel matrices commute and which Hankel matrices are normal. The corresponding necessary and sufficient conditions are more complicated than those in the analogous Toeplitz results. However, with our method, these theorems are just as easy to prove.

These algebraic results can all be formulated in terms of products of Toeplitz matrices. Given Toeplitz matrices A, B, C, and D, our main result determines if the matrix $AB - CD$ is Toeplitz. The necessary and sufficient condition is a rank two matrix equation involving tensor products of the vectors defining A, B, C, and D. A necessary and sufficient condition for $AB - CD = 0$ is also provided.

Our theorem is proved using the special structure of the displacement matrix of a Toeplitz matrix. Let Z be the matrix consisting of zeros except for ones along the subdiagonal. For any square matrix M we define the displacement matrix of M to be $\Delta(M) = M - ZMZ^*$. The matrix M is zero if and only if $\Delta(M)$ is zero (see Lemma 2.1). Furthermore, a matrix M is Toeplitz if and only if $\Delta(M)$ has a particular structure which is of at most rank two (see Lemma 2.2).

If A, B, C, and D are Toeplitz, then the displacement matrix of $AB - CD$ is the sum of six rank one matrices. In applications, relationships between A, B, C, and D simplify this sum so that necessary and sufficient conditions for $AB - CD$ to be

*Received by the editors August 22, 2000; accepted for publication (in revised form) by A. C. M. Ran July 30, 2002; published electronically January 31, 2003.
http://www.siam.org/journals/simax/24-3/37732.html

†Department of Mathematics, California Polytechnic State University, San Luis Obispo, CA 93405 (cgu@calpoly.edu, lpatton@calpoly.edu). The research of the first author was partially supported by NSF grant DMS-9706838 and the SFSG grant of California Polytechnic State University.

Toeplitz or $AB = CD$ can be obtained. We remark that our approach will also be useful for discovering algebraic relations of more general classes of structured matrices by using the appropriate displacement matrices. See [6] and [13] for many examples of structured matrices and their related computational issues.

Many similar algebraic properties of (infinite) Toeplitz operators were derived by Brown and Halmos in [1]. For instance, they found necessary and sufficient conditions for Toeplitz operators to be unitary or normal. The conditions needed for similar classifications of finite Toeplitz matrices and Hankel matrices are more complicated. For instance, Brown and Halmos proved that a product of two Toeplitz operators is zero if and only if one of the factors is zero. However, there exist simple examples of nonzero Toeplitz matrices whose product is zero. Here, for a given Toeplitz matrix A we parametrize all the Toeplitz matrices B whose product with A is zero.

The remainder of the paper is organized as follows. Notation and facts about Toeplitz matrices, Hankel matrices, and displacement matrices are presented in section 2. The main theorem with applications to Toeplitz matrices is shown in section 3; results about Hankel matrices can be found in section 4.

2. Notation and lemmas. Throughout the paper, we will number the rows and columns of $n \times n$ matrices from 0 to $n-1$. Thus we will use the notation $e_0, e_1, \ldots, e_{n-1}$ for the standard basis vectors of C^n. Assume all matrices are $n \times n$ unless otherwise specified.

Let I be the identity matrix and let Z denote the matrix consisting of zeros, except for ones along the subdiagonal. Finally, let

$$P = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Note that $P^2 = I$.

Our results refer separately to the vectors which define the upper and lower triangular parts of Toeplitz matrices, so our notation for the matrices also refers to these vectors.

If $a = \begin{pmatrix} 0 & a_1 & \cdots & a_{n-1} \end{pmatrix}^T$ and $\alpha = \begin{pmatrix} 0 & \alpha_1 & \cdots & \alpha_{n-1} \end{pmatrix}^T$ are vectors in C^n, then let $T(a, \alpha)$ denote the Toeplitz matrix

$$T(a,\alpha) = \begin{bmatrix} 0 & \overline{\alpha_1} & \overline{\alpha_2} & \cdots & \overline{\alpha_{n-1}} \\ a_1 & 0 & \overline{\alpha_1} & \cdots & \overline{\alpha_{n-2}} \\ a_2 & a_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \overline{\alpha_1} \\ a_{n-1} & a_{n-2} & \cdots & a_1 & 0 \end{bmatrix},$$

and let $H(a, \alpha)$ denote the Hankel matrix

$$H(a,\alpha) = \begin{bmatrix} a_1 & a_2 & \cdots & a_{n-1} & 0 \\ a_2 & a_3 & \iddots & 0 & \overline{\alpha_1} \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ a_{n-1} & 0 & \iddots & \overline{\alpha_{n-3}} & \overline{\alpha_{n-2}} \\ 0 & \overline{\alpha_1} & \cdots & \overline{\alpha_{n-2}} & \overline{\alpha_{n-1}} \end{bmatrix}.$$

Since our results involve commutants, in many cases it will suffice to consider these Toeplitz and Hankel matrices with zero diagonals. In other cases, $T(a,\alpha)+a_0 I$ and $H(a,\alpha)+a_n P$ will describe the most general Toeplitz and Hankel matrices, respectively. With this notation,

$$T(a,\alpha)^* = T(\alpha,a), \quad H(a,\alpha)^* = H(\overline{a},\overline{\alpha}).$$

If $a = \begin{pmatrix} 0 & a_1 & \cdots & a_{n-1} \end{pmatrix}^T$, then we define

$$\widetilde{a} = \begin{pmatrix} 0 & \overline{a_{n-1}} & \cdots & \overline{a_1} \end{pmatrix}^T.$$

We will use the displacement matrix of M defined by

$$\Delta M = M - ZMZ^*$$

to determine whether a difference of matrix products is Toeplitz. See [6] and [13] for other types of displacement matrices.

The matrix ZMZ^* has zeroth row and column consisting of all zeros. The upper left $(n-1)\times(n-1)$ block of M is shifted diagonally to the lower right $(n-1)\times(n-1)$ block of ZMZ^*. That is, if

$$M = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} : C^{n-1} \oplus C \longrightarrow C^{n-1} \oplus C, \tag{2.1}$$

then

$$ZMZ^* = \begin{bmatrix} 0 & 0 \\ 0 & U_{11} \end{bmatrix} : C \oplus C^{n-1} \longrightarrow C \oplus C^{n-1}. \tag{2.2}$$

As Lemma 2.2 describes below, ΔM is particularly simple if M is Toeplitz. In fact ΔM is of rank two.

LEMMA 2.1. *Let M be an $n \times n$ matrix:*

$$M = \sum_{i=0}^{n-1} Z^i \,(\Delta M)\, Z^{i*}.$$

Proof. By definition,

$$\begin{aligned} \sum_{i=0}^{n-1} Z^i \,(\Delta M)\, Z^{i*} &= \sum_{i=0}^{n-1} Z^i \,(M - ZMZ^*)\, Z^{i*} \\ &= \sum_{i=0}^{n-1} \left(Z^i M Z^{i*} - Z^{i+1} M Z^{i+1*} \right) = M - Z^n M Z^{n*} = M, \end{aligned}$$

since $Z^n = 0$. ∎

Thus to determine when $M = 0$, it is sufficient to study the much simpler equation $\Delta M = 0$.

Recall that if x and y are two vectors in C^n, then the tensor product $x \otimes y$ is the rank one $n \times n$ matrix defined by

$$(x \otimes y)\, u = \langle u, y \rangle\, x \qquad \text{for all } u \in C^n.$$

LEMMA 2.2. *An $n \times n$ matrix M is Toeplitz if and only if there exist vectors u and v in C^n such that*

$$\Delta M = u \otimes e_0 + e_0 \otimes v.$$

Proof. This lemma is an immediate consequence of (2.1) and (2.2). If the diagonal of a Toeplitz matrix M consists of zeros, then u is the zeroth column and $\overline{v}$ is the zeroth row of M. □

LEMMA 2.3. *Let $c = \begin{pmatrix} 0 & c_1 & \cdots & c_{n-1} \end{pmatrix}^T$, $\gamma = \begin{pmatrix} 0 & \gamma_1 & \cdots & \gamma_{n-1} \end{pmatrix}^T$, $d = \begin{pmatrix} 0 & d_1 & \cdots & d_{n-1} \end{pmatrix}^T$, and $\delta = \begin{pmatrix} 0 & \delta_1 & \cdots & \delta_{n-1} \end{pmatrix}^T$ be vectors in C^n. If $C = T(c, \gamma) + c_0 I$ and $D = T(d, \delta) + d_0 I$, then*

$$\Delta(CD) = c \otimes \delta - \widetilde{\gamma} \otimes \widetilde{d} + [Cd + d_0 c + c_0 d_0 e_0] \otimes e_0 + e_0 \otimes [ZD^* Z^* \gamma + \overline{c_0}\delta]. \tag{2.3}$$

Proof. Let $\widehat{C} = T(c, \gamma)$ and $\widehat{D} = T(d, \delta)$. Then we have

$$\begin{aligned}
\Delta(CD) &= \Delta\left[\left(\widehat{C} + c_0 I\right)\left(\widehat{D} + d_0 I\right)\right] \\
&= \Delta\left[\widehat{C}\widehat{D} + c_0\widehat{D} + d_0\widehat{C} + c_0 d_0 I\right] \\
&= \Delta\widehat{C}\widehat{D} + c_0\Delta\widehat{D} + d_0\Delta\widehat{C} + c_0 d_0 \Delta I \\
&= \Delta\widehat{C}\widehat{D} + c_0[d \otimes e_0 + e_0 \otimes \delta] + d_0[c \otimes e_0 + e_0 \otimes \gamma] + c_0 d_0 (e_0 \otimes e_0). \qquad (2.4)
\end{aligned}$$

We applied Lemma 2.2 to the terms $\Delta\widehat{D}$ and $\Delta\widehat{C}$ in the above. Now

$$\begin{aligned}
\Delta\widehat{C}\widehat{D} &= \widehat{C}\widehat{D} - Z\widehat{C}\widehat{D}Z^* \\
&= \widehat{C}\widehat{D} - \widehat{C}Z\widehat{D}Z^* + \widehat{C}Z\widehat{D}Z^* - Z\widehat{C}[Z^*Z + e_{n-1} \otimes e_{n-1}]\widehat{D}Z^* \\
&= \widehat{C}\Delta\widehat{D} + \Delta\widehat{C}\left(Z\widehat{D}Z^*\right) - Z\widehat{C}[e_{n-1} \otimes e_{n-1}]\widehat{D}Z^* \\
&= \widehat{C}[d \otimes e_0 + e_0 \otimes \delta] + [c \otimes e_0 + e_0 \otimes \gamma]\left(Z\widehat{D}Z^*\right) - Z\widehat{C}e_{n-1} \otimes Z\widehat{D}^* e_{n-1} \\
&= \widehat{C}d \otimes e_0 + c \otimes \delta + c \otimes Z\widehat{D}^* Z^* e_0 + e_0 \otimes Z\widehat{D}^* Z^* \gamma - \widetilde{\gamma} \otimes \widetilde{d}. \qquad (2.5)
\end{aligned}$$

Combining (2.4) and (2.5) yields

$$\Delta CD = c \otimes \delta - \widetilde{\gamma} \otimes \widetilde{d} + \left[\widehat{C}d + c_0 d + d_0 c + c_0 d_0 e_0\right] \otimes e_0 + e_0 \otimes \left[Z\widehat{D}^* Z^* \gamma + \overline{c_0}\delta + \overline{d_0}\gamma\right].$$

Note that $Z^* e_0 = 0$, so the third term of the right side of (2.5) vanished. Also note that

$$\widehat{C}d + c_0 d = Cd \quad \text{and} \quad Z\widehat{D}^* Z^* \gamma + \overline{d_0}\gamma = ZD^* Z^* \gamma.$$

This completes the proof. □

Remark 2.4. The first two terms of the right side of (2.3) involve only the first through $(n-1)$th rows and columns of $\Delta(CD)$, while the remaining terms involve only the zeroth row and column of $\Delta(CD)$.

In the next section, we will compare our results with those involving infinite Toeplitz operators. By a Toeplitz operator T, we mean a bounded linear operator on l_2 with matrix representation

$$T = (a_{i-j})_{i,j=0}^{\infty}.$$

3. Toeplitz case. The first theorem in this section describes when a difference of products of Toeplitz matrices is Toeplitz. The necessary and sufficient condition is an equation involving the matrices' defining vectors. This idea can be used to solve a variety of problems about commutants of Toeplitz-like matrices.

THEOREM 3.1. *Let $a, \alpha, b, \beta, c, \gamma, d,$ and δ be vectors in C^n with 0 in the zeroth component. Let $A = T(a,\alpha) + a_0 I$, $B = T(b,\beta) + b_0 I$, $C = T(c,\gamma) + c_0 I$, and $D = T(d,\delta) + c_0 I$.*

(i) *$AB - CD$ or, equivalently,*

$$T(a,\alpha)T(b,\beta) - T(c,\gamma)T(d,\delta),$$

is Toeplitz if and only if

$$a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} = c \otimes \delta - \widetilde{\gamma} \otimes \widetilde{d}.$$

(ii) *If $AB - CD$ is Toeplitz, then $AB = CD$ if and only if*

$$Ab + b_0 a + a_0 b_0 e_0 = Cd + d_0 c + c_0 d_0 e_0 \tag{3.1}$$

and

$$B^*\alpha + \overline{a_0}\beta + \overline{a_0}\overline{b_0}e_0 = D^*\gamma + \overline{c_0}\delta + \overline{c_0}\overline{d_0}e_0. \tag{3.2}$$

Proof. (i) By Lemma 2.3,

$$\begin{aligned}\Delta(AB) - \Delta(CD) = {} & a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} - c \otimes \delta + \widetilde{\gamma} \otimes \widetilde{d} \\ & + [Ab + b_0 a + a_0 b_0 e_0 - Cd - d_0 c - c_0 d_0 e_0] \otimes e_0 \\ & + e_0 \otimes [ZB^*Z^*\alpha + \overline{a_0}\beta - ZD^*Z^*\gamma - \overline{c_0}\delta]. \end{aligned} \tag{3.3}$$

Note that the first four terms on the right side of the above equation involve vectors with 0 in the zeroth component. By Lemma 2.2, $AB - CD$ is Toeplitz if and only if

$$a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} - c \otimes \delta + \widetilde{\gamma} \otimes \widetilde{d} = 0,$$

which is the desired result.

(ii) If $AB - CD$ is Toeplitz, then $AB = CD$ if and only if $\Delta(AB - CD) = 0$. The latter equation holds if and only if the vectors which form tensor products with e_0 in (3.3) are 0. That is, $AB = CD$ if and only if

$$Ab + b_0 a + a_0 b_0 e_0 = Cd + d_0 c + c_0 d_0 e_0,$$

$$ZB^*Z^*\alpha + \overline{a_0}\beta = ZD^*Z^*\gamma + \overline{c_0}\delta.$$

However, (3.1) and (3.2) are equivalent to the above two equations since the difference between (3.2) and the second equation above is

$$\Delta B^*\alpha + \overline{a_0}\overline{b_0}e_0 = \Delta D^*\gamma + \overline{c_0}\overline{d_0}e_0.$$

By Lemma 2.2, this is the same as

$$\langle \alpha, b \rangle e_0 + \overline{a_0}\overline{b_0}e_0 = \langle \gamma, d \rangle e_0 + \overline{c_0}\overline{d_0}e_0,$$

which, up to a complex conjugation, is the zeroth component relation of (3.1). ∎

Brown and Halmos [1] showed that if the product of two Toeplitz operators is zero, then one of the operators is zero. This is not true for finite Toeplitz matrices, as the following simple example illustrates:

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

The following result characterizes Toeplitz matrices whose product is zero.

THEOREM 3.2. *Let a, α, b, and β be vectors in C^n with 0 in the zeroth component. Let $A = T(a, \alpha) + a_0 I$ and $B = T(b, \beta) + b_0 I$. Assume A is not zero.*

(1) *If a and $\widetilde{\alpha}$ are linearly independent, then $AB = 0$ implies that $B = 0$.*

(2) *If $\alpha = 0$, then $AB = 0$ implies that either $B = 0$ or $a_0 = b_0 = 0$ and $\beta = 0$. In the latter case, $AB = 0$ if and only if either*

$$a = \begin{pmatrix} 0 & a_1 & \cdots & a_{n-1} \end{pmatrix}^T \text{ and } b = \begin{pmatrix} 0 & \cdots & 0 & b_{n-1} \end{pmatrix}^T$$

or

$$b = \begin{pmatrix} 0 & b_1 & \cdots & b_{n-1} \end{pmatrix}^T \text{ and } a = \begin{pmatrix} 0 & \cdots & 0 & a_{n-1} \end{pmatrix}^T.$$

Similarly, if $a = 0$, then $AB = 0$, and $B \neq 0$ implies that $a_0 = b_0 = 0$ and $b = 0$. In this case, $AB = 0$ if and only if either

$$\alpha = \begin{pmatrix} 0 & \alpha_1 & \cdots & \alpha_{n-1} \end{pmatrix}^T \text{ and } \beta = \begin{pmatrix} 0 & \cdots & 0 & \beta_{n-1} \end{pmatrix}^T$$

or

$$\beta = \begin{pmatrix} 0 & \beta_1 & \cdots & \beta_{n-1} \end{pmatrix}^T \text{ and } \alpha = \begin{pmatrix} 0 & \cdots & 0 & \alpha_{n-1} \end{pmatrix}^T.$$

(3) *If $a = \lambda\widetilde{\alpha}$ for some $\lambda \in C$, then $AB = 0$ if and only if $b = \lambda\widetilde{\beta}$ and*

$$Ab + b_0 a + a_0 b_0 e_0 = 0. \tag{3.4}$$

Proof. By the previous theorem, $AB = 0$ implies that

$$a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} = 0. \tag{3.5}$$

(1) If a and $\widetilde{\alpha}$ are linearly independent, then $\beta = \widetilde{b} = 0$. Therefore $B = b_0 I$. Thus $AB = 0$ implies that $B = 0$.

(2) If $\alpha = 0$, then either $a = 0$ or $\beta = 0$. In the case $a = 0$, $A = a_0 I$, and $AB = 0$ implies that $B = 0$. Therefore assume $\beta = 0$. That is, both A and B are lower triangular. If either a_0 or b_0 is not zero, then A or B is invertible. Thus $a_0 = b_0 = 0$. Now the first column of AB is

$$T(a,0)b = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_1 & 0 & 0 & \cdots & 0 \\ a_2 & a_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ a_{n-1} & a_{n-2} & \cdots & a_1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ a_1 b_1 \\ \vdots \\ a_{n-2}b_1 + \cdots + a_1 b_{n-2} \end{bmatrix}.$$

If $a_1 \neq 0$, then $T(a,0)b = 0$ implies that $b_i = 0$ for $i = 1, \ldots, n-2$. This corresponds to the first pair of equations; if $b_1 \neq 0$, then we have the second pair of equations. The proof for $a = 0$ is similar.

(3) If $a = \lambda\widetilde{\alpha}$ for some $\lambda \in C$, then (3.5) implies that $b = \lambda\widetilde{\beta}$. So by (3.1) and (3.2) in the previous theorem with $C = D = 0$, $AB = 0$ if and only if

$$Ab + b_0 a + a_0 b_0 e_0 = 0 \quad \text{and} \quad B^*\alpha + \overline{a_0}\beta + \overline{a_0}\overline{b_0}e_0 = 0.$$

However, this system of two equations is equivalent to the first equation. It is easy to see that the zeroth component of the left side of the second equation is the conjugate of the zeroth component of the left side of the first equation. Therefore we write

$$Ab + b_0 a + a_0 b_0 e_0 = c_0 e_0 + c, \quad B^*\alpha + \overline{a_0}\beta + \overline{a_0}\overline{b_0}e_0 = \overline{c_0}e_0 + \gamma,$$

where c and γ have 0 in the zeroth component. It is straightforward (though lengthy) to verify that under the conditions $a = \lambda\widetilde{\alpha}$ and $b = \lambda\widetilde{\beta}$, we have $c = \lambda\widetilde{\gamma}$. This completes the proof. □

By (3) of Theorem 3.2, to check the matrix equation $AB = 0$, we need only check the vector equation (3.4). If $a_0 = b_0 = 0$, then there are no nontrivial 2×2 or 3×3 Toeplitz matrices $A = T(a, \alpha)$ and $B = T(b, \beta)$ such that $AB = 0$. The following are examples of two 4×4 Toeplitz matrices with $a_0 = b_0 = 0$ such that their product is zero. This corresponds to case (3) of Theorem 3.2 with $\lambda = -1$:

$$\begin{bmatrix} 0 & 1 & \sqrt{2} & 1 \\ -1 & 0 & 1 & \sqrt{2} \\ -\sqrt{2} & -1 & 0 & 1 \\ -1 & -\sqrt{2} & -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & -\sqrt{2} & 1 \\ -1 & 0 & 1 & -\sqrt{2} \\ \sqrt{2} & -1 & 0 & 1 \\ -1 & \sqrt{2} & -1 & 0 \end{bmatrix} = 0.$$

We further note that for a given noninvertible Toeplitz matrix $A = T(a, \alpha) + a_0 I$ with $a = \lambda\widetilde{\alpha}$ for some $\lambda \neq 0$, there always exists a nonzero Toeplitz matrix B such that $AB = 0$. In fact we can parametrize all such B in the following way. The derivation of such a parametrization follows essentially from Theorem 3.2 and is omitted.

COROLLARY 3.3. *Let a and α be vectors in C^n with 0 in the zeroth component. Let $A = T(a, \alpha) + a_0 I$. Assume $a = \lambda\widetilde{\alpha}$ for some $\lambda \neq 0$ and A is noninvertible. Let also $b_1, \ldots, b_k$, vectors in C^n with 0 in the zeroth component, be a basis of solutions to the linear system*

$$Ab = 0.$$

Note that if r is the rank of the matrix formed by deleting the zeroth column of matrix A, then $k = n - 1 - r$.

(1) *If the zeroth column $a + a_0 e_0$ of A is not a linear combination of the first through $(n-1)$th columns of A, then $k > 0$ and every Toeplitz matrix B such that $AB = 0$ is given by the following:*

$$B = \lambda_1 T(b_1, \widetilde{b_1}/\overline{\lambda}) + \lambda_2 T(b_2, \widetilde{b_2}/\overline{\lambda}) + \cdots + \lambda_k T(b_k, \widetilde{b_k}/\overline{\lambda}), \qquad \lambda_1, \ldots, \lambda_k \in C.$$

(2) *If the zeroth column $a + a_0 e_0$ of A is a linear combination of the first through $(n-1)$th columns of A, let b, a vector in C^n with 0 in the zeroth component, be a solution to the linear system*

$$Ab + a + a_0 e_0 = 0.$$

Then every Toeplitz matrix B such that $AB = 0$ is given by the following:

$$B = \lambda_0 \left[T(b, \widetilde{b}/\overline{\lambda}) + I \right] + \lambda_1 T(b_1, \widetilde{b_1}/\overline{\lambda}) + \lambda_2 T(b_2, \widetilde{b_2}/\overline{\lambda}) + \cdots + \lambda_k T(b_k, \widetilde{b_k}/\overline{\lambda}),$$

where $\lambda_0, \lambda_1, \ldots, \lambda_k \in C$ *and* k *might be* 0.

Brown and Halmos showed that a Toeplitz operator is normal if and only if it is a linear combination of the identity and a Hermitian. Again there are finite normal Toeplitz matrices which are not linear combinations of the identity and a Hermitian. The characterization of normal Toeplitz matrices has been discussed in [2], [3], [8], [9], [10], [12]. Our procedure leads naturally to this general class of normal Toeplitz matrices, as the following simple proof shows. Our statement is in slightly more compact form than the ones in [2], [12].

THEOREM 3.4. *Let* a *and* α *be vectors in* C^n *with* 0 *in the zeroth component. Let* $A = T(a, \alpha) + a_0 I$. *A is normal* ($A^*A = AA^*$) *if and only if either* $a = \lambda\widetilde{\alpha}$ *for some* $|\lambda| = 1$ *or* $a = \lambda\alpha$ *for some* $|\lambda| = 1$.

Proof. The sufficiency is easy to verify, so we will prove only the necessity. Without loss of generality, assume $A = T(a, \alpha)$; then $A^* = T(\alpha, a)$. By Theorem 3.1, $AA^* = A^*A$ implies that

$$a \otimes a - \widetilde{\alpha} \otimes \widetilde{\alpha} = \alpha \otimes \alpha - \widetilde{a} \otimes \widetilde{a}. \tag{3.6}$$

If $a = \lambda\widetilde{\alpha}$, then $\widetilde{a} = \overline{\lambda}\alpha$ and the above equation becomes

$$\left(|\lambda|^2 - 1\right) \widetilde{\alpha} \otimes \widetilde{\alpha} = \left(1 - |\lambda|^2\right) \alpha \otimes \alpha.$$

Therefore $|\lambda| = 1$.

Now assume that a and $\widetilde{\alpha}$ are linearly independent. We claim $\widetilde{a}$ and a are linearly independent. If $\widetilde{a} = qa$, then (3.6) becomes

$$\left(|q|^2 + 1\right) a \otimes a - \widetilde{\alpha} \otimes \widetilde{\alpha} = \alpha \otimes \alpha.$$

This is impossible since the left side of the above equation is of rank two. Rewriting (3.6) as

$$a \otimes a + \widetilde{a} \otimes \widetilde{a} = \alpha \otimes \alpha + \widetilde{\alpha} \otimes \widetilde{\alpha},$$

we see that there exist $\lambda, r \in C$ with $\lambda \neq 0$ such that

$$\alpha = \lambda a + r\widetilde{a}.$$

Substituting the above equation into (3.6) yields

$$\left[\left(1 - |\lambda|^2 - |r|^2\right) a - 2r\overline{\lambda}\widetilde{a}\right] \otimes a + \left[\left(1 - |\lambda|^2 - |r|^2\right) \widetilde{a} - 2\overline{r}\lambda a\right] \otimes \widetilde{a} = 0.$$

The linear independence of $\widetilde{a}$ and a implies that $r = 0$ and $|\lambda| = 1$. That is, $a = \overline{\lambda}\alpha$ with $|\lambda| = 1$. This completes the proof. □

Remark 3.5. If $a = \lambda\alpha$ for some $|\lambda| = 1$, then $T(a, \alpha) + a_0 I$ is a linear combination of the identity and a Hermitian. In fact, $T(a, \alpha) + a_0 I = \mu T(\mu\alpha, \mu\alpha) + a_0 I$, where μ is any complex number satisfying $\mu^2 = \lambda$. If $a = \lambda\widetilde{\alpha}$ for some $|\lambda| = 1$, then $T(a, \alpha) + a_0 I$ is not, in general, a linear combination of the identity and a Hermitian.

We now give a description of unitary Toeplitz matrices. A Toeplitz operator is unitary only if it is a constant (of modulus one) multiple of the identity.

THEOREM 3.6. *Let* a *and* α *be vectors in* C^n *with* 0 *in the zeroth component. Let* $A = T(a, \alpha) + a_0 I$. *A is unitary* ($A^*A = I$) *if and only if* $a = \lambda\widetilde{\alpha}$ *for some* $|\lambda| = 1$ *and*

$$T(\alpha, a)a + a_0\alpha + \overline{a_0}a + \left(|a_0|^2 - 1\right) e_0 = 0. \tag{3.7}$$

Proof. Note that

$$A^*A - I = [T(\alpha, a) + \overline{a_0}I]\,[T(a, \alpha) + a_0 I] - I.$$

So by Theorem 3.1 with $C = D = I$, A is unitary if and only if $\alpha \otimes \alpha - \widetilde{a} \otimes \widetilde{a} = 0$, that is, $a = \lambda\widetilde{\alpha}$ for some $|\lambda| = 1$, and (3.1) and (3.2) hold. But in this case, (3.1) and (3.2) are the same as (3.7). This completes the proof. □

Remark 3.7. We note that under the condition $a = \lambda\widetilde{\alpha}$, where $|\lambda| = 1$, it follows that

$$\begin{aligned} &T(\alpha, a)a + a_0\alpha + \overline{a_0}a + \left(|a_0|^2 - 1\right) e_0 \\ &= \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_k & \lambda\overline{\beta_k} & \cdots & \lambda\overline{\beta_2} & \lambda\overline{\beta_1} \end{pmatrix}^T \end{aligned}$$

if $n = 2k + 1$, and for $n = 2k$,

$$\begin{aligned} &T(\alpha, a)a + a_0\alpha + \overline{a_0}a + \left(|a_0|^2 - 1\right) e_0 \\ &= \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_k & \lambda\overline{\beta_{k-1}} & \cdots & \lambda\overline{\beta_2} & \lambda\overline{\beta_1} \end{pmatrix}^T. \end{aligned}$$

Thus about half of the component equations of (3.7) are redundant.

Let $\alpha = \begin{pmatrix} 0 & \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix}^T$ and assume no α_i is zero. Then $T(\widetilde{\alpha}, \alpha)$ (corresponding to $\lambda = 1$ in Theorem 3.6) is unitary if and only if

$$\alpha_1 = \omega r e^{-i\frac{\pi}{2}}, \quad \alpha_2 = \pm\omega\sqrt{1 - 2r^2}\, e^{i\frac{\pi}{4}}, \quad \alpha_3 = \omega r,$$

or

$$\alpha_1 = \omega r e^{i\frac{\pi}{2}}, \quad \alpha_2 = \pm\omega\sqrt{1 - 2r^2}\, e^{-i\frac{\pi}{4}}, \quad \alpha_3 = \omega r,$$

for $0 < r < 1/\sqrt{2}$ and $|\omega| = 1$. For example, if $\omega = 1$ and $r = 1/\sqrt{3}$, we have that

$$\frac{1}{\sqrt{3}} \begin{pmatrix} 0 & i & e^{-i\frac{\pi}{4}} & 1 \\ 1 & 0 & i & e^{-i\frac{\pi}{4}} \\ e^{-i\frac{\pi}{4}} & 1 & 0 & i \\ i & e^{-i\frac{\pi}{4}} & 1 & 0 \end{pmatrix} \text{ is unitary.}$$

Huang and Cline [7] and Greville [5] used properties of persymmetric matrices to obtain the following result for which we now give a very simple proof.

THEOREM 3.8. *Let a and α be vectors in C^n with 0 in the zeroth component. Let $A = T(a, \alpha) + a_0 I$. Assume A is invertible. The inverse of A is also Toeplitz if and only if $\widetilde{\alpha} = \lambda a$ for some $\lambda \in C$.*

Proof. Assume A has a Toeplitz inverse $B = T(b, \beta) + b_0 I$, where b and β are vectors in C^n with 0 in the zeroth component. Applying Theorem 3.1 with $C = D = I$, $AB = I$ implies that

$$a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} = 0. \tag{3.8}$$

Therefore $\widetilde{\alpha} = \lambda a$ for some $\lambda \in C$.

To prove the converse, assume A is invertible and $\widetilde{\alpha} = \lambda a$ for some $\lambda \in C$. Let b_0 and $b = \begin{pmatrix} 0 & b_1 & \cdots & b_{n-1} \end{pmatrix}^T$ be the unique solution of

$$A \begin{pmatrix} b_0 & b_1 & \cdots & b_{n-1} \end{pmatrix}^T = e_0, \tag{3.9}$$

which is guaranteed by the invertibility of A. Set $\beta = \overline{\lambda}\widetilde{b}$ and $B = T(b,\beta) + b_0 I$. It is straightforward to check that $AB = I$. That is, A has a Toeplitz inverse. This completes the proof. □

See Shalom [14] for related results on invertible block Toeplitz matrices.

We will use Theorem 3.1 again in the next theorem to determine which pairs of Toeplitz matrices commute. Two Toeplitz operators commute if and only if they are both lower triangular or both upper triangular or one is a linear combination of the identity and the other one. The next theorem was proved by Gel'fgat [4] by an ingenious method which used the circulant and skew-circulant components of Toeplitz matrices. Here we see again that the result follows naturally from our approach.

THEOREM 3.9. *Let a, α, b, and β be vectors in C^n with 0 in the zeroth component. Let $A = T(a,\alpha) + a_0 I$ and $B = T(b,\beta) + b_0 I$ be nonzero Toeplitz matrices.*

(1) *If $a = \lambda\widetilde{\alpha}$ for some $\lambda \in C$, then $AB = BA$ if and only if $b = \lambda\widetilde{\beta}$.*

(2) *If a and $\widetilde{\alpha}$ are linearly independent, then $AB = BA$ if and only if*

$$B = qA + rI \quad \text{for some } q, r \in C. \tag{3.10}$$

Proof. The sufficiency of the conditions in both cases (1) and (2) is easy to verify. By Theorem 3.1 with $C = B$ and $D = A$, $AB - BA = 0$ implies that

$$a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} = b \otimes \alpha - \widetilde{\beta} \otimes \widetilde{a}. \tag{3.11}$$

If $a = \lambda\widetilde{\alpha}$ for some $\lambda \in C$, $\widetilde{a} = \overline{\lambda}\alpha$ and the above equation becomes

$$\widetilde{\alpha} \otimes \left[\overline{\lambda}\beta - \widetilde{b}\right] = \left[b - \lambda\widetilde{\beta}\right] \otimes \alpha.$$

So for some $s \in C$,

$$\overline{\lambda}\beta - \widetilde{b} = s\alpha \quad \text{and} \quad b - \lambda\widetilde{\beta} = \overline{s}\widetilde{\alpha}.$$

Thus

$$\overline{s}\widetilde{\alpha} = \widetilde{s\alpha} = \widetilde{\overline{\lambda}\beta - \widetilde{b}} = \lambda\widetilde{\beta} - b = -\overline{s}\widetilde{\alpha}.$$

Therefore $s = 0$. That is, $b = \lambda\widetilde{\beta}$. This is case (1).

Next we assume that a and $\widetilde{\alpha}$ are linearly independent. If $\beta = 0$, then $B = b_0 I$. If $\beta \neq 0$, then we claim that b and $\widetilde{\beta}$ are also linearly independent. If $b = s\widetilde{\beta}$ for some $s \in C$, then $\widetilde{b} = \overline{s}\beta$. Equation (3.11) becomes

$$[a - s\widetilde{\alpha}] \otimes \beta = \widetilde{\beta} \otimes [\overline{s}\alpha - \widetilde{a}].$$

Thus $a - s\widetilde{\alpha} = \lambda\widetilde{\beta}$ and $\overline{s}\alpha - \widetilde{a} = \overline{\lambda}\beta$ for some $\lambda \in C$. This implies that

$$a - s\widetilde{\alpha} = \widetilde{\overline{s}\alpha - \widetilde{a}} = s\widetilde{\alpha} - a.$$

Therefore $a - s\widetilde{\alpha} = 0$, a contradiction.

Now both sides of (3.11) are of rank two due to the linear independence conditions. Since the range of each side of (3.11) must be equal, we know b and $\widetilde{\beta}$ are both in the span of a and $\widetilde{\alpha}$; that is, there exist $q_{11}, q_{12}, q_{21}, q_{22} \in C$ such that

$$\begin{aligned} b &= q_{11}a + q_{12}\widetilde{\alpha}, \\ \widetilde{\beta} &= q_{21}a + q_{22}\widetilde{\alpha}. \end{aligned} \tag{3.12}$$

Substituting (3.12) back into (3.11) yields

$$[2q_{21}a + (q_{22} - q_{11})\,\widetilde{\alpha}] \otimes \widetilde{a} + [(q_{22} - q_{11})\,a - 2q_{12}\widetilde{\alpha}] \otimes \alpha = 0.$$

Linear independence of a and $\widetilde{\alpha}$ shows that $q_{21} = q_{12} = 0$ and $q_{22} = q_{11}$. Therefore

$$b = qa \quad \text{and} \quad \widetilde{\beta} = q\widetilde{\alpha};$$

consequently $T(b, \beta) = qT(a, \alpha)$, so $B = qT(a, \alpha) + b_0 I = qA + (b_0 - a_0 q)\, I$, which has the form (3.10). □

The following result is due to Ikramov and Chugunov [11] in a different formulation.

THEOREM 3.10. *Let a, α, b, and β be vectors in C^n with 0 in the zeroth component. Let $A = T(a, \alpha) + a_0 I$ and $B = T(b, \beta) + b_0 I$. The skew-symmetric part of AB is Toeplitz if and only if one of the following holds:*

(1) *If $a = \lambda\widetilde{\alpha}$ for some $\lambda \in C$, then $(\overline{\lambda}\beta - \widetilde{b}) = s\overline{\widetilde{\alpha}}$ for some $s \in C$.*

(2) *If a and $\widetilde{\alpha}$ are linearly independent, then there exist $q_{11}, q_{22}, q_{12} \in C$ such that*

$$\overline{\beta} = q_{11}a + q_{12}\widetilde{\alpha},$$
$$\overline{\widetilde{b}} = -q_{12}a + q_{22}\widetilde{\alpha}.$$

Proof. The skew-symmetric part of AB is

$$AB - (AB)^T = (T(a, \alpha) + a_0 I)\,(T(b, \beta) + b_0 I) - \left(T(\overline{\beta}, \overline{b}) + b_0 I\right)(T(\overline{\alpha}, \overline{a}) + a_0 I)\,.$$

By Theorem 3.1, the right side above is Toeplitz if and only if

$$a \otimes \beta - \widetilde{\alpha} \otimes \widetilde{b} = \overline{\beta} \otimes \overline{a} - \overline{\widetilde{b}} \otimes \overline{\widetilde{\alpha}}. \tag{3.13}$$

If $a = \lambda\widetilde{\alpha}$ for some $\lambda \in C$, then

$$\widetilde{\alpha} \otimes \left(\overline{\lambda}\beta - \widetilde{b}\right) = \left(\lambda\overline{\beta} - \overline{\widetilde{b}}\right) \otimes \overline{\widetilde{\alpha}}.$$

Thus $(\overline{\lambda}\beta - \widetilde{b}) = s\overline{\widetilde{\alpha}}$ for $s \in C$.

If a and $\widetilde{\alpha}$ are linearly independent, then by (3.13) there exist $q_{11}, q_{22}, q_{12}, q_{21} \in C$ such that

$$\overline{\beta} = q_{11}a + q_{12}\widetilde{\alpha},$$
$$\overline{\widetilde{b}} = q_{21}a + q_{22}\widetilde{\alpha}.$$

Substituting the above equations into (3.13) yields

$$-(q_{21} + q_{12})\,\widetilde{\alpha} \otimes \overline{a} + (q_{21} + q_{12})\,a \otimes \overline{\widetilde{\alpha}} = 0.$$

The linear independence of a and $\widetilde{\alpha}$ implies that $q_{21} = -q_{12}$. □

A characterization of when the skew-Hermitian part of the product of two Toeplitz matrices is Toeplitz can be obtained similarly.

4. Hankel case. In this section we discuss the same questions for Hankel matrices. The following two simple identities connecting Toeplitz and Hankel matrices allow us to reduce questions about Hankel matrices to similar questions about Toeplitz matrices:

$$P\,[H(a,\alpha)+a_nP] = T(\widetilde{\overline{a}},\alpha)+a_nI, \tag{4.1}$$

$$[H(a,\alpha)+a_nP]\,P = T(\overline{\alpha},\widetilde{a})+a_nI. \tag{4.2}$$

Indeed, it follows from above that

$$\begin{aligned}&[H(a,\alpha)+a_nP]\,[H(b,\beta)+b_nP]\\ &= [H(a,\alpha)+a_nP]\,PP\,[H(b,\beta)+b_nP]\\ &= [T(\overline{\alpha},\widetilde{a})+a_nI]\left[T(\widetilde{\overline{b}},\beta)+b_nI\right].\end{aligned}$$

Therefore the question of when a product of two Hankel matrices is zero is equivalent to the question of when a product of two Toeplitz matrices is zero. This question is solved by Theorem 3.2. The above identity also shows that the Hankel matrix $H(a,\alpha)+a_nP$ has a Hankel inverse if and only if the Toeplitz matrix $T(\overline{\alpha},\widetilde{a})+a_nI$ has a Toeplitz inverse. Thus the question of when a Hankel matrix has a Hankel inverse is answered by Theorem 3.8.

Similarly, by the fact that

$$\begin{aligned}&[H(a,\alpha)+a_nP]\,[H(a,\alpha)+a_nP]^*-I\\ &= [H(a,\alpha)+a_nP]\,[H(\overline{a},\overline{\alpha})+\overline{a_n}P]-I=0\end{aligned}$$

if and only if

$$\begin{aligned}&[T(\overline{\alpha},\widetilde{a})+a_nI]\,[T(\widetilde{a},\overline{\alpha})+\overline{a_n}I]-I\\ &= [T(\overline{\alpha},\widetilde{a})+a_nI]\,[T(\overline{\alpha},\widetilde{a})+a_nI]^*-I=0,\end{aligned}$$

we see that the Hankel matrix $H(a,\alpha)+a_nP$ is unitary if and only if the Toeplitz matrix $T(\overline{\alpha},\widetilde{a})+a_nI$ is unitary. Therefore a characterization of unitary Hankel matrices can be obtained by using Theorem 3.6.

We note that in these three questions only one product of two Hankel matrices is involved. The characterizations of normal and commuting Hankel matrices are more complicated, as we shall see. This is because two products of Hankel matrices are needed in the study of normal and commuting Hankel matrices. It is clear that a Hankel matrix is Hermitian only if it is real symmetric.

THEOREM 4.1. *Let a and α be vectors in C^n with 0 in the zeroth component. Let $A=H(a,\alpha)+a_nP$. A is normal if and only if one of the following holds:*

(1) $a=\lambda_1\overline{a}$ and $\alpha=\lambda_2\overline{\alpha}$ for some $\lambda_1,\lambda_2\in C$ with $|\lambda_1|=|\lambda_2|=1$ and $T(\overline{\alpha},\widetilde{a})\widetilde{a}+a_n\widetilde{a}+\overline{a_n\alpha}$ is a real vector.

(2) $\alpha=\lambda\overline{a}+ra$ for some $\lambda,r\in C$ satisfying $|\lambda|^2-|r|^2=1$ and $T(\overline{\alpha},\widetilde{a})\widetilde{a}+a_n\widetilde{a}+\overline{a_n\alpha}$ is a real vector.

Proof. By (4.1) and (4.2),

$$\begin{aligned}&AA^*-A^*A\\ &= [H(a,\alpha)+a_nP]\,[H(\overline{a},\overline{\alpha})+\overline{a_n}P]-[H(\overline{a},\overline{\alpha})+\overline{a_n}P]\,[H(a,\alpha)+a_nP]\\ &= [H(a,\alpha)+a_nP]\,PP\,[H(\overline{a},\overline{\alpha})+\overline{a_n}P]-[H(\overline{a},\overline{\alpha})+\overline{a_n}P]\,PP\,[H(a,\alpha)+a_nP]\\ &= [T(\overline{\alpha},\widetilde{a})+a_nI]\,[T(\widetilde{a},\overline{\alpha})+\overline{a_n}I]-\left[T(\alpha,\widetilde{\overline{a}})+\overline{a_n}I\right]\left[T(\widetilde{\overline{a}},\alpha)+a_nI\right].\end{aligned}$$

By Theorem 3.1, A is normal if and only if

$$\overline{\alpha} \otimes \overline{\alpha} - a \otimes a = \alpha \otimes \alpha - \overline{a} \otimes \overline{a} \tag{4.3}$$

and (3.1) and (3.2) hold. Note that in this case, (3.1) and (3.2) are the same as

$$T(\overline{\alpha}, \widetilde{a})\widetilde{a} + a_n\widetilde{a} + \overline{a_n\alpha} = T(\alpha, \overline{\widetilde{a}})\overline{\widetilde{a}} + \overline{a_n}\overline{\widetilde{a}} + a_n\alpha,$$

that is, $T(\overline{\alpha}, \widetilde{a})\widetilde{a} + a_n\widetilde{a} + \overline{a_n\alpha}$ is a real vector.

If $a = \lambda_1\overline{a}$, then $|\lambda_1| = 1$ and (4.3) becomes

$$\overline{\alpha} \otimes \overline{\alpha} = \alpha \otimes \alpha.$$

Thus $\alpha = \lambda_2\overline{\alpha}$ for some $\lambda_2 \in C$ with $|\lambda_2| = 1$. This is case (1).

Now assume a and $\overline{a}$ are linearly independent. Equation (4.3) implies that

$$\alpha = \lambda\overline{a} + ra.$$

Substituting the above equation for α into (4.3) yields

$$\left(1 - |\lambda|^2 + |r|^2\right)\overline{a} \otimes \overline{a} - \left(1 - |\lambda|^2 + |r|^2\right) a \otimes a = 0.$$

The linear independence of a and $\overline{a}$ shows that $|\lambda|^2 - |r|^2 = 1$. This is case (2). ∎

Next we give examples of normal Hankel matrices which correspond to case (2) in the above theorem with $r = 1$ and $\lambda = 1 + i$. Let $a = \begin{pmatrix} 0 & a_1 & a_2 & a_3 \end{pmatrix}^T$ and assume no a_i is zero. Set $\alpha = (1+i)\overline{a} + a$. $H(a, \alpha)$ is normal if and only if

$$H(a, \alpha) = \omega \begin{pmatrix} y(-1+i) & 1 & y(-1+i) & 0 \\ 1 & y(-1+i) & 0 & y(-1+i) \\ y(-1+i) & 0 & y(-1+i) & 2-i \\ 0 & y(-1+i) & 2-i & y(-1+i) \end{pmatrix}$$

for some nonzero real number y and complex number ω. We will return later to case (2) with $r = 0$. First we give a complete characterization of normal Hankel matrices for case (1) above with $a_n = 0$.

COROLLARY 4.2. *Assume $H(a, \alpha)$ is not a constant multiple of a real symmetric matrix. If $a = \lambda_1\overline{a}$ and $\alpha = \lambda_2\overline{\alpha}$ for some $\lambda_1, \lambda_2 \in C$ with $|\lambda_1| = |\lambda_2| = 1$, then $H(a, \alpha)$ is normal if and only if, for some real numbers $r_i, i = 1, \ldots, n-1$, and a complex number p_1, either*

$$a = p_1 \begin{pmatrix} 0 & r_1 & \cdots & r_{n-1} \end{pmatrix}^T \text{ and } \alpha = \begin{pmatrix} 0 & \cdots & 0 & \alpha_{n-1} \end{pmatrix}^T$$

or

$$\alpha = p_1 \begin{pmatrix} 0 & r_1 & \cdots & r_{n-1} \end{pmatrix}^T \text{ and } a = \begin{pmatrix} 0 & \cdots & 0 & a_{n-1} \end{pmatrix}^T.$$

Proof. Assuming $a = \lambda_1\overline{a}$ and $\alpha = \lambda_2\overline{\alpha}$ for some $\lambda_1, \lambda_2 \in C$ with $|\lambda_1| = |\lambda_2| = 1$, we can write

$$a = e^{i\theta}r = e^{i\theta}\begin{pmatrix} 0 & r_1 & \cdots & r_{n-1} \end{pmatrix}^T, \quad \alpha = e^{i\varphi}s = e^{i\varphi}\begin{pmatrix} 0 & s_1 & \cdots & s_{n-1} \end{pmatrix}^T$$

for some $\theta, \varphi \in [0, 2\pi)$ and real vectors r and s. Note that

$$T(\overline{\alpha}, \widetilde{a})\widetilde{a} = T(e^{-i\varphi}s, e^{-i\theta}\widetilde{r})e^{-i\theta}\widetilde{r} = e^{-i(\theta+\varphi)}T(s, 0)\widetilde{r} + T(0, \widetilde{r})\widetilde{r}$$

and $T(s,0)\widetilde{r}$, $T(0,\widetilde{r})\widetilde{r}$ are real vectors. If $T(\overline{\alpha},\widetilde{a})\widetilde{a}$ is real, then either $e^{-i(\theta+\varphi)}=\pm 1$ or $T(s,0)\widetilde{r}=0$. In the case $e^{-i(\theta+\varphi)}=\pm 1$, $H(a,\alpha)$ is a constant multiple of a real symmetric matrix. Assume now $T(s,0)\widetilde{r}=0$. However,

$$T(s,0)\widetilde{r}=\begin{bmatrix} 0 & 0 & 0 & \cdots & 0\\ s_1 & 0 & 0 & \cdots & 0\\ s_2 & s_1 & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ s_{n-1} & s_{n-2} & \cdots & s_1 & 0 \end{bmatrix}\begin{bmatrix} 0\\ r_1\\ r_2\\ \vdots\\ r_{n-1}\end{bmatrix}=\begin{bmatrix} 0\\ 0\\ s_1r_1\\ \vdots\\ s_{n-2}r_1+\cdots+s_1r_{n-2}\end{bmatrix}.$$

If $r_1\neq 0$, then $T(s,0)\widetilde{r}=0$ implies that $s_i=0$ for $i=1,\dots,n-2$. This corresponds to the first case. Similarly if $s_1\neq 0$, we have the other case. ∎

Next we give a more detailed analysis when $r=0$ in case (2) of Theorem 4.1.

COROLLARY 4.3. *If $\alpha=\lambda\overline{a}$ for some $|\lambda|=1$, then $H(a,\alpha)$ is normal if and only if one of the following holds:*

(1) *$\lambda=\pm 1$ and $T(\overline{\alpha},\widetilde{a})\widetilde{a}$ is real.*

(2) *If $\lambda\neq\pm 1$, then $T(\overline{\alpha},\widetilde{a})\widetilde{a}=|a|^2 e_0$. In this case, $H(a,\alpha)$ is a scalar multiple of a unitary.*

Proof. If $\alpha=\lambda\overline{a}$ for some $|\lambda|=1$, then by Theorem 4.1 $H(a,\alpha)$ is normal if and only if $T(\overline{\alpha},\widetilde{a})\widetilde{a}$ is a real vector. Write $a=\begin{pmatrix} 0 & a_1 & \cdots & a_{n-1}\end{pmatrix}^T$. Note that

$$\begin{aligned} T(\overline{\alpha},\widetilde{a})\widetilde{a}&=T(\overline{\lambda}a,\widetilde{a})\widetilde{a}=\overline{\lambda}T(a,0)\widetilde{a}+T(0,\widetilde{a})\widetilde{a}\\ &=\overline{\lambda}\begin{pmatrix} 0 & 0 & \delta_1 & \cdots & \delta_{n-3} & \delta_{n-2}\end{pmatrix}^T+\begin{pmatrix} |a|^2 & \overline{\delta_{n-2}} & \overline{\delta_{n-3}} & \cdots & \overline{\delta_1} & 0\end{pmatrix}^T,\end{aligned}$$

where $|a|^2$ is the squared norm of vector a and

$$\delta_i=a_i\overline{a_{n-1}}+a_{i-1}\overline{a_{n-2}}+\cdots+a_1\overline{a_{n-i}},\quad i=1,\dots,n-2.$$

Therefore $T(\overline{\alpha},\widetilde{a})\widetilde{a}$ is real if and only if

$$\overline{\lambda}\delta_{n-2},\overline{\delta_{n-2}},\overline{\lambda}\delta_i+\overline{\delta_{n-2-i}}\text{ for }i=1,\dots,n-3\text{ are real.}$$

Set

$$s_i=\overline{\lambda}\delta_i+\overline{\delta_{n-2-i}},\qquad i=1,\dots,n-3,$$

and note that

$$s_i=\overline{\lambda}\overline{s_{n-2-i}},\qquad i=1,\dots,n-3.$$

Therefore if $\lambda\neq\pm 1$, then $T(\overline{\alpha},\widetilde{a})\widetilde{a}$ being real implies $\delta_{n-2}=0$ and $s_i=0$ for $i=1,\dots,n-3$. In this case, $T(\overline{\alpha},\widetilde{a})\widetilde{a}=|a|^2 e_0$. It follows from Theorem 3.6 that $T(\widetilde{a},\overline{\alpha})$ is a scalar multiple of a unitary. Therefore $H(a,\alpha)$ is also a scalar multiple of a unitary since

$$\begin{aligned} H(a,\alpha)H(a,\alpha)^*&=H(a,\alpha)PPH(\overline{a},\overline{\alpha})\\ &=T(\overline{\alpha},\widetilde{a})T(\widetilde{a},\overline{\alpha})=T(\widetilde{a},\overline{\alpha})^*T(\widetilde{a},\overline{\alpha})=|a|^2 I.\end{aligned}$$ ∎

For $\lambda=-1$ as in part (1) of the above corollary we have, for example, that

$$\begin{pmatrix} a_1 & a_2 & a_3 & 0\\ a_2 & a_3 & 0 & -a_1\\ a_3 & 0 & -a_1 & -a_2\\ 0 & -a_1 & -a_2 & -a_3\end{pmatrix}\text{ is normal if and only if }a_1=a_3.$$

For $\lambda = 1$ as in the above corollary, we have, for example, that

$$\begin{pmatrix} a_1 & a_2 & a_3 & 0 \\ a_2 & a_3 & 0 & a_1 \\ a_3 & 0 & a_1 & a_2 \\ 0 & a_1 & a_2 & a_3 \end{pmatrix} \text{ is normal if and only if } a_2\overline{a_3} + a_1\overline{a_2} \text{ is real.}$$

We now describe when two Hankel matrices commute.

THEOREM 4.4. *Let a, α, b, and β be vectors in C^n with 0 in the zeroth component. Let $A = H(a, \alpha) + a_n P$ and $B = H(b, \beta) + b_n P$.*

(1) *If $\overline{\alpha} = \lambda a$ for some $\lambda \in C$, then $AB = BA$ if and only if $\left(\lambda\overline{\beta} - b\right) = \delta a$ for some $\delta \in C$ and*

$$T(\overline{\alpha}, \widetilde{a})\overline{\overline{b}} + a_n\overline{\overline{b}} + b_n\overline{\alpha} = T(\overline{\beta}, \widetilde{b})\overline{\overline{a}} + b_n\overline{\overline{a}} + a_n\overline{\beta}. \tag{4.4}$$

(2) *Assume a and $\overline{\alpha}$ are linearly independent. Let*

$$\begin{aligned} v_1 &= \overline{\overline{a}} - \overline{\alpha}, \quad v_2 = T(\overline{\alpha}, 0)\overline{\overline{a}} + a_n\overline{\alpha}, \\ v_3 &= T(\overline{\alpha}, \widetilde{a})\widetilde{\alpha} - T(-a, \overline{\widetilde{\alpha}})\overline{\overline{a}} + a_n\widetilde{\alpha} + a_n a. \end{aligned} \tag{4.5}$$

If $\dim\{v_1, v_2, v_3\} = 3$, then B commutes with A if and only if

$$B = \lambda A, \quad \lambda \in C.$$

(3) *Assume $\dim\{v_1, v_2, v_3\} = 2$. Assume also that no two vectors of v_1, v_2, v_3 are linearly dependent. (The case where two of v_1, v_2, v_3 are linearly dependent can be treated similarly.) Let $s, q, r \in C$ be such that*

$$sv_1 + qv_2 + rv_3 = 0. \tag{4.6}$$

The Hankel matrices B such that $BA = AB$ are parametrized by the following:

$$B = \lambda_1 A + \lambda_2 \left[H(-r\overline{\alpha}, \overline{ra} + \overline{q}\alpha) + sP\right], \quad \lambda_1, \lambda_2 \in C.$$

(4) *Assume $\dim\{v_1, v_2, v_3\} = 1$. Assume also $v_1 \neq 0$. (Cases $v_2 \neq 0$ and $v_3 \neq 0$ can be treated similarly.) Let $s, r \in C$ be such that*

$$v_2 = sv_1, \quad v_3 = rv_1. \tag{4.7}$$

The Hankel matrices B such that $BA = AB$ are parametrized by the following:

$$B = \lambda_1 A + \lambda_2 \left[H(\overline{\alpha}, -\overline{a}) + rP\right] + \lambda_3 \left[H(0, -\alpha) + sP\right], \quad \lambda_1, \lambda_2, \lambda_3 \in C.$$

Proof. By (4.1) and (4.2),

(4.8)

$$\begin{aligned} &AB - BA \\ &= \left[H(a, \alpha) + a_n P\right] PP \left[H(b, \beta) + b_n P\right] - \left[H(b, \beta) + b_n P\right] PP \left[H(a, \alpha) + a_n P\right] \\ &= \left(T(\overline{\alpha}, \widetilde{a}) + a_n I\right) \left(T(\overline{\widetilde{b}}, \beta) + b_n I\right) - \left(T(\overline{\beta}, \widetilde{b}) + b_n I\right) \left(T(\overline{\widetilde{a}}, \alpha) + a_n I\right). \end{aligned}$$

By Theorem 3.1, $AB = BA$ if and only if

$$\overline{\alpha} \otimes \beta - a \otimes \overline{b} = \overline{\beta} \otimes \alpha - b \otimes \overline{a}, \tag{4.9}$$

$$T(\overline{\alpha}, \widetilde{a})\overline{\widetilde{b}} + a_n\overline{\widetilde{b}} + b_n\overline{\alpha} - T(\overline{\beta}, \widetilde{b})\overline{\widetilde{a}} - b_n\overline{\widetilde{a}} - a_n\overline{\beta} = 0, \tag{4.10}$$

and

$$T(\beta, \overline{\widetilde{b}})\widetilde{a} + \overline{b_n}\widetilde{a} + \overline{a_n}\beta - T(\alpha, \overline{\widetilde{a}})\widetilde{b} - \overline{a_n}\widetilde{b} - \overline{b_n}\alpha = 0. \tag{4.11}$$

An inspection reveals that the left side of (4.11) is minus the conjugate of the left side of (4.10). Therefore only (4.10) is needed. We now divide the proof into several cases. In each case we assume $AB = BA$ and derive the necessary conditions for B.

(1) If $\overline{\alpha} = \lambda a$ for some $\lambda \in C$, then (4.9) becomes

$$a \otimes \left(\overline{\lambda}\beta - \overline{b}\right) = \left(\lambda\overline{\beta} - b\right) \otimes \overline{a}.$$

Equivalently, $\lambda\overline{\beta} - b = \delta a$ for some $\delta \in C$.

Now assume a and $\overline{\alpha}$ are linearly independent. Equating ranges of each side of (4.9) shows that there exist $q_{11}, q_{12}, q_{21}, q_{22} \in C$ such that

$$\begin{aligned} b &= q_{11}a + q_{12}\overline{\alpha}, \\ \overline{\beta} &= q_{21}a + q_{22}\overline{\alpha}. \end{aligned} \tag{4.12}$$

When (4.12) is substituted into (4.9), we obtain

$$(q_{12} + q_{21})\,\overline{\alpha} \otimes \overline{a} - (q_{21} + q_{12})\,a \otimes \alpha = 0.$$

The linear independence of the pair $\overline{a}$ and α shows that the above equation holds if and only if $q_{21} = -q_{12}$. Substituting (4.12) into (4.10) gives

$$\begin{aligned} 0 &= T(\overline{\alpha}, \widetilde{a})\overline{\widetilde{b}} + a_n\overline{\widetilde{b}} + b_n\overline{\alpha} - T(\overline{\beta}, \widetilde{b})\overline{\widetilde{a}} - b_n\overline{\widetilde{a}} - a_n\overline{\beta} \\ &= \left[T(\overline{\alpha}, \widetilde{a}) + a_nI\right](q_{11}\overline{\widetilde{a}} + q_{12}\widetilde{\alpha}) + b_n\overline{\alpha} \\ &\quad - \left[T(-q_{12}a + q_{22}\overline{\alpha}, \overline{q_{11}}\widetilde{a} + \overline{q_{12}}\widetilde{\overline{\alpha}}) + b_nI\right]\overline{\widetilde{a}} - a_n(-q_{12}a + q_{22}\overline{\alpha}) \\ &= q_{12}v_3 + (q_{11} - q_{22})\,v_2 + (a_nq_{11} - b_n)v_1, \end{aligned} \tag{4.13}$$

where v_1, v_2, and v_3 are defined as in (4.5).

(2) If $\dim\{v_1, v_2, v_3\} = 3$, then $q_{12} = 0$, $q_{11} - q_{22} = 0$, and $(a_nq_{11} - b_n) = 0$. That is, $B = H(b, \beta) + b_nI = q_{11}A$.

(3) If $\dim\{v_1, v_2, v_3\} = 2$ and $s, q, r \in C$ are as in (4.6), then by (4.13), there exists some $\lambda \in C$ such that

$$q_{12} = \lambda r, \quad q_{11} - q_{22} = \lambda q, \quad a_nq_{11} - b_n = \lambda s.$$

Therefore by (4.12), we have

$$\begin{aligned} B &= H(b, \beta) + b_nP \\ &= H\left[q_{11}a + \lambda r\overline{\alpha}, -\overline{\lambda r}\overline{a} + (\overline{q_{11}} - \overline{\lambda q})\alpha\right] + (a_nq_{11} - \lambda s)\,P \\ &= q_{11}\left[H(a, \alpha) + a_nP\right] - \lambda\left[H(-r\overline{\alpha}, \overline{r}\overline{a} + \overline{q}\alpha) + sP\right] \\ &= \lambda_1A + \lambda_2\left[H(-r\overline{\alpha}, \overline{r}\overline{a} + \overline{q}\alpha) + sP\right]. \end{aligned}$$

(4) If $\dim\{v_1, v_2, v_3\} = 1$, then substituting (4.7) into (4.13) gives

$$b_n - a_n q_{11} = s\,(q_{11} - q_{22}) + r q_{12}.$$

Therefore by (4.12), we have

$$\begin{aligned} B &= H(b,\beta) + b_n P \\ &= H\left[q_{11}a + q_{12}\overline{\alpha}, -\overline{q_{12}a} + \left(\overline{q_{22}} - \overline{q_{11}} + \overline{q_{11}}\right)\alpha\right] + \left[a_n q_{11} + s\,(q_{11} - q_{22}) + r q_{12}\right] P \\ &= q_{11}\left[H(a,\alpha) + a_n P\right] + q_{12}\left[H(\overline{\alpha}, -\overline{a}) + rP\right] + (q_{11} - q_{22})\left[H(0,-\alpha) + sP\right] \\ &= \lambda_1 A + \lambda_2\left[H(\overline{\alpha}, -\overline{a}) + rP\right] + \lambda_3\left[H(0,-\alpha) + sP\right]. \end{aligned}$$

This completes the proof. □

Remark 4.5. It is straightforward (though lengthy) to show that if $v_1 = v_2 = v_3 = 0$, then either $A = a_n P$ or $A = 0$ if n is odd and $A = H(a, \widetilde{a})$, where

$$a = \begin{pmatrix} 0 & 0 & \cdots & 0 & a_m & 0 & \cdots & 0 \end{pmatrix}^T$$

if $n = 2m$ is even.

For a given Hankel matrix $A = H(a,\alpha) + a_n P$, in general, case (2) above occurs; thus the Hankel matrix B that commutes with A is a scalar multiple of A. We now give an example of a 4×4 Hankel matrix $A = H(a,\alpha) + a_4 P$ which corresponds to case (3) or (4), where the Hankel matrices B that commute with A are readily described by our formulas.

If $a_4 = 0, a_1, a_2, a_3 \in C$,

$$a = \begin{pmatrix} 0 & a_1 & a_2 & a_3 \end{pmatrix}^T, \text{ and } \alpha = \begin{pmatrix} 0 & -\overline{a_3} & \overline{a_2} & -\overline{a_1} \end{pmatrix}^T,$$

then

$$\begin{aligned} v_1 &= \begin{pmatrix} 0 & 2a_3 & 0 & 2a_1 \end{pmatrix}^T, \quad v_2 = \begin{pmatrix} 0 & 0 & -a_3^2 & 0 \end{pmatrix}^T, \\ v_3 &= \begin{pmatrix} 0 & 0 & a_1^2 + 2a_1 a_3 - a_3^2 & 0 \end{pmatrix}^T. \end{aligned}$$

Thus $\dim\{v_1, v_2, v_3\} = 2$ unless both $a_1 = 0$ and $a_3 = 0$.

If $a_4 = 0, a_1, a_2, a_3 \in C$,

$$a = \begin{pmatrix} 0 & a_1 & a_2 & a_3 \end{pmatrix}^T, \text{ and } \alpha = \begin{pmatrix} 0 & \overline{a_3} & -\overline{a_2} & \overline{a_1} \end{pmatrix}^T,$$

then

$$\begin{aligned} v_1 &= \begin{pmatrix} 0 & 0 & 2a_2 & 0 \end{pmatrix}^T, \quad v_2 = \begin{pmatrix} 0 & 0 & a_3^2 & 0 \end{pmatrix}^T, \\ v_3 &= \begin{pmatrix} 0 & 0 & -a_1^2 + 2a_1 a_3 + a_3^2 & 0 \end{pmatrix}^T. \end{aligned}$$

Thus $\dim\{v_1, v_2, v_3\} = 1$ unless $a = 0$.

The following corollary gives a more detailed analysis of case (1) in the above theorem.

COROLLARY 4.6. *Let a, α, b, and β be vectors in C^n with 0 in the zeroth component. Let $A = H(a,\alpha) + a_n P$ and $B = H(b,\beta) + b_n P$.*

(1) *Assume $\overline{\alpha} = 0$ and $a \neq 0$. Let γ_0, a vector in C^n with 0 in the zeroth component, be a solution to the linear system*

$$\left(T(\overline{\overline{a}}, 0) + a_n I\right)\overline{\gamma_0} = \overline{\overline{a}}, \tag{4.14}$$

and let $\gamma_1, \ldots, \gamma_k$, *vectors in* C^n *with* 0 *in the zeroth component, be a basis of solutions to the linear system*

$$\left(T(\overline{\widetilde{a}}, 0) + a_n I\right) \overline{\gamma} = 0.$$

Every Hankel matrix B *such that* $BA = AB$ *is given by the following:*

$$B = \delta A + \lambda_0 \left[H(0, \gamma_0) - P\right] + \lambda_1 H(0, \gamma_1) + \cdots + \lambda_k H(0, \gamma_k),$$

where $\delta, \lambda_0, \lambda_1, \ldots, \lambda_k \in C$.

(2) *Assume* $\overline{\alpha} = \lambda a$ *for some* $\lambda \neq 0$, *where* $a = \begin{pmatrix} 0 & a_1 & \cdots & a_{n-1} \end{pmatrix}^T$. *Define* $h_a = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}^T$. *Set*

$$c = \begin{pmatrix} 0 & a_n & \cdots & a_2 \end{pmatrix}^T, \quad v = H(0, \overline{a}/\overline{\lambda}) h_a, \tag{4.15}$$

$$C = H(a, \alpha) + a_n P - T(c/\lambda, \widetilde{c}) - a_1 I. \tag{4.16}$$

Let $\eta_0 = \begin{pmatrix} \eta_{01} & \eta_{02} & \cdots & \eta_{0n} \end{pmatrix}^T$ *be a solution to the linear system*

$$C\eta = v, \tag{4.17}$$

and let $\eta_i = \begin{pmatrix} \eta_{i1} & \eta_{i2} & \cdots & \eta_{in} \end{pmatrix}^T$, $i = 1, \ldots, k$, *be a basis of solutions to the linear system*

$$C\eta = 0.$$

Set $\gamma_i = \begin{pmatrix} 0 & \eta_{i1} & \cdots & \eta_{in-1} \end{pmatrix}^T$ *for* $i = 0, 1, \ldots, k$. *Every Hankel matrix* B *such that* $BA = AB$ *is given by the following:*

$$\begin{aligned} B = {} & \lambda_0 \left[H(\gamma_0, (\overline{\gamma_0} + \overline{a})/\overline{\lambda}) + \eta_{0n} P\right] + \lambda_1 \left[H(\gamma_1, \overline{\gamma_1}/\overline{\lambda}) + \eta_{1n} P\right] \\ & + \cdots + \lambda_k \left[H(\gamma_k, \overline{\gamma_k}/\overline{\lambda}) + \eta_{kn} P\right], \end{aligned}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_k \in C$.

Proof. We will prove only case (2) since the proof of case (1) is similar. Write $b = \begin{pmatrix} 0 & b_1 & \cdots & b_{n-1} \end{pmatrix}$. Set

$$h_b = \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix}^T, \quad h_a = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}^T.$$

Recall that if $\overline{\alpha} = \lambda a$ for some $\lambda \neq 0$, then $AB = BA$ if and only if $\left(\lambda\overline{\beta} - b\right) = \delta a$ for some $\delta \in C$ and

$$T(\overline{\alpha}, \widetilde{a})\overline{\widetilde{b}} + a_n \overline{\widetilde{b}} + b_n \overline{\alpha} = T(\overline{\beta}, \widetilde{b})\overline{\widetilde{a}} + b_n \overline{\widetilde{a}} + a_n \overline{\beta}. \tag{4.18}$$

By (4.8), the left side in the above equation plus $a_n b_n e_0$ is the first column of AB, that is,

$$T(\overline{\alpha}, \widetilde{a})\overline{\widetilde{b}} + a_n \overline{\widetilde{b}} + b_n \overline{\alpha} + a_n b_n e_0 = \left[H(a, \alpha) + a_n P\right] h_b,$$

and similarly, by using $\left(\lambda\overline{\beta} - b\right) = \delta a$, we have, for c and v defined by (4.15),

$$\begin{aligned} & T(\overline{\beta}, \widetilde{b})\overline{\widetilde{a}} + b_n \overline{\widetilde{a}} + a_n \overline{\beta} + a_n b_n e_0 \\ & = \left[H(b, \beta) + b_n P\right] h_a = \left[H(b, \overline{b}/\overline{\lambda}) + b_n P\right] h_a + \delta H(0, \overline{a}/\overline{\lambda}) h_a \\ & = \left[H(b, 0) + b_n P\right] h_a + H(0, \overline{b}/\overline{\lambda}) h_a + \delta v \\ & = \left[T(0, \widetilde{c}) + a_1 I\right] h_b + T(c/\lambda, 0) h_b + \delta v \\ & = \left[T(c/\lambda, \widetilde{c}) + a_1 I\right] h_b + \delta v. \end{aligned}$$

In the second-to-last equality above, we use the fact that

$$[H(b,0)+b_nP]\,h_a=[T(0,\widetilde{c})+a_1I]\,h_b \text{ and } H(0,\overline{b}/\overline{\lambda})h_a=T(c/\lambda,0)h_b.$$

Now (4.18) becomes

$$[H(a,\alpha)+a_nP-T(c/\lambda,\widetilde{c})-a_1I]\,h_b=Ch_b=-\delta v.$$

Thus, to parametrize the Hankel matrices B such that $AB=BA$ is to parametrize the solutions h_b and δ satisfying the above equation. The result is the representation formula for B stated in the corollary. We omit the details of its derivation. □

Remark 4.7. If there is no γ_0 such that (4.14) (or (4.17)) holds, then we simply drop the corresponding term containing γ_0 in the representation of B. We also note that the matrix which is a scalar multiple of A is $\lambda_i\left[H(\gamma_i,\overline{\gamma_i}/\overline{\lambda})+\eta_{in}P\right]$ for some i in the representation of B.

REFERENCES

[1] A. Brown and P. R. Halmos, *Algebraic properties of Toeplitz operators,* J. Reine Angew. Math., 213 (1963), pp. 89–102.

[2] D. R. Farenick, M. Krupnik, N. Krupnik, and W. Y. Lee, *Normal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1037–1043.

[3] V. I. Gel'fgat, *A normality criterion for Toeplitz matrices*, Comput. Math. Math. Phys., 35 (1995), pp. 1147–1150.

[4] V. I. Gel'fgat, *Commutation criterion for Toeplitz matrices*, Comput. Math. Math. Phys., 38 (1998), pp. 7–10.

[5] T. N. E. Greville, *Toeplitz matrices with Toeplitz inverses revisited,* Linear Algebra Appl., 55 (1983), pp. 87–92.

[6] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Birkhäuser Verlag, Basel, 1984.

[7] N. M. Huang and R. E. Cline, *Inversion of persymmetric matrices having Toeplitz inverses,* J. ACM, 19 (1972), pp. 437–444.

[8] Kh. D. Ikramov, *On a description of normal Toeplitz matrices*, Comput. Math. Math. Phys., 34 (1994), pp. 399–404.

[9] Kh. D. Ikramov, *Classification of normal Toeplitz matrices with real elements*, Math. Notes, 57 (1995), pp. 463–469.

[10] Kh. D. Ikramov and V. N. Chugunov, *A criterion for the normality of a complex Toeplitz matrix,* Comput. Math. Math. Phys., 36 (1996), pp. 131–137.

[11] Kh. D. Ikramov and V. N. Chugunov, *On the skew-symmetric part of Toeplitz matrices*, Math. Notes, 63 (1998), pp. 124–127.

[12] K. Ito, *Every normal Toeplitz matrix is either of type* I *or of type* II, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 998–1006.

[13] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications,* SIAM Rev., 37 (1995), pp. 297–386.

[14] T. Shalom, *On algebras of Toeplitz matrices*, Linear Algebra Appl., 96 (1987), pp. 211–226.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 747–761

DISTRIBUTION OF SUBDOMINANT EIGENVALUES OF MATRICES WITH RANDOM ROWS*

G. GOLDBERG† AND M. NEUMANN‡

Abstract. In a previous paper, the behavior of the subdominant eigenvalue of matrices $B = (b_{i,j}) \in \mathbb{R}^{n,n}$ whose entries are independent random variables with an expectation $E(b_{i,j}) = 1/n$ and with a variance $\mathrm{Var}(b_{i,j}) \leq c_1/n^2$, for some constant $c_1 \geq 0$, was investigated. For such matrices it was shown that for large n, the subdominant eigenvalues of B are, with great probability, in a small neighborhood of 0. Here we replace the assumption that the *individual* entries of B are independent random variables with the weaker assumption that the *rows* of B are independent n-dimensional random variables but which, within each row, satisfy that $|\mathrm{Cov}(b_{i,j}, b_{i,k})| \leq c_2/n^3$ for some constant $c_2 \geq 0$. We show that under these conditions the subdominant eigenvalues of B continue to tend in probability to 0 as $n \to \infty$. Our assumptions are satisfied, for example, in the case that $B \in \mathbb{R}^{n,n}$ is a stochastic matrix whose rows are chosen from a certain simplex lying in $\mathbb{R}^n$ according to the symmetric Dirichlet distribution satisfying further certain stipulation. The n-dimensional uniform distribution arises as a special case of this stipulation.

Key words. random matrices, eigenvalues, stochastic matrices

AMS subject classifications. 15A52, 15A18, 15A51

PII. S0895479801389102

1. Introduction and main result.

Let $K \in \mathbb{R}^{n,n}$, the space of all real $n \times n$ matrices, and denote the spectral radius of K by $\rho(K)$. Let $(\lambda_1, \ldots, \lambda_n)$ be an arrangement of the eigenvalues of K in which $|\lambda_1| = \rho(K)$. Then a *subdominant eigenvalue* of K is any eigenvalue μ of K for which

$$|\mu| \;=\; \max_{2 \leq i \leq n} |\lambda_i|.$$

In linear iterative methods in which the powers of the iteration matrix converge to *a nonzero limit* so that, necessarily, the spectral radius of the iteration matrix is 1, it is well known that the *magnitude of a subdominant eigenvalue(s) determines the asymptotic rate of convergence of the process*; see, for instance, Berman and Plemmons [2, p. 199]. An important example of an application of such iterative methods occurs in the problem of finding the stationary distribution vector of a finite homogeneous Markov chain by iteration. We now describe this application in more detail. Suppose that $P = (p_{i,j})$ is a (row stochastic) transition matrix for a finite ergodic homogeneous Markov process on n states and let $\gamma(P)$ be the magnitude of a subdominant eigenvalue of P. Let e be the n-vector of all 1's and let v be the stationary distribution vector for the chain, in which case $v^T P = v^T$ and $v^T e = 1$. In Seneta [11, p. 9], it is shown that if $\gamma(P) \neq 0$, then, as $k \to \infty$,

$$P^k \;=\; ev^T + O\left(k^s \gamma^k(P)\right),$$

where s is one less than the largest multiplicity of any subdominant eigenvalue of P.

*Received by the editors May 8, 2001; accepted for publication (in revised form) by H. J. Werner May 13, 2002; published electronically January 31, 2003.
http://www.siam.org/journals/simax/24-3/38910.html

†Citicorp Diners Club Inc., 7958 South Chester Street, Englewood, CO 80112 (grigoriy.goldberg@citicorp.com).

‡Department of Mathematics, University of Connecticut, Storrs, CT 06269–3009 (neumann@math.uconn.edu).

For background material on the statistical concepts used in the paper, see Feller [6]; for background material concerning nonnegative matrices and applications to Markov chains, see Berman and Plemmons [2] and Campbell and Meyer [3]. The working manuscript by Edelman [5] describes applications of large scale random matrices in physics to quantum mechanics and to other disciplines. We also refer the reader to the list of some 200 papers on random matrices and their applications compiled by Edelman [4] which is available on the Web. This list includes works by Girko [7] and Bai [1] on the circular law for random matrices.

In a previous work Goldberg et al. [8] proved results concerning the distribution of the subdominant eigenvalues of $n \times n$ matrices $B = (b_{i,j})$ whose entries are independent random variables from *any distribution*, provided that the entries have an expectation $E(b_{i,j}) = 1/n$ and a variance bounded by c_1/n^2 for some constant $c_1 \geq 0$. In this paper the results in [8] will be extended to the case when the *individual entries* of B are *no longer* independent random variables. *Instead*, it will be assumed that the *rows of B are independent n-dimensional random variables*. To recover the results of [8], we shall need to assume that the elements within each row of B have a covariance $\mathrm{Cov}|(b_{i,k}, b_{j,k})| \leq c_2/n^3$.

An example for a class of random stochastic matrices $B = (b_{i,j}) \in \mathbb{R}^{n,n}$ whose elements satisfy the three basic conditions used in the paper, namely, that

(i) $E(b_{i,j}) = 1/n$,

(ii) $\mathrm{Var}E(b_{i,j}) \leq c_1/n^2$,

and

(iii) $|\mathrm{Cov}(b_{i,j}, b_{i,k})| \leq c_2/n^3$,

we give matrices $B = (b_{i,j}) \in \mathbb{R}^{n,n}$ whose rows are generated as a special case of the n-dimensional Dirichlet distribution. Recall first that the n-dimensional Dirichlet distribution has the *probability density function* given by

$$f(x_1, \ldots, x_n) \;=\; \frac{\Gamma(\nu_1 + \cdots + \nu_n)}{\Gamma(\nu_1) \cdots \Gamma(\nu_n)} x_1^{\nu_1 - 1} \cdots x_{n-1}^{\nu_{n-1} - 1} \left(1 - x_1 - \cdots - x_{n-1}\right)^{\nu_n - 1} \tag{1.1}$$

at any point in the simplex

$$\Sigma_n \;=\; \left\{ (x_1, \ldots, x_n) \mid x_i \geq 0,\ i = 1, \ldots, n,\ \sum_{i=1}^{n} x_i = 1 \right\},$$

where ν_i, $i = 1, \ldots, n$, are positive numbers. We comment that the probability density function in (1.1) is a simple transformation of the probability density function in the formula [13, eq. (7.7.1)] in Wilks's book which is the Dirichlet distribution over the simplex

$$\Sigma_n^{'} \;=\; \left\{ (x_1, \ldots, x_n) \mid x_i \geq 0,\ i = 1, \ldots, n,\ \sum_{i=1}^{n} x_i \leq 1 \right\}$$

to the Dirichlet distribution over Σ_n. In the particular case when $\nu_1 = \nu_2 = \ldots = \nu_n := a$, we obtain the n-dimensional *symmetric* Dirichlet distribution whose density function is given by

$$F_n(a) \;:=\; f(x_1, \ldots, x_n) \;=\; \frac{\Gamma(na)}{[\Gamma(a)]^n} \left(x_1 \cdots x_{n-1}\right)^{a-1} \left(1 - x_1 - \cdots - x_{n-1}\right)^{a-1}.$$

It is known (see Wilks [13, section 7.7]) that the symmetric Dirichlet distribution

satisfies that

(1.2)
$$\begin{cases} E(x_i) \;=\; \dfrac{1}{n}, \quad i=1,\dots,n, \\ \\ \mathrm{Var}^2(x_i) \;=\; \dfrac{(n-1)a}{a^2n(na+1)} \;\le\; \dfrac{1}{a^2}\dfrac{n-1}{n^3} \;<\; \dfrac{1}{a^2}\dfrac{1}{n^2}, \quad i=1,\dots,n, \\ \\ \mathrm{Cov}(x_i,x_j) \;=\; -\dfrac{1}{n^2a^2(na+1)} \;\Rightarrow |\mathrm{Cov}(x_i,x_j)| \;\le\; \dfrac{1}{a^3}\dfrac{1}{n^3},\ i\neq j,\ i,j=1,\dots,n. \end{cases}$$

Thus, if each row of a matrix $B=(b_{i,j})\in\mathbb{R}^{n,n}$ is generated randomly and according to the symmetric Dirichlet distribution, then the requirements (i)–(iii) are fulfilled. In particular, when $a=1$, we obtain that

$$F_n(1) \;=\; \frac{\Gamma(n)}{[\Gamma(1)]^n},$$

which corresponds to the n-dimensional *uniform* distribution function.

We note that in applying assumptions (i)–(iii) above it suffices to assume that $c_1=c_2:=c$.

As in [8], it will be convenient to rewrite the entries of B as follows:

$$b_{i,j} \;=\; \frac{1}{n}+a_{i,j}, \quad 1\le i,j\le n.$$

Then, obviously, $E(a_{i,j})=0$ and $\mathrm{Var}(a_{i,j})=c/3n^2$. The principal difficulty in obtaining a result similar to Theorem 1.1 in [8] is that under the assumptions here we can no longer obtain in a simple way a bound on $E(\det(A^2))$ which was possible there on using Laplace's expansion of the determinant in conjunction with the independence of the entries of A as random variables. As will be seen in section 2, we shall require various combinatorial inequalities to overcome this deficiency.

For the sake of brevity of the statements in the paper we shall now formulate a series of assumptions to which we shall refer throughout the paper as (A,n)-*conditions*.

Assumption (A,n)-*conditions.* A matrix $A=(a_{i,j})\in\mathbb{R}^{\ell,\ell}$ is said to satisfy the (A,n)-*conditions* if there exists a constant $c\ge 0$ such that we have the following:

(i) The entries of A are random variables and the rows of A are independent ℓ-dimensional random variables.

(ii) $E(a_{i,j}) \;=\; 0, \quad i,j=1,\dots,\ell.$

(iii) $\mathrm{Var}(a_{i,j}) \;\le\; c/n^2, \quad i,j=1,\dots,\ell.$

(iv) $|\mathrm{Cov}\,(a_{i,k},a_{i,m})| \;\le\; c/n^3, \quad i,k,m=1,\dots,\ell,\ k\neq m.$

We are now ready to state the main result of this paper.

THEOREM 1.1. *Let* $0<\delta<1$ *and* $0<p<1$. *Suppose that* $B=\left(\frac{1}{n}+a_{i,j}\right)\in\mathbb{R}^{n,n}$, *where* $A=(a_{i,j})\in\mathbb{R}^{n,n}$ *is a matrix satisfying the* (A,n)-*conditions. Then there is an integer* $N(\delta,p)$ *such that for any* $n>N(\delta,p)$ *and for any* r *such that* $1>r>\delta$, *with a probability of at least* p, $n-1$ *of the eigenvalues of* B *are in an open disc of radius* r *centered at the origin.*

For a graphical illustration of some of the results of the theorem, we used the MATLAB random number generator to create $n\times n$ random matrices, $n=2,\dots,300$, as follows. For each n in the range we generated 300 $n\times n$ matrices of the same size, with the rows of the matrix being randomly generated row by row and then scaled

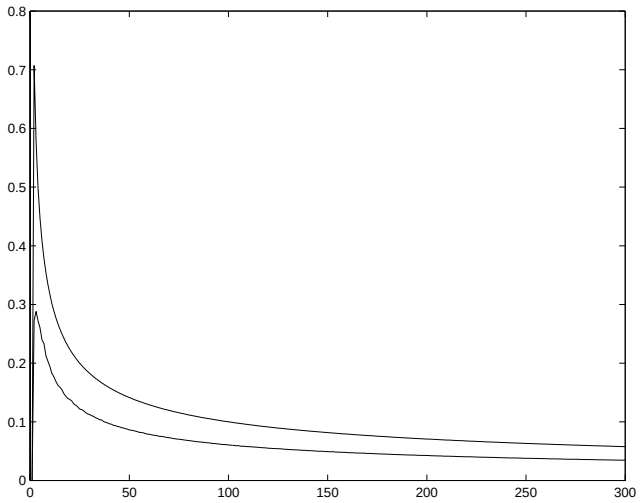


FIG. 1.

to have a unit 1-norm. For each n, the average of the moduli of the subdominant eigenvalue over the 300 matrices was computed. This average as a function of n is the lower curve in Figure 1. The upper curve is a plot of the function $1/\sqrt{n}, n = 2, \ldots, 300$.

We shall devote the next section to the proof of Theorem 1.1. The tools that we shall use to prove the theorem are similar to the tools developed to prove the main theorem in [8] which differs from the present Theorem 1.1 in that there each entry of the matrix is an independent random variable, rather than just the entire rows as here. However, as mentioned before, the absence of the entrywise independence requires that we overcome a variety of difficulties.

The proof of Theorem 1.1 relies on several results which are of independent interest. The main idea of the proof is to split the characteristic polynomial $p_B(\lambda)$ of B into two parts: the *principal part* which equals $\lambda^n - \lambda^{n-1}$ and the *remainder* $g_B(\lambda) := p_B(\lambda) - (\lambda^n - \lambda^{n-1})$. We then use (i) the reverse case of Chebyshev's inequality (which says that if X is a random variable, then $P(|X| < r) \geq 1 - E(X^2)/r^2$; see, for example, Manoukian [9, p. 11, (iv)–(v)]), (ii) Rouché's theorem (which says that if f and h are analytic functions in a domain containing the track and the interior of a closed Jordan contour γ and $|h(z)| < |f(z)|$ on γ, then f and $f + h$ have the same number of zeros inside γ; see, for example, Tall [12, p. 38]), and (iii) a sequence of estimations on the expected values of squares of sums of determinants to show that as $n \to \infty$, with great probability, the characteristic polynomial of B has in any disc of radius $r \neq 1$ as many roots as the polynomial $\lambda^n - \lambda^{n-1}$. From this it follows that for n large enough, with great probability, all the eigenvalues of B with the exception of spectral radius are in a small neighborhood of 0.

2. Proof of Theorem 1.1. As mentioned in the introduction, the proof of Theorem 1.1 is a consequence of a sequence of preliminary results, some of which are of independent interest. Recall that for the random matrix $B \in \mathbb{R}^{n,n}$, we introduced the splitting of its characteristic polynomial into

$$p_B(\lambda) \;=\; (\lambda^n - \lambda^{n-1}) + g_B(\lambda) \tag{2.1}$$

with $\lambda^n - \lambda^{n-1}$ being its *principal part* and with $g_B(\lambda)$ being its *remainder*. The results in this section through Lemma 2.11 have the purpose of allowing in Lemma 2.12 the estimation of $E(|g_B(\lambda)|^2)$ on the boundary of any disc $|\lambda| = r$ with $0 < r < 1$.

We begin with the following lemma.

LEMMA 2.1. *Let $A = (a_{i,j})$ be an $n \times n$ matrix whose entries are random variables with $E(a_{i,j}) = 0$ and such that its rows, as n-dimensional random variables, are independent. Let $X = A[\alpha, \beta]$ and $Y = A[\gamma, \delta]$, with $|\alpha| = |\beta| = \ell$, with $|\gamma| = |\delta| = k$, with $\alpha \neq \gamma$, and where α, β, γ, and δ are strictly increasing ordered subsets of $\{1, 2, \ldots, n\}$. Then*

$$E(\det(X)\det(Y)) \;=\; 0.$$

Proof. Since $\alpha \neq \gamma$, the entries in at least one row in X and the entries in at least one row in Y come from different rows in A. Suppose that X contains elements of the ith row of A and Y contains no elements of that row. Now $\det(X)$ is a sum of $\ell!$ numbers, each of which is up to a sign a product of ℓ elements of X. Similarly, $\det(Y)$ is a sum of $k!$ numbers, each of which is up to a sign a product of k elements of Y. Thus $\det(X)\det(Y)$ is a sum of $\ell! \times k!$ numbers each being equal, up to a sign a product of ℓ elements of X and k elements of Y. To complete the proof of the lemma we need only show that the expectation of every such product is 0. Let

$$\prod_{p=1}^{\ell} a_{j_p,s_p} \prod_{t=1}^{k} a_{q_t,r_t}$$

be such a product. This product contains exactly one element of the ith row of A. Suppose, without loss of generality, that a_{j_1,s_1} is an element of the ith row of A. Then a_{j_1,s_1} is independent from

$$\prod_{p=2}^{\ell} a_{j_p,s_p} \prod_{t=1}^{k} a_{q_t,r_t}$$

since rows of A are mutually independent and therefore

$$E\left(\prod_{p=1}^{\ell} a_{j_p,s_p} \prod_{t=1}^{k} a_{q_t,r_t}\right)$$

$$= \underbrace{E(a_{j_1,s_1})}_{=\,0} E\left(\prod_{p=2}^{\ell} a_{j_p,s_p} \prod_{t=1}^{k} a_{q_t,r_t}\right) \;=\; 0. \qquad \square$$

LEMMA 2.2. *Suppose that $A = (a_{i,j}) \in \mathbb{R}^{k,k}$ satisfies the (A, n)-conditions. If the rows of A are independent as k-dimensional random variables, then*

$$E\left(\det^2(A)\right) \;\leq\; \frac{gb^k k!}{n^{2k}} \tag{2.2}$$

for some nonnegative number g and where $b = 2c$.

Proof. We can write that

$$\begin{aligned}
& E((\det(A))^2) \\
= \; & E\left(\left(\sum_{\sigma\in S_k}\operatorname{sign}(\sigma)a_{1,\sigma(1)}\dots a_{k,\sigma(k)}\right)\left(\sum_{\tau\in S_k}\operatorname{sign}(\tau)a_{1\tau(1)}\dots a_{k\tau(k)}\right)\right) \\
= \; & E\left(\sum_{\sigma\in S_k,\tau\in S_k}\operatorname{sign}(\sigma)a_{1,\sigma(1)}\dots a_{k,\sigma(k)}\operatorname{sign}(\tau)a_{1,\tau(1)}\dots a_{k,\tau(k)}\right) \\
= \; & E\left(\sum_{\sigma\in S_k,\tau\in S_k}\operatorname{sign}(\sigma)\operatorname{sign}(\tau)\left(a_{1,\sigma(1)}a_{1,\tau(1)}\right)\dots\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right)\right) \\
= \; & \sum_{\sigma\in S_k,\tau\in S_k}E\left(\operatorname{sign}(\sigma)\operatorname{sign}(\tau)\left(a_{1,\sigma(1)}a_{1,\tau(1)}\right)\dots\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right)\right) \\
= \; & \sum_{\sigma\in S_k,\tau\in S_k}\operatorname{sign}(\sigma)\operatorname{sign}(\tau)E\left(a_{1\sigma,(1)}a_{1,\tau(1)}\right)\dots E\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right).
\end{aligned}$$

The last line in the display above follows because all the expressions appearing in parentheses in the line above it are mutually independent since they are made up from the elements of different rows of A.

Now fix a permutation σ and consider the sum

$$S_\sigma \; := \; \sum_{\tau\in S_k}\operatorname{sign}(\sigma)E\left(a_{1\sigma,(1)}a_{1,\tau(1)}\right)\dots E\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right).$$

We need to consider two cases.

The case of coincidence. For some $1\le i\le k$, $\sigma(i)=\tau(i)$. Then

$$E\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right) \; = \; \operatorname{Var}(a_{i,\sigma(i)}) \; \le \; \frac{c}{n^2}. \tag{2.3}$$

The case of displacement. $1\le i\le k$, but $\sigma(i)\ne\tau(i)$. Then

$$\left|E\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right)\right| \; = \; \left|\operatorname{Cov}(a_{i,\sigma(i)}a_{i,\tau(i)})\right| \; \le \; \frac{c}{n^3}. \tag{2.4}$$

Suppose now that there are exactly i cases of coincidence and $k-i$ cases of displacement between the permutations σ and τ. Then from (2.3) and (2.4) we have that

$$\left|E\left(a_{1\sigma,(1)}a_{1,\tau(1)}\right)\right|\cdots\left|E\left(a_{k,\sigma(k)}a_{k,\tau(k)}\right)\right| \; \le \; \frac{c^i}{n^{2i}}\,\frac{c^{k-i}}{n^{3(k-i)}} \; = \; \frac{c^k}{n^{3k-i}}. \tag{2.5}$$

Suppose now that $\Delta_{k,i}$ denotes the number of permutations τ which have i coincident indices with σ and $k-i$ displacement indices with σ. Then from (2.5) we have that

$$|S_\sigma| \; \le \; \sum_{i=0}^{k}\Delta_{k,i}\frac{c^k}{n^{3k-i}}. \tag{2.6}$$

We comment that it is known that

$$\Delta_{k,i} \; = \; \binom{k}{i}\Delta_{k-i}, \tag{2.7}$$

where for an integer m, Δ_m is called a *subfactorial* which is known to satisfy that

$$\Delta_m \ \leq \ g \cdot m! \tag{2.8}$$

with $g = 2e^{-1}$; see, for example, Riordan [10, pp. 59–60]. Using (2.6) and (2.7), we can now further estimate S_σ as follows:

$$|S_\sigma| \ \leq \ \sum_{i=0}^{k} \left(\begin{array}{c} k \\ \\ i \end{array} \right) \Delta_{k-i} \frac{c^k}{n^{3k-i}}. \tag{2.9}$$

From (2.8) and (2.9) we now have that

$$|S_\sigma| \ \leq \ \frac{c^k}{n^{2k}} \sum_{i=0}^{k} \left(\begin{array}{c} k \\ \\ i \end{array} \right) g(k-i)! \frac{1}{n^{k-i}}. \tag{2.10}$$

But as $(k-i)! \ < \ (k-i)^{(k-i)} \ < \ k^{k-i}$, (2.10) yields that

$$|S_\sigma| \ \leq \ \frac{gc^k}{n^{2k}} \sum_{i=0}^{k} \left(\begin{array}{c} k \\ \\ i \end{array} \right) \left(\frac{k}{n} \right)^{k-i} \ = \ \frac{gc^k}{n^{2k}} \left(1 + \frac{k}{n} \right)^k \ < \ \frac{gc^k 2^k}{n^{2k}} \tag{2.11}$$

because $1 + k/n \leq 2$. Now put $b = 2c$ and (2.2) follows. □

In our analysis of the behavior of the remainder $g_B(\lambda)$ of the characteristic polynomial, the following definition will be helpful.

DEFINITION 2.3. *Let $\mathcal{S}_k$ be the set of all subsets of $\{1, \ldots, n\}$ of cardinality k. Suppose that $A = (a_{i,j}) \in \mathbb{R}^{n,n}$. For $L \in \mathcal{S}_k$, let $H_L = (h_{i,j})$ be the $n \times n$ matrix defined by*

$$h_{i,j} \ = \ \begin{cases} -a_{i,j} & \text{if } j \in L, \\ \\ -\frac{1}{n} & \text{if } j \notin L \text{ and } i \neq j, \\ \\ \lambda - \frac{1}{n} & \text{if } j \notin L \text{ and } i = j. \end{cases} \tag{2.12}$$

Note that for a fixed $L \in \mathcal{S}_k$, there are exactly $k(n-k)$ sets T in $\mathcal{S}_k$ such that $|L \cap T| = k-1$. Denote these sets by L_i, $i = 1, \ldots, k(n-k)$. In what follows we shall require the next lemma.

LEMMA 2.4 (see [8, Lemma 2.6]). *Suppose that $A = (a_{i,j}) \in \mathbb{R}^{n,n}$ and let $L \in \mathcal{S}_k$ and H_L be as given in Definition* 2.3. *Then*

$$\det(H_L) \ = \ \lambda^{n-k-1} \left(\frac{n-k}{n} - \lambda \right) \xi_L + \frac{1}{n} \sum_{i=1}^{k(n-k)} \lambda^{n-k-1} \xi_{L_i},$$

where ξ_L is up to a sign $\det(A[L, L])$ and ξ_{L_i} is up to a sign $\det(A[L_i, L])$.

Based on Lemma 2.4, the following representation was obtained in [8] for the remainder of the characteristic polynomial of B.

LEMMA 2.5 (see [8, Lemma 2.7]). *Let $B = \left(\frac{1}{n} + a_{i,j}\right)$, where $A = (a_{i,j}) \in \mathbb{R}^{n,n}$. Then the remainder of the characteristic polynomial of B defined via* (2.1) *satisfies that*

$$g_B(\lambda) \ = \ \sum_{k=1}^{n} \left[B_k \left(\frac{n-k}{n} - \lambda \right) + C_k \right] \lambda^{n-k-1},$$

where

$$B_k \;=\; \sum_{i=1}^{\binom{n}{k}} Y_i \tag{2.13}$$

and where

$$C_n \;=\; 0 \;\; \textit{and} \;\; C_k \;=\; \frac{1}{n} \sum_{j=1}^{\binom{n}{k}k(n-k)} X_j, \quad k = 1, \ldots, n-1, \tag{2.14}$$

with the Y_i's and X_j's being up to a sign determinants of distinct $k \times k$ principal submatrices and almost principal submatrices of the matrix $A = (a_{i,j})$, respectively.

The purpose of the next few lemmas is to estimate the expected value of the squares of the B_k's and C_k's which appear in (2.13) and (2.14), respectively, and of other related quantities, all of which will be required in order to approximate $E(|g_B(\lambda)|^2)$ on discs of radius $0 < r < 1$ in Lemma 2.12.

LEMMA 2.6. *Let $A = (a_{i,j}) \in \mathbb{R}^{n,n}$ satisfy the (A, n)-conditions. Let B_k be given in* (2.13). *Then*

$$E\left(B_k^2\right) \;\leq\; g\frac{b^k}{n^k}$$

for some nonnegative number g.

Proof. Now

$$E\left(B_k^2\right) \;=\; E\left(\sum_{j=1}^{\binom{n}{k}} Y_j\right)^2 \;=\; \sum_{j=1}^{\binom{n}{k}} E\left(Y_j^2\right),$$

with the Y_j's being up to a sign the determinants of *different $k \times k$ principal submatrices* of the matrix $(-a_{i,j})$. But then, by Lemma 2.1,

$$E(Y_i Y_j) \;=\; 0$$

whenever $i \neq j$ so that

$$E\left(B_k^2\right) \;=\; E\left(\left(\sum_{j=1}^{\binom{n}{k}} Y_j\right)^2\right) \;=\; \sum_{j=1}^{\binom{n}{k}} E\left(\left(Y_j\right)^2\right).$$

Next from Lemma 2.2 we know that

$$E\left(Y_j^2\right) \;\leq\; g\frac{b^k k!}{n^{2k}}$$

and therefore

$$E\left(B_k^2\right) \;\leq\; \begin{pmatrix} n \\ \\ k \end{pmatrix} g\frac{b^k k!}{n^{2k}}.$$

But then, as $\binom{n}{k} < n^k/k!$, we obtain that

$$E\left(B_k^2\right) \;\le\; g\frac{b^k}{n^k}$$

and our proof is done. □

Our next lemma is concerned with an upper estimate on the expectation of C_k^2.

LEMMA 2.7. *Let $A = (a_{i,j}) \in \mathbb{R}^{n,n}$ satisfy the (A,n)-conditions. Let C_k be given in* (2.14). *Then*

$$E\left(C_k^2\right) \;\le\; g\frac{b^k}{n^k}\left(2k^2 - k + 1\right) \tag{2.15}$$

for some nonnegative number g.

Proof. Put $N := \binom{n}{k}k(n-k)$. Then, by (2.14), to estimate the expectation $E\left(C_k^2\right)$ we need to estimate the expectation $E(((1/n)\sum_{j=1}^N X_j)^2)$. Set

$$P \;:=\; \frac{1}{n^2}\sum_{j=1}^{N} E\left(X_j^2\right) \tag{2.16}$$

and

$$Q \;:=\; \frac{2}{n^2}\sum_{\substack{j=2\\ i<j}}^{N} E\left(X_i X_j\right) \tag{2.17}$$

so that $E\left(C_k^2\right) \;=\; P + Q$.

We begin by showing that P as defined in (2.16) satisfies that

$$P \;\le\; \frac{gb^k}{n^k}. \tag{2.18}$$

From (2.16) and (2.2) we have that

$$P \;\le\; \frac{1}{n^2}N\frac{gb^k k!}{n^{2k}}.$$

But then, as $\binom{n}{k} < n^k/k!$ and $k(n-k) < n^2$, it follows from the definition of N that

$$P \;\le\; \frac{1}{n^2}\frac{n^k}{k!}n^2\frac{gb^k k!}{n^{2k}} \;=\; \frac{gb^k}{n^k}, \tag{2.19}$$

thus establishing (2.18).

To obtain an upper bound on Q given in (2.17) we require several auxiliary lemmas which, for convenience, are collected together in the appendix. We note that there are $\binom{N}{2}$ terms in (2.17). The auxiliary lemmas show that only a small number of the pairs X_i, X_j are correlated.

Returning to the proof of our lemma (Lemma 2.7) and, in particular, to (2.15), we see that (2.15) is an immediate consequence of (2.19) and (3.2). □

In what follows we shall let $p(\cdot)$ be the quadratic given by

$$p(k) \;=\; 2k^2 - k + 1. \tag{2.20}$$

Note that p has no real roots and that it is an increasing function of k. We now have the following lemmas.

LEMMA 2.8. *Let $A = (a_{i,j}) \in \mathbb{R}^{n,n}$ satisfy the (A, n)-conditions and let B_k and C_k be as given in* (2.13) *and* (2.14)*, respectively. Then*

$$E\left(|B_i||B_j|\right) \;\leq\; \frac{gb^{(i+j)/2}}{n^{(i+j)/2}}, \tag{2.21}$$

$$E\left(|B_i||C_j|\right) \;\leq\; \frac{gb^{(i+j)/2}}{n^{(i+j)/2}}(p(j))^{1/2}, \tag{2.22}$$

and

$$E\left(|C_i||C_j|\right) \;\leq\; \frac{gb^{(i+j)/2}}{n^{(i+j)/2}}(p(i))^{1/2}(p(j))^{1/2}. \tag{2.23}$$

Proof. The proof follows from Lemmas 2.6 and 2.7 and the Cauchy–Schwarz inequality. □

LEMMA 2.9. *Let $A = (a_{i,j}) \in \mathbb{R}^{n,n}$ satisfy the (A, n)-conditions. Let B_k and C_k be as given in* (2.13) *and* (2.14)*, respectively. Set*

$$D_k \;=\; \sum_{i+j=k} \left(|B_i||B_j| + |B_i||C_j| + |C_i||C_j|\right). \tag{2.24}$$

Then

$$E(D_k) \;\leq\; g\frac{b^{k/2}}{n^{k/2}}k\left[1 + (p(k))^{1/2}\right]^2. \tag{2.25}$$

Proof. By Lemma 2.8,

$$E(D_k) \;\leq\; g\frac{b^{k/2}}{n^{k/2}}\sum_{i+j=k}\left[1 + (p(j))^{1/2} + (p(i))^{1/2} + (p(i))^{1/2}(p(j))^{1/2}\right].$$

But then, since p given in (2.20) is an increasing function so that $p(i) \leq p(k)$ for all $i = 1, \ldots, k$, we have that

$$E(D_k) \;\leq\; g\frac{b^{k/2}}{n^{k/2}}k\left[1 + 2(p(k))^{1/2} + p(k)\right] \;=\; E(D_k) \;=\; g\frac{b^{k/2}}{n^{k/2}}k\left[1 + (p(k))^{1/2}\right]^2$$

and the proof is done. □

LEMMA 2.10. *Let $B = (\frac{1}{n} + a_{i,j})$, where $A = (a_{i,j}) \in \mathbb{R}^{n,n}$. Then the remainder $g_B(\lambda)$ of the characteristic polynomial defined via* (2.1) *satisfies that*

$$\max_{|\lambda|=r}|g_B(\lambda)| \;\leq\; (1+r)\sum_{k=1}^{n-1}\left[|B_k| + |C_k|\right]r^{n-k-1} + |B_n|. \tag{2.26}$$

Proof. From Lemma 2.5 it follows that

$$g_B(\lambda) \;=\; \sum_{k=1}^{n-1}\left[\left(\frac{n-k}{n} - \lambda\right)B_k + C_k\right]\lambda^{n-k-1} + B_n.$$

Now from the definition of C_k given in (2.14), it follows that $C_n = 0$. Furthermore,

$$\left|\frac{n-k}{n} - \lambda\right| \;\le\; \left|1 - \frac{k}{n}\right| + |\lambda| \;\le\; 1 + r.$$

This completes the proof of the lemma. ∎

To state our next lemma we shall require the following notation. For D_k as given in (2.24) set

$$Z \;:=\; (1+r)^2 \sum_{k=2}^{2n-2} D_k r^{2n-2-k}. \tag{2.27}$$

LEMMA 2.11. *Let $B = \left(\frac{1}{n} + a_{i,j}\right)$, where $A = (a_{i,j}) \in \mathbb{R}^{n,n}$. The remainder $g_B(\lambda)$ of the characteristic polynomial defined via* (2.1) *satisfies that*

$$\left(\max_{|\lambda|=r} |g_B(\lambda)|\right)^2 \;\le\; 2(Z + B_n^2), \tag{2.28}$$

where Z is given in (2.27).

Proof. This follows from Lemma 2.10, the definition of D_k given in (2.24), and the inequality $(x+y)^2 \le 2(x^2+y^2)$. ∎

LEMMA 2.12. *Let $B = \left(\frac{1}{n} + a_{i,j}\right)$, where $A = (a_{i,j}) \in \mathbb{R}^{n,n}$ is a matrix satisfying the (A,n)-conditions. Define*

$$f(k) \;:=\; k\left[1 + (p(k))^{1/2}\right]^2,$$

where $p(k)$ is given in (2.20) *and where*

$$\gamma \;:=\; \left(\frac{b}{nr^2}\right)^{1/2}. \tag{2.29}$$

If $\gamma < 1$, then

$$E\left(\left(\max_{|\lambda|=r} |g_B(\lambda)|^2\right)\right) \;\le\; r^{2n-2}(1+r)^2 F(\gamma), \tag{2.30}$$

where

$$F(\gamma) \;:=\; 2g \sum_{k=2}^{\infty} \gamma^k f(k). \tag{2.31}$$

Proof. Set

$$K \;=\; E\left(\left(\max_{|\lambda|=r} |g_B(\lambda)|^2\right)\right).$$

Then from Lemmas 2.11, 2.9, and 2.2 we have that

$$K \;\le\; 2\left\{\left[\sum_{k=2}^{2n-2} \frac{g b^{k/2}}{n^{k/2}} f(k) r^{2n-k-2}\right](1+r)^2 + \frac{b^n}{n^n}\right\}.$$

Now from (2.29) we see that

$$\sum_{k=2}^{2n-2} \frac{gb^{k/2}}{n^{k/2}} f(k) r^{2n-k-2} \;=\; gr^{2n-2} \sum_{k=2}^{2n-2} \gamma^k f(k).$$

Observing that

$$\frac{b^n}{n^n} \;=\; r^{2n} \frac{b^n}{n^n r^{2n}} \;=\; r^{2n}\gamma^{2n} \;=\; r^{2n-2}(r^2\gamma^{2n}),$$

it follows that

$$g\gamma^{2n} f(2n) \;>\; r^2\gamma^{2n}$$

and we see that

$$K \;\leq\; 2gr^{2n-2}\left[\sum_{k=2}^{2n-2} \gamma^k f(k) + \gamma^{2n} f(2n)\right] \;<\; 2gr^{2n-2} \sum_{k=2}^{\infty} \gamma^k f(k).$$

Finally, when $\gamma < 1$, the series on the right-hand side of the inequality converges by Cauchy's comparison test and our proof is done. □

LEMMA 2.13. *Let $B = \left(\frac{1}{n} + a_{i,j}\right) \in \mathbb{R}^{n,n}$, where A is a matrix satisfying the (A, n)-conditions. Suppose that the rows of A are independent as n-dimensional random variables. Let $g_B(\lambda)$ be the remainder of the characteristic polynomial of B as defined via* (2.1). *Let $0 < r < 1$ be fixed. Then the probability that for all λ with $|\lambda| = r$, $|g_B(\lambda)|$ is strictly less than $|\lambda^n - \lambda^{n-1}|$ and tends to 1 as n tends to infinity.*

Proof. We begin by noting that $|\lambda^n - \lambda^{n-1}| > \frac{1}{2} r^{n-1}(1-r)$ when $|\lambda| = r$. Therefore,

$$P\left(\max_{|\lambda|=r} |g_B(\lambda)| \;<\; \min_{|\lambda|=r} \left|\lambda^n - \lambda^{n-1}\right|\right) \;\geq\; P\left(\max_{|\lambda|=r} |g_B(\lambda)| < \frac{1}{2} r^{n-1}(1-r)\right).$$

But then, by (the reverse case of) Markov's inequality,

$$P\left(\max_{|\lambda|=r} |g_B(\lambda)| \;<\; \frac{1}{2} r^{n-1}(1-r)\right) \;\geq\; 1 - \frac{E\left(\left(\max_{|\lambda|=r} |g_B(\lambda)|\right)^2\right)}{\left[\frac{1}{2} r^{n-1}(1-r)\right]^2}$$

$$\geq\; 1 - \frac{r^{2n-2}(1+r)^2 F(\gamma)}{\left[\frac{1}{2} r^{n-1}(1-r)^2\right]^2} \;=\; 1 - \frac{4(1+r)^2 F(\gamma)}{(1-r)^2}.$$

Now from the definition of $F(\gamma)$, it easily follows that $F(\gamma) \to 0$ as $\gamma \to 0$. Since r is fixed, we see from (2.29) that as $n \to \infty$, $\gamma \to 0$ and our proof is done. □

Proof of Theorem 1.1. We begin by noting that in the interior of every disc in the complex plane of radius $0 < \eta < 1$, the principal part of the characteristic polynomial, namely, $\lambda^n - \lambda^{n-1}$, has precisely $n-1$ zeros.

Fix a $0 < \delta < 1$ and consider the disc of radius $r = \delta$ centered at the origin of the complex plane. *If* on the boundary of this disc the inequality

$$|g_B(\lambda)| \;<\; |\lambda^n - \lambda^{n-1}|$$

would hold, then by Rouché's theorem stated in the introduction the characteristic polynomial of B, $p_B(\lambda)$ would (also) have $n-1$ roots in the interior of the disc. Now by Lemma 2.13 and (2.30) the above inequality being true tends to 1 as n tends to ∞. This proves the theorem for the specific choice of $r = \delta$. If $r > \delta$, then for each n the probability that the characteristic polynomial has $n-1$ roots in the interior of the disc of radius r is only greater than the probability that the characteristic polynomial has $n-1$ roots in the interior of the disc of radius δ. Allowing $n \to \infty$, our proof is done. □

3. Appendix. As mentioned within the proof of Lemma 2.7, we assemble in this appendix several auxiliary lemmas which are necessary to obtain an upper bound on Q given in (2.17). We note that there are $\binom{N}{2}$ terms in (2.17). The auxiliary lemmas show that only a small number of the pairs X_i, X_j are correlated. In the proof of some of these lemmas we shall make use of Lemma 2.1 which tells us that a pair X_i, X_j is correlated if and only if $X_i = X_j$.

AUXILIARY LEMMA 1. *Let $L, M \in \mathcal{S}_k$ with $L \neq M$ and where $\mathcal{S}_k$ is as in Definition 2.3. Suppose that L_i and M_j are subsets of $\mathcal{S}_k$ such that*

$$E\left(\det(A[L_i, L])\, \det(A[M_j, M])\right) \neq 0. \tag{3.1}$$

Then $|L \cap M|$ is equal to $k-1$ or $k-2$.

Proof of Auxiliary Lemma 1. Because of (3.1) we must have, according to Lemma 2.1, that $L_i = M_j := F$. Set $U = L \cap F$ and $V = M \cap F$. Now $|U| = k-1$ and $|V| = k-1$. Put $W = U \cap V$. Since $U, V \subset F$ and $|F| = k$, there are only two possibilities: $|W| = k-1$ or $|W| = k-2$. As $W = L \cap M$ our proof is done. □

AUXILIARY LEMMA 2. *The number N_1 of pairs of sets $M, L \in \mathcal{S}_k$ such that $|L \cap M| = k-1$ is given by*

$$N_1 = \begin{pmatrix} n \\ k-1 \end{pmatrix} \begin{pmatrix} n-(k-1) \\ 2 \end{pmatrix},$$

while the number N_2 of pairs of sets $M, L \in \mathcal{S}_k$ such that $|L \cap M| = k-2$ is given by

$$N_2 = \begin{pmatrix} 4 \\ 2 \end{pmatrix} \begin{pmatrix} n \\ k-2 \end{pmatrix} \begin{pmatrix} n-(k-2) \\ 4 \end{pmatrix}.$$

Proof of Auxiliary Lemma 2. Clearly, there are $\binom{n}{k-1}$ choices that $L \cap M$ can assume and, for each such choice, there are $\binom{n-(k-1)}{2}$ ways of completing $M \cap L$ to a set in $\mathcal{S}_k$. The second part of the lemma is proved in a similar way. □

AUXILIARY LEMMA 3. *Let $L, M \in \mathcal{S}_k$ be such that $|L \cap M| = k-1$. Let $\{L_j\}$ and $\{M_j\}$ be the subsets of $\mathcal{S}_k$ such that $|L \cap L_j| = k-1$ and $|M \cap M_j| = k-1$, respectively. Then there are exactly $n-2$ pairs of sets L_i and M_i from $\{L_j\}$ and $\{M_j\}$, respectively, such that $L_i = M_i$. Similarly, if $L, M \in \mathcal{S}_k$ are such that $|L \cap M| = k-2$ and $\{L_j\}$ and $\{M_j\}$ are the subsets of $\mathcal{S}_k$ for which $|L \cap L_j| = k-1$ and $|M \cap M_j| = k-1$, then there are exactly four pairs of sets L_i and M_i from $\{L_j\}$ and $\{M_j\}$, respectively, such that $L_i = M_i$.*

Proof of Auxiliary Lemma 3. Put $F = L \cap M$ so that for some integers $a, b \in \langle n \rangle$, with $a \neq b$, $L = F \cup \{a\}$, and $M = F \cup \{b\}$. Let $\mathcal{N} = \langle n \rangle \setminus F \cup \{a, b\}$. Then $|\mathcal{N}| = n - (k+1)$. Now for each $\alpha_i \in \mathcal{N}$ consider the sets $L_i = F \cup \{\alpha_i\}$ and

$M_i = F \cup \{\alpha_i\}$, in which case $L_i = M_i$. Clearly, there are $N = n - (k+1)$ such pairs of sets. Next, for each $\beta_i \in F$, let $L_i = (L \setminus \{\beta_i\}) \cup \{b\}$ and $M_i = (M \setminus \{\beta_i\}) \cup \{a\}$ and observe that both sets can be represented as $(F \cup \{a, b\}) \setminus \{\beta_i\}$. Furthermore, as $|F| = k - 1$, there are $k - 1$ such pairs of sets.

We now claim that there are no more pairs L_i and M_i which fulfill the conditions of the lemma. Because if $L_i = M_i$ with $a \in L_i$, but $b \notin L_i$, then $M_i = F \cup \{a\}$, while $L_i \neq \{a\} \cup F = L$ and so $M_i \neq L_i$.

Thus all told, there are $n-(k+1)+(k-1) = n-2$ pairs of subsets of $\mathcal{S}_k$ satisfying the assumptions of the first part of the lemma.

The proof of the second part of the lemma follows in a similar fashion. □

An auxiliary result which follows from Auxiliary Lemmas 2–3 follows.

AUXILIARY LEMMA 4. *Let $L, M \in \mathcal{S}_k$ and let $L_j \in \mathcal{S}_k$ and $M_j \in \mathcal{S}_k$ be* all *the sets for which $|L_j \cap L| = k - 1$ and $|M_j \cap M| = k - 1$. Then among the L_j's and M_j's there are*

$$N_3 \;\leq\; n^{k+2}\left[\frac{1}{2(k-1)!} + \frac{1}{(k-2)!}\right]$$

pairs such that $L_i = M_i$.

Proof of Auxiliary Lemma 4. By Auxiliary Lemmas 2 and 3, $N_3 = J_1 + J_2$, where

$$J_1 \;=\; (n-2)\begin{pmatrix} n \\ k-1 \end{pmatrix}\begin{pmatrix} n-(k-1) \\ 2 \end{pmatrix}$$

and

$$J_2 \;=\; 4\begin{pmatrix} n \\ k-2 \end{pmatrix}\begin{pmatrix} n-(k-2) \\ 4 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Since $\binom{n}{k} < n^k/k!$, we see that

$$J_1 \;<\; \frac{n^{k-1}}{(k-1)!}\frac{[n-(k-1)]^2}{2!}(n-2) \;<\; \frac{n^{k+2}}{2(k-1)!}$$

and

$$J_2 \;<\; \frac{n^{k-2}}{(k-2)^2}\frac{[n-(k-2)]^4}{4!}\frac{4!}{2!2!}\cdot 4 \;<\; \frac{n^{k+2}}{(k-2)!}$$

from which the proof follows. □

AUXILIARY LEMMA 5. *For Q given in* (2.17),

$$|Q| \;\leq\; \frac{gb^k}{n^k}\;(2k^2 - k). \tag{3.2}$$

Proof of Auxiliary Lemma 5. Each X_i is, up to a sign, the determinant of $A[L_i, L]$ for some $L \in \mathcal{S}_k$. By Lemma 2.1, $E(A[L_i, L]A[M_j, M]) \neq 0$ if and only if $L_i = M_j$. Therefore, the number of nonzero terms in (3.2) is precisely the number N_3 appearing

in Auxiliary Lemma 4. But then, by Lemma 2.2 and the Cauchy–Schwarz inequality,

$$\begin{aligned} |Q| &\le \frac{2}{n^2}\, M\, \frac{gb^k k!}{n^{2k}} \\ &\le \frac{2}{n^2} n^{k+2} \left[\frac{1}{2(k-1)} + \frac{1}{(k-2)!}\right] \frac{gb^k k!}{n^{2k}} \\ &= \frac{1}{n^k}\, g\, b^k\, [k + 2k(k-1)] \;=\; \frac{gb^k}{n^k}\, (2k^2 - k), \end{aligned}$$

and the proof is complete. □

REFERENCES

[1] Z. D. Bai, *Circular law*, Ann. Probab., 25 (1997), pp. 494–529.

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.

[3] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover, New York, 1991.

[4] A. Edelman, *Random Eigenvalue Bibliography*, http://www.math.berkeley.edu/~edelman, Mathematical Sciences Research Institute, University of California, Berkeley, CA, 1999.

[5] A. Edelman, *Random Eigenvalues*, Mathematical Sciences Research Institute, University of California, Berkeley, CA, 1999, manuscript.

[6] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, John Wiley, New York, 1966.

[7] V. L. Girko, *An Introduction to Statistical Analysis of Random Arrays*, VSP, Utrecht, The Netherlands, 1988.

[8] G. Goldberg, P. Okunev, M. Neumann, and H. Schneider, *Distribution of subdominant eigenvalues of random matrices*, Methodol. Comput. Appl. Probab., 2 (2000), pp. 137–151.

[9] E. B. Manoukian, *Modern Concepts and Theorems of Mathematical Statistics*, Springer Ser. Statist., Springer-Verlag, New York, 1985.

[10] J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley, New York, 1958.

[11] E. Seneta, *Non–negative Matrices and Markov Chains*, 2nd ed., Springer Ser. Statist., Springer-Verlag, New York, 1981.

[12] D. O. Tall, *Functions of a Complex Variable*, Vol. II, Routledge & Kegan Paul Ltd., London, 1970.

[13] S. S. Wilks, *Mathematical Statistics*, John Wiley, New York, 1962.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 762–767

A COUNTEREXAMPLE TO THE POSSIBILITY OF AN EXTENSION OF THE ECKART–YOUNG LOW-RANK APPROXIMATION THEOREM FOR THE ORTHOGONAL RANK TENSOR DECOMPOSITION*

TAMARA G. KOLDA†

Abstract. Earlier work has shown that no extension of the Eckart–Young SVD approximation theorem can be made to the strong orthogonal rank tensor decomposition. Here, we present a counterexample to the extension of the Eckart–Young SVD approximation theorem to the orthogonal rank tensor decomposition, answering an open question previously posed by Kolda [*SIAM J. Matrix Anal. Appl.*, 23 (2001), pp. 243–355].

Key words. singular value decomposition, principal components analysis, multidimensional arrays, higher-order tensor, multilinear algebra

AMS subject classifications. 15A69, 49M27, 62H25

PII. S0895479801394465

1. Introduction.

We consider the problem of whether or not we can extend the Eckart–Young result to tensors for a particular extension of the SVD known as the *orthogonal rank decomposition*. In other words, suppose a tensor A has an orthogonal rank decomposition of the form

$$A = \sum_{i=1}^{r} \sigma_i U_i.$$

Here, r is the minimal number of terms that can be used to represent A, the σ_i's are scalars such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and the U_i's are *decomposed tensors* (i.e., rank-1 tensors) with the property that any pair of the decomposed tensors are *orthogonal*. Notation and definitions are provided in section 2. The question is: Does the sum of the first k terms yield the best rank-k approximation?

In the case that A is a matrix, the orthogonal rank approximation is equivalent to the SVD approximation where each σ_i is equal to the ith singular value and each U_i is the outer product of the ith left singular vector with the ith right singular vector. For matrices, the Eckart–Young theorem [3] says that the best rank-k approximation to A is indeed given by the sum of the first k terms of the SVD.

Kolda [4] showed that the Eckart–Young approximation property does not hold for the strong orthogonal rank tensor decomposition, another extension of the SVD. Leibovici and Sabatier attempted to show that the Eckart–Young approximation property holds for the orthogonal rank tensor decomposition [5, Theorem 2]. The refutation

*Received by the editors August 29, 2001; accepted for publication (in revised form) by N. J. Higham August 29, 2002; published electronically January 31, 2003. This work was supported by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy, under contracts DE–AC05–96OR22464 with Lockheed Martin Energy Research Corporation and DE-AC04-94AL85000 with Sandia Corporation. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/simax/24-3/39446.html

†Computational Science and Mathematics Research Department, Sandia National Laboratories, Livermore, CA 94551-9217 (tgkolda@sandia.gov).

of that claim in [4] is incorrect,[1] so here we reconsider this issue and show that the Eckart–Young approximation property does not hold for the orthogonal rank tensor decomposition.

Our argument proceeds as follows. In section 3, we present an orthogonal rank decomposition of a tensor A. From the decomposition, we can determine that the orthogonal rank of A is 2. If the Eckart–Young extension hypothesis is true, then the first term of the decomposition should be the best rank-1 approximation of A. In section 4, however, we compute the best rank-1 approximation of A and find that it is not equal to the first term of the orthogonal decomposition presented in section 3. We know from Kolda [4] that the orthogonal rank decomposition is not unique, so in section 5, we consider whether or not we can extend the best rank-1 approximation of A to an orthogonal rank decomposition. We find that the best we can possibly do is produce a 3-term orthogonal decomposition, which is not a rank decomposition. Thus we conclude in section 6 that the Eckart–Young approximation theorem for the SVD cannot be extended to the orthogonal rank tensor decomposition.

2. Notation and definitions. We use the notation and definitions from Kolda [4], briefly summarized here. If A is an $m_1 \times m_2 \times \cdots \times m_n$ tensor, we say the *order* of A is n, and the jth *dimension* of A is m_j. The set of all tensors of size $m_1 \times m_2 \times \cdots \times m_n$ is denoted by $\mathcal{T}(m_1, m_2, \ldots, m_n)$.

Decomposed tensors are the building blocks of tensor decompositions. A *decomposed tensor* is a tensor $U \in \mathcal{T}(m_1, m_2, \ldots, m_n)$ that can be written as

$$U = u^{(1)} \otimes u^{(2)} \otimes \cdots \otimes u^{(n)},$$

where $\otimes$ denotes the outer product and each $u^{(j)} \in \mathbb{R}^{m_j}$ for $j = 1, \ldots, n$. The vectors $u^{(j)}$ are called the *components* of U. The set of all decomposed tensors of size $m_1 \times m_2 \times \cdots \times m_n$ is denoted by $\mathcal{D}(m_1, m_2, \ldots, m_n)$.

Let $U, V \in \mathcal{D}(m_1, m_2, \ldots, m_n)$. We say that U and V are *orthogonal* ($U \perp V$) if

$$U \cdot V = \prod_{j=1}^{n} u^{(j)} \cdot v^{(j)} = 0.$$

The *orthogonal rank* of A, denoted $\operatorname{rank}_{\perp}(A)$, is defined to be the minimal r such that A can be expressed as

$$A = \sum_{i=1}^{r} \sigma_i U_i,$$

where $U_i \perp U_j$ for all $i \neq j$ and $\|U_i\| = 1$ for all i. This decomposition is called the *orthogonal rank decomposition*. Other decompositions are described by Kolda [4], including the *strong orthogonal rank decomposition* mentioned in section 1.

3. An example tensor with orthogonal rank 2. Consider the following tensor $A \in \mathcal{T}(m, m, m)$ defined by

$$A = \sigma_1 \underbrace{a \otimes a \otimes a}_{U_1} + \sigma_2 \underbrace{b \otimes b \otimes \hat{a}}_{U_2}. \tag{3.1}$$

[1] It was also erroneous to refer to Remark 6.3 of [5] as a "result" rather than a "remark."

Let the vectors $a, \hat{a} \in \mathbb{R}^m$ be orthogonal (i.e., $a \perp \hat{a}$) with $\|a\| = \|\hat{a}\| = 1$. Define $b = \frac{1}{\sqrt{2}}(a + \hat{a})$. Let $\sigma_1, \sigma_2 \in \mathbb{R}$ with $\sigma_1 > \sigma_2 > 0$. Observe that $U_1 \perp U_2$, so $\operatorname{rank}_\perp(A) \leq 2$. Further, we can see that we cannot reduce this to a single decomposed tensor since the span in *every* component has dimension 2. Thus, we can conclude that

$$\operatorname{rank}_\perp(A) = 2$$

and that (3.1) is an orthogonal rank decomposition of A.

4. The best rank-1 approximation. We directly compute the best rank-1 approximation of A in (3.1), which we denote by

$$A_1 = \gamma\, x \otimes y \otimes z, \tag{4.1}$$

where $\gamma > 0$ and $\|x\| = \|y\| = \|z\| = 1$. Note that we may assume that γ is positive since its sign can be absorbed into, e.g., the x-vector without affecting the quality of the approximation. We proceed to solve for γ, x, y, z.

Consider the first component. Without loss of generality, we assume $x \in \operatorname{span}\{a, \hat{a}\}$. Let $\hat{x}$ be the orthogonal complement of x in the space defined by $\operatorname{span}\{a, \hat{a}\}$. Then we can define $\alpha_x, \beta_x \in \mathbb{R}$ such that

$$\begin{aligned} x &= \alpha_x a + \beta_x \hat{a}, \\ \hat{x} &= \beta_x a - \alpha_x \hat{a}, \\ a &= \alpha_x x + \beta_x \hat{x}, \qquad (4.2) \\ \hat{a} &= \beta_x x - \alpha_x \hat{x}. \qquad (4.3) \end{aligned}$$

Using these definitions, we can express b as

$$b = \frac{(\alpha_x + \beta_x)}{\sqrt{2}}\, x - \frac{(\alpha_x - \beta_x)}{\sqrt{2}}\, \hat{x}.$$

We can produce similar decompositions for the second and third components using y and z, respectively. We can then rewrite A in terms of x and $\hat{x}$ in the first component, y and $\hat{y}$ in the second component, and z and $\hat{z}$ in the third component; in other words, we can rewrite A as the sum of eight terms which are the combinations of $\{x, \hat{x}\} \otimes \{y, \hat{y}\} \otimes \{z, \hat{z}\}$ as follows:

$$\begin{aligned} A = &[\ \sigma_1\ \alpha_x\alpha_y\alpha_z + \tfrac{\sigma_2}{2}\ (\alpha_x + \beta_x)\,(\alpha_y + \beta_y)\,\beta_z\ ]\ x \otimes y \otimes z \\ &+ [\ \sigma_1\ \alpha_x\alpha_y\beta_z - \tfrac{\sigma_2}{2}\ (\alpha_x + \beta_x)\,(\alpha_y + \beta_y)\,\alpha_z\ ]\ x \otimes y \otimes \hat{z} \\ &+ [\ \sigma_1\ \alpha_x\beta_y\alpha_z - \tfrac{\sigma_2}{2}\ (\alpha_x + \beta_x)\,(\alpha_y - \beta_y)\,\beta_z\ ]\ x \otimes \hat{y} \otimes z \\ &+ [\ \sigma_1\ \alpha_x\beta_y\beta_z + \tfrac{\sigma_2}{2}\ (\alpha_x + \beta_x)\,(\alpha_y - \beta_y)\,\alpha_z\ ]\ x \otimes \hat{y} \otimes \hat{z} \\ &+ [\ \sigma_1\ \beta_x\alpha_y\alpha_z - \tfrac{\sigma_2}{2}\ (\alpha_x - \beta_x)\,(\alpha_y + \beta_y)\,\beta_z\ ]\ \hat{x} \otimes y \otimes z \\ &+ [\ \sigma_1\ \beta_x\alpha_y\beta_z + \tfrac{\sigma_2}{2}\ (\alpha_x - \beta_x)\,(\alpha_y + \beta_y)\,\alpha_z\ ]\ \hat{x} \otimes y \otimes \hat{z} \\ &+ [\ \sigma_1\ \beta_x\beta_y\alpha_z + \tfrac{\sigma_2}{2}\ (\alpha_x - \beta_x)\,(\alpha_y - \beta_y)\,\beta_z\ ]\ \hat{x} \otimes \hat{y} \otimes z \\ &+ [\ \sigma_1\ \beta_x\beta_y\beta_z - \tfrac{\sigma_2}{2}\ (\alpha_x - \beta_x)\,(\alpha_y - \beta_y)\,\alpha_z\ ]\ \hat{x} \otimes \hat{y} \otimes \hat{z}. \end{aligned} \tag{4.4}$$

The coefficient of the $x \otimes y \otimes z$ term is

$$\gamma = \sigma_1\alpha_x\alpha_y\alpha_z + \frac{\sigma_2}{2}(\alpha_x + \beta_x)(\alpha_y + \beta_y)\beta_z.$$

The best rank-1 approximation of the form in (4.1) is produced by maximizing γ [2]:

$$\begin{aligned} \max \quad & \sigma_1\alpha_x\alpha_y\alpha_z + \frac{\sigma_2}{2}(\alpha_x+\beta_x)(\alpha_y+\beta_y)\beta_z \\ \text{s.t.} \quad & \alpha_x^2+\beta_x^2=1, \\ & \alpha_y^2+\beta_y^2=1, \\ & \alpha_z^2+\beta_z^2=1. \end{aligned} \tag{4.5}$$

First observe that none of the α's can be zero because in that case we have

$$\gamma = \frac{\sigma_2}{2}(\alpha_x+\beta_x)(\alpha_y+\beta_y)\beta_z \le \frac{\sigma_2}{2}(\sqrt{2})(\sqrt{2})(1) = \sigma_2.$$

From the assumption that $\sigma_1 > \sigma_2$, we can get a larger objective value by simply choosing $\alpha_x = \alpha_y = \alpha_z = 1$ to yield $\gamma = \sigma_1$.

It also turns out that the β's are nonzero, but proving this is more difficult. We must consider the first-order necessary conditions for optimality for (4.5), which produces the following system of equations:

$$\sigma_1\alpha_y\alpha_z + \frac{\sigma_2}{2}(\alpha_y+\beta_y)\beta_z + 2\lambda_x\alpha_x = 0, \tag{4.6}$$

$$\sigma_1\alpha_x\alpha_z + \frac{\sigma_2}{2}(\alpha_x+\beta_x)\beta_z + 2\lambda_y\alpha_y = 0, \tag{4.7}$$

$$\sigma_1\alpha_x\alpha_y + 2\lambda_z\alpha_z = 0, \tag{4.8}$$

$$\frac{\sigma_2}{2}(\alpha_y+\beta_y)\beta_z + 2\lambda_x\beta_x = 0, \tag{4.9}$$

$$\frac{\sigma_2}{2}(\alpha_x+\beta_x)\beta_z + 2\lambda_y\beta_y = 0, \tag{4.10}$$

$$\frac{\sigma_2}{2}(\alpha_x+\beta_x)(\alpha_y+\beta_y) + 2\lambda_z\beta_z = 0. \tag{4.11}$$

Case I. We show $\beta_z \neq 0$ by contradiction. Suppose $\beta_z = 0$. Note that this implies $\alpha_z = \pm 1$ from the equality constraint in (4.5). From (4.9) and (4.10), we get $\lambda_x\beta_x = 0$ and $\lambda_y\beta_y = 0$. Suppose $\lambda_x = 0$. Then we get that $\alpha_y = 0$ from (4.6), but we know none of the α's are zero from the argument above, so this is a contradiction and $\lambda_x \neq 0$. Likewise, we can show $\lambda_y \neq 0$. So, we must have $\beta_x = \beta_y = 0$ and $\alpha_x = \alpha_y = \pm 1$, but then (4.11) yields a contradiction. Thus we conclude that $\beta_z \neq 0$.

Case II. We show $\beta_x \neq 0$ by contradiction. Suppose $\beta_x = 0$. Then from (4.9), we have $(\alpha_y + \beta_y)\beta_z = 0$. From Case I, we know that $\beta_z \neq 0$, so we must have $(\alpha_y + \beta_y) = 0$. Combining this with (4.11) and the fact that $\beta_z \neq 0$, we get $\lambda_z = 0$. Then from (4.8), we get $\alpha_y = 0$ since $\alpha_x = \pm 1$. Once again, since none of the α's can be zero, we have a contradiction. Hence, we must have $\beta_x \neq 0$.

Case III. Using an argument analogous to Case II, we can show that $\beta_y \neq 0$.

Thus we have that every α and β is nonzero, i.e.,

$$\alpha_x \neq 0, \quad \alpha_y \neq 0, \quad \alpha_z \neq 0, \quad \beta_x \neq 0, \quad \beta_y \neq 0, \quad \text{and } \beta_z \neq 0. \tag{4.12}$$

This implies that each component of A_1, the best rank-1 contribution to A, has contributions from both a and $\hat{a}$. Therefore, $A_1 \neq U_1$; i.e., A_1 is not the first term of the orthogonal rank decomposition given in (3.1). In the next section, we attempt to extend A_1 to an orthogonal rank decomposition.

Before we go on, let us show that we may, without loss of generality, assume that all the α's and β's are positive. The argument is as follows.

At any optima of (4.5), each term of γ must be nonnegative. If the first term were negative, we could reverse the sign of α_z, which is nonzero by (4.12), resulting in a larger objective value without affecting the other term nor violating the constraint. Likewise for the second term and β_z. Thus,

$$\alpha_x\alpha_y\alpha_z > 0 \quad \text{and} \quad (\alpha_x + \beta_x)(\alpha_y + \beta_y)\beta_z > 0. \tag{4.13}$$

Additionally, for any optima of (4.5), we must have

$$\operatorname{sign}(\alpha_x) = \operatorname{sign}(\beta_x) \quad \text{and} \quad \operatorname{sign}(\alpha_y) = \operatorname{sign}(\beta_y). \tag{4.14}$$

In this case, if α_x is positive and β_x is negative or vice versa, then reversing the sign of whichever one is not the same as their sum, $(\alpha_x + \beta_x)$, results in a larger objective value without affecting the other term nor violating the constraint. Note that here we assume that if the sum is negative, there is one other negative term in the product which enforces the positivity required by (4.13).

Finally, for any optima of (4.5), we must also have

$$\operatorname{sign}(\alpha_z) = \operatorname{sign}(\beta_z). \tag{4.15}$$

If α_x and α_y are both negative or both positive, then α_z must be positive to ensure that the first term of γ is positive from (4.13). Furthermore, this implies that $(\alpha_x+\beta_x)$ and $(\alpha_y + \beta_y)$ are both negative or both positive by (4.14), so once again β_z must be positive to ensure positivity of the second term of the objective. Likewise, both α_z and β_z must be negative if α_x and α_y have opposite signs.

From (4.14) and (4.15), each (α, β) pair must have the same sign. Now suppose that an (α, β) pair, say the one associated with x, is negative. Then we may *absorb* the minus sign by substituting $x = -x$ and $\hat{x} = -\hat{x}$ in (4.2) and (4.3). Therefore we may assume, without loss of generality, that

$$\alpha_x > 0, \quad \alpha_y > 0, \quad \alpha_z > 0, \quad \beta_x > 0, \quad \beta_y > 0, \quad \text{and } \beta_z > 0. \tag{4.16}$$

5. Extending the rank-1 approximation. Although the best rank-1 approximation to A is not the first term of the orthogonal decomposition in section 3, there is still the possibility that the best rank-1 approximation may be the first term of some *alternate* orthogonal rank decomposition of A since we know that the decomposition is not unique [4, Lemma 3.5]. Therefore we consider the problem of extending the best rank-1 approximation to an orthogonal rank decomposition, i.e., an orthogonal decomposition with only two terms.

Now consider the remainder tensor $R_1 = A - A_1$, consisting of the last seven terms from (4.4). In order for us to extend the best rank-1 approximation defined by A_1 to an orthogonal decomposition of rank 2, we must be able to rewrite R_1 as a *single decomposed tensor* for any choice of σ_1 and σ_2.

From (4.16), we know that all of the α- and β-terms are positive. Observe that as the ratio $\sigma_1/\sigma_2 \to +\infty$, we have $\alpha_x, \alpha_y, \alpha_z \to 1$. In other words, there exists σ_1 sufficiently larger than σ_2, such that

$$\alpha_x > \beta_x \quad \text{and} \quad \alpha_y > \beta_y. \tag{5.1}$$

If we choose σ_1 and σ_2 such that (5.1) holds, then the coefficients in R_1 corresponding to $x \otimes \hat{y} \otimes \hat{z}$ and $\hat{x} \otimes y \otimes \hat{z}$ must be positive. These two terms cannot be

reduced to a single rank-1 term because the span in the first two components has dimension two. Adding any additional nonzero terms from R_1 cannot reduce the number of orthogonal decomposed tensors in the sum.

So, if A_1 is the first term, we cannot express A as the sum of fewer than three decomposed tensors.

6. Conclusion. We conclude that the Eckart–Young approximation theorem cannot be extended to the orthogonal rank tensor decomposition. In section 3, we considered the orthogonal rank decomposition of A given by

$$A = \sigma_1 U_1 + \sigma_2 U_2.$$

If we can extend the Eckart–Young approximation theorem, then $\sigma_1 U_1$ should be the best rank-1 approximation, but we saw in section 4 that this is not the case. On the other hand, the orthogonal rank decomposition is not unique [4], so in section 5 we considered the alternate tack of extending A_1, the best rank-1 approximation, to an orthogonal rank decomposition. In this case, we found that we cannot express A using fewer than three terms whenever A_1 is the first term.

In other words, the best rank-1 decomposition is not nested in the best rank-2 decomposition. Thus we have derived a counterexample to the extension of the Eckart–Young matrix approximation theorem to the orthogonal rank tensor decomposition.

Acknowledgments. I am very grateful to Didier Leibovici for email exchanges that inspired the present work and for bringing to my attention related work on the higher-order SVD (HOSVD) by De Lathauwer, De Moor, and Vandewalle [1, 2]. I would also like to acknowledge the anonymous referees for their extremely helpful critiques and recommendations and to thank Nick Higham for his handling of this manuscript.

REFERENCES

[1] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Analysis, 21 (2000), pp. 1253–1278.

[2] L. De Lathauwer, B. De Moor, and J. Vandewalle, *On the best rank-*1 *and rank-*$(R_1, R_2, \ldots, R_n)$ *approximation of higher-order tensors*, SIAM J. Matrix Analysis, 21 (2000), pp. 1324–1342.

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

[4] T. G. Kolda, *Orthogonal tensor decompositions*, SIAM J. Matrix Analysis, 23 (2001), pp. 243–255.

[5] D. Leibovici and R. Sabatier, *A singular value decomposition of a k-way array for principal component analysis of multiway data, PTA-k*, Linear Algebra Appl., 269 (1998), pp. 307–329.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 768–786

A KRONECKER PRODUCT REPRESENTATION OF THE FAST GAUSS TRANSFORM*

XIAOBAI SUN† AND YUJUAN BAO†

Dedicated to G. W. Stewart on the occasion of his 60th birthday

Abstract. We present a matrix representation for the fast Gauss transform (FGT) originally proposed by Greengard and Strain. With the matrix representation we reveal the matrix structures explored and exploited in the FGT, relate the multidimensional FGT to the one-dimensional FGT via Kronecker products, and unify various FGT versions. Based on the unifying representation, we present also a framework of FGT algorithms that demonstrates an algorithmic approach to utilizing the revealed matrix factor structures and suggests computational varieties for adapting the FGT to architecture specifics as well as application specifics to achieve optimal performance.

Key words. multidimensional FGT, approximations of Gaussians, sparse or structured matrices, Kronecker product

AMS subject classifications. 15A23, 65F99, 65R10

PII. S0895479800380374

1. Introduction.

The d-dimensional *discrete Gauss transform* (DGT) evaluated at a *target* point $t \in \mathcal{R}^d$ may be defined as

$$u(t) = \sum_{j=1}^{n} e^{-\|t-s_j\|^2/\delta^2} q(s_j), \tag{1}$$

where $q(s)$ may be considered as a charge distribution function defined at n *source* points $s_j \in \mathcal{R}^d$, $\|t-s\|$ is the Euclidean distance between t and s, and $\delta > 0$ is the Gaussian parameter. The transform at target t due to a unit charge at single source s is described by the *Gaussian*

$$g(x) = e^{-\|x\|^2/\delta^2} \tag{2}$$

with $x = t - s$. The evaluation at m target points t_i due to n source points s_j can be cast as a matrix-vector product

$$u = Gq, \tag{3}$$

wherein the matrix and vector are defined elementwise as

$$G_{ij} = g(t_i - s_j), \quad q_j = q(s_j), \quad u_i = u(t_i), \quad i = 1:m,\ j = 1:n.$$

The DGT matrix G captures all pairwise Gaussian interactions between targets $\{t_i\}$ and sources $\{s_j\}$. The class of DGT matrices may be characterized, or may appear, in the equivalent exponential form

$$G_{ij} = \rho^{\|t_i - s_j\|^2}, \quad 0 < \rho < 1.$$

*Received by the editors November 2, 2000; accepted for publication (in revised form) by S. Van Huffel September 12, 2002; published electronically January 31, 2003. This work was supported in part by DARPA/DSO grant DABT63-98-1-0001, NSF/CISE grant CAD-9726370, and ARO grant DAAD19-00-0540.

http://www.siam.org/journals/simax/24-3/38037.html

†Department of Computer Science, Duke University, Durham, NC 27708 (xiaobai@cs.duke.edu, byj@cs.duke.edu).

The DGT (3) involves the construction of the DGT matrix G and the matrix-vector product $G \cdot q$, given the sources, targets, charges, and the Gaussian parameter δ (or the base ρ in the general exponential form). We measure the memory allocation requirement and the arithmetic complexity of an algorithm by the number of required data entries in floating-point format and the number of required floating-point operations (flops), respectively. When matrix G is provided or formed explicitly in all elements, the complexity of matrix-vector product $G \cdot q$ using the direct method is $2mn$. The arithmetic complexity of matrix construction with an existing implementation for evaluating exponential functions shall be $O(mn)$. The memory allocation requirement may vary from $O(m+n)$ for the source and target locations only to $O(mn)$ for the entire matrix with explicit elements, depending on how the matrix generation and the computation of matrix-vector products are arranged. The evaluated matrix elements are approximate, except for some special values of δ and special distributions of sources and targets. The approximation accuracy may depend on architecture-dependent precision, the underlying evaluation method, and a user-specified requirement.

Certain DGTs can be computed faster with conventional matrix computation techniques. The d-dimensional DGT matrix with the source and target points at the nodes of a d-dimensional tensor product grid is the Kronecker product of d one-dimensional DGT matrices, based on the simple fact that the d-dimensional Gaussian is the product of d one-dimensional Gaussians. Suppose for convenience that the one-dimensional DGT matrices are square and of same order n_1. Then $m = n = n_1^d$. By exploiting the Kronecker product structure of the DGT matrix we can reduce the arithmetic complexity for a matrix-vector product from $2n^2$ to $2dn^{1+1/d}$. In this approach the DGT matrix is represented by its Kronecker factor matrices. Consequently, the memory allocation requirement for the matrix in the compressed form is reduced from n^2 to dn_1^2, and the arithmetic complexity for matrix construction is reduced to $O(dn_1^2)$ as well. The complexities can be further reduced when the grid is square in each dimension. In this case, the one-dimensional DGT matrices are Toeplitz. The memory allocation requirement is $2dn_1$ only. Via the use of fast Fourier transform (FFT) algorithms [17] the arithmetic complexity of a matrix-vector product is $O(n \log(n))$. Unfortunately, these structures are distorted when either the targets or the sources are not regularly distributed. The desirable structures may be restored by embedding the irregular points on a fine enough tensor grid, but the embedding may increase the transform size substantially and diminish the potential gain in efficiency.

Fortunately, all large DGTs can be computed efficiently with the fast Gauss transform (FGT) of Greengard and Strain [6], which is a novel, relatively recent approach to exploiting the mathematical structures inherent in DGTs. The FGT is of $O(\log(\tau^{-1})(m+n))$ in both arithmetic complexity and memory allocation requirement, where τ, $0 < \tau < 1$, is a tolerance on absolute approximation errors in matrix entries. In other words, the complexities are linear in the total number of sources and targets and decrease as the error tolerance increases. Moreover, the FGT admits arbitrarily distributed source and target points without a trade-off in efficiency.

The FGT may offer additional benefits in many applications where the DGT arises. In nonparametric statistics [14], for instance, on-line kernel density estimation and on-line kernel regression require rapid calculation of (1) at query points, i.e., the target points, when the Gaussian kernel is used [10]. The number n of source points is large, and the dimensionality d may be greater than three. The source points and the query points are nonstatic. The FGT decomposes the transform into three

consecutive subtransforms: the local translation from the sources, the source-target translation, and the local translation to the targets. Once n sources s_j, a function q at s_j, and an estimate of the query range (or the target range) are provided, a set of m reference points within the target range is chosen and the first two transforms are carried out with $O(m+n)$ flops. An evaluation at an arbitrary point t within the query range is then obtained quickly by the local translation to the target t, taking only a constant number of flops instead of $O(n)$ flops. If a query point is out of the estimated target range, then the reference set is updated by the source-target translation with a constant number of flops. When an update in function q or in source distribution takes place, the FGT updates the first two transforms with a constant number of flops also. Such a fast adaptive property is not shared by the other approaches discussed above.

The DGT may also arise in numerical solution of certain differential or integral equations as a result of discretizing its continuous counterpart, the *Gauss transform*,[1]

$$G_{\delta,f}(t) = \int_{\Omega} e^{-\|t-s\|^2/\delta^2} f(s)ds,$$

where f is a function defined on $\Omega \subset \mathcal{R}^d$. For example, the Gauss transform is used in the solution of an initial or boundary value problem for the heat equation by means of potential theory [2, 7, 6, 11, 15], where δ is time dependent. The FGT makes the numerical evaluation at every time step efficient. Other applications of the DGT can be found in, for instance, [6, 9] and references therein.

Two versions of the FGT have appeared in the literature. In the first version, Greengard and Strain [6] explore and exploit the mathematical structure of the DGT by using a series expansion of the Gaussian (2) in Hermite functions. We refer to this version as the *H-version*. An alternative version, given by Greengard and Sun in [8], is based on the plane wave expansion, which we refer to as the *W-version*. A matrix interpretation of the FGT is described briefly in [8] also.

In this paper we elaborate on the matrix interpretation of the FGT. In particular, we reveal the Kronecker product structure at the level of matrix blocks introduced by the FGT, although such a structure may be lost at the matrix level when the sources and targets are not regularly distributed on a tensor product grid. The matrix representation of the FGT introduced here is useful in a few respects. It establishes a connection of the FGT to the conventional techniques and highlights the distinguished feature of the FGT from the viewpoint of matrix computation. It offers a simplified and systematic way of relating the multidimensional FGT to the one-dimensional FGT. Moreover, various FGT versions are unified by our representation approach. Based on the unifying matrix representation, we present also a framework of FGT algorithms that demonstrates a simple algorithmic approach to exploiting the revealed matrix structures and suggests algorithmic varieties for adapting the FGT to architecture specifics as well as application specifics to achieve optimal performance.

The following notation is used throughout this paper. Matrices are denoted by uppercase letters and vectors by lowercase letters. The colon notation, such as $i = 1 : n$, specifies an index enumeration, as in MATLAB. When the colon notation is used as a subscript of a matrix, such as $G(:, j)$, it refers to the whole range of rows and/or columns. Matrices or row vectors are often given by enumerating their elements within a pair of square brackets. The transpose, or Hermitian transpose, of A is denoted by

[1] The Gauss transform is also known as the Gauss–Weierstrass transform, Weierstrass transform, and Hille transform [18].

A^{T} or A^{H}, respectively. The function diag$(\cdot)$ is used to form a diagonal or block diagonal matrix. The symbol $\otimes$ stands for the Kronecker product of two matrices. The symbol $\odot$ stands for the elementwise multiplication of two matrices, also known as the Hadamard product. Any exception to the above notational conventions will be mentioned explicitly.

The rest of this paper is organized as follows. In section 2, we describe in detail the two versions of one-dimensional FGT, a unifying scheme for DGT decompositions, and a few FGT algorithm variants. In section 3, we present the Kronecker product representation of the multidimensional FGT and present an algorithm framework as a set of algorithm-building templates for the d-dimensional FGT. In section 4 we present experimental results. Concluding remarks are in section 5.

2. One-dimensional FGT. The FGT idea is novel in creating an elementwise expansion and inducing a desirable factorization of DGT matrix blocks from the elementwise expansion. Elementwise expansions and matrix factorizations go hand in hand. For example, if matrix $A = LDU$ is a factorization of A, then the elements of A have an expansion of the form $A(i,j) = \sum_{p,q} L(i,p)D(p,q)U(q,j)$. The factorization is exploitable in computing matrix-vector products if the arithmetic complexity and the memory allocation requirement are lower than that of the direct method. Such a factorization is often obtained numerically. However, a matrix factorization for an FFT algorithm is induced from an exact algebraic elementwise expansion and a matrix partition associated with an algebraic group factorization of the coincident source and target points. In contrast, the FGT creates an approximate elementwise expansion with a uniform bound on approximation errors. In this section we introduce first the respective elementwise expansions underlying the H-version and the W-version. Along with the elementwise expansion, we present a geometric partition of the DGT matrix into blocks and the associated reference points for the elementwise expansion in each block to ensure uniform approximation. Then we disclose the relationship between the two FGT versions and introduce other varieties as well.

2.1. The elementwise expansion for the H-version. The Gaussian (2) has the following series expansion about a reference point c:

$$e^{-x^2/\delta^2} = e^{-c^2} \sum_{k=0}^{\infty} \frac{H_k(c)}{k!}(x-c)^k = \sum_{k=0}^{\infty} \frac{h_k(c)}{k!}(x-c)^k, \tag{4}$$

where $H_k(x)$ and $h_k(x) = e^{-x^2} H_k(x)$ are the Hermite polynomials and the associated Hermite functions, respectively. The Hermite polynomials have the recurrence relation

$$H_k(x) = 2xH_{k-1}(x) - 2(k-1)H_{k-2}(x), \quad H_0(x) = 1, \quad H_1(x) = 2x.$$

Expansion (4) can be obtained directly from the Taylor expansion of e^{2xy-y^2} with respect to y. In other words, e^{2xy-y^2} is the generating function for Hermite polynomials; cf. [16]. The following theorem gives a truncated form of the expansion (4) and a bound on the truncation error.

THEOREM 1 (the H-version finite-term expansion). *Let B_s and B_t be a pair of source interval and target interval of length l, centered at s_c and t_c, respectively. Assume that l and δ satisfy the condition*

$$l \le \frac{\alpha\delta}{\sqrt{2}} \tag{5}$$

for some α, $0 < \alpha \le 1$. Then, for arbitrary $s \in B_s$, $t \in B_t$, and any positive integer p,

$$g(t-s) = g_h(t,s,p) + e_h(t,s,p)$$

with

$$g_h(t,s,p) = \sum_{i=0}^{p-1}\sum_{j=0}^{p-1} h_{i+j}(c)\frac{(dt)^i}{i!}\frac{(ds)^j}{j!}, \tag{6}$$

and

$$|e_h(t,s,p)| < 2.18\frac{\sqrt{p+1}}{\sqrt{p+1}-\alpha}\frac{\alpha^p}{\sqrt{e^{c^2}\,p!}}, \tag{7}$$

where $dt = (t-t_c)/\delta$, $ds = -(s-s_c)/\delta$, and $c = (t_c - s_c)/\delta$.

Proof. Substitute $x - c$ in (4) with $dt + ds$ and expand $(dt+ds)^k$ in binomial expansion. We get

$$g(t-s) = \sum_{k=0}^{\infty} h_k(c)\sum_{j=0}^{k}\frac{(dt)^j}{j!}\frac{(ds)^{k-j}}{(k-j)!} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} h_{i+j}(c)\frac{(dt)^i}{i!}\frac{(ds)^j}{j!}.$$

Let $g_h(t,s,p)$ be defined as in the theorem. Then

$$\begin{aligned} |e_h(t,s,p)| &< \sum_{i=0}^{\infty}\sum_{j=p}^{\infty}|h_{i+j}(c)|\left(\frac{|dt|^i}{i!}\frac{|ds|^j}{j!} + \frac{|ds|^i}{i!}\frac{|dt|^j}{j!}\right) \\ &\le 2\sum_{i=0}^{\infty}\sum_{j=p}^{\infty}|h_{i+j}(c)|\frac{1}{\sqrt{2^{i+j}}}\left(\frac{\alpha}{2}\right)^{i+j}\frac{1}{i!j!}. \end{aligned}$$

By Cramer's inequality for Hermite polynomials/functions [13],

$$|h_k(x)| < \gamma\sqrt{e^{-x^2}2^k k!}, \quad \gamma = 1.086435,$$

we have

$$\begin{aligned} |e_h(t,s,p)| &< 2\gamma e^{-c^2/2}\sum_{i=0}^{\infty}\sum_{j=p}^{\infty}\frac{(\alpha/2)^{i+j}}{\sqrt{(i+j)!}}\frac{(i+j)!}{i!j!} \\ &= 2\gamma e^{-c^2/2}\sum_{k=p}^{\infty}\frac{(\alpha/2)^k}{\sqrt{k!}}\sum_{j=0}^{k}\frac{k!}{(k-j)!j!} = 2\gamma e^{-c^2/2}\sum_{k=p}^{\infty}\frac{\alpha^k}{\sqrt{k!}}, \end{aligned}$$

and hence (7). ∎

There are three parameters, p, α, and c, affecting the truncation error. The error decreases as the number p of the retained expansion terms increases. It also decreases with α, which bounds the size of the source and the target intervals, and decreases as the distance c between the reference points t_c and s_c increases. For convenient extension to the multidimensional FGT, we call B_s and B_t the one-dimensional source box and target box, respectively.

The H-version expansion of (6) can be written in matrix-vector form

$$g_h(t,s) = v(dt)^{\mathrm{T}} H(c)\, v(ds), \tag{8}$$

where

$$v(\mu)^{\mathrm{T}} = \left[1, \mu, \frac{\mu^2}{2!}, \frac{\mu^3}{3!}, \ldots, \frac{\mu^{p-1}}{(p-1)!}\right] \tag{9}$$

is the pth Vandermonde vector evaluated at μ, scaled by the factorials, and $H(c)$ is a $p \times p$ Hankel matrix with

$$H(c)_{ij} = h_{i+j}(c), \quad i, j = 0 : p-1. \tag{10}$$

The vectors $v(dt)$ and $v(ds)$ are called the *local expansion vectors* of s and t about their respective box centers. In (8), they are coupled by $H(c)$, which we refer to as the *translation matrix* between source box B_s and target box B_t with respect to the truncated Gaussian interactions.

2.2. The elementwise expansion for the W-version. The expansion of the Gaussian in plane waves is obtained by discretizing the integral representation

$$e^{-x^2} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\tau^2} \cos(2x\tau) d\tau = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\tau^2} e^{2\imath x\tau} dt, \tag{11}$$

where $\imath = \sqrt{-1}$.

THEOREM 2 (the W-version finite-term expansion). *Let B_s and B_t be a pair of source interval and target interval of length l, centered at s_c and t_c, respectively. For any $L \geq 1$, let*

$$\lambda = \frac{\pi\delta}{L\delta + l + |t_c - s_c|}. \tag{12}$$

Then, for arbitrary $s \in B_s$, $t \in B_t$,

$$g(t,s) = g_w(t,s,L) + e_w(t,s,L)$$

with

$$g_w(t,s,L) = \frac{1}{\sqrt{\pi}} \sum_{j=-p}^{p} e^{-(j\lambda)^2} e^{2\imath j\lambda(t-s)/\delta}, \tag{13}$$

where $p = \lceil L/\lambda \rceil$, and

$$|e_w(t,s,L)| < 2\left(\sqrt{\pi} + \frac{\pi}{L}\right) e^{-L^2}. \tag{14}$$

Proof. By (11),

$$e^{-x^2} = \frac{1}{\sqrt{\pi}} \int_{-L}^{L} e^{-\tau^2} e^{\imath 2x\tau} d\tau + e_L(x),$$

where $e_L(x)$ is the error introduced by the reduction in the integration domain,

$$|e_L(x)| < \frac{2}{\sqrt{\pi}} \int_{L}^{\infty} e^{-\tau^2} d\tau < \frac{1}{\sqrt{\pi}L} e^{-L^2}.$$

The trapezoidal quadrature with $\lambda \leq \pi/(L + |x|)$ and $p = \lceil L/\lambda \rceil$ gives

$$\int_{-ph}^{ph} e^{-\tau^2} e^{\imath 2\tau x} d\tau \quad = \quad \sum_{j=-p}^{p} e^{-(jh)^2} e^{\imath 2xjh} + e_T(x)$$

with

$$|e_T(x)| < 2\left(\sqrt{\pi} + \frac{2\sqrt{2}}{L}\right) e^{-L^2}.$$

Substituting x with $(t-s)/\delta$ in the summation terms gives $g_h(t,s,L)$ in (13). Let the step size λ be as defined in (12). Then the bound on e_T holds for any $(t-s)/\delta$ with $t \in B_t$ and $s \in B_s$. The bound on the total truncation error e_h in (14) is the sum of the bounds on e_L and e_T. ∎

Unlike the H-version expansion, the distance between the source box and the target box does not appear explicitly in expansion (13). However, it affects the quadrature step size λ (or the quadrature sampling rate $1/\lambda$). As a matter of fact, for a specified error tolerance τ, the required number $(2p+1)$ of the wave terms increases with the distance between the box centers while the number of the terms in the H-versions decreases with the distance. The specifications of the quadrature step size and the error bound in Theorem 2 are new. Our numerical experiments show that the error control based on Theorem 2 is quite tight.

The truncated expansion in the plane wave form (13) can be written in the matrix-vector form

$$g_w(t,s,L) = w(t)^{\mathrm{H}} T(\lambda)\, w(s), \tag{15}$$

where

$$w(\mu)^{\mathrm{H}} = e^{2\imath\mu[-p:p]\lambda/\delta} \tag{16}$$

is called the *wave* vector of degree p evaluated at μ, and $T(\cdot)$ is a diagonal matrix of order $2p+1$,

$$T(\lambda) = \mathrm{diag}\left(e^{-(j\lambda)^2} | j = -p:p\right), \tag{17}$$

which is the translation matrix between source box B_s and target box B_s. The expression (15) of $g_w(t,s,L)$ can be written equivalently in the form with translation analogous to (8),

$$g_w(t,s,L) = w(dt)^{\mathrm{H}} \hat{T}(c,\lambda)\, w(ds),$$

where $dt = t - t_c$, $ds = s - s_c$, and $\hat{T}(c,\lambda) = T(\lambda) \odot \mathrm{diag}(w(s_c - t_c))$ is diagonal.

2.3. Blockwise factorizations. We are in a position to describe a DGT matrix factorization based on an elementwise expansion. Consider first the case that the matrix is defined on a pair consisting of a source box and a target box satisfying the condition of Theorem 1. Denote by $G(B_t, B_s)$ the DGT matrix defined on the sources in a source interval B_s and the targets in a target interval B_t. The matrix size is $|B_s| \times |B_s|$, where $|B|$ denotes the number of the points in box B. Such a matrix may be a submatrix of a larger DGT matrix. In one extreme case the elements of a DGT matrix are 1×1 submatrices $G(t,s)$.

Define the aggregation matrix of the local expansion vectors over points in a box B centered at x_c as follows:

$$V(B,p) = \left[v(dx_1), v(dx_2), \ldots, v(dx_{|B|})\right], \quad dx_j = (x_j - x_c)/\delta, \tag{18}$$

where $v(dx)$ is the pth Vandermonde vector at dx, as defined in (9). In particular, $V(B_s, p)$ and $V(B_t, p)$ are the aggregation matrices of the local Vandermonde vectors of the sources in B_s and the targets in B_t, respectively.

COROLLARY 3 (the H-version block factorization). *Let (B_t, B_s) be a pair of source and target boxes centered at s_c and t_c, respectively. Assume that (B_t, B_s) satisfies the condition of Theorem* 1. *Then, for any $p > 0$,*

$$\begin{aligned} G(B_t, B_s) &= G_h(B_t, B_s, p) + E_h(B_t, B_s, p), \\ G_h(B_t, B_s, p) &= V(B_t, p)^{\mathrm{T}} H(c) V(B_s, p), \end{aligned} \tag{19}$$

where $H(c)$ is the $p \times p$ Hankel matrix in (10) *and the error matrix $E_h(B_t, B_s, p)$ is uniformly bounded elementwise as in* (7).

A few remarks are in order. (i) The numerical values of the Hankel matrix $H(c)$ depend on c, the distance between box centers. The size of $H(c)$ depends on the number of the retained terms in the Hermite expansion. In other words, the translation matrix is independent of the number and the distribution of the points in B_s or B_t, as long as neither of the boxes is empty. (ii) When $|B_s|$ and $|B_t|$ are much larger than p, $G_h(B_t, B_s)$ is a low-rank approximation to $G(B_t, B_s)$. As $|B_s| + |B_t|$ increases, the rank of $G_h(B_t, B_s)$ remains constant. (iii) A matrix-vector product with matrix $G(B_t, B_s)$ requires $2p(|B_s| + |B_t| + p)$ flops. Thus, within a fixed accuracy requirement on elementwise approximation, the arithmetic complexity for the matrix-vector product is $O(|B_s| + |B_t|)$.

We now turn to blockwise factorizations in the W-version. Define the aggregation matrix of the wave vectors over points x_j in a box B centered at x_c similarly,

$$W(B, p) = \left[w(x_1), w(x_2), \ldots, w(dx_{|B|}) \right], \tag{20}$$

where $w(x)$ is the wave vector of pth degree as defined in (16). In particular, $W(B_s, p)$ and $W(B_t, p)$ are the aggregation matrices of the wave vectors of the sources in B_s and the targets in B_t, respectively.

COROLLARY 4 (the W-version block factorization). *Let (B_t, B_s) be a pair of target and source boxes centered at t_c and s_c, respectively. Let L, λ, and p be defined as in Theorem* 2. *Then,*

$$\begin{aligned} G(B_t, B_s) &= G_w(B_t, B_s, L) + E_w(B_t, B_s), \\ G_w(B_t, B_s) &= W(B_t, p)^{\mathrm{H}} T(\lambda) W(B_s, p), \end{aligned} \tag{21}$$

where $T(\lambda)$ is the diagonal matrix in (17), *and the error matrix E_w is uniformly bounded elementwise as in* (14).

We have a few comments. (i) Similarly to the H-version, the (diagonal) translation matrix in the W-version is independent of $|B_s|$ and $|B_t|$. We emphasize here that the distance between the box centers plays a major but implicit role in the discretization. (ii) When $|B_s|$ and $|B_t|$ are much larger than $2p + 1$, $G_w(B_t, B_s)$ is a low-rank approximation to $G(B_t, B_s)$ in the plane wave form. The matrix-vector product with $G(B_t, B_s)$ is of linear complexity in $|B_s| + |B_t|$ within a fixed accuracy requirement.

The two approximate block factorizations can be related as follows:

$$G_h(B_t, B_s, p) = G_w(B_t, B_s, p') + E_{h-w}.$$

The W-version factorization can be viewed as the result of diagonalizing approximately the Hankel translation matrix $H(c)$. The approximate diagonalization is obtained analytically instead of numerically. Obviously, $E_{h-w} = E_h - E_w$ is small when both components are small.

2.4. Factorizations of general DGT matrices. To obtain an approximate compressed representation of a general DGT matrix G with the errors bounded uniformly, the FGT uses a geometric scheme to partition the DGT matrix into blocks so that the elementwise expansion condition and the error bound are held on many of the blocks. The block partition and blockwise factorizations then jointly induce a factorization of the entire DGT matrix.

The block partition scheme is preceded by a bin-packing of the source and target particles. We may assume, without loss of generality and by adjusting the Gaussian parameter δ, that the sources and targets are in the normalized root box $[0,1]$. Partition $[0,1]$ into k nonoverlapping subboxes of equal size: $B_1 = [0, b_1]$, $B_j = (b_{j-1}, b_j]$ with $b_j = j/k$, $j = 2:k$. Denote by c_j the center of the jth subbox. Every source or target particle falls into a box. We use $B_{j,t}$ and $B_{j,s}$ to distinguish the target set and the source set in box B_j. Accordingly, the DGT matrix is partitioned into a block matrix

$$G = [\, G(B_{i,t}, B_{j,s}) \,]. \tag{22}$$

Block $G(B_{i,t}, B_{i,s})$ aggregates the Gaussian interactions between the sources and targets in the same box B_i and are therefore called the blocks on the geometric diagonal. The blocks on the geometric diagonal may not be on the diagonal in the block index. We may assume that the boxes are arranged by their centers in increasing order.

For the H-version, we fix the expansion length p for the moment. Apply the blockwise factorization to all the matrix blocks and extract the common factor from every row block and every column block. We have

$$\begin{aligned} &G = G_h + E_h, \qquad G_h = D_{V,t}^{\mathrm{T}}\, H\, D_{V,s}, \\ &D_{V,s} = \operatorname{diag}(V(B_{j,s}, p)), \quad D_{V,t} = \operatorname{diag}(V(B_{i,t}, p)), \quad H = [\, H(c_{ij}) \,], \end{aligned} \tag{23}$$

where the first term is in factor form and the factors on the source and target sides are block diagonal. The block matrix H in the middle is composed of the Hankel blocks, as shown in (10), with $c_{ij} = (c_i - c_j)/\delta$. We say H is the *aggregated translation matrix* of the H-version FGT. We determine the box-size parameter α, $0 < \alpha \le 1$, and the number k of subboxes so that condition (5) is met. The truncation error matrix E_h is bounded elementwise as in (7). Since the error bound is larger over the blocks on or close to the geometric diagonal, we determine the expansion length p so that a specified tolerance τ is satisfied on the diagonal blocks, and hence on all the other blocks.

We have, similarly for the W-version,

$$\begin{aligned} &G = G_w + E_w, \qquad G_w = D_{W,t}^{\mathrm{T}} T D_{W,s}, \\ &D_{W,s} = \operatorname{diag}(W(B_{j,s}, p)), \quad D_{W,t} = \operatorname{diag}(W(B_{i,t}, p)), \quad T = [1] \otimes T(\lambda), \end{aligned} \tag{24}$$

where $[1]$ denotes the matrix with all elements equal to 1. The aggregated translation matrix T is composed of $(2p+1) \times (2p+1)$ diagonal matrix blocks that are identical, as shown in (17). The truncation error matrix E_w is bounded elementwise as in (14).

The bound changes with c_{ij} also. We determine p so that a specified error tolerance is met on the blocks most distant from the geometric diagonal.

The FGT partition scheme utilizes the information on the spatial location of the sources and targets. Until the appearance of the fast multipole method [4, 5], such information had not been well exposed on irregularly distributed points.

2.5. The FGT complexity. We show in this section that the arithmetic complexity of the FGT is asymptotically linear in $m+n$. We introduce also a reduction in the constant factor associated with the linear term $m+n$. Consider first the approximate transform with matrix G_h in the factored form. The arithmetic complexity of a matrix-vector product with the block diagonal matrix on the source side is $2pn$, including the computation of the Vandermonde blocks at the same time. The complexity of a matrix-vector product with the block diagonal matrix on the target side is $2pm$. The complexities are indirectly dependent on the block partition because of its effect on the determination of p. The remaining complexity is the matrix-vector multiplication with the translation matrix H. When every box B_i contains both source and target particles, H is a $k \times k$ block matrix of $p \times p$ blocks. The complexity of a matrix-vector product with H is at most $2(pk)^2$, including the formation of the Hankel blocks. On the one extreme, δ is large enough so that no partition is needed and H has only one block, $k = 1$. On the other extreme, δ is sufficiently small so that k, the number of boxes, can be very large, although k is constant in asymptotic complexity. The order of k in the arithmetic complexity is reduced in the FGT by a simple scheme exploiting the decaying property of the Gaussian interaction, as the source and the target are farther apart.

Specifically, the DGT matrix is *split* into two components,

$$G = G_{\text{near}} + G_{\text{far}},$$

where G_{far} consists of the blocks $G(B_t, B_s)$ that are far enough from the geometric diagonal in the sense that

$$|t_c - s_c|/\delta > \sqrt{\ln(1/\tau)} + \frac{\alpha}{\sqrt{2}}.$$

That is, the elements of G_{far} are bounded above by τ. The matrix G_{far} may be empty when δ is big enough for a given τ. We then factorize the matrix G_{near} with the FGT factorization scheme

$$G_{\text{near}} = D_{V,t}^{\mathrm{T}} H_{\text{near}} D_{V,s}.$$

We may say that G_{near} is a banded block matrix with *block semibandwidth*

$$b = \left\lceil \frac{\sqrt{2\ln(1/\tau)}}{\alpha} \right\rceil + 1. \tag{25}$$

The splitting of G amounts to truncating the block translation matrix H into a block banded matrix H_{near} with at most $2b+1$ nonzero blocks in each block row or column. With a fixed threshold τ, the complexity of a matrix-vector multiplication with H_{near} is linear in bk. The total complexity for the H-version transform with G_{near} is then $2p(m+n) + p^2(2 + (2b+1)k)$ in the leading term or no more than $5p(m+n)$ when $k(2b+1) \le (m+n)/p - 2$.

The splitting offers a great benefit to the W-version as well. Recall that the expansion length is determined by the most distant off-diagonal blocks, and the distance increases as δ decreases. The splitting effectively eliminates the far-enough blocks and narrows the determination of the expansion length on the blocks within the bandwidth. The complexity of a matrix-vector product with the translation matrix T is $2(2p+1)(2(k-1)+k)$ by taking advantage of the structure that all the blocks of T are the same diagonal block. The total complexity for the W-version transform with G_{near} is $2\eta(2p+1)(m+n)+4(2p+1)(k+b)$, where η is a modest constant for the wave function evaluations.

The block semibandwidth is determined by the error tolerance τ and the box-size parameter α. Figure 1 illustrates the relationship between the bandwidth and $\tau = 10^{-j}$ as j increases, with $\alpha = 1/2, 1/3, 1/4, 1/8$.

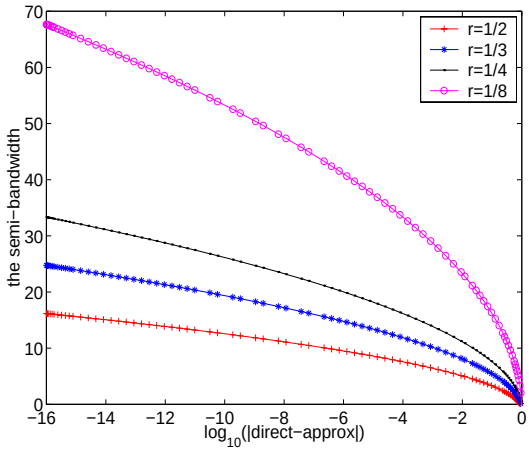


FIG. 1. *The semibandwidth b determined by α and τ.*

2.6. Summary and extensions. In summary, the FGT creates and uses G_{near} as the approximate transform matrix in compressed, factored form. The one-dimensional FGT matrix G_{near} is block banded with low-rank blocks as a result of expansion, partition, split, and factorization. The FGT introduces truncation errors, which are inevitable even with the explicit formation of the DGT matrix. With respect to this fact, the FGT permits higher efficiency when the error tolerance is bigger. On the other hand, the truncation errors introduced by the FGT can be made sufficiently small in comparison to the rounding errors in a given architecture of computing systems.

The FGT can be extended in a number of ways. First, the FGT factorization is not unique, as we have shown. Based on an elementwise expansion, an FGT factorization varies with the accuracy requirement and the partition scheme used. It won't be a surprise when a third expansion for the Gaussian emerges. A different elementwise expansion leads to a different block factorization. The above matrix representation framework can accommodate such emerging varieties.

Second, a hybrid version can be derived to reduce the number of expansion terms without refining the partition. We split a DGT matrix G into three components,

$$G = G_{\text{near1}} + G_{\text{near2}} + G_{\text{far}},$$

where the elements of G_{far} are below the tolerance, G_{near1} has the blocks close to the diagonal, and G_{near2} has the rest of the blocks. Apply the W-version to G_{near1} and the H-version to G_{near2}. Recall that the W-version requires fewer expansion terms in the blocks close to the diagonal and that the H-version requires fewer expansion terms in the blocks away from the diagonal. When G_{near1} (or G_{near2}) is empty, the hybrid method recovers the H-version (or the W-version).

Third, the H-version FGT can be employed at more than one partition level to keep the number of expansion terms low. Consider the case of two partition levels. For a given τ, determine p according to the blocks in G_{near2}. Fix p and reduce the box-size parameter α for G_{near1}. That is, we partition the blocks of G_{near1} into smaller blocks to meet the accuracy requirement in expansions. The local expansion vectors are computed at the finest level only. Let $v(dx)$ be the local expansion vector of x in a box $B(c)$ centered at c at the finer level. Let $v(dx')$ be the local expansion vector of x with respect to c'; the center of the box x resides at the next coarser level. It is easy to verify the following translation relationship between the two local expansion vectors:

$$v(dt') = \text{Toeplitz}(v(c - c'))v(dt),$$

where $\text{Toeplitz}(v(c - c'))$ is a lower triangular Toeplitz matrix with the first column equal to vector $v(c - c')$. There is no need to compute $v(dt')$ explicitly. We have the matrix decomposition

$$G_{\text{near1}} + G_{\text{near2}} = D_{V,t}^{\text{T}}(H_{\text{near1}} + D_{t,c}^{\text{T}} H_{\text{near2}} D_{s,c}) D_{V,s},$$

where $D_{s,c}$ is a block diagonal matrix and each diagonal is a row block matrix $[\text{Toeplitz}(v(c - c'))|B(c) \subset B(c')]$. The transform with the middle matrix $H_{\text{near1}} + D_{t,c}^{\text{T}} H_{\text{near2}} D_{s,c}$ depends only on the number of boxes and the number of expansion terms. The multi-level FGT is similar to the fast multipole method (FMM) [5], except that (i) the translation of the expansion vectors from center to center is more complex in the FMM, and (ii) the partition refinement ratio must obey the decaying rate of the transform kernel in question.

Finally, we extend the matrix representation scheme to multidimensional FGT in the next section.

3. Multidimensional FGT. In this section we present the Kronecker product representation of multidimensional FGT and provide a framework of FGT algorithms. The Kronecker product representation originates in the fact that the Gaussian (2) is separable in the spatial variables. For example, the two-dimensional Gaussian is the product of two one-dimensional Gaussians,

$$g(t, s) = g(t_x, s_x) \cdot g(t_y, s_y) = e^{-|t_x - s_x|^2} e^{-|t_y - s_y|^2}, \tag{26}$$

where $t = (t_x, t_y)$ and $s = (s_x, s_y)$ in Cartesian coordinates. We show that the product property in elements is preserved in the FGT blockwise factors, although such a structure may be lost at the matrix level. The Kronecker product structures in the FGT blockwise factors are a major factor responsible for the high efficiency of the multidimensional FGT.

3.1. The Kronecker product representation. Let B_s be a d-dimensional Cartesian source box centered at s_c. Let B_t be a target box centered at t_c. Assume that the condition of Theorem 1 is satisfied in each and every dimension. We let

$$g(t, s) = g_h(t, s) + e_h(t, s)$$

with

$$g_h(t,s) = g_h(t_{1:d-1}, s_{1:d-1}) \cdot g_h(t_d, s_d). \tag{27}$$

Then, by the fact that $g(t,s) = g(t_{1:d-1}, s_{1:d-1}) \cdot g(t_d, s_d)$ and Theorem 1,

$$|e_h(t,s)| \le |e_h(t_{1:d-1}, s_{1:d-1}) + e_h(t_d, s_d)| + 3|e_h(t_{1:d-1}, s_{1:d-1}) \cdot e_h(t_d, s_d)|.$$

We obtain a representation of $g_h(t,s)$ in the fashion of (8) by applying the following equality on the Kronecker product of matrices [17]:

$$(AB) \otimes (CD) = (A \otimes C) \cdot (B \otimes D), \tag{28}$$

where the ordinary matrix products AB and CD are defined. Notice that $\alpha\beta = \alpha \otimes \beta$ for any scalars α and β. We have

$$\begin{aligned} g_h(t,s) &= v(dt)^{\mathrm{T}} H(c) v(ds), \\ v(dt) &= v(dt_{1:d-1}) \otimes v(dt_d), \\ v(ds) &= v(ds_{1:d-1}) \otimes v(ds_d), \\ H(c) &= H(c_{1:d-1}) \otimes H(c_d), \end{aligned} \tag{29}$$

where $c = (t_c - s_c)/\delta = (c_{1:d-1}, c_d)$, $v(ds)$ is the local expansion vector of the source, and $v(dt)$ is the local expansion vector of the target. The local expansion vectors are of length p^d and coupled by the translation matrix $H(c)$, which is common to all source-target pairs from the source box and the target box.

Similarly to Corollary 3, we have the factorization of the block $G_h(B_t, B_s)$ associated with the pair of d-dimensional boxes,

$$G_h(B_t, B_s) = V(B_t)^{\mathrm{T}} H(c) V(B_s),$$

where $V(B)$, as defined in (18), is the aggregation matrix of local expansion vectors at the particles in box B, and the translation matrix $H(c)$ is independent of the number of sources and the number of targets. With the same algebraic approach we get a factorization of the matrix block from the plane wave expansion, with diagonal translation matrix $\otimes_{1:d} T(\lambda)$.

An approximate factorization of a d-dimensional DGT matrix can then be obtained from the blockwise factorization and a partition-and-split scheme, as for the one-dimensional DGT matrix. We may assume, by adjusting the Gaussian parameter δ, that the sources and the targets are in the unit root box $[0,1]^d$. Partition the unit box into subboxes as the tensor product of the one-dimensional subintervals in all dimensions, where each one-dimensional partition satisfies the condition in Corollary 3. Denote by B_j the jth subbox and by c_j its center. The subboxes may be ordered according to, for example, the lexicographic indexing scheme. The d-dimensional DGT matrix is thus expressed as a block matrix $G = [G(B_{i,t}, B_{j,s})]$. Split the DGT matrix into two parts

$$G = G_{\mathrm{near}} + G_{\mathrm{far}}.$$

A matrix block $G(B_t, B_s)$ is in G_{far} if the source box B_s and the target box B_t are far away in at least one dimension in the sense of (20). This split amounts to truncating the box-to-box translation matrix H into a block matrix H_{near}, with at most $(2b+1)^d$ nonzero blocks in each block row/column.

For the W-version, the block translation matrix T_{near} has a two-level Kronecker product structure. At the block level, every nonzero block is identical to $T_d(\lambda) = \otimes_{1:d} T(\lambda)$. At the matrix level, $T_{\text{near}} = T_{\text{template}} \otimes T_d(\lambda)$; here T_{template} is a binary matrix with values 0 and 1, marking the box-to-box translation between nonempty source and target sets within the banded neighborhood. The template matrix T_{template} has at most $(2b+1)^d$ nonzero elements in each row/column.

3.2. A framework for FGT algorithms. The DGT matrix is specified by the given data δ, $\{t_i\}$, $\{s_j\}$. For an FGT algorithm, an elementwise expansion also is provided. One must determine the partition scheme first, based on the given data, the expansion, and an error tolerance τ. The partition of the DGT matrix is realized by the spatial partition (bin-packing) of the sources and the targets. Then, the FGT approximates the discrete Gauss transform $G \cdot q$ by $G_{\text{near}} \cdot q$, with G_{near} in the factored form (23) or (24). In particular, the computation consists of three stages in the H-version,

$$\begin{aligned} q_c &= \operatorname{diag}(V(B_{j,s}, p))\, q, \\ u_c &= H_{\text{near}}\, q_c, \\ u &= \operatorname{diag}(V(B_{i,t}, p))^{\mathrm{T}} u_c. \end{aligned}$$

The Kronecker product structure in each transform factor exposes the common expressions and suggests certain computational schemes to respect and preserve the structure. In fact, each of the successive transforms consists of a d-sweep operation, one sweep in each dimension. The basic data structure suggested by the Kronecker product structures is the d-dimensional array, *dcube*, where the index in each dimension ranges in $\{0 : p-1\}$ for the H-version and in $\{-p : p\}$ for the W-version. The W-version can be performed either in complex operations or in real operations with additional computation arrangement.

Transform 1 is the translation transform from the source points to the source centers, called the *local-translation* from the sources. We illustrate the transform with one diagonal block associated with a source box B_s. For every source $s(1:d) \in R^d$, we use $dcube[ds]$ to represent the local expansion vector $v(ds)$. With $d = 3$, for example, $dcube[ds](i,j,k) = q(s) ds(1)^i ds(2)^j ds(3)^k / (i!j!k!)$. The cube can be filled by d perfectly nested loops. By setting $dcube[s](0,0,0) = q(s)$, the multiplication with $q(s)$ at every cube cell can be completely avoided. The complexity for filling a *dcube* is p^d. We use the same data structure for the translated vector q_c.

LOCAL TRANSLATION FROM THE SOURCES
Initialize the cube q_c with zeros;
for each source $s \in B_s$
form $dcube[s]$ and update $q_c += dcube[s]$.

For the case of multiple source boxes, we use $q_c[s_c]$ to denote the segment of q_c corresponding to the source box centered at s_c. Let n be the total number of sources. The arithmetic complexity is $2p^d n$, including the formation of the block diagonal matrix on the fly. The same algorithm applies to the W-version.

Transform 2 is the translation transform from the source centers to the target centers, called the *source-target translation*. Two basic operations are involved. The first is the translation from s_c to t_c (box-to-box translation),

$$temp = H(c) q_c = (\otimes_{j=1:d} H(c(j)))\, q_c[s_c], \qquad c = (t_c - s_c)/\delta,$$

where *temp* is a vector in a *dcube* data structure. The algorithm is as follows.

BOX-TO-BOX TRANSLATION : $H(c)q_c[s_c]$
  $temp = q_c[s_c]$;
  for each dimension $j \in \{1, \ldots, d\}$
    for all $tuple = (i_1, \ldots, i_{j-1}, [0 : p-1], i_{j+1}, \ldots, i_d)$
      $temp(tuple) = H(c(j)) \times temp(tuple)$.

Each item of the tuple not in the jth dimension ranges in $\{0, \ldots, p-1\}$. The complexity is at most $2dp^{d+1}$. For the W-version, the computation of $(\otimes_{j=1:d} T(\lambda))q_c[s_c]$ is simplified by forming a *dcube* for the translation block $\otimes T(\lambda)$ once and doing elementwise multiplication with $q_c[t_c]$.

The second operation is to perform the matrix-vector multiplication at the block level, exploiting the structure of the block translation matrix H that has at most $(2b+1)^d$ nonzero blocks in each block row/column. Denote by $u_c[t_c]$ the segment of the vector u_c corresponding to the target box centered at t_c. The algorithm in rowwise operations is as follows.

SOURCE-TARGET TRANSLATION
  for each target center t_c
    set $u_c[t_c]$ with zeros;
    for each source s_c within the neighborhood of b boxes
      $u_c[t_c] += H(c)q_c[s_c]$; % box-to-box translation.

The complexity of the source-target translation is $2dkp^{d+1}(2b+1)^d$ with the H-version.

The nonzero blocks of H_{near} within a block row, and the corresponding segments of q_c, can be visited in a fashion of d-sweeps also. For the W-version, the translation blocks are identically the same and the two consecutive row operations (associated with two target boxes) differ only at the *boundary* blocks [8]. By boundary blocks we mean those associated with the source boxes that are not shared by the consecutive target boxes. The total number of boundary boxes can be minimized by choosing an appropriate ordering among the boxes. For example, it is minimized at the Hilbert ordering [12] in the case where there is no empty target set or empty source set in any partitioned box. The complexity of the source-target translation is $2(2p+1)^d((2b+1)^d + 2(k-1)(2b+1)^{d-1})$ with the W-version.

Transform 3 is the translation transform from the target centers to the targets and is called the local translation to the targets. We illustrate the computation of u_t with one diagonal block $V^{\mathrm{T}}(B_t)$, a transposed aggregation matrix of the local expansion vectors at the targets in box B_t centered at t_c.

LOCAL TRANSLATION TO THE TARGETS
  for every target $t \in B_t$
    $u(t) = 0$; form $dcube[dt]$;
    $u(t) += \mathrm{sum}(dcube[dt] \odot u_c[t_c])$.

$\mathrm{Sum}(A)$ is the sum over all the elements of matrix A. The arithmetic complexity of transform 3 is the same as that of transform 1.

In an application problem where a target point or an evaluation point is not known in advance, such as the on-line kernel density estimation, the FGT computation can be carried out to the intermediate result u_c. The evaluation at a new point t requires only a constant number of operations in first locating the reference point t_c and then translating from $u_c[t_c]$.

We would like to note that the above templates specify only the partial ordering in serial computations. In other words, the templates expose the freedom in the total ordering of computations, which shall be taken advantage of in FGT implementation

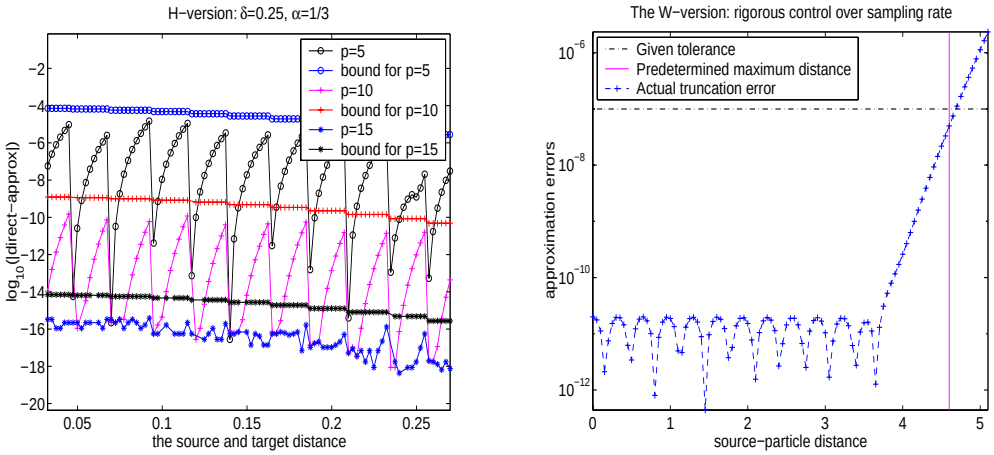


FIG. 2. *Error control over elementwise expansions.*

for any particular application problem and for performance enhancement in a specific computation environment.

4. Experiments. We present in this section three sets of experimental results for the FGT in both the H-version and the W-version. The first set is on the control over the truncation errors in elementwise expansions; see Figure 2. The plot to the left displays the dependence of the truncation error on the expansion length p and the source-target distance, with $\alpha = 1/3$. The stairway-like curves are the error bounds over the partitioned boxes, according to Theorem 1. The plot to the right indicates that the sampling rate and the error bound given in Theorem 2 leave little room for improvement in the error estimation.

The second set is on the overall FGT performance in the accuracy aspect; see Figure 3. The experiment is on a two-dimensional FGT with $m = n = 80{,}000$ random source and target particles in uniform distribution within $[0, 1]^2$. The box-size parameter α is set equal to 1. The charges at the sources are random in $[0, 1]$. The error is estimated by the difference between the computational results via FGT and via the naive direct method. The error is scaled by $\sqrt{n}$, the largest 2-norm of q.

The third set provides some snapshots of the FGT performance in the temporal aspect. The experiments are carried out on an Ultra Sparc 10 with vendor-provided C and FORTRAN compilers. The experiment set-up for the results in Figure 4 is the same as that for Figure 3. The plot to the left shows that the H-version is more time-consuming when δ is small. In this case, most of the H-version time is spent in the source-target translation. The plot to the right shows that the dominant computation is in the local translations when δ is big and there are fewer boxes. Both versions are much faster than the direct method. We use the data in the following table as a comparison reference. The table lists the time comparison in μs for the direct method, the W-version FGT, and the H-version FGT. The experiment set-up remains the same except that the matrix size is made much smaller ($m = n = 10{,}000$) and δ is set equal to $1/\sqrt{20}$. For the case $\tau = 1.0e{-}8$, the computation is carried out

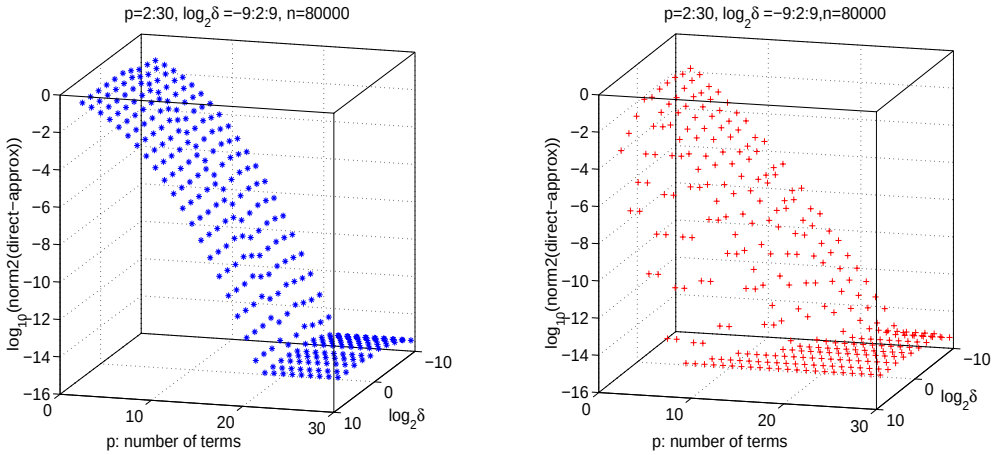


FIG. 3. *Accumulated approximation errors.*

in single precision.

τ	Direct method	W-version FGT	H-version FGT
$1.0e-15$	50080570	2071252	3226422
$1.0e-8$	48351065	499614	236577

One may exploit the flexibility in choosing the parameter values to minimize arithmetic complexity subject to a specific accuracy requirement τ. We have described in section 2.5 the dependency of the box partition and banded truncation on α and δ and the dependency of the arithmetic complexity on k the total number of boxes, b the bandwidth, p the number of expansion terms, and (m, n) the numbers of targets and sources. For one-dimensional FGT ($d = 1$) in the H-version, for instance,

$$k = \left(\frac{\sqrt{2}}{\alpha\,\delta}\right)^d, \quad b = \left\lceil \frac{\sqrt{2\ln(1/\tau)}}{\alpha} \right\rceil + 1.$$

We may determine δ, α, and p to minimize the arithmetic complexity

$$2p(m+n) + p^2[(2b+1)k + 2]$$

subject to the accuracy requirement

$$|e_h(t,s,p)| < 2.18\frac{\sqrt{p+1}}{\sqrt{p+1}-\alpha}\frac{\alpha^p}{\sqrt{e^{c^2}\,p!}} < \tau.$$

Without loss of generality, we set the Gaussian parameter δ to be the maximal so that the targets and sources are in the unit box. The arithmetic complexity can then be minimized over integer p and real number α, $0 < \alpha < 1$. The partition parameter α may be determined off-line if dynamic partition and sorting are not preferred. One shall notice that k may be very large when δ is small and the dimension is high. We consider the case where $m + n$ is sufficiently large so that the minimal complexity is smaller than $2mn$, the cost for the direct method. When an error bound is substantially overestimated, one may obtain reliable sharper bounds from a well-designed numerical experiment. This numerical approach is first proposed and used in

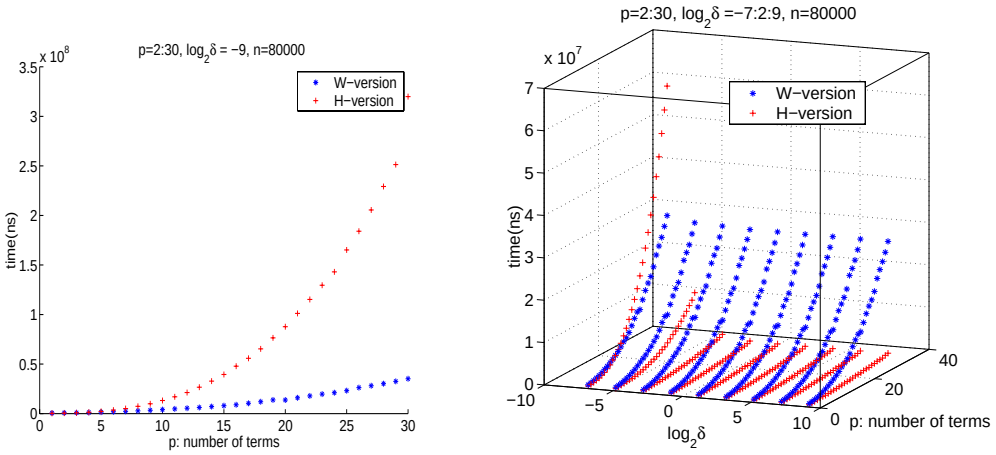


FIG. 4. *Temporal performance on Ultra Sparc* 10.

a code for the FGT in the H-version developed by Florence and Van Loan at Cornell University [1]. As mentioned earlier, our error bounds for the W-versions are shown as very tight.

5. Concluding remarks. We have introduced a unifying scheme for revealing, representing, and exploring the structures of a DGT matrix. The matrix interpretation of the FGT is helpful in understanding the novel ideas behind the FGT from the viewpoint of matrix computation. It has helped establish a simplified and systematic approach to constructing FGT algorithms. The representation scheme consists of elementwise expansion, a geometric matrix partition governed by the expansion condition, matrix split according to the decaying rate and accuracy requirement, and block factorization. The multidimensional FGT can be constructed from the one-dimensional FGT using Kronecker product operations at the level of blockwise factors. The numerical aspect of the FGT is beyond the scope of this paper.

We showed that the higher efficiency of low-accuracy FGT should be exploited whenever possible. We have discussed a few approaches to reducing the number of expansion terms while keeping the same truncation accuracy. Some application problems do not demand high accuracy. A low-accuracy FGT also may be used to speed up computation for the inverse DGT problem with iterative methods. The accuracy for the FGT may be set low in early iteration steps and increase gradually as the iteration proceeds.

The computational framework we present here makes it feasible to create an adaptive software architecture for discrete Gauss transforms, following the FFTW ideas [3]. Because the FGT permits nonregularly distributed sources and targets and adapts to arbitrary approximation requirements, the conventional benchmark notion of the crossover point in temporal performance comparison must be modified. There is no single fixed crossover point in matrix size for all circumstances. The selection of an algorithm or algorithm parameters should be comprehensive because the parameters are not independent of each other, and both application specifics and architecture specifics should be considered.

Acknowledgments. The authors thank the anonymous referees for their careful reading of the manuscript and for suggestions on improving the presentation quality.

The authors thank Nikos Pitsianis and Pau'l Pauca for their discussions and comments throughout the work, and thank Pau'l Pauca for providing the one-dimensional FGT code in C and FORTRAN.

REFERENCES

[1] A. Florence and C. Van Loan, *Fast Gauss Transform C Implementation*, Cornell University, Ithaca, NY, 2000. Available online at http://www.cs.cornell.edu/aflorenc/research/fgt_code.html.
[2] A. Friendman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[3] M. Frigo and S. Johnson, *FFTW: An adaptive software architecture for the FFT*, in Proc. IEEE Internat. Conf. on ASSP, Vol. 3, 1998, pp. 1381–1384.
[4] L. Greengard, *The Rapid Evaluation of Potential Fields in Particle Systems*, ACM Distinguished Dissertations, MIT Press, Cambridge, MA, 1988.
[5] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.
[6] L. Greengard and J. Strain, *The fast Gauss transform*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 79–94.
[7] L. Greengard and J. Strain, *A fast algorithm for evaluating heat potentials*, Comm. Pure Appl. Math., 43 (1990), pp. 949–963.
[8] L. Greengard and X. Sun, *A new version of the fast Gauss transform*, in Proc. Internat. Congress Math., Vol. III (Berlin, 1998), Doc. Math., Extra Vol. III, University Bielefeld, Bielefeld, Germany, 1998, pp. 575–584 (electronic).
[9] K. N. Kudin and G. E. Scuseria, *Linear-scaling density-functional theory with Gaussian orbitals and periodic boundary conditions: Efficient evaluation of energy and forces via the fast multipole method*, Phys. Rev. B (3), 61 (2000), pp. 16440–16453.
[10] C. Lambert, S. Harrington, C. Harvey, and A. Glodjo, *Efficient on-line nonparametric kernel density estimation*, Algorithmica, 25 (1999), pp. 37–57.
[11] W. Pogorzelski, *Integral Equations and Their Applications*, Pergamon Press, Oxford, 1966.
[12] W. T. Rankin, J. A. Board, Jr., and V. L. Henderson, *The impact of data ordering strategies on a distributed hierarchical multipole algorithm*, in Proc. Ninth SIAM Conf. on Parallel Processing for Scientific Computing, B. Hendrickson, K. A. Yelick, C. H. Bischof, I. S. Duff, A. S. Edelman, G. A. Geist, M. T. Heath, M. A. Henoux, C. Koebel, R. S. Schrieber, R. S. Sincovec, and M. F. Wheeler, eds., SIAM, Philadelphia, 1999, CD-ROM.
[13] G. Sansone, *Orthogonal Functions*, Dover, New York, 1991.
[14] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York, 1986.
[15] J. Strain, *The fast Gauss transform with variable scales*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 1131–1139.
[16] G. Szegö, *Orthogonal Polynomials*, 3rd ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1967.
[17] C. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM, Philadelphia, 1992.
[18] A. H. Zemanian, *Generalized Integral Transformations*, Dover, New York, 1987.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 787–801

A SUBSPACE ERROR ESTIMATE FOR LINEAR SYSTEMS*

YANG CAO† AND LINDA PETZOLD†

Abstract. This paper proposes a new method for estimating the error in the solution of linear systems. A condition number is defined for a linear function of the solution components. This definition of the condition number is quite versatile. It reduces to the component condition number proposed by Chandrasekaran and Ipsen [*SIAM J. Matrix Anal. Appl.*, 16 (1995), pp. 93–112] and to Skeel's definition of condition number [*J. ACM*, 26 (1979), pp. 494–526] in some special cases, and it can be used to estimate the error in a subspace. The estimate is based on the adjoint equation in combination with small sample statistical theory. It can be implemented simply and is inexpensive to compute. Numerical examples are presented which illustrate the power and effectiveness of this error estimate.

Key words. condition number, adjoint method, linear system, subspace error estimate

AMS subject classifications. 65F35, 15A12

PII. S0895479801390649

1. Conditioning and error estimation for linear systems.

Perturbation theory for linear systems has been studied for many years. The basic question is, How sensitive is the solution to perturbations in the data? First order analysis is often used in estimating errors; for instance, for stability analysis of algorithms or for the condition number of an eigenvalue.

Consider the linear system

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

where $A \in R^{n\times n}$. The basic question of perturbation theory is, How much will x change if A and b are perturbed? Suppose we are solving a perturbed linear system $(A+\Delta A)\tilde{x} = b+\Delta b$. We would like to estimate the relative error $\|x-\tilde{x}\|/\|x\|$. Here we skip the details of which norm we are using and what kind of perturbation we are assuming. Traditionally the relative error is estimated using the condition number $K(A) = \|A\|\|A^{-1}\|$ and the backward error. The following results are well known [6, p. 133].

If $\frac{\|\Delta A\|}{\|A\|} < \mu$, $\frac{\|\Delta b\|}{\|b\|} < \mu$, and $\mu K(A) < 1$, then

$$\frac{\|x-\tilde{x}\|}{\|x\|} \le \frac{2\mu K(A)}{1-\mu K(A)}. \tag{1.2}$$

Here we take μ to be a multiple of the relative machine precision ϵ_{mach}. The error estimate can be given in terms of the residual $r = A\tilde{x} - b$ by

$$\frac{\|x-\tilde{x}\|}{\|x\|} \le \frac{K(A)\|r\|}{\|A\|\|x\|}. \tag{1.3}$$

*Received by the editors January 9, 2002; accepted for publication (in revised form) by I. C. F. Ipsen August 27, 2002; published electronically January 31, 2003. This work was supported by grants NSF/ITR ACI-0086061, NSF/KDI ATM-9873133, DOE DE-F603-00ER25430, and LLNL ISCR 00-15.

http://www.siam.org/journals/simax/24-3/39064.html

†Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA 93106 (ycao@cs.ucsb.edu, petzold@engineering.ucsb.edu).

When the condition number $K(A)$ is very large, the system is considered to be ill-conditioned and the solution may not be accurate. We call $K(A)$ the standard condition number in the following.

Many examples have demonstrated that the standard condition number may lead to an overly pessimistic estimate for the overall error and that it may underestimate the relative error for some components. Consider the following problems.

Example 1. Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \delta \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ \delta \end{bmatrix},$$

where δ is very small. The solution is $\binom{1}{1}$ The condition number is $\frac{1}{\delta}$. Although for small ϵ the condition number is very large, the solution is accurate. In fact, the solution always has a high relative accuracy for any right-hand side b (assuming a relative perturbation in A and b).

Example 2. Let

$$A = \begin{bmatrix} 1 & 1+\delta \\ 1-\delta & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1+\delta+\delta^2 \\ 1 \end{bmatrix},$$

where δ is a small parameter. Choosing $\delta = 10^{-5}$, the estimate (1.2) will not produce a warning in Matlab [9]. However, the true value of x_2 is 10^{-5} and the result computed by Matlab is 8.8818×10^{-6}, which has relative error of 0.112. There is not even one digit of accuracy! On the other hand, when $\delta = 10^{-4}$, the computed result is 1.000888×10^{-4}, with a relative error of 8.89×10^{-5}. The computed result has four digits of accuracy. The discrepancy can be explained using sensitivity analysis of individual solution components [2].

Example 3. The numerical solution [12] of certain high-index differential-algebraic equations (DAEs) by a fully implicit method yields an ill-conditioned system of linear equations to be solved at each time step. But the propagation of error to future time steps depends only on a well-conditioned subspace. Consider the following simple index-2 DAE system:

$$\begin{cases} \dot{x}_1 &= x_3 + 1, \\ \dot{x}_2 &= x_3 + 2, \\ 0 &= x_1 + x_2 - 1. \end{cases} \tag{1.4}$$

Discretization by the backward Euler method yields a linear system with the matrix

$$A = \begin{bmatrix} 1 & 0 & -h \\ 0 & 1 & -h \\ 1 & 1 & 0 \end{bmatrix}. \tag{1.5}$$

The stepsize h at each time step may be very small. The condition number of A is $O(\frac{1}{h})$ [12, p. 144]. Thus the linear system can be very poorly conditioned for small stepsizes. However, the propagation of error to future time steps for this DAE depends only on errors in the lower-index variables x_1 and x_2. Thus, it is much more critical to get an accurate solution for these variables than for the higher-index variable x_3. In computation, we find that the linear system is solved quite accurately (using Gaussian elimination (GE) with partial pivoting) for x_1 and x_2, and it is only the variable x_3 that is affected by the ill-conditioning. The standard condition number cannot distinguish between the error in the two subspaces.

Many other definitions of condition number have been proposed. References [6], [7], [14], and [16] give some historical review. A precise analysis was given by Skeel [15], leading to a componentwise definition of the condition of the linear system,

$$\text{cond}(A, x) = \frac{\| \, |A^{-1}||A||x| \, \|_\infty}{\|x\|_\infty}, \tag{1.6}$$

where $|A| = \{|a_{ij}|\}$, and the condition number of A,

$$\text{cond}(A) = \||A^{-1}||A|\|_\infty. \tag{1.7}$$

This definition applies to componentwise relative perturbations. It can deal with Example 1 easily and leads to a well-conditioned matrix A for that example. This definition is also a special case of the componentwise analysis described in [6, p. 135]. Unfortunately, the cost to compute A^{-1} is large. In practice, the 1-norm of A^{-1} is estimated [6, p. 290]. In [2], the concept of a componentwise condition number, which yields a condition number for each component of the solution x, was proposed. Thus, for Example 2 we can compute the condition number for x_2 directly and obtain a better error estimate. Example 3 could be handled by computing the component condition number, but for larger DAE systems this could become awkward and expensive.

In this paper we will define a condition number that is applicable in even more general situations. From our experience with solving DAE systems and optimal control problems, we believe that whether or not a solution is acceptable depends on the requirements of the problem. In Example 2, if we are concerned only with the accuracy of component x_2, then the solution is unacceptable. The normwise condition number of the vector x cannot discern this. In Example 3, since we are mainly concerned with the accuracy of x_1 and x_2 but not of x_3, the solution is acceptable although the standard condition number may be very high. This suggests for us to define a condition number that can vary with different requirements. We will use the concept of "derived function" introduced in section 2 to derive such a condition number.

To estimate the condition number, it is not necessary to compute A^{-1} exactly. Typically, one only wants to know the condition number within a factor of 10. Condition estimators with $O(n^2)$ cost, based on the use of random vectors, have been proposed in a number of papers [1], [3], [5], [9], [10], [11]. A detailed review can be found in [6, Chap. 14]. Generally, these estimators yield poorer estimates than the standard condition number but are cheaper to compute. In this paper we will also propose a method that makes use of random vectors to perform error estimation. Our method makes use of the idea and analysis for small sample statistical estimate in [8] although we will not estimate A^{-1} directly. In [4], the complexity of computing error bounds for linear systems is analyzed. The analysis reveals that $O(n^2)$ condition estimators cannot be free of counterexamples. In particular, our $O(n^2)$ condition estimator has low-probability counterexamples which arise from some choices of the random vectors.

The main contribution of this paper has two parts. First, we define a condition number that resolves many of the problems with the standard condition number. It reduces to the component condition number in some special cases and it can be used to estimate the error in a subspace. A subspace condition number is proposed, which helps to separate a well-conditioned subspace from an ill-conditioned system. Second, we provide a means, using small sample statistical theory, of accurately and efficiently computing this condition.

This paper is organized as follows. In section 2 we introduce the concept of an error estimate for a derived function. In section 3 we present our condition estimator for general derived functions and apply it to some examples. In section 4, numerical results are presented which compare this definition with the standard condition number, Skeel's definition [15], and Matlab's condition estimator (which uses Higham's modification of Hager's method [6]). The numerical tests are based on randomly generated dense or banded matrices and right-hand side vectors.

2. Estimating the error of a derived function. Given the linear system (1.1), a derived function is a function $g(x)$ of the solution. We are concerned with the relative error in the derived function, $\|g(x) - g(\tilde{x})\|/\|g(x)\|$. But since the true solution x can only be approximated by the numerical solution $\tilde{x}$, it is more practical to compute $\|g(x) - g(\tilde{x})\|/\|g(\tilde{x})\|$. Before we begin our discussion, we need to specify the norm and what kind of perturbation we are concerned with. In the following, if we do not state a particular norm, the vector norm can be any monotone norm, which satisfies the requirement that if $|x| \leq |y|$, then $\|x\| \leq \|y\|$. For example, the p-norm and the ∞-norm meet this requirement. The matrix norm takes the operator norm.

Generally speaking, we cannot talk about errors without specifying some assumption about the corresponding numerical methods or perturbations. A bad numerical method will result in a large backward error even for a well-conditioned system. For example, it is well known that Cramer's rule gives a large backward error [6, p. 15]. GE without pivoting may lead to a large growth of perturbations for general matrices as well. In this paper we do not want to dig into the details of the numerical methods. Instead we will make some simple but reasonable assumptions about the size of the perturbations. There are two major types of assumption: normwise and componentwise. Normwise analysis assumes $\|\Delta A\| \leq \epsilon\|E\|$ and $\|\Delta b\| \leq \epsilon\|f\|$, while componentwise analysis assumes $|\Delta A| \leq \epsilon E$ and $|\Delta b| \leq \epsilon f$, where E and f are assumed to have nonnegative entries. Different choices of E and f result in different error bounds. As stated in [6, p. 134], the most common choice of tolerance is $E = |A|$ and $f = |b|$. This choice is satisfied by QR factorization [6, p. 369], where $|\Delta A| \leq f(n)\epsilon G|A|$ and $|\Delta b| \leq f(n)\epsilon G|b|$. For LU factorization, $E = |L||U|$ should be used [6, p. 175]. Some special classes of matrices have LU factorization with $|L||U| = |A|$ or $|L||U| \leq 3|A|$ [6, p. 184]. In this paper, we will present a componentwise analysis by taking $E = |A|$ and $f = |b|$.

Different derived functions lead to different condition numbers. When we choose $g(x) = x$ we will obtain the traditional condition number. When we choose $g(x) = x_i$ we will obtain the component condition number. The derived function reflects the requirements of the application. For example, in the application of condition estimate for the linear system generated in a DAE solver, as in Example 3, the derived function is defined via the projection of the solution onto the space of the lower index variables. Thus we will refer to the corresponding error estimate as a *subspace error estimate*. Usually we define the derived function as a linear function of the solution x. Of course we could define a nonlinear derived function, but so far in our applications we have needed only the linear one. Thus we will write the derived function as $g(x) = Lx$, where $L : R^n \longrightarrow R^k$ is a linear function. We assume $\text{rank}(L) = k$.

Consider the perturbed linear system

$$(A + \Delta A)\tilde{x} = b + \Delta b, \tag{2.1}$$

where $|\Delta A| < \epsilon|A|$, $|\Delta b| < \epsilon|b|$. We have

$$A(x - \tilde{x}) = \Delta A\tilde{x} - \Delta b,$$

hence

$$x - \tilde{x} = A^{-1}(\Delta A\tilde{x} - \Delta b), \tag{2.2}$$

and

$$g(x) - g(\tilde{x}) = LA^{-1}(\Delta A\tilde{x} - \Delta b). \tag{2.3}$$

Thus we have the estimate

$$\begin{aligned} \frac{\|g(x) - g(\tilde{x})\|}{\|g(\tilde{x})\|} &\leq \frac{\||LA^{-1}|(|\Delta A||\tilde{x}| + |\Delta b|)\|}{\|L\tilde{x}\|} \\ &\leq \epsilon \frac{\||LA^{-1}|(|A||\tilde{x}| + |b|)\|}{\|L\tilde{x}\|}, \end{aligned} \tag{2.4}$$

and we obtain the condition number

$$\mathrm{cond}_L(A, \tilde{x}) = \frac{\||LA^{-1}|(|A||\tilde{x}| + |b|)\|}{\|L\tilde{x}\|}. \tag{2.5}$$

Supposing that $|b| \leq |A||x|$, and assuming that x is closely approximated by $\tilde{x}$, yields

$$\mathrm{cond}_L(A, \tilde{x}) \leq \frac{2\||LA^{-1}||A||\tilde{x}|\|}{\|L\tilde{x}\|}. \tag{2.6}$$

When we take $g(x) = x$, L is the identity operator, and this definition reduces to the condition number introduced by Skeel [15]. When we take $g(x) = x_i$, this definition reduces to the component condition number defined in [2]. Thus the relative error in the derived function is bounded by

$$\frac{\|g(x) - g(\tilde{x})\|}{\|g(\tilde{x})\|} \leq \mathrm{cond}_L(A, \tilde{x})\epsilon. \tag{2.7}$$

It is easy to generalize the properties of the standard condition number using this definition. We remind the reader that (2.6) and (2.7) are approximate in that they are based on the assumption that x is closely approximated by $\tilde{x}$.

For Example 3, we have

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad A^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2h} & -\frac{1}{2h} & \frac{1}{2h} \end{bmatrix}.$$

Using (2.6), we have

$$\mathrm{cond}_L(A, x) \leq \frac{\sqrt{(|x_1| + |hx_3|)^2 + (|x_2| + |hx_3|)^2 + (|x_1| + |x_2|)^2}}{\sqrt{x_1^2 + x_2^2}}$$

in the 2-norm. The subspace condition number is $O(1)$ even in the case of inconsistent initial conditions for the index-2 DAE (the index-2 variable x_3 can be $O(\frac{1}{h})$ in this case because it is approximating an impulse). This corresponds well with DAE theory [12, p. 144].

3. Condition estimate. Just changing the definition of the condition number doesn't give us much benefit since in practice we may not be able to afford to compute A^{-1} or LA^{-1}. The natural question is, How can we efficiently compute this condition number? We will first give a method based on a scalar derived function, $Lx = l^T x$, where $l \in R^n$, and then extend the estimate for the case of a vector derived function.

3.1. Scalar derived function. For a scalar derived function $g(x) = l^T x$, we can efficiently compute the condition number by first computing the adjoint variable λ which solves

$$A^T \lambda = l \tag{3.1}$$

so that $\lambda^T = l^T A^{-1}$. Assuming that we have the LU or QR decomposition of A, this equation can be solved in $O(n^2)$ cost. Then the condition number becomes

$$\mathrm{cond}_l(A, x) = \frac{|\lambda^T|(|A||x| + |b|)}{|l^T x|}. \tag{3.2}$$

It is the condition number in a particular direction, so we will call it a *directional condition number*. When the direction is toward a single component, this becomes the component condition number.

3.2. Vector derived function. A direct extension of the above defined error estimate to the case of a vector derived function can be quite expensive to compute. Thus we will *estimate* a measure of the vector error by making use of a scalar derived function. To accomplish that, we introduce the small-sample statistical method for estimating the 2-norm (details can be found in [8]). In the following, the norm is the 2-norm.

For any vector $l \in R^n$, if z is selected uniformly and randomly from the unit sphere S_{n-1} in n dimensions, the expected value of $|l^T z|$ is given by

$$E(|l^T z|) = \|l\| E_n,$$

where $E_1 = 1$, $E_2 = \frac{2}{\pi}$, and for $n > 2$,

$$E_n = \frac{1 \cdot 3 \cdot 5 \cdots (n-2)}{2 \cdot 4 \cdot 6 \cdots (n-1)} \quad \text{for } n \text{ odd},$$

$$E_n = \frac{2}{\pi} \cdot \frac{2 \cdot 4 \cdot 6 \cdots (n-2)}{1 \cdot 3 \cdot 5 \cdots (n-1)} \quad \text{for } n \text{ even}.$$

E_n can be estimated by $\sqrt{\frac{2}{\pi(n-\frac{1}{2})}}$. Thus we use $\xi = \frac{|l^T z|}{E_n}$ to estimate $\|l\|$. The estimate satisfies

$$Pr\left(\frac{\|l\|}{w} \le \xi \le w\|l\|\right) \ge 1 - \frac{2}{\pi w} + O\left(\frac{1}{w^2}\right),$$

where $Pr()$ denotes the probability, and $w > 0$ is a real number. The bound does not depend on the vector l. In condition number estimation, usually we are interested in finding an estimate that is accurate to a factor of 10 ($w = 10$).

For a more accurate estimate, we can use more orthogonal random vectors. Suppose we have k orthogonal random vectors $z_1, z_2, \ldots, z_k$. Let

$$\xi_i = \frac{|l^T z_i|}{E_n}.$$

Then the estimate for $\|l\|$ is given by

$$\xi(k) = E_k \sqrt{\xi_1^2 + \cdots + \xi_k^2}. \tag{3.3}$$

Usually, at most two or three random vectors are required in practice. The corresponding probabilities satisfy [8]

$$Pr\left(\frac{\|l\|}{w} \le \xi(2) \le w\|l\|\right) \approx 1 - \frac{\pi}{4w^2},$$

$$Pr\left(\frac{\|l\|}{w} \le \xi(3) \le w\|l\|\right) \approx 1 - \frac{32}{3\pi^2 w^3}.$$

We will use this tool to construct a subspace error estimate for the linear system. To estimate $\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|}$, where L is a linear function from R^n to R^k, we select a vector z uniformly and randomly from the unit sphere S_{k-1}. Let $g_z(x) = z^T L x$. Then $|g_z(x) - g_z(\tilde{x})| = |z^T L(x - \tilde{x})|$. Defining $K_1 = \frac{|z^T L(x-\tilde{x})|}{E_k \|L\tilde{x}\|}$, we have

$$Pr\left(\frac{1}{w}\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|} \le K_1 \le w\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|}\right) \approx 1 - \frac{2}{\pi w}.$$

Taking λ to solve the adjoint equation

$$A^T \lambda = L^T z, \tag{3.4}$$

we have from (2.3),

$$|z^T L(x - \tilde{x})| \le \epsilon |\lambda|^T (|A||\tilde{x}| + |b|).$$

We define

$$e_1 = \frac{|\lambda|^T (|A||\tilde{x}| + |b|)}{E_k \|L\tilde{x}\|},$$

where λ solves (3.4). We have $K_1 \le e_1 \epsilon$. The condition estimate is given by e_1. The relative error is estimated by $e_1\epsilon$. When $L = I$, this differs from the traditional relative error bound by a factor of E_k. Note that K_1 approximates the relative error with a high probability, and $e_1\epsilon$ is an upper bound for K_1. Thus $e_1\epsilon$ is usually larger than the relative error.

Numerical experiments show that this estimate, using one random vector, gives a good result for most cases. But for some random vectors, it may produce a large error. In this situation, using more random orthogonal vectors improves the result. To keep the computational cost low, we use at most two or three random orthogonal vectors. Given orthogonal vectors $z_i \in R^k$, define

$$K_2 = \frac{E_2 \sqrt{(z_1^T L(x-\tilde{x}))^2 + (z_2^T L(x-\tilde{x}))^2}}{E_k \|L\tilde{x}\|},$$

$$K_3 = \frac{E_3 \sqrt{(z_1^T L(x-\tilde{x}))^2 + (z_2^T L(x-\tilde{x}))^2 + (z_3^T L(x-\tilde{x}))^2}}{E_k \|L\tilde{x}\|}.$$

Then

$$Pr\left(\frac{1}{w}\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|} \le K_2 \le w\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|}\right) \approx 1 - \frac{\pi}{4w^2},$$

$$Pr\left(\frac{1}{w}\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|} \le K_3 \le w\frac{\|L(x-\tilde{x})\|}{\|L\tilde{x}\|}\right) \approx 1 - \frac{32}{3\pi^2 w^3}.$$

For a condition estimate, we usually require the magnitude of the estimate to be within a ratio of 10. Letting $w = 10$, the probability of an acceptable estimate for K_1 is 93.6%, while for K_2 it is 99.2% and for K_3 it is 99.9%.

Let λ_i solve

$$A^T\lambda_i = L^T z_i.$$

Defining

$$v_i = |\lambda_i|^T(|A||\tilde{x}| + |b|),$$

we obtain

$$e_2 = \frac{E_2\sqrt{(v_1^2 + v_2^2)}}{E_k\|L\tilde{x}\|} \tag{3.5}$$

and

$$e_3 = \frac{E_3\sqrt{(v_1^2 + v_2^2 + v_3^2)}}{E_k\|L\tilde{x}\|}. \tag{3.6}$$

Thus $K_2 \le e_2\epsilon$, $K_3 \le e_3\epsilon$. e_1, e_2, and e_3 are the corresponding condition estimates.

This method is especially useful for obtaining a subspace condition estimate. Let L be a projection from R^n to R^k. The above method gives a relative error estimate for the subspace of the solution under the projection.

To summarize, the algorithm for the subspace error estimate is given as follows. We suggest using three random vectors for the estimate.

Subspace Error Estimate Algorithm. *Suppose we have an LU or QR decomposition of A and the numerical solution $\tilde{x}$. The condition number is estimated as follows:*

Step 1. *Determine the subspace or the components for which one wants to estimate the error. Let k be the dimension of the subspace and L be the projection from R^n to the subspace.*

Step 2. *Randomly choose three orthogonal vectors z_1, z_2, z_3 from the unit sphere S_{k-1}. Solve* (3.4) *for the corresponding λ_1, λ_2, λ_3.*

Step 3. *Compute*

$$v_i = |\lambda_i^T|(|A||\tilde{x}| + |b|).$$

Then the subspace condition estimate is given by

$$e_3 = \frac{E_3\sqrt{(v_1^2 + v_2^2 + v_3^2)}}{E_k\|L\tilde{x}\|}, \tag{3.7}$$

and the subspace relative error estimate is given by $e_3\epsilon$.

3.3. Examples. Here we demonstrate how the proposed method resolves the problems in Examples 1–3 of section 1.

Example 4.

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \delta \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

The solution is $\tilde{x} = (b_1, b_2/\delta)$. Recall that, for this example, the solution has high relative accuracy for any right-hand side, assuming a relative perturbation. To compute the error estimate, let the random vector be z, where $\|z\|_2 = 1$. Solving the adjoint equation (3.4) yields $\lambda = (z_1, z_2/\delta)^T$. Then the relative error is estimated using e_1 by

$$\frac{2(|b_1||z_1| + |b_2||z_2|/|\delta|)}{E_2\sqrt{b_1^2 + b_2^2/\delta^2}}\epsilon \leq \frac{2\epsilon}{E_2}.$$

Regardless of the random vector chosen, this method always yields a small condition number. Of course, e_2 will yield the exact condition number since the problem has just two dimensions and we choose orthogonal random vectors.

Example 5.

$$A = \begin{bmatrix} 1 & 1+\delta \\ 1-\delta & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1+\delta+\delta^2 \\ 1 \end{bmatrix}.$$

Suppose our goal is an accurate x_2. Then we let $g(x) = x_2$. Since $g(x)$ is a scalar function, we do not need a random vector here. Solving the adjoint equation (3.1), we have $\lambda = \frac{1}{\delta^2}[-(1-\delta), 1]^T \approx \frac{1}{\delta^2}[-1, 1]^T$. The relative error in x_2 is estimated by

$$\frac{|\lambda^T|(|A||\tilde{x}| + |b|)\epsilon}{|\tilde{x}_2|} \approx \frac{4}{|\delta|^3}\epsilon.$$

With a good numerical method like GE with partial pivoting (GEPP) or QR, ϵ is just a multiple of the relative machine precision ϵ_{mach} for this two-dimensional problem. For Matlab we get $\epsilon_{mach} \approx 10^{-16}$. We can see from our estimate that when $\epsilon = 10^{-5}$, the solution for x_2 will have a relative error of 0.1 (the computed result yields an error of 0.112). When $\epsilon = 10^{-4}$, the estimate predicts four digits of accuracy in x_2. Thus, the estimate accurately predicts the results obtained by Matlab (described in section 1), while the standard condition number underestimates the error.

Example 6.

$$A = \begin{bmatrix} 1 & 0 & -h \\ 0 & 1 & -h \\ 1 & 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

For the subspace condition number, we choose a random vector $z = [r_1,\ r_2,\ 0]^T$ of norm 1. Solving the adjoint equation (3.4) yields $\lambda = [\frac{1}{2}(r_1 - r_2),\ \frac{1}{2}(r_2 - r_1),\ \frac{1}{2}(r_1 + r_2)]^T$. The condition is estimated using one random vector and (2.6) to be

$$\frac{|\lambda_1|(|x_1| + |hx_3| + |x_1 - hx_3|) + |\lambda_2|(|x_2| + |hx_3| + |x_2 - hx_3|) + |\lambda_3|(|x_1| + |x_2| + |x_1 + x_2|)}{E_3\sqrt{x_1^2 + x_2^2}}$$
$$\leq \frac{(|\lambda_1| + |\lambda_2| + |\lambda_3|)(|x_1| + |x_2| + |hx_3|)}{E_3\sqrt{x_1^2 + x_2^2}}.$$

Thus the condition estimated is $O(1)$, as we would expect from DAE theory [12, p. 144]) for the condition of the low-index subspace.

4. Numerical results. The numerical experiments were performed in Matlab on a Linux computer. We chopped the data for a round-off error of 10^{-8} to avoid any possibility that differences in the conclusions could be caused by different machine precisions.

We compare our error estimate with Skeel's condition estimate (1.6), the standard condition number, and the condition estimate provided by Matlab for randomly generated data. We first generate the random matrix A. Then a real x is generated randomly, and b is determined by $b = Ax$. We chop the data of A and b to get a relative error of 10^{-8}. Then we solve $A\tilde{x} = b$ for $\tilde{x}$. We compare the estimates and the actual relative error $\frac{\|x-\tilde{x}\|_2}{\|\tilde{x}\|_2}$. Skeel's condition number and the standard condition number have been computed accurately without approximation of $|A^{-1}|$. For the relative error of x, our definition reduces to Skeel's definition. But the statistical estimate used in our method is different from the estimate used in the suggested implementation of Skeel's method. Our estimate uses the small sample statistical method and the adjoint equation to estimate $\||A^{-1}|(|A\tilde{x}| + |b|)\|$ for the whole space, or $\||LA^{-1}|(|A\tilde{x}| + |b|)\|$ for some subspace, using several random orthogonal vectors, while the suggested implementation of Skeel's method approximates the matrix $|A^{-1}|$ directly [6, section 14.5]. The latter is much more complicated and expensive and is limited to matrices of a particular structure. For the three orthogonal random vectors on the unit sphere, we first generate three random vectors r_1, r_2, r_3 uniformly in $R^k([-1,1]) = \{x \in R^k | x_i \in [-1,1]\}$ and then make them orthogonal by setting

$$z_1 = \frac{r_1}{\|r_1\|}, \quad z_2 = \frac{r_2 - z_1^T r_2 z_1}{\|r_2 - z_1^T r_2 z_1\|}, \quad z_3 = \frac{r_3 - z_1^T r_3 z_1 - z_2^T r_3 z_2}{\|r_3 - z_1^T r_3 z_1 - z_2^T r_3 z_2\|}.$$

Note that although this is not exactly uniform on the unit sphere, it is cheaper to generate, and from our practice we feel it works quite well.

4.1. Scalar function g. Our first numerical test is to estimate the relative error for a scalar function. Here we let $g(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Since $g(x)$ is a scalar function, we can use the directional condition estimate. Other definitions do not provide a good estimate because they have not been designed to deal with this type of derived function. The corresponding results are shown in Figures 4.1 and 4.2 and Table 4.1. We show both the overestimate ratio $\frac{estimate}{real\ error}$ and the underestimate ratio $\frac{real\ error}{estimate}$. It can be seen that the standard condition definition and Skeel's definition result in a much greater overestimate than our method.

4.2. Vector error estimate. Next we compared the relative error $\frac{\|x-\tilde{x}\|}{\|\tilde{x}\|}$ with the estimates for 10,000 randomly generated dense matrices A and vectors x of dimension 100. The results are shown in Figures 4.3 and 4.4. The underestimates and overestimates for our method are displayed in Figure 4.3. Figure 4.4 shows the overestimate ratio for Skeel's definition, the standard condition number, and Matlab's estimator. Table 4.2 compares the mean and max value of those ratios for each method. From the results, we can see that there is a potential for a substantial overestimate for all the definitions and estimators. Our estimator is, with high probability, within a factor of 10 of the standard condition estimate, as shown in Figure 4.5. If we take an overestimate larger than 100 as a bad estimate, in 10,000 random tests, our method generates 142 bad estimates (1.42%), Skeel's condition number generates 195 bad estimates (1.95%), the standard condition number generates 405 bad estimates (4.05%), and the Matlab estimator generates 2124 bad estimates (21.24%).

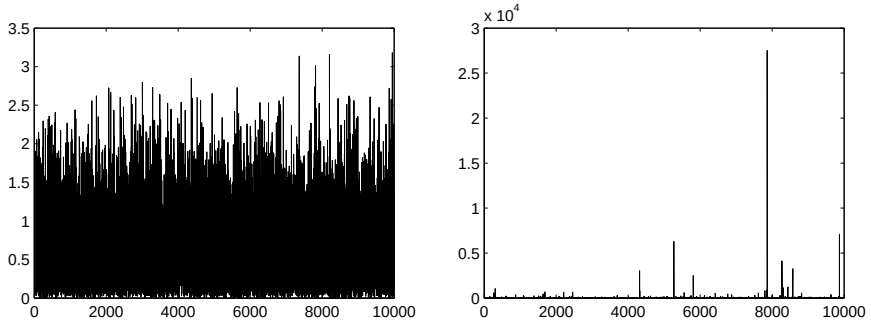


FIG. 4.1. *The plot on the left shows* $\frac{real\ error}{our\ estimate}$, *the amount by which our method underestimates the error of the mean function* $g(x)$, *for* 10,000 *randomly generated dense matrices* A *and vectors* x *of dimension* 100. *The plot on the right shows the amount of overestimate* $\frac{our\ estimate}{real\ error}$.

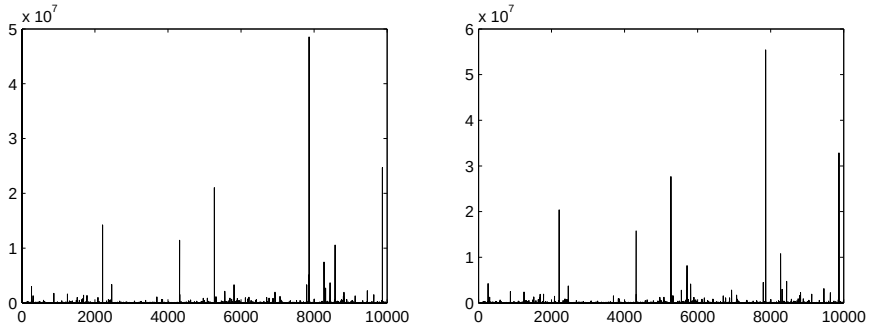


FIG. 4.2. *Overestimate of error of the mean function by Skeel's definition (left) and by the standard condition estimate (right) for* 10,000 *randomly generated dense matrices* A *and vectors* x *of dimension* 100. *(Note that, since Skeel's definition and the standard condition estimate are not designed for the computation of the condition of a scalar derived function, for these definitions we are using the estimate of the full vector.)*

TABLE 4.1

Comparison of ratios of overestimate and underestimate of error of the mean function using different condition estimates for dense matrices. For our method, the maximum of the overestimate and the underestimate is shown.

	Our method	Skeel	Standard	Matlab
MEAN	12.48	3.58×10^4	4.65×10^4	1.16×10^5
MAX	2.75×10^4	4.85×10^7	5.54×10^7	1.22×10^8

4.3. Ill-conditioned matrices. Another group of experiments was done for the (ill-conditioned) Hilbert matrix of dimension 10, where $a_{ij} = \frac{1}{i+j}$. The results are shown in Figures 4.6 and 4.7 for 10,000 randomly generated vectors x. Here we can see that all the methods can give a substantial overestimate to the actual error. Our method yields a result which is comparable to Skeel's estimate and to the standard condition estimate. For the number of overestimates by a factor of more than 100, in 10,000 random tests our method generated 259 (2.59%), Skeel's condition number generated 628 (6.28%), the standard condition number generated 4056 (40.56%), and Matlab's estimator generated 7394 (73.94%).

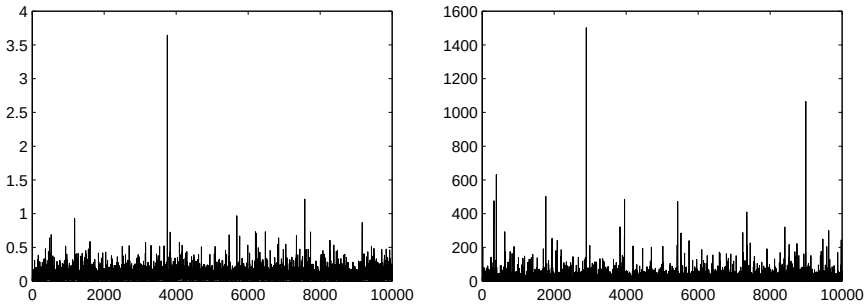


FIG. 4.3. *Underestimate of vector error* $\frac{real\ error}{our\ estimate}$ *(left) and overestimate of vector error* $\frac{our\ estimate}{real\ error}$ *(right) by our method for* 10,000 *randomly generated dense matrices* A *and vectors* x *of dimension* 100.

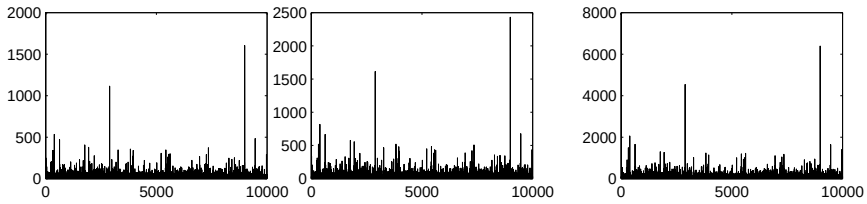


FIG. 4.4. *Overestimate of vector error by Skeel's condition estimate (left), by the standard condition estimate (middle), and by Matlab's condition estimate (right) for* 10,000 *randomly generated dense matrices* A *and vectors* x *of dimension* 100.

TABLE 4.2
Comparison of ratios of overestimate and underestimate of vector error using different condition estimates for dense matrices.

	Our method	Skeel	Standard	Matlab
MEAN	21	25	33	83
MAX	1500	1604	2432	6389

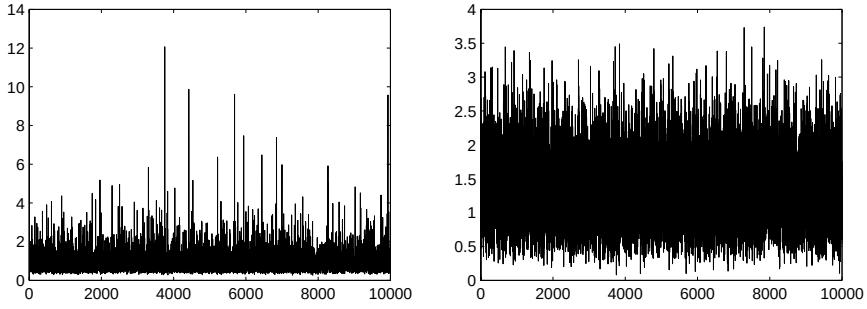


FIG. 4.5. *Underestimate of standard condition number* $\frac{standard\ condition\ number}{our\ estimate}$ *(left) and overestimate of standard condition number* $\frac{our\ estimate}{standard\ condition\ number}$ *(right) by our method for* 10,000 *randomly generated dense matrices* A *and vectors* x *of dimension* 100.

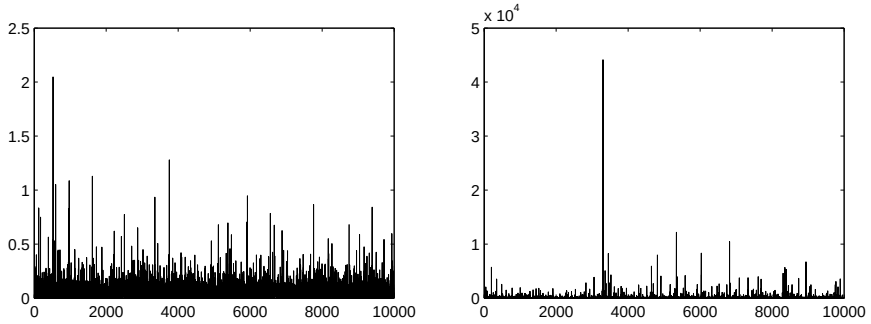


FIG. 4.6. *Underestimate (left) and overestimate (right) of the vector error by our method for the Hilbert matrix of dimension* 10 *with* 10,000 *randomly generated vectors* x.

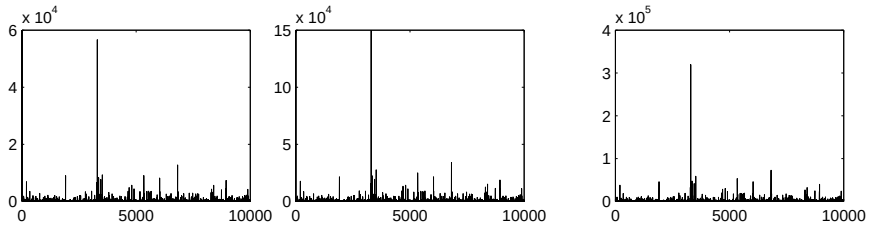


FIG. 4.7. *Overestimate of the vector error by Skeel's condition estimate (left), the standard condition estimate (middle), and Matlab's condition estimate (right) for the Hilbert matrix of dimension* 10 *with* 10,000 *randomly generated vectors* x.

TABLE 4.3
Comparison of condition numbers for Example 3 *in section* 1.

Stepsize h	Ours (full space)	Skeel's	Standard	Ours (subspace)
10^{-6}	2.04×10^6	1.41×10^6	1.5×10^6	3.30
10^{-8}	2.13×10^8	1.41×10^8	1.5×10^8	3.30
10^{-12}	2.40×10^{12}	1.41×10^{12}	1.5×10^{12}	3.30

4.4. DAE examples. We take Example 3 in section 1 as our first DAE example. We choose different stepsizes $h = 10^{-6}, 10^{-8}, 10^{-12}$ and random right-hand sides b. The corresponding condition numbers are listed in Table 4.3. With the stepsize decreasing, the condition number for the full solution space grows as $O(\frac{1}{h})$ for all these definitions. But for the subspace of only the first two components, the subspace condition number remains at 3.30. This indicates that this subspace is well-conditioned, although the system is ill-conditioned in the full solution space.

Another DAE example comes from an application in mechanics. It is of interest for the computation of the elliptic Fekete points [13]. The problem is of the form

$$M\frac{dy}{dt} = f(y(t)), \quad y(0) = y_0, \quad y'(0) = y_0', \tag{4.1}$$

with $y, f \in R^{2N}$ and $0 \le t \le t_{end}$. Here, $t_{end} = 1000$, $N = 20$, and M is the mass matrix given by

$$M = \begin{pmatrix} I_{6N} & 0 \\ 0 & 0 \end{pmatrix},$$

TABLE 4.4
Comparison of condition numbers for the Fekete problem.

Stepsize h	Ours (full space)	Skeel's	Standard	Ours (subspace)
10^{-6}	7.07×10^6	1.00×10^6	2.5×10^{11}	24.37
10^{-8}	5.93×10^8	1.00×10^8	3.44×10^{15}	24.37
10^{-12}	1.15×10^9	1.00×10^{12}	1.42×10^{27}	24.37

where I_{6N} is the identity matrix of dimension $6N$. The details of this problem can be found in [13] and also on the website http://hilbert.dm.uniba.it/∼testset/descrip.htm. Since we are concerned only with the linear system generated in the solution process, we extract the linear system for different stepsizes $h = 10^{-6}$, 10^{-8}, 10^{-12} and randomly generate the right-hand sides b. The subspace with the first 120 components is what we are concerned with here. The numerical results are shown in Table 4.4. The condition number of the full solution space grows when the stepsize decreases, while the condition number for the subspace remains the same at 24.37. This subspace condition number shows that the solution to the linear system can be computed safely for the first 120 components.

5. Conclusion. In this paper we proposed a new definition of condition number and a new method for error and condition estimation based on the adjoint equation and the small-sample statistical method. This new definition can produce a subspace error estimate, which is useful in some applications. For a vector measure of the error, the new definition, estimated as outlined by the small-sample statistical method, has low ($3n^2$) cost (assuming direct solution of dense linear systems where the matrix has already been factorized) and probability of 99.9% for the accuracy of the error estimate to be within a factor of 10. The method easily allows for the use of different derived functions (measures of the error) that may be relevant for different problems.

Acknowledgments. The authors would like to thank Charles Kenney, Shiv Chandrasekaran, and the referees for their insightful comments.

REFERENCES

[1] L. S. BLACKFORD, J. CHOI, A. CLEARY, E. D'AZEVEDO, J. DEMMEL, I. DHILLON, J. DONGARRA, S. HAMMARLING, G. HENRY, A. PETITET, K. STANLEY, D. WALKER, AND R. C. WHITNEY, *ScaLAPACK Users' Guide*, SIAM, Philadelphia, 1997.

[2] S. CHANDRASEKARAN AND I. C. F. IPSEN, *On the sensitivity of solution components in linear systems of equations*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 93–112.

[3] A. K. CLINE, C. B. MOLER, G. W. STEWART, AND J. H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.

[4] J. DEMMEL, B. DIAMENT, AND G. MALAJOVICH, *On the complexity of computing error bounds*, Found. Comput. Math., 1 (2001), pp. 101–125.

[5] J. D. DIXON, *Estimating extremal eigenvalues and condition numbers of matrices*, SIAM J. Numer. Anal., 20 (1983), pp. 812–814.

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

[7] W. KAHAN, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.

[8] C. S. KENNEY AND A. J. LAUB, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput., 15 (1994), pp. 36–61.

[9] C. S. KENNEY, A. J. LAUB, AND M. S. REESE, *Statistical condition estimation for linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 566–583.

[10] J. KUCZYŃSKI AND H. WOŹNIAKOWSKI, *Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1094–1122.

[11] J. Kuczyński and H. Woźniakowski, *Probabilistic bounds on the extremal eigenvalues and condition number by the Lanczos algorithm*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 672–691.
[12] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Applied Mathematics 14, SIAM, Philadelphia, 1996.
[13] P. M. Pardalos, *An open global optimization problem on the unit sphere*, J. Global Optim., 6 (1995), p. 213.
[14] J. R. Rice, *A theory of condition*, SIAM J. Numer Anal., 3 (1966), pp. 187–310.
[15] R. D. Skeel, *Scaling for numerical stability in Gaussian elimination*, J. ACM, 26 (1979), pp. 494–526.
[16] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 802–808

LEAST SQUARES SOLUTION OF MATRIX EQUATION $AXB^* + CYD^* = E^*$

SANG-YEUN SHIM† AND YU CHEN†

Abstract. We present an efficient algorithm for the least squares solution (X, Y) of the matrix equation $AXB^* + CYD^* = E$ with arbitrary coefficient matrices A, B, C, D and the right-hand side E. This method determines the least squares solution (X, Y) with the least norm. It relies on the SVD and generalized SVD of the coefficient matrices and has complexity proportional to the cost of these SVDs.

Key words. least norm solution, matrix equation, singular value decomposition

AMS subject classifications. 15A24, 65F20, 65F22, 65K10

PII. S0895479802401059

1. Introduction. Let m, m_1, m_2; n, n_1, n_2 be six positive integers, and let $E \in \mathbb{C}^{m\times n}$, $A \in \mathbb{C}^{m\times m_1}$, $B \in \mathbb{C}^{n\times n_1}$, $C \in \mathbb{C}^{m\times m_2}$, and $D \in \mathbb{C}^{n\times n_2}$. We consider the linear matrix equation

$$AXB^* + CYD^* = E \tag{1.1}$$

for $X \in \mathbb{C}^{m_1\times n_1}$ and $Y \in \mathbb{C}^{m_2\times n_2}$. The least squares solution of (1.1) with the least norm is essential to the inverse scattering problem for the Helmholtz equation, where E is the scattering matrix for a domain D partitioned into two nonoverlapping subdomains D_1 and D_2, and X and Y are the scattering matrices for the two subdomains. The determination of the two scattering matrices (X, Y) from the parent scattering matrix E is known as matrix splitting, and the least norm solution is crucial to the stability of splitting.

In terms of generalized inverse, generalized SVD, and canonical correlation decomposition (CCD), respectively, solution formulae for (1.1) are established in [3], [4], and [5], provided that (1.1) is consistent. Least squares solutions are also given in [5] via CCD if (1.1) is not consistent. It appear that there is no method that determines the least squares solution with the least norm at a cost proportional to that for the SVDs of the coefficient matrices A, B, C, D.

In this paper, we develop such an efficient method for the least squares solution of (1.1) with the least norm. In section 2, we will start with the normal equation of (1.1) and construct least squares solutions to (1.1). Our approach differs from [5]; it requires only SVDs of the coefficient matrices A, B, C, D. The resulting formula for the least squares solutions also differs from that of [5], and it enables us to construct the least norm solution in section 3.

As is well known, the use of the normal equation leads to the squaring of the condition number. This does not seem to cause any practical problem to our intended application where the linear equation (1.1) originates from an inverse scattering problem and thus has a high condition number; see [7, sections 4.5 and 5]. It appears that the squaring of a high condition number does not have adverse effects on the regularization.

*Received by the editors January 15, 2002; accepted for publication (in revised form) by L. Eldén August 14, 2002; published electronically January 31, 2003.
http://www.siam.org/journals/simax/24-3/40105.html
†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 (syshim@cims.nyu.edu, yuchen@cims.nyu.edu).

2. Least squares solutions. The pair (X, Y) is referred to as the least squares solution of (1.1) if it minimizes the Frobenius norm of the residual

$$\|AXB^* + CYD^* - E\|_F^2. \tag{2.1}$$

To construct the least squares solution of (1.1), we first consider its normal equation in section 2.1. We then reduce the normal equation to the two equations (2.10) and (2.11), which are always consistent and are equivalent to the normal equation. Finally, we solve (2.10) in section 2.2 and (2.11) in section 2.3.

2.1. The normal equation. In this section we will reformulate the least squares problem for the linear equation (1.1) as the solution of its normal equation. We will require the following two lemmas on the normal equation.

LEMMA 2.1. *The normal equation of the linear equation* (1.1) *is*

$$\begin{aligned} A^*AXB^*B + A^*CYD^*B &= A^*EB, \\ C^*AXB^*D + C^*CYD^*D &= C^*ED, \end{aligned} \tag{2.2}$$

and it is always consistent.

Proof. Let $\mathcal{L}$ be a linear mapping $C^{m_1\times n_1} \times C^{m_2\times n_2}$ to $C^{m\times n}$, given by

$$\mathcal{L}(X, Y) = AXB^* + CYD^*, \tag{2.3}$$

so that (2.1) can be written as

$$\mathcal{L}(X, Y) = E. \tag{2.4}$$

Then, the conjugate linear map $\mathcal{L}^*$, mapping $C^{m\times n}$ to $C^{m_1\times n_1} \times C^{m_2\times n_2}$, is of the form (see section 7 of [4])

$$\mathcal{L}^*(P) = (A^*PB, C^*PD). \tag{2.5}$$

We get the normal equation (2.2) by applying the conjugate linear map $\mathcal{L}^*$ on both sides of (2.4), namely,

$$\mathcal{L}^*(\mathcal{L}(X, Y)) = \mathcal{L}^*(E). \tag{2.6}$$

Note that the normal equation is always consistent (see [6, p. 223]). Thus (2.2), which is the normal equation of the linear equation (1.1), is consistent.

LEMMA 2.2. *The pair* (X, Y) *is a least squares solution of the linear equation* (1.1) *if and only if it is a solution of the normal equation* (2.2).

Proof. See [6, p. 220].

Therefore, the remainder of this section is devoted to the solution of the normal equation (2.2). Two steps are required to simplify (2.2).

Step 1. Take the reduced SVDs of the coefficient matrices A, B, C, D,

$$A = U_A D_A V_A^*, \quad B = U_B D_B V_B^*, \quad C = U_C D_C V_C^*, \quad D = U_D D_D V_D^*, \tag{2.7}$$

where D_A, D_B, D_C, D_D are square, diagonal matrices with full rank. Substituting (2.7) into (2.2), we obtain a system of equations, which is equivalent to (2.2),

$$\begin{aligned} D_A V_A^* X V_B D_B + (U_A^* U_C) D_C V_C^* Y V_D D_D (U_D^* U_B) &= U_A^* E U_B, \\ (U_C^* U_A) D_A V_A^* X V_B D_B (U_B^* U_D) + D_C V_C^* Y V_D D_D &= U_C^* E U_D. \end{aligned} \tag{2.8}$$

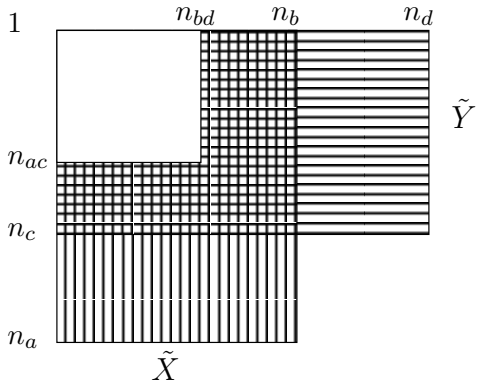


FIG. 2.1. *Overlaying the matrices $\tilde{X}$ and $\tilde{Y}$; $\tilde{X}$ is n_a-by-n_b, $\tilde{Y}$ is n_c-by-n_d.*

REMARK 2.3. *The singular values of $U_A^*U_C$ and $U_B^*U_D$ are bounded by* 1 *because U_A, U_B, U_C, U_D all have orthonormal columns.*

Step 2. Take the full SVD of the matrices $U_A^*U_C$ and $U_B^*U_D$ in (2.8),

$$U_A^*U_C = U_{AC}D_{AC}V_{AC}^*, \quad U_B^*U_D = U_{BD}D_{BD}V_{BD}^*. \tag{2.9}$$

We rewrite (2.8) as

$$\begin{aligned} \tilde{X} + D_{AC}\tilde{Y}D_{BD}^* &= U_{AC}^*U_A^*EU_BU_{BD}, \\ D_{AC}^*\tilde{X}D_{BD} + \tilde{Y} &= V_{AC}^*U_C^*EU_DV_{BD} \end{aligned} \tag{2.10}$$

with new variables

$$\tilde{X} = U_{AC}^*D_AV_A^*XV_BD_BU_{BD}, \quad \tilde{Y} = V_{AC}^*D_CV_C^*YV_DD_DV_{BD}. \tag{2.11}$$

REMARK 2.4. *The linear equations* (2.10) *and* (2.11) *for (X, Y) are equivalent to* (2.8) *because the procedures leading to* (2.10) *and* (2.11) *are reversible. Therefore, it remains that we solve* (2.10) *for $(\tilde{X}, \tilde{Y})$ and then* (2.11) *for (X, Y) in order to construct the least squares solutions of* (1.1).

2.2. Solution of (2.10) for $(\tilde{X}, \tilde{Y})$. The coefficient matrices of (2.10) are all diagonal (they may not be square), and therefore (2.10) is decoupled into 1-by-2, 2-by-2, and 1-by-1 scalar equations.

Let $n_a = \text{rank}(A)$, $n_b = \text{rank}(B)$, $n_c = \text{rank}(C)$, $n_d = \text{rank}(D)$, let n_{ac} be the number of unit singular values in D_{AC}, and let n_{bd} be the number of unit singular values in D_{BD}. Note that matrix $\tilde{X}$ and the first equation in (2.10) both have dimensions n_a-by-n_b; matrix $\tilde{Y}$ and the second equation in (2.10) both have dimensions n_c-by-n_d. Depending on how the two equations in (2.10) overlay (see, for example, Figure 2.1 for a possible configuration), we group the decoupled equations into four cases.

Case 1. The rectangular domain of entries (i, j) of dimensions n_{ac}-by-n_{bd} inside the overlapping area of $\tilde{X}$ and $\tilde{Y}$; see the unshaded area in Figure 2.1. In this area, the (i, j)th entry of the matrices $\tilde{X}$, $\tilde{Y}$ are multiplied by the unit singular values $(D_{AC})_{ii}$ and $(D_{BD})_{jj}$, and the two equations in (2.10) are identical:

$$\tilde{X}_{ij} + \tilde{Y}_{ij} = (U_{AC}^*U_A^*EU_BU_{BD})_{ij} \tag{2.12}$$

for $1 \le i \le n_{ac}$, $1 \le j \le n_{bd}$.

Case 2. The overlapping area of $\tilde{X}$ and $\tilde{Y}$ that is doubly shaded in Figure 2.1, where $i \le \min(n_a, n_c)$ and $j \le \min(n_b, n_d)$ and $\{n_{ac} < i \text{ or } n_{bd} < j\}$. In this area, at least one of the two singular values $(D_{AC})_{ii}, (D_{BD})_{jj}$ is less than 1 (see Remark 2.3 and Case 1), and the (i,j)th entries of $\tilde{X}$, $\tilde{Y}$ are uniquely determined by the pair of equations

$$\begin{aligned} \tilde{X}_{ij} + (D_{AC})_{ii}(D_{BD})_{jj}\tilde{Y}_{ij} &= (U_{AC}^* U_A^* E U_B U_{BD})_{ij}, \\ (D_{AC})_{ii}(D_{BD})_{jj}\tilde{X}_{ij} + \tilde{Y}_{ij} &= (V_{AC}^* U_C^* E U_D V_{BD})_{ij}. \end{aligned} \tag{2.13}$$

Case 3. The singly shaded area of $\tilde{X}$, if it exists at all, where $\tilde{X}_{ij}$ is given by

$$\tilde{X}_{ij} = (U_{AC}^* U_A^* E U_B U_{BD})_{ij} \tag{2.14}$$

for $\{\ n_c < i \le n_a,\ 1 \le j \le n_b\ \}$ or $\{\ 1 \le i \le n_a,\ n_d < j \le n_b\ \}$.

Case 4. The singly shaded area of $\tilde{Y}$, if it exists at all, where $(\tilde{Y})_{ij}$ is given by

$$\tilde{Y}_{ij} = (V_{AC}^* U_C^* E U_D V_{BD})_{ij} \tag{2.15}$$

for $\{\ n_a < i \le n_c,\ 1 \le j \le n_b\ \}$ or $\{\ 1 \le i \le n_a,\ n_b < j \le n_d\ \}$.

Evidently, matrices $\tilde{X}$, $\tilde{Y}$ each can be uniquely partitioned into 2-by-2 blocks

$$\tilde{X} = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}, \tag{2.16}$$

where the matrices Φ_{11}, Ψ_{11} are dimensioned n_{ac}-by-n_{bd} and are solutions to (2.12) in Case 1. The general solutions to (2.12) are of the form

$$\Phi_{11} = R, \quad \Psi_{11} = [U_{AC}^* U_A^* E U_B U_{BD}]\,(1{:}n_{ac},\ 1{:}n_{bd}) - R, \tag{2.17}$$

where R is an arbitrary n_{ac}-by-n_{bd} matrix. We choose a special solution to be

$$\hat{\Phi}_{11} = 0, \quad \hat{\Psi}_{11} = [U_{AC}^* U_A^* E U_B U_{BD}]\,(1{:}n_{ac},\ 1{:}n_{bd}). \tag{2.18}$$

The remaining six blocks in (2.16) appear only in (2.13)–(2.15) and are uniquely determined. Therefore,

$$\tilde{X}_s = \begin{pmatrix} \hat{\Phi}_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}, \quad \tilde{Y}_s = \begin{pmatrix} \hat{\Psi}_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix} \tag{2.19}$$

is a special solution of (2.10), and

$$\tilde{X} = \begin{pmatrix} \hat{\Phi}_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} + \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} \hat{\Psi}_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix} - \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix} \tag{2.20}$$

is the general solution.

2.3. Solution of (2.11) for (X, Y). With $(\tilde{X}, \tilde{Y})$ obtained in section 2.2, we solve (2.11) for (X, Y). Since $U_{AC}, U_{BD}, V_{AC}, V_{BD}$ are unitary and D_A, D_B, D_C, D_D are invertible, (2.11) can be rewritten as

$$\begin{aligned} V_A^* X V_B &= D_A^{-1} U_{AC} \tilde{X} U_{BD}^* D_B^{-1}, \\ V_C^* Y V_D &= D_C^{-1} V_{AC} \tilde{Y} V_{BD}^* D_D^{-1}. \end{aligned} \tag{2.21}$$

The following lemma from [3] is directly useful for the solution of (2.21).

LEMMA 2.5. *Let A^+ and B^+ be pseudoinverses of A and B. The linear equation*

$$AZB = C \tag{2.22}$$

for matrix Z is consistent if and only if

$$AA^+CB^+B = C. \tag{2.23}$$

Furthermore, if (2.22) *is consistent, its general solution is given by*

$$Z = A^+CB^+ + U - A^+AUBB^+ \tag{2.24}$$

with U an arbitrary matrix. Finally,

$$\|Z\|_F^2 = \|A^+CB^+\|_F^2 + \|U - A^+AUBB^+\|_F^2. \tag{2.25}$$

To apply Lemma 2.5 for the solution of (2.21), we note that

$$(V_A^*)^+ = V_A, \quad (V_B)^+ = V_B^*, \quad (V_C^*)^+ = V_C, \quad (V_D)^+ = V_D^*$$

and that (2.21) is trivially consistent. It follows immediately from (2.24) that the solutions of (2.21) are

$$\begin{aligned} X &= V_A D_A^{-1} U_{AC} \tilde{X} U_{BD}^* D_B^{-1} V_B^* + R_X - V_A V_A^* R_X V_B V_B^*, \\ Y &= V_C D_C^{-1} V_{AC} \tilde{Y} V_{BD}^* D_D^{-1} V_D^* + R_Y - V_C V_C^* R_Y V_D V_D^*, \end{aligned} \tag{2.26}$$

where arbitrary matrices R_X and R_Y are n_a-by-n_b and n_c-by-n_d, respectively.

THEOREM 2.6. *Let $(\tilde{X}_s, \tilde{Y}_s)$ be the special solution* (2.19) *to* (2.10). *Furthermore, let*

$$C_1 = -D_A^{-1} U_{AC}\, \tilde{X}_s\, U_{BD}^* D_B^{-1}, \quad C_2 = D_C^{-1} V_{AC}\, \tilde{Y}_s\, V_{BD} D_D^{-1}. \tag{2.27}$$

Finally, let

$$\begin{aligned} \tilde{U}_{AC} &= U_{AC}(:, 1{:}n_{ac}), \quad \tilde{U}_{BD} = U_{BD}(:, 1{:}n_{bd}), \\ \tilde{V}_{AC} &= V_{AC}(:, 1{:}n_{ac}), \quad \tilde{V}_{BD} = V_{BD}(:, 1{:}n_{bd}). \end{aligned}$$

Then the least squares solutions of (1.1) *are given by the formula*

$$\begin{aligned} X &= V_A(D_A^{-1} \tilde{U}_{AC}\, R\, \tilde{U}_{BD}^* D_B^{-1} - C_1) V_B^* + R_X - V_A V_A^* R_X V_B V_B^*, \\ Y &= V_C(C_2 - D_C^{-1} \tilde{V}_{AC}\, R\, \tilde{V}_{BD}^* D_D^{-1}) V_D^* + R_Y - V_C V_C^* R_Y V_D V_D^*, \end{aligned} \tag{2.28}$$

where arbitrary matrices R_X, R_Y, and R are n_a-by-n_b, n_c-by-n_d, and n_{ac}-by-n_{bd}, respectively.

3. The least norm solution. Denote by $\mathcal{C}$ the set of least squares solutions of (1.1); see Theorem 2.6. A pair $(X, Y) \in \mathcal{C}$ is referred to as a least norm solution if it minimizes

$$\|X\|_F^2 + \|Y\|_F^2 \tag{3.1}$$

over $\mathcal{C}$. Since the Frobenius norm of a matrix is the standard 2-norm of the vector formed by columns of the matrix, there is a unique least norm solution to (1.1). In

this section, we construct the least norm solution by minimizing (3.1) over the three arbitrary matrices R_X, R_Y, and R in (2.28).

Step 1. Eliminate R_X and R_Y. It follow from (2.25) that

$$\begin{aligned} \|X\|_F^2 &= \|V_A Z_X V_B^*\|_F^2 + \|R_X - V_A V_A^* R_X V_B V_B^*\|_F^2, \\ \|Y\|_F^2 &= \|V_C Z_Y V_D^*\|_F^2 + \|R_Y - V_C V_C^* R_Y V_D V_D^*\|_F^2, \end{aligned} \tag{3.2}$$

where

$$Z_X = D_A^{-1}\tilde{U}_{AC}\, R\, \tilde{U}_{BD}^* D_B^{-1} - C_1, \quad Z_Y = C_2 - D_C^{-1}\tilde{V}_{AC}\, R\, \tilde{V}_{BD}^* D_D^{-1}. \tag{3.3}$$

It is evident from (3.2) that the least norm solution (X, Y) requires

$$\|R_X - V_A V_A^* R_X V_B V_B^*\|_F^2 = 0, \quad \|R_Y - V_C V_C^* R_Y V_D V_D^*\|_F^2 = 0, \tag{3.4}$$

which is attainable by setting

$$R_X = 0, \quad R_Y = 0. \tag{3.5}$$

Step 2. Minimize (3.1) over matrix R in Z_X, Z_Y. Combining (3.2) and (3.5), and observing that V_A, V_B, V_C, V_D are unitary, we obtain

$$\begin{aligned} \min_{R,R_X,R_Y} \Big(\|X\|_F^2 + \|Y\|_F^2 \Big) &= \min_R \Big(\|V_A Z_X V_B^*\|_F^2 + \|V_C Z_Y V_D^*\|_F^2 \Big) \\ &= \min_R \Big(\|Z_X\|_F^2 + \|Z_Y\|_F^2 \Big); \end{aligned} \tag{3.6}$$

therefore, it remains to minimize

$$\|D_A^{-1}\tilde{U}_{AC}\, R\, \tilde{U}_{BD}^* D_B^{-1} - C_1\|_F^2 + \|D_C^{-1}\tilde{V}_{AC}\, R\, \tilde{V}_{BD}^* D_D^{-1} - C_2\|_F^2 \tag{3.7}$$

over arbitrary R. This is possible via generalized singular value decomposition (GSVD); we use the version given in [1, p. 466]. Following [2], we take GSVDs of the pair $D_A^{-1}\tilde{U}_{AC},\ D_C^{-1}\tilde{V}_{AC}$,

$$D_A^{-1}\tilde{U}_{AC} = U_1 D_1 X_{AC}, \quad D_C^{-1}\tilde{V}_{AC} = U_3 D_3 X_{AC}, \tag{3.8}$$

and of the pair $D_B^{-1}\tilde{U}_{BD},\ D_D^{-1}\tilde{V}_{BD}$,

$$D_B^{-1}\tilde{U}_{BD} = U_2 D_2 X_{BD}, \quad D_D^{-1}\tilde{V}_{BD} = U_4 D_4 X_{BD}, \tag{3.9}$$

where X_{AC}, X_{BD} are nonsingular, U_i is orthonormal, D_i is real and diagonal, $1 \le i \le 4$.

REMARK 3.1. *With* (3.5) *and the GSVDs* (3.8), (3.9), *we may update* (2.28)*:*

$$\begin{aligned} X &= V_A(U_1 D_1 (X_{AC}\, R\, X_{BD}^*) D_2 U_2^* - C_1) V_B^*, \\ Y &= -V_C(U_3 D_3 (X_{AC}\, R\, X_{BD}^*) D_4 U_4^* - C_2) V_D^*. \end{aligned} \tag{3.10}$$

Substituting (3.8), (3.9) into (3.7), we have

$$\begin{aligned} \|X\|_F^2 + \|Y\|_F^2 &= \|U_1 D_1 (X_{AC}\, R\, X_{BD}^*) D_2 U_2^* - C_1\|_F^2 \\ &\quad + \|U_3 D_3 (X_{AC}\, R\, X_{BD}^*) D_4 U_4^* - C_2\|_F^2 \\ &= \|D_1 (X_{AC}\, R\, X_{BD}^*) D_2 - U_1^* C_1 U_2\|_F^2 \\ &\quad + \|C_1 - U_1 U_1^* C_1 U_2 U_2^*\|_F^2 \\ &\quad + \|D_3 (X_{AC}\, R\, X_{BD}^*) D_4 - U_3^* C_2 U_4\|_F^2 \\ &\quad + \|C_2 - U_3 U_3^* C_2 U_4 U_4^*\|_F^2. \end{aligned} \tag{3.11}$$

LEMMA 3.2. *Let D_1, D_3 be k-by-k real diagonal matrices, and let D_2, D_4 be ℓ-by-ℓ real diagonal matrices. Furthermore, let G, H be k-by-ℓ matrices. Finally, let P be a k-by-ℓ matrix defined by*

$$P_{ij} = \begin{cases} 0 & \text{if } (D_1)_{ii}^2(D_2)_{jj}^2 + (D_3)_{ii}^2(D_4)_{jj}^2 = 0, \\ \left[(D_1)_{ii}^2(D_2)_{jj}^2 + (D_3)_{ii}^2(D_4)_{jj}^2\right]^{-1} & \text{otherwise.} \end{cases} \tag{3.12}$$

Then the minimization

$$\min_W \|D_1 W D_2 - G\|_F^2 + \|D_3 W D_4 - H\|_F^2 \tag{3.13}$$

has a solution

$$W = P \circ (D_1 G D_2 + D_3 H D_4), \tag{3.14}$$

where $\circ$ is the entrywise (or Hadamard) matrix multiplication so that $(B \circ C)_{ij} = B_{ij}C_{ij}$.

A proof of the lemma can be found in [5, p. 96]. It follows immediately that (3.11) is minimized if the product $T = X_{AC}\, R\, X_{BD}^*$ in (3.11) is chosen

$$T = X_{AC}\, R\, X_{BD}^* = P \circ (D_1 U_1^* C_1 U_2 D_2 + D_3 U_3^* C_2 U_4 D_4), \tag{3.15}$$

where P is defined by (3.12) with $k = n_{ac}$, $\ell = n_{bd}$. Our main result follows immediately from (3.15) and (3.10).

THEOREM 3.3. *The least squares solution of the matrix equation* (1.1)*, which minimizes the residual and has the least Frobenius norm, is*

$$\begin{aligned} X &= V_A(U_1 D_1 T D_2 U_2^* - C_1)V_B^*, \\ Y &= -V_C(U_3 D_3 T D_4 U_4^* - C_2)V_D^*, \end{aligned} \tag{3.16}$$

where T is given in (3.15)*, and C_1, C_2 in* (2.27)*.*

To summarize, we have presented an efficient procedure for the least squares solution of the matrix equation $AXB^* + CYD^* = E$ with arbitrary matrices A, B, C, D, and E. The algorithm uses the SVD and GSVD on the coefficient matrices and determines the least squares solution with the least norm at a cost proportional to that for the SVDs of the coefficient matrices. If all the matrices in the equation are n-by-n, our method constructs the least squares solution (X, Y) in $O(n^3)$ flops.

REFERENCES

[1] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1997.

[2] C. F. VAN LOAN, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100.

[3] J. K. BAKSALARY AND R. KALA, *The matrix equation $AXB+CYD = E$*, Linear Algebra Appl., 30 (1980), pp. 141–147.

[4] K. E. CHU, *Singular value and generalized singular value decompositions and the solution of linear matrix equations*, Linear Algebra Appl., 87 (1987), pp. 83–98.

[5] G. XU, M. WEI, AND D. ZHENG, *On solutions of matrix equation $AXB + CYD = F$*, Linear Algebra Appl., 279 (1998), pp. 93–109.

[6] P. H. GILL, W. MURRAY, AND M. H. WRIGHT, *Numerical Linear Algebra and Optimization*, Addison–Wesley, Redwood City, CA, 1991.

[7] Y. CHEN, *Inverse scattering via Heisenberg's uncertainty principle*, Inverse Problems, 13 (1997), pp. 253–282.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 809–821

SHIFTED FOURIER MATRICES AND THEIR TRIDIAGONAL COMMUTORS*

STUART CLARY† AND DALE H. MUGLER†

Abstract. It is known that, for $n \geq 3$, the $n \times n$ Fourier matrix $F = n^{-1/2}[e^{2\pi i \mu\nu/n}]$ $(0 \leq \mu < n,$ $0 \leq \nu < n)$ commutes with a nonscalar tridiagonal matrix T and also with another matrix X that is "almost" tridiagonal. These matrices are important in selecting eigenvectors for the Fourier matrix itself. The purpose of this paper is to generalize those results to matrices $F_n(\tau, \alpha)$, variants of the Fourier matrix depending on a base-choosing parameter τ and a shift parameter α. These *shifted Fourier matrices* are defined by $F_n(\tau, \alpha) = n^{-1/2}[e^{2\pi i \tau(\mu - m + a)(\nu - m + a)}]$, where $m = (n-1)/2$ and $a = \alpha/\tau$. We show that $F_n(\tau, \alpha)$ commutes with a nonscalar tridiagonal matrix $T_n(\tau, \alpha)$ for all values of the shift parameter α, as long as the base $q = e^{2\pi i \tau}$ determined by τ is an nth root of unity, and also for all values of τ in the "centered" case corresponding to $\alpha = 0$. Furthermore, we show that, in certain more specialized cases, $F_n(\tau, \alpha)$ also commutes with a matrix $X_n(\tau, \alpha)$ that has ± 1 in the upper-right and lower-left corners and is otherwise tridiagonal. In most cases, $T_n(\tau, \alpha)$ and $X_n(\tau, \alpha)$ are essentially the only matrices of their band-structure that commute with $F_n(\tau, \alpha)$.

Key words. shifted Fourier matrix, generalized Fourier matrix, centered Fourier matrix, tridiagonal commutor, extended-tridiagonal matrix

AMS subject classifications. 15A27, 65T50

PII. S0895479800372754

1. Introduction.

Given a square matrix A, it is a rare and valuable phenomenon for there to exist a *tridiagonal* matrix B, not a scalar multiple of the identity matrix, that commutes with A. When this happens, we call B a nonscalar tridiagonal commutor of A, coining the word *commutor* as a convenient way to refer to any matrix that commutes with a given matrix. Grünbaum [8] has shown that, for $n \geq 3$, the $n \times n$ Fourier matrix, in its traditional form, has a nonscalar tridiagonal commutor. The purpose of this paper is to generalize that remarkable discovery.

There are two principal reasons why it is valuable to find a tridiagonal commutor B for a given matrix A, both reasons deriving from the theorem that commuting matrices have the same eigenvectors. (Some care is required when multiple eigenvalues are involved.) One reason is computational—eigenvectors can be computed much more efficiently for tridiagonal matrices than for matrices in general. The other reason is that if A has eigenvalues of high multiplicity and B does not, as happens with many of the most important Fourier matrices, then B can be used to standardize the selection of the eigenvectors of A, resolving the ambiguity caused by the degeneracy in the eigenvalues.

In this paper, we investigate the existence of tridiagonal commutors for a class of matrices that we call *shifted Fourier matrices*. These matrices are modified in two ways from the traditional form of Fourier matrix. Except for a scalar multiplier, the entries of the traditional $n \times n$ Fourier matrix are powers of the nth root of unity $q = e^{2\pi i/n}$ and are arranged so that the row and column containing the zeroth power of q intersect in the upper-left corner, which we call the *anchor point* in the matrix. The two modifications allowed in this paper are, first, to permit the base q to be any

*Received by the editors May 23, 2000; accepted for publication (in revised form) by D. Calvetti April 29, 2002; published electronically February 4, 2003.
http://www.siam.org/journals/simax/24-3/37275.html
†Department of Theoretical and Applied Mathematics, The University of Akron, Akron, OH 44325-4002 (clary@uakron.edu, dmugler@uakron.edu).

nonzero complex number and, second, to shift the location of the anchor point to any point on the main diagonal of the matrix. We measure these two modifications with parameters τ and α, respectively, and we denote the resulting shifted Fourier matrix by $F_n(\tau, \alpha)$. See section 2 for the details.

Perhaps the biggest surprise in this work is the prominent role played by the *centered Fourier matrices*, those for which the anchor point is at the center of the matrix. We have become convinced that centered Fourier matrices have better properties than traditional Fourier matrices in almost all respects, and, therefore, we break with tradition and define the shift parameter α so that $\alpha = 0$ corresponds to the centered case.

We also broaden our investigation beyond the truly tridiagonal commutors. This is motivated by Dickinson and Steiglitz's discovery [4] that the traditional Fourier matrix has an interesting commutor that is *almost* tridiagonal and whose entries are even simpler than those in Grünbaum's commutor. Upon observing that the identity matrix together with certain products of the Grünbaum and Dickinson–Steiglitz commutors can be used to produce a basis for the set of *all* matrices that commute with the traditional Fourier matrix, we considered it imperative to include generalizations of the Dickinson–Steiglitz matrix in our investigation.

The Dickinson–Steiglitz matrix has all of its nonzero entries confined to the three central diagonals and the two off-diagonal corners. Since the upper-right and lower-left corners can be regarded as cyclic extensions of the subdiagonal and the superdiagonal, we call such a matrix an *extended-tridiagonal* matrix.

It is possible for a shifted Fourier matrix to have no tridiagonal or extended-tridiagonal commutors other than scalar matrices. The principal result reported in this paper is that, for a shifted Fourier matrix $F_n(\tau, \alpha)$ of order $n \geq 3$, if the base-choosing parameter τ is chosen so that the base q is any nonreal nth root of unity or if the shift parameter α is chosen so that the anchor point is at the center of the matrix, then there definitely exists a nonscalar tridiagonal commutor (a generalization of the Grünbaum matrix), and sometimes there even exists a second commutor of extended-tridiagonal form (a generalization of the Dickinson–Steiglitz matrix). Although these commutors have many interesting properties, we defer a discussion of those properties to another occasion. In this paper, we concentrate on the existence and essential uniqueness of these commutors, including explicit formulas for them, with only brief mention of their properties.

We have dealt with some of these topics in earlier papers [11, 12], with an emphasis on the properties of the eigenvalues and eigenvectors in particular cases. We intend to devote a future paper to further questions regarding the eigenvalues and eigenvectors of the more general matrices discussed here.

The idea of shifting the anchor point of a Fourier matrix (while keeping q equal to $e^{\pm 2\pi i/n}$) appears in a limited way in Grünbaum's paper [8] and in various other papers [2, 3, 10, 15], and it has important connections with the discrete cosine transform and the discrete sine transform [10, 14, 17]. The idea of allowing q to be arbitrary, but without shifting the anchor point from the upper-left corner, was investigated by Bailey and Swarztrauber [1]. In their paper, "fractional Fourier transform" refers to the arbitrariness of q and is unrelated to the more common use of that term to mean a fractional power of a Fourier transform [12, 13].

A few interesting matrices other than Fourier matrices have been found to have nonscalar tridiagonal commutors. Grünbaum [6, 7] has classified all of the Toeplitz matrices that have nonscalar tridiagonal commutors; included as limiting cases are

matrices representing the Karhunen–Loève transform. The corresponding problem for Hankel matrices is, to the best of our knowledge, still unsolved, although Sawyer [16] and Grünbaum [9] have shown that the Hilbert matrix is a Hankel matrix having a nonscalar tridiagonal commutor, and we have discovered some others.

2. Shifted Fourier matrices. Let n be a positive integer, and let τ and α be complex constants. As abbreviations, let $m = (n-1)/2$ and $a = \alpha/\tau$ (if $\tau \neq 0$). We define two forms of the *shifted Fourier matrix* specified by these parameters. The first form (the periodic form) is the $n \times n$ matrix

$$\widetilde{F}_n(\tau,\alpha) = \frac{1}{\sqrt{n}}\Big[e^{2\pi i\tau(\mu-m)(\nu-m)}e^{2\pi i\alpha(\mu+\nu-2m)}\Big]_{\substack{0\le\mu<n\\0\le\nu<n}}. \tag{2.1}$$

The second form (the nonperiodic form) is

$$F_n(\tau,\alpha) = \frac{1}{\sqrt{n}}\Big[e^{2\pi i\tau(\mu-m+a)(\nu-m+a)}\Big]_{\substack{0\le\mu<n\\0\le\nu<n}}. \tag{2.2}$$

The second form is undefined if $\tau = 0$. Otherwise, we have

$$F_n(\tau,\alpha) = e^{2\pi i\alpha^2/\tau}\widetilde{F}_n(\tau,\alpha). \tag{2.3}$$

Of the two forms $\widetilde{F}_n(\tau,\alpha)$ and $F_n(\tau,\alpha)$, the latter explains why these matrices are called shifted Fourier matrices, since it shows an explicit shift of the entries of the matrix by the quantity $a = \alpha/\tau$, which counts the number of rows (and columns) that the anchor point moves from the center of the matrix. However, the form $\widetilde{F}_n(\tau,\alpha)$ is probably more fundamental because of the periodicity properties

$$\widetilde{F}_n(\tau+2,\alpha) = -\widetilde{F}_n(\tau,\alpha) \quad \text{if } n \text{ is even}, \tag{2.4}$$

$$\widetilde{F}_n(\tau+1,\alpha) = \widetilde{F}_n(\tau,\alpha) \quad \text{if } n \text{ is odd}, \tag{2.5}$$

and

$$\widetilde{F}_n(\tau,\alpha+1) = \widetilde{F}_n(\tau,\alpha). \tag{2.6}$$

(The wavy shape of the tilde in the notation $\widetilde{F}_n(\tau,\alpha)$ is intended as a reminder of periodicity.) For the purposes of this paper, in which our goal is to discuss the existence and uniqueness of tridiagonal and extended-tridiagonal commutors, it makes almost no difference which form we use, since (for $\tau \neq 0$) (2.3) shows that they have exactly the same commutors. We will usually favor $\widetilde{F}$ because the formulas are simpler and $\tau = 0$ is not an exceptional case. For future investigations involving the eigenvalues and eigenvectors of the shifted Fourier matrices and their commutors, it will be convenient to have both forms.

We call τ the *base-choosing parameter* and α the *shift parameter*. The quantities $e^{2\pi i\tau}$ and $e^{2\pi i\alpha}$ occur so often that we abbreviate them as $q = e^{2\pi i\tau}$ and $z = e^{2\pi i\alpha}$, with the understanding that a power q^s is always to be interpreted as an abbreviation for $e^{2\pi i\tau s}$. In most signal-processing work, $\tau = \pm 1/n$, corresponding to $q = e^{\pm 2\pi i/n}$.

It is sometimes useful to introduce yet another parameter, β, to parametrize a shift *perpendicular* to the principal diagonal. We do not indulge in this extra generality, however, because, for the purposes of this paper, it is essentially trivial. Any new

matrix $\widetilde{F}_n(\tau, \alpha, \beta)$ so obtained is similar to the matrix $\widetilde{F}_n(\tau, \alpha)$ (that is, with $\beta = 0$) that is already under consideration, and the similarity is implemented by a *diagonal* matrix. Since any similarity transformation induced by a diagonal matrix preserves band structure (for example, tridiagonal matrices map to tridiagonal matrices, and extended-tridiagonal matrices map to extended-tridiagonal matrices), the results of this paper can easily be transferred to the more general context.

We call $\widetilde{F}_n(\tau, 0)$—or $F_n(\tau, 0)$, which is the same thing when $\tau \neq 0$—a *centered Fourier matrix*, and we call the especially important special case $F_n(1/n, 0)$ the *primal* centered Fourier matrix of order n, where the adjective "primal" means that q is equal to the first nth root of unity, counting counterclockwise from 1. Regardless of the parity of n, the anchor point of a centered Fourier matrix is at the center of the matrix, although, when n is even, that point does not correspond to a genuine entry in the matrix. For example, in the 5×5 case the centered Fourier matrix is

$$F_5(\tau, 0) = \frac{1}{\sqrt{5}} \begin{bmatrix} q^4 & q^2 & 1 & q^{-2} & q^{-4} \\ q^2 & q & 1 & q^{-1} & q^{-2} \\ 1 & 1 & 1 & 1 & 1 \\ q^{-2} & q^{-1} & 1 & q & q^2 \\ q^{-4} & q^{-2} & 1 & q^2 & q^4 \end{bmatrix}, \tag{2.7}$$

while in the 6×6 case the centered Fourier matrix is

$$F_6(\tau, 0) = \frac{1}{\sqrt{6}} \begin{bmatrix} q^{25/4} & q^{15/4} & q^{5/4} & q^{-5/4} & q^{-15/4} & q^{-25/4} \\ q^{15/4} & q^{9/4} & q^{3/4} & q^{-3/4} & q^{-9/4} & q^{-15/4} \\ q^{5/4} & q^{3/4} & q^{1/4} & q^{-1/4} & q^{-3/4} & q^{-5/4} \\ q^{-5/4} & q^{-3/4} & q^{-1/4} & q^{1/4} & q^{3/4} & q^{5/4} \\ q^{-15/4} & q^{-9/4} & q^{-3/4} & q^{3/4} & q^{9/4} & q^{15/4} \\ q^{-25/4} & q^{-15/4} & q^{-5/4} & q^{5/4} & q^{15/4} & q^{25/4} \end{bmatrix}. \tag{2.8}$$

In the literature, Fourier matrices have usually been chosen to have the anchor point in the upper-left corner. Since the quantity $a = \alpha/\tau$ measures the offset of the anchor point from the center of the matrix, we can put $F_n(\tau, \alpha)$ into that form by taking $a = (n-1)/2$, that is, $\alpha = (n-1)\tau/2$, obtaining the matrix

$$\frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & q & q^2 & \cdots & q^{n-1} \\ 1 & q^2 & q^4 & \cdots & q^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & q^{n-1} & q^{2(n-1)} & \cdots & q^{(n-1)^2} \end{bmatrix}, \tag{2.9}$$

which we call a *traditional Fourier matrix*. In the special case in which $\tau = 1/n$ and $\alpha = (n-1)/(2n)$, we call the matrix $F_n(\tau, \alpha)$ the primal traditional Fourier matrix of order n. These are the matrices that have traditionally been used for discrete Fourier synthesis, and their inverses have been used for discrete Fourier analysis.

It is for these primal traditional Fourier matrices that Grünbaum and Dickinson and Steiglitz originally discovered their tridiagonal and extended-tridiagonal commutors. Let F denote the primal case ($q = e^{2\pi i/n}$) of the matrix in (2.9). Let $c_k = \cos(\pi k/n)$ and $s_k = \sin(\pi k/n)$. Then the nonscalar tridiagonal commutor of F

discovered by Grünbaum [8] is the matrix

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 2c_1 s_1^2 & -s_1 s_2 & 0 & \cdots & 0 & 0 \\ 0 & -s_1 s_2 & 2c_1 s_2^2 & -s_2 s_3 & \cdots & 0 & 0 \\ 0 & 0 & -s_2 s_3 & 2c_1 s_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2c_1 s_{n-2}^2 & -s_{n-2} s_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & -s_{n-2} s_{n-1} & 2c_1 s_{n-1}^2 \end{bmatrix} \tag{2.10}$$

(although actually Grünbaum divided through by $s_1 s_2$), and the extended-tridiagonal commutor of Dickinson and Steiglitz [4] is

$$X = \begin{bmatrix} 2c_0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 2c_2 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 2c_4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2c_{2n-4} & 1 \\ 1 & 0 & 0 & \cdots & 1 & 2c_{2n-2} \end{bmatrix}. \tag{2.11}$$

In this paper, we generalize T and X to matrices $T_n(\tau, \alpha)$ and $X_n(\tau, \alpha)$, which we call Grünbaum matrices and Dickinson–Steiglitz matrices, respectively, in honor of the pioneering discoveries of Grünbaum and of Dickinson and Steiglitz.

The tridiagonal matrices $T_n(\tau, \alpha)$—the Grünbaum matrices—are defined in section 3. The matrix $T_n(\tau, \alpha)$ is defined for all values of the parameters τ and α, but it does not always commute with $\widetilde{F}_n(\tau, \alpha)$. Theorem 3.1 gives necessary and sufficient conditions for the two matrices to commute. There is an annoying complication in finding the "right" standardization of the Grünbaum matrix, and so we actually work with two slightly different forms, $T_n^{\mathrm{A}}(\tau, \alpha)$ and $T_n^{\mathrm{B}}(\tau, \alpha)$.

The extended-tridiagonal matrices $X_n(\tau, \alpha)$—the Dickinson–Steiglitz matrices—are defined in section 4. Again, $X_n(\tau, \alpha)$ is defined for all values of τ and α, but it does not always commute with the corresponding shifted Fourier matrix. Theorem 4.1 gives necessary and sufficient conditions for the two matrices to commute.

Finally, in section 5, we describe the extent to which the Grünbaum matrices and the Dickinson–Steiglitz matrices provide all of the tridiagonal and extended-tridiagonal commutors of $\widetilde{F}_n(\tau, \alpha)$. All proofs are omitted in section 5. The proofs are elementary (based on recurrence relations, on row and column reduction to forms somewhat analogous to the Smith normal form, and on similarity transformations), but they are quite long and intricate, and presenting them would tend to obscure the essential simplicity of the results presented in this paper.

3. Grünbaum matrices. We define the first form of the shifted version of the Grünbaum matrix to be the tridiagonal matrix $T_n^{\mathrm{A}}(\tau, \alpha)$ whose diagonal entries are

$$-2\cos\bigl(\pi(n\tau - 2\alpha)\bigr)\sin(\pi\mu\tau)\sin\bigl(\pi((n-\mu-1)\tau - 2\alpha)\bigr) \quad \text{for } 0 \le \mu \le n-1 \tag{3.1}$$

and whose subdiagonal entries and superdiagonal entries are

$$\sin(\pi\mu\tau)\sin\bigl(\pi((n-\mu)\tau - 2\alpha)\bigr) \quad \text{for } 1 \le \mu \le n-1. \tag{3.2}$$

For example, when $n = 5$ the matrix $T_n^{\mathrm{A}}(\tau, \alpha)$ is

$$\begin{bmatrix} 0 & s_1 s_4^{\mathrm{A}} & 0 & 0 & 0 \\ s_1 s_4^{\mathrm{A}} & -2c^{\mathrm{A}} s_1 s_3^{\mathrm{A}} & s_2 s_3^{\mathrm{A}} & 0 & 0 \\ 0 & s_2 s_3^{\mathrm{A}} & -2c^{\mathrm{A}} s_2 s_2^{\mathrm{A}} & s_3 s_2^{\mathrm{A}} & 0 \\ 0 & 0 & s_3 s_2^{\mathrm{A}} & -2c^{\mathrm{A}} s_3 s_1^{\mathrm{A}} & s_4 s_1^{\mathrm{A}} \\ 0 & 0 & 0 & s_4 s_1^{\mathrm{A}} & -2c^{\mathrm{A}} s_4 s_0^{\mathrm{A}} \end{bmatrix}, \tag{3.3}$$

where $c^{\mathrm{A}} = \cos\bigl(\pi(n\tau - 2\alpha)\bigr)$ and $s_k = \sin(\pi k\tau)$ and $s_k^{\mathrm{A}} = \sin\bigl(\pi(k\tau - 2\alpha)\bigr)$. We define $T_n^{\mathrm{A}}(\tau, \alpha)$ only when $n \geq 2$.

THEOREM 3.1. *Let $n \geq 2$. Then $T_n^{\mathrm{A}}(\tau, \alpha)$ commutes with $\widetilde{F}_n(\tau, \alpha)$ if and only if $n\tau$ is an integer or 2α is an integer, that is, if and only if $q^n = 1$ or $z^2 = 1$.*

Proof. Let $F = \sqrt{n}\,\widetilde{F}_n(\tau, \alpha)$, let $T = T_n^{\mathrm{A}}(\tau, \alpha)$, and let W denote their commutator $W = FT - TF$. A direct calculation shows that W has the form

$$W = \eta(q^n - 1)(z^2 - 1) \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -h \\ 0 & 0 & 0 & \cdots & 0 & 0 & -h^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & -h^{n-3} \\ 0 & 0 & 0 & \cdots & 0 & 0 & -h^{n-2} \\ 1 & h & h^2 & \cdots & h^{n-3} & h^{n-2} & 0 \end{bmatrix}, \tag{3.4}$$

where $h = q^{(n+1)/2} z$ and where η is the nonzero quantity $\eta = \frac{1}{4} q^{-(n^2+2n-1)/4}$. Thus T commutes with F if and only if $(q^n - 1)(z^2 - 1) = 0$, that is, if and only if $q^n = 1$ or $z^2 = 1$. ∎

A special case of Theorem 3.1 tells us that every centered Fourier matrix ($\alpha = 0$) has a nonscalar tridiagonal commutor, regardless of whether q is an nth root of unity. (Some care is required if the Grünbaum matrix vanishes; see the remarks below.) This is one of the many properties that are better for centered Fourier matrices than for traditional Fourier matrices. Another special case tells us that every primal Fourier matrix ($\tau = 1/n$) has a nonscalar tridiagonal commutor, regardless of the amount of shifting. Other special cases are quite peculiar. For example, $F_n\bigl(1/(n-1), 1/2\bigr)$ is a traditional Fourier matrix, since its anchor point is in the upper-left corner, but it is based on a root of unity of the wrong order, an $(n-1)$th root of unity instead of an nth root of unity. Yet it commutes with the corresponding Grünbaum matrix because 2α is an integer even though $n\tau$ is not. For example, in the 5×5 case the matrices are as follows:

$$F_5(1/4, 1/2) = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & \mathring{\imath} & -1 & -\mathring{\imath} & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -\mathring{\imath} & -1 & \mathring{\imath} & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \tag{3.5}$$

$$T_5^{\mathrm{A}}(1/4, 1/2) = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.6}$$

It is sometimes possible for $T_n^{\mathrm{A}}(\tau,\alpha)$ to vanish when 2τ is an integer, that is, when $q=\pm 1$. The details are in our next theorem.

THEOREM 3.2. *Let $n\geq 2$. Then $T_n^{\mathrm{A}}(\tau,\alpha)=0$ if and only if either $\tau\equiv 0 \pmod 1$ or else both $\tau\equiv 1/2 \pmod 1$ and $2\alpha\equiv(n-1)/2 \pmod 1$. If $T_n^{\mathrm{A}}(\tau,\alpha)$ is not zero, then it is a nonscalar tridiagonal matrix.*

Proof. Assume that τ is not an integer, and look at the upper-left 2×2 submatrix of $T_n^{\mathrm{A}}(\tau,\alpha)$. The entry in position $(0,0)$ is zero. We will prove that, if both of the entries in positions $(1,0)$ and $(1,1)$ are zero, then $\tau\equiv 1/2 \pmod 1$ and $2\alpha\equiv(n-1)/2 \pmod 1$. Those two entries are

$$\sin(\pi\tau)\sin\bigl(\pi((n-1)\tau-2\alpha)\bigr)$$

and

$$-2\cos\bigl(\pi(n\tau-2\alpha)\bigr)\sin(\pi\tau)\sin\bigl(\pi((n-2)\tau-2\alpha)\bigr),$$

respectively. Since $\sin(\pi\tau)\neq 0$, assuming that these two quantities vanish is equivalent to assuming that $(n-1)\tau-2\alpha$ is an integer and at that least one of $n\tau-2\alpha-1/2$ and $(n-2)\tau-2\alpha$ is an integer. If $(n-1)\tau-2\alpha$ and $(n-2)\tau-2\alpha$ were both integers, then τ would be an integer, contrary to assumption. Therefore, we conclude that $(n-1)\tau-2\alpha$ and $n\tau-2\alpha-1/2$ are both integers. Thus their difference, $\tau-1/2$, must be an integer. Therefore, $n\tau-2\alpha-1/2$ and $\tau-1/2$ are both integers; that is, $n(\tau-1/2)+(n-1)/2-2\alpha$ and $\tau-1/2$ are both integers. It follows that $(n-1)/2-2\alpha$ is an integer. We have proven that $\tau\equiv 1/2 \pmod 1$ and $2\alpha\equiv(n-1)/2 \pmod 1$.

Conversely, it is obvious that $T_n^{\mathrm{A}}(\tau,\alpha)=0$ if τ is an integer. And $T_n^{\mathrm{A}}(\tau,\alpha)$ is also zero if $\tau\equiv 1/2 \pmod 1$ and $2\alpha\equiv(n-1)/2 \pmod 1$ because the cosine factor shared by all entries on the main diagonal vanishes, and, for each entry on the subdiagonal, one or the other of the two sine factors vanishes. ∎

When the Grünbaum matrix $T_n^{\mathrm{A}}(\tau,\alpha)$ is zero, the theorems of section 5 imply that $\widetilde{F}_n(\tau,\alpha)$ nevertheless does have a nonscalar tridiagonal commutor. If τ is an integer, then the space of tridiagonal commutors is high-dimensional (Proposition 5.3). If $\tau\equiv 1/2 \pmod 1$ and $2\alpha\equiv(n-1)/2 \pmod 1$, then a nonscalar tridiagonal commutor can always be found by using $\partial T_n^{\mathrm{A}}(\tau,\alpha)/\partial\tau$ (if n is odd) or $\partial T_n^{\mathrm{A}}(\tau,\alpha)/\partial\alpha$ (if n is even) in place of $T_n^{\mathrm{A}}(\tau,\alpha)$. This can be seen by varying τ or α so that (τ,α) follows a line in the $\tau\alpha$-plane along which the commutator $\widetilde{F}_n(\tau,\alpha)T_n^{\mathrm{A}}(\tau,\alpha)-T_n^{\mathrm{A}}(\tau,\alpha)\widetilde{F}_n(\tau,\alpha)$ is identically zero but $T_n^{\mathrm{A}}(\tau,\alpha)$ is not identically zero. The existence of such a line follows from Theorems 3.1 and 3.2, and the line is parallel to the τ-axis if n is odd and is parallel to the α-axis if n is even.

The Grünbaum matrix $T_n^{\mathrm{A}}(\tau,\alpha)$ has the sines arranged so that the entry in the upper-left corner is always zero. By arranging things in reverse, we obtain a second standardization, $T_n^{\mathrm{B}}(\tau,\alpha)$, in which the entry in the lower-right corner is always zero. That matrix is the tridiagonal matrix whose diagonal entries are

$$-2\cos\bigl(\pi(n\tau+2\alpha)\bigr)\sin\bigl(\pi(n-\mu-1)\tau\bigr)\sin\bigl(\pi(\mu\tau+2\alpha)\bigr)\quad\text{for } 0\leq\mu\leq n-1 \tag{3.7}$$

and whose subdiagonal entries and superdiagonal entries are

$$\sin\bigl(\pi(n-\mu)\tau\bigr)\sin\bigl(\pi(\mu\tau+2\alpha)\bigr)\quad\text{for } 1\leq\mu\leq n-1. \tag{3.8}$$

For example, when $n = 5$ the matrix $T_n^{\mathrm{B}}(\tau, \alpha)$ is

$$\begin{bmatrix} -2c^{\mathrm{B}}s_4s_0^{\mathrm{B}} & s_4s_1^{\mathrm{B}} & 0 & 0 & 0 \\ s_4s_1^{\mathrm{B}} & -2c^{\mathrm{B}}s_3s_1^{\mathrm{B}} & s_3s_2^{\mathrm{B}} & 0 & 0 \\ 0 & s_3s_2^{\mathrm{B}} & -2c^{\mathrm{B}}s_2s_2^{\mathrm{B}} & s_2s_3^{\mathrm{B}} & 0 \\ 0 & 0 & s_2s_3^{\mathrm{B}} & -2c^{\mathrm{B}}s_1s_3^{\mathrm{B}} & s_1s_4^{\mathrm{B}} \\ 0 & 0 & 0 & s_1s_4^{\mathrm{B}} & 0 \end{bmatrix}, \tag{3.9}$$

where $c^{\mathrm{B}} = \cos\bigl(\pi(n\tau + 2\alpha)\bigr)$, $s_k = \sin(\pi k\tau)$, and $s_k^{\mathrm{B}} = \sin\bigl(\pi(k\tau + 2\alpha)\bigr)$.

Theorems much like Theorems 3.1 and 3.2 also hold for $T_n^{\mathrm{B}}(\tau, \alpha)$. We shall also see in Theorem 4.2 that $T_n^{\mathrm{A}}(\tau, \alpha)$ and $T_n^{\mathrm{B}}(\tau, \alpha)$ differ by a scalar matrix if and only if they commute with $\widetilde{F}_n(\tau, \alpha)$.

In order not to favor one corner of the matrix (upper-left or lower-right) over the other, it may be preferable to use $(T^{\mathrm{A}} + T^{\mathrm{B}})/2$ as the standard Grünbaum matrix or even to adjust that matrix by adding a scalar multiple of the identity matrix so that the adjusted matrix has its trace equal to zero.

In the centered case ($\alpha = 0$), the two forms $T_n^{\mathrm{A}}(\tau, 0)$ and $T_n^{\mathrm{B}}(\tau, 0)$ are always equal, and they always commute with $\widetilde{F}_n(\tau, 0)$. In the primal centered case ($\tau = 1/n$, $\alpha = 0$), the matrix $T_n^{\mathrm{A}}(1/n, 0) + 2\sin^2\bigl(\pi/(2n)\bigr)I_n$ has a rather simple form involving squares of cosines, as illustrated by the 5×5 case

$$\begin{bmatrix} 2p_4 & p_3 & 0 & 0 & 0 \\ p_3 & 2p_2 & p_1 & 0 & 0 \\ 0 & p_1 & 2p_0 & p_1 & 0 \\ 0 & 0 & p_1 & 2p_2 & p_3 \\ 0 & 0 & 0 & p_3 & 2p_4 \end{bmatrix}, \tag{3.10}$$

where $p_k = \cos^2(\pi k/10)$.

4. Dickinson–Steiglitz matrices. For $n \geq 3$, we define the shifted version of the Dickinson–Steiglitz matrix to be the extended-tridiagonal matrix $X_n(\tau, \alpha; u, v)$ whose diagonal entries are

$$2\cos\bigl(\pi((2\mu - n + 1)\tau + 2\alpha)\bigr) \quad \text{for } 0 \leq \mu \leq n - 1, \tag{4.1}$$

whose subdiagonal entries and superdiagonal entries are all 1, and whose entries in the upper-right and lower-left corners are u and v, respectively. For example, in the 4×4 case the matrix $X_4(\tau, \alpha; u, v)$ is

$$\begin{bmatrix} 2c_{2\alpha-3\tau} & 1 & 0 & u \\ 1 & 2c_{2\alpha-\tau} & 1 & 0 \\ 0 & 1 & 2c_{2\alpha+\tau} & 1 \\ v & 0 & 1 & 2c_{2\alpha+3\tau} \end{bmatrix}, \tag{4.2}$$

where $c_x = \cos(\pi x)$. The case in which $n = 2$ is of some use in Theorem 4.2 below, and we extend the definition to that case by simply ignoring u and v so that

$$X_2(\tau, \alpha; u, v) = \begin{bmatrix} 2\cos\bigl(\pi(2\alpha - \tau)\bigr) & 1 \\ 1 & 2\cos\bigl(\pi(2\alpha + \tau)\bigr) \end{bmatrix}. \tag{4.3}$$

The off-diagonal corner entries u and v are ± 1 in the most interesting cases, as shown in Theorem 4.1 below. With that in mind, we abbreviate $X_n(\tau, \alpha; u, v)$ to

TABLE 4.1
The sign of the corner entries.

n	$n\tau$	$2n\alpha$	u and v
even	even	even	$+1$
even	even	odd	-1
even	odd	even	-1
even	odd	odd	$+1$
odd	even	even	$+1$
odd	even	odd	-1
odd	odd	even	$+1$
odd	odd	odd	-1

$X_n(\tau,\alpha)$ when $u = e^{-\pi i((n-1)n\tau+2n\alpha)}$ and $v = e^{\pi i((n-1)n\tau+2n\alpha)}$. This reduces to the original matrix discovered by Dickinson and Steiglitz [4] in the primal traditional case when $\tau = 1/n$, $\alpha = (n-1)/(2n)$, and $u = v = 1$.

THEOREM 4.1. *Let $n \geq 4$. Then $X_n(\tau,\alpha;u,v)$ commutes with $\widetilde{F}_n(\tau,\alpha)$ if and only if $n\tau$ is an integer, $2n\alpha$ is an integer, and $u = v = (-1)^{(n-1)n\tau}(-1)^{2n\alpha}$.*

The cases $n = 2$ and $n = 3$ are exceptional: $X_2(\tau,\alpha;u,v)$ commutes with $\widetilde{F}_2(\tau,\alpha)$ if and only if 2α is an integer; and $X_3(\tau,\alpha;u,v)$ commutes with $\widetilde{F}_3(\tau,\alpha)$ if and only if either 3τ is an integer, 6α is an integer, and $u = v = (-1)^{6\alpha}$ or 2α is an integer and $u = v = (-1)^{2\alpha}$.

For commutativity when $n \geq 4$, $n\tau$ and $2n\alpha$ must be integers, and the off-diagonal corner entries of X must both be $+1$ or both be -1, depending on the parities of n, $n\tau$, and $2n\alpha$ according to Table 4.1.

Proof. We prove only the case in which $n \geq 4$, leaving the exceptional cases to any interested reader. Let $F = \sqrt{n}\,\widetilde{F}_n(\tau,\alpha)$, let $X = X_n(\tau,\alpha;u,v)$, and let W denote their commutator $W = FX - XF$. A direct calculation shows that W has the form

$$W = \begin{bmatrix} w_{0,0} & w_{0,1} & w_{0,2} & \cdots & w_{0,n-3} & w_{0,n-2} & w_{0,n-1} \\ w_{1,0} & 0 & 0 & \cdots & 0 & 0 & w_{1,n-1} \\ w_{2,0} & 0 & 0 & \cdots & 0 & 0 & w_{2,n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ w_{n-3,0} & 0 & 0 & \cdots & 0 & 0 & w_{n-3,n-1} \\ w_{n-2,0} & 0 & 0 & \cdots & 0 & 0 & w_{n-2,n-1} \\ w_{n-1,0} & w_{n-1,1} & w_{n-1,2} & \cdots & w_{n-1,n-3} & w_{n-1,n-2} & w_{n-1,n-1} \end{bmatrix}, \tag{4.4}$$

where the corner entries are

$$\begin{aligned} w_{0,0} &= -q^{-(n-1)^2/4}(u-v), \\ w_{0,n-1} &= -q^{-(n^2-1)/4}(z-z^{-1}) - q^{(n-1)^2/4}(z^{n-1}-z^{-(n-1)})u, \\ w_{n-1,0} &= q^{-(n^2-1)/4}(z-z^{-1}) + q^{(n-1)^2/4}(z^{n-1}-z^{-(n-1)})v, \\ w_{n-1,n-1} &= q^{-(n-1)^2/4}(u-v) \end{aligned}$$

and the noncorner edge entries are

$$\begin{aligned} w_{0,\nu} &= q^{(n-1)(2\nu-n+1)/4}z^{\nu-n}\bigl(q^{n(n-1)/2-n\nu} - z^n u\bigr), \\ w_{\mu,0} &= -q^{(n-1)(2\mu-n+1)/4}z^{\mu-n}\bigl(q^{n(n-1)/2-n\mu} - z^n v\bigr), \\ w_{n-1,\nu} &= q^{(n-1)(n-2\nu-1)/4}z^{\nu+1}\bigl(q^{n\nu-n(n-1)/2} - z^{-n} v\bigr), \\ w_{\mu,n-1} &= -q^{(n-1)(n-2\mu-1)/4}z^{\mu+1}\bigl(q^{n\mu-n(n-1)/2} - z^{-n} u\bigr). \end{aligned}$$

In terms of q and z, the conditions that are claimed to be necessary and sufficient for commutativity when $n \geq 4$ are (1) $q^n = 1$, (2) $z^{2n} = 1$, and (3) $u = v = q^{(n-1)n/2}z^n$. It is clear that if these three conditions are satisfied, then $W = 0$. (Some care is required because q^n and $q^{(n-1)n/2}$ are abbreviations for $e^{2\pi i n\tau}$ and $e^{2\pi i(n-1)n\tau/2}$, and so $q^n = 1$ does not necessarily imply $q^{(n-1)n/2} = 1$ by raising both sides to the $(n-1)/2$ power.)

It remains to prove the converse. Therefore, assume that $n \geq 4$ and that $W = 0$. We will prove conditions (1), (2), and (3) by looking at $w_{0,0}$, $w_{0,1}$, $w_{0,2}$, and $w_{n-2,n-1}$.

First, $w_{0,1} = 0$ implies that $z^n u = q^{n(n-3)/2}$, and $w_{0,2} = 0$ implies that $z^n u = q^{n(n-5)/2}$, the condition $n \geq 4$ ensuring that $w_{0,2}$ is not the upper-right entry of W. Since $z^n u$ is equal to both $q^{n(n-3)/2}$ and $q^{n(n-5)/2}$, we have $q^n = 1$, showing that condition (1) is satisfied.

Next, $w_{n-2,n-1} = 0$ implies that $u = q^{n(n-3)/2}z^n$. Since we have already seen that $u = q^{n(n-3)/2}z^{-n}$, it follows that $z^{2n} = 1$, showing that condition (2) is satisfied.

Finally, since we know that $q^n = 1$ and that $u = q^{n(n-3)/2}z^n$, it follows that $u = q^{(n-1)n/2}z^n$. Since $w_{0,0} = -q^{-(n-1)^2/4}(u - v) = 0$, it also follows that $u = v$. Thus condition (3) is satisfied. ∎

The matrix $X_n(\tau, \alpha; 0, 0)$ is tridiagonal, and Theorem 4.1 shows that it never commutes with $\widetilde{F}_n(\tau, \alpha)$ except when $n = 2$ and 2α is an integer. It is, however, just what is needed to describe the connection between our two different standardizations of the Grünbaum matrix, as stated in the following theorem.

THEOREM 4.2. *If $n \geq 2$, then the difference between $T_n^{\mathrm{A}}(\tau, \alpha)$ and $T_n^{\mathrm{B}}(\tau, \alpha)$ is a linear combination of I_n and $X_n(\tau, \alpha; 0, 0)$, namely,*

$$T_n^{\mathrm{A}}(\tau, \alpha) - T_n^{\mathrm{B}}(\tau, \alpha) = c_1 I_n + c_2 X_n(\tau, \alpha; 0, 0), \tag{4.5}$$

where $c_1 = \sin\bigl(\pi(2n-1)\tau\bigr)\sin(4\pi\alpha)$ and $c_2 = -\sin(\pi n\tau)\sin(2\pi\alpha)$.

Proof. The proof is a straightforward calculation with trigonometric identities. ∎

Among all of the extended-tridiagonal commutors of $\widetilde{F}_n(\tau, \alpha)$ when $n\tau$ and $2n\alpha$ are integers, the Dickinson–Steiglitz matrix $X_n(\tau, \alpha)$ stands out because of the simplicity of its subdiagonal and superdiagonal. There is a second extended-tridiagonal matrix $Y_n(\tau, \alpha)$ of simple form—having all zeros on its main diagonal—that also commutes with $\widetilde{F}_n(\tau, \alpha)$ whenever $n\tau$ and $2n\alpha$ are integers. In order to include the noncommuting cases gracefully, it is convenient to have two slightly different variants,

$Y_n^{\mathrm{A}}(\tau,\alpha)$ and $Y_n^{\mathrm{B}}(\tau,\alpha)$. The first of these is

$$Y_n^{\mathrm{A}}(\tau,\alpha) = \begin{bmatrix} 0 & y_1 & 0 & \cdots & 0 & 0 & uy_0 \\ y_1 & 0 & y_2 & \cdots & 0 & 0 & 0 \\ 0 & y_2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & y_{n-2} & 0 \\ 0 & 0 & 0 & \cdots & y_{n-2} & 0 & y_{n-1} \\ vy_0 & 0 & 0 & \cdots & 0 & y_{n-1} & 0 \end{bmatrix}, \tag{4.6}$$

where

$$y_\mu = \cos\bigl(\pi((2\mu - n)\tau + 2\alpha)\bigr), \tag{4.7}$$

$$u = e^{-\pi i((n-1)n\tau+2n\alpha)}, \tag{4.8}$$

$$v = e^{\pi i((n-1)n\tau+2n\alpha)}. \tag{4.9}$$

The second variant, $Y_n^{\mathrm{B}}(\tau,\alpha)$, is the same except that the off-diagonal corner entries are uy_n and vy_n instead of uy_0 and vy_0. The matrix $Y_n^{\mathrm{A}}(\tau,\alpha)$ is a linear combination of $T_n^{\mathrm{A}}(\tau,\alpha)$, $X_n(\tau,\alpha)$, and the identity matrix, namely,

$$Y_n^{\mathrm{A}}(\tau,\alpha) = 2T_n^{\mathrm{A}}(\tau,\alpha) + y_0 X_n(\tau,\alpha) - 2y_0 y_{1/2} I_n, \tag{4.10}$$

and there is the alternate linear relationship

$$Y_n^{\mathrm{B}}(\tau,\alpha) = 2T_n^{\mathrm{B}}(\tau,\alpha) + y_n X_n(\tau,\alpha) - 2y_n y_{n-1/2} I_n, \tag{4.11}$$

both of which hold regardless of whether the matrices involved commute with $\widetilde{F}_n(\tau,\alpha)$.

5. Uniqueness. If a square matrix A has a nonscalar tridiagonal commutor B, then it has infinitely many—just look at $c_1 I + c_2 B$, where c_1 and c_2 are scalars and $c_2 \neq 0$. However, these are related to each other in such a trivial way that they should be regarded as essentially the same. In particular, they have the same eigenspaces. From this point of view, we can say that a tridiagonal commutor is *essentially unique* if and only if the space of all tridiagonal commutors is two-dimensional. We shall see that, if a shifted Fourier matrix $\widetilde{F}_n(\tau,\alpha)$ has a nonscalar tridiagonal commutor, then in most cases that commutor is essentially unique and is the Grünbaum matrix $T_n^{\mathrm{A}}(\tau,\alpha)$.

We do not fare quite so well with the Dickinson–Steiglitz commutors. When a shifted Fourier matrix commutes with the corresponding Dickinson–Steiglitz matrix, it always also commutes with a nonscalar tridiagonal matrix (usually the corresponding Grünbaum matrix), and so we have to expect the set of extended-tridiagonal commutors to be at least three-dimensional; in most cases it is exactly three-dimensional. Consequently, the eigenspaces are not uniquely determined simply by asking for an extended-tridiagonal matrix that is not tridiagonal.

Let $\mathcal{T} = \mathcal{T}_n(\tau,\alpha)$ denote the set of all tridiagonal commutors of $\widetilde{F}_n(\tau,\alpha)$, and let $\mathcal{E} = \mathcal{E}_n(\tau,\alpha)$ denote the set of all extended-tridiagonal commutors of $\widetilde{F}_n(\tau,\alpha)$. In this section, we state theorems giving the dimensions of $\mathcal{T}$ and $\mathcal{E}$ in all cases. For simplicity, we give the dimensions only, without giving a list of basis matrices, and we omit the proofs, which are quite lengthy. In most cases when n is not too small and $q \neq \pm 1$, the basis matrices can be found among the identity matrix, the Grünbaum commutor (if there is one), and the Dickinson–Steiglitz commutor (if there is one).

The statement of the $\mathcal{E}$-part of Theorem 5.1 benefits from the use of an abbreviation for what is, in effect, the world's most general characteristic function. We adopt the "bracket" version of Iverson's notation [5, pp. 24–25] for this, writing $[\mathcal{B}]$, where $\mathcal{B}$ is any Boolean-valued expression in any set of variables, to mean 1 if $\mathcal{B}$ is true and 0 if $\mathcal{B}$ is false.

THEOREM 5.1. *Let n be an integer with $n \geq 6$, and let τ and α be complex constants such that 2τ is not an integer. Then $\mathcal{T}_n(\tau,\alpha)$ is a subspace of $\mathbb{C}^{n\times n}$ whose dimension is exactly 2 if $n\tau$ is an integer or 2α is an integer (that is, if $q^n = 1$ or $z^2 = 1$) and is exactly 1 otherwise. Furthermore, $\mathcal{E}_n(\tau,\alpha)$ is a subspace of $\mathbb{C}^{n\times n}$ whose dimension is*

$$\begin{aligned}\dim[\mathcal{E}_n(\tau,\alpha)] &= 1 + [n\tau \in \mathbb{Z} \vee 2\alpha \in \mathbb{Z}] + [n\tau \in \mathbb{Z} \wedge 2n\alpha \in \mathbb{Z}] + [(n-1)\tau \in \mathbb{Z}] \\ &= 1 + [q^n = 1 \vee z^2 = 1] + [q^n = 1 \wedge z^{2n} = 1] + [q^{n-1} = 1].\end{aligned}$$

In Theorem 5.1, τ is not an integer, and so $n\tau$ and $(n-1)\tau$ cannot both be integers; so $\dim(\mathcal{E}) \leq 3$. When $q^{n-1} = 1$, one of the basis matrices for $\mathcal{E}$ can be chosen to be a matrix all of whose noncorner entries are zero.

Theorem 5.1 completely determines the dimensions of $\mathcal{T}$ and $\mathcal{E}$ when $n \geq 6$ and 2τ is not an integer. The situation can be somewhat more complicated in the other cases, that is, when either $n \leq 5$ or $q^2 = 1$. For completeness, we now give the dimensions of $\mathcal{T}$ and of $\mathcal{E}$ in all those exceptional cases, leaving the task of finding basis matrices to any interested reader. We express the results in terms of q and z instead of in terms of τ and α. When n is 1 or 2, $\mathcal{T}$ and $\mathcal{E}$ are the same thing, namely, the entire commutant of $\widetilde{F}_n(\tau,\alpha)$. When $n = 3$, $\mathcal{E}$ is the entire commutant of $\widetilde{F}_n(\tau,\alpha)$.

PROPOSITION 5.2. *Let $n \leq 2$. Then $\dim(\mathcal{T}) = \dim(\mathcal{E}) = n$.*

PROPOSITION 5.3. *Let $n \geq 3$ and let $q^2 = 1$.*

1. *If $q = 1$, then $\dim(\mathcal{T}) = n$ and $\dim(\mathcal{E}) = n + 2$.*
2. *If $q = -1$ and n is odd, then $\dim(\mathcal{T}) = 2$ and $\dim(\mathcal{E}) = 3$.*
3. *If $q = -1$ and n is even, then $\dim(\mathcal{T}) = 2 + b$ and $\dim(\mathcal{E}) = 4 + b$, where*

$$b = [z^{2n} = 1 \wedge z^4 \neq 1].$$

PROPOSITION 5.4. *Let $3 \leq n \leq 5$ and let $q^2 \neq 1$.*

1. *If $n = 3$, then $\dim(\mathcal{T}) = 2 + b$ and $\dim(\mathcal{E}) = 3 + 2b$, where*

$$b = [q = -1/2 \wedge z^2 = 1].$$

2. *If $n = 4$, then $\dim(\mathcal{T}) = 1 + [q^4 = 1 \vee z^2 = 1]$ and*

$$\dim(\mathcal{E}) = 2 + [(q^4 = 1 \wedge z^8 = 1) \vee z^2 = 1].$$

3. *If $n = 5$, then $\dim(\mathcal{T}) = 1 + [q^5 = 1 \vee z^2 = 1]$ and*

$$\begin{aligned}\dim(\mathcal{E}) = 1 &+ [q^5 = 1 \vee z^2 = 1] + [q^5 = 1 \wedge z^{10} = 1] + [q^4 = 1] \\ &+ [z^2 = -1 \wedge q^5 \neq 1 \wedge q^4 \neq 1] \\ &+ [q^3 = 1 \wedge z \text{ is a primitive } 12\text{th root of unity}].\end{aligned}$$

Acknowledgments. We are grateful to the anonymous referees for suggestions for clarifying and simplifying the presentation.

REFERENCES

[1] D. H. Bailey and P. N. Swarztrauber, *The fractional Fourier transform and applications*, SIAM Rev., 33 (1991), pp. 389–404.

[2] G. Bongiovanni, P. Corsini, and G. Frosini, *One-dimensional and two-dimensional generalized discrete Fourier transforms*, IEEE Trans. Acoust. Speech Signal Process., ASSP-24 (1976), pp. 97–99.

[3] G. Bonnerot and M. Bellanger, *Odd-time odd-frequency discrete Fourier transform for symmetric real-valued series*, Proc. IEEE, 64 (1976), pp. 392–393.

[4] B. W. Dickinson and K. Steiglitz, *Eigenvectors and functions of the discrete Fourier transform*, IEEE Trans. Acoust. Speech Signal Process., ASSP-30 (1982), pp. 25–31.

[5] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed., Addison–Wesley, Reading, MA, 1994.

[6] F. A. Grünbaum, *Eigenvectors of a Toeplitz matrix: Discrete version of the prolate spheroidal wave functions*, SIAM J. Algebraic Discrete Methods, 2 (1981), pp. 136–141.

[7] F. A. Grünbaum, *Toeplitz matrices commuting with tridiagonal matrices*, Linear Algebra Appl., 40 (1981), pp. 25–36.

[8] F. A. Grünbaum, *The eigenvectors of the discrete Fourier transform: A version of the Hermite functions*, J. Math. Anal. Appl., 88 (1982), pp. 355–363.

[9] F. A. Grünbaum, *A remark on Hilbert's matrix*, Linear Algebra Appl., 43 (1982), pp. 119–124.

[10] S. A. Martucci, *Symmetric convolution and the discrete sine and cosine transforms*, IEEE Trans. Signal Process., 42 (1994), pp. 1038–1051.

[11] D. H. Mugler and S. Clary, *Discrete Hermite functions*, in Proceedings of the International Conference on Scientific Computing and Mathematical Modeling, Milwaukee, WI, 2000, pp. 318–321.

[12] D. H. Mugler and S. Clary, *Discrete Hermite functions and the fractional Fourier transform*, in Proceedings of the International Conference on Sampling Theory and Applications, Orlando, FL, 2001, pp. 303–308.

[13] H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing*, John Wiley and Sons, New York, 2001.

[14] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press, New York, 1990.

[15] R. O. Rowlands, *The odd discrete Fourier transform*, in Conference Record of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, 1976, pp. 130–133.

[16] W. W. Sawyer, *On the matrix with elements* $1/(r+s-1)$, Canad. Math. Bull., 17 (1974), pp. 297–298.

[17] G. Strang, *The discrete cosine transform*, SIAM Rev., 41 (1999), pp. 135–147.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 822–851

A BLOCK CONSTANT APPROXIMATE INVERSE FOR PRECONDITIONING LARGE LINEAR SYSTEMS*

PH. GUILLAUME†, A. HUARD†, AND C. LE CALVEZ†

Abstract. A new class of approximate inverses is presented for preconditioning large linear systems issued from the discretization of elliptic boundary value problems. At the intersection of multipole, multigrid, and sparse approximate inverse (SAI) methods, they consist in approximating the inverse of a matrix by a block constant matrix instead of a sparse matrix like in SAI methods. They do not require more storage, or even less, and are well adapted to parallel computing, both for the construction of the preconditioner and for matrix-vector products. Numerical examples are provided and compared with SAI and incomplete Cholesky factorization preconditioners.

Key words. preconditioning, approximate inverse, fast multipole methods, multigrid methods

AMS subject classifications. 65F10, 65F50, 65N30, 65N55, 65Y05, 35J50

PII. S0895479802401515

1. Introduction. Recently, multipole methods [29, 17, 5, 13] have dramatically improved the solution of scattering problems in electromagnetism. The basic idea behind multipole methods consists in a low-rank approximation of far field interactions. The matrix A obtained when using integral equations may be thought of as an approximation of the Green function $g(x, y)$ of the problem, i.e., $A_{ij} \simeq g(x_i, x_j)$ if the x_i's are the discretization points. The matrix A is full, and, for large problems, it becomes impossible to store the whole matrix. When two points x^* and y^* are distant from each other, the starting point of the multipole approach is a separate variables approximation

$$g(x, y) \simeq \sum_{k=1}^{r} u_k(x) v_k(y),$$

which holds for x close to x^* and y close to y^*. It leads to a low-rank approximation of the associated block of the matrix A:

$$A_{IJ} \simeq UV^T.$$

Here I and J are sets of indices of points x_i, $i \in I$, x_j, $j \in J$, respectively, close to x^* and y^*. The matrices U and V are given by $U_{ik} \simeq u_k(x_i)$, $i \in I, V_{jk} \simeq v_k(x_j)$, $j \in J$. The sizes of A_{IJ}, U, and V are, respectively, $|I| \times |J|$, $|I| \times r$, and $|J| \times r$. Hence both memory and computational time are saved if $r \ll |I|$ and $r \ll |J|$. When x^* and y^* are close to each other, the above approximation does not hold anymore, and $g(x^*, y^*)$ must be computed more carefully, i.e., by the usual techniques derived from the integral equations theory. This approach is general and relies on the fact that the Green function associated to a pseudodifferential operator is singular on the diagonal but regular outside.

*Received by the editors January 28, 2002; accepted for publication by A. Wathen May 31, 2002; published electronically February 4, 2003.
http://www.siam.org/journals/simax/24-3/40151.html
†MIP, UMR 5640, INSA, Département de Mathématiques, 135 Avenue de Rangueil, 31077 Toulouse Cedex 4, France (guillaum@gmm.insa-tlse.fr, huard@gmm.insa-tlse.fr, lecalvez@gmm.insa-tlse.fr).

The situation looks quite different at first when considering finite element methods: the matrix A issued from the discretization of a PDE is sparse, and there is a priori no need for using a low-rank approximation of almost zero blocks (see, however, [4]). However, its inverse A^{-1} is usually a dense matrix: inverses of irreducible matrices are structurally dense [11, 9]. Like the matrix issued from an integral equation, it is associated to a Green function $g(x, y)$ with $(A^{-1})_{ij} \simeq g(x_i, x_j)$, which is singular on the diagonal $x = y$ but smooth outside the diagonal. Here we seek an approximate inverse of A for preconditioning an iterative method. Following the multipole strategy, one can approximate off-diagonal blocks of A^{-1} by low-rank matrices, as, for example, in [20], where some knowledge of the Green function is used. When the Green function is not known, since low-rank off-diagonal blocks can lead to a number of unknowns significantly larger than in the original problem and because we do not need such a good approximation as in the case of integral equations, we can go even further and simply approximate off-diagonal blocks $(A^{-1})_{IJ}$ by constant blocks. Their size can vary; they are smaller when close to the diagonal and get larger away from it. This approach is well adapted to nonoscillatory Green functions associated to elliptic equations like Poisson's equation or elasticity equations. It relies on the fact that piecewise constant functions can well approximate the Green function. It would not be adapted, e.g., to Helmholtz equations (or matrices of the form $K-\omega^2 M$, where K and M are symmetric and positive definite (SPD)) unless the size of the blocks is small enough with respect to wavelength.

Like in the multipole method (see, e.g., [10, 27]), a crucial point is the ordering of the unknowns. They need to be sorted by proximity: unknowns associated to neighboring points have to be grouped together and vice-versa. An analogy can also be found with the nested grids decomposition [26]. When the nodal table which has been used for assembling the matrix is available, a simple way to achieve this is to use a recursive coordinates bisection [28], but other methods can be used as well, like recursive graph bisection or recursive spectral bisection [28, 31, 30], which do not require the nodal table.

In this paper, we focus our attention on the solution of systems issued from the discretization of elliptic boundary value problems, leading to a sparse SPD matrix A of size n. Section 2 describes how to construct a block constant approximate inverse (BCAI) of A together with its connection to multipole, multigrid [14, 23, 16, 25] and sparse approximate inverse (SAI) methods [2, 7, 22]. A connection between the SAI method and multigrid methods can be found in [32] with an opposite objective: the SAI method is proposed as a smoother, whereas the BCAI presented here works more like the coarse grid correction itself. Next, an interpretation of the BCAI in terms of minimizing the distance (for the energy norm) to the discrete Green function is given in section 3. In view of practical implementation, section 4 describes a data structure associated to block constant matrices (BCMs). This structure is used for computing the preconditioner and performing matrix-vector products, both of which can be done in parallel. A particular case of BCAI is described in section 5, where classical properties of multigrid methods are recovered. Finally, numerical experiments are reported in section 6. It is shown that the condition number of the preconditioned system is order $1/h$ (h is the mesh size) when using a number of constants less than n, whereas it is of order $1/h^2$ for the original system. This condition number is comparable to the one obtained by using a modified incomplete Cholesky decomposition [12, 19] but requires less storage. The block constant preconditioner (BCP) obtained from the BCAI is compared to an SAI having the same sparsity pattern as A [8] and to the in-

complete Cholesky factorization with no fill-in IC(0). These two preconditioners have been chosen for comparison because they use in a similar way the sparsity pattern of A. The reported experiments show that the BCP takes the advantage when the size of the system increases, both in terms of memory requirements and number of iterations, the operation counts per iteration being comparable.

2. The BCP. We consider linear systems of the form

$$Ax = b, \quad x, b \in \mathbb{R}^n,$$

where $A \in \mathcal{M}_n(\mathbb{R})$ is throughout this paper an SPD matrix. When the dimension n is large, an iterative method like the preconditioned conjugate gradient is often used for solving such a system. It consists in applying the conjugate gradient algorithm [21, 24] to a system of the form $MAx = Mb$. Here M is an explicit preconditioner and should also be SPD. This paper describes a class of such left preconditioners. Right preconditioners associated to a system $AMu = b$, $x = Mu$, can be derived in a similar way and will not be discussed here. A BCP M is of the form

$$M = C + \omega I, \quad \omega > 0, \tag{1}$$

where I is the identity matrix and C is a BCM, consisting in rectangular blocks of variable size whose elements are constant in each block. For an appropriate choice of the constants, C becomes a BCAI of the matrix A.

As mentioned in the introduction, the unknowns must be ordered by proximity, like in multipole methods. In what follows, we suppose that this reordering has been done. For example, in the case of the discretization of Poisson's equation in a square using the five-point finite difference stencil, Figure 1 shows the sparsity pattern of the matrix with the natural ordering (left) and the one obtained (right) after a recursive coordinates bisection (cf. also section 6).

After reordering, the steps for computing a BCP for the matrix A are the following:

- choose a BCM pattern, i.e., location and size of the different constant blocks;
- compute the constants associated to this pattern by minimizing a Frobenius norm of $C - A^{-1}$ over the set of matrices having the same pattern;
- choose the parameter ω.

An example of a BCM pattern is given in Figure 2. It is associated to the reordered matrix in Figure 1. Each color corresponds to a constant block. One can observe that the blocks are smaller at the locations where the matrix A has nonzero elements.

The way of computing C resembles SAI methods [9, 18, 6, 15]: it minimizes in a given class of matrices a Frobenius norm of $C - A^{-1}$, and its computation as well as the matrix-vector preconditioning operation can be done in parallel. The difference lies in the fact that the approximation of the Green function $g(x, y)$ by a sum of discrete Dirac functions (SAI methods without factorization) is replaced by a piecewise constant function, which is likely to offer a better approximation outside the diagonal $x = y$. For some particular cases of pattern of the BCAI, the method becomes very close to a two-grid method with volume agglomeration [23] (cf. section 5). In the general case, the difference lies in an attempt to simulate a cycle over several grids in one single operation, like in the Bramble, Pasciak, and Xu (BPX) algorithm [3]. Hence the BCP method is at the intersection of multipole, multigrid, and SAI methods.

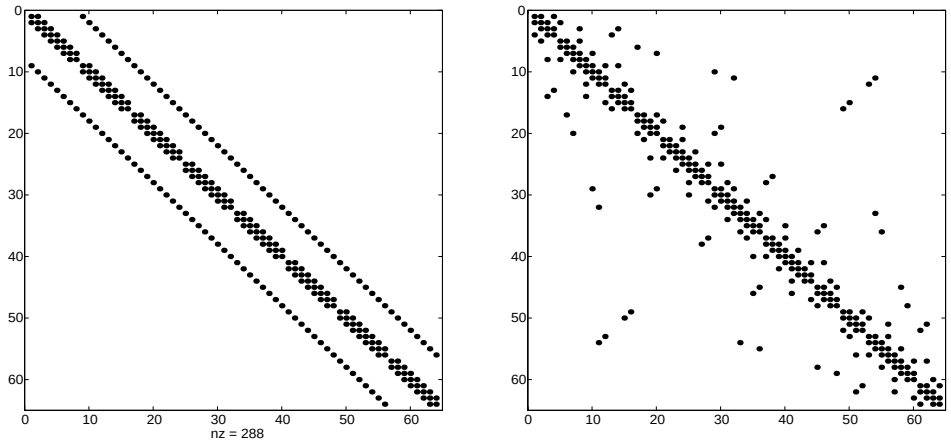


FIG. 1. *Five-point difference matrix before and after reordering.*

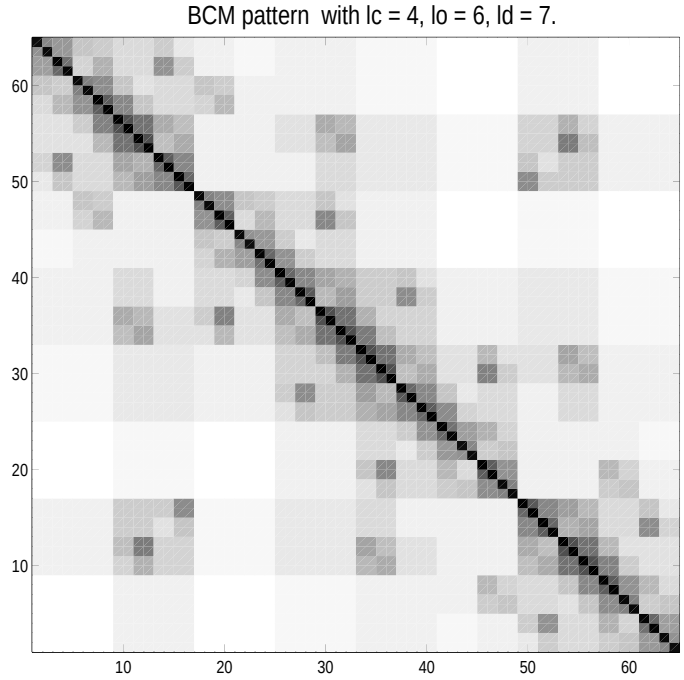


FIG. 2. *BCM pattern example.*

2.1. Determination of a BCM pattern. A pattern is obtained by a recursive splitting of an initial block of size $n \times n$. The depth of recursiveness (the level of refinement) is determined by three parameters l_c, l_d, and l_o:

- l_c is the coarsest level of refinement and fixes the size of the largest blocks of the BCM;
- l_o is an intermediate level of refinement and determines the size of the smallest off-diagonal blocks of the BCM;
- l_d is the finest level of refinement and determines the size of the blocks of the

BCM containing the diagonal elements.

Let d be the integer defined by $2^{d-2} < n \leq 2^{d-1}$. It will always be supposed that

$$1 \leq l_c \leq l_o \leq l_d \leq d.$$

The two extreme situations correspond to $l_d = 1$ (in this case, the whole BCM consists in one single constant block) and to $l_c = d$ (in that case, each block of the BCM is of size 1×1, and the BCAI is equal to A^{-1}).

Different values of these parameters lead to different sequences of refinement and consequently to different patterns. Each refinement consists in splitting a block K into four blocks of equal size if possible. The algorithm is the following.

ALGORITHM 2.1 (determination of the BCM pattern).

1. *lev* $= 1$*; K is an* $n \times n$ *block of level* 1*;*
2. *for lev* $= 2 : l_c$
   - *split each* $k \times l$ *block K of level lev* $- 1$ *into four blocks of size* $k_i \times l_j$*, where* $k_1 \geq k_2$*,* $k_1 + k_2 = k$*,* $k_1 - k_2 \leq 1$*, and* $l_1 \geq l_2$*,* $l_1 + l_2 = l$*,* $l_1 - l_2 \leq 1$*;*

   endfor
3. *for lev* $= l_c + 1 : l_d$
   - *for each* $k \times l$ *block K of level lev* $- 1$
     - *if*
       - *K is a diagonal block and lev* $\leq l_d$
       - *or*
       - *K is not a diagonal block and lev* $\leq l_o$
     - *and*
       - *the corresponding block of A has nonzero elements,*
     - *then*
       - *split the block K into four blocks following the previous rule as long as possible. If* $\min(k, l) = 1$*, just split it into two blocks when* $k \neq l$ *and of course do not split it if* $k = l = 1$*;*
     - *endif*
   - *endfor*

   endfor

Observe that the obtained pattern is symmetric if A is symmetric.

Following this algorithm, the BCM pattern in Figure 2 has been obtained from the classical Poisson square matrix of size 64 (Figure 1, left) reordered by a recursive coordinates bisection method (Figure 1, right). With $l_c = 4$, $l_o = 6$, and $l_d = 7$, the largest blocks are 8×8, the smallest off-diagonal blocks are 2×2, and the smallest diagonal blocks are 1×1. (As all A_{ii} are nonzero, all diagonal blocks are in fact of size 1×1.) The parameter l_d allows a finer refinement on the diagonal, where the Green function is singular. As in SAI methods, many variants can be proposed. For example, instead of using the sparsity pattern of A when deciding whether a block should be split or not, one can use the sparsity pattern of A^m for some integer m.

Another example is given by the following matrix A. The matrix C shows the pattern of the associated BCM obtained by using the previous algorithm with $l_c = l_o = 2$ and $l_d = 4$ (cf. section 4.2 for the ordering of the c_i's). Of course, C will never

be explicitly formed.

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 6 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & 0 & 6 & 1 & 2 \\ 1 & 2 & 0 & 4 & 1 & 2 & 1 & 0 & 0 \\ 0 & 4 & 5 & 1 & 1 & 0 & 2 & 1 & 0 \\ 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 1 & 4 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2}$$

$$C = \left[\begin{array}{ccc|cc|cccc} c_2 & c_2 & c_2 & c_3 & c_3 & c_1 & c_1 & c_1 & c_1 \\ c_2 & c_2 & c_2 & c_3 & c_3 & c_1 & c_1 & c_1 & c_1 \\ c_2 & c_2 & c_2 & c_3 & c_3 & c_1 & c_1 & c_1 & c_1 \\ \hline c_4 & c_4 & c_4 & c_5 & c_6 & c_1 & c_1 & c_1 & c_1 \\ c_4 & c_4 & c_4 & c_7 & c_8 & c_1 & c_1 & c_1 & c_1 \\ \hline c_9 & c_9 & c_9 & c_9 & c_9 & c_{10} & c_{10} & c_{10} & c_{10} \\ c_9 & c_9 & c_9 & c_9 & c_9 & c_{10} & c_{10} & c_{10} & c_{10} \\ c_9 & c_9 & c_9 & c_9 & c_9 & c_{10} & c_{10} & c_{10} & c_{10} \\ c_9 & c_9 & c_9 & c_9 & c_9 & c_{10} & c_{10} & c_{10} & c_{10} \end{array}\right].$$

2.2. Definition of the BCP. Once the BCM pattern has been determined, it remains to compute the values c_i of the constants in each block. Let $\mathcal{C} \subset \mathcal{M}_n(\mathbb{R})$ denote the linear space of matrices which correspond to the obtained pattern. Its dimension is equal to the number n_b of blocks. A basis of this space consists in matrices E_i which have one single nonzero block (corresponding to the given pattern) with value 1. In such a basis, a matrix $C \in \mathcal{C}$ can be written

$$C = \sum_{i=1}^{n_b} c_i E_i.$$

Observe that each E_i can be put under the form $E_i = f_i g_i^T$ with $f_i, g_i \in \{0,1\}^n$. For example, the space $\mathcal{C} \subset \mathcal{M}_9(\mathbb{R})$ associated to (2) is of dimension 10, and the two first E_i's are

$$E_1 = \left[\begin{array}{c|cccc} 0_{5\times 5} & \begin{matrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{matrix} \\ \hline 0_{4\times 5} & 0_{4\times 4} \end{array}\right], \quad E_2 = \left[\begin{array}{ccc|c} \begin{matrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{matrix} & & & 0_{3\times 6} \\ \hline \multicolumn{3}{c|}{0_{6\times 3}} & 0_{3\times 3} \end{array}\right].$$

The space $\mathcal{M}_n(\mathbb{R})$ is equipped with the Frobenius norm defined by its square

$$\|A\|_F^2 = \sum_{i,j=1}^{n} |A_{ij}|^2.$$

This norm is associated to the inner product

$$B : C = \sum_{i,j=1}^{n} B_{ij} C_{ij} = \operatorname{tr}(BC^T), \quad B, C \in \mathcal{M}_n(\mathbb{R}).$$

The subspace $\mathcal{C}$ is equipped with the same norm. A useful property is the following:

$$AC^T : B = A : BC = B^T A : C, \quad A, B, C \in \mathcal{M}_n(\mathbb{R}). \tag{3}$$

DEFINITION 2.1. *The BCP for solving the linear system $Ax = b$ is defined by*

$$M = \frac{1}{2}(C + C^T) + \omega I, \tag{4}$$

where the BCM $C \in \mathcal{C}$ is the solution to the residual norm minimization problem

$$\min_{C \in \mathcal{C}} \|(C - A^{-1})A^{1/2}\|_F^2 \tag{5}$$

and $\omega > 0$ is chosen in such a way that M is positive definite. The matrix $(C+C^T)/2$ itself is called the BCAI of the matrix A.

If $l_c = l_d$ and if A is symmetric, we will see (see (23)) that C is also symmetric, and thus $M = C + \omega I$. In all cases where A is symmetric, $(C + C^T)/2$ is a BCM having the same pattern as C and will be denoted by the same letter C if there is no confusion. When $l_d = d$, it may not be necessary to add a diagonal term ωI because, in that case, $\mathcal{C}$ already contains the diagonal matrices. The larger l_d is, the larger the number n_b of unknowns involved in the BCAI is; however, the numerical results of section 6 show that conclusive results can be obtained with $l_d < d$ and a reasonable number $n_b < n$ of constant blocks. When $l_d < d$, adding a diagonal term becomes necessary because, as in multigrid methods, CA is not invertible, and some smoother must complete the preconditioning operation. A simple way of choosing ω is to make M diagonal dominant.

More general definitions can be proposed, such as, for example, $M = C + D$ with D diagonal and M minimizing $\|(M - A^{-1})A^s\|_F$. The value $s = 1/2$ has a precise meaning only for SPD matrices, although the optimality condition (7) can be used for any kind of matrices without the guarantee of a minimum. A simple choice for non-SPD matrices is to take $s = 1$, which leads to the standard minimization of $\|MA - I\|_F$ also used in SAI methods.

Remark 2.1. Another possibility for using C is to define the preconditioner M by an iteration matrix of the form

$$I - MA = (I - GA)(I - CA), \tag{6}$$

where G is a smoother (for example, Jacobi or Gauss–Seidel), as in multigrid methods. This could be applied to more general matrices than SPD and is at present under study. For SPD matrices, (4) provides a shortcut for the smoothing operation $I - GA$ in (6), and numerical experiments (not presented here) show comparable performances.

2.3. Computation of the BCAI. The computation of C follows from the next proposition.

PROPOSITION 2.2. *If the matrix A is SPD, then problem* (5) *has a unique solution, which is also the unique solution to the linear system of equations*

$$CA : V = I : V \quad \forall\, V \in \mathcal{C}. \tag{7}$$

For $C = \sum_{i=1}^{n_b} c_i E_i$, this system reads

$$Lc = z,\ L \in \mathcal{M}_{n_b}(\mathbb{R}),\ c,\, z \in \mathbb{R}^{n_b}, \tag{8}$$

where, for $i,\ j = 1, \ldots, n_b$,

$$L_{ij} = A : E_j^T E_i = (g_j^T A g_i)(f_j^T f_i), \quad z_i = \operatorname{tr}(E_i). \tag{9}$$

Proof. Equation (7) is the optimality condition associated to the convex minimization problem (5). It is sufficient to prove that a solution H to the square and homogeneous system associated to (7) must be zero. Such a solution satisfies $0 = HA : H = \|HA^{1/2}\|_F^2$, and hence $H = 0$. Finally, system (8) is obtained from $E_i = f_i g_i^T$ and

$$\sum_{j=1}^{n_b} c_j E_j A : E_i = I : E_i, \quad 1 \le i \le n_b. \qquad \square$$

The matrix L is block diagonal with $p = 2^{l_c - 1}$ sparse blocks L_k, $1 \le k \le p$, on the diagonal. (Its sparsity pattern will be discussed in section 4.2.) In particular, its computation as well as the solution of (8) can be parallelized. Depending on the distribution of the matrix A among the processors, the parallelization may be complicated but still remains possible. Similarly, the matrix-vector preconditioning operation is parallelizable (cf. section 4.3).

3. Interpretation via the Green function. We suppose in this section that the matrix A is issued from the discretization of a PDE by using the finite element method. Let $u \in H_0^1(\Omega)$ be the solution to the equation

$$-\Delta u = f \quad \text{in } \Omega, \tag{10}$$

where Ω is a bounded and open subset of $\mathbb{R}^N$ and $f \in L^2(\Omega)$. More general operators can be considered, but this example is sufficient to give an interpretation of the BCAI (which holds also for the SAI). Let $g(x, y)$ be the Green function of the Laplacean operator associated to the Dirichlet boundary condition. It is symmetric in x, y, singular on the diagonal $x = y$, and smooth outside and belongs to the Sobolev space $W^{1,p}(\Omega \times \Omega)$ for $1 \le p < N/(N-1)$. The solution u is given by

$$u(x) = \int_\Omega g(x, y) f(y) dy.$$

Let $\mathcal{V}_h \subset H_0^1(\Omega)$ be a finite element space, and let $(\varphi_i)_{i=1}^n$ be its standard basis. Throughout this section, we use Einstein's summation on repeated indices. The finite element approximation $u_h(x) = \varkappa_i \varphi_i(x)$ of $u(x)$ is defined by

$$\int_\Omega \nabla u_h . \nabla \varphi_i \, dx = \int_\Omega f \, \varphi_i \, dx \quad \forall i = 1, 2, \dots, n,$$

i.e., with $A_{ij} = \int_\Omega \nabla \varphi_j . \nabla \varphi_i \, dx$ and $b_i = \int_\Omega f \, \varphi_i \, dx$:

$$A\varkappa = b.$$

An equivalent way of defining the approximation u_h is to use the discrete Green function $g_h \in \mathcal{V}_h \otimes \mathcal{V}_h$:

$$g_h(x, y) = (A^{-1})_{ij} \varphi_i(x) \varphi_j(y),$$
$$u_h(x) = \int_\Omega g_h(x, y) f(y) dy = (A^{-1} b)_i \varphi_i(x).$$

Finally, the BCAI and SAI methods consist in choosing a subspace $\mathcal{W}_h \subset \mathcal{V}_h \otimes \mathcal{V}_h$ and $c_h = C_{ij} \varphi_i \otimes \varphi_j \in \mathcal{W}_h$ (with $(\varphi_i \otimes \varphi_j)(x, y) = \varphi_i(x) \varphi_j(y)$), which yields an approximation

$$\widetilde{u}_h(x) = \int_\Omega c_h(x, y) f(y) dy = (Cb)_i \varphi_i(x).$$

In the BCAI method, $\mathcal{W}_h$ is the space constructed from a BCM space $\mathcal{C}$,

$$\mathcal{W}_h = \{C_{ij}\varphi_i \otimes \varphi_j\,; \quad C \in \mathcal{C}\}\,,$$

whereas, in the SAI method, a space of matrices having a given sparsity pattern is used instead of $\mathcal{C}$. In the latter case, $\mathcal{W}_h$ consists in sums of discrete Dirac functions, localized around the nodes corresponding to the nonzero entries of C.

At this point, a natural question is how to choose $c_h \in \mathcal{W}_h$ (or, equivalently, C). We can obtain a response in observing how g_h itself is obtained. Although we cannot say that g_h minimizes $\int\int_{\Omega\times\Omega} |\nabla(g - g_h)|^2\, dydx$ because this integral is usually infinite, the discrete Green function is the solution to the associated optimality conditions:

$$\int\int_{\Omega\times\Omega} \nabla(g - g_h).\nabla v_h\, dydx = 0 \quad \forall v_h \in \mathcal{V}_h \otimes \mathcal{V}_h. \tag{11}$$

This can be checked by substituting $g_h = (A^{-1})_{ij}\varphi_i \otimes \varphi_j$ and $v_h = V_{ij}\varphi_i \otimes \varphi_j$ in the integral (see also (17)) and by using the definition of the Green function: $-\Delta_x g(.,y) = \delta_y$, $g = 0$, on $\partial(\Omega\times\Omega)$ (δ_y is the Dirac mass at the point y), which gives

$$\begin{aligned}\int\int_{\Omega\times\Omega} \nabla g.\nabla v_h\, dydx &= \int\int_{\Omega\times\Omega} \nabla_x g.\nabla_x v_h\, dxdy + \int\int_{\Omega\times\Omega} \nabla_y g.\nabla_y v_h\, dydx \\ &= \int_\Omega v_h(y,y)\, dy + \int_\Omega v_h(x,x)\, dx \\ &= 2\int_\Omega v_h(x,x)\, dx. \qquad (12)\end{aligned}$$

Hence, following the variational formulation guideline, i.e., substituting $\mathcal{W}_h$ for $\mathcal{V}_h \otimes \mathcal{V}_h$ in (11), we obtain the following approximation $c_h \in \mathcal{W}_h$ of the Green function g: it should be the solution to

$$\int\int_{\Omega\times\Omega} \nabla(g - c_h).\nabla v_h\, dydx = 0 \quad \forall v_h \in \mathcal{W}_h.$$

Alternatively, it follows from (11) that this formulation is equivalent to

$$\int\int_{\Omega\times\Omega} \nabla c_h.\nabla v_h\, dydx = \int\int_{\Omega\times\Omega} \nabla g_h.\nabla v_h\, dydx \quad \forall v_h \in \mathcal{W}_h, \tag{13}$$

which will be used here as a definition for c_h. The latter formulation is the optimality condition associated to the minimization problem

$$\min_{c_h \in \mathcal{W}_h} \int\int_{\Omega\times\Omega} |\nabla(g_h - c_h)|^2\, dydx. \tag{14}$$

Let B be the mass matrix: $B_{ij} = \int_\Omega \varphi_j\varphi_i\, dx$. We denote by $\mathcal{C}_S$ the subspace of $\mathcal{C}$ whose elements are symmetric matrices. Then we have the following result.

PROPOSITION 3.1. *The two formulations* (13)–(14) *are equivalent. Equation* (13) *has a unique solution* $c_h = C_{ij}\varphi_i \otimes \varphi_j \in \mathcal{W}_h$, *where* $C \in \mathcal{C}$ *is symmetric, is the unique solution to the Sylvester matrix equations*

$$(ACB + BCA) : V = 2B : V \quad \forall V \in \mathcal{C}, \tag{15}$$

and is the minimizer on $\mathcal{C}_S$ of the function

$$J(C) = \|B^{1/2}(C - A^{-1})A^{1/2}\|_F^2. \tag{16}$$

Proof. Using $c_h = C_{ij}\varphi_i \otimes \varphi_j$, $v_h = V_{kl}\varphi_k \otimes \varphi_l$ in (13), and the symmetry of A and B, we get on the one hand

$$\begin{aligned}\int\int_{\Omega\times\Omega} \nabla_x c_h.\nabla_x v_h \, dydx &= \int\int_{\Omega\times\Omega} C_{ij}V_{kl}\nabla\varphi_i(x).\nabla\varphi_k(x)\,dx\,\varphi_j(y)\varphi_l(y)\,dy \\ &= A_{ki}C_{ij}V_{kl}B_{lj} = AC : VB = ACB : V. \end{aligned} \tag{17}$$

The same computation on the integrand with ∇_y leads to

$$\int\int_{\Omega\times\Omega} \nabla c_h.\nabla v_h \, dydx = (ACB + BCA) : V.$$

On the other hand, it follows from (11), (12), and the symmetry of B that

$$\begin{aligned}\int\int_{\Omega\times\Omega} \nabla g_h.\nabla v_h \, dydx &= \int\int_{\Omega\times\Omega} \nabla g.\nabla v_h \, dydx = 2\int_\Omega v_h(x,x)\,dx \\ &= 2\int_\Omega V_{kl}\varphi_k(x)\varphi_l(x)\,dx = 2V_{kl}B_{lk} = 2B : V.\end{aligned}$$

Hence a solution c_h to (13) has a coefficients matrix C which equivalently solves (15). A solution $\widetilde{c_h}$ to the homogeneous system associated to (13) will satisfy, in particular (for $v_h = \widetilde{c_h}$),

$$\int\int_{\Omega\times\Omega} |\nabla\widetilde{c_h}|^2 dydx = 0;$$

thus $\widetilde{c_h}$ is constant, and, due to the boundary condition, $\widetilde{c_h} = 0$. This proves that (13) has a unique solution $c_h = C_{ij}\varphi_i \otimes \varphi_j \in \mathcal{W}_h$, where C is the unique solution to (15). Substituting C^T for C in (15) shows that C^T is also the solution, and hence C is symmetric. Equation (15) is the optimality condition of the minimization problem

$$\min_{C\in\mathcal{C}} \|A^{1/2}(C - A^{-1})B^{1/2}\|_F^2 + \|B^{1/2}(C - A^{-1})A^{1/2}\|_F^2,$$

which reduces on $\mathcal{C}_S$ to minimize $\|B^{1/2}(C - A^{-1})A^{1/2}\|_F^2$. ∎

The above formulation leads to a symmetric approximate Green function c_h. A nonsymmetric formulation is obtained when minimizing on the whole space $\mathcal{C}$ the same function

$$J(C) = \|B^{1/2}(C - A^{-1})A^{1/2}\|_F^2 = \int\int_{\Omega\times\Omega} |\nabla_y(g_h - c_h)|^2 \, dydx,$$

where the gradient is taken only on the variable y. We can now interpret the BCAI construction in terms of the approximate Green function: it is the nonsymmetric formulation of the minimization on $\mathcal{C}$ of the distance to g_h for the norm $\|v_h\|^2 = \int\int_{\Omega\times\Omega} |\nabla_y v_h|^2 \, dydx$ if the mass matrix B is replaced by the identity matrix (lumping process). Unfortunately, the BCAI $(C + C^T)/2$ obtained by symmetrization of the solution to the nonsymmetric formulation does not coincide in general with the solution to the symmetric formulation.

The advantage of the nonsymmetric formulation is to provide a block diagonal matrix L in the resulting system (8): the solution of this system requires fewer operations, and it is more adapted to parallel computing. Observe, however, that, when solving, for example, a two-dimensional system (10), i.e., finding $u_h \in \mathcal{V}_h$ which minimizes $\int_\Omega |\nabla(u-u_h)|^2\,dx$, it would not be recommended (if ever possible) to seek instead the function which minimizes $\int_\Omega (\partial_{x_2}(u-u_h))^2\,dx$. Hence it could be interesting to explore the potentialities of the symmetric formulation. In particular, the latter leads to the following geometric criterion, which guarantees an approximate inverse C to be positive semidefinite (and thus the preconditioner $C+\omega I$ to be SPD for $\omega > 0$). We consider here an arbitrary linear subspace $\mathcal{E} \subset \mathcal{M}_n(\mathbb{R})$ of symmetric matrices and denote $\mathcal{P} \subset \mathcal{E}$ the closed and convex cone of positive semidefinite matrices of $\mathcal{E}$. The polar subset of $\mathcal{P}$ with respect to the inner product $(V,W) = BVA : W$ is defined by $\mathcal{P}^\circ = \{V \in \mathcal{E};\ (V,W) \le 0 \text{ for all } W \in \mathcal{P}\}$.

PROPOSITION 3.2. *If $\mathcal{E}$ has the property*

$$-\mathcal{P}^\circ \subset \mathcal{P}, \tag{18}$$

then the approximate inverse C of A which minimizes (16) *on $\mathcal{E}$ is positive semidefinite.*

Proof. For all $W \in \mathcal{P}$, we have $BCA : W = B : W = \mathrm{tr}(B^{1/2} W B^{1/2}) \ge 0$, and hence $C \in -\mathcal{P}^\circ$, and, due to (18), C is positive semidefinite. ∎

In the general case, it is not clear whether property (18) can be easily checked. This property is true for all subspaces of the form $\{PSP^T;\ S \in \mathcal{M}_m(\mathbb{R}), S \text{ symmetric}\}$, where $P \in \mathcal{M}_{n,m}(\mathbb{R})$, $m \le n$, is an arbitrary matrix (see, in particular, section 5). It is usually not true for (sub)spaces containing nonsymmetric matrices.

4. Implementation. First, we describe a simple data structure used for storing a BCM and computing BCM-vector products. Next, an ordering of the unknown constants is proposed which is well adapted for solving the preconditioner system. Then we describe a basic algorithm for computing BCM-vector products. The storage of a BCM requires $\simeq 3(n+n_b)$ integers and n_b real numbers (recall that n_b is the number of constants in the BCM), while the BCM-vector product involves $\simeq n+2n_b$ arithmetic operations for large n and n_b of order n.

4.1. BCM data structure. The unknowns are organized by grids in d levels (recall that d satisfies $2^{d-2} < n \le 2^{d-1}$), the level 1 corresponding to the coarsest grid and the level d to the finest one. Four tables are used to describe a BCM:

- A table $J \in \mathcal{M}_{d,n}(\mathbb{N})$, whose row l describes the level l, and is defined by

$$\begin{aligned} &J(1,1) = n, \quad J(1,k) = 0, \quad 2 \le k \le n, \\ &J(l,k) = J(l+1,2k-1) + J(l+1,2k), \text{ and} \\ &\frac{J(l,k)}{2} - 1 < J(l+1,2k) \le \frac{J(l,k)}{2}, \quad 2 \le l \le d-1,\ 1 \le k \le n/2, \\ &J(d,k) = 1, \quad 1 \le k \le n. \end{aligned}$$

  Notice that each row l except the last one has exactly 2^{l-1} nonzero elements and that, for those elements, one has necessarily $J(l,k) \ge 2$ if $1 \le l \le d-2$.

For example, for $n = 9$, we have $d = 5$ and

$$J = \begin{bmatrix} 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{19}$$

Row 3 indicates that, at the third level, a vector $x \in \mathbb{R}^n$ will be partitioned into four groups having, respectively, $3, 2, 2, 2$ elements. Observe that $\sum_{k=1}^{n} J(l,k) = n$ for $1 \le l \le d$.

- A table $T \in \mathcal{M}_{3,n_b}(\mathbb{N})$ which describes the level and the position of the blocks of the BCM. Each column describes one block. The coefficient $T(1,k)$ defines the level of the kth block, $i = T(2,k)$ is the row number of this block, and $j = T(3,k)$ is its column number *relative to its level* $T(1,k)$. In example (2), the corresponding table T is

$$T = \begin{bmatrix} 2 & 3 & 3 & 3 & 4 & 4 & 4 & 4 & 2 & 2 \\ 1 & 1 & 1 & 2 & 3 & 3 & 4 & 4 & 2 & 2 \\ 2 & 1 & 2 & 1 & 3 & 4 & 3 & 4 & 1 & 2 \end{bmatrix},$$

  and $l_c = \min_k T(1,k)$. (Recall that l_c is the coarsest level of refinement of the BCM.)
- A vector $c \in \mathbb{R}^{n_b}$ which contains the value of each block. In the same example, it is

$$c = (c_1, c_2, \ldots, c_{10}).$$

  The size of each block is obtained from the table J. For example, the block containing c_4 has $J(T(1,4), T(2,4)) = 2$ rows and $J(T(1,4), T(3,4)) = 3$ columns.
- A vector $N_B \in \mathbb{N}^p$ which gives the number $N_B(i)$ of constant blocks in each group C_i of rows (defined below) of the matrix C which can be used for independent matrix-vector multiplication:

$$C = \begin{bmatrix} C_1 \\ \vdots \\ C_p \end{bmatrix}, \quad Cx = \begin{bmatrix} C_1 x \\ \vdots \\ C_p x \end{bmatrix}. \tag{20}$$

  Number p is defined by $p = 2^{l_c - 1}$. The submatrix C_i contains $J(l_c, i)$ rows. In the example, $l_c = 2$, $N_B = [8,\ 2]$, and one has two groups of rows, C_1 and C_2. The blocks (the c_i's and, by the way, the corresponding E_i's) are ordered in such a way that the first $N_B(1)$ blocks are in C_1, the next $N_B(2)$ blocks are in $C_2 \ldots$, and $n_b = \sum_{i=1}^{p} N_B(i)$. In particular, the size p of N_B can determine the number of processors used for parallelization.

4.2. Sparsity pattern of L. The sparsity pattern of each diagonal block L_k, $k = 1, \ldots, p$, of L (cf. section 2.3) depends on the ordering of the matrices E_i and on the sparsity pattern of A itself. The ordering of the matrices E_i can be done simultaneously with the determination of the BCM pattern in the following way. As mentioned in the previous section, the E_i's are partitioned into p groups, each of

which is associated to one of the C_k's in (20). We discuss here the numbering inside one of these groups. At the end of step 2 of Algorithm 2.1, each block K initially receives the value $s(K) = 3^{l_d - l_c}$. Then, in step 3 of the same algorithm, each K is split into (usually) four blocks:

$$K = \begin{bmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{bmatrix},$$

and the blocks $K_{i,j}$ receive the following values, where $l_c < lev \leq l_d$:

$$s(K_{1,1}) = s(K_{1,2}) = s(K) + 3^{l_d - lev}, \quad s(K_{2,1}) = s(K_{2,2}) = s(K) + 2 \times 3^{l_d - lev}.$$

Each E_i is associated to one of the blocks K obtained at the end of the algorithm and receives the value $s(E_i) := s(K)$, which is of the form

$$s(E_i) = \sum_{k=0}^{m(i)} \gamma_k 3^{l_d - l_c - k}, \quad \gamma_k = 1 \text{ or } 2, \quad 0 \leq m(i) \leq l_d - l_c.$$

Then one can check that $L_{ij} = (g_j^T A g_i)(f_j^T f_i)$ with $i \leq j$ in (9) may be nonzero only if

$$s(E_j) \leq s(E_i) < s(E_j) + 3^{l_d - l_c - m(j)}.$$

A similar relation holds for $i \geq j$. Based on this observation, the matrices E_i are decreasingly sorted with respect to $s(E_i)$, i.e., in such a way that $i \leq j$ iff $s(E_j) \leq s(E_i)$ for all i and j. In case of the equality $s(E_i) = s(E_j)$, the ordering is done from left to right. For example, with $n = 144$, $l_c = l_o = 2$, and $l_d = 6$, Figure 3 (left) shows the obtained pattern of L, where nonzero entries may be found (here $p = 2$). One can observe that this pattern is well suited for a Cholesky or LU factorization of each block L_k. Then, depending on the sparsity structure of A, some more zeros may appear, as shown in Figure 3 (right), where Poisson's matrix is used for A. Taking into account the sparsity pattern of A itself when constructing L requires more work and will not be discussed here.

The sparsity pattern of each diagonal block L_k depends on the number of levels, i.e., on the difference $l_d - l_c$. For example, Figure 4 shows the sparsity patterns obtained with $n = 144$, $l_c = l_o = 4$, and $l_d = 6$, where it can be observed that more fill-in occurs in each diagonal block (here $p = 8$). The number of operations required for computing the preconditioner will depend on p and on this fill-in. In the case in which the blocks L_k are full (which occurs if $l_c = l_d$ and A is dense), its exact computation requires $O(n_b^3/p^2)$ operations. Each subsystem of the form $L_k u = v$ can also be solved by an iterative method. Then an interesting question is to determine the accuracy needed for obtaining an efficient preconditioner.

4.3. BCM-vector product. We describe here a basic implementation of the BCM-vector product

$$y = Cx.$$

It requires $\simeq n + n_b + 2^{l_d} - 2^{l_c}$ additions and n_b multiplications. For example, the values of l_c and l_d chosen for the numerical experiments in section 6 lead for large n to $n_b \lesssim n/8$ and $2^{l_d} - 2^{l_c} = O(\sqrt{n})$. Hence, in that case, the total number of arithmetic operations is approximately $5n/4$. The algorithm is the following. For simplicity, we use MATLAB notation.

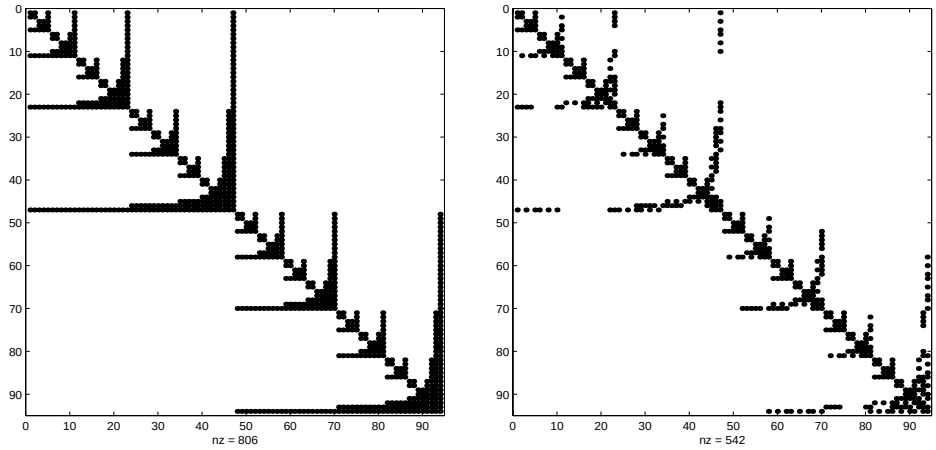


FIG. 3. *Structure of L, before and after taking into account the sparsity of A.*

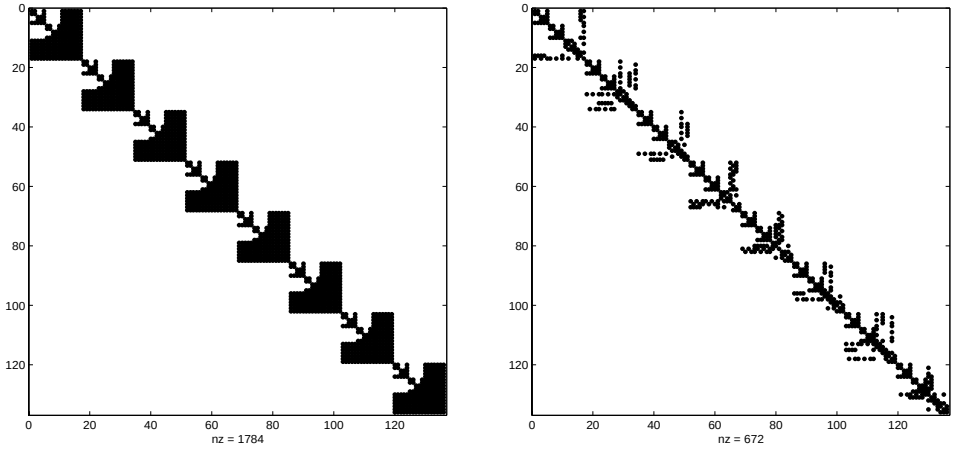


FIG. 4. *Structure of L, before and after taking into account the sparsity of A.*

ALGORITHM 4.1 (BCM-vector product).

1. *Compute the partial sums of each block of* x*, relative to each level, and, using table* J*, store them in a sparse matrix* X *of size* $d \times n$*. For example, with* $n = 9$ *and* $S_i^j = \sum_{k=i}^{j} x_k$*, the result is*

$$X = \begin{bmatrix} S_1^9 & & & & & & & & \\ S_1^5 & S_6^9 & & & & & & & \\ S_1^3 & S_4^5 & S_6^7 & S_8^9 & & & & & \\ S_1^2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & & \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{bmatrix}. \tag{21}$$

   Observe that $\sum_j X(i,j) = S_1^n$ *for all* i*.*

2. *Let* Y *be a sparse matrix of size* $(l_d - l_c + 1) \times n$ *initialized with* 0.

   `For k = 1:nb`

   - `c = T(k,4);`
   - `lev = T(1,k); i = T(2,k); j = T(3,k);`

- `w = c*X(lev,j);`
- `Y(lev,i) = Y(lev,i)+w;`

`end`

3. `For lev = lc+1:ld`
   - *transfer and add the values from row* `Y(lev-1,:)` *to row* `Y(lev,:);`

   `end`

Step 1 involves $n-1$ additions. Step 2 involves n_b multiplications and n_b additions, and step 3 involves $2^{l_d}-2^{l_c}$ additions in the usual case, where $l_d < d$. At this stage, the vector y is represented as a block constant vector in row `Y(ld,1)`, which can then be converted to full format if necessary.

5. Description of a particular case. The complete analysis of the BCP is yet to be developed. Here we study a particular case which almost reduces to a coarse grid preconditioner but which will give insight into how the method works. The two basic operations of a two-grid cycle (solution on the coarse grid followed by a residual smoother) are here performed in one single shot corresponding to the preconditioning operation $v = Mu = (C+\omega I)u$, although other choices for the smoother are possible (cf. Remark 2.1).

We consider BCMs of size $n = qm$, $q = 2^{l-1}$, where the coarsest level is equal to the finest one: $l_c = l_o = l_d = l$. The space of such BCMs made of $q \times q$ constant square blocks of size m is still denoted $\mathcal{C}$. Let P be the rectangular block diagonal matrix defined by

$$P = \begin{bmatrix} \zeta & & \\ & \ddots & \\ & & \zeta \end{bmatrix} \in \mathcal{M}_{n,q}(\mathbb{R}),$$
$$\zeta = (1,\dots,1)^T/\sqrt{m} \in \mathcal{M}_{m,1}(\mathbb{R}),$$

and let $\mathcal{P} = \mathrm{Ran}(P)$. For all $C \in \mathcal{C}$, one has $\mathrm{Ran}(C) \subseteq \mathcal{P}$. The columns of P form an orthonormal basis of $\mathcal{P}$ with $P^TP = I_q$ (the identity matrix of size q), and one has

$$\mathcal{C} = \left\{ P\,T\,P^T, T \in \mathcal{M}_{q,q}(\mathbb{R}) \right\}.$$

The reason why the BCP improves the condition number of the original system is that the space $\mathcal{P}$ captures "low frequency" components, i.e., that $\mathcal{P}$ is a good approximation of the space spanned by the eigenvectors associated to the smallest eigenvalues of A. This property relies on the fact that A is issued from the discretization of an elliptic PDE and depends on the way A has been reordered; low-frequency components can be correctly captured only if the unknowns have been grouped by proximity like in the multipole approach (cf. experiments of section 6.4).

Problem (5) becomes here

$$\min_{T \in \mathcal{M}_{q,q}(\mathbb{R})} \|(P\,T\,P^T - A^{-1})A^{1/2}\|_F^2,$$

and its unique solution is characterized by

$$(P\,T\,P^TA - I_n) : P\,S\,P^T = 0 \quad \forall S \in \mathcal{M}_{q,q}(\mathbb{R}). \tag{22}$$

Using $P^TP = I_q$, we have $(P\,T\,P^TA - I_n) : P\,S\,P^T = (\,T\,P^TAP - I_q) : S$, and (22) becomes $T\,P^TAP : S = I_q : S$ for all $S \in \mathcal{M}_{q,q}(\mathbb{R})$, from which it follows that $T = (P^TAP)^{-1}$ and the BCAI of A is

$$C = P\,(P^TAP)^{-1}\,P^T. \tag{23}$$

Finally, the BCP is given by

$$M = P\left(P^T AP\right)^{-1} P^T + \omega I_n, \tag{24}$$

where ω is related to the smoothing operation (cf. Remark 2.1). The next properties are well known in multigrid methods [25]. They derive directly from $\mathrm{Ran}(C) \subseteq \mathcal{P}$ and (23) (see also Proposition 3.2).

PROPOSITION 5.1. *If A is SPD, then the following properties hold:*

- *The subspace $\mathcal{P}$ is invariant under CA, with $CAx = x$ for all $x \in \mathcal{P}$.*
- *If $\mathcal{Q} \subset \mathbb{R}^n$ denotes the orthogonal space to $\mathcal{P}$, then $CAx = 0$ for all $x \in A^{-1}\mathcal{Q}$.*
- *The eigenvalues of CA are 1 (with multiplicity q) and 0 (with multiplicity $n - q$).*
- *The matrix C is SPD, and, for all $\omega > 0$, M is SPD, and the eigenvalues of MA are positive.*

For estimating the condition number of MA, let $P_A \in \mathcal{M}_{n,q}(\mathbb{R})$ and $Q_A \in \mathcal{M}_{n,n-q}(\mathbb{R})$ be two matrices whose columns form, respectively, an orthonormal basis of $A^{1/2}\mathcal{P}$ and $A^{-1/2}\mathcal{Q}$; let $R_A = [P_A, Q_A]$,

$$R_A^T R_A = I_n; \tag{25}$$

and consider the similar matrix

$$M_A = (A^{-1/2}R_A)^{-1} MA (A^{-1/2}R_A).$$

It follows from the above proposition and (25) that

$$\begin{aligned}(A^{-1/2}R_A)^{-1} CA (A^{-1/2}R_A) &= (A^{-1/2}R_A)^{-1} CA [A^{-1/2}P_A,\, A^{-1/2}Q_A] \\ &= (A^{-1/2}R_A)^{-1} [A^{-1/2}P_A,\, 0] \\ &= R_A^T\, [P_A,\, 0] \\ &= \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix}.\end{aligned}$$

We also have

$$\begin{aligned}(A^{-1/2}R_A)^{-1} A (A^{-1/2}R_A) &= R_A^{-1} A R_A = R_A^T A R_A \\ &= \begin{bmatrix} P_A^T AP_A & P_A^T AQ_A \\ Q_A^T AP_A & Q_A^T AQ_A \end{bmatrix} \\ &= \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{1,2}^T & A_{2,2} \end{bmatrix},\end{aligned}$$

where $A_{1,1}$ and $A_{2,2}$ are both SPD. (Recall that A is supposed to be SPD.) Hence the matrix of the preconditioned system is similar to

$$M_A = \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} + \omega \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{1,2}^T & A_{2,2} \end{bmatrix}.$$

We use standard notation for the Euclidean inner product $x.y$, the associated 2-norm, and the induced norm on matrices.

LEMMA 5.2. *We have the following inequality: $\|A_{1,2}\|^2 \leq \|A_{1,1}\|\,\|A_{2,2}\|$.*

Proof. It follows from $R_A^T A R_A x.x \geq 0$ for all $x = (tu, v) \in \mathbb{R}^n$ that

$$t^2 A_{1,1}u.u + 2tA_{1,2}v.u + A_{2,2}v.v \geq 0 \quad \forall (t, u, v) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^{n-q}.$$

The discriminant of this polynomial is nonpositive; thus

$$(A_{1,2}v.u)^2 \leq (A_{1,1}u.u)(A_{2,2}v.v) \quad \forall (u,v) \in \mathbb{R}^q \times \mathbb{R}^{n-q},$$

which yields the required inequality. ∎

The (spectral) condition number $\lambda_{\max}(MA)/\lambda_{\min}(MA)$ of MA is denoted $\mathrm{cond}(MA)$.

PROPOSITION 5.3. *Let $0 < \lambda \leq \|A\|$ be the largest eigenvalue of $A_{1,1}$, and let $0 < \gamma \leq \|A\|$ be the smallest eigenvalue of $A_{2,2}$. Then*

$$\mathrm{cond}(MA) \leq 1 + \frac{2}{\gamma} + \frac{4\lambda}{\gamma^2}.$$

In particular, if there exist $c_1, c_2 > 0$, and $0 \leq r \leq s$ such that $\gamma \geq c_1 h^s$ and $\lambda \leq c_2 h^r$, then there exists $c > 0$ such that, for all $0 < h \leq 1$,

$$\mathrm{cond}(MA) \leq c h^{r-2s}.$$

Proof. For convenience, we can suppose that the SPD matrix A has been normalized; i.e., $\|A\| = 1$, which has no incidence on $\mathrm{cond}(MA)$. First, we estimate a lower bound for the smallest eigenvalue of M_A. Let $\beta = \|A_{1,2}\|$. For $x = (u,v) \in \mathbb{R}^q \times \mathbb{R}^{n-q}$, we have

$$\begin{aligned} M_A x.x &= \|u\|^2 + \omega(A_{1,1}u.u + 2A_{1,2}v.u + A_{2,2}v.v) \\ &\geq \|u\|^2 - 2\omega\beta \|u\| \|v\| + \omega\gamma \|v\|^2. \end{aligned}$$

For $\eta > 0$, the classical inequality $2ab \leq a^2/\eta + \eta b^2$ leads to

$$M_A x.x \geq \left(1 - \frac{\omega\beta}{\eta}\right) \|u\|^2 + \omega(\gamma - \eta\beta) \|v\|^2.$$

Next, choosing η such that $\eta\beta = \gamma/2$ and ω such that $1 - \omega\beta/\eta = \omega(\gamma - \eta\beta)$, i.e., $\omega = 2\gamma/(4\beta^2 + \gamma^2)$, yields the lower bound

$$\begin{aligned} M_A x.x &\geq \frac{\gamma^2}{4\beta^2 + \gamma^2}(\|u\|^2 + \|v\|^2) \\ &= \frac{\gamma^2}{4\beta^2 + \gamma^2} \|x\|^2. \end{aligned}$$

An upper bound for the largest eigenvalue of M_A is given by

$$\begin{aligned} M_A x.x &= \|u\|^2 + \omega R_A^T A R_A x.x \\ &\leq (1 + \omega) \|x\|^2. \end{aligned}$$

Gathering these two bounds, we obtain the next upper bound for $\mathrm{cond}(MA) = \mathrm{cond}(M_A)$:

$$\begin{aligned} \mathrm{cond}(MA) &\leq (1 + \omega) \frac{4\beta^2 + \gamma^2}{\gamma^2} \\ &= \frac{4\beta^2 + \gamma^2 + 2\gamma}{\gamma^2}. \end{aligned}$$

It follows from Lemma 5.2 that $\beta^2 = \|A_{1,2}\|^2 \le \|A_{1,1}\| \, \|A_{2,2}\| \le \|A_{1,1}\| = \lambda$. Hence

$$\begin{aligned} \mathrm{cond}(MA) &\le \frac{4\lambda + \gamma^2 + 2\gamma}{\gamma^2} \\ &\le 1 + \frac{4c_2 h^r + 2c_1 h^s}{c_1^2 h^{2s}} \le c h^{r-2s} \end{aligned}$$

with $c = 1 + (4c_2 + 2c_1)/c_1^2$. ∎

This proposition expresses the fact that the better $\mathrm{Ran}(P_A)$ and $\mathrm{Ran}(Q_A)$ approach the eigenspaces associated, respectively, to the smallest and largest eigenvalues of A, the better will be the condition number of MA. The ideal case would correspond to $\mathrm{Ran}(P_A) = \mathrm{span}\{v_1, v_2, \dots, v_m\}$ and $\mathrm{Ran}(Q_A) = \mathrm{span}\{v_{m+1}, v_{m+2}, \dots, v_n\}$, where $(v_i)_{i=1}^n$ denotes an orthonormal basis of eigenvectors associated to the eigenvalues of A increasingly ordered $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. In that case, if m was chosen such that $\lambda_m \simeq c_1 h^r$, for example, $m \simeq c_2 n^{1-r/2}$ for the two dimensional Poisson equation, then we would have $\lambda = \lambda_m$, $\gamma = \lambda_{m+1}$, $s = r$, and $\mathrm{cond}(MA) \le c h^{-r}$, which is the best that we can hope.

Finally, we examine the case of the five-point finite difference matrix A_n of the Laplacean operator in a square with a homogeneous Dirichlet boundary condition. The mesh size is $h = 1/(\widetilde{n}+1)$, and there are $\widetilde{n}^2 = n$ interior nodes $(x_i, y_j) = (ih, jh)$. We suppose that $\widetilde{n} = \widetilde{q}\widetilde{m}$ with $q = \widetilde{q}^2$, $m = \widetilde{m}^2$. The recursive coordinate bisection leads simply to grouping together the m nodes (x_i, y_j), $k_1\widetilde{m} + 1 \le i \le (k_1+1)\widetilde{m}$, $k_2\widetilde{m} + 1 \le j \le (k_2+1)\widetilde{m}$ for $0 \le k_1, k_2 \le \widetilde{q} - 1$. The reordering of the nodes is given by the order in which they appear when executing the four loops on k_2, k_1, j, and i respectively, from the outer-most to the inner-most loop. The reordered matrix A is of the form $A = \Pi^T A_n \Pi$, where Π is the permutation matrix associated to this reordering. With this choice of Π, the columns of P span the space of discrete functions, which are constant on each subset of the form $\{(x_i, y_j);\, k_1\widetilde{m} + 1 \le i \le (k_1+1)\widetilde{m}, k_2\widetilde{m} + 1 \le j \le (k_2+1)\widetilde{m}\}$.

PROPOSITION 5.4. *The BCAI of A is given by*

$$C = \widetilde{m} P A_q^{-1} P^T,$$

where A_q is the five-point finite difference matrix for the Laplacean, issued from a coarser mesh with mesh size $\widetilde{h} = 1/(\widetilde{q}+1) \simeq \widetilde{m}h$.

Proof. The exact expression of ΠP is given by

$$\begin{aligned} \widetilde{\zeta} &= (1, \dots, 1)^T/\widetilde{m} \in \mathbb{R}^{\widetilde{m}}, \\ \widetilde{P} &= \begin{bmatrix} \widetilde{\zeta} & & \\ & \ddots & \\ & & \widetilde{\zeta} \end{bmatrix} \in \mathcal{M}_{\widetilde{n},\widetilde{q}}(\mathbb{R}), \\ \widetilde{U} &= (\widetilde{P}^T, \dots, \widetilde{P}^T)^T \in \mathcal{M}_{\widetilde{m}\widetilde{n},\widetilde{q}}(\mathbb{R}), \\ \Pi P &= \begin{bmatrix} \widetilde{U} & & \\ & \ddots & \\ & & \widetilde{U} \end{bmatrix} \in \mathcal{M}_{n,q}(\mathbb{R}). \end{aligned}$$

A direct computation of $P^T A P = (\Pi P)^T A_n (\Pi P)$ gives $P^T A P = A_q/\widetilde{m}$, and the result follows from (23). ∎

Remark 5.1. One has $P^TAP = A_q/\widetilde{m}$ and $P_A^TAP_A = A_{1,1}$. The two matrices P and P_A (whose columns are orthonormal) span, respectively, $\mathcal{P}$ and $A^{1/2}\mathcal{P}$, which are close each other. The eigenvalues of $A_q/\widetilde{m}$ are given by

$$\lambda_{jk} = [4(\sin^2(j\pi\widetilde{h}/2) + \sin^2(k\pi\widetilde{h}/2))]/\widetilde{m}, \quad 1 \le j,k \le \widetilde{q},$$

and, in particular, $\lambda_{\max}(A_q/\widetilde{m}) < 8/\widetilde{m}$. Suppose now that $\widetilde{q} = \widetilde{m}$. Then $\lambda_{\max}(A_q/\widetilde{m}) \simeq 8h^{1/2}$, and one can expect that $\lambda = \lambda_{\max}(A_{1,1}) \simeq c_2h^{1/2}$. Similarly, we have numerically observed that $\gamma = \lambda_{\min}(A_{2,2}) \simeq c_2h$. Then, having here $r \simeq 1/2$ and $s \simeq 1$, it follows from Proposition 5.3 that $\mathrm{cond}(MA) = O(h^{-3/2})$, which is confirmed by numerical tests in section 6.5. In that case, the number of constants involved in the BCAI is $q^2 = n$.

6. Numerical examples. The numerical examples presented in this section illustrate the general method described in section 2. They aim to give a partial answer to the following questions:

- How does the BCP compare with other standard preconditioners?
- How do variable or discontinuous coefficients in the PDE operator affect the BCP?
- What is the influence of reordering before computing the BCP?
- What is the effect of the variable block size in the BCP?

At the present time, the BCP has only been written in the interpreted language MATLAB; thus CPU time information is not relevant. For that reason, we will indicate only memory storage and the number of iterations. That should foreshadow what will happen when implementing the BCP in a compiled language.

In our experiments, the nodes of the mesh were reordered by using the recursive coordinate bisection algorithm [28]; i.e., they were partitioned by proximity into two sets along the x axis, and each of them was partitioned into 2 sets along the y axis, and then again along the x axis until all sets contained a single point. The corresponding and simultaneous renumbering consists at each step in numbering the nodes of the first set before those of the second set. Unless otherwise specified, the following parameters were used in the BCP:

```
omega = 1.5/normest(A);
d = ceil(log2(n))+1;
lc = ceil(d/2)-2;
lo = lc+3; ld = lo+2.
```

The MATLAB command `normest(A)` computes an approximation of $||A||$, `ceil(x)` is the smallest integer $i \ge x$, and `omega, d, lc, ld,` and `lo` denote, respectively, ω, d, l_c, l_d, and l_o (cf. section 2).

6.1. Poisson's equation in a square. We start with Poisson's equation in a square with a homogeneous Dirichlet boundary condition, solved by using the five-point finite difference stencil. Figure 5 illustrates the convergence history (residual 2-norms) for four matrices of increasing size $n = 900,\ 10000,\ 40000,\ 160000$. The BCP is compared to the SAI(1) preconditioner [18] (the same pattern as A for the SAI, symmetrized in the same way as the BCP) and the IC(0) preconditioner. The incomplete Cholesky factorization is of the form $A = U^TU + R$. The three preconditioners are used with the conjugate gradient algorithm. At the top of each figure is indicated the amount of storage (number of nonzero elements) which was used by each preconditioner.

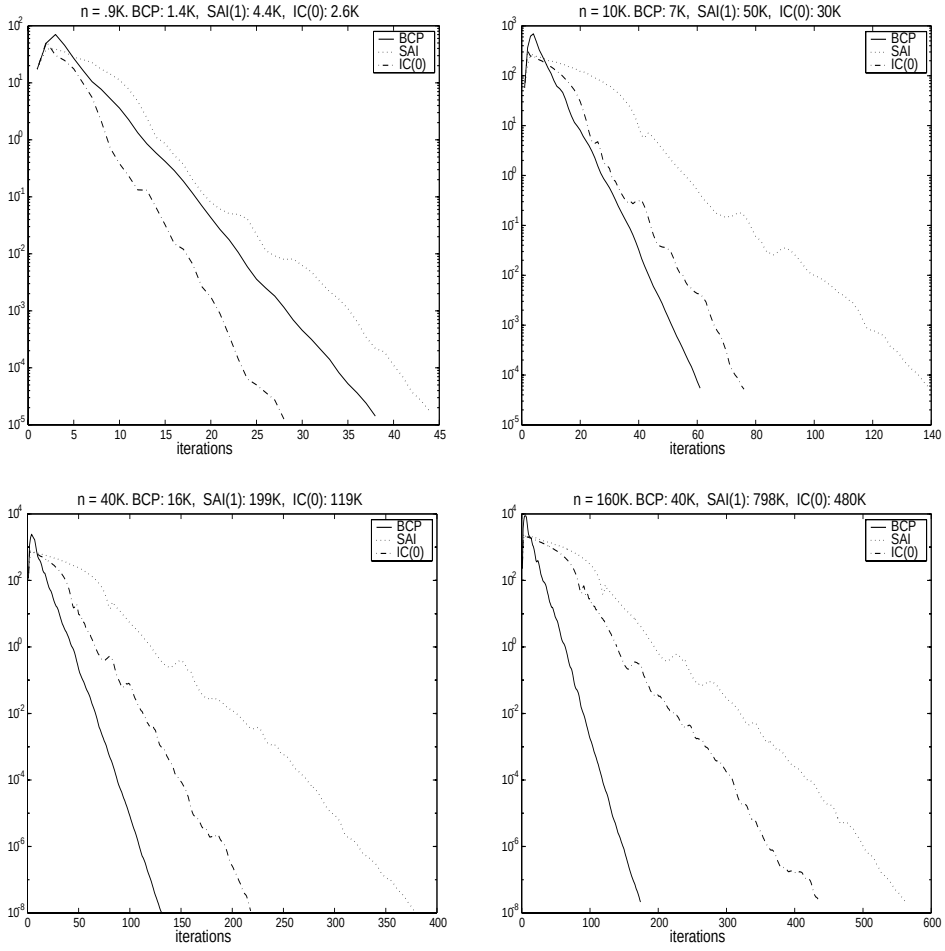


FIG. 5. *Comparison of BCP, SAI, IC(0) for* $n = 900$, 10000, 40000, 160000.

One observes that the number of iterations with the BCP seems to be of order

$$O(h^{-1/2}) = O(n^{1/4}).$$

This is confirmed by Figure 6, which plots the number of iterations needed for obtaining a given precision (relative residual norm $\leq 10^{-6}$) versus the size of the system. The results with the BCP match the line $y = 6\,n^{1/4}$, and the results with IC(0) match the line $y = 2.5\,n^{0.35}$. The associated values of n_b (number of constant blocks) and `nnz(U)` (number of nonzero elements in the incomplete Cholesky factor) are given in Table 1. They indicate the amount of storage used by the preconditioners, and they satisfy asymptotically

$$n_b \lesssim \frac{n}{8}, \quad \texttt{nnz(U)} \simeq 3\,n.$$

Table 2 reports the condition numbers $\lambda_{\max}/\lambda_{\min}$ of the matrices A, MA, and $U^{-T}AU^{-1}$ for different values of n, with $M = C + \omega I$ and $U = \mathrm{IC}(0)$ as defined above. We used MATLAB's function `eigs` for computing the eigenvalues. One can observe that

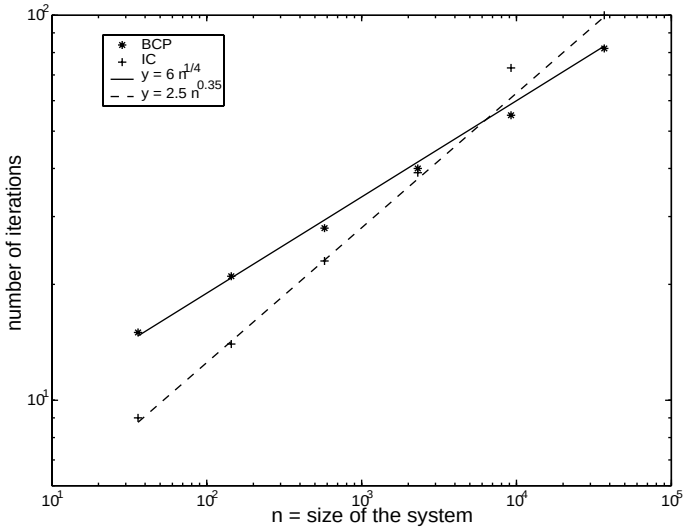


FIG. 6. *Number of iterations with respect to the size n.*

TABLE 1
Storage used by the preconditioners.

n	36	144	576	2300	9200	36900
n_b	208	616	1300	2900	6300	15000
`nnz(U)`	90	408	1700	6900	27600	111000

TABLE 2
Condition numbers.

n	n_b	cond(A)	cond(MA)	$\sqrt{n}$	cond($U^{-T}AU^{-1}$)
625	1480	273	23	25	25
2500	3220	1053	38	50	93
10000	7024	4133	65	100	366
40000	16048	16373	119	200	1448

the condition number of MA is of order $O(\sqrt{n}) = O(h^{-1})$. Finally, Figure 7 shows the distribution of their eigenvalues for $n = 900$. They are increasingly sorted and normalized in such a way as to have the maximum value 1.

6.2. Effect of a variable coefficient in the operator. Now we solve the equation with variable coefficient $a(x)$,

$$\begin{cases} -\text{div}\,(a(x)\nabla u(x)) = f & \text{on} \quad \Omega, \\ u = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{26}$$

on the domain $\Omega \subset]-1.5, 1.5[\times]-1, 1[$ described in Figure 8 by using Lagrange's finite elements of degree 1. The right-hand side f was chosen randomly.

Figure 9 reports the results obtained with a constant coefficient $a(x) = 1$ (left) and a variable coefficient (right) defined by

$$a(x) = 1 + 100\, x_1^2 + x_2^2.$$

As before, the sizes of the preconditioners are indicated at the top of each figure.

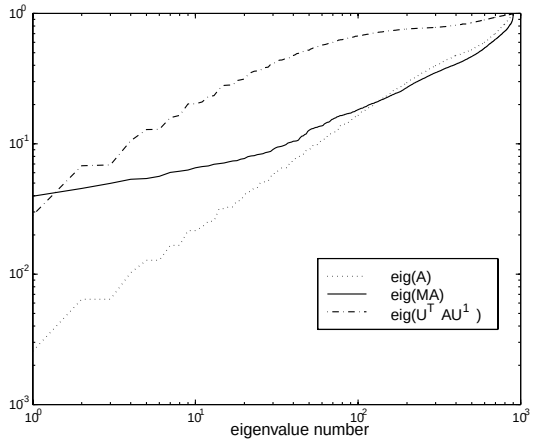


FIG. 7. *Eigenvalue distribution.*

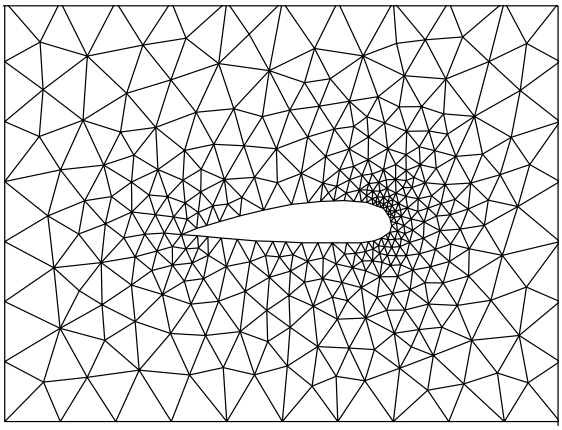


FIG. 8. *Domain* Ω.

The effect of a variable coefficient is negligible in this example for all three preconditioners. However, for the BCP, it is related to the fact that, before computing the preconditioners, the matrix has been first "diagonal-normalized," i.e., modified into the matrix $D^{-1}AD^{-1}$ with D diagonal, $D_{ii} = \sqrt{A_{ii}}$. Hence the resulting matrix has 1's on its main diagonal. When the matrix is not normalized (in the case of a variable coefficient $a(x)$), then the current implementation of BCP does not work very well, as illustrated by Figure 10. In that case, it could be better to use a BCP of the form $C + D$ with D diagonal; this technique is not very far from normalizing the matrix.

6.3. Effect of a discontinuous coefficient in the operator. Now we solve (26) on the square $\Omega =]-1, 1[^2$ with $f = 1$ and a discontinuous coefficient

$$a(x) = \begin{cases} 1 & \text{if } x_1 < 0 \text{ and } x_2 < 0, \\ 10^3 & \text{if } x_1 x_2 < 0, \\ 10^6 & \text{if } x_1 > 0 \text{ and } x_2 > 0. \end{cases} \tag{27}$$

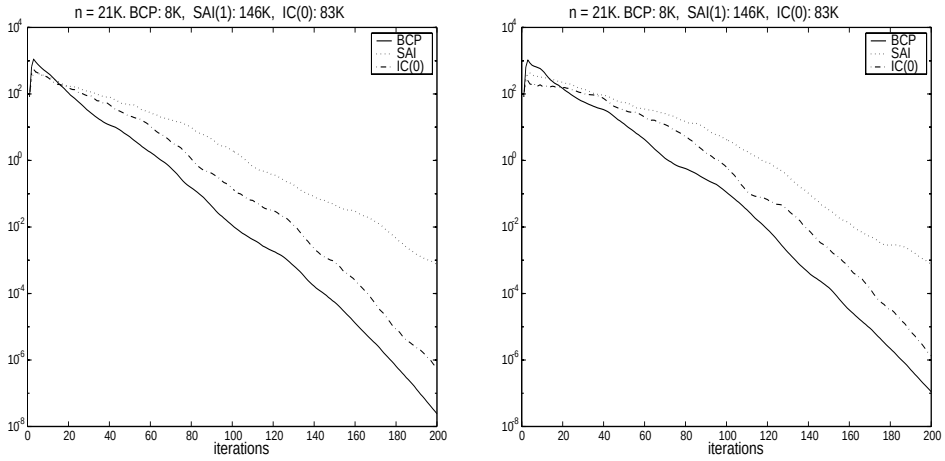


FIG. 9. *Left: constant a. Right: variable* $a(x)$.

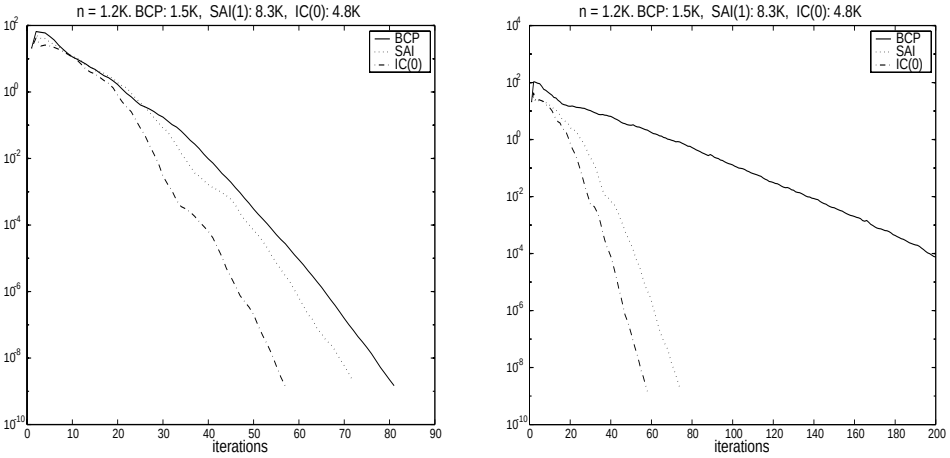


FIG. 10. *Variable* $a(x)$. *Left: with normalization. Right: without normalization.*

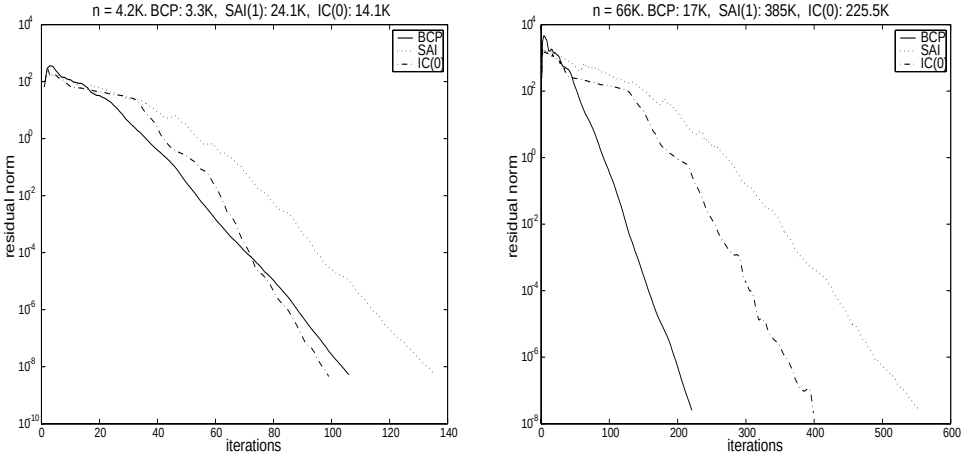


FIG. 11. *Discontinuous* $a(x)$ *from* (27).

Figure 11 shows the results obtained with two mesh sizes, with, respectively, $n = 4200$ and $n = 66000$. The matrix A has been normalized as described in the previous section. The ordering produced by the recursive coordinate bisection groups together nodes associated to the same value $a(x)$ (with some few exceptions), which explains the good convergence of the BCP.

If we change the location of the discontinuity by taking, for example,

$$a(x) = \begin{cases} 1 & \text{if } x_1 + x_2 < 0 \text{ and } x_1 - x_2 < 0, \\ 10^3 & \text{if } (x_1 + x_2)(x_1 - x_2) < 0, \\ 10^6 & \text{if } x_1 + x_2 > 0 \text{ and } x_1 - x_2 > 0, \end{cases} \tag{28}$$

which corresponds to a rotation of $\pi/4$, then the recursive coordinate bisection produces groups of nodes associated to two different values of $a(x)$ (along the two main diagonals of Ω), and BCP is less competitive, as illustrated by Figure 12.

6.4. Effect of reordering. We continue with the same test case (Figure 8) with a constant coefficient $a(x) = 1$. No reordering was used in the IC(0) preconditioner. Reordering does not affect the SAI preconditioner, but, as expected from its conception, it has a great influence on the BCP efficiency, as shown by Figure 13. When the nodes are reordered by proximity, low-frequency components are correctly captured by the preconditioner, and the convergence is driven by the remaining eigenvalues. Without reordering, the preconditioner captures some intermediate eigenvalues (in the two-dimensional case and for a natural ordering, modes which are low frequency in one direction and high frequency in the other), which does not help the convergence very much, unless precisely only those modes are present in the solution.

6.5. Effect of refining the blocks' size. In order to measure the influence of refining the size of the BCP blocks, we compare the results of a refined BCP ($l_c < l_d$) to the one obtained without refinement ($l_c = l_d$). We still use the same test case (Figure 8) with $a(x) = 1$. For a fair comparison, the parameters $l_c \leq l_o \leq l_d$ have been adjusted in such a way that the number of constants involved in each situation is about the same. This implies that level l_c is smaller with refinement than without refinement.

On the top of each figure from Figure 14, `m1` and `m2` denote, respectively, the number n_b of constant blocks used without and with refinement. The case without refinement is almost the two-grid method with a mesh size of order $\sqrt{h}$ for the coarse grid if h is the size of the mesh elements of the fine grid (cf. section 5). The four figures use the successive values h, $h/2$, $h/4$, and $h/8$. One can observe that the BCP with refinement performs better than that without and that the difference increases together with the dimension of the system.

This difference of convergence between the two variants of the BCP which involve approximately the same storage and computation time, is related to different asymptotic condition numbers $\lambda_{\max}/\lambda_{\min}$, illustrated by Figure 15, which also shows for reference the condition numbers of A and IC(0). As before, the extreme eigenvalues were computed with MATLAB's function `eigs`.

The condition number asymptotic regime for the matrices A, $U^{-T}AU^{-1}$, M_1A without refinement ($l_c = l_d$) and M_2A with refinement ($l_c < l_d$) is also reported in Table 3. The condition number of MA is significantly improved when allowing the blocks' size to vary, i.e., when using the multipole approach.

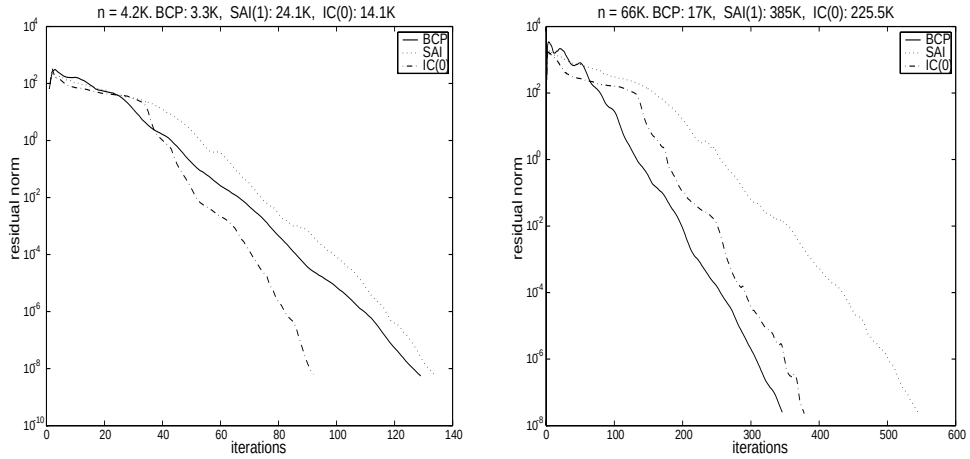


FIG. 12. *Discontinuous $a(x)$ from* (28).

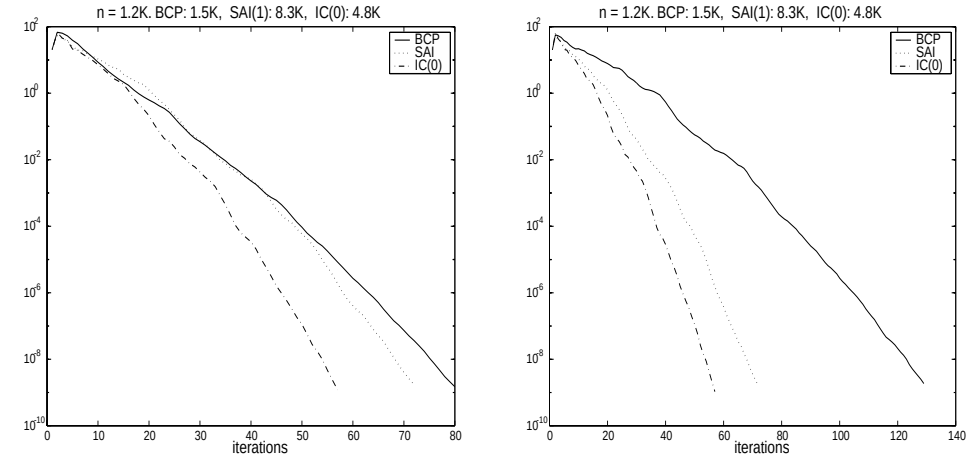


FIG. 13. *Left: with reordering. Right: without reordering.*

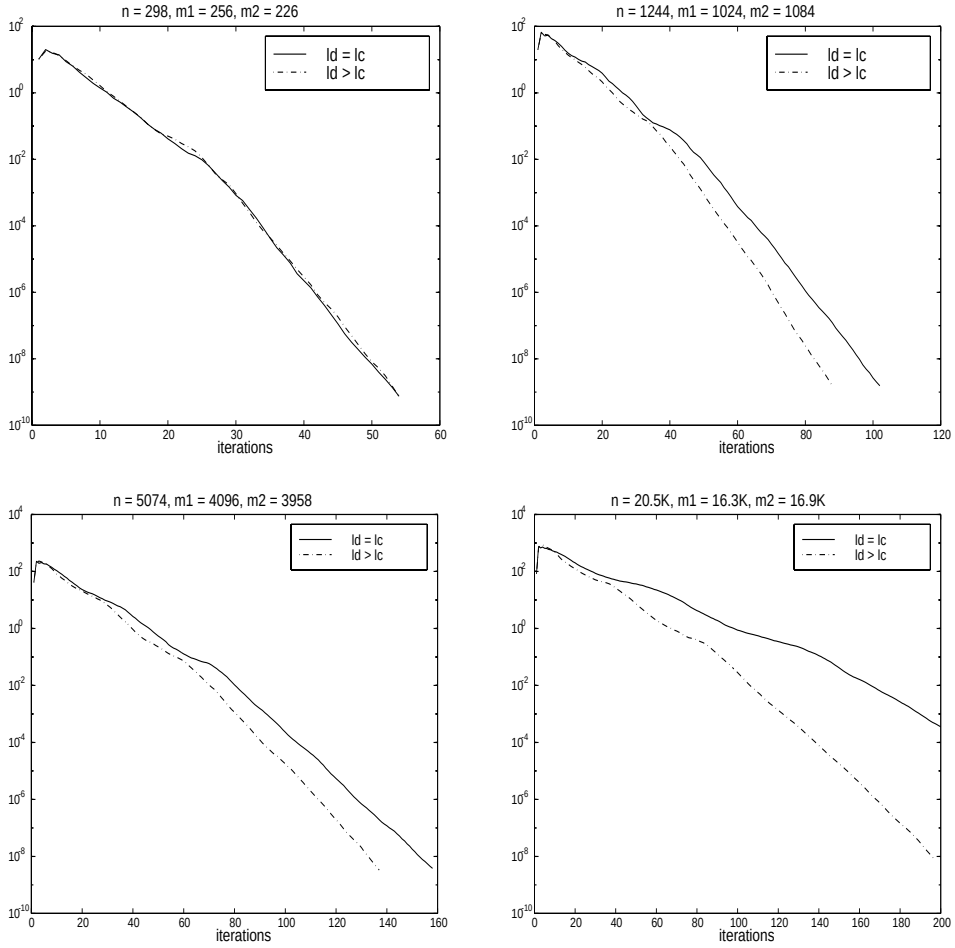


FIG. 14. *Without* `(m1)` *or with* `(m2)` *refining the blocks for* h, $h/2$, $h/4$, $h/8$.

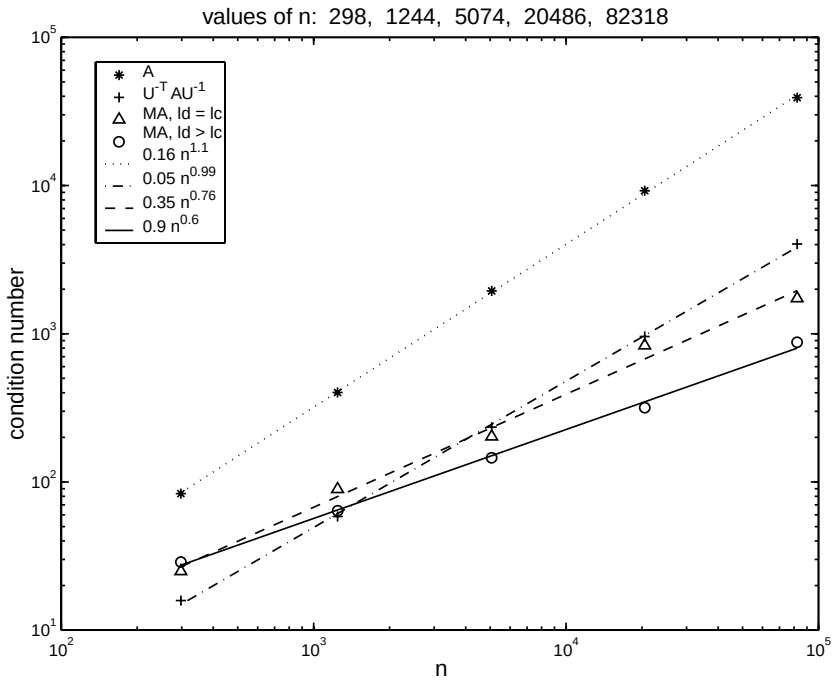


FIG. 15. *Condition number without or with refining the blocks.*

TABLE 3
Condition numbers.

$\mathrm{cond}(A)$	$\mathrm{cond}(U^{-T}AU^{-1})$	$\mathrm{cond}(M_1A)$	$\mathrm{cond}(M_2A)$
$O(n)$	$O(n)$	$O(n^{3/4})$	$O(n^{0.6})$

6.6. Example in mechanics. We finally consider a PDE system instead of a scalar PDE, namely, the two-dimensional plain stress equations. In solid and structural mechanics, the displacement field u is the solution to the equilibrium equation

$$\left\{\begin{array}{lll} -\operatorname{div}\sigma(u) = f & \text{in} & \Omega, \\ u = 0 & \text{on} & \Gamma_D, \\ \sigma(u)\mathbf{n} = g & \text{on} & \Gamma_N, \end{array}\right. \tag{29}$$

where f and g are, respectively, volume and boundary forces and $\mathbf{n}$ is the unit outward normal to the boundary $\Gamma = \Gamma_D \cup \Gamma_N$ of a domain Ω.

In the linear homogeneous isotropic elasticity model, the stress tensor $\sigma(u)$ is related to the strain tensor $\varepsilon(u) = (Du + Du^T)/2$ by

$$\sigma(u) = 2\mu\varepsilon(u) + \lambda \mathrm{div}(u) I,$$

where λ and $\mu > 0$ are the Lamé coefficients and I denotes the identity matrix of size 2.

We consider here the case of a rectangular cantilever clamped on the left side, submitted to the load $g = (0, -1)$ on the right side (see Figure 16), and to the homogeneous Neumann boundary condition on the rest of the boundary.

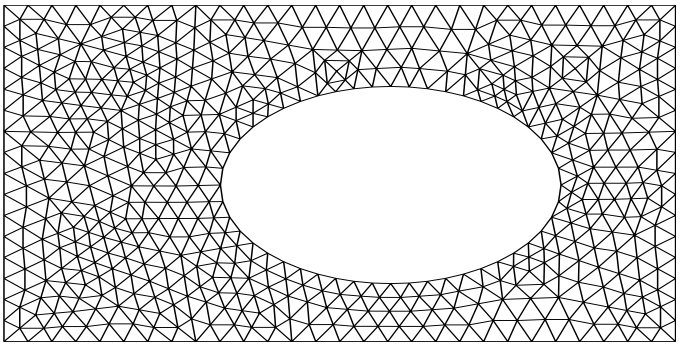


FIG. 16. *Perforated cantilever.*

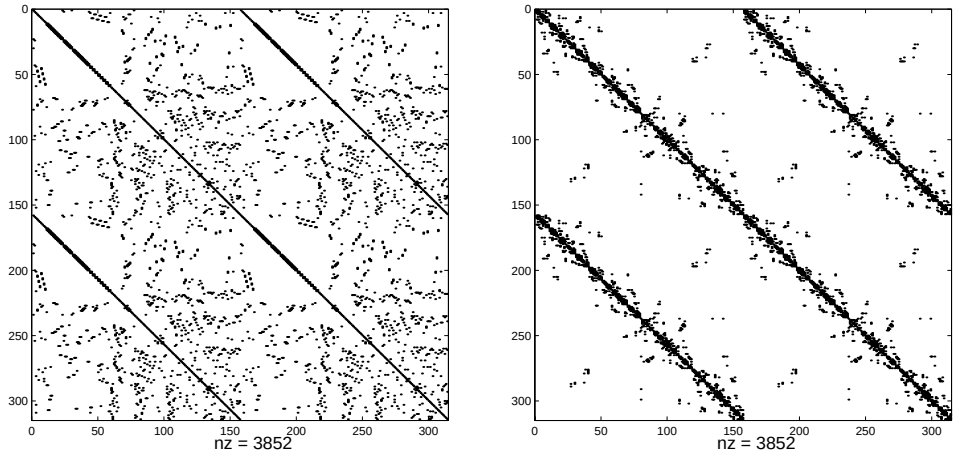


FIG. 17. *Plain stress matrix before and after reordering.*

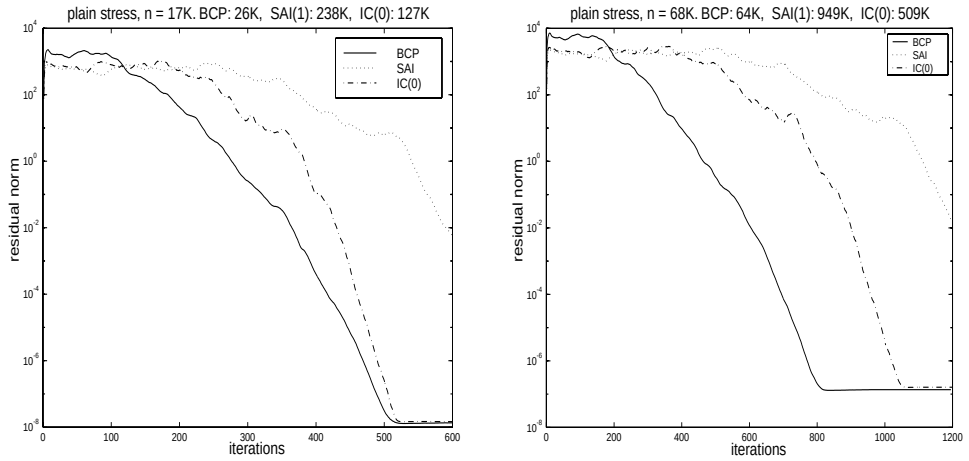


FIG. 18. *Comparison of BCP, SAI, IC(0) for* $n = 17000, 68000$.

The same reordering as in the previous sections is applied to the first component v of the displacement field $u = (v, w)$. The obtained permutation p is also applied to the second component w. The resulting reordered displacement field is given by $((v_{p(i)})_{i=1}^{n/2}, (w_{p(i)})_{i=1}^{n/2})$ if n is the total number of unknowns. It would be ineffective to use the alternate reordering $(v_{p(i)}, w_{p(i)})_{i=1}^{n/2}$ because there is no reason for v and w to have the same value on a given node. Figure 17 shows the sparsity pattern of the matrix before and after reordering. Figure 18 compares the BCP, SAI, and IC(0) for $n = 17000$ and $n = 68000$. The results are comparable to those obtained with Poisson's equation.

Conclusion and prospect. The new BCP is at the intersection of multipole, multigrid, and SAI methods. Its computation as well as its application on a vector is highly parallelizable. Our experiments show that the BCP takes the advantage over two basic preconditioners when the size of the system increases, both in terms of memory requirements and number of iterations, the costs per iterations being comparable. The next step will be to implement the method in a compiled language. The complete analysis of the condition number of the preconditioned system remains to be developed and will be the object of a future work. Finally, it could be interesting to explore piecewise linear matrices instead of BCMs (i.e., on each block of the form $c_{ij} = ai + bj + c$ or $a.x_i + b.x_j + c$ if the nodal table is available, instead of $c_{ij} = c$), which may be a good approach for oscillatory Green functions arising, for example, from Helmholtz equations.

Acknowledgment. The authors are grateful to the referees for their valuable comments and suggestions, which have helped to improve the presentation of this work.

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.
[2] M. W. Benson and P. O. Frederickson, *Iterative solution of large sparse linear systems arising in certain multidimensional approximation problems*, Util. Math., 22 (1982), pp. 127–140.
[3] J. H. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.
[4] R. Bramley and V. Meñkov, *Low Rank Off-Diagonal Block Preconditioners for Solving Sparse Linear Systems on Parallel Computers*, Tech. Rep. 446, Department of Computer Science, Indiana University, Bloomington, IN, 1996.
[5] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 669–686.
[6] E. Chow and Y. Saad, *Approximate inverse preconditioners via sparse-sparse iterations*, SIAM J. Sci. Comput., 19 (1998), pp. 995–1023.
[7] P. Concus, G. Golub, and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 220–252.
[8] J. D. F. Cosgrove and J. C. Dias, *Fully parallel preconditionings for sparse systems of equations*, in Proceedings of the Second Workshop on Applied Computing, S. Uselton, ed., The University of Tulsa, Tulsa, OK, 1988, pp. 29–34.
[9] J. D. F. Cosgrove, J. C. Dias, and A. Griewank, *Approximate inverse preconditionings for sparse linear systems*, Int., J. Comput. Math., 44 (1992), pp. 91–110.
[10] E. Darve, *Méthodes multipôles rapides: Résolution des équations de Maxwell par formulations intégrales*, Thèse de doctorat, l'Université Paris 6, Paris, France, 1999.
[11] I. S. Duff, A. M. Erisman, C. W. Gear, and J. K. Reid, *Sparsity structure and Gaussian elimination*, ACM SIGNUM Newsletters, 23 (1988), pp. 2–8.
[12] T. F. Dupont, R. P. Kendall, and H. H. Rachford, *An approximate factorization procedure for solving self-adjoint elliptic difference equations*, SIAM J. Numer. Anal., 5 (1968), pp. 559–573.

[13] N. Engheta, W. D. Murphy, V. Rokhlin, and M. S. Vassiliou, *The fast multipole methode (fmm) for electromagnetic scattering problems*, IEEE Trans. Antennas Propag., 40 (1992), pp. 634–641.
[14] R. P. Fedorenko, *On the speed of convergence of an iteration process*, USSR Comput. Math. Math. Phys., 4 (1964), pp. 227–235.
[15] L. Giraud, *On the Numerical Solution of Partial Differential Equations: Iterative Solvers for Parallel Computers*, Habilitation à Diriger des Recherches, I.N.P. de Toulouse, Toulouse, France, 2000.
[16] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, 1997.
[17] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.
[18] M. Grote and H. S. Simon, *Parallel preconditioning and approximate inverse on the Connection Machine*, in Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing, R. Sincovec, ed., SIAM, Philadelphia, 1993, pp. 519–523.
[19] I. Gustafsson, *A class of 1st order factorization methods*, BIT, 18 (1978), pp. 142–156.
[20] W. Hackbusch and B. N. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic: General complexity estimates*, J. Comput. Appl. Math., 125 (2000), pp. 479–501.
[21] M. R. Hestenes and E. L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[22] L. Yu. Kolotina and A. Yu. Yeremin, *On a family of two-level preconditionings of the incomplete block factorization type*, Sov. J. Numer. Anal. Math. Modelling, 1 (1986), pp. 292–320.
[23] M. H. Lallemand, H. Steve, and A. Dervieux, *Unstructured multigridding by volume agglomeration: Current status*, Comput. & Fluids, 21 (1992), pp. 397–433.
[24] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.
[25] G. Meurant, *Computer Solution of Large Linear Systems*, North–Holland, Amsterdam, 1999.
[26] A. van de Ploeg, E. F. F. Botta, and F. W. Wubs, *Nested grids ILU-decomposition (NGILU)*, J. Comput. Appl. Math., 66 (1996), pp. 515–526.
[27] J. R. Poirier, *Modélisation électromagnétique des effets de rugosité surfacique*, Thèse de doctorat, l'INSA Toulouse, Toulouse, France, 2000.
[28] A. Pothen, H. D. Simon, and K.-P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.
[29] V. Rokhlin, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.
[30] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, New York, 1996.
[31] H. D. Simon and S.-H. Teng, *How good is recursive bisection?*, SIAM J. Sci. Comput., 18 (1997), pp. 1436–1445.
[32] W.-P. Tang and W. L. Wan, *Sparse approximate inverse smoother for multigrid*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1236–1252.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 852–876

EFFICIENT AND STABLE SOLUTION OF M-MATRIX LINEAR SYSTEMS OF (BLOCK) HESSENBERG FORM*

LUCA GEMIGNANI† AND GRAZIA LOTTI‡

Abstract. The problem of solving large linear systems whose coefficient matrix is a sparse M-matrix in block Hessenberg form has recently received much attention, especially for applications in Markov chains and queueing theory. Stewart proposed a recursive algorithm which is shown to be backward stable. Although the theoretical derivation of such an algorithm is very simple, its efficient implementation is logically rather involved. An analysis of its computational cost in the case where the initial coefficient matrix satisfies quite general sparsity properties can be found in [P. Favati et al., *Acta Tecnica Acad. Sci. Hungar.,* 108 (1997–1999), pp. 89–105].

In this paper we devise a different divide-and-conquer strategy for the solution of block Hessenberg linear systems. Our approach follows from a recursive application of the block Gaussian elimination algorithm. For dense matrices, the present method has a computational cost comparable to that of Stewart's algorithm; for sparse matrices it is more efficient and more robust.

Key words. block Hessenberg matrices, M-matrices, recursive Gaussian elimination, sparse matrices, error analysis

AMS subject classifications. 65F05, 65G05

PII. S0895479801387085

1. Introduction.

In this paper we are concerned with the efficient and stable solution of large sparse linear systems of the form

$$AX = B, \tag{1.1}$$

where $A = (A_{i,j})$ is an $N \times N$ nonsingular M-matrix in block Hessenberg form, with the blocks $A_{i,j} \in \mathbf{R}^{d\times d}$, $N = dk$, and $X, B \in \mathbf{R}^{N\times d}$.

Systems of this type are of practical interest since they can be derived in the analysis of the steady state of Markov chains (see, for instance, [4] and [20] and the references given therein). In typical applications, the coefficient matrix A is obtained by truncation of an infinite matrix and, therefore, is very large. In this case, the design of effective solution algorithms requires the exploitation of all the specific features of the problem itself.

An efficient recursive algorithm for solving systems of the form (1.1) was proposed by Stewart [19, 17, 16]. For dense matrices, this method has an operation count comparable to Gaussian elimination. However, it takes advantage of the sparsity properties of the coefficient matrix because it does not modify the blocks of the original matrix and the fill-in does not occur. In addition, the algorithm has been proven to be stable for M-matrices [17].

Our contribution is to devise a different recursive approach which outperforms Stewart's algorithm with respect to both the computational cost and the error propagation.

*Received by the editors April 2, 2001; accepted for publication (in revised form) by Z. Strakoš July 22, 2002; published electronically February 4, 2003.

http://www.siam.org/journals/simax/24-3/38708.html

†Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti 2, 56127 Pisa, Italy (gemignan@dm.unipi.it).

‡Dipartimento di Matematica, Università di Parma, Via D'Azeglio 85, 43100 Parma, Italy (lotti@prmat.math.unipr.it).

More precisely, we propose to use a divide-and-conquer version of the block Gaussian elimination algorithm in order to split the original system (1.1), with coefficient matrix A, into two smaller systems of the same type of about half the size and to which the same reduction technique can be applied. Then, we show that this recursive block variant of block Gaussian elimination can be conveniently used to solve linear systems of the form (1.1). In particular, we prove that the effectiveness of our approach depends on such things as the dimension of the blocks and the sparsity patterns within each block.

Block variants of the naive implementation of block Gaussian elimination have been widely studied because they can be organized in such a way that matrix multiplication becomes the dominant operation [14]. In the block Hessenberg case, this important feature of block methods can be combined with a sparsity preserving property. Specifically, we are able to show that our process almost maintains at each step the sparsity pattern of the initial coefficient matrix A. In fact, it has the property of producing low fill-in because it generates a sequence of Schur complements which differ from the corresponding matrices only in the last few columns.

By putting these facts together, we prove that for sparse matrices the new algorithm roughly allows an $O(\log k)$ savings in the computational cost with respect to Stewart's. For dense matrices, the two algorithms have a comparable operation count. Computational savings can also be obtained under the assumption that the original system (1.1) has some additional structure. Typical examples where A is block Hessenberg, block Toeplitz, or Toeplitz in block Hessenberg form arise from the solution of computational problems of queueing theory and Markov chains [18, 20], from the numerical treatment of difference and differential equations [10, 9, 8] and, moreover, from approximate factorization problems for polynomials and analytic functions [3, 11, 2]. In these cases, since our recursive scheme proceeds merely by computing Schur complements, it can easily be seen that both scalar and block Toeplitz-like structures are maintained at any intermediate step of the computation. In [10] an application to Toeplitz Hessenberg linear systems modified by a band perturbation is provided. These systems arise frequently when we apply a computational scheme based on the use of difference equations for the computation of many special functions and quantities occurring in engineering and physics [21, 22].

Concerning the stability issues, we present an error analysis showing that our algorithm is backward stable in the sense that the computed solution X is the exact solution of a nearby linear system,

$$AX = B + \Delta B. \tag{1.2}$$

Specifically, we provide upper bounds on the norm of the residual ΔB that are proportional to the condition number of A or of a scaled version of A, where the proportionality constant is linearly increasing (up to a logarithmic factor) with size N. A quite similar conclusion also holds for Stewart's algorithm [17], but the error estimates given in [17] are worse. In fact, for linear systems with M-matrices the proportionality constant grows at about N^2. Moreover, since they depend on the $\log N$-power of certain quantities, Stewart's algorithm can be impractical for general block Hessenberg linear systems even if all the matrices generated in the binary tree are relatively well conditioned.

We implemented our algorithm in C++ and then compared the resulting implementation, both in cost and in accuracy, with Stewart's algorithm. From many numerical experiments performed on a Pentium 550 workstation with the Linux system, our algorithm has been proved to be more robust and faster than Stewart's.

Experimental comparisons with more general solvers for sparse matrices also have been carried out. For sparse Hessenberg M-matrices of size $N = 8192$ with randomly generated entries, our program has been about 10 times faster than a sparse solver provided by MATLAB based on Gaussian elimination with partial pivoting applied to AP, where P is a certain permutation matrix found by computing a minimum degree ordering for the columns of A. To compare our algorithm with another popular strategy for sparse Gaussian elimination, we considered subroutine MA28 distributed by the HSL (formerly the Harwell Subroutine Library) Archive and discussed in [6]. This implements a variant of Gaussian elimination for sparse systems, where pivotal interchanges are determined according to a suitable Markowitz strategy combined with a threshold pivoting technique for considerations of numerical stability. We tested both our algorithm and MA28 for the solution of nonsingular M-matrix linear systems of Hessenberg form which are very sparse; i.e., the number of nonzeros is $O(N)$. We have found that in such cases MA28 generally performs better than our approach from the point of view of computational efficiency. In fact, fill-ins are ordinarily $O(N)$ in MA28, whereas in our implementation they are of order $O(N \log N)$ due to the additional entries generated in the recursive phase by Schur complement operations. However, the scenario completely changes when we take into account the accuracy of the computed solutions. The residuals generated by our methods are usually of an order of machine precision; conversely, in many cases the threshold pivoting strategy of MA28 is catastrophic, as it destroys the M-structure of the coefficient matrix and produces an exponential growth of the entries encountered during the factorization process. Although many more sophisticated sparse block oriented codes could be taken into account, we believe that the issues revealed by the numerical comparisons shown are quite typical when using a general purpose sparse direct solver. Aggressive pivoting strategies are sparseness preserving but usually lead to accuracy problems. On the contrary, conservative procedures are numerically robust but time consuming, as they substantially increase the fill-in of the initial coefficient matrix.

This paper is organized in the following way. In section 2 we describe the divide-and-conquer algorithm for the solution of problem (1.1) and derive its computational cost. In section 3 we analyze its behavior in finite precision arithmetic and, finally, in section 4 we present and discuss numerical experiments confirming the effectiveness of the proposed method.

The programs presented are available at the first author's website http://www.dm.unipi.it/~gemignan/ric.html.

2. The algorithm and its computational cost. In the first part of this section we describe our recursive scheme for solving linear system (1.1). For a general known vector B, the resulting algorithm requires a preprocessing phase, where the same recursive procedure is applied to the solution of a linear system with the same coefficient matrix but a different known vector E formed from the first d columns of the $N \times N$ identity matrix I_N. For this reason, in the derivation of the algorithm, we consider first the solution of $AX = E$, where E is given as above, and then that of $AX = B$ for a general block vector B.

In the second part, we carry out the cost analysis of our method by showing that, under some auxiliary assumptions on the sparsity of the initial coefficient matrix, it behaves better than Stewart's algorithm (for a cost analysis of Stewart's algorithm see [7]).

2.1. Derivation of the algorithm. Our first aim is to solve the block linear system

$$AX = E, \tag{2.1}$$

where A is a real $N \times N$ nonsingular M-matrix of block upper Hessenberg form having blocks of size d,

$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1,k-1} & A_{1,k} \\ A_{21} & A_{22} & \dots & A_{2,k-1} & A_{2,k} \\ O & A_{32} & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & A_{k-1,k-1} & A_{k-1,k} \\ O & \dots & O & A_{k,k-1} & A_{k,k} \end{bmatrix}, \ A_{i,j} \in \mathbf{R}^{d\times d}, \ N = kd, \tag{2.2}$$

and $E \in \mathbf{R}^{N\times d}$ is the matrix containing the first d columns of the $N \times N$ identity matrix I_N. For the sake of simplicity, we restrict ourselves to the case $k = 2^p$, with p a positive integer.

In the description of the algorithms we make use of a binary string notation for the indices of the recursion level and of the submatrices. For a finite binary string ω, we denote by $|\omega|$ the length of ω, that is, the number of bits of ω, and by $n(\omega)$ the natural number whose binary representation is ω. We denote by λ the empty string such that $|\lambda| = 0$ and $n(\lambda) = 0$. Moreover, $\omega 0$, $\omega 1$ denote the strings obtained by concatenating ω with 0 or 1, respectively.

At the ith level of recursion, $i \le p$, systems either of the form

$$A_\omega X_\omega = E_\omega \tag{2.3}$$

or

$$\widehat{A}_\omega X_\omega = E_\omega \tag{2.4}$$

are to be solved, where A_ω (resp., $\widehat{A}_\omega$) is a nonsingular M-matrix of block upper Hessenberg form of order $m_\omega = k_\omega d$ with k_ω^2 blocks of order d, ω is any string such that $|\omega| = i$, and E_ω contains the first d columns of the $m_\omega \times m_\omega$ identity matrix.

At the zero level of recursion we have to solve the system

$$A_\lambda X_\lambda = E_\lambda,$$

with $A_\lambda = A$, $E_\lambda = E$, $m_\lambda = N$, and $k_\lambda = k$. At the pth level of recursion, the k resulting systems of order d are solved by applying the Gaussian elimination algorithm.

Assume, without loss of generality, we are going to solve systems (2.3). The recursion is applied by halving at each step the number of blocks in the coefficient matrix. Matrix A_ω is partitioned as

$$A_\omega = \begin{bmatrix} A_{\omega 0} & R_\omega \\ T_\omega & A_{\omega 1} \end{bmatrix}, \tag{2.5}$$

where $A_{\omega 0}$ and $A_{\omega 1}$ have $k_{\omega 0}^2 = k_\omega/2$ and $k_{\omega 1}^2 = k_\omega/2$ blocks and are of order $m_{\omega 0} = m_\omega/2$ and $m_{\omega 1} = m_\omega/2$, respectively. Hence $k_\omega = k/2^{|\omega|}$.

Since A has a block upper Hessenberg form, matrix T_ω has just one nonzero block of order d, say $C_\omega = A_{r+1,r}$, for a suitable index r, in the upper right corner. Thus, the matrix T_ω can be written as

$$T_\omega = E_{\omega 1} C_\omega G_{\omega 0}, \tag{2.6}$$

where

$$E_{\omega 1} = \begin{bmatrix} I_d \\ O \end{bmatrix}, \quad G_{\omega 0} = \begin{bmatrix} O & I_d \end{bmatrix}.$$

Matrix $E_{\omega 1}$ has size $m_{\omega 1} \times d$ and contains the first d columns of the $m_{\omega 1} \times m_{\omega 1}$ identity matrix; matrix $G_{\omega 0}$ has size $d \times m_{\omega 0}$ and contains the last d rows of the $m_{\omega 0} \times m_{\omega 0}$ identity matrix. Matrix I_d stands for the identity matrix of size d.

Consider the following factorization of matrix A_ω:

$$A_\omega = \begin{bmatrix} \widehat{A}_{\omega 0} & R_\omega \\ O & A_{\omega 1} \end{bmatrix} \begin{bmatrix} I_{m_{\omega 0}} & O \\ A_{\omega 1}^{-1} T_\omega & I_{m_{\omega 1}} \end{bmatrix}, \tag{2.7}$$

where

$$\widehat{A}_{\omega 0} = A_{\omega 0} - R_\omega A_{\omega 1}^{-1} E_{\omega 1} C_\omega G_{\omega 0}. \tag{2.8}$$

Note that since A_ω is a nonsingular M-matrix, both $A_{\omega 1}$ and $\widehat{A}_{\omega 0}$ (the Schur complement of $A_{\omega 0}$) are nonsingular M-matrices [1]. Moreover, observe that, since $\widehat{A}_{\omega 0}$ differs from $A_{\omega 0}$ only in the last d columns, the matrices $\widehat{A}_\omega$ can also be partitioned as

$$\widehat{A}_\omega = \begin{bmatrix} A_{\omega 0} & \widehat{R}_\omega \\ T_\omega & \widehat{A}_{\omega 1} \end{bmatrix}.$$

From (2.7), one finds that the solution of (2.3) can be carried out by the following steps:

1. Solve
$$A_{\omega 1} X_{\omega 1} = E_{\omega 1}. \tag{2.9}$$
2. Compute $\widehat{A}_{\omega 0}$, according to (2.8).
3. Solve
$$\widehat{A}_{\omega 0} X_{\omega 0} = E_{\omega 0}, \tag{2.10}$$
where $E_{\omega 0} = E_{\omega 1}$.
4. Compute
$$Y = -X_{\omega 1} C_\omega G_{\omega 0} X_{\omega 0}. \tag{2.11}$$
5. Set
$$X_\omega = \begin{bmatrix} X_{\omega 0} \\ Y \end{bmatrix}.$$

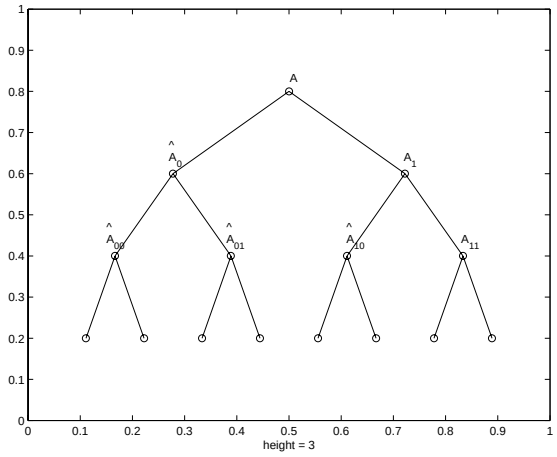


FIG. 2.1. *Binary tree representing* **function** `base`.

Systems (2.9) and (2.10) can be solved by recursively applying the technique used above to solve system (2.3). In this way, one obtains a recursive function, hereafter called **function** `base`, which takes on input the entries of A and then returns in output the solution X of (2.1) together with all the solution vectors X_ω computed at the intermediate steps,

function `base`(ω, A_ω);
begin
  if $|\omega| = p$ `then begin`
    $X_\omega := A_\omega^{-1}$;
    `set`(ω, X_ω);
    `base` $:= X_\omega$;
  `end`
  `else begin`
    $X_{\omega 1} :=$ `base`$(\omega 1, A_{\omega 1})$;
    $W := X_{\omega 1}\, C_\omega\, G_{\omega 0}$;
    $\widehat{A}_{\omega 0} := A_{\omega 0} - R_\omega\, W$;
    $X_{\omega 0} :=$ `base`$(\omega 0, \widehat{A}_{\omega 0})$;
    $Y := -W\, X_{\omega 0}$;
    $X_\omega = \left[X_{\omega 0}^T, Y^T\right]^T$;
    `set`(ω, X_ω);
    `base` $:= X_\omega$;
  `end`
`end`;

where `set`(ω, V) is a function which allocates matrix V in a data structure at the position denoted by ω.

Such a process can also be described by a binary tree, where each node represents either linear system (2.3) or (2.4) to be solved, as shown in Figure 2.1.

More precisely, the root of the tree stands for the system (2.1) with $\omega = \lambda$, each leaf of the tree represents linear systems that are not divided any further but directly solved, and each intermediate node labeled by A_ω ($\widehat{A}_\omega$) is related to the linear system

(2.3) (resp., (2.4)). Under the hypothesis $k = 2^p$, the tree turns out to be a complete binary tree with height p.

Once the matrices X_ω have been computed, at each recursion level of the previous algorithm, and stored in a data structure, the computation of system (1.1),

$$A_\lambda X_\lambda = B_\lambda,$$

for a general known vector $B = B_\lambda$ can start. For the sake of simplicity, assume that at the ith level of recursion, $i \le p$, we are going to solve the system

$$A_\omega Z_\omega = B_\omega, \tag{2.12}$$

where

$$B_\omega = \begin{bmatrix} B_{\omega 0} \\ B_{\omega 1} \end{bmatrix}.$$

From (2.7), one has that

$$A_\omega = \begin{bmatrix} \widehat{A}_{\omega 0} & R_\omega \\ O & A_{\omega 1} \end{bmatrix} \begin{bmatrix} I_{m_{\omega 0}} & O \\ X_{\omega 1} C_\omega G_{\omega 0} & I_{m_{\omega 1}} \end{bmatrix},$$

where

$$\widehat{A}_{\omega 0} = A_{\omega 0} - R_\omega X_{\omega 1} C_\omega G_{\omega 0}, \tag{2.13}$$

and $X_{\omega 1}$ is already available from **function** `base`. By using this block triangular factorization of A_ω, we may easily determine the solution vector Z_ω as follows.

1. Solve
$$A_{\omega 1} Z_{\omega 1} = B_{\omega 1}. \tag{2.14}$$
2. Compute
$$\widehat{B}_{\omega 0} = B_{\omega 0} - R_\omega Z_{\omega 1}.$$
3. Compute $\widehat{A}_{\omega 0}$, according to (2.13).
4. Solve
$$\widehat{A}_{\omega 0} Z_{\omega 0} = \widehat{B}_{\omega 0}. \tag{2.15}$$
5. Compute
$$U = Z_{\omega 1} - X_{\omega 1} C_\omega G_{\omega 0} Z_{\omega 0}.$$
6. Set
$$Z_\omega = \begin{bmatrix} Z_{\omega 0} \\ U \end{bmatrix}.$$

Again, systems (2.14) and (2.15) can be solved by recursively applying the technique used above to solve system (2.12). The resulting recursive scheme for the solution of (1.1) is called **function** `solve`,

function `solve`$(\omega, A_\omega, B_\omega)$;
begin
  if $|\omega| = p$; `then begin`
    $X_\omega := \texttt{get}(\omega)$;
    $Z_\omega := X_\omega\, B_\omega$;
    `solve` $:= Z_\omega$;
  `end`
  `else begin`
    $Z_{\omega 1} := \texttt{solve}(\omega 1, A_{\omega 1}, B_{\omega 1})$;
    $\widehat{B}_{\omega 0} := B_{\omega 0} - R_\omega Z_{\omega 1}$;
    $X_{\omega 1} := \texttt{get}(\omega 1)$;
    $W := X_{\omega 1}\, C_\omega\, G_{\omega 0}$;
    $\widehat{A}_{\omega 0} := A_{\omega 0} - R_\omega\, W$;
    $Z_{\omega 0} := \texttt{solve}(\omega 0, \widehat{A}_{\omega 0}, \widehat{B}_{\omega 0})$;
    $U := Z_{\omega 1} - W\, Z_{\omega 0}$;
    $Z_\omega = \left[Z_{\omega 0}^T, U^T\right]^T$;
    `solve` $:= Z_\omega$;
  `end`
`end`;

where `get`(ω) is a function which returns matrix X_ω. The main program calls `base`(λ, A) and then calls `solve`(λ, A, B).

REMARK 2.1. *The proposed algorithm* **function** `base` *could be simplified (without any computational savings in the order) by noting that all the information needed by function* **function** `solve`*, namely, matrices $X_{\omega 1}$, can be obtained by solving the linear system*

$$AX = \mathbf{e}, \tag{2.16}$$

where $\mathbf{e}$ *is the first column of the matrix I_N, instead of system* (2.1).

REMARK 2.2. *Note that although the version of the algorithm* **function** `base` *proposed here works on block Hessenberg systems with uniform block size, it can be generalized to the case of diagonal blocks with different size. According to Remark* 2.1*, let*

$$A_\lambda X_\lambda = E_\lambda,$$

with $E_\lambda = \mathbf{e}$, be the system at the zero level of recursion. Then, with d_i denoting the size of the ith diagonal block, at the ith level of recursion, $i \le p$ systems either of the form

$$A_\omega X_\omega = E_\omega \tag{2.17}$$

or

$$\widehat{A}_\omega X_\omega = E_\omega \tag{2.18}$$

are to be solved, where $|\omega| = i$, $E_\omega = \mathbf{e}$, when $n(\omega) = 0$, and E_ω contains the first $d_{k_\omega n(\omega)+1}$ columns of the $m_\omega \times m_\omega$ identity matrix otherwise.

Obviously, similar simplifications and/or modifications can also be applied to **function** `solve`. Moreover, it is worth noting that a nonrecursive version of **function** `solve` dealing with diagonal blocks of possibly different sizes can also be developed.

The resulting iterative algorithm is thus particularly suited to solving increasing dimension problems similar to the ones encountered in the treatment of infinite systems by truncation.

In what follows, a cost analysis of **function** `base` and **function** `solve` will be performed. The interesting case where the initial coefficient matrix satisfies quite general sparsity properties will be considered in great detail.

2.2. Computational cost of the algorithm. So far we have described two recursive algorithms, **function** `base` and **function** `solve`, for the solution of system (1.1) in the case where E is formed by the first d columns of I_N and in the general case, respectively. The latter function makes use of certain quantities calculated by the former one at the intermediate steps of recursion. However, if we assume that such quantities are already available, then one easily deduces that the two functions have computational costs of the same order.

Denote by $\#S_A$ the number of nonzero entries of the superdiagonal part S_A of the matrix A:

$$\#S_A = \sum_{r=1}^{k-1} \sum_{s=r+1}^{k} \#A_{rs}.$$

The cost of Stewart's algorithm has been analyzed in [7] under the condition $\#S_A = O(d^2 k \log k)$. The key property used in [7] for the derivation of the complexity estimates of Stewart's algorithm is the following observation. The descent method of Stewart solves the linear system (1.1) by recursively dividing it into two block Hessenberg systems of the same form of half the size, both having a right-hand side known vector with $2d$ columns. In this way, at the bottom of the corresponding binary tree one needs to solve $k \log k$ systems of order d instead of k.

More generally, along this line the following upper bound to the computational cost of Stewart's algorithm can be easily derived:

$$\begin{cases} O\big(\log k(\#S_A\ d + k\ d^3\ \log k)\big) & \text{if} \quad \#S_A = O\big(d^2k^2/\log k\big); \\ O\big(k^2\ d^3\big) \qquad \text{if otherwise} & \#S_A = \Omega\big(d^2k^2/\log k\big), \end{cases} \tag{2.19}$$

where the notation $f = \Omega(g)$ means $g = O(f)$. It is worth noting that the bound (2.19) is tight since it can be reached for some distribution of the nonzero entries of the coefficient matrix. An experimental confirmation of this claim is provided by Figure 2.2, where we compare the arithmetic cost of Stewart's algorithm with respect to that of **function** `solve`.

We denote by γ_ω the cost of solving system (2.3) or (2.4). We have

$$\gamma_\omega \le \gamma_{\omega 0} + \gamma_{\omega 1} + C_1 + C_2 \quad \text{for} \quad |\omega| \le p-1, \tag{2.20}$$

where

- C_1 is the cost of computing Y by means of (2.11);
- C_2 is the cost of computing matrix $\widehat{A}_{\omega 0}$ by means of (2.8).

It holds that

$$C_1 = O\big(m_{\omega 0} d^2\big).$$

The computation of $\widehat{A}_{\omega 0}$ requires two matrix multiplications. One of them involves either R_ω or $\widehat{R}_\omega$, depending on the type of system ((2.3) or (2.4)) we are going

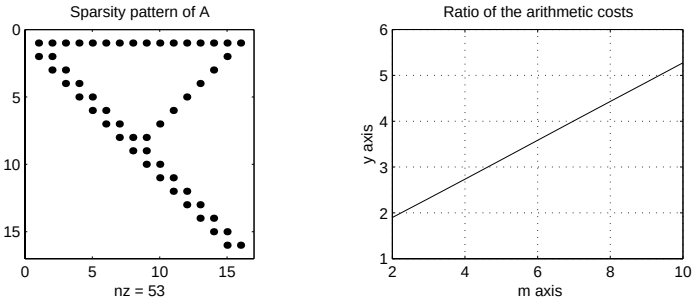


FIG. 2.2. *Plot of the ratio $y = c_s(m)/c_{solve}(m)$ between the arithmetic cost of Stewart's algorithm and that of our implementation of* **function** `solve` *applied to a matrix A of size 2^m.*

to solve at that level of recursion. Noting that

$$\#\widehat{R}_\omega \le \#R_\omega + d\, m_{\omega 0},$$

we get

$$C_2 = O(\#R_\omega\ d + d^2\ m_{\omega 0}).$$

Hence, from (2.20) we have

$$\gamma_\omega = \ \gamma_{\omega 0} + \gamma_{\omega 1} + O(\#R_\omega\ d + d^2\ m_{\omega 0}) \quad \text{for} \quad |\omega| \le p-1.$$

Performing p steps of recursion we get

$$\gamma_\lambda = \sum_{|\omega|=p} \gamma_\omega + O\left(\sum_{r=0}^{p-1} \sum_{|\omega|=r} \#R_\omega d + d^2\ m_{\omega 0}\right). \tag{2.21}$$

For any ω such that $|\omega| = p$, γ_ω is the cost of inverting one of the k final blocks of order d. Hence, we find that

$$\sum_{|\omega|=p} \gamma_\omega = O(kd^3). \tag{2.22}$$

Moreover, it follows that

$$\sum_{r=0}^{p-1} \sum_{|\omega|=r} d\ \#R_\omega = d\ \#S_A \tag{2.23}$$

and

$$\sum_{r=0}^{p-1} \sum_{|\omega|=r} d^2 m_{\omega 0} = d^3\ k\ p/2. \tag{2.24}$$

By substituting (2.22), (2.23), and (2.24) into (2.21), the following asymptotic upper bound to the cost $\gamma = \gamma_\lambda$ of the algorithm **function** `base` is obtained:

$$\gamma = O\big(\#S_A\ d + k\ d^3\ \log k\big).$$

The same upper bound is attained by the computational cost of **function** `solve`, provided that matrices X_ω are available in a data structure. In fact, the recurrent relation for the cost σ_ω of solving system (2.12) results in

$$\sigma_\omega = \sigma_{\omega 0} + \sigma_{\omega 1} + O(\#R_\omega\, d + d^2\, m_{\omega 0}) \quad \text{for} \quad |\omega| \le p-1,$$

since the dominant operation at each level of recursion is still the computation of matrix $\widehat{A}_{\omega 0}$ by means of (2.8).

Therefore, we may conclude that the new algorithm outperforms Stewart's algorithm up to an $O(\log k)$ factor for sufficiently sparse matrices. To support this claim numerically, in Figure 2.2 we compare the arithmetic costs $c_s(m)$ and $c_{solve}(m)$ of Stewart's algorithm and of our **function** `solve`, respectively, when they are applied to matrices of size $N = 2k = 2^m$, $m = 2, 3, \ldots, 10$, having a sparsity pattern as the one shown in the left side of Figure 2.2. Both algorithms have been implemented in a nonrecursive way using MATHEMATICA, and only the multiplicative operations involving nonzero entries of A are counted. The reported values $c_s(m)$ and $c_{solve}(m)$ are found by averaging on 100 numerical experiments with random nonzero entries. Since the plot on the right is linear with respect to m, this means that the cost of Stewart's algorithm is about $m = \log k$ times the cost of our method up to a suitable proportionality constant.

The storage requirement of both algorithms is of order $O\big(d^2 k \log k\big)$ if we do not consider the space needed to allocate the coefficient matrix A. This is a reasonable assumption since in practice A is generally sparse and structured and, therefore, the allocation of A can be made as efficiently as possible by using techniques which exploit such structural and sparsity properties. Note that, similarly to the approach used in [19], the algorithm proposed here solves system (1.1) by performing a two-phase computation, namely, a factorization phase and a solution phase. In [7] a bound to the storage requirement of Stewart's algorithm of order $O\big(d^2 k\big)$ was obtained by grouping together the computation of the two phases. It is easy to see that the same reduction can be reached by applying the same technique to our algorithm.

3. Backward error analysis. In this section we present a backward error analysis of our recursive modification of block Gaussian elimination for the solution of linear system (1.1). Our error analysis is also of a recursive nature and this is because the classical stability analysis of LU factorization schemes, which is extended to the block case in [5], does not apply to our recursive variants, as we explain below.

Indeed, it is easy to show that the RL decomposition (2.7) is equivalent to a more standard LU factorization scheme applied to the matrix $J_{m_\omega} A_\omega^T J_{m_\omega}$, where J_{m_ω} is the block permutation matrix given by

$$J_{m_\omega} = \begin{bmatrix} 0 & I_{m_\omega/2} \\ I_{m_\omega/2} & 0 \end{bmatrix}.$$

In this way, we deduce that **function** `base` performs a recursive variant of the standard LU block factorization method applied to the initial coefficient matrix $A' = J_{m_\lambda} A_\lambda^T J_{m_\lambda}$, which is still a block upper Hessenberg matrix. This resulting description of **function** `base` was used by the first author in [10] to devise a fast and stable solver for certain structured Hessenberg systems. Now, let us assume that

$$A' = \begin{bmatrix} I_{N/2} & 0 \\ L_{21} & I_{N/2} \end{bmatrix} \begin{bmatrix} A'_{11} & A'_{12} \\ 0 & S \end{bmatrix}.$$

The basic assumption underlying the error analysis of block LU factorization methods presented in [5] is that only S is to be recursively factored. This "one-way" assumption allows us to view L_{21} as a block entry of the L factor of the final decomposition of $A' = LR$ and, hence, to bound its size by that of L. Clearly, this is not possible in the case where also A'_{11} should be further factored in the LU fashion. Roughly speaking, this means that recursive variants of block LU factorization methods are more suited to preserve sparsity but, at the same time, stability results are weaker than the ones established for their nonrecursive counterparts.

In what follows we describe an error analysis for both **function** `base` and **function** `solve`, where it is assumed that matrix operations are computed in a conventional way. Without loss of generality, we may always suppose that the known vector B has nonzero entries of constant sign. If this is not the case, one can split B into two different vectors B_1 and B_2 formed from the nonnegative and negative parts of B and then solve the systems whose known vectors are B_1 and B_2. Under this auxiliary assumption, it will be proved that **function** `solve` is backward stable when it is applied to a nonsingular M-matrix A and, moreover, the given upper bounds on the residual are always better than the ones provided for Stewart's algorithm.

Since **function** `solve` makes use of the output generated by **function** `base`, the first step of our analysis is to show that similar stability properties also hold for this latter procedure. Thus, we first investigate the numerical behavior of the recursive algorithm **function** `base` when it is applied to solving (1.1) in the case where $B = E$ is formed from the first d columns of the identity matrix I_N.

We assume that we will work with the infinity norm; however, one might use any induced matrix norm satisfying the following properties. It is mutually consistent with itself, that is,

$$\| FG \| \leq \| F \| \| G \|,$$

where F is m-by-p and G is p-by-n. In addition, it satisfies

$$\| F \| = \| |F| \|$$

and

$$\| F \| \leq \| G \| \quad \text{if} \quad |F| \leq |G|.$$

For instance, these conditions also hold for the one-norm.

In the derivation of upper bounds on the rounding errors generated in the computation of **function** `base` and **function** `solve`, the crucial observation is that the condition number of both matrices $\widehat{A}_{\omega 0}$ and $A_{\omega 1}$ of (2.7) can be suitably bounded from above in terms of the condition number of the input coefficient matrix A.

PROPOSITION 3.1. *We have that*

$$\| A_{\omega 1}^{-1} \| \leq \| A_{\omega}^{-1} \| .$$

Proof. Recall that A_ω is a nonsingular M-matrix and, therefore, the same holds for its trailing principal submatrix $A_{\omega 1}$. Hence, if we denote by $\mathcal{D}(A_{\omega 0})$ the diagonal part of $A_{\omega 0}$ of (2.5), then one finds that the block diagonal matrix $\tilde{A}_\omega$,

$$\tilde{A}_\omega = \begin{bmatrix} \mathcal{D}(A_{\omega 0}) & O \\ O & A_{\omega 1} \end{bmatrix},$$

is still a nonsingular M-matrix such that $\tilde{A}_\omega \geq A_\omega$. From a well-known comparison result for M-matrices [15], this implies that

$$O \leq (\tilde{A}_\omega)^{-1} \leq A_\omega^{-1},$$

from which the proposition follows. ∎

Concerning the estimation of the condition number of the Schur complement $\widehat{A}_{\omega 0}$, we first recall an important property of M-matrices [1].

PROPOSITION 3.2. *Let $A = (a_{i,j})$, $1 \leq i, j \leq N$, be an $N \times N$ nonsingular* M-*matrix. There exists a positive diagonal matrix D, $D = \mathrm{Diag}(d_{1,1}, \ldots, d_{N,N})$, such that AD is strictly diagonally dominant, that is,*

$$a_{i,i}d_i > \sum_{j \neq i} |a_{i,j}|d_j, \quad 1 \leq i \leq N.$$

Now, let D be a positive diagonal matrix such that AD is strictly diagonally dominant. If we apply the recursive procedure **function** `base` to the initial coefficient matrix AD instead of A, then at the ith level of recursion we have to solve systems of either form

$$A_\omega D_\omega X_\omega = E_\omega$$

or

$$\widehat{A}_\omega D_\omega X_\omega = E_\omega,$$

where D_ω are suitable principal submatrices of D. All matrices $A_\omega D_\omega$ and $\widehat{A}_\omega D_\omega$ are still nonsingular M-matrices. In addition, since they are generated by repeated Schur complement operations, they are strictly diagonally dominant, too. From this, one obtains the following.

PROPOSITION 3.3. *It holds that*

$$\| \widehat{A}_{\omega 0} D_{\omega 0} \| \leq 2 \, \| \mathcal{D}(A_{\omega 0} D_{\omega 0}) \|,$$

where $\mathcal{D}(A)$ gives the diagonal part of the matrix A, and $D_{\omega 0}$ is the leading principal submatrix of order $m_{\omega 0}$ of D_ω.

Proof. Note that both $\widehat{A}_{\omega 0} D_{\omega 0}$ and $A_{\omega 0} D_{\omega 0}$ are strictly diagonally dominant M-matrices. The theorem is immediately established by observing that the nonnegative diagonal entries of $\widehat{A}_{\omega 0} D_{\omega 0}$ are less than the corresponding entries of $A_{\omega 0} D_{\omega 0}$. ∎

By combining Propositions 3.1 and 3.3, we are able to prove that every matrix $A_\omega D_\omega$ and $\widehat{A}_\omega D_\omega$ generated by **function** `base` has a condition number bounded from above by 2 times that of AD.

PROPOSITION 3.4. *For every matrix $A_\omega D_\omega$ and $\widehat{A}_\omega D_\omega$ generated by* **function** `base` *applied to AD, we have*

$$\| (A_\omega D_\omega)^{-1} \| \leq \| (AD)^{-1} \|, \quad \| (\widehat{A}_\omega D_\omega)^{-1} \| \leq \| (AD)^{-1} \|$$

and

$$\| A_\omega D_\omega \| \leq \| AD \|, \quad \| \widehat{A}_\omega D_\omega \| \leq 2 \, \| AD \| .$$

Proof. The matrices $A_\omega D_\omega$ are trailing principal submatrices of AD and, therefore, the norms of their inverses can be bounded from above by means of Proposition 3.1.

The matrices $\widehat{A}_\omega D_\omega$ are trailing principal submatrices of Schur complements of trailing principal submatrices of AD. Thus, the estimate on the norm of $\widehat{A}_\omega D_\omega$ follows from Proposition 3.3. In addition, by recalling that the inverse of a Schur complement defines a principal submatrix of the inverse, one gets that one application of Proposition 3.1 implies the upper bound on the inverse of $\widehat{A}_\omega D_\omega$. □

By making use of these results, we are able to express the upper bound $\zeta_{|\omega|}$ on the total error incurred in the solution of (2.3) in terms of the upper bounds $\zeta_{|\omega 1|} = \zeta_{|\omega 0|}$ on the total errors affecting the solution of the systems of order $m_{\omega 0} = m_{\omega 1}$ with coefficient matrices $A_{\omega 1}$ and $\widehat{A}_{\omega 0}$. The subsequent backward error analysis is of recursive type as that of [17].

In view of the inductive assumption, the solution $fl(X_{\omega 1})$ computed at step 1 of our recursive procedure **function** `base` satisfies

$$A_{\omega 1}\ fl(X_{\omega 1}) = E_{\omega 1} + \Delta E_{\omega 1}, \tag{3.1}$$

where $\Delta E_{\omega 1} \in \mathbf{R}^{m_{\omega 1} \times d}$ is such that

$$|\Delta E_{\omega 1}| \leq |H_1 D_{\omega 1}||D_{\omega 1}^{-1} fl(X_{\omega 1})| + O(\epsilon^2), \quad \| H_1 D_{\omega 1} \| \leq \zeta_{|\omega 1|}\epsilon \| A_{\omega 1} D_{\omega 1} \| . \tag{3.2}$$

Here ϵ is the machine precision, H_1 denotes a certain matrix of order $m_{\omega 1}$ and, moreover, $D_{\omega 1}$ is a certain principal submatrix of order $m_{\omega 1}$ of the diagonal matrix D defined by Proposition 3.2.

Since $A_{\omega 1}$ is a nonsingular M-matrix, its inverse is nonnegative and, therefore, the same holds for $X_{\omega 1}$. Hence, without loss of generality, we may assume that $fl(X_{\omega 1})$ has nonnegative entries, too. If this is true, then one easily deduces the following representation for the computed vector $fl(U_{\omega 1})$, where $U_{\omega 1}$ is given by

$$R_\omega X_{\omega 1} C_\omega G_{\omega 0} = U_{\omega 1} G_{\omega 0}.$$

We have

$$fl(U_{\omega 1}) = R_\omega fl(X_{\omega 1}) C_\omega + \Delta U_{\omega 1}, \tag{3.3}$$

where

$$|\Delta U_{\omega 1}| \leq c\epsilon |R_\omega D_{\omega 1}||D_{\omega 1}^{-1} fl(X_{\omega 1}) C_\omega| + O(\epsilon^2),$$

and c is a positive constant linearly increasing with size $m_{\omega 1}$ of $U_{\omega 1}$.

The evaluation of the entries of the Schur complement $\widehat{A}_{\omega 0}$ at step 2 produces a floating point matrix $fl(\widehat{A}_{\omega 0})$ such that

$$fl(\widehat{A}_{\omega 0}) = A_{\omega 0} - fl(U_{\omega 1}) G_{\omega 0} + \Delta \widehat{A}_{\omega 0}, \tag{3.4}$$

where

$$|\Delta \widehat{A}_{\omega 0}| \leq \epsilon |A_{\omega 0} - fl(U_{\omega 1}) G_{\omega 0}|. \tag{3.5}$$

From the inductive assumption, it follows that the computed solution $fl(X_{\omega 0})$ of the linear system of order $m_{\omega 0}$ at step 3 satisfies

$$fl(\widehat{A}_{\omega 0})\ fl(X_{\omega 0}) = E_{\omega 0} + \Delta E_{\omega 0}, \tag{3.6}$$

with

$$\begin{array}{c} |\Delta E_{\omega 0}| \le |H_2 D_{\omega 0}||D_{\omega 0}^{-1} fl(X_{\omega 0})| + O(\epsilon^2), \\ \| H_2 D_{\omega 0} \| \le \zeta_{|\omega 0|}\epsilon \, \| fl(\widehat{A}_{\omega 0}) D_{\omega 0} \| . \end{array} \tag{3.7}$$

Finally, the computed vector solution $fl(X_\omega)$,

$$fl(X_\omega) = \left[\begin{array}{c} fl(X_{\omega 0}) \\ fl(Y) \end{array} \right],$$

is such that

$$fl(Y) = -fl(X_{\omega 1}) C_\omega (fl(X_{\omega 0}))_{k_{\omega 0}} + \Delta Y, \tag{3.8}$$

with

$$|\Delta Y| \le d\epsilon |fl(X_{\omega 1}) C_\omega (fl(X_{\omega 0}))_{k_{\omega 0}}| + O(\epsilon^2), \tag{3.9}$$

where $(Y)_i$ denotes the ith block entry of a block vector Y.

From (3.8) and (3.9), one obtains that

$$\left[\begin{array}{cc} I_{m_{\omega 0}} & O \\ fl(X_{\omega 1}) C_\omega G_{\omega 0} & I_{m_{\omega 1}} \end{array} \right] \left[\begin{array}{c} fl(X_{\omega 0}) \\ fl(Y) \end{array} \right] = \left[\begin{array}{c} fl(X_{\omega 0}) \\ \Delta Y \end{array} \right]. \tag{3.10}$$

By multiplying each side of (3.10) by the block upper triangular factor

$$\left[\begin{array}{cc} fl(\widehat{A}_{\omega 0}) & R_\omega \\ O & A_{\omega 1} \end{array} \right],$$

we have that

$$\left(A_\omega + \left[\begin{array}{cc} \Delta \widehat{A}_{\omega 0} - \Delta U_{\omega 1} G_{\omega 0} & O \\ \Delta E_{\omega 1} C_\omega G_{\omega 0} & O \end{array} \right] \right) \left[\begin{array}{c} fl(X_{\omega 0}) \\ fl(Y) \end{array} \right] = \left[\begin{array}{c} E_{\omega 0} + \Delta E_{\omega 0} + R_\omega \Delta Y \\ A_{\omega 1} \Delta Y \end{array} \right]. \tag{3.11}$$

This means that the computed vector $fl(X)$ is the exact solution of a perturbed linear system,

$$A_\omega fl(X_\omega) = E_\omega + \left[\begin{array}{c} \Delta E_{\omega 0} + R_\omega \Delta Y - (\Delta \widehat{A}_{\omega 0} - \Delta U_{\omega 1} G_{\omega 0}) fl(X_{\omega 0}) \\ A_{\omega 1} \Delta Y - \Delta E_{\omega 1} C_\omega G_{\omega 0} fl(X_{\omega 0}) \end{array} \right]. \tag{3.12}$$

Since we suppose that the computed and exact quantities agree in sign, then we have that $fl(X_{\omega 1})$ and $fl(X_\omega)$ have nonnegative entries. Moreover, in view of (3.9), one finds that

$$|\Delta Y| \le \frac{d\epsilon}{1 - d\epsilon} |fl(Y)| + O(\epsilon^2) \le 2d\epsilon |fl(Y)| + O(\epsilon^2).$$

On using these facts, it is easily seen that the modulus of the perturbing term on the right of (3.12) can be upper bounded by

$$|H_3 D_\omega| \left[\begin{array}{c} |D_{\omega 0}^{-1} fl(X_{\omega 0})| \\ |D_{\omega 1}^{-1} fl(Y)| \end{array} \right] + O(\epsilon^2),$$

where the matrix $|H_3 D_\omega|$ is a 2×2 block diagonal matrix with diagonal blocks, respectively, given by

$$|H_2 D_{\omega 0}| + d\epsilon |fl(U_{\omega 1})| G_{\omega 0} D_{\omega 0} + |\Delta \widehat{A}_{\omega 0}| D_{\omega 0} + |\Delta U_{\omega 1}| G_{\omega 0} D_{\omega 0}$$

and

$$2d\epsilon |A_{\omega 1} D_{\omega 1}| + |H_1 D_{\omega 1}|.$$

From (3.5) it follows that

$$|\Delta \widehat{A}_{\omega 0}| D_{\omega 0} \leq \epsilon |(A_{\omega 0} - fl(U_{\omega 1}) G_{\omega 0}) D_{\omega 0}| = \epsilon |\widehat{A}_{\omega 0} D_{\omega 0}| + O(\epsilon^2).$$

Analogously, (3.3) implies that

$$|fl(U_{\omega 1})| \leq |R_\omega D_{\omega 1}| D_{\omega 1}^{-1} A_{\omega 1}^{-1} E_{\omega 1} |C_\omega| + O(\epsilon)$$

and

$$|\Delta U_{\omega 1}| \leq c\epsilon |R_\omega D_{\omega 1}| D_{\omega 1}^{-1} A_{\omega 1}^{-1} E_{\omega 1} |C_\omega| + O(\epsilon^2).$$

In this way, one finds that

$$|H_3 D_\omega| = |H_4 D_\omega| + O(\epsilon^2), \tag{3.13}$$

where $|H_4 D|$ is a 2×2 block diagonal matrix with diagonal blocks, respectively, given by

$$|H_2 D_{\omega 0}| + (c+d)\epsilon |R_\omega D_{\omega 1}| D_{\omega 1}^{-1} A_{\omega 1}^{-1} E_{\omega 1} |C_\omega| G_{\omega 0} D_{\omega 0} + \epsilon |\widehat{A}_{\omega 0}| D_{\omega 0}$$

and

$$2d\epsilon |A_{\omega 1} D_{\omega 1}| + |H_1 D_{\omega 1}|.$$

Hence, in view of (3.2) and (3.7), the norm of the matrix $|H_4 D_\omega|$ can be upper bounded by

$$\| H_4 D_\omega \| \leq 2\epsilon \big(\zeta_{|\omega 1|} + (c+d)\gamma \| D_\omega^{-1} A_\omega^{-1} \| \| D_\omega \| + d + 1 \big) \| A_\omega D_\omega \|,$$

where

$$\gamma = \max_{1 \leq i \leq N-1} \| A_{i+1,i} \|,$$

from which one finally gets that

$$\zeta_{|\omega|} \leq 2\big(\zeta_{|\omega 1|} + (c+d)\gamma \| D^{-1} A^{-1} \| \| D \| + d + 1 \big).$$

Summing up, we arrive at the following result showing that the solution computed by means of **function** `base` at the top level of the binary tree is the exact solution of a nearby linear system.

PROPOSITION 3.5. *The block vector $fl(X)$ computed at the top level of our recursive procedure* **function** `base` *for solving system* (2.1) *is the exact solution of a linear system*

$$A fl(X) = E + \Delta E,$$

with

$$\| \Delta E \| \leq \bar{c}\epsilon(1 + \gamma \| D^{-1}A^{-1} \| \| D \|) \| AD \| \| D^{-1} fl(X) \| + O(\epsilon^2),$$

where $\bar{c}$ behaves as the size of A by the height of the binary tree.

The stability analysis of **function** `solve` can be performed in a similar way under the auxiliary assumption that the nonzero entries of B are of constant sign, say, nonnegative. In particular, this condition ensures that $Z_{\omega 1} \geq 0$ and $Z_{\omega 0} \geq 0$. Thus, the following relations can easily be established:

$$A_{\omega 1}\ fl(Z_{\omega 1}) = B_{\omega 1} + \Delta B_{\omega 1},$$

where

$$|\Delta B_{\omega 1}| \leq |\widehat{H}_1 D_{\omega 1}||D_{\omega 1}^{-1} fl(Z_{\omega 1})| + O(\epsilon^2), \quad \| \widehat{H}_1 D_{\omega 1} \| \leq \widehat{\zeta}_{|\omega 1|}\epsilon \| A_{\omega 1} D_{\omega 1} \|;$$

$$P_{\omega 1} = R_\omega Z_{\omega 1},\ fl(P_{\omega 1}) = R_\omega fl(Z_{\omega 1}) + \Delta P_{\omega 1},\ |\Delta P_{\omega 1}| \leq c\epsilon |R_\omega| fl(Z_{\omega 1}) + O(\epsilon^2);$$

$$fl(\widehat{B}_{\omega 0}) = B_{\omega 0} - fl(P_{\omega 1}) + \Delta \widehat{B}_{\omega 0},\ |\Delta \widehat{B}_{\omega 0}| \leq \epsilon |B_{\omega 0} - fl(P_{\omega 1})|;$$

$$fl(W) = fl(X_{\omega 1}) C_\omega G_{\omega 0} + \Delta W,\ |\Delta W| \leq d\epsilon |fl(X_{\omega 1})||C_\omega| G_{\omega 0} + O(\epsilon^2);$$

$$fl(\widehat{A}_{\omega 0})\ fl(Z_{\omega 0}) = fl(\widehat{B}_{\omega 0}) + \Delta fl(\widehat{B}_{\omega 0}),$$

where

$$|\Delta fl(\widehat{B}_{\omega 0})| \leq |\widehat{H}_2 D_{\omega 0}||D_{\omega 0}^{-1} fl(Z_{\omega 0})| + O(\epsilon^2)$$

and

$$\| \widehat{H}_2 D_{\omega 0} \| \leq \widehat{\zeta}_{|\omega 0|}\epsilon \| fl(\widehat{A}_{\omega 0}) D_{\omega 0} \|;$$

$$S_{\omega 1} = W Z_{\omega 0}, \quad fl(S_{\omega 1}) = fl(W) fl(Z_{\omega 0}) + \Delta S_{\omega 1},$$

where

$$|\Delta S_{\omega 1}| \leq d\epsilon |fl(W)| fl(Z_{\omega 0}) + O(\epsilon^2);$$

$$fl(U) = fl(Z_{\omega 1}) - fl(S_{\omega 1}) + \Delta U,$$

where

$$|\Delta U| \leq \epsilon |fl(Z_{\omega 1}) - fl(S_{\omega 1})|.$$

Since we may always assume that the computed and exact quantities agree in sign, then we obtain that

$$fl(Z_{\omega 1}) \leq fl(U).$$

By combining these error estimates, we obtain that

$$\left(A_\omega + \begin{bmatrix} \Delta\widehat{A}_{\omega 0} - \Delta U_{\omega 1} G_{\omega 0} & O \\ \Delta E_{\omega 1} C_\omega G_{\omega 0} & O \end{bmatrix}\right) \begin{bmatrix} fl(Z_{\omega 0}) \\ fl(U) \end{bmatrix}$$
$$= \begin{bmatrix} B_{\omega 0} + \Delta\widehat{B}_{\omega 0} + \Delta fl(\widehat{B}_{\omega 0}) - \Delta P_{\omega 1} + R_\omega(\Delta U - \Delta S_{\omega 1} - \Delta W fl(Z_{\omega 0})) \\ B_{\omega 1} + \Delta B_{\omega 1} + A_{\omega 1}\Delta U - A_{\omega 1}\Delta S_{\omega 1} - A_{\omega 1}\Delta W fl(Z_{\omega 0}) \end{bmatrix}.$$

Under the reasonable assumption $\epsilon \le 0.1$, one has

$$|\Delta U| \le \frac{\epsilon}{1-\epsilon}|fl(U)| + O(\epsilon^2) \le 1.2\epsilon|fl(U)| + O(\epsilon^2)$$

and, analogously,

$$|\Delta\widehat{B}_{\omega 0}| \le \frac{\epsilon}{1-\epsilon}|fl(\widehat{B}_{\omega 0})| + O(\epsilon^2) \le 1.2\epsilon|fl(\widehat{B}_{\omega 0})| + O(\epsilon^2).$$

Hence, it follows that

$$A_\omega \begin{bmatrix} fl(Z_{\omega 0}) \\ fl(U) \end{bmatrix} = \begin{bmatrix} B_{\omega 0} \\ B_{\omega 1} \end{bmatrix} + \Delta B_\omega,$$

where

$$|\Delta B_\omega| \le |\widehat{H}_3 D_\omega| \left| \begin{bmatrix} D_{\omega 0}^{-1} fl(Z_{\omega 0}) \\ D_{\omega 1}^{-1} fl(U) \end{bmatrix} \right| + O(\epsilon^2).$$

The matrix $\widehat{H}_3$ is a 2×2 block matrix,

$$\widehat{H}_3 = \begin{bmatrix} \widehat{H}_{1,1} & \widehat{H}_{1,2} \\ \widehat{H}_{2,1} & \widehat{H}_{2,2} \end{bmatrix},$$

whose block entries are such that

$$|\widehat{H}_{1,1}| \le |\Delta\widehat{A}_{\omega 0}| D_{\omega 0} + |\Delta U_{\omega 1}| G_{\omega 0} D_{\omega 0} + |\widehat{H}_2 D_{\omega 0}|$$
$$+ 1.2\epsilon|\widehat{A}_{\omega 0} D_{\omega 0}| + 2d\epsilon|R_\omega| D_{\omega 1} D_{\omega 1}^{-1} fl(X_{\omega 1})|C_\omega| G_{\omega 0} D_{\omega 0},$$

$$|\widehat{H}_{1,2}| \le (1.2 + c)\epsilon|R_\omega D_{\omega 1}|,$$

$$\widehat{H}_{2,1} = (2d\epsilon|A_{\omega 1} D_{\omega 1}| + |H_1 D_{\omega 1}|) D_{\omega 1}^{-1} fl(X_{\omega 1})|C_\omega| G_{\omega 0} D_{\omega 0}$$

and, moreover,

$$|H_{2,2}| \le 1.2\epsilon|A_{\omega 1} D_{\omega 1}| + |\widehat{H}_1 D_{\omega 1}|.$$

These inequalities lead to the following recurrences involving the parameters $\widehat{\zeta}_j$ and ζ_j which are used to bound the errors. Specifically, we have

$$\widehat{\zeta}_\omega \le 2(\alpha_1 c + \alpha_2 + \alpha_3\gamma \parallel D \parallel\!\parallel D^{-1}A^{-1} \parallel + \widehat{\zeta}_{\omega 1} + \widehat{\zeta}_{\omega 0} + \zeta_{\omega 1}\gamma \parallel D \parallel\!\parallel D^{-1}A^{-1} \parallel),$$

TABLE 3.1
Comparison of the accuracy of Stewart's algorithm with respect to that of **function** `solve`.

m	$\| X_s - X_g \|_\infty$	$\| X_{solve} - X_g \|_\infty$
5	1.1e-11	1.8e-12
7	2.3e-8	1.7e-10
9	2.6e-5	1.7e-8
11	2.4e-2	2.7e-6

where α_i, $i = 1, 2, 3$, are positive constants of small size. By using Proposition 3.5, this relation can be further simplified as follows:

$$\begin{aligned}\widehat{\zeta}_\omega \le 2(&\alpha_1 c + \alpha_2 + \alpha_3\gamma \| D \| \| D^{-1}A^{-1} \| + \widehat{\zeta}_{\omega 1} + \widehat{\zeta}_{\omega 0} \\ &+ c_{m_{\omega 1}}(1 + \gamma \| D^{-1}A^{-1} \| \| D \|)\gamma \| D \| \| D^{-1}A^{-1} \|),\end{aligned} \tag{3.14}$$

where $c_{m_{\omega 1}}$ is upper bounded by $\alpha_4 m_{\omega 1} \log m_{\omega 1}$. Hence, we finally arrive at the next result showing that **function** `solve` is backward stable when it is applied to nonsingular M-matrices.

PROPOSITION 3.6. *The block vector $fl(X)$ computed at the top level of our recursive procedure* **function** `solve` *for solving system* (1.1) *is the exact solution of a linear system*

$$Afl(X) = B + \Delta B,$$

with

$$\| \Delta B \| \le \tilde{c}\epsilon(1 + (\gamma \| D^{-1}A^{-1} \| \| D \|)^2) \| AD \| \| D^{-1} fl(X) \| + O(\epsilon^2),$$

where $\tilde{c}$ behaves as the size of A by the square of the height of the binary tree.

For comparison, let us recall that in [17] Stewart's algorithm was shown to be backward stable when applied to nonsingular M-matrices; however, the upper bound on the norm of the residual term given there is worse than the estimate of Proposition 3.6. In particular, it grows at about N^2 and, therefore, it can deteriorate when the size N of the considered system increases. An illustration of this drawback is given in Table 3.1, where we compare the numerical behavior of Stewart's algorithm and of **function** `solve` applied to the solution of linear system (1.1) in the case where $d = 1$, $N = 2^m$, $E = [1, \ldots, N]^T$ and the coefficient matrix A is the $N \times N$ symmetric tridiagonal matrix with 1 and -0.5 on the main diagonal and on the first superdiagonal, respectively. For $m = 5, 7, 9, 11$, in Table 3.1 we report the absolute errors $\| X_s - X_g \|_\infty$ and $\| X_{solve} - X_g \|_\infty$, where X_s and X_{solve} are the approximate solutions computed by a nonrecursive version of Stewart's algorithm and of **function** `solve`, respectively, implemented by using MATHEMATICA. The "exact" solution X_g is determined by using the MATHEMATICA function *LinearSolve* run at high precision.

In the next section we will report the results of many other numerical experiments and comparisons that confirm the effectiveness of our algorithm when applied in finite precision arithmetic.

4. Numerical experiments. In our implementation of the algorithm, the sparse matrix is stored in compressed sparse column format with the nonzero entries of each column in reverse order. In the following we refer to **a**, **b**, **r** as the vectors constituting the structure; vectors **a**, **b** contain the nonzero entries and their corresponding row indices; and vector **r** gives the locations in the other vectors of the last element in

each column. The structure is not changed during the computation. For the sake of simplicity, blocks $A_{i,i}$ and $A_{i+1,i}$, $i = 1, \ldots, k$, are assumed to be full matrices of order d.

The main operations on the structure are the retrieval of both the last block column of $A_{\omega 0}$ and the entries of R_ω, $|\omega| = r$, $r = 0, \ldots, p-1$. It is easy to see that this operational overhead does not affect the overall computational cost of the algorithm.

The first operation takes $O(\#SA + kd^2)$ comparisons between integers and double assignment operations. In fact the nonzero entries of the jth column of $A_{\omega 0}$, with $m_{\omega 0} - d < j \leq m_{\omega 0}$, are stored at the locations l of array $\mathbf{a}$ with l in the range $r_{h(j)} + d \leq l < r_{h(j)+1}$ and such that $p < b_l \leq p + m_{\omega 0}$, where $p = m_{\omega 0} n(\omega 0)$ and $h(j) = p + j$.

The retrieval of the nonzero entries of all the matrices R_ω can be accomplished with an operational overhead of order $O(\#SA)$. In fact, information on the position in the structure of the nonzero entries of R_ω can be obtained during the searching of the nonzero entries of $R_{\omega 1}$ and stored in an auxiliary array **ind** of order N. After the retrieval of the nonzero entries in the jth column of $R_{\omega 1}$ is completed, vector **ind** is updated and $\mathbf{ind}(i)$, $i = j + m_{\omega 1} n(\omega 1)$, contains the position in the structure at which to start the search of the nonzero entries of the $(m_\omega - m_{\omega 1} + j)$th column of R_ω.

Finally, we show how to efficiently perform the computation of $A_{\omega 0}$. Recall that only the last block column of $R_\omega X_{\omega 1} C_\omega G_{\omega 0}$ is required. Let $R_\omega = [\mathbf{t}_1, \ldots, \mathbf{t}_{m_{\omega 1}}]$, $C_\omega^T X_{\omega 1}^T = [\mathbf{y}_1, \ldots, \mathbf{y}_{m_{\omega 1}}]$, where $\mathbf{t}_j$ and $\mathbf{y}_j$, $j = 1, \ldots, m_{\omega 1}$, are $m_{\omega 0}$-vectors and d-vectors, respectively. Then the last block column of $R_\omega X_{\omega 1} C_\omega G_{\omega 0}$ can be expressed as

$$\sum_{j=1}^{m_{\omega 1}} \mathbf{t}_j \mathbf{y}_j^T,$$

and the computation of $A_{\omega 0}$, $|\omega| = r$, $i = 0, \ldots, p-1$, can be done with $O(d\#SA + kd^3 \log k)$ double arithmetic operations, as mentioned in section 2.

To check the stability properties of our method numerically, we performed numerical experiments on a Pentium 550 workstation with the Linux system by using the standard IEEE 53-bit floating point arithmetic, where all the floating point variables in the program have been declared as double.

As a test suite we considered both matrices specifically chosen in order to illustrate our error bounds and matrices with random entries to test the computational effectiveness of our approach when it is compared with quite general methods for sparse matrices.

The first class consists of matrices A of the form

$$A = \left[\begin{array}{cc|cc|cc|cc} \alpha & \gamma & \lambda & & & & & \\ \rho & \beta & \delta & \mu & & & & \\ \hline & \varrho & \alpha & \gamma & \lambda & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ \hline & & & \ddots & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots & \mu \\ \hline & & & & & \varrho & \alpha & \gamma \\ & & & & & & \rho & \beta \end{array}\right], \tag{4.1}$$

TABLE 4.1
Residuals for the first class of test matrices.

	$\| AX - B \|_\infty$		$\frac{\| AX - B \|_\infty}{\| A \|_\infty \| X \|_\infty}$	
N	**function** `solve`	Gaussian elimination	**function** `solve`	Gaussian elimination
256	8.5e-14	5.6e-14	1.6e-18	1.1e-18
512	4.5e-13	2.2e-13	1.8e-18	9.4e-19
1024	9.1e-12	3.6e-12	2.7e-18	1.1e-18
2048	9.3e-10	4.7e-10	1.5e-18	7.9e-19
4096	4.5e-5	4.5e-5	2.6e-18	8.8e-19

TABLE 4.2
Residuals for the second class of test matrices.

	$\| AX - B \|_\infty$		$\frac{\| AX - B \|_\infty}{\| A \|_\infty \| X \|_\infty}$	
N	**function** `solve`	Gaussian elimination	**function** `solve`	Gaussian elimination
256	7.8e-11	4.3e-11	1.9e-16	1.1e-16
512	2.5e-10	2.1e-10	1.6e-16	1.3e-16
1024	1.6e-9	9.1e-10	2.4e-16	1.4e-16
2048	7.8e-9	3.7e-9	3.0e-16	1.4e-16
4096	2.9e-8	1.6e-8	2.8e-16	1.6e-16

where $\alpha, \beta, \gamma, \lambda, \delta, \mu, \rho$, e, ϱ stand for real parameters. Their values are determined as in [17] to generate sets of matrices with different properties. In particular, the following sets are considered:

1. $\alpha = 0.01$, $\beta = 1$, $\gamma = 0$, $\delta = 1$, $\lambda = 1$, $\mu = 0$, $\rho = 0$, $\varrho = 1$. This is a critical situation for Stewart's algorithm since certain quantities which occur in the upper bounds on the residuals generated by that procedure can be very large. The resulting coefficient matrix A is neither diagonally dominant nor an M-matrix.
2. $\alpha = 1$, $\beta = 1$, $\gamma = 0.01$, $\delta = 0.99$, $\lambda = 0$, $\mu = 0$, $\rho = \gamma$, $\varrho = \delta$. In this case, A is a symmetric diagonally dominant matrix and Stewart's algorithm is proven to be backward stable.
3. $\alpha = 1$, $\beta = 1$, $\gamma = -490$, $\delta = -0.00049$, $\lambda = 0$, $\mu = 0$, $\rho = \delta$, $\varrho = \gamma$. In this case, A is an M-matrix which can be reduced to be diagonally dominant by multiplication on the right by the diagonal matrix

$$D = \mathrm{Diag}(1000, 1, \ldots, 1000, 1).$$

For each considered coefficient matrix $A \in \mathbf{R}^{N \times N}$, our implementation of **function** `solve` computes the solution X of the linear system $AX = B$, where B is the Nth vector with entries $(B)_i = i$, $i = 1, \ldots, N$. By using MATLAB, we also evaluate the absolute and relative residuals $\| AX - B \|_\infty$ and $\| AX - B \|_\infty / (\| A \|_\infty \| X \|_\infty)$ and, moreover, we compare them with the ones generated by Gaussian elimination (backslash operator in MATLAB).

Tables 4.1, 4.2, and 4.3 report the results of our numerical experiments. These results clearly confirm the robustness and effectiveness of our method when it is applied to the solution of linear systems with M-matrices in (block) Hessenberg form. The computed residuals are of an order comparable with the ones produced by Gaussian elimination with partial pivoting. Moreover, Tables 4.1 and 4.2 show that accurate results are still found for diagonally dominant matrices and even for more general block

TABLE 4.3
Residuals for the third class of test matrices.

	$\| AX - B \|_\infty$		$\frac{\| AX - B \|_\infty}{\| A \|_\infty \| X \|_\infty}$	
N	**function** `solve`	Gaussian elimination	**function** `solve`	Gaussian elimination
256	1.5e-9	2.2e-10	2.7e-19	3.9e-20
512	2.6e-9	4.4e-10	2.2e-19	3.8e-20
1024	7.9e-9	8.9e-10	3.3e-19	3.7e-20
2048	1.5e-8	1.8e-9	3.1e-19	3.6e-20
4096	3.3e-8	3.5e-9	3.4e-19	3.6e-20

Hessenberg matrices. Furthermore, Table 4.3 shows that the estimate of Proposition 3.6 is generally too pessimistic and that small residuals can be obtained also in the cases where A is ill-conditioned—for any N the estimated condition number of A is about $2.4e+7$—or where AD is strictly diagonally dominant for a matrix D of large norm. Obviously, in these situations the computed solution can be highly inaccurate.

Concerning numerical experiments with random matrices, our primary interest is the comparison of the computational efficiency of our method with respect to more general solvers for sparse matrices based on LU factorization. These latter algorithms produce a solution of a sparse linear system $AX = B$ at the cost of at least $(nz(U+L-I))$ arithmetic operations, where $nz(U+L-I)$ denotes the number of nonzero entries in $U + L - I$. Permutations are profitably introduced since a suitable reordering of columns and/or rows of A can often make its LU factors sparser. However, the number of nonzeros in A gives a lower bound to the number of nonzeros in the factorization, that is, $nz(L - I + U) \geq nz(A)$.

Assuming that d systems with the same coefficient matrix A must be solved, the following considerations can be made. Any LU-based solver requires at least $d\, nz(A)$ arithmetic operations, while our algorithm solves the problem with a cost of order $O(d\#SA + kd^3 \log k)$, including the operational overhead, as it follows from the discussion made at the beginning of this section. In the case of weak sparsity, i.e., for $\#S_A = \Omega(kd^2 \log k)$, $nz(A)$ and $\#S_A$ are of the same order, and our algorithm has a cost of order $O(d\, \#S_A)$, that is, it reaches in the order the best possible performance of any LU-based solver.

Several effective packages for the numerical treatment of sparse matrices are available [13]; in particular, in our experiments we have considered both a method provided by MATLAB and subroutine MA28 distributed by the HSL (formerly the Harwell Subroutine Library) Archive and discussed in [6].

We generated M-matrices $A = (a_{i,j}) \in \mathbf{R}^{N \times N}$ in upper Hessenberg form with random entries according to the following rules:

$$A = H + R,$$

where $H = (h_{i,j})$ has a sparsity pattern like that shown in Figure 2.2 with nonzero entries $h_{i,i} = N$ and $h_{i,j} = -1$ if $i \neq j$ and, moreover, R is a strictly upper triangular matrix with $O(N \log N)$ nonzero entries equal to -1 generated on the fly. Since $\#S_A = O(N \log N)$, it follows that **function** `solve` computes the solution of $AX = B$ at the cost of $O(N \log N)$ arithmetic operations. We compared the results of our method with results returned by a MATLAB routine which applies Gaussian elimination with partial pivoting to the matrix AP, where P is a certain permutation matrix generated by the MATLAB function $P = colmmd(A)$ which determines a

TABLE 4.4
Values of the parameter BIG reported by MA28.

N	BIG
32	27776
64	1.9E+09
128	8.8E+18

minimum degree ordering for the columns of A [12, 13]. Typically, the nonzeros in the computed upper triangular factor U are $\Omega(N \log^2 N)$ and, therefore, the sparse linear solver by MATLAB behaves worse than our method. In particular, for $N = 8192$ our implementation is about 10 times faster than MATLAB.

MA28 solves a sparse system of linear equations using a completely different approach based on the Markowitz strategy. This chooses a pivot which minimizes $(r_i - 1)(c_j - 1)$ subject to a threshold pivot tolerance, where r_i and c_j are the row and column counts in the reduced matrix in Gaussian elimination. It has been observed that this method is quite effective for keeping fill-in low in LU factors. To verify this claim experimentally we have carried out extensive numerical experiments with upper Hessenberg M-matrices A of the form $A = H + R$, where R has $O(N)$ nonzeros entries only. In these cases our algorithm is not optimal since its computational cost is $O(N \log N)$ due to fill-ins generated in the recursive process by Schur complement operations. On the contrary, MA28 generally preserves the sparsity of the input coefficient matrix, and fill-ins are ordinarily of order N. However, serious problems are encountered with MA28 when checking the accuracy of the computed solutions. The value $u = 0.1$ is suggested in MA28 for the parameter u which controls the pivot choice, and this means that MA28 usually accepts a maximum growth factor of about $\simeq 10$ per step in the Gaussian elimination process. The resulting pivoting strategy destroys the M-structure of the initial matrix and in many experiments performed it caused an exponential growth of the entries in the computed triangular factors leading to quite inaccurate results.

To clearly show this drawback, we consider an upper Hessenberg matrix $A = (a_{i,j}) \in \mathbf{R}^{N \times N}$ generated according to the following rules: $a_{i,i} = 1$ for $1 \le i \le N$; $a_{i+1,i} = -0.5$ for $1 \le i \le N-1$; $a_{1,i+1} = -1/N$ for $1 \le i \le N-1$; $a_{i,n-i+1} = -0.5$ for $2 \le i \le N/2$; moreover, $a_{i,i+1} = -0.5$ for $N/2 + 1 \le i \le N - 1$. We applied both our method and MA28 for the solution of $AX = B$, where $B = [1/N, 0, \ldots, 0, 0.5]^T$. Our method was tested up to $N = 8192$ and in all cases it returned a computed solution whose residual is of order of machine precision. On the contrary, MA28 produces results which are completely inaccurate for $N = 128$. Table 4.4 reports the values of the parameter BIG, generated as output by MA28, which provides the absolute value of the largest entry computed during the factorization phase.

The same values are found if u is set to 0.4. The growth factor remains under control for $u \ge 0.5$.

5. Conclusion and further developments. In this paper a recursive variant of block Gaussian elimination has been developed for the solution of large sparse linear systems with M-matrices in block Hessenberg form. We have shown that our approach compares favorably with respect to the other existing methods in taking advantage of the sparsity of the initial coefficient matrix A. In particular, it outperforms Stewart's algorithm, which is a different variant of Gaussian elimination especially suited to the considered class of matrices, both in cost and numerical stability. However, Stewart's algorithm can be implemented in parallel because it is based on the Sherman–Morrison

update, while our algorithm is not well suited to parallel implementation. The search for an effective parallel algorithm for the solution of M-matrix linear systems of block Hessenberg form is a topic of future research.

More recently, the theoretical and computational properties of Stewart's approach have been investigated under the assumption that the original system has some additional structure. Typical examples where A is block Hessenberg, block Toeplitz, or Toeplitz in block Hessenberg form arise from the solution of computational problems of queueing theory and Markov chains, from the numerical treatment of difference and differential equations and, moreover, from approximate factorization problems for polynomials and analytic functions. In these cases, since our recursive scheme proceeds merely by computing Schur complements, it can easily be seen that scalar and block Toeplitz-like structures are maintained at any intermediate step of the computation of **function** `solve`. Moreover, it should be possible to devise a polynomial version of our algorithm where matrix operations are replaced by polynomial ones in such a way to obtain a further speed up of computations. Finally, in particular cases of interest for applications it can be shown that our algorithm has more strong stability properties, depending on the sign distribution of the entries of A. A numerical comparison between our method and Stewart's when applied to the solution of sparse and structured linear systems might therefore be very interesting.

REFERENCES

[1] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] D. A. Bini, L. Gemignani, and B. Meini, *Computations with infinite Toeplitz matrices and polynomials*, Linear Algebra Appl., 343/344 (2002), pp. 21–61.

[3] D. A. Bini, L. Gemignani, and B. Meini, *Factorization of analytic functions by means of Koenig's theorem and Toeplitz computations*, Numer. Math., 89 (2001), pp. 49–82.

[4] J. P. C. Blanc, *A numerical approach to cyclic-service queueing models*, Queueing Systems Theory Appl., 6 (1990), pp. 173–188.

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

[6] I. S. Duff and J. K. Reid, *Some design features of a sparse matrix code*, ACM Trans. Math. Softw., 5 (1979), pp. 18–35.

[7] P. Favati, G. Lotti, O. Menchi, and F. Romani, *Efficient solution of sparse block Hessenberg systems*, Acta Tecnica Acad. Sci. Hungar., 108 (1997–1999), pp. 89–105.

[8] P. Favati, G. Lotti, O. Menchi, and F. Romani, *Solution of infinite linear systems by automatic adaptive iterations*, Linear Algebra Appl., 318 (2000), pp. 209–225.

[9] L. Gemignani, *On a generalization of Poincaré's theorem for matrix difference equations arising from root-finding problems*, in Structured Matrices in Mathematics, Computer Science and Engineering, Vol. II, AMS, Providence, RI, 2001, pp. 265–278.

[10] L. Gemignani, *Efficient and stable solution of structured Hessenberg linear systems arising from difference equations*, Numer. Linear Algebra Appl., 7 (2000), pp. 319–335.

[11] L. Gemignani, *Polynomial factors, lemniscates and structured matrix technology*, in Structured Matrices: Recent Developments in Theory and Computation, D. Bini, P. Yalamov, and E. Tyrtyshnikov, eds., Nova Science Publishers, Huntington, NJ, 2000, pp. 117–134.

[12] A. George and J. W. H. Liu, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19.

[13] 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.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[15] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[16] G. Latouche and G. W. Stewart, *Numerical methods for M/G/1 type queues*, in Computations with Markov Chains, W. J. Stewart, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995, pp. 571–581.

[17] U. Von Matt and G. W. Stewart, *Rounding errors in solving block Hessenberg systems*, Math. Comp., 65 (1996), pp. 115–135.
[18] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York, 1989.
[19] G. W. Stewart, *On the solution of block Hessenberg systems*, Numer. Linear Algebra Appl., 2 (1995), pp. 287–296.
[20] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1994.
[21] J. Wimp, *Computation with Recurrence Relations*, Pitman Press, Boston, 1984.
[22] J. Wimp, B. Kline, A. Galardi, and D. Colton, *Some preliminary observations on an algorithm for the computation of moment integrals*, J. Comput. Appl. Math., 19 (1987), pp. 117–124.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 3, pp. 877–897

A CHART OF BACKWARD ERRORS FOR SINGLY AND DOUBLY STRUCTURED EIGENVALUE PROBLEMS*

FRANÇOISE TISSEUR†

Abstract. We present a chart of structured backward errors for approximate eigenpairs of singly and doubly structured eigenvalue problems. We aim to give, wherever possible, formulae that are inexpensive to compute so that they can be used routinely in practice. We identify a number of problems for which the structured backward error is within a factor $\sqrt{2}$ of the unstructured backward error. This paper collects, unifies, and extends existing work on this subject.

Key words. eigenvalue, eigenvector, symmetric matrix, Hermitian matrix, skew-symmetric matrix, skew-Hermitian matrix, symplectic matrix, conjugate symplectic matrix, Hamiltonian matrix, backward error, condition number

AMS subject classifications. 65F15, 65F20, 65H10, 65L15, 65L20, 15A18, 15A57

PII. S089547980139995X

1. Introduction.

Bunse-Gerstner, Byers, and Mehrmann [8] present a chart of numerical methods for structured eigenvalue problems for which the matrices have more than one of the properties defined as follows:

$A \in \mathbb{C}^{m\times m}$ is	$A \in \mathbb{R}^{m\times m}$ is
Hermitian if $A^* = A$,	symmetric if $A^T = A$,
skew-Hermitian if $A^* = -A$,	skew-symmetric if $A^T = -A$,
unitary if $A^*A = I$,	orthogonal if $A^TA = I$,
conjugate symplectic if $m = 2n$ and $A^*JA = J$,	symplectic if $m = 2n$ and $A^TJA = J$,
Hamiltonian if $m = 2n$ and $(JA) = (JA)^*$,	J-symmetric if $m = 2n$ and $(JA) = (JA)^T$,
skew-Hamiltonian if $m = 2n$ and $(JA) = -(JA)^*$,	J-skew symmetric if $m = 2n$ and $(JA) = -(JA)^T$,

where $J = \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$, I_n being the $n\times n$ identity matrix. Structured eigenvalue problems occur in numerous applications and we refer to [8] for a list of them and pointers to the relevant literature. In this paper we present a chart of computable backward errors for approximate eigenpairs and condition numbers for simple eigenvalues of matrices having one or two of these special structures.

The importance of condition numbers for characterizing the sensitivity of solutions to problems and backward errors for assessing the stability and quality of numerical algorithms is widely appreciated. A backward error of an approximate eigenpair (x, λ) of a matrix A is a measure of the smallest perturbation E such that $(A+E)x = \lambda x$. This backward error has two main uses. First, it can be used to determine if (x, λ) solves a nearby problem by comparing the backward error with the size of any

* Received by the editors December 21, 2001; accepted for publication (in revised form) by V. Mehrmann August 27, 2002; published electronically February 4, 2003.
http://www.siam.org/journals/simax/24-3/39995.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur/). The work of this author was supported by Engineering and Physical Sciences Research Council grant GR/R45079/01 and by Nuffield Foundation grant NAL/00216/G.

uncertainties in the data matrix A. Second, a bound on the forward error can be obtained in terms of the backward error and an appropriate condition number.

A natural definition of the normwise backward error of an approximate eigenpair (x, λ) is

$$\eta(x,\lambda) = \min\left\{\alpha^{-1}\|E\| : \ (A+E)x = \lambda x \ \right\}, \tag{1.1}$$

where α is a positive parameter that allows freedom in how the perturbations are measured and $\|\cdot\|$ denotes any vector norm and the corresponding subordinate matrix norm. Deif [9] derived the explicit expression for the 2-norm (also valid for any subordinate norm and the Frobenius norm),

$$\eta(x,\lambda) = \alpha^{-1}\|(A-\lambda I)x\|/\|x\|,$$

showing that the normwise backward error is a scaled residual. Also of interest is the backward error of a set of approximate eigenpairs $(x_j, \lambda_j)_{j=1}^k$, which we collect into matrices $X_k = [x_1, x_2, \ldots, x_k]$ and $\Lambda_k = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$. For a measure of the backward error we use the natural generalization of (1.1),

$$\eta(X_k,\Lambda_k) = \min\left\{\alpha^{-1}\|E\| : (A+E)X_k = X_k\Lambda_k\right\}, \tag{1.2}$$

for which an explicit expression is available for any unitarily invariant norm if $\mathrm{rank}(X_k) = k$ [26, Thm. 2.4.2],

$$\eta(X_k,\Lambda_k) = \alpha^{-1}\|R_k X_k^+\|, \tag{1.3}$$

where $R_k = X_k\Lambda_k - AX_k$ is the residual matrix and X_k^+ is the pseudoinverse of X_k.

The measure η is not entirely appropriate for our structured eigenvalue problems, as it does not respect any structure in A. Similar remarks can be made about condition numbers. Standard condition numbers are derived without requiring that perturbations preserve structure. As a consequence, standard condition numbers usually exceed the actual condition number for an eigenvalue problem subject to structured perturbation. In the last few years, efforts have been concentrated on deriving new structure-preserving algorithms for the solution of structured eigenvalue problems for both the dense case [1], [4], [12], [14], [23] and the large and sparse case [2], [3], [5], [21], to cite just a few articles. It is therefore of interest to develop backward errors and condition numbers that fully respect the inherent structure of these problems.

Let $A \in \mathcal{C}_{\mathbb{K}} \subset \mathbb{K}^{m\times m}$ ($\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$) be a singly or doubly structured matrix, where $\mathcal{C}_{\mathbb{K}}$ is the set of matrices having the structure of interest. We extend the definition of the normwise backward error for a set of eigenpairs (X_k, Λ_k) in (1.2) to the structured case by

$$\eta_{\mathbb{K}}(X_k,\Lambda_k) = \min\left\{\alpha^{-1}\|E\|_F : \ (A+E)X_k = X_k\Lambda_k, \ A+E \in \mathcal{C}_{\mathbb{K}}\right\}. \tag{1.4}$$

The contribution of this work is to unify and extend explicit expressions of backward errors for singly and doubly structured eigenproblems. These expressions allow structured backward errors to be computed more efficiently than if (1.4) were treated as a general nonlinear optimization problem.

In section 2 we recall some basic properties of the structured matrices under consideration and give some useful lemmas. We recall in the first part of section 3 that for linear structures a Kronecker product approach can be used to rewrite the minimization problem in (1.4) in terms of the minimal 2-norm solution to an

TABLE 2.1
Eigenvalue properties of the singly structured matrices $A \in \mathbb{C}^{m\times m}$ under consideration.

Class of matrices	Eigenvalues	Class of matrices	Eigenvalues
$A^* = A$	real eigenvalues	$A^T = A$	arbitrary
$A^* = -A$	purely imaginary	$A^T = -A$	0 and/or pairs μ, $-\mu$, $(\mu \neq 0)$
$A^*A = I$	$\lvert\mu\rvert = 1$	$A^TA = I$	± 1 and/or pairs $\mu, 1/\mu$, $(\mu^2 \neq 1)$
$A^*JA = J$	pairs $\mu, 1/\bar{\mu}$	$A^TJA = J$	pairs $\mu, 1/\mu$
$(JA) = (JA)^*$	pairs $\mu, -\bar{\mu}$	$(JA) = (JA)^T$	pairs $\mu, -\mu$
$(JA) = -(JA)^*$	pairs $\mu, \bar{\mu}$	$(JA) = -(JA)^T$	double eigenvalues

underdetermined system. The dimension of the underdetermined system may make the computation of backward error expensive. Fortunately, there are particular classes of linear structured problems for which we can characterize the set of solutions to the constraints in (1.4) and identify the solution of minimal Frobenius norm. This yields backward error formulae that are cheaper to compute and easier to analyze and understand than with the Kronecker product approach. As a result we show that, in some instances, forcing the backward error matrix to have a particular structure has little effect on its norm.

Backward errors for eigenproblems with nonlinear structure are harder to derive. Sun [25] characterizes the complete set of solutions to the constraints in (1.4) for the class of unitary matrices and derives a structured backward error for this class of problems. We use his approach and extend it to the classes of Hermitian unitary, symplectic unitary, and symmetric orthogonal matrices. Many problems remain open. Following the presentation in [8], we give in the second part of section 3 a chart of structured backward errors. For each class of matrices, we either recall an existing known explicit formula for the structured backward error, or derive a new explicit formula, or identify obtaining such a formula as an open problem. We aim to provide formulae that are cheap to compute so that they can be used in the course of a computation. We identify several cases in which the structured backward error is within a factor $\sqrt{2}$ of the unstructured backward error. For completeness, we recall in section 4 how to compute structured condition numbers of simple eigenvalues of matrices depending linearly on a set of parameters.

2. Basics.

2.1. Background material and definitions. We summarize in Table 2.1 the properties of the eigenvalues of the singly structured matrices considered in this paper. If the matrix is real, then its spectrum is symmetric with respect to the real axis. For doubly structured matrices the eigenvalue properties combine. For example, the eigenvalues of a real Hamiltonian matrix come in quadruples $(\lambda, \bar{\lambda}, -\lambda, -\bar{\lambda})$ if $\mathrm{Re}(\lambda) \neq 0$, and the eigenvalues of a Hermitian Hamiltonian matrix come in pairs $(\lambda, -\lambda)$ with λ real. For $A \in \mathbb{C}^{m\times k}$ with $m \geq k$, there exists a matrix $U \in \mathbb{C}^{m\times k}$ with orthonormal columns, and a unique Hermitian positive semidefinite matrix $H \in \mathbb{C}^{k\times k}$, such that $A = UH$. This is called the *polar decomposition* of A.

For a Hermitian matrix A, we define $\mathrm{sign}(A)$ by $\mathrm{sign}(A) = Q\,\mathrm{sign}(D)Q^*$, where $A = QDQ^*$ is the eigendecomposition of A with $Q^*Q = I$, $\mathrm{sign}(D) = \mathrm{diag}(\mathrm{sign}(d_i))$, and $\mathrm{sign}(0) = 1$.

We define the symplectic quasi-QR factorization of a $2n \times k$ matrix A by

$$A = QT, \quad T = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix},$$

where Q is unitary conjugate symplectic, $T_1 \in \mathbb{C}^{n\times k}$ is upper trapezoidal, and $T_2 \in \mathbb{C}^{n\times k}$ is strictly upper trapezoidal. This factorization is discussed in [7, Cor. 4.5(ii)] and [27].

We make frequent use of the following lemmas.

LEMMA 2.1. *Let $A \in \mathbb{C}^{m\times m}, Y_1 \in \mathbb{C}^{m\times k}, m \geq k$, and $Y = [Y_1, Y_2]$ be unitary and let $B \in \mathbb{C}^{k\times k}$. Then*

$$\|Y_1B - AY_1\|_F^2 = \|B - Y_1^*AY_1\|_F^2 + \|Y_2^*AY_1\|_F^2.$$

Proof. The proof is immediate using $Y_1Y_1^* + Y_2Y_2^* = I$ and

$$Y_1B - AY_1 = Y \begin{bmatrix} B - Y_1^*AY_1 \\ -Y_2^*AY_1 \end{bmatrix}. \qquad \square$$

LEMMA 2.2 ([25, Lem. 2.4]). *Let $A \in \mathbb{C}^{m\times m}$ be unitary, $Y_1 \in \mathbb{C}^{m\times k}$ with $2k \leq m$, $Y = [Y_1, Y_2]$ be unitary, and let H_1 and H_2 be the Hermitian polar factors of $Y_1^*AY_1$ and $Y_2^*AY_2$, respectively. Then for any unitarily invariant norm,*

$$\|I - H_1\| = \|I - H_2\| \quad \text{and} \quad \|Y_1^*AY_2\| = \|Y_2^*AY_1\|.$$

Proof. By the CS decomposition [22] there are unitary matrices $U = \operatorname{diag}(U_1, U_2)$ and $V = \operatorname{diag}(V_1, V_2)$ with $U_1, V_1 \in \mathbb{C}^{k\times k}$ such that

$$U^*Y^*AYV = \begin{bmatrix} C & -S & 0 \\ S & C & 0 \\ 0 & 0 & I \end{bmatrix},$$

where C, S are $k \times k$ diagonal matrices with nonnegative diagonal elements and $C^2 + S^2 = I$. Then

$$Y_1^*AY_1 = U_1CV_1^*, \qquad Y_2^*AY_2 = U_2 \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} V_2^*$$

so that $H_1 = V_1CV_1^*$ and $H_2 = V_2 \operatorname{diag}(C, I)V_2^*$. Hence, $\|I - H_2\| = \|I - C\| = \|I - H_1\|$. The second equality follows from

$$Y_2^*AY_1 = U_2 \begin{bmatrix} S \\ 0 \end{bmatrix} V_1, \quad Y_1^*AY_2 = U_1 \begin{bmatrix} -S & 0 \end{bmatrix} V_2. \qquad \square$$

2.2. Structured matrix problems. Before deriving structured backward errors, we need some results on the following structured matrix problem: *Given a class of structured matrices $\mathcal{C}_{\mathbb{K}} \subset \mathbb{K}^{m\times m}$, where $\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$, characterize*

1. *pairs of matrices $Y, B \in \mathbb{K}^{m\times k}$ for which there exists a matrix $A \in \mathcal{C}_{\mathbb{K}}$ such that $AY = B$;*
2. *the set $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} = \{A \in \mathcal{C}_{\mathbb{K}} : \ AY = B\}$.*

The lemmas in this section give a solution to this problem for several classes of structured matrices and give, whenever possible, the optimal solution A_{opt} defined by

$$\|A_{\text{opt}}\|_F = \min\{\|A\|_F : \ A \in \mathcal{S}_{\mathcal{C}_{\mathbb{K}}}\}.$$

First, we need to set the notation. Define the full and reduced singular value decompositions of Y by

$$Y = U \begin{bmatrix} \Sigma_Y & 0 \\ 0 & 0 \end{bmatrix} V^* = U_1 \Sigma_Y V_1^*, \tag{2.1}$$

where $U = [U_1, U_2]$ and $V = [V_1, V_2]$ are unitary with $U_1 \in \mathbb{K}^{m\times r}$, $V_1 \in \mathbb{K}^{k\times r}$, and $\Sigma_Y = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_i > 0$, $i = 1\colon r$, $r = \operatorname{rank}(Y)$. In what follows, Y^+ denotes the pseudoinverse of Y, $P_Y = YY^+ = U_1U_1^*$ is the orthogonal projector onto range(Y), and $P_Y^\perp = I - P_Y$.

The first result is from [24, Lem. 1.4] and concerns the class of Hermitian matrices when $\mathbb{K} = \mathbb{C}$ and the class of symmetric matrices when $\mathbb{K} = \mathbb{R}$. We give the proof for completeness.

LEMMA 2.3. *Let $Y, B \in \mathbb{K}^{m\times k}$, $m \geq k$, be given and let*

$$\mathcal{C}_{\mathbb{K}} = \{A \in \mathbb{K}^{m\times m} : A = A^*\}.$$

Then $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} \neq \emptyset$ if and only if $BP_{Y^*} = B$ and $P_Y BY^+ \in \mathcal{C}_{\mathbb{K}}$, and if $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} \neq \emptyset$, then

$$\begin{aligned} \mathcal{S}_{\mathcal{C}_{\mathbb{K}}} &= \{BY^+ + (BY^+)^* P_Y^\perp + P_Y^\perp H P_Y^\perp : \ H \in \mathcal{C}_{\mathbb{K}}\}, \\ A_{\text{opt}} &= BY^+ + (BY^+)^* P_Y^\perp. \end{aligned}$$

Proof. Substituting (2.1) for Y in $AY = B$ and letting

$$U^*AU = \widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}, \quad U^*BV = \begin{bmatrix} \widetilde{B}_{11} & \widetilde{B}_{12} \\ \widetilde{B}_{21} & \widetilde{B}_{22} \end{bmatrix}, \tag{2.2}$$

with $\widetilde{A}_{11}, \widetilde{B}_{11} \in \mathbb{K}^{r\times r}$, we obtain

$$\begin{bmatrix} \widetilde{A}_{11}\Sigma_Y & 0 \\ \widetilde{A}_{21}\Sigma_Y & 0 \end{bmatrix} = \begin{bmatrix} \widetilde{B}_{11} & \widetilde{B}_{12} \\ \widetilde{B}_{21} & \widetilde{B}_{22} \end{bmatrix}. \tag{2.3}$$

Hence, solutions to $AY = B$ exist if and only if

$$U^*BV = \begin{bmatrix} \widetilde{B}_{11} & 0 \\ \widetilde{B}_{21} & 0 \end{bmatrix}, \quad (\widetilde{B}_{11}\Sigma_Y^{-1})^* = \widetilde{B}_{11}\Sigma_Y^{-1}.$$

The first condition is equivalent to

$$B = U \begin{bmatrix} \widetilde{B}_{11} & 0 \\ \widetilde{B}_{21} & 0 \end{bmatrix} V^* = [\, U_1 \quad U_2 \,] \begin{bmatrix} U_1^*BV_1 & 0 \\ U_2^*BV_1 & 0 \end{bmatrix} \begin{bmatrix} V_1^* \\ V_2 \end{bmatrix} = BV_1V_1^* = BY^+Y = BP_{Y^*}.$$

The second condition is equivalent to $(P_Y BY^+)^* = P_Y BY^+$.

We now prove that $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} = \widetilde{\mathcal{S}}_{\mathcal{C}_{\mathbb{K}}}$, where $\widetilde{\mathcal{S}}_{\mathcal{C}_{\mathbb{K}}} = \{BY^+ + (BY^+)^* P_Y^\perp + P_Y^\perp H P_Y^\perp : H \in \mathcal{C}_{\mathbb{K}}\}$. First, we assume that $A \in \mathcal{S}_{\mathcal{C}_{\mathbb{K}}} \neq \emptyset$. Then from (2.3) we have

$$\begin{aligned} A &= U \begin{bmatrix} \widetilde{B}_{11}\Sigma_Y^{-1} & \Sigma_Y^{-1}\widetilde{B}_{21}^* \\ \widetilde{B}_{21}\Sigma_Y^{-1} & \widetilde{A}_{22} \end{bmatrix} U^* \\ &= U_1U_1^*BV_1\Sigma_Y^{-1}U_1^* + U_2U_2^*BV_1\Sigma_Y^{-1}U_1^* + U_1\Sigma_Y^{-1}V_1^*B^*U_2U_2^* + U_2U_2^*AU_2U_2^* \\ &= BV_1\Sigma_Y^{-1}U_1^* + Y^{+*}B^*(I - U_1U_1^*) + (I - U_1U_1^*)A(I - U_1U_1^*) \\ &= BY^+ + (BY^+)^* P_Y^\perp + P_Y^\perp A P_Y^\perp \end{aligned}$$

so that $A \in \widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}}$ and $\mathcal{S}_{\mathcal{C}_\mathbb{K}} \subset \widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}}$. Now it is easy to verify that if $BP_{Y^*} = B$ and $P_Y BY^+$ is Hermitian, then any $A \in \widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}}$ satisfies $AY = B$ and $A^* = A$ so that $\widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}} \subset \mathcal{S}_{\mathcal{C}_\mathbb{K}}$, which completes the proof of the first part of the lemma.

For the second part, we have

$$\begin{aligned}\|A\|_F^2 &= \|\widetilde{A}\|_F^2 \\ &= \left\| \begin{bmatrix} \widetilde{B}_{11}\Sigma_Y^{-1} \\ \widetilde{B}_{21}\Sigma_Y^{-1} \end{bmatrix} \right\|_F^2 + \| \begin{bmatrix} \widetilde{B}_{11}\Sigma_Y^{-1} & \Sigma_Y^{-1}\widetilde{B}_{21}^* \end{bmatrix} \|_F^2 - \|\widetilde{B}_{11}\Sigma_Y^{-1}\|_F^2 + \|\widetilde{A}_{22}\|_F^2 \\ &= 2\|BY^+\|_F^2 - \|U_1^* B V_1 \Sigma_Y^{-1}\|_F^2 + \|\widetilde{A}_{22}\|_F^2.\end{aligned}$$

Hence $\|A\|_F$ is minimized by setting $\widetilde{A}_{22} = 0$, which implies $P_Y^\perp A P_Y^\perp = 0$. The expression for A_{opt} follows. ∎

The result of Lemma 2.3 can be extended to other classes of matrices.

LEMMA 2.4. *Let $Y, B \in \mathbb{K}^{m\times k}$, $m \ge k$, be given.*

1. *Let $\mathcal{C}_\mathbb{K} = \{A \in \mathbb{K}^{m\times m} : A = -A^*\}$. Then $\mathcal{S}_{\mathcal{C}_\mathbb{K}} \neq \emptyset$ if and only $BP_{Y^*} = B$ and $P_Y BY^+ = -(P_Y BY^+)^*$, and if $\mathcal{S}_{\mathcal{C}_\mathbb{K}} \neq \emptyset$, then*

$$\begin{aligned}\mathcal{S}_{\mathcal{C}_\mathbb{K}} &= \{BY^+ - (BY^+)^* P_Y^\perp + P_Y^\perp H P_Y^\perp : \ H \in \mathcal{C}_\mathbb{K}\}, \\ A_{\mathrm{opt}} &= BY^+ - (BY^+)^* P_Y^\perp.\end{aligned}$$

2. *Let $\mathcal{C}_\mathbb{C} = \{A \in \mathbb{C}^{m\times m} : A = A^T\}$. Then $\mathcal{S}_{\mathcal{C}_\mathbb{C}} \neq \emptyset$ if and only $BP_{Y^*} = B$ and $P_{\overline{Y}} BY^+ = (P_{\overline{Y}} BY^+)^T$, and if $\mathcal{S}_{\mathcal{C}_\mathbb{C}} \neq \emptyset$, then*

$$\begin{aligned}\mathcal{S}_{\mathcal{C}_\mathbb{C}} &= \{BY^+ + (BY^+)^T P_Y^\perp + P_{\overline{Y}}^\perp H P_Y^\perp : \ H \in \mathcal{C}_\mathbb{C}\}, \\ A_{\mathrm{opt}} &= BY^+ + (BY^+)^T P_Y^\perp.\end{aligned}$$

3. *Let $\mathcal{C}_\mathbb{C} = \{A \in \mathbb{C}^{m\times m} : A = -A^T\}$. Then $\mathcal{S}_{\mathcal{C}_\mathbb{C}} \neq \emptyset$ if and only $BP_{Y^*} = B$ and $P_{\overline{Y}} BY^+ = -(P_{\overline{Y}} BY^+)^T$, and if $\mathcal{S}_{\mathcal{C}_\mathbb{C}} \neq \emptyset$, then*

$$\begin{aligned}\mathcal{S}_{\mathcal{C}_\mathbb{C}} &= \{BY^+ - (BY^+)^T P_Y^\perp + P_{\overline{Y}}^\perp H P_Y^\perp : \ H \in \mathcal{C}_\mathbb{C}\}, \\ A_{\mathrm{opt}} &= BY^+ - (BY^+)^T P_Y^\perp.\end{aligned}$$

Proof. All these results are proved in a similar way to Lemma 2.3. For the symmetric or skew-symmetric case, the matrices $\widetilde{A}$ and $\widetilde{B}$ in (2.2) are defined by $\widetilde{A} = U^T A U$ and $\widetilde{B} = U^T B V$. ∎

Note that Lemma 2.3 solves the Hamiltonian structured matrix problem since JA is Hermitian, and for similar reasons Lemma 2.4 solves the skew-Hamiltonian, J-symmetric, and J-skew-symmetric structured matrix problems.

In the next lemma, we extend a result of Kahan, Parlett, and Jiang [19]. Here, Y and X do not have to have orthonormal columns, we do not require X^*Y to be nonsingular, and Y and X may have different ranks.

LEMMA 2.5. *Let $Y, X, B, C \in \mathbb{K}^{m\times k}$, $m \ge k$, be given with* $\mathrm{rank}(Y) = r$ *and* $\mathrm{rank}(X) = s$, *and let $\mathcal{S}_\mathbb{K} = \{A \in \mathbb{K}^{m\times m} : \ AY = B, \ A^*X = C\}$. If $C^*Y = X^*B$, then*

$$\begin{aligned}\mathcal{S}_\mathbb{K} &= \{BY^+ + (CX^+)^* P_Y^\perp + P_X^\perp H P_Y^\perp, \ H \in \mathbb{K}^{m\times m}\} \\ &= \{(C^*X^+)^* + P_X^\perp BY^+ + P_X^\perp H P_Y^\perp, \ H \in \mathbb{K}^{m\times m}\}, \\ A_{\mathrm{opt}} &= BY^+ + (CX^+)^* P_Y^\perp = (C^*X^+)^* + P_X^\perp BY^+.\end{aligned}$$

Proof. Let $\widetilde{\mathcal{S}}^1_{\mathbb{K}} = \{BY^+ + (CX^+)^* P_Y^\perp + P_X^\perp H P_Y^\perp,\ H \in \mathbb{K}^{m\times m}\}$ and $\widetilde{\mathcal{S}}^2_{\mathbb{K}} = \{(C^*X^+)^* + P_X^\perp BY^+ + P_X^\perp H P_Y^\perp,\ H \in \mathbb{K}^{m\times m}\}$. First, we assume that $A \in \mathcal{S}_{\mathbb{K}}$. Let

$$Y = U \begin{bmatrix} \Sigma_Y & 0 \\ 0 & 0 \end{bmatrix} V^* = U_1 \Sigma_Y V_1^*, \quad X = W \begin{bmatrix} \Sigma_X & 0 \\ 0 & 0 \end{bmatrix} Z^* = W_1 \Sigma_X Z_1^*$$

be the full and reduced singular value decompositions of Y and X, $U = [U_1, U_2]$, $W = [W_1, W_2]$ with $U_1 \in \mathbb{K}^{m\times r}$, $W_1 \in \mathbb{K}^{m\times s}$. Partition $V = [V_1, V_2]$ and $Z = [Z_1, Z_2]$ accordingly to U and W and let

$$\widetilde{A} = W^* A U = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}.$$

Then

$$\begin{aligned} &\widetilde{A}_{11} = W_1^* A U_1 = W_1^* B V_1 \Sigma_Y^{-1} = \Sigma_X^{-1} Z_1^* C^* U_1, && \widetilde{A}_{12} = W_1^* A U_2 = \Sigma_X^{-1} Z_1^* C^* U_2, \\ &\widetilde{A}_{21} = W_2^* A U_1 = W_2^* B V_1 \Sigma_Y^{-1}, && \widetilde{A}_{22} = W_2^* A U_2. \end{aligned}$$

Now,

$$A = W \widetilde{A} U^* = W_1 \widetilde{A}_{11} U_1^* + W_1 \widetilde{A}_{12} U_2^* + W_2 \widetilde{A}_{21} U_1^* + W_2 \widetilde{A}_{22} U_2^*.$$

Then replacing $\widetilde{A}_{11}$, $\widetilde{A}_{12}$, and $\widetilde{A}_{21}$ by the expressions above yields, for $\widetilde{A}_{11} = W_1^* B V_1 \Sigma_Y^{-1}$,

$$A = BY^+ + (CX^+)^* P_Y^\perp + P_X^\perp A P_Y^\perp,$$

and for $\widetilde{A}_{11} = \Sigma_X^{-1} Z_1^* C^* U_1$,

$$A = (C^*X^+)^* + P_X^\perp BY^+ + P_X^\perp A P_Y^\perp.$$

Hence $\mathcal{S}_{\mathbb{K}} \subset \widetilde{\mathcal{S}}^1_{\mathbb{K}}$ and $\mathcal{S}_{\mathbb{K}} \subset \widetilde{\mathcal{S}}^2_{\mathbb{K}}$. It is easy to verify that if $C^*Y = X^*B$, then any $A \in \widetilde{\mathcal{S}}^1_{\mathbb{K}}$ and any $A \in \widetilde{\mathcal{S}}^2_{\mathbb{K}}$ satisfy $AY = B$ and $A^*X = C$ so that $\widetilde{\mathcal{S}}^1_{\mathbb{K}} = \mathcal{S}_{\mathbb{K}} = \widetilde{\mathcal{S}}^2_{\mathbb{K}}$.

We have

$$\begin{aligned} \|A\|_F^2 &= \|\widetilde{A}\|_F^2 \\ &= \left\| \begin{bmatrix} \widetilde{A}_{11} \\ \widetilde{A}_{21} \end{bmatrix} \right\|_F^2 + \| \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \end{bmatrix} \|_F^2 - \|\widetilde{A}_{11}\|_F^2 + \|\widetilde{A}_{22}\|_F^2 \\ &= \|BV_1\Sigma_Y^{-1}\|_F^2 + \|\Sigma_X^{-1} Z_1^* C^*\|_F^2 - \|W_1^* B V_1 \Sigma_Y^{-1}\|_F^2 + \|\widetilde{A}_{22}\|_F^2. \end{aligned}$$

Hence $\|A\|_F$ is minimized by setting $\widetilde{A}_{22} = 0$, which implies $P_X^\perp A P_Y^\perp = 0$, and the expressions for A_{opt} follow. ∎

This last result is from [25, Lem. 2.2]. We give the proof for completeness.

LEMMA 2.6. *Let $Y, B \in \mathbb{K}^{m\times k}$, $m \geq k$, be given and let $\mathcal{C}_{\mathbb{K}} = \{A \in \mathbb{K}^{m\times m} : A^*A = I\}$. Then, $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} \neq \emptyset$ if and only if $Y^*Y = B^*B$, and if $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} \neq \emptyset$, then*

$$\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} = \{BY^+ + QP_Y^\perp :\ Q \in \mathcal{C}_{\mathbb{K}},\ QP_Y = P_B Q\}.$$

Proof. If $\mathcal{S}_{\mathcal{C}_{\mathbb{K}}} \neq \emptyset$, then $Y^*Y = B^*B$. Now assume that $Y^*Y = B^*B$. Substituting Y by (2.1) into $Y^*Y = B^*B$ gives $BV_2 = 0$ and $BV_1 = Q_1\Sigma$, where $Q_1 \in \mathbb{K}^{m\times r}$ with $Q_1^*Q_1 = I$. Hence

$$B = Q \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^*,$$

where $Q = [Q_1, Q_2]$ is unitary. Then $A = QU^* \in \mathcal{S}_{\mathcal{C}_\mathbb{K}}$ and therefore $\mathcal{S}_{\mathcal{C}_\mathbb{K}} \neq \emptyset$.

Let $\widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}} = \{BY^+ + Q(I - YY^+) : \ Q^*Q = I, \ QP_Y = P_BQ\}$. First, we assume that $A \in \mathcal{S}_{\mathcal{C}_\mathbb{K}}$. We can rewrite A as $A = BY^+ + A(I - YY^+)$. Note that since A is unitary, $Y^+ = (A^*B)^+ = B^+A$. Also, $AP_Y = BY^+ = P_BA$ so that $A \in \widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}}$ and $\mathcal{S}_{\mathcal{C}_\mathbb{K}} \subset \widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}}$.

Assume that $A \in \widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}}$. Hence $A = BY^+ + QP_Y^\perp$ for some unitary Q such that $QYY^+ = BB^+Q$. From $AY = B$, $Y^+ = B^+A$, and $YY^+ = (YY^+)^*$ it is easy to show that $Y^+Y = B^+B$. We have

$$\begin{aligned} A^*A &= ((BY^+)^* + (I - P_Y)Q^*)(BY^+ + Q(I - P_Y)) \\ &= (BY^+)^*BY^+ + (BY^+)^*Q(I - P_Y) + (I - P_Y)Q^*BY^+ + I - P_Y. \end{aligned}$$

First,

$$(BY^+)^*BY^+ = Y^{+*}B^*BY^+ = Y^{+*}Y^*YY^* = (YY^+)^*(YY^+) = P_Y,$$

and second,

$$\big((BY^+)^*Q(I - P_Y)\big)^* = (I - P_Y)Q^*BY^+ = Q^*(I - P_B)BY^+ = 0.$$

Hence $A^*A = P_Y + 0 + 0 + I - P_Y = I$. Also $AY = BY^+Y + Q(I - YY^+)Y = BB^+B = B$ so that $A \in \mathcal{S}_{\mathcal{C}_\mathbb{K}}$ and $\widetilde{\mathcal{S}}_{\mathcal{C}_\mathbb{K}} \subset \mathcal{S}_{\mathcal{C}_\mathbb{K}}$, which completes the proof. ∎

3. Structured normwise backward errors.

3.1. Kronecker product approach. Assume that A depends linearly on $t \leq m^2$ free parameters and that every element of A is a multiple of a single parameter. We write this dependence as $A = A[p]$ with $p \in \mathbb{K}^t$. Higham and Higham [15], [16] extend the notion of componentwise backward error to allow dependence of the perturbations on a set of parameters, and they define structured componentwise backward errors. We use their approach to rewrite the constraint $A + E \in \mathcal{C}_\mathbb{K}$ in (1.4) as $A + E = A[p + \Delta p]$ or, equivalently, $E = E[\Delta p]$, where Δp is a t-vector of perturbed parameters. Note that if any sparsity of A is included in the structure, then E will have the same sparsity as A.

Applying the vec operator (which stacks the columns of a matrix into one long vector) to the constraints in (1.4) gives

$$(X_k^T \otimes I_m)\,\mathrm{vec}(E) = \mathrm{vec}(R_k), \quad \mathrm{vec}(E) = B\Delta p, \tag{3.1}$$

where $\otimes$ denotes the Kronecker product, $B \in \mathbb{K}^{m^2 \times t}$ is of full rank, and R_k is the residual matrix. We refer to Lancaster and Tismenetsky [20, Chap. 12] for properties of the vec operator and the Kronecker product. Let D be a diagonal matrix such that

$$\|E\|_F = \|D\Delta p\|_2,$$

and let $y = D\Delta p$, $M_k = (X_k^T \otimes I_m)BD^{-1} \in \mathbb{K}^{km \times t}$, and $s_k = \mathrm{vec}(R_k)$. Then we can rewrite (3.1) as the linear system $M_k y = s_k$ and therefore

$$\eta_\mathbb{K}(X_k, \Lambda_k) = \alpha^{-1} \min_{y \in \mathbb{K}^t} \big\{ \|y\|_2 \ : \ M_k y = s_k \big\}.$$

This shows that the structured normwise backward error is given in terms of the minimal 2-norm solution to an overdetermined system if $t < km$ or an underdetermined

system otherwise. There may be no solution to the system if M_k is rank deficient or if the system is overdetermined. If the system is underdetermined and consistent, then the minimal 2-norm solution is given in terms of the pseudoinverse by $y = M_k^+ s_k$. In this case

$$\eta_{\mathbb{K}}(X_k, \Lambda_k) = \alpha^{-1} \|M_k^+ s_k\|_2. \tag{3.2}$$

When the data A, X_k, Λ_k are all real, then Δp is automatically real. In certain circumstances it is appropriate to restrict Δp to be real even though the data A, X_k, Λ_k are complex. This happens when the constraints on A's structure involve conjugation of its coefficients or, in the case of real structured backward error, when A is real and λ or x is complex. In these cases, the backward error derivation must be modified by taking real and imaginary parts in the constraint $(A + E)x = \lambda x$ to obtain a real system of equations. For example, consider a 2×2 skew-Hermitian matrix E and a single eigenpair (x, λ) $(k = 1)$. Taking real and imaginary parts in the constraint $Ex = r$ yields

$$[F, G] \begin{bmatrix} \operatorname{Re}(x) & \operatorname{Im}(x) \\ -\operatorname{Im}(x) & \operatorname{Re}(x) \end{bmatrix} = [\operatorname{Re}(r), \operatorname{Im}(r)],$$

where $F = \operatorname{Re}(E)$ is skew-symmetric and $G = \operatorname{Im}(E)$ is symmetric. The 2×2 skew-Hermitian E can be parameterized by

$$E = \begin{bmatrix} 0 & -\Delta p_1 \\ \Delta p_1 & 0 \end{bmatrix} + i \begin{bmatrix} \Delta p_2 & \Delta p_3 \\ \Delta p_3 & \Delta p_4 \end{bmatrix}, \quad \Delta p_j \in \mathbb{R}, \quad j = 1\colon 4,$$

so that

$$\operatorname{vec}([F, G]) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta p_1 \\ \Delta p_2 \\ \Delta p_3 \\ \Delta p_4 \end{bmatrix} := B \Delta p.$$

In this case, M_1 in (3.2) is given by

$$M_1 = \left(\begin{bmatrix} \operatorname{Re}(x) & \operatorname{Im}(x) \\ -\operatorname{Im}(x) & \operatorname{Re}(x) \end{bmatrix}^T \otimes I_2 \right) B D^{-1} \in \mathbb{R}^{4 \times 4},$$

with $D = \operatorname{diag}([\|b_1\|_1, \ldots, \|b_4\|_1]) = \operatorname{diag}([2, 1, 2, 1])$ and with b_j being the jth column of B.

Generally, the size of M_k makes the computation of $\eta_{\mathbb{K}}(X_k, \Lambda_k)$ expensive. Thus (3.2) is not a formula we would evaluate routinely in the course of solving a problem. Nevertheless, it is useful as a tool when testing the stability of newly developed structure-preserving algorithms, as shown in [27], or to illustrate instability of well-known algorithms.

As we shall see in the next section, for certain classes of structured matrices it is possible to express the structured backward error in a form that is much less expensive to evaluate than (3.2). We also consider some nonlinear structures that are not covered by this Kronecker product approach.

3.2. A chart of structured backward errors. This section provides a chart of structured backward errors for a set of approximate eigenpairs $(x_j, \lambda_j)_{j=1}^k$ for the singly and doubly structured matrices under consideration. We aim to give, whenever possible, formulae that are cheap to compute so that they can be used routinely in practice. We give an expression for $E_{\rm opt}$, the solution of minimal Frobenius norm to the constraints in (1.4). We assume that for each class of problems the set of eigenvalues $\{\lambda_j\}_{j=1}^k$ satisfies the relevant eigenvalue properties listed in Table 2.1, since otherwise $\eta_{\mathbb{K}}(X_k, \Lambda_k) = \infty$. For the structured backward error to exist, we may also need to impose some restrictions on X_k.

The first chart, in Table 3.1, covers the complex case, and the second chart, in Table 3.3, covers the real case. They both list structured backward errors that may be applied to the corresponding structured eigenvalue problems. Question marks indicate cases for which explicit expressions for the structured backward errors are not yet known. An **X** or X indicates that an explicit expression for $\eta_{\mathbb{K}}(X_k, \Lambda_k)$ exists. The symbol **X** emphasizes that the structured backward error is at most a factor $\sqrt{2}$ larger than the corresponding unstructured backward error (never smaller). Finally, entries marked with $\otimes$ indicate that an explicit expression for $\eta_{\mathbb{K}}(X_k, \Lambda_k)$ is obtained via the Kronecker product approach described in section 3.1 (which we recall is applicable to linear structure only). Table 3.2 provides the block structure of the corresponding doubly structured matrices together with the matrix properties of the blocks and is useful in forming the matrix B in (3.1).

In the following, "W-trick"[1] refers to the unitary similarity transformation

$$W^*AW = \frac{1}{2}\begin{bmatrix} A_{11}+A_{22}+i(A_{12}-A_{21}) & A_{11}-A_{22}-i(A_{12}+A_{21}) \\ A_{11}-A_{22}+i(A_{12}+A_{21}) & A_{11}+A_{22}-i(A_{12}-A_{21}) \end{bmatrix},$$

where $W = 2^{-\frac{1}{2}}\left[\begin{smallmatrix} I & I \\ iI & -iI \end{smallmatrix}\right]$ and $A = \left[\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right] \in \mathbb{K}^{2n\times 2n}$. We define

$$Y_k = \begin{bmatrix} Y_{k,1} \\ Y_{k,2} \end{bmatrix} := W^*X_k \quad \text{and} \quad S_k = \begin{bmatrix} S_{k,1} \\ \mathcal{S}_{k,2} \end{bmatrix} := W^*R_k.$$

The superscript (i,j) in $\eta_{\mathbb{K}}^{(i,j)}$ refers to the class of matrices in position (i,j) of the complex chart if $\mathbb{K} = \mathbb{C}$ and of the real chart if $\mathbb{K} = \mathbb{R}$. Recall that $R_k = X_k\Lambda_k - AX_k$.

3.2.1. Complex chart ($\mathbb{K} = \mathbb{C}$).

Position (1,1): $\mathcal{C}_{\mathbb{C}}^{(1,1)} = \{A \in \mathbb{C}^{m\times m} : A^* = A\}$ is the class of Hermitian matrices. First, we assume that X_k has orthonormal columns,[2] since otherwise $\eta_{\mathbb{C}}^{(1,1)}(X_k, \Lambda_k) = \infty$. We have $X_k^+ = X_k^*$ so that $R_kP_{X_k^*} = R_k$ and $P_{X_k}R_kX_k^* = X_k\Lambda_kX_k^* - X_kX_k^*AX_kX_k^*$ is Hermitian. Hence, from Lemma 2.3 the optimal solution to the constraints in (1.4) is given by

$$E_{\rm opt} = R_kX_k^* + (X_kR_k^*)P_{X_k}^{\perp}$$

so that

$$\eta_{\mathbb{C}}^{(1,1)}(X_k, \Lambda_k) = \alpha^{-1}\sqrt{\operatorname{trace}(E_{\rm opt}^*E_{\rm opt})} = \alpha^{-1}\sqrt{2\|R_k\|_F^2 - \|X_k^*R_k\|_F^2},$$

[1] The term "X-trick" is used in [8]. We use W-trick to avoid confusion with our notation.

[2] In practice, if X_k has columns that are close to being orthonormal, then one can replace them by the unitary factor of either its QR factorization or its polar decomposition.

TABLE 3.1
Summary of the structured backward errors.

		1 $A^* = A$	2 $A^* = -A$	3 $A^*A = I$	4 $A^*JA = J$	5 $JA = (JA)^*$	6 $JA = -(JA)^*$	7 $A^T = A$	8 $A^T = -A$	9 $A^TA = I$	10 $A^TJA = J$	11 $JA = (JA)^T$	12 $JA = -(JA)^T$
1	$A^* = A$	**X**	$\emptyset$	**X**	?	X	**X**	**X**	**X**	?	?	$\otimes$	$\otimes$
2	$A^* = -A$		**X**	**X**	?	**X**	X	**X**	**X**	?	?	$\otimes$	$\otimes$
3	$A^*A = I$			**X**	**X**	?	?	?	?	?	?	?	?
4	$A^*JA = J$				?	?	?	?	?	?	?	?	?
5	$JA = (JA)^*$					X	$\emptyset$	$\otimes$	$\otimes$	?	?	X	X
6	$JA = -JA^*$						X	$\otimes$	$\otimes$	?	?	X	X
7	$A^T = A$							X	$\emptyset$	?	?	X	**X**
8	$A^T = -A$								X	?	?	X	X
9	$A^TA = I$									?	?	?	?
10	$A^TJA = J$										?	?	?
11	$JA = (JA)^T$											X	$\emptyset$
12	$JA = -(JA)^T$												X

X: explicit expression for $\eta_{\mathbb{C}}(X_k, \Lambda_k)$ is available and within a factor $\sqrt{2}$ of the unstructured backward error.
X: explicit expression for $\eta_{\mathbb{C}}(X_k, \Lambda_k)$ is available.
?: no explicit expression known for $\eta_{\mathbb{C}}(X_k, \Lambda_k)$.
$\emptyset$: no nontrivial matrices with the prescribed pair of structures.
$\otimes$: expression available from Kronecker product approach. See Table 3.2 for the block structure of the matrices.

TABLE 3.2
Block structure and block property of some doubly structured matrices.

	$(JA) = (JA)^*$	$(JA) = -(JA)^*$
$A^T = A$	$\begin{bmatrix} A_1 & A_2 \\ \bar{A}_2 & -\bar{A}_1 \end{bmatrix}$, $A_1 = A_1^T$, $\bar{A}_2 = A_2^T$	$\begin{bmatrix} A_1 & A_2 \\ -\bar{A}_2 & \bar{A}_1 \end{bmatrix}$, $A_1 = A_1^T$, $A_2^T = -\bar{A}_2$
$A^T = -A$	$\begin{bmatrix} A_1 & A_2 \\ -\bar{A}_2 & \bar{A}_1 \end{bmatrix}$, $A_1 = -A_1^T$, $A_2 = A_2^*$	$\begin{bmatrix} A_1 & A_2 \\ \bar{A}_2 & -\bar{A}_1 \end{bmatrix}$, $A_1 = -A_1^T$, $A_2 = -A_2^*$

	$(JA) = (JA)^T$	$(JA) = -(JA)^T$
$A^* = A$	$\begin{bmatrix} A_1 & A_2 \\ A_2^* & -A_1^T \end{bmatrix}$, $A_1 = A_1^*$, $A_2 = A_2^T$	$\begin{bmatrix} A_1 & A_2 \\ -\bar{A}_2 & \bar{A}_1 \end{bmatrix}$, $A_1 = A_1^*$, $A_2 = -A_2^T$
$A^* = -A$	$\begin{bmatrix} A_1 & A_2 \\ -A_2^* & -A_1^T \end{bmatrix}$, $A_1 = -A_1^*$, $A_2 = A_2^T$	$\begin{bmatrix} A_1 & A_2 \\ -A_2^* & A_1^T \end{bmatrix}$, $A_1 = -A_1^*$, $A_2 = -A_2^T$

where the second equality follows after some algebra. The expression for $\eta_{\mathbb{C}}^{(1,1)}(X_k, \Lambda_k)$ was obtained in [26, Thm. 2.5.9]. If $\eta(X_k, \Lambda_k)$ denotes the unstructured backward error in (1.3), then

$$\eta(X_k, \Lambda_k) \le \eta_{\mathbb{C}}^{(1,1)}(X_k, \Lambda_k) \le \sqrt{2}\,\eta(X_k, \Lambda_k).$$

The first inequality is due to the fact that the class of admissible perturbations is larger for the unstructured case than for the structured case. These inequalities show, as for the structured backward error for Hermitian linear systems [6], [18, Prob. 7.12], that forcing the backward error matrix to be Hermitian has little effect on its norm. Note that for a single eigenpair (x, λ) with x of unit 2-norm and $r = (\lambda I - A)x$ being the residual, E_{opt} is given by

$$E_{\mathrm{opt}} = rx^* + xr^* - (r^*x)xx^*,$$

which is a well-known result in the fields of nonlinear equations and optimization [10], [11, p. 171] and numerical linear algebra [6], [19]. In this case,

$$\eta_{\mathbb{C}}^{(1,1)}(x, \lambda) = \alpha^{-1}\sqrt{2\|r\|_2^2 - (\lambda - x^*Ax)^2}.$$

Position (1,3): $\mathcal{C}_{\mathbb{C}}^{(1,3)} = \{A \in \mathbb{C}^{m\times m} : A^* = A,\ A^*A = I\}$ is the class of Hermitian unitary matrices. We assume that the columns of X_k are orthonormal and that $\Lambda_k = \mathrm{diag}(\pm 1)$. The derivation of $\eta_{\mathbb{C}}$ is along the same lines as that for the class of unitary matrices (see position (3,3)) but with an extra constraint in the minimization problem. Therefore, we give just an outline and refer to position (3,3) for a detailed derivation. Let $X = [X_k, \widetilde{X}]$ be unitary. From (3.5) below we have that

$$\begin{aligned} \alpha^2\eta_{\mathbb{C}}^{(1,3)}(X_k, \Lambda_k)^2 &= \|R\|_F^2 + \min_{\substack{\widetilde{Z}^*\widetilde{Z} = I \\ Z^* = Z}} \|\widetilde{X}\widetilde{Z} - A\widetilde{X}\|_F^2 \\ &= \|R\|_F^2 + \|X_k^*A\widetilde{X}\|_F^2 + \min_{\substack{\widetilde{Z}^*\widetilde{Z} = I \\ Z^* = Z}} \|\widetilde{Z} - \widetilde{X}^*A\widetilde{X}\|_F^2, \end{aligned} \tag{3.3}$$

where the second equality is obtained using Lemma 2.1. The minimization problem in (3.3) is a nearness problem whose solution is given in terms of $\operatorname{sign}(\widetilde{X}^*A\widetilde{X})$ by [17]

$$\min_{\substack{\widetilde{Z}^*\widetilde{Z}=I\\ \widetilde{Z}^*=\widetilde{Z}}} \|\widetilde{Z}-\widetilde{X}^*A\widetilde{X}\|_F = \|\operatorname{sign}(\widetilde{X}^*A\widetilde{X})-\widetilde{X}^*A\widetilde{X}\|_F.$$

Let U_kH_k and $\widetilde{U}\widetilde{H}$ be the polar decompositions of $X_k^*AX_k$ and $\widetilde{X}^*A\widetilde{X}$, respectively. If $\widetilde{X}^*A\widetilde{X}=QDQ^*$ is the eigendecomposition of $\widetilde{X}^*A\widetilde{X}$ with Q unitary and D real diagonal, then $\widetilde{X}^*A\widetilde{X}=Q\operatorname{sign}(D)Q^*Q|D|Q^*=\operatorname{sign}(\widetilde{X}^*A\widetilde{X})Q|D|Q^*$ with $\operatorname{sign}(\widetilde{X}^*A\widetilde{X})$ unitary and $Q|D|Q^*$ Hermitian positive definite. Hence, we have $\widetilde{U}=\operatorname{sign}(\widetilde{X}^*A\widetilde{X})$ and $\widetilde{H}=Q|D|Q^*$ so that

$$\|\operatorname{sign}(\widetilde{X}^*A\widetilde{X})-\widetilde{X}^*A\widetilde{X}\|_F=\|\widetilde{U}-\widetilde{U}\widetilde{H}\|_F=\|I-\widetilde{H}\|_F.$$

By Lemmas 2.2 and 2.1 we have

$$\|I-\widetilde{H}\|_F^2=\|I-H_k\|_F^2=\|U_k-X_k^*AX_k\|_F^2=\|X_kU_k-AX_k\|_F^2-\|X_k^*A\widetilde{X}\|_F^2.$$

Then replacing the minimization problem in (3.3) by the above expression yields

$$\eta_{\mathbb{C}}^{(1,3)}(X_k,\Lambda_k)=\alpha^{-1}\sqrt{\|R_k\|_F^2+\|X_kU_k-AX_k\|_F^2}.$$

Note that $\eta_{\mathbb{C}}^{(1,3)}(X_k,\Lambda_k)=\eta_{\mathbb{C}}^{(3,3)}(X_k,\Lambda_k)$, where $(3,3)$ refers to the class of unitary matrices. Let $\eta(X_k,\Lambda_k)$ be the unstructured backward error. Then, as in position (3,3), we have

$$\eta(X_k,\Lambda_k)\le\eta_{\mathbb{C}}^{(1,3)}(X_k,\Lambda_k)\le\sqrt{2}\,\eta(X_k,\Lambda_k),$$

showing that forcing $A+E$ to be Hermitian and unitary has little effect on its norm.

Position (1,5): $\mathcal{C}_{\mathbb{C}}^{(1,5)}=\{A\in\mathbb{C}^{2n\times 2n}:\ A^*=A,\ JA=(JA)^*\}$ is the class of Hermitian Hamiltonian matrices. Note that $A\in\mathcal{C}_{\mathbb{C}}$ has the form $\left[\begin{smallmatrix}A_1 & A_2\\ A_2 & -A_1\end{smallmatrix}\right]$ with $A_1=A_1^*$ and $A_2=A_2^*$ and that the W-trick gives

$$W^*AW=\begin{bmatrix}0 & \widetilde{A}\\ \widetilde{A}^* & 0\end{bmatrix},\quad \widetilde{A}=A_1-iA_2.$$

Hence, using the W-trick, the constraints in (1.4) can be rewritten as

$$\widetilde{E}Y_{k,2}=S_{k,1},\qquad \widetilde{E}^*Y_{k,1}=S_{k,2},\qquad \widetilde{E}=E_1-iE_2\in\mathbb{C}^{n\times n},$$

because E is transformed in the same way as A. If $S_{k,2}^*Y_{k,2}=Y_{k,1}^*S_{k,1}$, then

$$\eta_{\mathbb{C}}^{(1,5)}(X_k,\Lambda_k)=\frac{\sqrt{2}}{\alpha}\|E_{\mathrm{opt}}\|_F,\quad \widetilde{E}_{\mathrm{opt}}=S_{k,1}Y_{k,2}^{+}+(S_{k,2}Y_{k,1}^{+})^*P_{Y_{k,2}}^{\perp},$$

using Lemma 2.5.

Note that if X_k and Λ_k are such that $X_k^*X_k=I$, $X_k^*JX_k=J$ and $J\Lambda_k=(J\Lambda_k)^*$ is Hamiltonian, then we can show that the assumption $S_{k,2}^*Y_{k,2}=Y_{k,1}^*S_{k,1}$ is satisfied.

Position (1,6): $\mathcal{C}_{\mathbb{C}}^{(1,6)}=\{A\in\mathbb{C}^{2n\times 2n}:\ A^*=A,\ JA=-(JA)^*\}$ is the class of Hermitian skew-Hamiltonian matrices. Note that $A\in\mathcal{C}_{\mathbb{C}}$ has the form $\left[\begin{smallmatrix}A_1 & A_2\\ -A_2 & A_1\end{smallmatrix}\right]$

with $A_1 = A_1^*$ and $A_2 = -A_2^*$ and that the W-trick diagonalizes A, $\widetilde{A} = W^*AW = \operatorname{diag}(\widetilde{A}_1, \widetilde{A}_2)$, where $\widetilde{A}_1 = A_1 + iA_2$ and $\widetilde{A}_2 = A_1 - iA_2$ are Hermitian. Hence, the $2n \times 2n$ skew-Hermitian problem can be reduced to two $n \times n$ Hermitian eigenproblems that can be solved independently. We refer to position (1,1) for the corresponding backward error.

Positions (1,7), (1,8): $A \in \mathbb{C}^{m\times m}$, $A^* = A$, and $A^T = A$ (or $A^T = -A$) imply that A is real symmetric (or iA is real skew-symmetric). Hence

$$\eta_{\mathbb{C}}^{(1,7)}(X_k, \Lambda_k) = \eta_{\mathbb{R}}^{(1,1)}(X_k, \Lambda_k), \qquad \eta_{\mathbb{C}}^{(1,8)}(X_k, \Lambda_k) = \eta_{\mathbb{R}}^{(2,2)}(X_k, i\Lambda_k).$$

Positions $(2, k)$, $k = 2\colon 12$: Each of these classes consists of matrices which are the scalar i times matrices in the corresponding classes in row 1. Hence

$$\begin{aligned}
\eta_{\mathbb{C}}^{(2,2)}(X_k, \Lambda_k) &= \eta_{\mathbb{C}}^{(1,1)}(X_k, i\Lambda_k), \qquad & \eta_{\mathbb{C}}^{(2,3)}(X_k, \Lambda_k) &= \eta_{\mathbb{C}}^{(1,3)}(X_k, i\Lambda_k),\\
\eta_{\mathbb{C}}^{(2,5)}(X_k, \Lambda_k) &= \eta_{\mathbb{C}}^{(1,6)}(X_k, i\Lambda_k), \qquad & \eta_{\mathbb{C}}^{(2,6)}(X_k, \Lambda_k) &= \eta_{\mathbb{C}}^{(1,5)}(X_k, i\Lambda_k),\\
\eta_{\mathbb{C}}^{(2,7)}(X_k, \Lambda_k) &= \eta_{\mathbb{C}}^{(1,7)}(X_k, i\Lambda_k), \qquad & \eta_{\mathbb{C}}^{(2,8)}(X_k, \Lambda_k) &= \eta_{\mathbb{C}}^{(1,8)}(X_k, i\Lambda_k).
\end{aligned}$$

Position (3,3): $\mathcal{C}_{\mathbb{C}}^{(3,3)} = \{A \in \mathbb{C}^{m\times m} : A^*A = I\}$ is the class of unitary matrices. We use Sun's approach [25] to derive $\eta_{\mathbb{C}}(X_k, \Lambda_k)$. First, we assume that the columns of X_k are orthonormal. As $X_k^*X_k = (X_k\Lambda_k)^*X_k\Lambda_k = I_k$, then from Lemma 2.6, solutions of $(A + E)X_k = X_k\Lambda_k$ with $A + E$ unitary exist and have the form

$$A + E = X_k\Lambda_kX_k^* + Q(I - X_kX_k^*) \tag{3.4}$$

with $Q \in \mathcal{C}_{\mathbb{C}}^{(3,3)}$ such that $QX_kX_k^* = X_kX_k^*Q$. Substituting $X_kX_k^* = X \operatorname{diag}(I_k, 0)X^*$, where $X = [X_k, \widetilde{X}]$ is unitary, into $QX_kX_k^* = X_kX_k^*Q$, yields

$$X^*QX \operatorname{diag}(I_k, 0) = \operatorname{diag}(I_k, 0)X^*QX$$

which implies $\widetilde{X}^*QX_k = 0$ and $X_k^*Q\widetilde{X} = 0$ or, equivalently,

$$Q = X \begin{bmatrix} Z_k & 0 \\ 0 & \widetilde{Z} \end{bmatrix} X^*, \quad Z = \operatorname{diag}(Z_k, \widetilde{Z}) \in \mathcal{C}_{\mathbb{C}}^{(3,3)}.$$

Hence, from (3.4)

$$E = X_k\Lambda_kX_k^* + \widetilde{X}\widetilde{Z}\widetilde{X}^* - A = [(X_k\Lambda_k - AX_k), (\widetilde{X}\widetilde{Z} - A\widetilde{X})]X^*$$

so that

$$\alpha^2\eta_{\mathbb{C}}^{(3,3)}(X_k, \Lambda_k)^2 = \|R\|_F^2 + \min_{\widetilde{Z}^*\widetilde{Z}=I} \|\widetilde{X}\widetilde{Z} - A\widetilde{X}\|_F^2. \tag{3.5}$$

Let U_kH_k and $\widetilde{U}\widetilde{H}$ be the polar decompositions of $X_k^*AX_k$ and $\widetilde{X}^*A\widetilde{X}$, respectively. The minimization problem in (3.5) is a well-known Procrustes problem [13, p. 149] whose solution is given by

$$\min_{\widetilde{Z}^*\widetilde{Z}=I} \|\widetilde{X}\widetilde{Z} - A\widetilde{X}\|_F^2 = \|\widetilde{X}\widetilde{U} - A\widetilde{X}\|_F^2.$$

By applying Lemma 2.1, then Lemma 2.2, and finally Lemma 2.1 again, we have

$$\begin{aligned}
\|\widetilde{X}\widetilde{U} - A\widetilde{X}\|_F^2 &= \|X_k^*A\widetilde{X}\|_F^2 + \|\widetilde{U} - \widetilde{X}^*A\widetilde{X}\|_F^2\\
&= \|\widetilde{X}^*AX_k\|_F^2 + \|U_k - X_k^*AX_k\|_F^2\\
&= \|X_kU_k - AX_k\|_F^2.
\end{aligned}$$

Hence

$$\eta_{\mathbb{C}}^{(3,3)}(X_k,\Lambda_k)=\alpha^{-1}\sqrt{\|R_k\|_F^2+\|X_kU_k-AX_k\|_F^2}\le\alpha^{-1}\sqrt{2}\,\|R_k\|_F=\sqrt{2}\,\eta(X_k,\Lambda_k),$$

where the inequality follows from

$$\|X_kU_k-AX_k\|_F=\min_{Z_k^*Z_k=I}\|X_kZ_k-AX_k\|_F\le\|X_k\Lambda_k-AX_k\|_F=\|R_k\|_F.$$

This is another example in which forcing the backward error matrix to be unitary has little effect on its norm.

Position (3,4): $\mathcal{C}_{\mathbb{C}}^{(3,4)}=\{A\in\mathbb{C}^{2n\times 2n}:\ A^*A=I,\ A^*JA=J\}$ is the class of symplectic unitary matrices. Matrices in this class have the form $A=\left[\begin{smallmatrix}A_1 & -A_2\\ A_2 & A_1\end{smallmatrix}\right]$ and are diagonalized by the W-trick, $W^*AW=\operatorname{diag}(\widetilde{A}_1,\widetilde{A}_2)$ with $\widetilde{A}_1=A_1-iA_2$, $\widetilde{A}_2=A_1+iA_2$ unitary. Hence, the $2n\times 2n$ original eigenvalue problem can be reduced to two $n\times n$ unitary eigenproblems that can be solved independently. Position (3,3) provides an explicit expression of the corresponding structured backward error.

Position (5,5): $\mathcal{C}_{\mathbb{C}}^{(5,5)}=\{A\in\mathbb{C}^{2n\times 2n}:\ JA=(JA)^*\}$ is the class of Hamiltonian matrices. The constraints in (1.4) can be rewritten as $JEX_k=JR_k$ with JE Hermitian. If

$$JR_kP_{X_k^*}=JR_k\quad\text{and}\quad P_{X_k}(JR_k)X_k^+\text{ is Hermitian,}\tag{3.6}$$

then

$$\eta_{\mathbb{C}}^{(5,5)}(X_k,\Lambda_k)=\alpha^{-1}\|E_{\mathrm{opt}}\|_F\quad\text{with}\quad E_{\mathrm{opt}}=R_kX_k^+ +(JR_kX_k^+J)^*P_{X_k}^{\perp}$$

using Lemma 2.3.

For a single approximate eigenpair (x,λ), the assumptions in (3.6) are always satisfied and, for x of unit 2-norm,

$$\eta_{\mathbb{C}}^{(5,5)}(x,\lambda)=\alpha^{-1}\sqrt{2\|r\|_2^2-\|x^*Jr\|_2^2}\le\sqrt{2}\,\eta(x,\lambda).$$

Hence, for a single eigenpair, forcing the backward error matrix to be Hamiltonian has little effect on its norm.

For a set of k approximate eigenpairs (X_k,Λ_k), if Λ_k is Hamiltonian, which implies that $k=2r$ is even and $\Lambda_k=\operatorname{diag}(\widetilde{\Lambda}_r,\widetilde{\Lambda}_r^*)$, and if $X_k^*JX_k=J$ with X_k of full rank, then we can show that the assumptions in (3.6) are satisfied and therefore $\eta_{\mathbb{C}}^{(5,5)}(X_k,\Lambda_k)$ is guaranteed to be finite.

Position (5,11): $\mathcal{C}_{\mathbb{C}}^{(5,11)}=\{A\in\mathbb{C}^{2n\times 2n}:\ JA=(JA)^*,\ JA=(JA)^T\}$. Matrices in this class are real and therefore

$$\eta_{\mathbb{C}}^{(5,11)}(X_k,\Lambda_k)=\eta_{\mathbb{R}}^{(5,5)}(X_k,\Lambda_k),$$

where $\eta_{\mathbb{R}}^{(5,5)}$ refer to position (5,5) of the real chart (see Table 3.3).

Position (5,12): $\mathcal{C}_{\mathbb{C}}^{(5,12)}=\{A\in\mathbb{C}^{2n\times 2n}:\ JA=(JA)^*,\ JA=-(JA)^T\}$. $A\in\mathcal{C}_{\mathbb{C}}$ implies that (iA) is real and satisfies $(J(iA))=(J(iA))^T$. Hence

$$\eta_{\mathbb{C}}^{(5,12)}(X_k,\Lambda_k)=\eta_{\mathbb{R}}^{(5,5)}(X_k,i\Lambda_k).$$

Positions $(6,j)$, $j=6:12$: Each of these classes consists of matrices which are the scalar i times matrices in the corresponding classes in row 5.

Position (7,7): $\mathcal{C}_{\mathbb{C}}^{(7,7)} = \{A \in \mathbb{C}^{m\times m} : \ A^T = A\}$ is the class of complex symmetric matrices. From Lemma 2.4 if $R_kX_k^+X_k = R_k$ and $\bar{X}_k\bar{X}_k{}^+R_kX_k^+ = (\bar{X}_k\bar{X}_k{}^+R_kX_k^+)^T$, then

$$\eta_{\mathbb{C}}^{(7,7)}(X_k,\Lambda_k) = \alpha^{-1}\|E_{\rm opt}\|_F \quad \text{with} \quad E_{\rm opt} = R_kX_k^+ + (R_kX_k^+)^TP_{X_k}^{\perp}.$$

Position (7,11): $\mathcal{C}_{\mathbb{C}}^{(7,11)} = \{A \in \mathbb{C}^{m\times m} : \ A^T = A, \ JA = (JA)^T\}$. Matrices in this class have the form $\left[\begin{smallmatrix} A_1 & A_2 \\ A_2 & -A_1 \end{smallmatrix}\right]$ with A_1, A_2 complex symmetric. The W-trick gives

$$W^*AW = \begin{bmatrix} 0 & \widetilde{A}_1 \\ \widetilde{A}_2 & 0 \end{bmatrix}, \quad \widetilde{A}_1 = \widetilde{A}_1^T, \ \widetilde{A}_2 = \widetilde{A}_2^T.$$

Hence using the W-trick, the constraints in (1.4) can be rewritten as

$$\widetilde{E}_1Y_{k,2} = S_{k,1}, \quad \widetilde{E}_2Y_{k,1} = S_{k,2}, \qquad \widetilde{E}_1 = \widetilde{E}_1^T, \ \widetilde{E}_2 = \widetilde{E}_2^T \in \mathbb{C}^{n\times n}. \tag{3.7}$$

If $S_{k,1}P_{Y_{k,2}^*} = S_{k,1}$ and $S_{k,2}P_{Y_{k,1}^*} = S_{k,2}$, and if $P_{\overline{Y}_{k,2}}S_{k,1}Y_{k,2}^+$ and $P_{\overline{Y}_{k,1}}S_{k,2}Y_{k,1}^+$ are complex symmetric, then

$$\eta_{\mathbb{C}}^{(7,11)}(X_k,\Lambda_k) = \alpha^{-1}\sqrt{\|\widetilde{E}_{1_{\rm opt}}\|_F^2 + \|\widetilde{E}_{2_{\rm opt}}\|_F^2},$$

where, using Lemma 2.4,

$$\widetilde{E}_{1_{\rm opt}} = S_{k,1}Y_{k,2}^+ + (S_{k,1}Y_{k,2}^+)^TP_{Y_{k,2}}^{\perp}, \quad \widetilde{E}_{1_{\rm opt}} = S_{k,2}Y_{k,1}^+ + (S_{k,2}Y_{k,1}^+)^TP_{Y_{k,1}}^{\perp}.$$

Position (7,12): $\mathcal{C}_{\mathbb{C}}^{(7,12)} = \{A \in \mathbb{C}^{m\times m} : \ A^T = A, \ JA = -(JA)^T\}$. Matrices in this class have the form $\left[\begin{smallmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{smallmatrix}\right]$ with A_1 complex symmetric and A_2 complex skew-symmetric, and they are diagonalized by the W-trick, $W^*AW = \operatorname{diag}(\widetilde{A}_1, \widetilde{A}_1^T)$. Then the $2n \times 2n$ original eigenvalue problem is reduced to one $n \times n$ eigenproblem with $\widetilde{A}_1$ of no particular structure. Hence, one can use the formula for the unstructured backward error in (1.3).

Position (8,8): $\mathcal{C}_{\mathbb{C}}^{(8,8)} = \{A \in \mathbb{C}^{m\times m} : \ A^T = -A\}$ is the class of complex skew-symmetric matrices. From Lemma 2.4, if $R_kX_k^+X_k = R_k$ and $\overline{X}_k\overline{X}_k^+R_kX_k^+ = -(\overline{X}_k\overline{X}_k^+R_kX_k^+)^T$, then the optimal solution to the constraints in (1.4) is given by

$$E_{\rm opt} = R_kX_k^+ - (R_kX_k^+)^TP_{X_k}^{\perp}$$

and then

$$\eta_{\mathbb{C}}^{(8,8)}(X_k,\Lambda_k) = \alpha^{-1}\|E_{\rm opt}\|_F.$$

Position (8,11): $\mathcal{C}_{\mathbb{C}}^{(8,11)} = \{A \in \mathbb{C}^{m\times m} : \ A^T = -A, \ JA = (JA)^T\}$. Matrices in this class have the form $\left[\begin{smallmatrix} A_1 & A_2 \\ -A_2 & A_1 \end{smallmatrix}\right]$ with $A_1^T = -A_1$ and $A_2^T = -A_2$. They are diagonalized by the W-trick, $W^*AW = \operatorname{diag}(\widetilde{A}_1, \widetilde{A}_2)$ with $\widetilde{A}_1$, $\widetilde{A}_2$ complex skew-symmetric. Hence, the $2n \times 2n$ original eigenvalue problem can be reduced to two $n \times n$ complex skew-symmetric eigenvalue problems that can be solved independently. We refer to position (8,8) for an explicit expression of the corresponding structured backward error.

TABLE 3.3
Summary of the structured backward errors for real matrices.

	1 $A^T = A$	2 $A^T = -A$	3 $A^TA = I$	4 $A^TJA = J$	5 $JA = (JA)^T$	6 $JA = -(JA)^T$
1 $A^T = A$	**X**	$\emptyset$	**X**	?	S,$\otimes$	S,$\otimes$
2 $A^T = -A$		**X**	?	?	S,$\otimes$	$\otimes$
3 $A^TA = I$			**X**	?	?	?
4 $A^TJA = J$				?	?	?
5 $JA = (JA)^T$					X	$\emptyset$
6 $JA = -(JA)^T$						X

X: explicit expression for $\eta_{\mathbb{R}}(X_k, \Lambda_k)$ is available and within a factor $\sqrt{2}$ of the unstructured backward error.
X: explicit expression for $\eta_{\mathbb{R}}(X_k, \Lambda_k)$ is available.
S: explicit backward error available for a single eigenpair (x, λ).
?: no explicit backward error known.
$\emptyset$: no nontrivial matrices with the prescribed pair of structures.
$\otimes$: expression available from Kronecker product approach. See Table 3.2 for the block structure of the matrices.

Position (8,12): $\mathcal{C}_{\mathbb{C}}^{(8,12)} = \{A \in \mathbb{C}^{2n\times 2n} : \ A^T = -A, \ JA = -(JA)^T\}$. Matrices in this class have the form $\left[\begin{smallmatrix} A_1 & A_2 \\ A_2 & -A_1 \end{smallmatrix}\right]$ with A_1, A_2 complex skew-symmetric. Using the W-trick, the constraints in (1.4) become

$$\widetilde{E}_1 Y_{k,2} = S_{k,1}, \qquad \widetilde{E}_2 Y_{k,1} = S_{k,2}, \qquad \widetilde{E}_1^T = -\widetilde{E}_1, \ \widetilde{E}_2^T = -\widetilde{E}_2 \in \mathbb{C}^{n\times n}$$

with $\widetilde{E}_1 = E_1 - iE_2$ and $\widetilde{E}_2 = E_1 + iE_2$. If the assumptions in Lemma 2.4 are satisfied, then

$$\eta_{\mathbb{C}}^{(8,12)}(X_k, \Lambda_k) = \sqrt{\|S_{k,1}Y_{k,2}^+ - (S_{k,1}Y_{k,2}^+)^T P_{Y_{k,2}}^{\perp}\|_F^2 + \|S_{k,2}Y_{k,1}^+ - (S_{k,2}Y_{k,1}^+)^T P_{Y_{k,1}^{\perp}}\|_F^2}.$$

Position (11,11): $\mathcal{C}_{\mathbb{C}}^{(11,11)} = \{A \in \mathbb{C}^{2n\times 2n} : \ JA = (JA)^T\}$ is the class of J-symmetric Hamiltonian matrices. If $R_k P_{X_k^*} = R_k$ and $P_{\bar{X}_k} JR_k X_k^+ = (P_{\bar{X}_k} JR_k X_k^+)^T$, then from Lemma 2.4

$$\eta_{\mathbb{C}}^{(11,11)}(X_k, \Lambda_k) = \alpha^{-1}\|JR_k X_k^+ + (JR_k X_k^+)^T P_{X_k^{\perp}}\|_F.$$

Position (12,12): $\mathcal{C}_{\mathbb{C}}^{(12,12)} = \{A \in \mathbb{C}^{2n\times 2n} : \ JA = -(JA)^T\}$ is the class of J-symmetric Hamiltonian matrices. If $R_k P_{X_k^*} = R_k$ and $P_{\bar{X}_k} JR_k X_k^+ = -(P_{\bar{X}_k} JR_k X_k^+)^T$, then from Lemma 2.4

$$\eta_{\mathbb{C}}^{(12,12)}(X_k, \Lambda_k) = \alpha^{-1}\|JR_k X_k^+ - (JR_k X_k^+)^T P_{X_k^{\perp}}\|_F.$$

3.2.2. Real chart ($\mathbb{K} = \mathbb{R}$). When the matrix of the structured eigenvalue problem is real, it is natural to consider perturbation matrices E that are real, too. This problem is addressed in this section and the results are summarized in Table 3.3. The W-trick cannot be used since the transformation with W would send our real problem to the complex space. We have to use the Kronecker product approach instead.

Position (1,1): $\mathcal{C}_{\mathbb{R}}^{(1,1)} = \{A \in \mathbb{R}^{m\times m} : A^T = A\}$ is the class of real symmetric matrices. For Λ_k and X_k real and such that $X_k^T X_k = I$, we have $X_k^+ = X_k^T$ so that $R_k P_{X_k^T} = R_k$ and $P_{X_k} R_k X_k^T = X_k \Lambda_k X_k^T - X_k X_k^T A X_k X_k^T$ is symmetric. Hence, from Lemma 2.3 applied with $\mathbb{K} = \mathbb{R}$ the optimal solution to $EX_k = R_k$ with $E^T = E$ is given by

$$E_{\text{opt}} = R_k X_k^T + (X_k R_k^T) P_{X_k}^{\perp}$$

so that

$$\eta_{\mathbb{R}}^{(1,1)}(X_k, \Lambda_k) = \alpha^{-1}\sqrt{\text{trace}(E_{\text{opt}}^T E_{\text{opt}})} = \alpha^{-1}\sqrt{2\|R_k\|_F^2 - \|X_k^T R_k\|_F^2}$$

and, as in the complex case,

$$\eta(X_k, \Lambda_k) \le \eta_{\mathbb{R}}^{(1,1)}(X_k, \Lambda_k) \le \sqrt{2}\,\eta(X_k, \Lambda_k).$$

Position (1,3): $\mathcal{C}_{\mathbb{R}}^{(1,3)} = \{A \in \mathbb{R}^{m\times m} : \ A^T = A, A^T A = I\}$ is the class of symmetric unitary matrices. As all the eigenvalues are ± 1, we can take X_k real and apply Lemma 2.6 with $\mathbb{K} = \mathbb{R}$. The derivation for the backward error for position (1,2) of the complex chart remains valid in real arithmetic and, therefore,

$$\eta_{\mathbb{R}}^{(1,3)}(X_k, \Lambda_k) = \alpha^{-1}\sqrt{\|R\|_F^2 + \|X_k U_k - A X_k\|_F^2},$$

where U_k is the orthogonal factor of the polar factorization of $X_k^T A X_k$.

Positions (1,5): $\mathcal{C}_{\mathbb{R}}^{(1,5)} = \{A \in \mathbb{R}^{2n\times 2n} : \ A^T = A, \ JA = (JA)^T\}$ is the class of symmetric Hamiltonian matrices. The backward error for this problem is considered in [27], where it is shown that for a single eigenpair (x, λ) with x of unit 2-norm,

$$\eta_{\mathbb{R}}^{(1,5)}(x, \lambda) = \alpha^{-1}\sqrt{2\|r\|_2^2 + 2(e_2^T Q^T r)^2},$$

with $e_2 = [0, 1, 0, \ldots, 0]^T$, $r = (\lambda I - A)x$, and Q the orthogonal factor in the symplectic quasi-QR factorization of $[x \ r]$. For a set of eigenpairs, an explicit expression for $\eta_{\mathbb{R}}^{(1,5)}(X_k, \Lambda_k)$ is obtained through the Kronecker product approach.

Positions (1,6): $\mathcal{C}_{\mathbb{R}}^{(1,6)} = \{A \in \mathbb{R}^{2n\times 2n} : \ A^T = A, \ JA = -(JA)^T\}$ is the class of symmetric skew-Hamiltonian matrices. The structured backward error for this class of problems is also considered in [27], where it is shown that for a single eigenpair (x, λ) with x of unit 2-norm,

$$\eta_{\mathbb{R}}^{(1,5)}(x, \lambda) = \alpha^{-1}\sqrt{2\|r\|_2^2 + 2(e_2^T \widetilde{Q}^T r)^2},$$

with $e_2 = [0, 1, 0, \ldots, 0]^T$, $r = (\lambda I - A)x$, and $\widetilde{Q}$ the orthogonal factor in the symplectic quasi-QR factorization of $[Jx \ r]$. For a set of eigenpairs, we need to use the Kronecker product approach.

Position (2,2): $\mathcal{C}_{\mathbb{R}}^{(2,2)} = \{A \in \mathbb{R}^{m\times m} : \ A^T = -A\}$ is the class of real skew-symmetric matrices. We assume that the spectrum of Λ_k is symmetric with respect to the real axis and that X_k has orthonormal columns. There exists a $k\times k$ unitary matrix N such that $Y_k = X_k N \in \mathbb{R}^{m\times k}$ and $\Omega_k = N^* \Lambda_k N \in \mathbb{R}^{k\times k}$ is block diagonal with 1×1 blocks equal to 0 and 2×2 blocks of the form $\left[\begin{smallmatrix} 0 & -\omega_i \\ \omega_i & 0 \end{smallmatrix}\right]$. We have $\eta_{\mathbb{R}}^{(2,2)}(X_k, \Lambda_k) =$

$\eta_{\mathbb{R}}^{(2,2)}(Y_k,\Omega_k)$. Let $\widetilde{R}_k = Y_k\Omega_k - AY_k$. Since $\widetilde{R}_k P_{Y_k^T} = \widetilde{R}_k$ and $P_{Y_k}\widetilde{R}_k Y_k^T$ is skew-symmetric, Lemma 2.4 applies with $\mathbb{K} = \mathbb{R}$. The optimal solution to $EY_k = \widetilde{R}_k$ with $E^T = -E$ is given by

$$E_{\text{opt}} = \widetilde{R}_k Y_k^T + (Y_k \widetilde{R}_k^T) P_{Y_k}^{\perp}$$

so that

$$\eta_{\mathbb{R}}^{(2,2)}(X_k,\Lambda_k) = \alpha^{-1}\sqrt{2\|\widetilde{R}_k\|_F^2 - \|Y_k^T\widetilde{R}_k\|_F^2} = \alpha^{-1}\sqrt{2\|R_k\|_F^2 - \|X_k^T R_k\|_F^2}.$$

Hence $\eta_{\mathbb{R}}^{(2,2)}(X_k,\Lambda_k) \le \sqrt{2}\,\eta(X_k,\Lambda_k)$, showing that forcing the backward error matrix to be real skew-symmetric has little effect on its norm.

Positions (3,3): $\mathcal{C}_{\mathbb{R}}^{(3,3)} = \{A \in \mathbb{R}^{m\times m} : \ A^TA = I\}$ is the class of orthogonal matrices. We assume that the spectrum of Λ_k is symmetric with respect to the real axis and that X_k has orthonormal columns. There exists a unitary $k \times k$ matrix N such that $Y_k = X_kN \in \mathbb{R}^{m\times k}$ and $\Omega_k = N^*\Lambda_k N \in \mathbb{R}^{k\times k}$ is block diagonal with 1×1 blocks equal to ± 1 and 2×2 blocks of the form $\left[\begin{smallmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{smallmatrix}\right]$, $\sin\theta \neq 0$. We have $\eta_{\mathbb{R}}^{(3,3)}(X_k,\Lambda_k) = \eta_{\mathbb{R}}^{(3,3)}(Y_k,\Omega_k)$. With (X_k,Λ_k) replaced by (Y_k,Ω_k), the technique described in position $(3,3)$ constructs a real solution E_{opt} of minimal Frobenius norm to the constraints in (1.4). Finally, we end up with $\eta_{\mathbb{R}}^{(3,3)}(X_k,\Lambda_k) = \eta_{\mathbb{C}}^{(3,3)}(X_k,\Lambda_k)$.

Positions (2,5): $\mathcal{C}_{\mathbb{R}}^{(2,5)} = \{A \in \mathbb{R}^{2n\times 2n} : \ A^T = -A,\ JA = (JA)^T\}$ is the class of skew-symmetric Hamiltonian matrices. We assume that λ is purely imaginary and $x = \left[\begin{smallmatrix}x_1\\ x_2\end{smallmatrix}\right]$ with $x_2 = \pm i x_1$ has unit 2-norm. It is shown in [27] that

$$\eta_{\mathbb{R}}^{(2,5)}(x,\lambda) = \alpha^{-1}\sqrt{2\|s\|_2^2 + 2(e_2^TQ^Ts)^2},$$

where $e_2 = [0,1,0,\ldots,0]^T$ and Q is the orthogonal factor in the symplectic quasi-QR factorization of $[w\ s] = [\,I \quad -(A+\sigma i\lambda I\,]\left[\begin{smallmatrix}\operatorname{Re}(x_1)\\ -\sigma\operatorname{Im}(x_1)\end{smallmatrix}\right]$ with $\sigma = 1$ if $x_2 = ix_1$ or $\sigma = -1$ otherwise. The computation of $\eta_{\mathbb{R}}^{(2,5)}(x,\lambda)$ can be done in $O(n^2)$ operations. For a set of eigenpairs, we refer to the Kronecker product approach.

Position (5,5): $\mathcal{C}_{\mathbb{R}}^{(5,5)} = \{A \in \mathbb{R}^{2n\times 2n} : \ JA = (JA)^T\}$ is the class of Hamiltonian matrices. We assume that $k \le n$. The constraints in (1.4) can be rewritten as $JE\widetilde{X}_{2k} = J\widetilde{R}_{2k}$, $JE = (JE)^T$, where $\widetilde{X}_{2k} = [\operatorname{Re}(X_k)\ \operatorname{Im}(X_k)]$, $\widetilde{R}_{2k} = [\operatorname{Re}(R_k)\ \operatorname{Im}(R_k)]$. If

$$\widetilde{R}_{2k}P_{\widetilde{X}_{2k}}^T = \widetilde{R}_{2k} \quad \text{and} \quad P_{\widetilde{X}_{2k}}J\widetilde{R}_{2k}\widetilde{X}_{2k}^T = (P_{\widetilde{X}_{2k}}J\widetilde{R}_{2k}\widetilde{X}_{2k}^T)^T,$$

then $\eta_{\mathbb{R}}^{(5,5)}(X_k,\Lambda_k) = \alpha^{-1}\|E_{\text{opt}}\|_F$ where, using Lemma 2.3,

$$E_{\text{opt}} = \widetilde{R}_{2k}\widetilde{X}_{2k}^{+} + J(\widetilde{R}_{2k}\widetilde{X}_{2k}^{+})^TJP_{\widetilde{X}_{2k}}^{\perp} \in \mathbb{R}^{2n\times 2n}.$$

Position (6,6): $\mathcal{C}_{\mathbb{R}}^{(6,6)} = \{A \in \mathbb{R}^{2n\times 2n} : \ JA = -(JA)^T\}$ is the class of skew-Hamiltonian matrices. We assume that $k \le n$. The constraints in (1.4) can be rewritten as $JE\widetilde{X}_{2k} = J\widetilde{R}_{2k}$, $JE = (JE)^T$, where

$$\widetilde{X}_{2k} = [\operatorname{Re}(X_k),\ \operatorname{Im}(X_k)], \qquad \widetilde{R}_{2k} = [\operatorname{Re}(R_k),\ \operatorname{Im}(R_k)].$$

If $\widetilde{R}_{2k}P_{\widetilde{X}_{2k}}^T = \widetilde{R}_{2k}$, and if $P_{\widetilde{X}_{2k}}J\widetilde{R}_{2k}\widetilde{X}_{2k}^T$ is skew-symmetric, then using Lemma 2.4 we obtain

$$E_{\text{opt}} = \widetilde{R}_{2k}\widetilde{X}_{2k}^{+} - J(\widetilde{R}_{2k}\widetilde{X}_{2k}^{+})^TJP_{\widetilde{X}_{2k}}^{\perp} \quad \text{and} \quad \eta_{\mathbb{R}}^{(6,6)}(X_k,\Lambda_k) = \alpha^{-1}\|E_{\text{opt}}\|_F.$$

4. Structured normwise condition numbers. The condition number characterizes the sensitivity of solutions to problems. If λ is a simple, nonzero eigenvalue of a singly or doubly structured matrix $A \in \mathcal{C}_{\mathbb{K}}$, with corresponding right eigenvector x and left eigenvector y, then a structured normwise condition number of λ can be defined as follows:

$$\kappa_{\mathbb{K}}(\lambda) := \lim_{\epsilon \to 0} \sup \left\{ \frac{|\Delta\lambda|}{\epsilon|\lambda|} : (A+E)(x+\Delta x) = (\lambda+\Delta\lambda)(x+\Delta x), \right. \\ \left. A + E \in \mathcal{C}_{\mathbb{K}},\ \|E\|_F \le \epsilon\alpha \right\}, \tag{4.1}$$

where α is a positive parameter. The forward error, condition number, and backward error are related by the inequality (correct to first order in the backward error)

$$\textit{forward error} \ \le \ \textit{condition number} \ \times \ \textit{backward error}.$$

In this section, we consider only linear structure in A. Expanding the first constraint in (4.1) and premultiplying by y^* lead to

$$\Delta\lambda = \frac{y^*Ex}{y^*x} + O(\epsilon^2).$$

To evaluate $\kappa_{\mathbb{K}}(\lambda)$ we need to obtain a sharp bound for the first term in this expansion. If the structure is linear, then with the same notation as in section 3.1 we have

$$Ex = \operatorname{vec}(Ex) = (x^T \otimes I_m)\operatorname{vec}(E) = (x^T \otimes I_m)B\Delta p = MD\Delta p,$$

where $\operatorname{vec}(E) = B\Delta p$, $M = (x^T \otimes I_m)BD^{-1}$, and D is such that $\|E\|_F = \|D\Delta p\|_2$. Hence,

$$|y^*Ex| = \|y^*MD\Delta p\|_2 \le \|y^*M\|_2\|E\|_F = \|y^*M\|_2\|D\Delta p\|_2.$$

Equality is obtainable for a suitable Δp because equality is always possible in the Cauchy–Schwarz inequality. Therefore

$$\kappa_{\mathbb{K}}(\lambda) = \alpha\frac{\|y^*M\|_2}{|\lambda||y^*x|}. \tag{4.2}$$

Acknowledgments. I thank the referees for valuable suggestions that improved the paper.

REFERENCES

[1] P. Benner, R. Byers, and E. Barth, *Algorithm* 800: *Fortran* 77 *subroutines for computing the eigenvalues of Hamiltonian matrices* I: *The square-reduced method*, ACM Trans. Math. Softw., 26 (2000), pp. 49–77.

[2] P. Benner and H. Fassbender, *An implicitly restarted symplectic Lanczos method for the Hamiltonian eigenvalue problem*, Linear Algebra Appl., 263 (1997), pp. 75–111.

[3] P. Benner and H. Fassbender, *The symplectic eigenvalue problem, the butterfly form, the SR algorithm, and the Lanczos method*, Linear Algebra Appl., 275/276 (1998), pp. 19–47.

[4] P. Benner, V. Mehrmann, and H. Xu, *A new method for computing the stable invariant subspace of a real Hamiltonian matrix*, J. Comput. Appl. Math., 86 (1997), pp. 17–43.

[5] P. Benner, V. Mehrmann, and H. Xu, *A numerically stable, structure preserving method for computing the eigenvalues of real Hamiltonian or symplectic pencils*, Numer. Math., 78 (1998), pp. 329–358.
[6] 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.
[7] A. Bunse-Gerstner, *Matrix factorizations for symplectic QR-like methods*, Linear Algebra Appl., 83 (1986), pp. 49–77.
[8] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *A chart of numerical methods for structured eigenvalue problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 419–453.
[9] A. Deif, *A relative backward perturbation theorem for the eigenvalue problem*, Numer. Math., 56 (1989), pp. 625–626.
[10] J. E. Dennis, Jr. and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.
[11] J. E. Dennis, Jr. and Robert B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[12] H. Fassbender, D. S. Mackey, and N. Mackey, *Hamilton and Jacobi come full circle: Jacobi algorithms for structured Hamiltonian problems*, Linear Algebra Appl., 332/334 (2001), pp. 37–80.
[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[14] W. B. Gragg and L. Reichel, *A divide and conquer method for unitary and orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.
[15] 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.
[16] 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.
[17] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 (1994), pp. 3–20.
[18] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
[19] W. Kahan, B. N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal., 19 (1982), pp. 470–484.
[20] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, London, 1985.
[21] V. Mehrmann and D. Watkins, *Structure-preserving methods for computing eigenpairs of large sparse skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Sci. Comput., 22 (2001), pp. 1905–1925.
[22] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, London, 1990.
[23] M. Stewart, *Stability properties of several variants of the unitary Hessenberg QR algorithm*, in Structured Matrices in Mathematics, Computer Science, and Engineering. II (Boulder, CO, 1999), V. Olshevsky, ed., AMS, Providence, RI, 2001, pp. 57–72.
[24] J.-G. Sun, *Backward perturbation analysis of certain characteristic subspaces*, Numer. Math., 65 (1993), pp. 357–382.
[25] J.-G. Sun, *Backward Errors for the Unitary Eigenproblem*, Tech. Report UMINF-97.25, Department of Computing Science, University of Umeå, Umeå, Sweden, 1997.
[26] J.-G. Sun, *Stability and Accuracy: Perturbation Analysis of Algebraic Eigenproblems*, Tech. Report UMINF 98-07, Department of Computing Science, University of Umeå, Umeå, Sweden, 1998.
[27] F. Tisseur, *Stability of structured Hamiltonian eigensolvers*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 103–125.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 899–913

FAST LOCAL RECONSTRUCTION METHODS FOR NONUNIFORM SAMPLING IN SHIFT-INVARIANT SPACES*

KARLHEINZ GRÖCHENIG† AND HARALD SCHWAB‡

Abstract. We present a new method for the fast reconstruction of a function f from its samples $f(x_j)$ under the assumption that f belongs to a shift-invariant space $V(\varphi)$. If the generator φ has compact support, then the reconstruction is local, quite in contrast to methods based on band-limited functions. Using frame theoretic arguments, we show that the matrix of the corresponding linear system of equations is a positive-definite banded matrix. This special structure makes possible the fast local reconstruction algorithm in $O(S^2J)$ operations, where J is the number of samples and S is the support length of the generator φ. Further optimization can be achieved by means of data segmentation. Ample numerical simulation is provided.

Key words. shift-invariant space, nonuniform sampling, banded matrix, localization, data segmentation, denoising

AMS subject classifications. 41A15, 42C15, 46A35, 46E15, 46N99, 47B37

PII. S0895479802409067

1. Introduction.

Shift-invariant spaces serve as a universal model for uniform and nonuniform sampling of functions. The objective of the so-called sampling problem is either to recover a signal (function) f from its samples $\{f(x_j) : j \in \mathbb{Z}\}$ or to approximate a data set (x_j, y_j) by a suitable function f satisfying $f(x_j) \approx y_j$. Obviously this problem is ill-posed, and so a successful reconstruction requires some a priori information about the signal. Usually it is assumed that f is contained in the span of integer translates of a given generator φ. In technical terms, the original function f has the form $f(x) = \sum_{k\in\mathbb{Z}} c_k\varphi(x-k)$ and belongs to the shift-invariant space $V(\varphi)$.

Until recently the only choice for φ was the cardinal sine function $\varphi(x) = \frac{\sin \pi\alpha x}{\pi\alpha x}$, since in this case $V(\varphi)$ coincides with the band-limited functions of bandwidth 2α. Despite the existence of fast numerical methods [9], this model has some drawbacks because it is nonlocal, and the behavior of f at a point x also depends on samples far away from x. For this reason, one works with truncated versions of the cardinal sine. This idea leads naturally to work in shift-invariant spaces with a generator φ of compact support.

The concept of shift-invariant spaces first arose in approximation theory and wavelet theory [5, 6, 15]. Its potential for the systematic treatment of sampling problems was recognized much later. We refer to [1] for a detailed survey of the state-of-the-art and an extensive list of references.

Our goal is the investigation of specific numerical issues of nonuniform sampling in shift-invariant spaces. Usually the reconstruction algorithms are based on general frame methods, or they follow simple iterative schemes. Of course, these can be applied successfully in the context of shift-invariant spaces as well; see sections 6

*Received by the editors May 24, 2002; accepted for publication by M. Hanke June 10, 2002; published electronically February 12, 2003. This research was supported by the Austrian Science Fund project FWF P-14485.
http://www.siam.org/journals/simax/24-4/40906.html
†Department of Mathematics, The University of Connecticut, Storrs, CT 06269-3009 (groch@math.uconn.edu).
‡NUHAG, Department of Mathematics, University of Vienna, Strudlhofg. 4, A-1090 Vienna, Austria (harald.schwab@univie.ac.at).

and 7 of [1]. We adopt the philosophy that the optimal numerical algorithms always have to use the special structure of the problem. Hence the general purpose algorithms have to be fine-tuned to achieve their optimal performance. Here we use the peculiar structure of shift-invariant spaces with compactly supported generator to solve the sampling problem.

We want to reconstruct a function $f \in V(\varphi)$ from a *finite* number of samples $f(x_j)$ taken from an interval $x_j \in [M_0, M_1]$. In our derivation of the special structure we use a frame theoretic argument and combine it with the fact that the generator φ has compact support. The resulting algorithm is local in the sense that the complete reconstruction of a function in $V(\varphi)$ on the interval $[M_0, M_1]$ requires samples only from $[M_0, M_1]$ (quite in contrast to band-limited functions). By comparing reconstructions in different spline spaces we find that the algorithm can be used as an efficient denoising procedure for noisy samples.

We will assume that the reader is familiar with the short section on frames in [8] or in [5], so we will not define explicitly these standard concepts.

The paper is organized as follows: In section 2 we present the precise technical details of shift-invariant spaces, state the well-known equivalence of sampling problems with the construction of certain frames, and discuss some general reconstruction techniques associated to frames. In section 3 we exploit the special structure of shift-invariant spaces to derive a local reconstruction algorithm of order $O(J)$ and discuss the numerical issues involved. Section 4 explains the results of the numerical simulations and provides the pseudocode of the main algorithm.

2. Shift-invariant spaces and sampling: General theory.

2.1. Shift-invariant spaces, frames, and sampling. Let φ be a continuous function with compact support of size S so that

$$\text{supp } \varphi \subseteq [-S, S]\,. \tag{1}$$

For convenience we assume that S is a positive integer and thus $\varphi(\pm S) = 0$. Then the shift-invariant space $V(\varphi)$ is defined as

$$V(\varphi) = \Big\{ f \in L^2(\mathbb{R}) : f(x) = \sum_{k\in\mathbb{Z}} c_k \varphi(x-k) \text{ for } (c_k) \in \ell^2(\mathbb{Z}) \Big\}\,. \tag{2}$$

To guarantee the stability of these representations, we assume that the generator φ is stable, which means that there exists a constant $C > 0$ such that

$$C^{-1}\,\|c\|_{\ell^2} \le \Big\| \sum_{k\in\mathbb{Z}} c_k \varphi(\cdot - k) \Big\|_2 \le C\,\|c\|_{\ell^2} \tag{3}$$

for all finite sequences $c = (c_k)_{k\in\mathbb{Z}}$, or, equivalently, the translates $\phi(.-k)$, $k \in \mathbb{Z}$, form a Riesz basis for $V(\varphi)$. As a consequence, $V(\varphi)$ is a closed subspace of $L^2(\mathbb{R})$ and inherits the inner product $\langle .\,, . \rangle$ of $L^2(\mathbb{R})$.

The sampling problem in $V(\varphi)$ is as follows: Given a set of sampling points $X = \{x_j : j \in \mathbb{Z}\}$ arranged in increasing order $x_j < x_{j+1}$ and a sequence of samples $\{f(x_j) : j \in \mathbb{Z}\}$ of a function $f \in V(\varphi)$, we would like to recover the original function f in a stable and numerically efficient way. Here stability means that there exist (possibly unspecified) constants $A, B > 0$ such that

$$A\|f\|_2 \le \Big(\sum_{j\in\mathbb{Z}} |f(x_j)|^2 \Big)^{1/2} \le B\|f\|_2 \qquad \forall f \in V(\varphi)\,. \tag{4}$$

A sampling set satisfying (4) is called a *set of stable sampling.*

Obviously, for (4) to be valid, we need point evaluations in $V(\varphi)$ to be well-defined. This is guaranteed by the following lemma [1].

LEMMA 1. *If φ is continuous and satisfies the condition $\sum_{k\in\mathbb{Z}} \max_{x\in[0,1]} |f(x+k)| < \infty$, in particular, if φ is continuous with compact support, then for all $x \in \mathbb{R}$ there exists a function $K_x \in V(\varphi)$ such that $f(x) = \langle f, K_x\rangle$ for $f \in V(\varphi)$. We say that $V(\varphi)$ is a reproducing kernel Hilbert space.*

Explicit formulas for K_x are known (see [1]), but we do not need them here. Note that with the kernels K_x the sampling inequality (4) can be formulated as $A\|f\|_2 \le (\sum_{j\in\mathbb{Z}} |\langle f, K_{x_j}\rangle|^2)^{1/2} \le B\|f\|_2$, which is equivalent to saying that the set $\{K_{x_j} : j \in \mathbb{Z}\}$ is a frame for $V(\varphi)$.

Let U be the infinite matrix with entries

$$U_{jk} = \varphi(x_j - k), \qquad j,k \in \mathbb{Z}\,. \tag{5}$$

Then the sampling problem in $V(\varphi)$ can be formulated in several distinct ways [2, Prop. 1.3].

LEMMA 2. *If φ satisfies the condition of Lemma* 1*, then the following are equivalent:*

(i) *$X = \{x_j : j \in \mathbb{Z}\}$ is a set of sampling for $V(\varphi)$.*

(ii) *There exist $A, B > 0$ such that*

$$A\|c\|_{\ell^2} \le \|Uc\|_{\ell^2} \le B\|c\|_{\ell^2} \qquad \forall c \in \ell^2(\mathbb{Z})\,.$$

(iii) *The set of reproducing kernels $\{K_{x_j} : j \in \mathbb{Z}\}$ is a frame for $V(\varphi)$.*

Remark. It is difficult to characterize sets of sampling for $V(\varphi)$. If φ is a B-spline of order N, i.e., $\varphi = \chi_{[0,1]} * \cdots * \chi_{[0,1]}$ ($N+1$ convolutions), then the main result of [2] implies that the maximum gap condition $\sup_{j\in\mathbb{Z}}(x_{j+1} - x_j) = \delta < 1$ is sufficient for the conditions of Lemma 2 to hold.

2.2. General purpose reconstructions. Lemma 2 leads to some general reconstruction techniques that are always applicable.

1. *Linear algebra solution.* One could simply try to solve the (infinite) system of linear equations

$$\sum_{k\in\mathbb{Z}} c_k \varphi(x_j - k) = f(x_j) \qquad \forall j \in \mathbb{Z} \tag{6}$$

for the coefficients (c_k), or, in the notation of (5) with $f|_X = (f(x_j))_{j\in\mathbb{Z}}$,

$$Uc = f|_X\,. \tag{7}$$

2. *The normal equations.* Frequently it is better to consider the associated system of normal equations [10]

$$U^*U\,c = U^* f|_X\,. \tag{8}$$

This approach has the advantage that the matrix $T := U^*U$ is a positive operator on $\ell^2(\mathbb{Z})$. Furthermore, if the input $y = (y_j)_{j\in\mathbb{Z}}$ does not consist of a sequence of exact samples of $f \in V(\varphi)$, then the function $f = \sum_{k\in\mathbb{Z}} c_k \varphi(\cdot - k)$ corresponding to the solution $c = (U^*U)^{-1}U^*y$ solves the least squares problem

$$\sum_{j\in\mathbb{Z}} |y_j - f(x_j)|^2 = \min_{h\in V(\varphi)} \sum_{j\in\mathbb{Z}} |y_j - h(x_j)|^2\,. \tag{9}$$

3. *Frame approach.* Lemma 2(iii) suggests using versions of the frame algorithm to find a reconstruction of f. By (iii) the frame operator which is defined as

$$Sf(x) = \sum_{j\in\mathbb{Z}} \langle f, K_{x_j}\rangle K_{x_j}(x) = \sum_{j\in\mathbb{Z}} f(x_j) K_{x_j}(x) \tag{10}$$

is invertible and its inverse defines the dual frame $\widetilde{K_{x_j}} = S^{-1}K_{x_j}$, $j \in \mathbb{Z}$. Then the reconstruction is given by

$$f(x) = \sum_{j\in J} \langle f, K_{x_j}\rangle \widetilde{K_{x_j}} = \sum_{j\in J} f(x_j)\widetilde{K}_{x_j}(x)\,. \tag{11}$$

We observe that the linear algebra solution (7) and the frame method are equivalent. By definition the vector of samples is given by $Uc = f|_X$. The sampled energy of $f \in V(\varphi)$ is

$$\sum_{j\in\mathbb{Z}} |f(x_j)|^2 = \langle f|_X, f|_X\rangle_{\ell^2} = \langle Uc, Uc\rangle_{\ell^2} = \langle U^*Uc, c\rangle_{\ell^2}\,. \tag{12}$$

Thus X is a set of sampling if and only if U^*U is invertible on $\ell^2(\mathbb{Z})$.

4. *Iterative frame methods.* In nonuniform sampling problems it is usually difficult to calculate the entire dual frame; therefore one often resorts to iterative methods. Since the Richardson–Landweber iteration in the original paper of Duffin and Schaeffer [8] is slow and requires good estimates of the frame bounds, we recommend the conjugate gradient acceleration of the frame algorithm for all problems without additional structure [11]. It converges optimally and does not require the estimate of auxiliary parameters.

3. Exploiting the structure of the problem.

So far we have discussed general purpose methods for the reconstruction of the function. These could be applied in any situation involving frames and do not take into consideration the particular structure of the sampling problem in shift-invariant spaces.

3.1. A localization property.

We now exploit the special structure of shift-invariant spaces. The following lemma is simple but crucial. It is a consequence of the assumption that the generator of $V(\varphi)$ has compact support.

LEMMA 3. *If* supp $\varphi \subseteq [-S, S]$, *then* $T = U^*U$ *is a band matrix of (upper and lower) bandwidth* $2S$.

Proof. By definition the entries of U^*U are

$$(U^*U)_{kl} = \sum_{j\in\mathbb{Z}} \overline{U_{jk}} U_{jl} = \sum_{j\in\mathbb{Z}} \overline{\varphi(x_j - k)}\ \varphi(x_j - l)\,.$$

Since φ has compact support, the sum is always locally finite, and its convergence does not pose any problem. Since $\varphi(x_j - k) = 0$ if $|x_j - k| \geq S$, we find that $(U^*U)_{kl}$ can be nonzero only if both $|x_j - k| < S$ and $|x_j - l| < S$. In other words, $(U^*U)_{kl} \neq 0$ implies that

$$|k - l| \leq |k - x_j| + |x_j - l| < 2S\,.$$

This means that only $4S - 1$ diagonals of U^*U contain nonzero entries. □

Remarks. 1. Banded matrices and the resulting numerical advantages occur in a number of related problems. For instance, in the interpolation of scattered data

by radial functions with compact support, the interpolation matrix is banded; see [3] and the references therein. Likewise, the calculation of the optimal smoothing spline on a finite set of arbitrary nodes requires the inversion of a banded matrix [12].

2. Lemma 3 combined with a result of Demko, Moss, and Smith [7] or of Jaffard [13] implies that the inverse matrix possesses exponential decay off the diagonal, i.e., there exist $C, A > 0$ such that

$$|(U^*U)^{-1}_{kl}| \leq Ce^{-A|k-l|} \qquad \forall k, l \in \mathbb{Z}\,.$$

To make our treatment more realistic, we take into account that in any real problem only a finite (albeit large) number of samples is given. It turns out that the model of shift-invariant spaces with compactly supported generator possesses excellent localization properties. These are quantified in the next lemma.

LEMMA 4. *The restriction of $f \in V(\varphi)$ to the interval $[M_0, M_1]$ is determined completely by the coefficients c_k for $k \in (M_0 - S, M_1 + S) \cap \mathbb{Z}$.*

Proof. Since $\varphi(x-k) = 0$ for $|x-k| \geq S$ and $x \in [M_0, M_1]$, we obtain that

$$M_0 - S \leq x - S < k < x + S \leq M_1 + S\,.$$

Consequently, as $S \in \mathbb{N}$, we have

$$\begin{aligned} f(x) &= \sum_{k\in\mathbb{Z}} c_k\varphi(x-k) = \sum_{|x-k|<S} c_k\varphi(x-k) \\ &= \sum_{k=M_0-S+1}^{M_1+S-1} c_k\varphi(x-k)\,. \qquad \square \end{aligned}$$

In other words, the exact reconstruction of $f \in V(\varphi)$ on $[M_0, M_1]$ requires only the $M_1 - M_0 + 2S - 1$ unknown coefficients c_k with $k \in (M_0 - S, M_1 + S) \cap \mathbb{Z}$. By counting dimensions, we find that we need at least $M_1 - M_0 + 2S - 1$ samples in $[M_0, M_1]$ for the coefficients to be determined uniquely. Usually the length $M_1 - M_0$ is large compared to S; therefore the additional $2S-1$ coefficients amount to a negligible oversampling.

Lemma 4 demonstrates an important theoretical and practical advantage of shift-invariant spaces with compactly supported generators. *A function $f \in V(\varphi)$ can be reconstructed exactly on an arbitrary interval solely from samples in that interval.* In contrast, the restriction of a band-limited function to an interval is *not* uniquely determined by any finite number of samples in that interval but can only be approximated by these samples. The localization property expressed in Lemma 4 is one of the main reasons for working with shift-invariant spaces with compactly supported generators as a sampling model!

Finally we remark that uniform sampling at critical density is not local and may even be unstable in this model. If $f \in V(\varphi)$ is sampled at $\xi + k$, $k \in \mathbb{Z}$ for some $\xi \in [0, 1)$, then there exists an interpolating function ψ_ξ of exponential decay such that $f(x) = \sum_{k\in\mathbb{Z}} f(\xi+k)\psi_\xi(x-k)$ [14]. In this case the restriction of f to $[M_0, M_1]$ is not determined exclusively by the $M_1 - M_0$ values $f(\xi + k)$ for $\xi + k \in [M_0, M_1]$. Moreover, if φ is continuous, then there always exists a $\xi \in [0, 1)$ such that the reconstruction $\{f(\xi + k)\} \to f$ is unstable. Janssen's results in [14] indicate that a small amount of oversampling is an essential hypothesis in guaranteeing the locality and the stability of its reconstruction.

3.2. A local reconstruction algorithm. In practice we perform the calculations with a truncated version of the matrices U and T. We now combine Lemmas 3 and 4 to a first version of an efficient numerical reconstruction algorithm.

ALGORITHM 1.

Input. We assume that finitely many sampling points $x_1, \dots, x_J \in [M_0, M_1]$ are given with associated sampling vector $y = (y_1, \dots, y_J) \in \mathbb{R}^J$. Assume that $J \geq M_1 - M_0 + 2S - 1$ and that the truncated matrix T defined below is invertible.

Step 0. First we define and compute the truncated matrices $\mathsf{U} = \mathsf{U}^{M_0,M_1}$ and $\mathsf{T} = \mathsf{T}^{M_0,M_1} = \mathsf{U}^*\mathsf{U}$, given by their entries

$$\begin{aligned} \mathsf{U}_{jk} &= \varphi(x_j - k)\,, \\ \mathsf{T}_{kl} &= \sum_{j=1}^{J} \overline{\varphi(x_j - k)}\varphi(x_j - l) \end{aligned} \tag{13}$$

for $j = 1, \dots, J$ and $k, l = M_0 - S + 1, \dots, M_1 + S - 1$.

Step 1. Compute $b = \mathsf{U}^* y$, i.e.,

$$b_k = \sum_{j=1}^{J} \overline{\varphi(x_j - k)}\, y_j \qquad \text{for } k = M_0 - S + 1, \dots, M_1 + S - 1\,. \tag{14}$$

Step 2. Solve the system of equations

$$c = \mathsf{T}^{-1} b\,. \tag{15}$$

Step 3. Compute the restriction of f to $[M_0, M_1]$ by

$$f(x) = \sum_{k=M_0-S+1}^{M_1+S-1} c_k \varphi(x - k) \qquad \text{for } x \in [M_0, M_1]\,. \tag{16}$$

Then f is the (unique) least squares approximation of the given data vector y in the sense that

$$\sum_{j=1}^{J} |y_j - f(x_j)|^2 = \min_{h \in V(\varphi)} \sum_{j=1}^{J} |y_j - h(x_j)|^2\,. \tag{17}$$

If y arises as the sampled vector of an $f \in V(\varphi)$, i.e., $y_j = f(x_j)$, then this algorithm provides the exact reconstruction of f.

Proof. The least squares property (17) is clear, since this is exactly the property of the solution of the system of normal equations $\mathsf{U}^*\mathsf{U}\, c = \mathsf{U}^* y$. See [10] for details. □

In the case of B-splines a sufficient condition on the sampling density can be extracted from the proofs of Theorems 2.1 and 2.2 of [2]. *Assume that* $x_{j+1} - x_j \leq \delta$ *and that*

$$\delta \leq \frac{M_1 - M_0}{M_1 - M_0 + 2S - 1} < 1\,. \tag{18}$$

Then T *is invertible.* Condition (18) guarantees that there are at least $M_1 - M_0 + 2S - 1$ samples in $[M_0, M_1]$. Then the Schoenberg–Whitney theorem [16, p. 167] implies that T is invertible. See [2] for the detailed arguments.

3.3. Data segmentation. A further optimization of the reconstruction procedure is possible by data segmentation. Instead of solving the large system of equations

$$\mathsf{T}^{M_0,M_1} = (\mathsf{U}^{M_0,M_1})^* y$$

with a band matrix of dimension $M_1 - M_0 + 2S - 1$, we will solve t systems of smaller size. For this purpose we partition the large interval $[M_0, M_1]$ into t smaller intervals $[m_r, m_{r+1}]$, $r = 0, \ldots, t-1$ with $M_0 = m_0$ and $M_1 = m_t$.

Now we apply Algorithm 1 to each interval separately. More precisely, given the data (x_j, y_j) where $x_j \in [m_r, m_{r+1}]$, we set up the matrices $\mathsf{U}^{m_r,m_{r+1}}$ and $\mathsf{T}^{m_r,m_{r+1}}$ and solve t equations

$$\mathsf{T}^{m_r,m_{r+1}} c^{(r)} = (\mathsf{U}^{m_r,m_{r+1}})^* y^{(r)}\,, \tag{19}$$

where the vector $y^{(r)}$ consists of those data y_j for which $x_j \in [m_r, m_{r+1}]$ and the coefficient vector $c^{(r)} = (c_{m_r-S+1}, \ldots, c_{m_{r+1}+S-1})$.

The segmentation technique has a number of practical advantages.

1. The dimension of vectors and matrices can be reduced drastically. Using data segmentation, we solve t small systems of size $(M_1 - M_0)/t + 2S - 1$ instead of the large system of size $M_1 - M_0 + 2S - 1$.

2. Parallel processing can be applied because nonadjacent intervals can be handled simultaneously.

3. The function can be reconstructed on specified subintervals at smaller cost. See Figure 6.

On the other hand, data segmentation also comes with some caveats:

1. The coefficients c_k with indices $k \in [m_r - S + 1, m_r + S - 1]$ are computed at least twice because of overlap. Heuristically it has proved best to take averages of the multiply computed coefficients.

2. For a successful execution of the segmentation method it is necessary that each of the small matrices $\mathsf{T}^{m_r,m_{r+1}}$ is invertible. Again by dimension counts we find that the number of data in the interval $[m_r, m_{r+1}]$ should exceed the number of variables, i.e.,

$$\#\Big(X \cap [m_r, m_{r+1}]\Big) \geq m_{r+1} - m_r + 2S - 1\,.$$

Obviously this condition imposes an upper bound for the possible number of segmentations.

3.4. Implementation issues. 1. In Algorithm 1 the most expensive step is the calculation of the matrix U because it requires the point evaluations of φ. However, if the sampling points x_j are given, then U and T can be computed in advance and stored. Thus Step 0 can be taken care of before solving the reconstruction problem.

We handle the pointwise evaluation of φ by "quantizing" the generator. This means that for $\delta > 0$ sufficiently small we create a vector ψ consisting of entries $\varphi(\frac{l}{N})$ for $l = -NS, \ldots, NS$ such that

$$\left|\varphi(x) - \varphi\left(\frac{l}{N}\right)\right| < \delta \qquad \text{for } \left|x - \frac{l}{N}\right| < \frac{1}{2N}\,.$$

Thus building the matrix $\mathsf{U}_{jk} = \varphi(x_j - k)$ amounts to selecting the appropriate entries of ψ. This approximation of U works remarkably well and fast in the numerical simulations.

2. For the solution of the banded system (15) a number of fast algorithms is available. Golub and van Loan [10, Chap. 4.3] offer several efficient algorithms for this task; other options for the inversion of a banded matrix are mentioned in [12]. Since T is assumed to be positive-definite, the band Cholesky algorithm seems to be a good choice that minimizes the operation count for Step 2. MATLAB provides the commands SPARSE and CHOL to deal with this task.

3. Usually f is reconstructed on a grid $G = \{\frac{l}{N} : l = M_0 N, \dots, M_1 N\}$. Then (16) amounts to a discrete convolution, and thus Step 3 can be performed quickly. Again, since φ has compact support, we can use the banded structure of the associated matrix to perform this step.

3.5. Operation count. We estimate the number of multiplications for Algorithm 1. Recall that J is the number of samples, and $D = M_1 - M_0 + 2S - 1$ is the dimension of the problem.

(a) According to (14) each of the D entries of the vector b requires $\#\{j : |x_j - k| < S\}$ multiplications. Consequently Step 1 requires

$$\begin{aligned}\sum_{k=M_0-S+1}^{M_1+S-1} \#\{j : |x_j - k| < S\} &= \sum_{k=M_0-S+1}^{M_1+S-1} \sum_{j=1}^{J} \chi_{(k-S,k+S)}(x_j) \\ &= \sum_{j=1}^{J} \sum_{k=M_0-S+1}^{M_1+S-1} \chi_{(k-S,k+S)}(x_j) \\ &\le \sum_{j=1}^{J} 2S = 2SJ\end{aligned}$$

operations, because a point x is in at most $2S$ translates of the open interval $(-S, S)$.

(b) Likewise, calculating an entry of T requires $\#\big(\{j : |x_j - k| < S\} \cap \{j : |x_j - l| < S\}\big)$ multiplications; see (13). As in (a) we estimate the number of operations to set up the matrix T by

$$\begin{aligned}&\sum_{k=M_0-S+1}^{M_1+S-1} \sum_{l=M_0-S+1}^{M_1+S-1} \#\Big(\{j : |x_j - k| < S\} \cap \{j : |x_j - l| < S\}\Big) \\ &= \sum_{k=M_0-S+1}^{M_1+S-1} \sum_{l=M_0-S+1}^{M_1+S-1} \sum_{j=1}^{J} \chi_{(k-S,k+S)}(x_j)\chi_{(l-S,l+S)}(x_j) \\ &= \sum_{j=1}^{J} \Big(\sum_{k=M_0-S+1}^{M_1+S-1} \chi_{(k-S,k+S)}(x_j) \Big) \Big(\sum_{l=M_0-S+1}^{M_1+S-1} \chi_{(l-S,l+S)}(x_j) \Big) \\ &\le J \cdot (2S)^2 .\end{aligned}$$

(c) For the solution of the banded system $\mathsf{T}\, c = b$ by means of the band Cholesky algorithm we need at most

$$D\big((2S)^2 + 16S + 1\big) = (M_1 - M_0 + 2S - 1)\big((2S)^2 + 16S + 1\big) \le J(4S^2 + 16S + 1)$$

operations (and no square roots); see [10, Chap. 4.3.6].

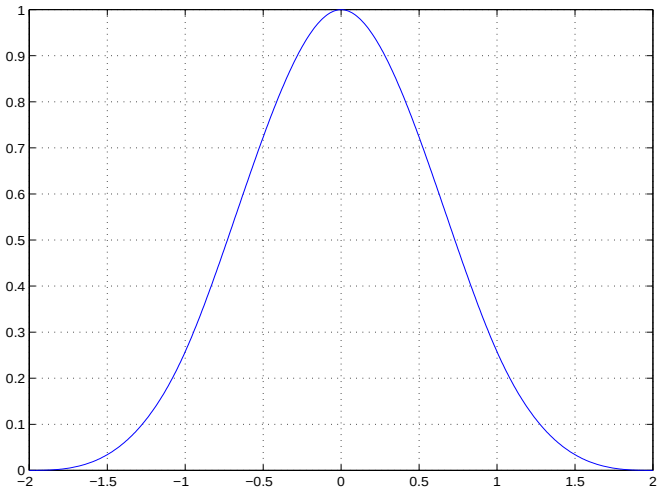


FIG. 1. *Generator φ: a B-spline of order* 3 *with* supp $\varphi \subseteq [-2, 2]$.

(d) To compute the reconstruction f on a grid $\{\frac{l}{N} : l = M_0N, \dots, M_1N\}$ we need to calculate $(M_1 - M_0)N$ point evaluations of f via (16). Since $\#\{k \in \mathbb{Z} : |x - k| < S\} \leq 2S$, each point evaluation requires at most $2S$ multiplications. Thus for the reconstruction on the grid we need at most

$$(M_1 - M_0)N \cdot 2S \leq J \cdot 2SN$$

multiplications.

Combining these estimates, we find that *Algorithm* 1 *requires*

$$\mathsf{O}\left(J(S^2 + SN)\right) \tag{20}$$

operations. In other words, the cost of the algorithm is linear in the number of data and quadratic in the size of the generator!

4. Numerical simulations. In our simulation we have used MATLAB. We used the shift-invariant spline spaces with the B-spline of order 3

$$\varphi = \underbrace{\chi_{[-1/2,1/2]} * \cdots * \chi_{[-1/2,1/2]}}_{4 \text{ times}}$$

as the generator of $V(\varphi)$. Thus supp $\varphi \subseteq [-2, 2]$ and $S = 2$. See Figure 1.

Figure 2 is a plot of the operation count as a function of the number of sampling points. We have reconstructed examples of size $114, 226, 444, 667, 887, 1085$ and used the MATLAB function FLOPS to count the number of operations.

The example in Figure 3 uses a signal on the interval $[0, 128]$. Since $S = 2$, we need at least $M_1 - M_0 + 2S - 1 = 131$ samples. The actual sampling set of Figure 3 consists of about 200 points and satisfies the maximum gap condition $\max_j(x_{j+1} - x_j) \approx 0.67 < 1$.

To make the example more realistic, we have added white noise to the sampled values of a given function $f \in V(\varphi)$. Instead of using the correct values $f(x_j)$ in the reconstruction algorithm, we use the noisy values $f_{err}(x_j) = f(x_j) + e_j$ so that

$$f_{err}|_X = f|_X + e\,.$$

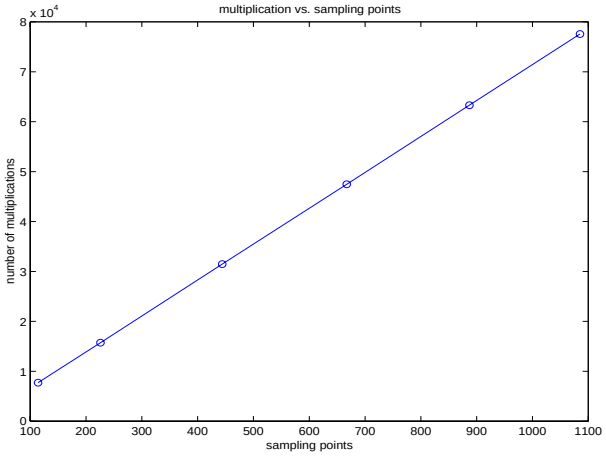


FIG. 2. *Number of multiplications for reconstruction problems of different size. The operation count is* linear *in the number of samples.*

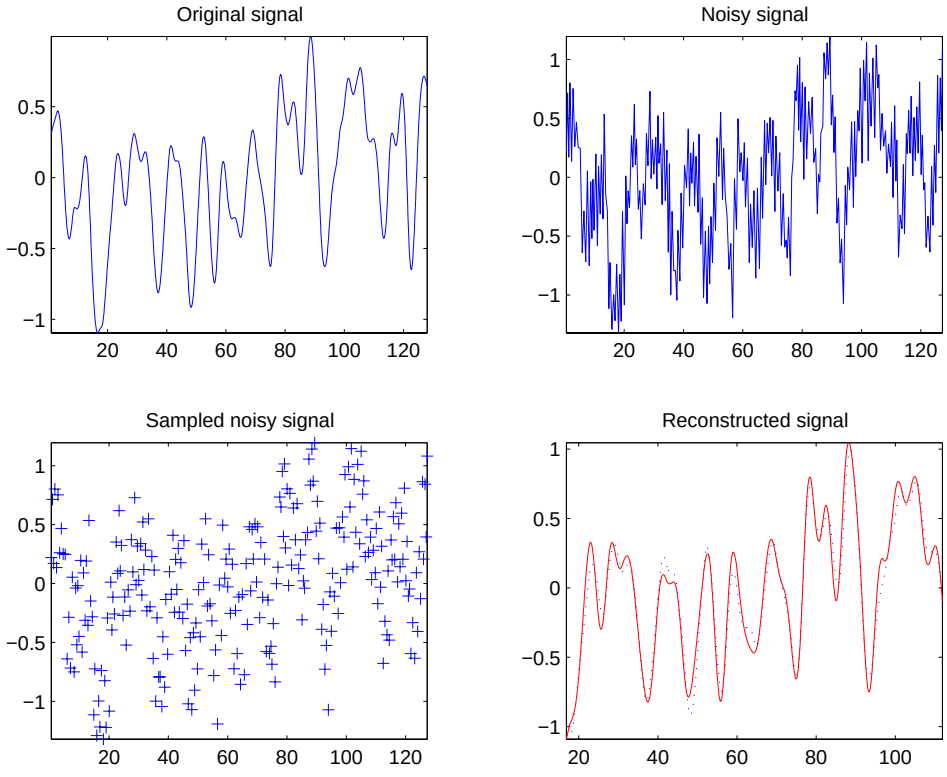


FIG. 3. *Reconstruction with noisy nonuniform samples: The top right plot shows the signal with additive* $err_{samp} = 63.8\%$ *noise. Bottom left shows the noisy signal sampled on a nonuniform grid with maximal gap* ≈ 0.67. *Bottom right shows the reconstructed function (continuous line) and original function (dotted line).*

The relative error between the original signal and the noisy signal is measured by

$$err_{samp} = \frac{\|f_{err}|_X - f|_X\|_2}{\|f|_X\|_2} = \left(\frac{\sum_{j=1}^J |e_j|^2}{\sum_{j=1}^J |f(x_j)|^2} \right)^{1/2}.$$

In our example $err_{samp} = 63.8\%$.

Figure 3 shows the plots of the original signal (top left), of the noisy signal (top right), and the plot of the noisy samples, which looks rather chaotic (bottom left). The last plot (bottom right) displays the reconstruction (continuous line) means of Algorithm 1. For comparison we have added the original function as a dotted line. The relative error err_{rec} of the reconstruction measured at the sampling points with respect to the correct samples is now

$$err_{rec} = \frac{\|f_{rec}|_X - f|_X\|_2}{\|f|_X\|_2} = 18.5\%\,.$$

The noise reduction is thus

$$err_{samp} = 63.8\% \to err_{rec} = 18.5\%\,.$$

In Figure 4 we investigate the dependence of the reconstruction on the generator φ. In each subplot the generator is a B-spline of order N, i.e., $\varphi_N = \chi_{[-1/2,1/2]} * \cdots * \chi_{[-1/2,1/2]}$ ($N+1$-fold convolution). The data set $(x_j, y_j)_{j=1,\ldots,J}$ is generated by sampling a function $f \in V(\varphi_5)$, and then we have added noise. The top left picture shows the original signal and the noisy sampled data. Then each subplot depicts the optimal approximation of these data in the spline space $V(\varphi_N)$, $N = 0, \ldots, 6$, starting with an approximation by a step function $f_{rec} \in V(\varphi_0)$ via an approximation by a piecewise linear function $f_{rec} \in V(\varphi_1)$ and ending with a smooth approximation $f_{rec} \in V(\varphi_6)$.

In each case we have also plotted the original function f (dotted line) for comparison. In addition, the relative error err_{samp} is indicated. The dependency of this error of N is a typical L-curve as it occurs in regularization procedures. In all our examples the best approximation is obtained in the correct space in which f was originally generated. This observation is consistent with the extended literature on smoothing splines, e.g., [4,12,17]. The main difference between those methods and the algorithm of section 3.2 is in the underlying function space. The reconstruction Algorithm 1 finds the best local reconstruction in the shift-invariant spline space $V(\varphi)$, whereas the smoothing spline of [4] is based on the nodes x_j and does not belong to a fixed function space.

Figure 5 displays the associated banded matrix T of the linear system (15). White squares correspond to zero entries, dark squares signify large entries of T, the shading being proportional to the size. The banded structure is clearly visible.

Figure 6 exhibits the power of the method of data segmentation. Instead of reconstructing the entire signal f, we have reconstructed only the restriction to two disjoint intervals. In the absence of noise, the reconstruction is exact. Since supp $\varphi \subset [-S, S]$, the calculation for the two intervals can be done locally and simultaneously. This property can be used for parallel processing.

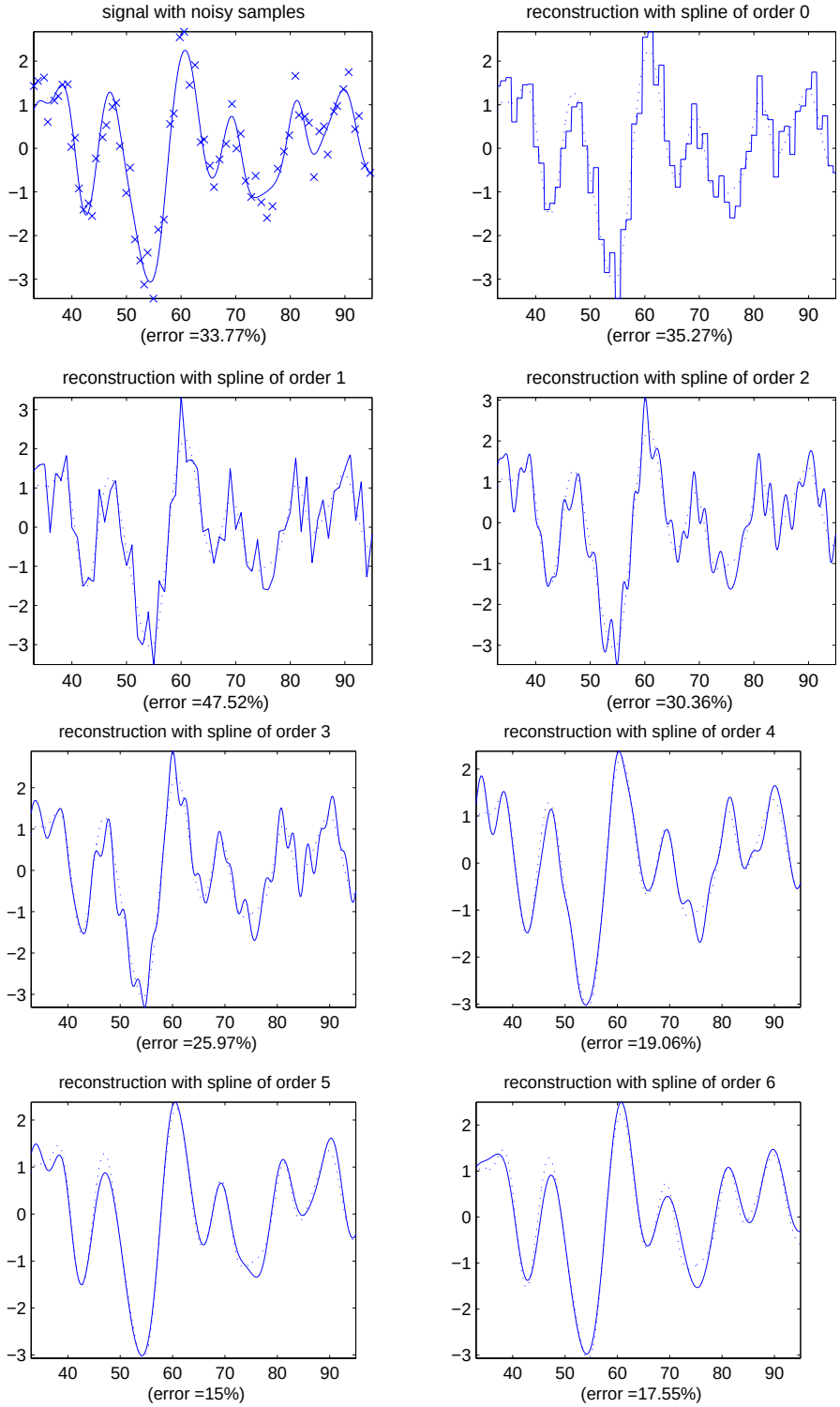


FIG. 4. *Reconstruction of a signal from noisy samples in shift-invariant spaces with B-splines of different orders as generator. Top left: Original signal (continuous line) and the noisy samples are marked (×). In the other plots the original signal is represented by the dotted line and the reconstruction is represented by the solid line.*

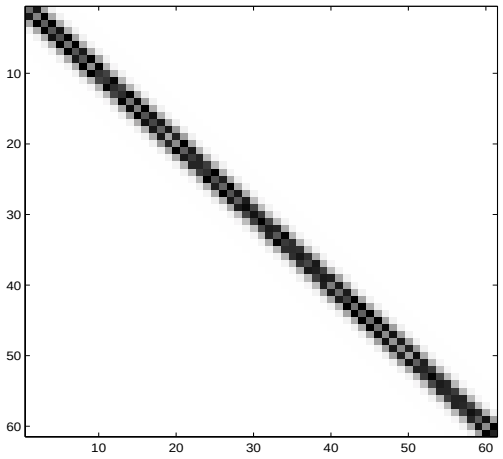


FIG. 5. *Structure of the banded matrix* T .

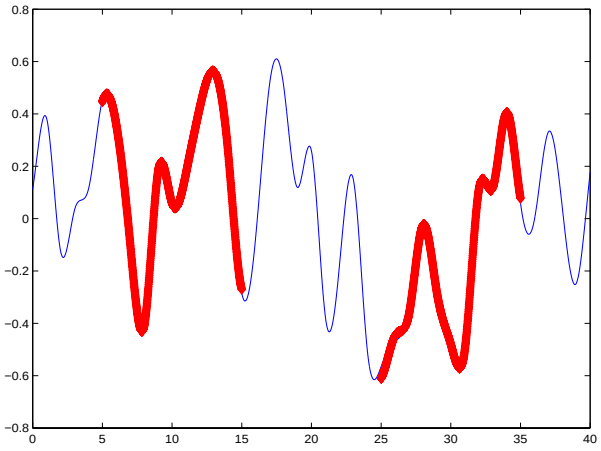


FIG. 6. *Reconstruction of the function on disjoint intervals (without noise).*

Appendix. Pseudocode.

```
function f_rec = reconstruction(xp, xs, x_rec, t);
% xp    ... sampling positions
% xs    ... sampling values
% x_rec ... positions, where the function should be reconstructed
% gen   ... generator with support supp(gen)=[-S,S]
% t     ... number of segmentation

step = round((max(x_rec)-min(x_rec))/t);
for k=min(x_rec):step:max(x_rec)
 xp_rel = {xp: k-S < xp < k+step+S} %relevant sampling positions
                                    %for the interval [k,k+step]
 xp_min = min(xp_rel);
 xp_max = max(xp_rel);
 J = length(xp_rel);                %number of sampling points
```

```
% calculation of the left side b:
 for j=1:J
   mi=ceil(xp(j))-S;
   ma=floor(xp(j))+S;
   for l = mi : ma
     b(l-mi+1) = b(l-mi+1) + xs(j)*gen(xp(j)-l);
   end
 end

% calculation of the matrix T:
 T=zeros(xp_max-xp_min+1+2*S,xp_max-xp_min+1+2*S);
 for j=1:J
   mi=ceil(xp(j))-S;
   ma=floor(xp(j))+S;
   for k=mi:ma
     for l=mi:ma
       T(k-mi+1,l-mi+1) = T(k-mi+1,l-mi+1) + gen(xp(j)-l)*gen(xp(j)-l);
     end
   end
 end

% calculation of the coefficients
  c_part = chol(T,b);       % solving the system T*c_part=b with a
                            %  banded Cholesky algorithm
  c(xp_min-S:xp_max+S) = c(xp_min-S:xp_max+S) + c_part;
  n(xp_min-S:xp_max+S) = n(xp_min-S:xp_max+S) + ones(xp_max-xp_min+1+2*S);
            %n ... normalization of coefficients because of overlapping
end

c = c ./ n;     %normalization of coefficients because of overlapping
% calculation of the reconstruction
for i = 1 : length(x_rec)
 for j = floor(x_rec(i)-S) : ceil(x_rec(i)+S)   % |x_rec-j| <= S
  f_rec(i) = f_rec(i) + gen(x_rec(i)-j) * c(j);
 end
end
```

Acknowledgment. We would like to thank Thomas Strohmer, University of California at Davis, for his valuable comments and several references.

REFERENCES

[1] A. Aldroubi and K. Gröchenig, *Nonuniform sampling and reconstruction in shift-invariant spaces*, SIAM Rev., 43 (2001), pp. 585–620.

[2] A. Aldroubi and K. Gröchenig, *Beurling-Landau-type theorems for non-uniform sampling in shift invariant spline spaces*, J. Fourier Anal. Appl., 6 (2000), pp. 93–103.

[3] M.D. Buhmann, *Radial basis functions*, in Acta Numerica, 2000, Cambridge University Press, Cambridge, UK, 2000, pp. 1–38.

[4] P. Craven and G. Wahba, *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math., 31 (1978/79), pp. 377–403.

[5] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.
[6] C. de Boor, R.A. DeVore, and A. Ron, *The structure of finitely generated shift-invariant spaces in $L_2(\mathbf{R}^d)$*, J. Funct. Anal., 119 (1994), pp. 37–78.
[7] S. Demko, W.F. Moss, and P.W. Smith, *Decay rates for inverses of band matrices*, Math. Comp., 43 (1984), pp. 491–499.
[8] R.J. Duffin and A.C. Schaeffer, *A class of nonharmonic Fourier series*, Trans. Amer. Math. Soc., 72 (1952), pp. 341–366.
[9] H.G. Feichtinger, K. Gröchenig, and T. Strohmer, *Efficient numerical methods in non-uniform sampling theory*, Numer. Math., 69 (1995), pp. 423–440.
[10] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[11] K. Gröchenig, *Acceleration of the frame algorithm*, IEEE Trans. Signal Process., 41 (1993), pp. 3331–3340.
[12] M.F. Hutchinson and F.R. de Hoog, *Smoothing noisy data with spline functions*, Numer. Math., 47 (1985), pp. 99–106.
[13] S. Jaffard, *Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 461–476.
[14] A.J.E.M. Janssen, *The Zak transform and sampling theorems for wavelet subspaces*, IEEE Trans. Signal Process., 41 (1993), pp. 3360–3365.
[15] R.-Q. Jia, *Stability of the shifts of a finite number of functions*, J. Approx. Theory, 95 (1998), pp. 194–202.
[16] L.L. Schumaker, *Spline Functions: Basic Theory*, Pure Appl. Math, Wiley-Interscience, New York, 1981.
[17] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, PA, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 914–931

SOLVING THE INDEFINITE LEAST SQUARES PROBLEM BY HYPERBOLIC QR FACTORIZATION*

ADAM BOJANCZYK†, NICHOLAS J. HIGHAM‡, AND HARIKRISHNA PATEL‡

Abstract. The indefinite least squares (ILS) problem involves minimizing a certain type of indefinite quadratic form. We develop perturbation theory for the problem and identify a condition number. We describe and analyze a method for solving the ILS problem based on hyperbolic QR factorization. This method has a lower operation count than one recently proposed by Chandrasekaran, Gu, and Sayed that employs both QR and Cholesky factorizations. We give a rounding error analysis of the new method and use the perturbation theory to show that under a reasonable assumption the method is forward stable. Our analysis is quite general and sheds some light on the stability properties of hyperbolic transformations. In our numerical experiments the new method is just as accurate as the method of Chandrasekaran, Gu, and Sayed.

Key words. indefinite least squares problem, downdating, hyperbolic rotation, hyperbolic QR factorization, rounding error analysis, forward stability, perturbation theory, condition number

AMS subject classifications. 65F20, 65G05

PII. S0895479802401497

1. Introduction.

The indefinite least squares problem (ILS) takes the form

$$\text{ILS}: \qquad \min_x (b-Ax)^T J(b-Ax), \tag{1.1}$$

where $A \in \mathbb{R}^{m\times n}$, $m \geq n$, and $b \in \mathbb{R}^m$ are given and J is the signature matrix

$$J = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}, \qquad p+q = m. \tag{1.2}$$

For $p = 0$ or $q = 0$ we have the standard least squares (LS) problem and the quadratic form is definite, while for $pq > 0$ the problem is to minimize a genuinely indefinite quadratic form. Chandrasekaran, Gu, and Sayed [3] discuss the application of the ILS problem to the solution of total least squares problems [18] and to the area of optimization known as H^∞ smoothing [8], [14].

*Received by the editors January 24, 2002; accepted for publication (in revised form) by L. Eldén June 25, 2002; published electronically February 12, 2003. This work was supported by Engineering and Physical Sciences Research Council Visiting Fellowship GR/R22414. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defence Advanced Research Project Agency (DARPA), Rome Laboratory, or the U.S. Government. This work was performed by an employee of the U.S. Government or under U.S. government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

http://www.siam.org/journals/simax/24-4/40149.html

†School of Electrical and Computer Engineering, Cornell University, 335 Theory Center and Engineering, Ithaca, NY 14853-3801 (adamb@ee.cornell.edu, http://www.ee.cornell.edu/~adamb/). This author was sponsored by the Defence Advanced Research Project Agency (DARPA) and Rome Laboratory, Air Force Material Command, USAF, under agreement F30602-97-1-0292.

‡Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/; hpatel@ma.man.ac.uk, http://www.ma.man.ac.uk/~hpatel/). The second author's work was supported by Engineering and Physical Sciences Research Council grant GR/R22612. The third author's work was supported by an Engineering and Physical Sciences Research Council Ph.D. Studentship.

The normal equations for (1.1), which are first order conditions for optimality, are

$$A^TJ(b - Ax) = 0. \tag{1.3}$$

Since the Hessian matrix of the quadratic to be minimized in (1.1) is $2A^TJA$, it follows that the ILS problem has a unique solution if and only if

$$A^TJA \text{ is positive definite.} \tag{1.4}$$

We will assume throughout this paper that (1.4) holds. Note that (1.4) implies $p \geq n$ and that $A(1\colon p, 1\colon n)$ (and hence A) has full rank. For a genuinely indefinite LS problem we therefore need $m > n$.

We note in passing that (1.3) gives $x = M^{-1}A^TJb$, where $M = A^TJA$, and the matrix $X = M^{-1}A^TJ$ is a pseudoinverse of A but not the Moore–Penrose pseudoinverse ($XA = I$, but AX is not symmetric).

One way of solving the ILS problem is to form the normal equations and solve them with the aid of a Cholesky factorization. Since this method has poor numerical stability properties for the standard LS problem it is clearly not a good choice for the ILS problem, except perhaps when A^TJA is well conditioned.

Chandrasekaran, Gu, and Sayed [3] propose a method for solving the ILS problem based on a QR factorization of A,

$$A = QR = \begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \overset{n}{} \\ \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \end{matrix} R, \qquad R \in \mathbb{R}^{n\times n}.$$

This factorization yields

$$A^TJA = R^T(Q_1^TQ_1 - Q_2^TQ_2)R,$$

which, in view of (1.4), implies that R is nonsingular and $Q_1^TQ_1 - Q_2^TQ_2$ is positive definite. Hence the normal equations (1.3) can be rewritten as

$$(Q_1^TQ_1 - Q_2^TQ_2)Rx = Q^TJb. \tag{1.5}$$

Using the Cholesky factorization

$$Q_1^TQ_1 - Q_2^TQ_2 = U^TU,$$

(1.5) becomes

$$U^TURx = Q^TJb.$$

This system can be solved for x by one forward and two backward substitutions. We will refer to this method as the "QR-Cholesky" method. It is shown in [3] that this method produces a computed solution $\widehat{x}$ that solves the problem

$$\min_x \big(b + \Delta b - (A + \Delta A)x)^TJ(b + \Delta b - (A + \Delta A)x\big),$$

where

$$\|\Delta A\|_F \leq c_{m,n}u\|A\|_F, \qquad \|\Delta b\|_2 \leq c_{m,n}u\|b\|_2,$$

with $c_{m,n}$ a constant depending on the problem dimensions and u the unit roundoff; in other words, the QR-Cholesky method is backward stable.

In this work we investigate the solution of the ILS problem via hyperbolic QR factorization. This approach has a lower operation count than the QR-Cholesky method but, in view of the use of hyperbolic transformations, its stability is questionable. We give rounding error analysis and perturbation analysis that combine to show that the method is forward stable under a reasonable assumption and hence of practical interest.

We begin, in the next section, with the perturbation analysis. The hyperbolic QR factorization method is described in section 3, its error analysis is given in section 4, and numerical experiments are presented in section 5. It is an important fact that obtaining useful error bounds for the application of a product of hyperbolic transformations to a vector is much more difficult than when the transformations are orthogonal. In section 4.1 we show how such products can be analyzed under a natural assumption on the form of the hyperbolic transformations.

2. Perturbation theory. In this section we derive normwise and componentwise perturbation bounds for the solution x and a residual r of the ILS problem. Our approach is based on that used by Cox and Higham [4] to obtain perturbation bounds for the equality constrained LS problem. We let $\widetilde{x}$ be the solution of the perturbed ILS problem

$$\min_x \bigl(b + \Delta b - (A + \Delta A)x\bigr)^T J \bigl(b + \Delta b - (A + \Delta A)x\bigr) \tag{2.1}$$

and define

$$\widetilde{r} = b + \Delta b - (A + \Delta A)\widetilde{x}, \qquad r = b - Ax$$

to be the residuals of the perturbed and unperturbed problems, respectively. We assume that $A + \Delta A$ satisfies the uniqueness condition (1.4), which will always be the case for ΔA sufficiently small in norm. The perturbations to the data will be measured by the smallest ϵ for which

$$\|\Delta A\|_F \le \epsilon \|\mathbf{A}\|_F, \quad \|\Delta b\|_2 \le \epsilon \|\mathbf{b}\|_2, \tag{2.2}$$

where $\mathbf{A}$ and $\mathbf{b}$ are a matrix and vector of tolerances.

The normal equations (1.3) can be rewritten as the augmented system (with $r = b - Ax$)

$$\begin{bmatrix} I & A \\ A^T J & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

It is convenient to define $s = Jr$ and rewrite the system with a symmetric coefficient matrix:

$$\begin{bmatrix} J & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} s \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{2.3}$$

The perturbed augmented system corresponding to (2.3) is

$$\begin{bmatrix} J & A + \Delta A \\ (A + \Delta A)^T & 0 \end{bmatrix} \begin{bmatrix} \widetilde{s} \\ \widetilde{x} \end{bmatrix} = \begin{bmatrix} b + \Delta b \\ 0 \end{bmatrix}. \tag{2.4}$$

Writing

$$\widetilde{s} = s + \Delta s, \quad \widetilde{x} = x + \Delta x$$

and subtracting (2.3) from (2.4), we obtain

$$\begin{bmatrix} J & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta s \\ \Delta x \end{bmatrix} = \begin{bmatrix} \Delta b - \Delta A\widetilde{x} \\ -\Delta A^T \widetilde{s} \end{bmatrix}. \tag{2.5}$$

It is straightforward to verify that the inverse of the matrix on the left-hand side of (2.5) is

$$\begin{bmatrix} J - JAM^{-1}A^TJ & JAM^{-1} \\ M^{-1}A^TJ & -M^{-1} \end{bmatrix}, \qquad \text{where } M = A^TJA.$$

Premultiplying by the inverse and expanding the right-hand side, we obtain

$$\Delta s = (J - JAM^{-1}A^TJ)(\Delta b - \Delta A\widetilde{x}) - (JAM^{-1})\Delta A^T\widetilde{s}, \tag{2.6a}$$

$$\Delta x = M^{-1}A^TJ(\Delta b - \Delta A\widetilde{x}) + M^{-1}\Delta A^T\widetilde{s}. \tag{2.6b}$$

If we put $J = I$, then we recover perturbation expressions for the standard LS problem.

Since the perturbations Δs and Δx are of order ϵ, we can substitute $s = Jr$ and x for their perturbed counterparts to obtain first order expressions. Then, taking norms, we deduce

$$\begin{aligned} \|\Delta r\|_2 &\le \epsilon\Big[\|I - JAM^{-1}A^T\|_2(\|\mathbf{b}\|_2 + \|\mathbf{A}\|_F\|x\|_2) + \|AM^{-1}\|_2\|\mathbf{A}\|_F\|r\|_2\Big] \\ &\quad + O(\epsilon^2), \\ \|\Delta x\|_2 &\le \epsilon\Big[\|M^{-1}A^T\|_2(\|\mathbf{b}\|_2 + \|\mathbf{A}\|_F\|x\|_2) + \|M^{-1}\|_2\|\mathbf{A}\|_F\|r\|_2\Big] + O(\epsilon^2). \end{aligned} \tag{2.7}$$

Hence, provided $x \neq 0$,

$$\begin{aligned} \frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon\Bigg[&\|M^{-1}A^T\|_2\|\mathbf{A}\|_F\left(\frac{\|\mathbf{b}\|_2}{\|\mathbf{A}\|_F\|x\|_2} + 1\right) \\ &+ \|M^{-1}\|_2\|A\|_F^2\frac{\|\mathbf{A}\|_F}{\|A\|_F}\frac{\|r\|_2}{\|A\|_F\|x\|_2}\Bigg] + O(\epsilon^2). \end{aligned} \tag{2.8}$$

This bound shows that the sensitivity of the ILS problem is bounded in terms of $\|M^{-1}A^T\|_2\|\mathbf{A}\|_F$ when the residual is zero or small and $\|M^{-1}\|_2\|A\|_F^2$ otherwise; note that for $\mathbf{A} = A$ the former quantity is no larger than the latter and is potentially much smaller.

Now we examine whether (2.8) is attainable for some ΔA and Δb. The three terms in brackets in (2.7) are

$$E_1 = \|M^{-1}A^T\|_2\|\mathbf{b}\|_2, \quad E_2 = \|M^{-1}A^T\|_2\|\mathbf{A}\|_F\|x\|_2, \quad E_3 = \|M^{-1}\|_2\|\mathbf{A}\|_F\|r\|_2,$$

and they result from the perturbations Δb, ΔA, and ΔA^T, respectively. It follows that the bound (2.7) can fail to be achieved for some Δb and ΔA only if $E_1 < E_2 \approx E_3$ and there is substantial cancellation in the expression $-M^{-1}A^TJ\Delta Ax + M^{-1}\Delta A^TJr$ for all ΔA. We can show in various special cases that these circumstances cannot

arise (for example, when r is small, or when $|r^T Jr| \approx \|r\|_2^2$), but we have been unable to establish attainability of the bound (2.8) in general.

A natural definition of the condition number of the ILS problem is

$$\kappa_{\mathrm{ILS}}(A,b) = \lim_{\epsilon\to 0}\sup\left\{\frac{\|x-\widetilde{x}\|_2}{\|x\|_2} : (2.2)\text{–}(2.4)\text{ hold}\right\}. \tag{2.9}$$

Without a guarantee of sharpness, the bound (2.8) does not provide an estimate of $\kappa_{\mathrm{ILS}}(A,b)$ to within a readily identifiable constant factor. Therefore we take a different approach in which we combine the two ΔA terms in (2.6b) before taking norms. To do this, we use the vec operator, which stacks the columns of a matrix into one long column vector, together with the Kronecker product $A\otimes B = (a_{ij}B)$, which for $A\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{p\times q}$ is the block matrix $(a_{ij}B)\in\mathbb{R}^{mp\times nq}$ (see [9], [11, Chap. 4]). Applying the vec operator to (2.6b) and using the relation $\mathrm{vec}(AXB) = (B^T\otimes A)\mathrm{vec}(X)$, we obtain

$$\Delta x = M^{-1}A^T J\Delta b - (x^T\otimes M^{-1}A^T J)\mathrm{vec}(\Delta A) + (r^T J\otimes M^{-1})\mathrm{vec}(\Delta A^T) + O(\epsilon^2).$$

Using the relation $\mathrm{vec}(\Delta A^T) = \Pi\mathrm{vec}(\Delta A)$, where Π is the vec-permutation matrix [9], gives

$$\Delta x = M^{-1}A^T J\Delta b - \left[(x^T\otimes M^{-1}A^T J) - (r^T J\otimes M^{-1})\Pi\right]\mathrm{vec}(\Delta A) + O(\epsilon^2).$$

Now we take 2-norms. Using (2.2) and the fact that $\|\mathrm{vec}(\Delta A)\|_2 = \|\Delta A\|_F$, we deduce that

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \psi\epsilon + O(\epsilon^2), \tag{2.10}$$

where

$$\psi = \left(\|M^{-1}A^T\|_2\|\mathbf{b}\|_2 + \|(x^T\otimes M^{-1}A^T J) - (r^T J\otimes M^{-1})\Pi\|_2\|\mathbf{A}\|_F\right)/\|x\|_2,$$

and we have

$$\kappa_{\mathrm{ILS}}(A,b) \le \psi \le 2\kappa_{\mathrm{ILS}}(A,b).$$

In extensive numerical comparisons between the first order terms of the bounds (2.8) and (2.10), including with direct search optimization, we have found these terms always to be within a small factor of each other. We believe that (2.8) is nearly attainable and, because this bound is much easier to work with than (2.10), we will use it when we investigate the stability of hyperbolic QR factorization for solving the ILS problem.

To end this section, we note that we can also use (2.6) to obtain componentwise perturbation bounds for the ILS problem. For the solution, we obtain

$$|\Delta x| \le \epsilon|M^{-1}A^T|(\mathbf{b} + \mathbf{A}|x|) + |M^{-1}|\mathbf{A}^T|r| + O(\epsilon^2),$$

where inequalities and the absolute value are interpreted componentwise and ϵ has been redefined as the smallest value for which $|\Delta A| \le \epsilon\mathbf{A}$, $|\Delta b| \le \epsilon\mathbf{b}$, where $\mathbf{A}$ and $\mathbf{b}$ are now assumed to have nonnegative entries.

3. Hyperbolic QR factorization method. We define a matrix $Q \in \mathbb{R}^{m\times m}$ to be J-orthogonal if

$$Q^T J Q = J,$$

or, equivalently, $QJQ^T = J$, where J is defined in (1.2). Suppose we can find a J-orthogonal matrix Q such that

$$Q^T A = Q^T \begin{array}{c} {\scriptstyle p} \\ {\scriptstyle q} \end{array} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{array}{c} {\scriptstyle n} \\ {\scriptstyle m-n} \end{array} \begin{array}{c} {\scriptstyle n} \\ \end{array} \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{3.1}$$

where $R \in \mathbb{R}^{n\times n}$ is upper triangular. We refer to this factorization as a *hyperbolic QR factorization*. Then

$$Q^T(b - Ax) = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} - \begin{bmatrix} R \\ 0 \end{bmatrix} x = \begin{bmatrix} d_1 - Rx \\ d_2 \end{bmatrix}, \qquad \begin{array}{c} {\scriptstyle n} \\ {\scriptstyle m-n} \end{array} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = Q^T b,$$

and so

$$\begin{aligned} (b - Ax)^T J (b - Ax) &= (b - Ax)^T Q J Q^T (b - Ax) \\ &= \begin{bmatrix} d_1 - Rx \\ d_2 \end{bmatrix}^T J \begin{bmatrix} d_1 - Rx \\ d_2 \end{bmatrix} \\ &= \|d_1 - Rx\|_2^2 + d_2^T J(n+1\colon m, n+1\colon m) d_2, \end{aligned} \tag{3.2}$$

recalling that (1.4) implies $p \geq n$ in (1.2). Hence the ILS solution is obtained by solving $Rx = d_1$. This method is an analogue of Golub's method for the LS problem [7].

The matrix Q can be constructed as a product of hyperbolic rotations and orthogonal matrices. A 2×2 hyperbolic rotation has the form

$$H = \begin{bmatrix} c & -s \\ -s & c \end{bmatrix}, \qquad c^2 - s^2 = 1,$$

and it is so named because $|c| = \cosh\theta$ and $s = \sinh\theta$ for some θ. It is easy to check that H is J-orthogonal for $J = \operatorname{diag}(1, -1)$. We will choose H to effect the zeroing operation

$$\begin{bmatrix} c & -s \\ -s & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix},$$

which requires that $cx_2 = sx_1$. The latter equation has a real solution only when $|x_1| > |x_2|$, in which case

$$c = \frac{x_1}{\sqrt{x_1^2 - x_2^2}}, \qquad s = \frac{x_2}{\sqrt{x_1^2 - x_2^2}}. \tag{3.3}$$

In practice a rescaling of these formulas is desirable to reduce the risk of overflow.

```
function [c, s] = Hrotate(x1, x2)
% Compute c and s defining hyperbolic rotation H such that
% Hx has zero second element.
if |x1| > |x2|
   t = x2/x1, c = 1/sqrt(1 - t^2), s = ct
else
   No real rotation exists—abort.
end
```

Unlike for orthogonal rotations, how hyperbolic rotations are applied to a vector is crucial to the stability of the computation [1], [15]. Consider the computation of $y = Hx$:

$$\begin{aligned} y_1 &= cx_1 - sx_2, \\ y_2 &= -sx_1 + cx_2. \end{aligned} \tag{3.4}$$

The first equation gives

$$x_1 = \frac{y_1}{c} + \frac{s}{c}x_2, \tag{3.5}$$

which allows the second to be rewritten as

$$\begin{aligned} y_2 &= -\frac{s}{c}y_1 + \left(-\frac{s^2}{c} + c\right)x_2 \\ &= -\frac{s}{c}y_1 + \frac{x_2}{c}. \end{aligned} \tag{3.6}$$

We will apply hyperbolic rotations using (3.4) and (3.6). As noted by Park and Eldén [13], this way of forming the product $y = Hx$ corresponds to use of the rescaled LU factorization

$$H = \begin{bmatrix} c & -s \\ -s & c \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -s/c & 1/c \end{bmatrix} \begin{bmatrix} c & -s \\ 0 & 1 \end{bmatrix}.$$

That this way of forming y is advantageous for stability was proved in [1] in the context of downdating a Cholesky factorization. We express the formation as follows:

function $B = \text{Happly}(c, s, B)$
% Apply hyperbolic rotation defined by c and s to $2 \times n$ matrix B.
for $j = 1\colon n$
    $B(1,j) = cB(1,j) - sB(2,j)$
    $B(2,j) = -(s/c)B(1,j) + B(2,j)/c$
end

For later use we note that (3.5) and (3.6) can be expressed together in the form

$$\begin{bmatrix} x_1 \\ y_2 \end{bmatrix} = G \begin{bmatrix} y_1 \\ x_2 \end{bmatrix}, \tag{3.7}$$

where

$$G = \begin{bmatrix} 1/c & s/c \\ -s/c & 1/c \end{bmatrix} \equiv \begin{bmatrix} \widetilde{c} & \widetilde{s} \\ -\widetilde{s} & \widetilde{c} \end{bmatrix}, \qquad \widetilde{c}^{\,2} + \widetilde{s}^{\,2} = 1.$$

The matrix G is a Givens rotation. Hence function Happly can be interpreted as forming the first row of the product HB by a hyperbolic rotation and the second row by a Givens rotation.

Our algorithm for computing the triangular factor R in (3.1) begins by computing the QR factorization

$$A_1 = Q_1 R_1, \qquad Q_1 \in \mathbb{R}^{p\times p}, \quad R_1 \in \mathbb{R}^{p\times n},$$

where Q_1 is orthogonal. Defining $\widetilde{Q} = \text{diag}(Q_1^T, I_q)$ we have

$$A^{(1)} = \widetilde{Q}A = \begin{bmatrix} R_1 \\ A_2 \end{bmatrix},$$

and $\widetilde{Q}$ is trivially J-orthogonal. We now zero A_2 with the aid of hyperbolic rotations. This can be done entirely with hyperbolic rotations or with a mix of hyperbolic and orthogonal rotations. Since hyperbolic rotations do not preserve the norms of vectors to which they are applied, we will use the minimum number, n, of them.

From a 2×2 hyperbolic rotation we build an $m \times m$ rotation in the (i, j) plane, $H_{i,j}$, defined to be the identity matrix modified according to $h_{ii} = h_{jj} = c$ and $h_{ij} = h_{ji} = -s$. Note that, provided the indices satisfy $i \leq p$ and $j > p$, H_{ij} is J-orthogonal. The parameters c and s are chosen to zero the jth element of the vector to which H_{ij} is applied.

Consider the first column of $A^{(1)}$. We first zero the elements in positions $(p+2, 1)$, $(p+3, 1), \ldots, (m, 1)$ using a Householder transformation, P_1, acting on rows $p+1\colon m$. Then we eliminate the $(p+1, 1)$ element, which is the sole remaining subdiagonal element in column 1, by a hyperbolic rotation $H_{1,p+1}$. It is clear that these operations do not disturb the existing zeros in positions $(2\colon p, 1)$ of $A^{(1)}$. At this point we have formed

$$A^{(2)} := H_{1,p+1} P_1 A^{(1)} =: Q^{(1)} A, \tag{3.8}$$

where $A^{(2)}(2\colon m, 1) = 0$. The matrix $Q^{(1)}$ is a product of J-orthogonal matrices and so is J-orthogonal. Elements below the diagonal in the remaining columns are eliminated in an analogous way, with the hyperbolic rotation used for the jth column being in the $(j, p+1)$ plane. The complete algorithm for solving the ILS problem is summarized as follows.

ALGORITHM 1. *This algorithm solves the ILS problem* (1.1) *using Householder QR factorization and hyperbolic rotations.*

```
Compute the Householder QR factorization A(1:p,:) = Q_1 R_1
    (Q_1 ∈ R^{p×p}, R_1 ∈ R^{p×n}), overwriting A(1:p,:) with R and
    b(1:n) with Q(1:n,:)^T b(1:n).
for j = 1:min(m-1, n)
    Construct a Householder transformation H_j such that
        H_j A(p+1:m, j) = σ_j e_1.
    A(p+1:m, j:n) = H_j A(p+1:m, j:n)
    b(p+1:m) = H_j b(p+1:m)
    % Eliminate sole remaining subdiagonal element in column j by a
    % hyperbolic rotation.
    [c, s] = Hrotate(A(j, j), A(p+1, j))
    A([j p+1], j:n) = Happly(c, s, A([j p+1], j:n))
    b([j p+1]) = Happly(c, s, b([j p+1]))
end
R = A(1:n,:)
Solve Rx = b(1:n) by substitution.
```

The operation count of Algorithm 1 is the same as that for solution of the standard LS problem by Householder QR factorization (essentially because the hyperbolic rotations contribute only to the lower order terms in the operation count). Table 3.1 compares the cost of Algorithm 1 with the cost of forming and solving the normal equations (1.3) and the cost of the QR-Cholesky method. Algorithm 1 requires fewer operations than the QR-Cholesky method by a factor 2.5–3.

It remains to show that the desired hyperbolic rotations exist. Suppose the algorithm has succeeded in eliminating the first $k-1$ columns of A_2, yielding $A_2^{(k)}$, and

TABLE 3.1
Operation counts for methods for solving the ILS problem.

	Normal equations	Hyperbolic QR	QR-Cholesky
	$n^2(m+n/3)$	$2n^2(m-n/3)$	$n^2(5m-n)$
$m \approx n$	$4n^3/3$	$4n^3/3$	$4n^3$
$m \gg n$	mn^2	$2mn^2$	$5mn^2$

define $C \in \mathbb{R}^{n\times q}$ and $R_1^{(k)} \in \mathbb{R}^{n\times n}$ by

$$R_1^{(k)^T} C \equiv A_1^{(k)^T} \begin{bmatrix} C \\ 0 \end{bmatrix} = A_2^{(k)^T}. \tag{3.9}$$

Since

$$\begin{bmatrix} A_1^{(k)} \\ A_2^{(k)} \end{bmatrix} = Q^{(k-1)} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

for a J-orthogonal matrix $Q^{(k-1)}$ and (1.4) holds, the matrix

$$A_1^{(k)^T} A_1^{(k)} - A_2^{(k)^T} A_2^{(k)} = R_1^{(k)^T} (I - CC^T) R_1^{(k)}$$

is positive definite. Since $R_1^{(k)}$ is upper triangular and nonsingular, it follows that $I - CC^T$ is positive definite and hence that $|c_{ij}| < 1$ for all i and j. Now $A_1^{(k)}$ is upper triangular and $A_2^{(k)}(1\colon q, 1\colon k-1) = 0$, so, using (3.9),

$$1 > |c_{ki}| = \left| \frac{a_{p+i,k}^{(k)}}{a_{k,k}^{(k)}} \right|, \qquad i = 1\colon q,$$

which ensures the existence of the hyperbolic rotation required on the $(k+1)$st stage.

4. Rounding error analysis. We now give a rounding error analysis of Algorithm 1. First, we note that from (3.1) we have $R^TR = A_1^TA_1 - A_2^TA_2$. Hence if A_1 is upper trapezoidal, then the hyperbolic QR factor R is the result of (block) downdating a Cholesky factorization. Various algorithms, both hyperbolic and nonhyperbolic, are known for downdating Cholesky factorizations, and error analysis is available; see, for example, [1], [2], [5], [6], [15], [17]. While we could invoke some of the earlier results in the part of the analysis that does not involve the right-hand side, b, we have chosen to give an independent development, aiming to make clear how the various errors combine and provide building blocks that should be of use in future analyses. In particular, we emphasize the high-level features of the analysis and thereby provide new insight into what is required of a sequence of hyperbolic transformations in order for satisfactory error bounds to be obtainable.

We use the standard model of floating point arithmetic [10, sect. 2.2]:

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta)^{\pm 1}, \qquad |\delta| \le u, \quad \operatorname{op} = +, -, *, /,$$

where u is the unit roundoff. Our bounds are expressed in terms of the constants

$$\gamma_k = \frac{ku}{1-ku}, \qquad \widetilde{\gamma}_k = \frac{cku}{1-cku}, \tag{4.1}$$

where c denotes a small integer constant whose exact value is unimportant. We also employ the relative error counter, $\langle k \rangle$:

$$\langle k \rangle = \prod_{i=1}^{k} (1+\delta_i)^{\rho_i}, \qquad \rho_i = \pm 1, \quad |\delta_i| \le u. \tag{4.2}$$

We use the fact that $|\langle k \rangle - 1| \le \gamma_k = ku/(1-ku)$ [10, Lem. 3.1].

Given an error bound for a single orthogonal transformation it is relatively easy to obtain a useful error bound for a product of several orthogonal transformations, as first shown by Wilkinson in the 1960s. The situation is quite different for a product of hyperbolic transformations, $y = H_p \dots H_2 H_1 x$, say. It is possible to mimic the analysis for orthogonal transformations and write, for example, $\widehat{y} = (H_p + \Delta H_p) \dots (H_2 + \Delta H_2)(H_1 + \Delta H_1)x$, with each ΔH_j bounded relative to H_j. However, this expression does not lead to a satisfactory forward or backward error bound, because the H_j are unbounded in norm. A better approach is to exploit the following equivalence between orthogonal and hyperbolic transformations.

Let

$$A = \begin{matrix} p \\ q \end{matrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{matrix} n \\ \end{matrix} = \begin{matrix} p \\ q \end{matrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{matrix} p & q \\ \end{matrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{matrix} p \\ q \end{matrix} \begin{matrix} n \\ \end{matrix} = QB, \tag{4.3}$$

where Q is J-orthogonal, for J in (1.2). Then $Q_{11}^T Q_{11} = I + Q_{21}^T Q_{21}$, and hence Q_{11} is nonsingular. It is not hard to show that

$$\begin{bmatrix} B_1 \\ A_2 \end{bmatrix} = \operatorname{exc}(Q) \begin{bmatrix} A_1 \\ B_2 \end{bmatrix}, \tag{4.4}$$

where the matrix

$$\operatorname{exc}(Q) = \begin{bmatrix} Q_{11}^{-1} & -Q_{11}^{-1} Q_{12} \\ Q_{21} Q_{11}^{-1} & Q_{22} - Q_{21} Q_{11}^{-1} Q_{12} \end{bmatrix}$$

is orthogonal. Moreover, if P is an orthogonal matrix partitioned in the same way as Q and its $(1,1)$ block is nonsingular, then $\operatorname{exc}(P)$ is J-orthogonal. In fact, the exchange operator is involutary: $\operatorname{exc}(\operatorname{exc}(P)) = P$. Note that (3.7) is a special case of (4.4). For proofs of these properties, see [12, Lem. 1], [17, sect. 2].

The advantage of (4.4) is that because the transformation matrix is orthogonal error terms can be moved around in the equation without changing their norm. The disadvantage is that it is hard to analyze more than one transformation. For example, let $C = PA$, where P is J-orthogonal. Then $C = PQB$ and corresponding to (4.4) we have

$$\begin{bmatrix} A_1 \\ C_2 \end{bmatrix} = \operatorname{exc}(PQ) \begin{bmatrix} C_1 \\ A_2 \end{bmatrix}. \tag{4.5}$$

Despite the elegance of this relation, $\operatorname{exc}(PQ)$ is a complicated function of P and Q. In practice the equations $A = QB$ and $C = PA$ must be modified to include rounding error terms, and these terms appear to preclude a suitably perturbed version of (4.5) with satisfactory bounds on the perturbations.

The gist of this analysis is that it is unclear how to obtain useful error bounds for the product of two or more *arbitrary* hyperbolic transformations. Fortunately, the transformations in Algorithm 1 are far from arbitrary, and in the next two sections we show that by exploiting their structure we can make useful progress.

4.1. Combining two hyperbolic transformations. We now analyze a product of two hyperbolic transformations that satisfy one key assumption: that the two transformations are "nonoverlapping" in components $1\!:\!p$. Nonoverlapping means that for $i = 1\!:\!p$ at least one of the two transformations agrees with the identity matrix in row i and column i. Without loss of generality, we consider a transformation $H_1(2,3)$ agreeing with the identity matrix in rows and columns $1\!:\!t$ and a transformation $H_2(1,3)$ agreeing with the identity matrix in rows and columns $t+1\!:\!p$, where $1 \le t < p$. Let

$$H_1(2,3)\begin{bmatrix} R \\ S \\ X \end{bmatrix}\begin{matrix} t \\ p-t \\ q \end{matrix} =: \begin{bmatrix} R \\ S_1 \\ X_1 \end{bmatrix}\begin{matrix} t \\ p-t \\ q \end{matrix}, \qquad H_2(1,3)\begin{bmatrix} R \\ S_1 \\ X_1 \end{bmatrix} =: \begin{bmatrix} R_1 \\ S_1 \\ X_2 \end{bmatrix},$$

or, overall,

$$H_2(1,3)H_1(2,3)\begin{bmatrix} R \\ S \\ X \end{bmatrix} = \begin{bmatrix} R_1 \\ S_1 \\ X_2 \end{bmatrix}.$$

We know from (4.3) and (4.4) that these two operations can be rewritten in terms of orthogonal transformations G_i as follows, where we now express the relations in terms of the affected components only:

$$G_1\begin{bmatrix} S_1 \\ X \end{bmatrix} = \begin{bmatrix} S \\ X_1 \end{bmatrix}, \tag{4.6a}$$

$$G_2\begin{bmatrix} R_1 \\ X_1 \end{bmatrix} = \begin{bmatrix} R \\ X_2 \end{bmatrix}. \tag{4.6b}$$

These two relations can be rewritten as

$$\begin{bmatrix} R_1 \\ S_1 \\ X \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix}\begin{bmatrix} R_1 \\ S \\ X_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix}\widetilde{G}_2^T\begin{bmatrix} R \\ S \\ X_2 \end{bmatrix} \equiv G\begin{bmatrix} R \\ S \\ X_2 \end{bmatrix}, \tag{4.7}$$

where $\widetilde{G}_2([1\!:\!t, p{+}1\!:\!m], [1\!:\!t, p{+}1\!:\!m]) = G_2$ and elsewhere $\widetilde{G}_2$ agrees with the identity matrix, and G is orthogonal. This relation shows that $\operatorname{exc}(H_2(1,3)H_1(2,3)) = G$ is of a relatively simple form given the no-overlap assumption.

Now we incorporate errors into the analysis. Consider the perturbed versions of (4.6),

$$G_1\begin{bmatrix} S_1 + E_1 \\ X + E_2 \end{bmatrix} = \begin{bmatrix} S \\ X_1 \end{bmatrix}, \tag{4.8a}$$

$$G_2\begin{bmatrix} R_1 + F_1 \\ X_1 + F_2 \end{bmatrix} = \begin{bmatrix} R \\ X_2 \end{bmatrix}, \tag{4.8b}$$

where

$$\max_{i=1,2}\|E_i\|_2 \le \mu\max(\|S_1\|_2, \|X\|_2), \quad \max_{i=1,2}\|F_i\|_2 \le \mu\max(\|R_1\|_2, \|X_1\|_2). \tag{4.9}$$

We will show below that perturbations of this form model rounding errors in Algorithm 1.

We now obtain an analogue of (4.7) for the perturbed quantities. We have

$$\begin{aligned}\begin{bmatrix} R_1 + F_1 \\ S_1 + E_1 \\ X + E_2 \end{bmatrix} &= \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix} \begin{bmatrix} R_1 + F_1 \\ S \\ X_1 \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & G_1^T \end{bmatrix} \left(\widetilde{G}_2^T \begin{bmatrix} R \\ S \\ X_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ F_2 \end{bmatrix} \right).\end{aligned}$$

This may be rewritten as

$$\begin{bmatrix} R_1 \\ S_1 \\ X \end{bmatrix} + \Delta = G \begin{bmatrix} R \\ S \\ X_2 \end{bmatrix}, \tag{4.10}$$

where, using $\|X_1\|_2 \le 2\max(\|S_1\|_2, \|X\|_2) + O(\mu)$,

$$\Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}, \qquad \max_i \|\Delta_i\|_2 \le 3\mu \max(\|R_1\|_2, \|X\|_2, \|S_1\|_2) + O(\mu^2).$$

The key fact is that the error bound for the two transformations combined is commensurate with that for the individual transformations. Because G is orthogonal the relation (4.10) can, if desired, be rewritten so that the 3×1 block matrix on the right is perturbed instead of the one on the left, as in the assumptions (4.8a) and (4.8b).

4.2. One rotation. Now we analyze the application of a hyperbolic rotation. We make the simplifying assumption that c and s in (3.3) are computed exactly.

The computed quantities from (3.4) and (3.6) satisfy

$$\widehat{y}_1\langle 1\rangle = cx_1\langle 1\rangle - sx_2\langle 1\rangle,$$

that is,

$$x_1 = \frac{\widehat{y}_1}{c}\langle 2\rangle + \frac{s}{c}x_2\langle 2\rangle,$$

and

$$\widehat{y}_2 = -\frac{s}{c}\widehat{y}_1\langle 3\rangle + \frac{x_2}{c}\langle 2\rangle.$$

Hence the analogue of (3.7) for the computed quantities is

$$\begin{bmatrix} x_1 \\ \widehat{y}_2 \end{bmatrix} = \begin{bmatrix} g_{11}\langle 2\rangle & g_{12}\langle 2\rangle \\ g_{21}\langle 3\rangle & g_{22}\langle 2\rangle \end{bmatrix} \begin{bmatrix} \widehat{y}_1 \\ x_2 \end{bmatrix} = (G + \Delta G) \begin{bmatrix} \widehat{y}_1 \\ x_2 \end{bmatrix}, \qquad |\Delta G| \le \gamma_3 |G|,$$

where G is orthogonal. This result can be rewritten as

$$\begin{bmatrix} x_1 \\ \widehat{y}_2 \end{bmatrix} = G \begin{bmatrix} \widehat{y}_1 + e_1 \\ x_2 + e_2 \end{bmatrix},$$

where

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = G^T \Delta G \begin{bmatrix} \widehat{y}_1 \\ x_2 \end{bmatrix},$$

so that

$$\max(|e_1|, |e_2|) \le \gamma_3(1 + 2|\widetilde{c}\,||\widetilde{s}|)\max(|\widehat{y}_1|, |x_2|) \le \gamma_6 \max(|\widehat{y}_1|, |x_2|).$$

This is a mixed backward–forward error result, since one element of each of the input and output vectors is perturbed. Importantly, this result is of the form (4.8).

4.3. Hyperbolic QR factorization. As before, we partition

$$A = \begin{matrix} p \\ q \end{matrix} \overset{n}{\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}}.$$

The first stage of Algorithm 1 computes the Householder QR factorization $A_1 = Q_1 \widetilde{R}_1$, where $\widetilde{R}_1 \in \mathbb{R}^{n\times n}$ is upper trapezoidal. We know from standard error analysis that the computed $\widetilde{R}_1$ is the exact factor of $A_1 + \Delta_1$, with $\|\Delta_1\|_F \le \widetilde{\gamma}_{pn}\|A_1\|_F$ [10, Thm. 19.4]. To simplify the notation, we will assume for the moment that A_1 is already in upper trapezoidal form and will introduce the error Δ_1 at the end.

Consider the jth column of A,

$$A(:,j) = \begin{matrix} p \\ q \end{matrix} \overset{n}{\begin{bmatrix} a_j^{(1)} \\ a_j^{(2)} \end{bmatrix}}.$$

It undergoes n Householder transformations in the last q components, intertwined with n hyperbolic rotations in the planes $(1, p+1), \dots, (n, p+1)$; the jth pair of these transformations introduce the required zeros in this column. The final $n-j$ pairs of transformations leave the column unchanged.

Consider a Householder transformation and the subsequent hyperbolic rotation. The Householder transformation agrees with the identity in rows and columns $1\colon p$ and its application is described by a standard backward stability result [10, Lem. 19.2]. It satisfies (4.8a) and (4.9) with $E_1 = 0$ and $\mu = \widetilde{\gamma}_q$. The hyperbolic rotation satisfies the bound of section 4.2 and is nonoverlapping with the Householder transformation. Therefore the analysis of section 4.1 can be applied to these two transformations. Importantly, all the subsequent pairs of Householder and hyperbolic rotations are mutually nonoverlapping and so the result of section 4.1 can be applied inductively.

The overall finding relating the jth columns of A and the final upper trapezoidal factor R_1 is that

$$\begin{matrix} p \\ q \end{matrix} \begin{bmatrix} \widehat{r}_j \\ a_j^{(2)} \end{bmatrix} + h_j = G \begin{bmatrix} a_j^{(1)} \\ 0 \end{bmatrix} \begin{matrix} p \\ q \end{matrix}, \qquad \|h_j\|_2 \le \widetilde{\gamma}_{qj} \max(\|\widehat{r}_j\|_2, \|a_j^{(2)}\|_2)$$

for some exactly orthogonal G that is *independent of* j. Importantly, $h_j(n+1\colon p) = 0$, because after the initial Householder QR factorization rows $n+1\colon p$ of A rest untouched. Putting these equations together for $j = 1\colon n$ and incorporating the error from the initial QR factorization of A_1 gives

$$\begin{bmatrix} \widehat{R}_1 + \Delta_3 \\ A_2 + \Delta_2 \end{bmatrix} = G \begin{bmatrix} A_1 + \Delta_1 \\ 0 \end{bmatrix}, \tag{4.11}$$

where $\Delta_3(n+1\colon p, :) = 0$ and

$$\|\Delta_1\|_F \le \widetilde{\gamma}_{pn}\|A_1\|_F, \tag{4.12a}$$

$$\|\Delta_i\|_F \le \widetilde{\gamma}_{qn} \max(\|\widehat{R}_1\|_F, \|A_2\|_F) \le \widetilde{\gamma}_{qn}\|A_1\|_F, \quad i = 2\colon 3. \tag{4.12b}$$

Certainly, $\max_i \|\Delta_i\|_F \le \widetilde{\gamma}_{mn}\|A\|_F$.

Note that in view of the equivalence (4.3) and (4.4), as long as G has a nonsingular $(1,1)$ block this result is equivalent to

$$\begin{bmatrix} A_1 + \Delta_1 \\ A_2 + \Delta_2 \end{bmatrix} = Q \begin{bmatrix} \widehat{R}_1 + \Delta_3 \\ 0 \end{bmatrix} \tag{4.13}$$

for a J-orthogonal Q. Both (4.11) and (4.13) are mixed backward–forward error results, because both the original data A and the trapezoidal factor R_1 are perturbed. We can obtain a genuine backward error result with the aid of the following lemma (for a proof, see [16, pp. 302–304]).

LEMMA 4.1. *Let $m = p + q$ and $n \geq p$. Given a full rank matrix $A \in \mathbb{R}^{p\times n}$ and $E \in \mathbb{R}^{q\times n}$ there exists an orthogonal $Q \in \mathbb{R}^{m\times m}$ such that*

$$Q \begin{bmatrix} A \\ E \end{bmatrix} = \begin{bmatrix} A + F \\ 0 \end{bmatrix}, \tag{4.14}$$

where, for small $\|E\|_2$,

$$\|F\|_2 \leq \frac{\|E\|_2^2}{2\sigma_{\min}(A)} + O(\|E\|_2^4). \qquad \square$$

Rewriting (4.11) as

$$\begin{bmatrix} \widehat{R}_1 \\ A_2 + \Delta_2 \end{bmatrix} = G \begin{bmatrix} A_1 + \Delta_1 + \widetilde{\Delta}_1 \\ \widetilde{\Delta}_2 \end{bmatrix}, \qquad \widetilde{\Delta} = -G^T \begin{bmatrix} \Delta_3 \\ 0 \end{bmatrix}$$

and applying Lemma 4.1 to the right-hand side leads to the conclusion that

$$\begin{bmatrix} \widehat{R}_1 \\ A_2 + \Delta_2 \end{bmatrix} = \widetilde{G} \begin{bmatrix} A_1 + \overline{\Delta}_1 \\ 0 \end{bmatrix},$$

where $\widetilde{G}$ is orthogonal and

$$\begin{aligned} \|\Delta_2\|_F &\leq \widetilde{\gamma}_{mn} \|A\|_F, \\ \|\overline{\Delta}_1\|_F &\leq \frac{\tilde{\gamma}_{mn}^2 \|A_1\|_F^2}{2\sigma_{\min}(A_1 + \Delta_1 + \widetilde{\Delta}_1)} + O(u^4) \\ &\leq \frac{\sqrt{n}}{2} \big(\kappa_2(A_1)\widetilde{\gamma}_{mn}\big)\widetilde{\gamma}_{mn} \|A\|_F + O(u^3). \end{aligned}$$

We conclude that backward stability of the factorization is guaranteed if $\kappa_2(A_1)u$ is of order 1. Thus the factorization is only conditionally backward stable, although the condition is quite weak. To relate the condition of A_1 to the sensitivity of the ILS problem, we note that

$$\kappa_2(A_1) \leq \big(\|M^{-1}\|_2 \|A\|_2^2\big)^{1/2},$$

where $M = A^T J A = A_1^T A_1 - A_2^T A_2$, from which it follows that if the perturbation bound (2.8) is small and the residual is not small, then A_1 must be well conditioned.

4.4. Solving the ILS problem. In solving the ILS problem we also transform the right-hand side $b = [b_1^T\ b_2^T]^T$ to $d = [d_1^T\ d_2^T]^T$. The above analysis gives

$$\begin{matrix} p \\ q \end{matrix} \begin{bmatrix} \widehat{d}_1 + \delta_3 \\ b_2 + \delta_2 \end{bmatrix} = G \begin{bmatrix} b_1 + \delta_1 \\ \widehat{d}_2 \end{bmatrix} \begin{matrix} p \\ q \end{matrix}, \tag{4.15}$$

where $\delta_3(n+1\colon p) = 0$ and

$$\|\delta_1\|_2 \le \widetilde{\gamma}_p \|b_1\|_2, \qquad \|\delta_i\|_2 \le \widetilde{\gamma}_q \max(\|\widehat{d}_1(1\colon n)\|_2, \|b_2\|_2), \quad i = 2\colon 3.$$

The ensuing analysis is simpler if $\widehat{d}_1$ is not perturbed, so we rewrite this relation as

$$\begin{bmatrix} \widehat{d}_1 \\ b_2 + \overline{\delta}_2 \end{bmatrix} = G \begin{bmatrix} b_1 + \overline{\delta}_1 \\ \widehat{d}_2 + \overline{\delta}_3 \end{bmatrix}, \tag{4.16}$$

where $\overline{\delta}_2 = \delta_2$ and

$$\max_{i=1\colon 3} \|\overline{\delta}_i\|_2 \le \widetilde{\gamma}_m \max(\|\widehat{d}_1(1\colon n)\|_2, \|b\|_2). \tag{4.17}$$

In Algorithm 1 the final step is to solve the triangular system $Rx = d_1$, where $R = R_1(1\colon n, :)$. The computed solution $\widehat{x}$ satisfies $(\widehat{R}+\Delta R)\widehat{x} = \widehat{d}_1(1\colon n)$, $|\Delta R| \le \gamma_n |\widehat{R}|$ [10, Thm. 8.5]; that is, the rounding errors in the substitution correspond to a further small perturbation of $\widehat{R}$.

We now consider the forward error of the computed solution $\widehat{x}$. First, let z_1 be the solution of the perturbed ILS problem with data

$$A + \Delta A := \begin{bmatrix} A_1 + \Delta_1 \\ A_2 + \Delta_2 \end{bmatrix}, \qquad b + \Delta b := \begin{bmatrix} b_1 + \overline{\delta}_1 \\ b_2 + \overline{\delta}_2 \end{bmatrix},$$

for which we know from (4.11) that the exact upper triangular R-factor is $\widehat{R} + \widetilde{\Delta}_3$, where $\widetilde{\Delta}_3 = \Delta_3(1\colon n, :)$. Then, in view of (4.16),

$$(\widehat{R} + \widetilde{\Delta}_3) z_1 = \widehat{d}_1(1\colon n).$$

Write

$$x - \widehat{x} = (x - z_1) + (z_1 - \widehat{x}).$$

Using the bounds on ΔA and Δb in (4.12) and (4.17), we have, from (2.8),

$$\begin{aligned} \frac{\|x - z_1\|_2}{\|x\|_2} \le \widetilde{\gamma}_{mn} \Bigg[& \|M^{-1}A^T\|_2 \|A\|_F \left(\frac{\max(1,\theta)\|b\|_2}{\|A\|_F \|x\|_2} + 1 \right) \\ & + \|M^{-1}\|_2 \|A\|_F^2 \frac{\|r\|_2}{\|A\|_F \|x\|_2} \Bigg] + O(u^2), \end{aligned} \tag{4.18}$$

where

$$\theta = \frac{\|d_1(1\colon n)\|_2}{\|b\|_2}. \tag{4.19}$$

The quantity θ measures the growth in the leading n components of the right-hand side as a result of the transformations that reduce A to triangular form. We now show

that even though θ can be large, it is innocuous. Suppose that $\theta \gg 1$. Note first that, since $\|d_1\|_2^2 + \|b_2\|_2^2 = \|b_1\|_2^2 + \|d_2\|_2^2$, we have $\|d_2\|_2 \approx \|d_1\|_2 \gg \|b_1\|_2$. Note also that $b_1(n+1\colon p)$ is not subjected to hyperbolic rotations and hence $\|d_1(n+1\colon p)\|_2 \le \|b\|_2$. Hence, from (3.2),

$$\|r\|_2^2 \ge |(b-Ax)^T J(b-Ax)| = |\,\|d(n+1\colon p)\|_2^2 - \|d_2\|_2^2\,| \approx \|d_2\|_2^2 \approx \|d_1\|_2^2.$$

Therefore $\theta \lesssim \|r\|_2/\|b\|_2$ and it follows that the first term in (4.18) is no larger than the second, showing that a large θ does not worsen the bound. Therefore (4.18) is essentially the same as (2.8) with $\epsilon = \widetilde{\gamma}_{mn}$.

From standard perturbation theory for square linear systems, the term $\|z_1 - \widehat{x}\|_2/\|x\|_2$ is bounded by

$$\begin{aligned}\phi &= \kappa_2(R)\left(\gamma_n + \widetilde{\gamma}_{qn}\frac{\max(\|R\|_F, \|A_2\|_F)}{\|R\|_2}\right)\\ &= \|R^{-1}\|_2\big(\gamma_n\|R\|_2 + \widetilde{\gamma}_{qn}\max(\|R\|_F, \|A_2\|_F)\big)\\ &\le \widetilde{\gamma}_{qn}\|R^{-1}\|_2\|A\|_F. \qquad (4.20)\end{aligned}$$

Now from the exact arithmetic analogue of (4.11) we have

$$\begin{bmatrix} R \\ A_2 \end{bmatrix} = G \begin{bmatrix} A_1 \\ 0 \end{bmatrix},$$

where G is orthogonal. Postmultiplying by $R^{-1}R^{-T}$ and transposing gives

$$[\,R^{-1} \quad R^{-1}R^{-T}A_2^T\,] = [\,R^{-1}R^{-T}A_1^T \quad 0\,]\,G^T.$$

Recalling that $M = A^TJA = R^TR$, it follows that

$$\begin{aligned}\|R^{-1}\|_2 &\le \|\,[\,R^{-1}R^{-T}A_1^T \quad 0\,]\,\|_2\\ &\le \|R^{-1}R^{-T}\,[\,A_1^T \quad A_2^T\,]\,\|_2\\ &= \|M^{-1}A^T\|_2.\end{aligned}$$

Hence

$$\phi \le \widetilde{\gamma}_{qn}\|M^{-1}A^T\|_2\|A\|_F,$$

which is smaller than the first term in (4.18). Our overall conclusion is that $\|x - \widehat{x}\|_2/\|x\|_2$ has an upper bound no larger than (2.8) with $\epsilon = \widetilde{\gamma}_{mn}$.

Recall that a method for solving the ILS problem is forward stable if it produces a computed solution with forward error similar to that for a backward stable method. If we make the reasonable assumption that the perturbation bound (2.8) is approximately attainable, then our rounding error analysis has shown that the hyperbolic QR factorization method for solving the ILS problem is forward stable.

It is unclear whether the hyperbolic QR factorization method is mixed backward–forward stable, or even backward stable. It is an open problem to determine a computable formula for the backward error of an arbitrary approximate solution to the ILS problem, and without such a formula it is difficult to test numerically for backward instability.

TABLE 5.1

Errors $\|x - \widehat{x}\|_2/\|x\|_2$ *for the three methods. In every case* $\theta \le 0.78$, $\|r\|_2/(\|A\|_2\|x\|_2) \approx u$, $\|Q\|_2 = 1.73$, $\|Q^T A - [R^T\ 0]^T\|_2/\|A\|_2 \approx u$, *and* $\|Q^T JQ - J\|_2 \le 10u$.

κ	Hyperbolic QR	QR-Cholesky	Normal equations	ψu
10^2	4.9e-15	4.0e-15	2.0e-13	2.1e-14
10^6	3.0e-11	1.7e-11	2.6e-5	1.9e-10
10^{10}	9.7e-8	1.5e-7	2.4e0	1.3e-6
10^{12}	4.8e-4	9.8e-3	6.4e0	1.4e-2

TABLE 5.2

Errors $\|x - \widehat{x}\|_2/\|x\|_2$ *for the three methods. In every case* $\|r\|_2/(\|A\|_2\|x\|_2) \approx 10^{-1}$ *and* $\|Q^T JQ - J\|_2 \approx$ *the error for hyperbolic QR.*

μ	Hyperbolic QR	QR-Cholesky	Normal equations	ψu	θ	$\|Q\|_2$	$\frac{\|Q^T A - [R^T\ 0]^T\|_2}{\|A\|_2}$
10	4.4e-13	1.9e-13	1.6e-12	2.0e-12	3.1e1	4.3e1	1.7e-14
10^2	1.6e-12	3.1e-12	1.5e-12	1.1e-11	1.1e2	1.5e2	2.8e-13
10^3	1.0e-10	5.4e-11	9.7e-11	7.8e-10	8.8e2	1.2e3	5.6e-12
10^4	2.2e-5	6.7e-5	5.0e-5	3.0e-4	5.6e5	8.0e5	3.2e-9
10^5	1.3e-1	5.4e-2	3.3e-2	3.1e-1	1.8e7	2.6e7	2.7e-7

5. Numerical experiments. We have carried out MATLAB experiments to compare the forward errors $\|x - \widehat{x}\|_2/\|x\|_2$ from Algorithm 1, the normal equations method (which forms and solves (1.3)), and the QR-Cholesky method. We approximated the exact solution by forming and solving the normal equations in 100-digit arithmetic using MATLAB's Symbolic Math Toolbox. We report results with $m = 16$, $n = 8$, and $p = 10$.

We formed the first class of test problems as

$$A = \begin{matrix} p \\ q \end{matrix} \begin{bmatrix} Q_1 DU \\ \frac{1}{2} Q_2 DU \end{bmatrix}, \tag{5.1}$$

where U, Q_1, and Q_2 are random orthogonal matrices and D is diagonal with diagonal elements distributed exponentially from κ^{-1} to 1. We have $A^T JA = (3/4)U^T D^2 U$, so A satisfies (1.4). The solution x is chosen from the random $N(0,1)$ distribution and $b := Ax$. Table 5.1 shows some results. In the table ψu is the first order term in (2.10) with $\epsilon = u$, $\mathbf{A} = A$, and $\mathbf{b} = b$; thus ψu is a first order bound for the forward error for a backward stable method. Recall that θ is defined in (4.19). For the statistics shown in the caption we explicitly formed Q by accumulating all the orthogonal and hyperbolic transformations.

In the second set of tests we generated A as in (5.1) and then premultiplied it by a random J-orthogonal matrix that is the product of 5 random hyperbolic rotations of norm approximately μ; this gives a Q factor of norm depending on μ in the hyperbolic QR factorization. Then we defined b as the right singular vector corresponding to the largest singular value of Q^T, which tends to make θ in (4.19) large. Results are shown in Table 5.2.

In all the tests the relative difference between ψu and the first order term from (2.8) was at most 0.1.

Three main conclusions can be drawn from the results shown. First, as expected, the normal equations method is not forward stable. Second, Algorithm 1 behaves in

a forward stable way in these tests and is just as accurate as the backward stable QR-Cholesky method, even when θ and $\|Q\|_2$ are large. The latter behavior adequately summarizes more extensive experiments that we have carried out. Third, the last column of Table 5.2 is consistent with the fact that we have proved our algorithm for computing the hyperbolic QR factorization to be mixed backward–forward stable and only conditionally backward stable.

Acknowledgments. Tony Cox contributed to the analysis in section 2. The first author would like to thank the Department of Mathematics, The University of Manchester, UK, where this work was carried out. He would also like to thank his hosts Professor Nick Higham and Dr. Françoise Tisseur for their great hospitality.

REFERENCES

[1] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[2] A. W. Bojanczyk and A. O. Steinhardt, *Stability analysis of a Householder-based algorithm for downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1255–1265.

[3] S. Chandrasekaran, M. Gu, and A. H. Sayed, *A stable and efficient algorithm for the indefinite linear least-squares problem*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 354–362.

[4] A. J. Cox and N. J. Higham, *Accuracy and stability of the null space method for solving the equality constrained least squares problem*, BIT, 39 (1999), pp. 34–50.

[5] L. Eldén and H. Park, *Perturbation analysis for block downdating of a Cholesky decomposition*, Numer. Math., 68 (1994), pp. 457–467.

[6] L. Eldén and H. Park, *Perturbation and error analyses for block downdating of a Cholesky decomposition*, BIT, 36 (1996), pp. 247–263.

[7] G. H. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.

[8] B. Hassibi, A. H. Sayed, and T. Kailath, *Linear estimation in Krein spaces—Part* I: *Theory*, IEEE Trans. Automat. Control, 41 (1996), pp. 18–33.

[9] H. V. Henderson and S. R. Searle, *The vec-permutation matrix, the vec operator and Kronecker products: A review*, Linear Multilinear Algebra, 9 (1981), pp. 271–288.

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

[11] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[12] C.-T. Pan and R. J. Plemmons, *Least squares modifications with inverse factorizations: Parallel implications*, J. Comput. Appl. Math., 27 (1989), pp. 109–127.

[13] H. Park and L. Eldén, *Stability analysis and fast algorithms for triangularization of Toeplitz matrices*, Numer. Math., 76 (1997), pp. 383–402.

[14] A. H. Sayed, B. Hassibi, and T. Kailath, *Inertia properties of indefinite quadratic forms*, IEEE Signal Process. Lett., 3 (1996), pp. 57–59.

[15] G. W. Stewart, *On the stability of sequential updates and downdates*, IEEE Trans. Signal Process., 43 (1995), pp. 2642–2648.

[16] G. W. Stewart, *Matrix Algorithms. Volume* I: *Basic Decompositions*, SIAM, Philadelphia, PA, 1998.

[17] M. Stewart and G. W. Stewart, *On hyperbolic triangularization: Stability and pivoting*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 847–860.

[18] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 932–942

APPROXIMATE PROJECTORS IN SINGULAR SPECTRUM ANALYSIS*

V. MOSKVINA† AND K. M. SCHMIDT†

Abstract. Singular spectrum analysis (SSA) is a method of time-series analysis based on the singular value decomposition of an associated Hankel matrix. We present an approach to SSA using an effective and numerically stable high-degree polynomial approximation of a spectral projector, which also provides a means of time-series forecasting. Several numerical examples illustrating the algorithm are given.

Key words. Hankel matrix, singular value decomposition, spectral projector, polynomial approximation

AMS subject classifications. 15A60, 41A10, 65F15, 62-07

PII. S0895479801398967

1. Introduction.

Singular spectrum analysis (SSA) is a well-established method of time-series analysis (see [2], [3], [15], [16], and the recent monographs [4] and [5]). The main idea of SSA is to select a number of significant principal components from the singular value decomposition (SVD) of the so-called trajectory matrix of a given time series, and hence to reconstruct a time series showing characteristic traits, e.g., the trend, periodicities, or signal (as opposed to random noise) of the original series.

SVD is a fundamental and very well studied process of numerical linear algebra with a long history (cf. [12], [1]). However, it is computationally expensive and thus problematic in real-time signal processing; therefore truncated forms of the SVD which only provide partial information (see [14], [17]) and alternative methods (see [7], [11], [13], and, specifically for Hankel-type matrices, [9]) have been proposed. Recently, combinations of SSA/SVD with a wavelet transform have attracted some attention [6], [18], [19].

The procedure of classical SSA is as follows. Let $x_1, x_2, \dots, x_N \in \mathbb{R}$, $N \in \mathbb{N}$, be (part of) a time series, let $M \le N/2$ be a positive integer, and let $K = N-M+1$. Set

$$\mathbf{X} = (x_{ij})_{i,j=1}^{M,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & x_{M+2} & \dots & x_N \end{pmatrix}. \tag{1.1}$$

$\mathbf{X}$ is called the *trajectory matrix*. Obviously $x_{ij} = x_{i+j-1}$, so that the matrix $\mathbf{X}$ has identical entries on the diagonals $i + j = \text{const}$, i.e., it is a Hankel matrix. One can consider $\mathbf{X}$ as multivariate data with M characteristics and $K = N - M + 1$ observations $X_1, X_2, \dots, X_K$, where

$$X_j = \begin{pmatrix} x_j \\ \vdots \\ x_{j+M-1} \end{pmatrix} \in \mathbb{R}^M \qquad (j \in \{1, \dots, K\}).$$

*Received by the editors November 29, 2001; accepted for publication (in revised form) by A. H. Sayed August 23, 2002; published electronically February 12, 2003.
http://www.siam.org/journals/simax/24-4/39896.html
†School of Mathematics, Cardiff University, Senghennydd Rd., Cardiff, UK, CF24 4YH (MoskvinaV1@cardiff.ac.uk, SchmidtKM@cardiff.ac.uk).

The SVD of $\mathbf{X}$ is based on the spectral decomposition of the *lag-covariance matrix* $\mathbf{R} = \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{M\times M}$. Note that $\mathbf{R}$ is symmetric and positive semidefinite. Therefore, it has a complete set of eigenvectors and can be diagonalized in the form

$$\mathbf{R} = U\Lambda U^T, \tag{1.2}$$

where Λ is the diagonal $M \times M$ matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0$, and

$$U = (U_1, U_2, \ldots, U_M) = \begin{pmatrix} u_{11} & u_{21} & \ldots & u_{M1} \\ u_{12} & u_{22} & \ldots & u_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1M} & u_{2M} & \ldots & u_{MM} \end{pmatrix}$$

is an orthogonal matrix of eigenvectors of the matrix $\mathbf{R}$. Denoting $d = \max\{i \in \{1,\ldots,M\} \mid \lambda_i > 0\}$ and $V_i = X^T U_i/\sqrt{\lambda_i}$ $(i = 1,\ldots,d)$, we can write the SVD of the trajectory matrix $\mathbf{X}$,

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T = \mathbf{X}_1 + \cdots + \mathbf{X}_d, \tag{1.3}$$

where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ are rank-one biorthogonal matrices; we have $\operatorname{rank} \mathbf{X} = d$.

Now a subset of the *SVD components* $\mathbf{X}_1, \ldots, \mathbf{X}_d$ is selected by choosing a set of indices $I \subset \{1,\ldots,d\}$, resulting in the decomposition

$$\mathbf{X} = \mathbf{X}_I + \mathbf{X}_{\bar{I}}, \quad \text{where } \mathbf{X}_I = \sum_{i\in I} \mathbf{X}_i \quad \text{and} \quad \mathbf{X}_{\bar{I}} = \sum_{i\notin I} \mathbf{X}_i.$$

If I is suitably chosen, $\mathbf{X}_I$ will represent some characteristic feature of the original time series which can be exhibited by removing $\mathbf{X}_{\bar{I}}$. Unfortunately, however, $\mathbf{X}_I$ itself is not in general the trajectory matrix of some time series, as it does not necessarily have Hankel structure. This obstacle is overcome by *diagonal averaging* over the diagonals $i + j = \text{const}$, which allows us to extract a time series $\tilde{x}_k$ $(k \in \{1,\ldots,N\})$ from any $M \times K$ matrix Y by the formula

$$\tilde{x}_k = \begin{cases} \dfrac{1}{k} \displaystyle\sum_{i=1}^{k} y_{i,\,k-i+1} & \text{for } 1 \leq k \leq M-1, \\ \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} y_{i,\,k-i+1} & \text{for } M \leq k \leq K, \\ \dfrac{1}{N-k+1} \displaystyle\sum_{i=k-K+1}^{M} y_{i,\,k-i+1} & \text{for } K+1 \leq k \leq N. \end{cases} \tag{1.4}$$

Applying this to $\mathbf{X}_I$ to construct a time series (z_t), we obtain the SSA decomposition of the original series

$$x_t = z_t + \varepsilon_t\,, \qquad t \in \{1,\ldots,N\}. \tag{1.5}$$

(It is not difficult to verify that the residual series (ε_t) results from diagonal averaging of $\mathbf{X}_{\bar{I}}$.)

An interesting practical application of SSA is the extraction of a signal from a time series perturbed by noise. Since one expects, in light of such asymptotic results as Corollary 6.1 in [5], that the signal will correspond to larger eigenvalues of the lag-covariance matrix, while eigenvalues associated with noise components should be small, this means that the first SVD components will be selected, cutting off at a certain eigenvalue size $\lambda_{\text{cut}} > 0$. Thus the index set will have the structure $I = \{1, \dots, l\}$ with some $l \in \{1, \dots, d\}$ such that $\lambda_l \geq \lambda_{\text{cut}} > \lambda_{l+1}$. Ideally, the series (z_t) in (1.5) can then be associated with signal and the residual series (ε_t) with noise. For an extensive discussion of the problem of choosing the values for the two SSA parameters, viz. the lag M and the number l of SVD components included in the reconstruction, see [5]. A common choice for M is the maximal value $M = \lfloor N/2 \rfloor$. The value of l (or, equivalently, of the cut-off point λ_{cut}) must depend on the properties of the given time series. If l is too small (underfitting), then we miss part of the signal; alternatively, if l is too large (overfitting), then we approximate a part of noise together with the signal.

In the present paper we develop a method of computing $\mathbf{X}_I$ (and hence the reconstructed series (x_t)) in this situation without actually performing the spectral decomposition of the lag-covariance matrix, i.e., without calculating its eigenvalues and eigenvectors. For large time series and correspondingly large matrices, this will offer a faster alternative and open the way for noise-reduction applications of the SSA method.

This paper is organized as follows. We first observe that the selection of the part $\mathbf{X}_I$ from the SVD of the trajectory matrix can be replaced by applying a spectral projector of the lag-covariance matrix; in the case at hand, this will be the orthogonal projector onto the eigenspace for eigenvalues in the interval $[\lambda_{\text{cut}}, \infty)$. In section 3, we then proceed to find a polynomial approximation of the characteristic function of this interval, which permits a direct approximate calculation of the spectral projector. We use an iterative method which avoids the problems inherent in a naive evaluation of the approximating polynomial, which would be inefficient and highly unstable. Section 4 presents a geometric forecasting algorithm based on the approximate spectral projector. The examples studied in section 5 demonstrate the workings and the practical applicability of our method. It turns out that even a relatively rough and inexpensive approximation, corresponding to an SSA with a "fuzzy cut-off," can yield a very high degree of noise suppression and an excellent reconstruction of the signal.

2. Polynomial approximation of the spectral projector. For the construction of the matrix $\mathbf{X}_I$ after choosing the index set I, it is sufficient to find the orthogonal projector P onto the subspace of $\mathbb{R}^M$ spanned by the eigenvectors U_j with $j \in I$,

$$P = \sum_{j \in I} U_j U_j^T. \tag{2.1}$$

Indeed, one then has, using (1.3) and the orthonormality of the eigenvectors,

$$P\mathbf{X} = \sum_{j \in I} \sum_{i=1}^{d} \sqrt{\lambda_i} U_j U_j^T U_i V_i^T = \sum_{j \in I} \sqrt{\lambda_j} U_j V_j^T = \mathbf{X}_I.$$

Thus, $\mathbf{X}_I$ is obtained by a simple matrix multiplication once P is known. The matrix P, on the other hand, can be represented as a function of the matrix $\mathbf{R}$. Generally,

given a function $f : \mathbb{R} \to \mathbb{R}$, one can define the matrix

$$f(\mathbf{R}) = \sum_{j=1}^{M} f(\lambda_j) U_j U_j^T. \tag{2.2}$$

(This holds for general symmetric matrices and, in analogous form, for self-adjoint operators in Hilbert space, provided f is measurable with respect to the spectral measure; see [10, Theorem VII.2].)

If $S \subset \mathbb{R}$ contains the eigenvalues with indices in I but no other, then clearly $P = \chi_S(\mathbf{R})$, where

$$\chi_S(\lambda) = \begin{cases} 1, & \lambda \in S, \\ 0, & \lambda \notin S \end{cases}$$

is the characteristic function of the set S.

Of course, rewriting the definition of P in this way seems of little benefit. Note, however, that for a polynomial function p, $p(\mathbf{R})$ can be evaluated directly, interpreting the powers $\mathbf{R}^n$ in the sense of matrix multiplication. Using the spectral representation

$$\mathbf{R} = \sum_{j=1}^{M} \lambda_j U_j U_j^T,$$

it is not hard to see that this calculation, for which no knowledge of the eigenvalues and eigenvectors of $\mathbf{R}$ is required, gives the same result as formula (2.2).

Even if f is not a polynomial, we can use this to find approximations for $f(\mathbf{R})$ based on the following observation. Let $(p_n)_{n\in\mathbb{N}}$ be a sequence of polynomials such that

$$\lim_{n\to\infty} \sup_{j\in\{1,\dots,M\}} |f(\lambda_j) - p_n(\lambda_j)| = 0;$$

then $\lim_{n\to\infty} p_n(\mathbf{R}) = f(\mathbf{R})$ in the Euclidean operator norm. Indeed, denoting by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^M$, we have for all $v \in \mathbb{R}^M$

$$\begin{aligned} \|(p_n(\mathbf{R}) - f(\mathbf{R}))v\|^2 &= \left\| \sum_{j=1}^{M} (p_n(\lambda_j) - f(\lambda_j)) U_j U_j^T v \right\|^2 \\ &= \sum_{j=1}^{M} |p_n(\lambda_j) - f(\lambda_j)|^2 \left| U_j^T v \right|^2 \\ &\le \sup_{j\in\{1,\dots,M\}} |p_n(\lambda_j) - f(\lambda_j)|^2 \|v\|^2 \to 0 \qquad (n \to \infty). \end{aligned}$$

In the situation at hand, we wish to omit eigenvalues below λ_{cut} and include all others, so here $f = \chi_{[\lambda_{\text{cut}},\infty)}$. In order to obtain an approximation of the corresponding spectral projector P, we replace f by an approximating polynomial p. Of course, we can expect to find only a good polynomial approximation on a compact interval, as the polynomial will grow rapidly near $\pm\infty$. However, we need only to approximate f at the eigenvalues of $\mathbf{R}$ or, as these are unknown, on an interval which contains them. We already know that the eigenvalues are nonnegative; furthermore, if $\|\cdot\|$ denotes any matrix norm satisfying

$$\|\mathbf{R}v\| \le \|\mathbf{R}\| \, \|v\| \qquad (v \in \mathbb{R}^M)$$

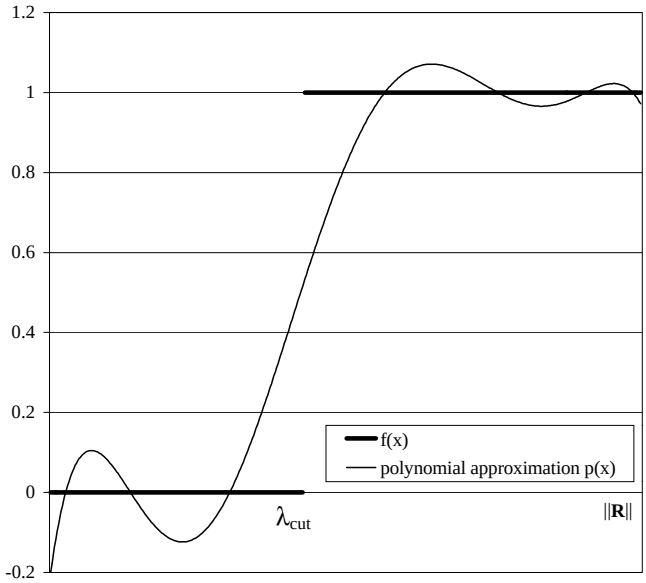


FIG. 2.1. *Function $f(x)$ and a polynomial approximation $p(x)$ of degree* 6.

for some norm $\|\cdot\|$ on $\mathbb{R}^M$, it follows that $\lambda_1 \leq \|\mathbf{R}\|$. Hence it is sufficient to approximate f on the interval $[0, \|\mathbf{R}\|]$ (see Figure 2.1).

Then we substitute the matrix $\tilde{P} = p(\mathbf{R})$ for the actual projector P to construct $\tilde{\mathbf{X}}_I = \tilde{P}X$, and hence reconstruct the time series without noise. In general, $\tilde{P}$ will not be a projector, but if p is a good approximation to f on $[0, \|\mathbf{R}\|]$, it will be close to the projector P. Moreover, $\tilde{P}$ can always be interpreted as a weighted sum of the spectral projectors $U_j U_j^T$ onto the eigenspaces for the individual λ_j by means of (2.2),

$$\tilde{P} = \sum_{j=1}^{M} p(\lambda_j) U_j U_j^T.$$

(In our approximation, we shall actually have weights $p(\lambda_j) \in [0, 1]$—see below.) Thus, even if p is only a coarse approximation of f, the resulting reconstructed time series is meaningful as a mixture of SVD components with a "fuzzy cut-off," which may even have advantages over the usual SSA reconstruction that uses a sharp selection of SVD components.

The number of components included in the projector P is $l = \operatorname{tr} P$; similarly we take $\operatorname{tr} \tilde{P}$ as an indicator of how much of the original series is included in the reconstruction. The size of $\operatorname{tr} \tilde{P}$ and its closeness to an integer can be used to assess whether the cut-off point λ_{cut} was suitably chosen. As a further guide to the choice of the cut-off point, one can obtain a rough estimate of the spectrum of $\mathbf{R}$ by performing either standard SSA or the approximate procedure outlined in the present paper with a small lag M; if this smaller window covers a stable structure (periodics) in the original time series, the corresponding eigenvalues will not change very much with increasing M and can thus give an idea of the spectral structure of $\mathbf{R}$. However, this preliminary step will miss out on large-scale structures, which will become apparent only when the approximate SSA with the large lag M is calculated.

It is common to express the eigenvalues of the lag-covariance matrix as a frac-

tion of their sum, called *eigenvalue share* (see [5] for details). Fortunately, the sum of the eigenvalues of P is known without calculating them, since by virtue of the diagonalization (1.2) we have

$$\sum_{j=1}^{M} \lambda_j = \operatorname{tr} \mathbf{R},$$

and the trace can easily be obtained as the sum of diagonal elements of $\mathbf{R}$.

3. The iterative approximation procedure. The approximation of functions of a real variable by polynomials is notoriously problematic. An accurate approximation requires polynomials of high degree, which in turn leads to strong oscillations and quick growth outside the interval of approximation, and to costly and numerically unstable computations. Therefore, one often prefers the use of localized substitutes for polynomial approximation, e.g., splines. In some situations, however, such as the present case, polynomials are the only type of function whose values can be calculated.

The following observation provides a numerically stable iterative method of calculating highly accurate polynomial approximations of a characteristic function (see Figure 3.1).

PROPOSITION 3.1. *Let $p_1(x) = 3x^2 - 2x^3$ and let p_n be the nth iteration of p_1, i.e., $p_n(x) = p_1(p_1(\dots p_1(x))\dots)$ (p_1 applied n times). Then for $n \in \mathbb{N}$, p_n is strictly increasing on $[0,1]$ and fixes the points $0, \frac{1}{2}$, and 1. Furthermore,*

$$\lim_{n\to\infty} p_n(x) = \begin{cases} 0, & x \in \left(-\frac{\sqrt{3}-1}{2}, \frac{1}{2}\right), \\ \frac{1}{2}, & x = \frac{1}{2}, \\ 1, & x \in \left(\frac{1}{2}, \frac{1+\sqrt{3}}{2}\right); \end{cases}$$

the convergence is uniform on $[-\frac{\sqrt{3}-1}{2} + \varepsilon, \frac{1}{2} - \varepsilon] \cup [\frac{1}{2} + \varepsilon, \frac{1+\sqrt{3}}{2} - \varepsilon]$ for any $\varepsilon > 0$.

The base polynomial p_1 is characterized as the lowest-degree polynomial fixing the points $0, \frac{1}{2}$, and 1, and with zero derivative at 0 and 1. The steepness of the flank for the nth iteration is

$$p_n'\left(\frac{1}{2}\right) = \left(\frac{3}{2}\right)^n.$$

The above proposition gives an approximation of the characteristic function $\chi_{[1/2,\infty)}$ on the interval $[0,1]$. By a suitable scaling transformation, we can always assume that we are dealing with a matrix with eigenvalues in $[0,1]$ and a cut-off point at $\frac{1}{2}$. Indeed, we can replace the matrix $\mathbf{R}$ by the matrix

$$B = \begin{cases} \dfrac{1}{2\lambda_{\text{cut}}}\mathbf{R} & \text{if } \lambda_{\text{cut}} \geq \|\mathbf{R}\|/2, \\ \dfrac{1}{2(\|\mathbf{R}\| - \lambda_{\text{cut}})}(\mathbf{R} + (\|\mathbf{R}\| - 2\lambda_{\text{cut}})\mathbf{I}) & \text{otherwise;} \end{cases}$$

then $\chi_{[\lambda_{\text{cut}},\infty)}(\mathbf{R}) = \chi_{[\frac{1}{2},\infty)}(B)$. (Here $\mathbf{I}$ is the identity $M \times M$ matrix.)

Remark. One may be tempted to consider calculating and storing the coefficients of the polynomial p_n instead of applying the above iterative procedure to the individual matrix B (or number x when evaluating $p_n(x)$). However, this alternative approach has severe disadvantages which make it wholly impractical. Indeed, p_n is

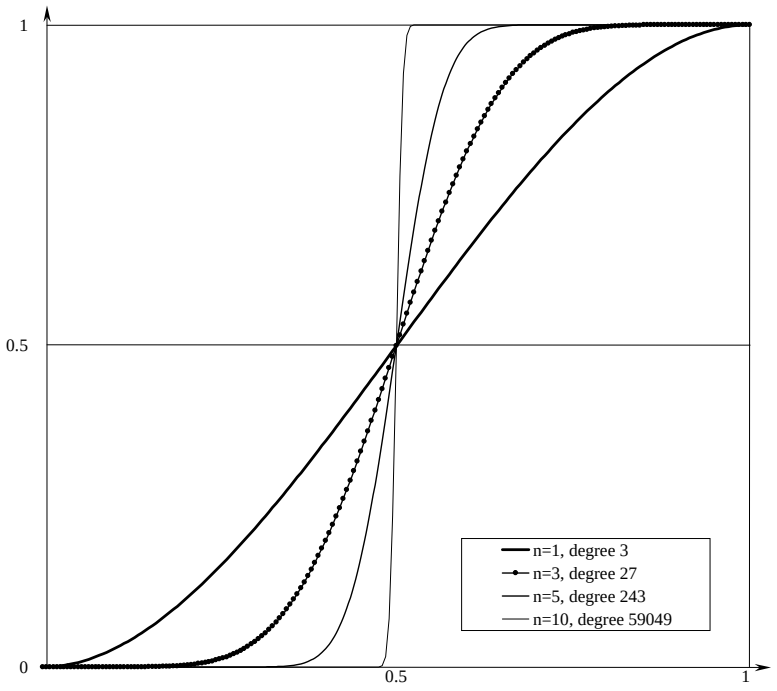


FIG. 3.1. *The approximating polynomials p_n.*

a polynomial of degree 3^n, and hence the calculation of $p_n(B)$ by the Horner scheme requires $3^n - 1$ matrix multiplications, as compared to the mere $2n$ matrix multiplications in the iterative method. Furthermore, the evaluation of the polynomial is highly unstable in view of its very large coefficients; thus even for numbers $x \in [0, 1]$, $p_4(x)$ cannot be correctly calculated in double precision (in single precision, the problem already appears in $p_3(x)$).

In contrast, the iterative method, in addition to its numerical stability, provides the possibility of monitoring the progress of the approximation, stopping when the desired accuracy is reached, rather than fixing the degree of the approximating polynomial in advance.

4. Geometric forecasting. The (approximate) projector P can also be used to forecast the given time series.

A simple geometric method of forecasting a time series $x_1, x_2, \ldots, x_N$ using components $\mathbf{X}_i$, $i \in I$, of the SVD of its trajectory matrix is based on the principle of choosing the next term x_{N+1} in the series in such a way that the vector $\mathbf{x} = (x_{N+2-M}, \ldots, x_N, x_{N+1})^T$ is closest to the subspace of $\mathbb{R}^M$ spanned by the eigenvectors U_j, $j \in I$. Taking the Euclidean norm in $\mathbb{R}^M$ as a measure for the closeness, we can express this in terms of the orthogonal projector P (see (2.1)) as the problem of minimizing the norm of the difference vector between $\mathbf{x}$ and its orthogonal projection $P\mathbf{x}$,

$$\|(\mathbf{I} - P)\mathbf{x}\|^2 \to \min,$$

varying x_{N+1}.

The minimum satisfies

$$0 = \frac{d}{dx_{N+1}} \|(\mathbf{I} - P)\mathbf{x}\|^2 = 2\mathbf{x}^T(\mathbf{I} - P)(\mathbf{I} - P)_M,$$

where $(\mathbf{I} - P)_M$ denotes the last column of the matrix $(\mathbf{I} - P)$. In other words, the minimizing vector $\mathbf{x}$ is orthogonal to the *forecast vector* $f = (\mathbf{I} - P)(\mathbf{I} - P)_M$. If $f_M \neq 0$, we thus find the recurrence for x_{N+1},

$$x_{N+1} = -\frac{1}{f_M} \sum_{j=1}^{M-1} f_j \, x_{N+1-M+j}.$$

This recurrence formula is meaningful and can be used to forecast the time series even when the orthogonal projector P is replaced by the approximate projector $\tilde{P}$ calculated according to the method outlined in section 3. Note that the resulting approximate forecast vector converges to the exact one as $\tilde{P}$ approaches P in the Euclidean operator norm.

5. Numerical examples. Let us study two examples to illustrate the approximate SSA algorithm and geometric forecasting described above.

First consider the simulated time series

$$z_t = \sin(0.1t) + e_t,$$

where the e_t are independent identically distributed random variables $e_t \sim N(0, 16)$ for $t = 1, \ldots, 2000$ (white noise).

Figure 5.1 shows the result of the approximate SSA with maximal lag $M = 1000$, cut-off point $\lambda_{\text{cut}} = 1\%$, and an approximating polynomial of degree 243 (i.e., 5 iterations). The trace of the approximate projector $\tilde{P}$ is 21.56, indicating that SSA components with a total weight corresponding to about 22 eigenvectors are included in the reconstruction.

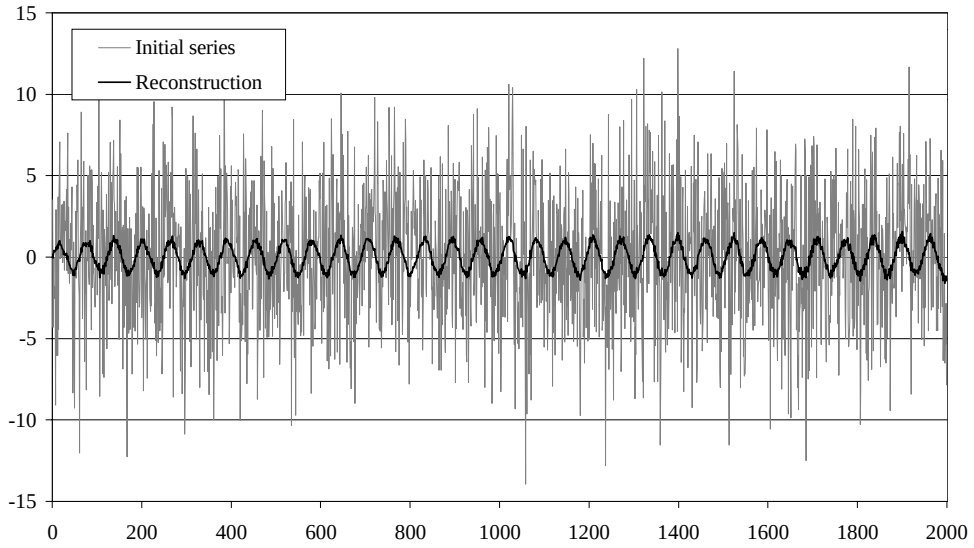


FIG. 5.1. $\sin(0.1t) + e_t$, $e_t \sim N(0, 16)$ *and its reconstruction.*

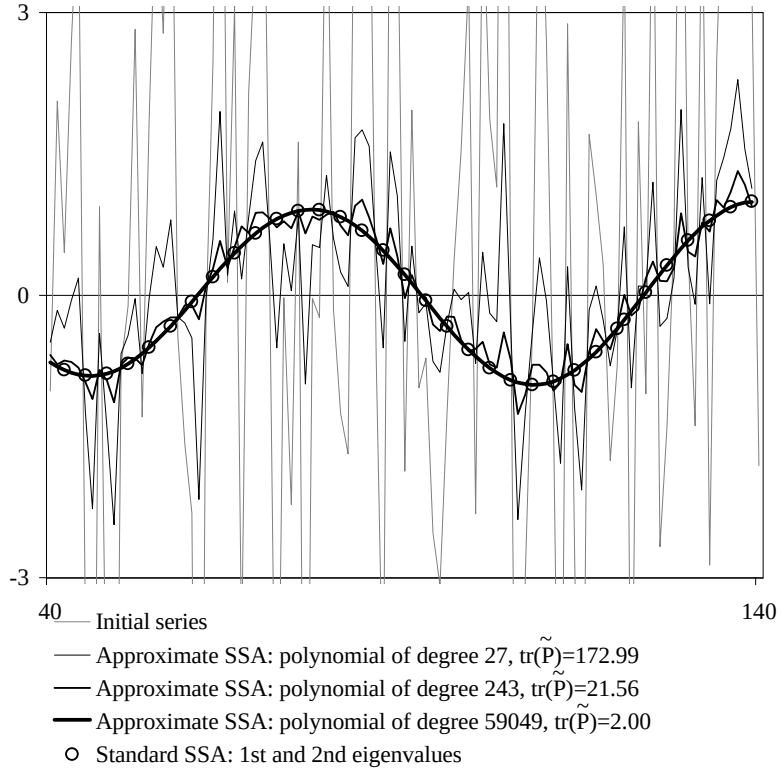


FIG. 5.2. *Comparison of reconstructions using different polynomials.*

In Figure 5.2 a zoomed part of the time series is presented. One can clearly observe how the degree of the approximating polynomial, and hence the sharpness of the cut-off in the calculation of the approximate projector, influences the quality of the reconstruction. In the case of the polynomial of degree 59049 (i.e., 10 iterations) the reconstruction by our approximate method reproduces the result of standard SSA; then the number of principal components included in the latter coincides (within some tolerance) with $\mathrm{tr}(\tilde{P})$.

Note, however, that already the reconstruction based on the coarser approximation of Figure 5.1, with $\tilde{P}$ still far from the actual projector, clearly picks out the signal from the very noisy time series.

Our second example is based on the well-known real-life data (see [8]) of monthly averages of hotel rooms occupied from 1963–1976; this example was studied in detail in [5]. We demonstrate by this example how different choices of the cut-off point λ_{cut} exhibit various features of the time series in the resulting reconstructions (see Figure 5.3).

To pick up the first eigenvalue (corresponding to the trend) we have chosen lag $M = 84$ and $\lambda_{\mathrm{cut}} = 2\%$, and 15 iterations were enough to separate it from the rest of the spectrum ($\mathrm{tr}\,\tilde{P} = 1.0000$). With $\lambda_{\mathrm{cut}} = 0.51\%$ and 19 iterations, we also include the second and third eigenvalues ($\mathrm{tr}\,\tilde{P} = 3.0000$), corresponding to the one year cycle. For a more precise reconstruction of the initial data we have taken the cut-off point $\lambda_{\mathrm{cut}} = 0.11\%$ and 23 iterations ($\mathrm{tr}\,\tilde{P} = 5.0000$). By this choice we include exactly the first five components, which describe the main structure of the series: linear

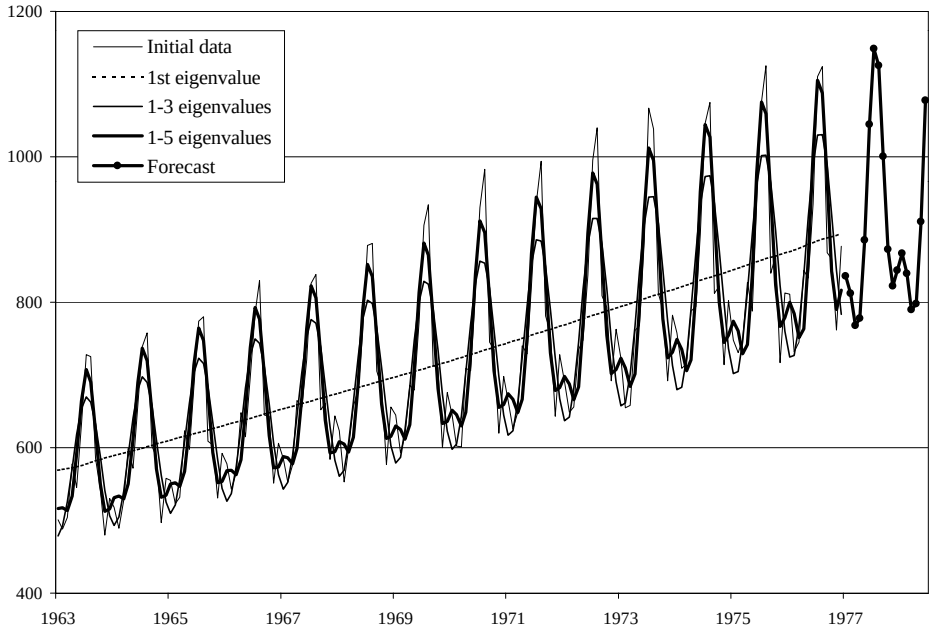


FIG. 5.3. *Occupied hotel rooms (average per month),* 1963–1976.

trend, one-year and half-year periodicities. The geometric forecast based on these components does not contradict the structure of the series.

Acknowledgments. The authors thank V. Nekrutkin and A. Zhigljavsky for their interest and advice.

REFERENCES

[1] E. BIGLIERI AND K. YAO, *Some properties of singular value decomposition and their applications to digital signal processing*, Signal Process., 18 (1989), pp. 277–289.

[2] D.S. BROOMHEAD, R. JONES, AND G.P. KING, *Topological dimension and local coordinates from time series data*, J. Phys. A, 20 (1987), pp. L563–L569.

[3] D.S. BROOMHEAD AND G.P. KING, *Extracting qualitative dynamics from experimental data*, Phys. D, 20 (1986), pp. 217–236.

[4] J. ELSNER AND A. TSONIS, *Singular Spectrum Analysis: A New Tool in Time Series Analysis*, Plenum Press, New York, 1996.

[5] N. GOLJANDINA, V. NEKRUTKIN, AND A. ZHIGLJAVSKY, *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman and Hall, London, 2001.

[6] R. KAKARALA AND P.O. OGUNBONA, *Signal analysis using a multiresolution form of the singular value decomposition*, IEEE Trans. Image Process., 10 (2001), pp. 723–735.

[7] T. MORITA AND T. KANADE, *A sequential factorization method for recovering shape and motion from image streams*, IEEE Trans. Pattern Anal. and Machine Intell., 19 (1997), pp. 858–867.

[8] T.M. O'DONOVAN, *Short Term Forecasting: An Introduction to the Box-Jenkins Approach*, Wiley, New York, 1983.

[9] H. PARK, L. ZHANG, AND J.B. ROSEN, *Low rank approximation of a Hankel matrix by structured total least norm*, BIT, 39 (1999), pp. 757–779.

[10] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics* I. *Functional Analysis*, Academic Press, New York, 1980.

[11] G.X. RITTER AND P. SUSSNER, *The minimax eigenvalue transform*, in Image Algebra and Morphological Image Processing, III (San Diego, CA, 1992), Proc. SPIE 1769, SPIE, Bellingham, WA, 1992, pp. 276–282.

[12] G.W. Stewart, *On the early history of the singular value decomposition*, SIAM Rev., 35 (1993), pp. 551–566.
[13] A.-J. van der Veen, *A Schur method for low-rank approximation*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 139–160.
[14] S. van Huffel, *Partial singular value decomposition algorithm*, J. Comput. Appl. Math., 33 (1990), pp. 105–112.
[15] R. Vautard and M. Ghil, *Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series*, Phys. D, 35 (1989), pp. 395–424.
[16] R. Vautard, P. Yiou, and M. Ghil, *Singular-spectrum analysis: A toolkit for short, noisy chaotic signals*, Phys. D, 58 (1992), pp. 95–126.
[17] C.R. Vogel and J.G. Wade, *Iterative SVD-based methods for ill-posed problems*, SIAM J. Sci. Comput., 15 (1994), pp. 736–754.
[18] F. Vogt and M. Tacke, *Fast principal component analysis of large data sets*, Chemometrics and Intell. Lab. Systems, 59 (2001), pp. 1–18.
[19] P. Yiou, D. Sornette, and M. Ghil, *Data-adaptive wavelets and multi-scale singular-spectrum analysis*, Phys. D, 142 (2000), pp. 254–290.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 943–962

REDUCTION TO VERSAL DEFORMATIONS OF MATRIX PENCILS AND MATRIX PAIRS WITH APPLICATION TO CONTROL THEORY*

M. I. GARCÍA-PLANAS† AND A. A. MAILYBAEV‡

Abstract. Matrix pencils under the strict equivalence and matrix pairs under the state feedback equivalence are considered. It is known that a matrix pencil (or a matrix pair) smoothly dependent on parameters can be reduced locally to a special typically more simple form, called the versal deformation, by a smooth change of parameters and a strict equivalence (or feedback equivalence) transformation. We suggest an explicit recurrent procedure for finding the change of parameters and equivalence transformation in the reduction of a given family of matrix pencils (or matrix pairs) to the versal deformation. As an application, this procedure is applied to the analysis of the uncontrollability set in the space of parameters for a one-input linear dynamical system. Explicit formulae for a tangent plane to the uncontrollability set at its regular point and the perturbation of the uncontrollable mode are derived. A physical example is given and studied in detail.

Key words. versal deformation, matrix pencil, matrix pair, feedback equivalence, controllability

AMS subject classifications. 15A21, 93B05, 93B52

PII. S0895479801392016

1. Introduction. The Arnold technique of constructing a local canonical form, called versal deformation, of a differentiable family of square matrices under conjugation [1, 2] has been generalized by several authors to matrix pencils under the strict equivalence [4, 10], pairs or triples of matrices under the action of the general linear group [18], pairs of matrices under the feedback similarity [6], and triples or quadruples of matrices representing linear dynamical systems under the equivalence derived from standard transformations (the change of basis in state, input, and output spaces, state feedback, and output injection) [8, 9]. Versal deformations provide a special parametrization of matrix spaces, which can be effectively applied to perturbation analysis and investigation of complicated objects like singularities and bifurcations in multiparameter dynamical systems [1, 2, 3, 4, 5, 12, 14, 15].

The general notion of versality is the following. Let $\mathcal{M}$ be a differential manifold with the equivalence relation defined by the action of a Lie group $\mathcal{G}$. The $\mathcal{G}$-action is described by the mapping $x \longrightarrow g \circ x$, where $x, g \circ x \in \mathcal{M}$ and $g \in \mathcal{G}$. The classical example is the space of square complex matrices $\mathcal{M} = M_{m\times m}(\mathbb{C})$ with the Lie group $\mathcal{G} = \mathrm{GL}(m, \mathbb{C})$ determining the similarity transformation (the change of basis) $A \longrightarrow C^{-1}AC$, where $A \in M_{m\times m}(\mathbb{C})$ and $C \in \mathrm{GL}(m, \mathbb{C})$. Let us consider a smooth mapping $x : \mathcal{U}_0 \longrightarrow \mathcal{M}$, where $\mathcal{U}_0$ is a neighborhood of the origin of the space $\mathbb{F}^\ell$; $\mathbb{F}$ stands for the space of real or complex numbers. The mapping $x(\gamma)$ is called a deformation of $x_0 = x(0)$ with the parameter vector $\gamma \in \mathbb{F}^\ell$. Introducing a change of parameters $\phi : \mathcal{U}_0' \longrightarrow \mathcal{U}_0$, where $\mathcal{U}_0'$ is a neighborhood of the origin in $\mathbb{F}^k$, such that $\phi(0) = 0$, we obtain the deformation $x(\phi(\xi))$ of x_0 with the parameter vector

*Received by the editors July 3, 2001; accepted for publication (in revised form) by P. Van Dooren August 5, 2002; published electronically February 12, 2003. This work was supported by INTAS Young Scientists Fellowship 00-58.

http://www.siam.org/journals/simax/24-4/39201.html

†Dept. de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Minería 1, Esc. C, 1-3, 08038 Barcelona, Spain (maria.isabel.garcia@upc.es).

‡Institute of Mechanics, Moscow State University, Michurinsky pr. 1, 117192 Moscow, Russia (mailybaev@imec.msu.ru).

$\xi \in \mathcal{U}_0' \subset \mathbb{F}^k$. Applying the equivalence transformation $g(\xi)$, where $g : \mathcal{U}_0' \longrightarrow \mathcal{G}$ is a smooth mapping such that $g(0) = e$ is the unit element of $\mathcal{G}$, we get the deformation

$$z(\xi) = g(\xi) \circ x(\phi(\xi)) \tag{1.1}$$

of $z(0) = e \circ x_0 = x_0$. Then $x(\gamma)$ is called a versal deformation of x_0 if any deformation $z(\xi)$ of x_0 can be represented in the form (1.1) in some neighborhood of the origin $\mathcal{U}_0'' \subset \mathbb{F}^k$. This definition implies that a versal deformation generates all deformations of x_0 and, hence, possesses properties (invariant under the equivalence transformation) of all deformations of the given element $x_0 \in \mathcal{M}$.

The theorem given by Arnold [1, 2] says that the deformation $x(\gamma)$ of x_0 is versal if and only if it is transversal to the orbit of x_0 under the action of $\mathcal{G}$. This theorem reduces the problem of finding a versal deformation to solving a specific linear equation determined by x_0. This method allows finding versal deformations $x(\gamma)$ having simple form, which can be treated as local canonical forms. For the reduction of a given deformation $z(\xi)$ to this form, one needs to find the change of parameters $\gamma = \phi(\xi)$ and the equivalence transformation $g(\xi)$ smoothly depending on ξ, which satisfy locally equality (1.1).

In this paper versal deformations of matrix pencils under the strict equivalence and pairs of matrices under the feedback equivalence are considered. The method of finding the change of basis $\gamma = \phi(\xi)$ and the equivalence transformation $g(\xi)$, which reduce a given deformation $z(\xi)$ to the versal deformation, is developed. The mappings $\phi(\xi)$ and $g(\xi)$ are represented in the form of Taylor series, whose coefficients are found from the explicit recurrent procedure. This approach is the generalization to these particular cases of the one presented by Mailybaev [12, 13] for spaces of square matrices under conjugation; see also [5, 17] for related problems.

A pair of matrices $(F, G) \in M_{m\times m}(\mathbb{R}) \times M_{m\times n}(\mathbb{R})$ determines the linear dynamical system $\dot{\psi} = F\psi + G\nu$ with the state vector $\psi \in \mathbb{R}^m$ and input vector $\nu \in \mathbb{R}^n$. The controllability of this system (the possibility of reaching any state ψ by choosing an appropriate input vector $\nu(t)$) is an invariant property under the feedback equivalence transformation. Using this fact, we apply the method presented in this paper to study the uncontrollability set of a multiparameter one-input linear dynamical system. As a result, explicit formulae for the tangent plane to the uncontrollability set at its regular point and the perturbation of the uncontrollable mode (the generalized eigenvalue) are derived. Note that this approach provides a simple and systematic way for the perturbation analysis of the uncontrollability set, while the classical controllability condition related to the rank of a certain matrix (called the controllability matrix) is difficult to use for multiparameter perturbation analysis.

The organization of the paper is as follows. In section 2 the case of matrix pencils under the strict equivalence is considered. The local structure of the orbit and stabilizer of a matrix pencil is described by a specific linear function (differential of the equivalence transformation mapping) and its adjoint. Using this information, a versal deformation $x(\gamma)$ is determined. Then the change of basis $\gamma = \phi(\xi)$ and the equivalence transformation $g(\xi)$ for the reduction of a given deformation $z(\xi)$ to this versal deformation are found in the form of Taylor series. Section 3 studies the case of pairs of matrices under the feedback equivalence. In section 4 the obtained results are applied to the perturbation analysis of the uncontrollability set for a one-input linear dynamical system dependent on parameters. A physical example is given and studied in detail. The conclusion discusses applicability issues of the presented method and its importance for the versal deformation theory.

2. Matrix pencils and their deformations. Let us consider a space of matrix pencils $\mathcal{M} = \{A - \lambda B \mid A, B \in M_{m\times n}(\mathbb{F})\}$, where $M_{m\times n}(\mathbb{F})$ is a set of $m \times n$ matrices with real or complex elements, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. In this space we consider the following equivalence relation [7]: two pencils $A_1 - \lambda B_1$ and $A_2 - \lambda B_2$ are (strict) equivalent if and only if

$$A_2 - \lambda B_2 = P^{-1}(A_1 - \lambda B_1)Q \tag{2.1}$$

for some nonsingular square matrices $P \in \mathrm{Gl}(m;\mathbb{F})$, $Q \in \mathrm{Gl}(n;\mathbb{F})$.

2.1. Equivalence as a Lie group action. Equivalence relation (2.1) may be seen as induced by the action of a Lie group $\mathcal{G} = \{(P, Q) \mid P \in \mathrm{Gl}(m;\mathbb{F}), Q \in \mathrm{Gl}(n;\mathbb{F})\}$. Using the short notation $g = (P, Q) \in \mathcal{G}$ and $x = A - \lambda B \in \mathcal{M}$, we define multiplication in $\mathcal{G}$, action of the group $\mathcal{G}$, and equivalence condition (2.1) as follows:

$$\begin{gathered} g_1 g_2 = (P_1 P_2, Q_1 Q_2) \in \mathcal{G}, \\ g \circ x = P^{-1}(A - \lambda B)Q \in \mathcal{M}, \\ x_2 = g \circ x_1. \end{gathered} \tag{2.2}$$

Multiplication in the group corresponds to successive equivalence transformations: $g_2 \circ (g_1 \circ x) = (g_1 g_2) \circ x$. The unit element of $\mathcal{G}$ has the form $e = (I_m, I_n)$, where I_m and I_n are the identity matrices.

Let us fix a pencil $x_0 = A_0 - \lambda B_0 \in \mathcal{M}$ and define the mapping

$$f_{x_0}(g) = g \circ x_0. \tag{2.3}$$

The equivalence class of the pencil x_0 with respect to the action of $\mathcal{G}$ is the range of the function f_{x_0}. It is called the *orbit* of x_0 and denoted by

$$\mathcal{O}(x_0) = \operatorname{Im} f_{x_0} = \{g \circ x_0 \mid g \in \mathcal{G}\}. \tag{2.4}$$

The *stabilizer* of x_0 under the $\mathcal{G}$-action is a null-space of the function $f_{x_0} - x_0$. We denote it by

$$\mathcal{S}(x_0) = \operatorname{Ker}(f_{x_0} - x_0) = \{g \in \mathcal{G} \mid g \circ x_0 = x_0\}. \tag{2.5}$$

The mapping f_{x_0} is differentiable, and $\mathcal{O}(x_0)$ and $\mathcal{S}(x_0)$ are smooth submanifolds of $\mathcal{M}$ and $\mathcal{G}$, respectively.

Let us use the notation $T_e\mathcal{G}$ for a tangent space to the manifold $\mathcal{G}$ at the unit element e. Since $\mathcal{G}$ is an open subset of $M_{m\times m}(\mathbb{F}) \times M_{n\times n}(\mathbb{F})$, we have

$$T_e\mathcal{G} = \{(U, V) \mid U \in M_{m\times m}(\mathbb{F}), V \in M_{n\times n}(\mathbb{F})\}$$

and, since $\mathcal{M}$ is a linear space,

$$T_{x_0}\mathcal{M} = \mathcal{M}.$$

The Euclidean scalar products in the spaces $\mathcal{M}$ and $T_e\mathcal{G}$ considered in this paper are defined as follows:

$$\begin{aligned} &\langle x_1, x_2\rangle_1 = \operatorname{trace}(A_1 A_2^*) + \operatorname{trace}(B_1 B_2^*), \quad x_i = A_i - \lambda B_i \in \mathcal{M}, \\ &\langle y_1, y_2\rangle_2 = \operatorname{trace}(U_1 U_2^*) + \operatorname{trace}(V_1 V_2^*), \quad y_i = (U_i, V_i) \in T_e\mathcal{G}, \end{aligned} \tag{2.6}$$

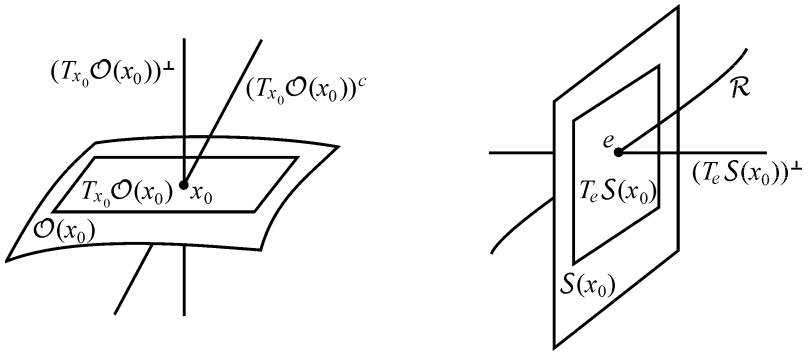


FIG. 1. *Local structure of the orbit $\mathcal{O}(x_0)$ and stabilizer $\mathcal{S}(x_0)$.*

where A^* denotes the conjugate transpose of a matrix A.

Let $df_{x_0} : T_e\mathcal{G} \longrightarrow \mathcal{M}$ be the differential of f_{x_0} at the unit element e. Using expressions (2.2) and (2.3), we find [4]

$$df_{x_0}(y) = (A_0V - UA_0) - \lambda(B_0V - UB_0) \in \mathcal{M}, \quad y = (U, V) \in T_e\mathcal{G}. \tag{2.7}$$

The adjoint linear mapping $df^*_{x_0} : \mathcal{M} \longrightarrow T_e\mathcal{G}$ is defined by the relation

$$\langle df_{x_0}(y), z\rangle_1 = \langle y, df^*_{x_0}(z)\rangle_2, \quad y \in T_e\mathcal{G}, \quad z \in \mathcal{M}. \tag{2.8}$$

Using expressions (2.6) and (2.7) in (2.8), it is straightforward to find

$$df^*_{x_0}(z) = (-XA_0^* - YB_0^*, A_0^*X + B_0^*Y) \in T_e\mathcal{G}, \quad z = X - \lambda Y \in \mathcal{M}. \tag{2.9}$$

The mappings df_{x_0} and $df^*_{x_0}$ provide a simple description of the tangent spaces $T_{x_0}\mathcal{O}(x_0)$, $T_e\mathcal{S}(x_0)$ and their normal complements $(T_{x_0}\mathcal{O}(x_0))^{\perp}$, $(T_e\mathcal{S}(x_0))^{\perp}$; see Figure 1.

THEOREM 2.1. *The tangent spaces to the orbit and stabilizer of the matrix pencil x_0 and the corresponding normal complementary subspaces with respect to $\mathcal{M}$ and $T_e\mathcal{G}$ can be found in the following form:*

1. $T_{x_0}\mathcal{O}(x_0) = \operatorname{Im} df_{x_0} \subset \mathcal{M}$.
2. $(T_{x_0}\mathcal{O}(x_0))^{\perp} = \operatorname{Ker} df^*_{x_0} \subset \mathcal{M}$.
3. $T_e\mathcal{S}(x_0) = \operatorname{Ker} df_{x_0} \subset T_e\mathcal{G}$.
4. $(T_e\mathcal{S}(x_0))^{\perp} = \operatorname{Im} df^*_{x_0} \subset T_e\mathcal{G}$.

Proof. Assertions 1 and 3 follow from (2.4), (2.5), and the definition of df_{x_0} as the differential of the function f_{x_0} at e. Then assertions 2 and 4 follow from properties of the adjoint function $df^*_{x_0}$ [7]. ∎

COROLLARY 2.2. *The mappings df_{x_0} and $df^*_{x_0}$ define one-to-one correspondences between the subspaces $T_{x_0}\mathcal{O}(x_0)$ and $(T_e\mathcal{S}(x_0))^{\perp}$:*

$$T_{x_0}\mathcal{O}(x_0) \underset{df_{x_0}}{\overset{df^*_{x_0}}{\rightleftarrows}} (T_e\mathcal{S}(x_0))^{\perp}.$$

Example 2.1. Let us consider a matrix pencil

$$x_0 = A_0 - \lambda B_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.10}$$

According to Theorem 2.1, the elements $z \in (T_{x_0}\mathcal{O}(x_0))^{\perp}$ can be found by solving the linear system $df^*_{x_0}(z) = 0$ with $df^*_{x_0}$ given by expression (2.9). As a result, we obtain a general element of $(T_{x_0}\mathcal{O}(x_0))^{\perp}$ in the form

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ \gamma_1 & 0 & \gamma_3 & 0 \\ \gamma_2 & \gamma_2 & 0 & \gamma_4 \end{pmatrix} - \lambda \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\gamma_2 & -\gamma_2 & 0 & -\gamma_4 \end{pmatrix}, \tag{2.11}$$

where $\gamma_1, \ldots, \gamma_4 \in \mathbb{F}$ are arbitrary; $\dim(T_{x_0}\mathcal{O}(x_0))^{\perp} = 4$. Using (2.11), it is straightforward to find a general element of the space $T_{x_0}\mathcal{O}(x_0)$ as follows:

$$\begin{pmatrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ 0 & \mu_5 & 0 & \mu_6 \\ \mu_7 + \mu_9 & \mu_8 - \mu_9 & \mu_{10} & \mu_{11} \end{pmatrix} - \lambda \begin{pmatrix} \mu_{12} & \mu_{13} & \mu_{14} & \mu_{15} \\ \mu_{16} & \mu_{17} & \mu_{18} & \mu_{19} \\ \mu_7 - \mu_9 & \mu_8 + \mu_9 & \mu_{20} & \mu_{11} \end{pmatrix}, \tag{2.12}$$

where $\mu_1, \ldots, \mu_{20} \in \mathbb{F}$ are arbitrary; $\dim T_{x_0}\mathcal{O}(x_0) = 20$. Using (2.12) in Corollary 2.2, we find a general element of the space $(T_e\mathcal{S}(x_0))^{\perp} = df^*_{x_0}(T_{x_0}\mathcal{O}(x_0))$ in the form

$$\left(\begin{pmatrix} -\mu_2 - \mu_{12} & -\mu_{14} & -\mu_4 - \mu_{15} \\ -\mu_5 - \mu_{16} & -\mu_{18} & -\mu_6 - \mu_{19} \\ 2\mu_9 - \mu_8 - \mu_7 & -\mu_{20} & -2\mu_{11} \end{pmatrix}, \begin{pmatrix} \mu_{12} & \mu_{13} & \mu_{14} & \mu_{15} \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \mu_{16} & \mu_{17} & \mu_{18} & \mu_{19} \\ 2\mu_7 & 2\mu_8 & \mu_{10} + \mu_{20} & 2\mu_{11} \end{pmatrix} \right). \tag{2.13}$$

Finally, we obtain elements of the space $T_e\mathcal{S}(x_0)$ from the equation $df_{x_0}(y) = 0$ as follows:

$$\left(\begin{pmatrix} \nu_1 & \nu_2 & \nu_3 \\ 0 & \nu_4 & 0 \\ 0 & 0 & \nu_5 \end{pmatrix}, \begin{pmatrix} \nu_1 & 0 & \nu_2 & \nu_3 \\ 0 & \nu_1 & 0 & \nu_3 \\ 0 & 0 & \nu_4 & 0 \\ 0 & 0 & 0 & \nu_5 \end{pmatrix} \right), \tag{2.14}$$

where $\nu_1, \ldots, \nu_5 \in \mathbb{F}$ are arbitrary; $\dim T_e\mathcal{S}(x_0) = 5$.

2.2. Versal deformation. Let $\mathcal{U}_0$ be a neighborhood of the origin of $\mathbb{F}^{\ell}$. A *deformation* $x(\gamma)$ of x_0 is a smooth mapping

$$x : \mathcal{U}_0 \longrightarrow \mathcal{M}$$

such that $x(0) = x_0$. The vector $\gamma = (\gamma_1, \ldots, \gamma_{\ell}) \in \mathcal{U}_0$ is called the parameter vector. The deformation $x(\gamma)$ is also called the *family* of matrix pencils. The deformation $x(\gamma)$ of x_0 is called *versal* if any deformation $z(\xi)$ of x_0, where $\xi = (\xi_1, \ldots, \xi_k) \in \mathcal{U}'_0 \subset \mathbb{F}^k$ is the parameter vector, can be represented in some neighborhood of the origin in the following form:

$$z(\xi) = g(\xi) \circ x(\phi(\xi)), \quad \xi \in \mathcal{U}''_0 \subset \mathcal{U}'_0, \tag{2.15}$$

where $\phi : \mathcal{U}''_0 \longrightarrow \mathbb{F}^{\ell}$ and $g : \mathcal{U}''_0 \longrightarrow \mathcal{G}$ are differentiable mappings such that $\phi(0) = 0$ and $g(0) = e$. Expression (2.15) means that any deformation $z(\xi)$ of x_0 can be obtained from the versal deformation $x(\gamma)$ of x_0 by an appropriate smooth change of parameters $\gamma = \phi(\xi)$ and equivalence transformation $g(\xi)$ smoothly dependent on parameters. The versal deformation with minimal possible number of parameters ℓ is called *miniversal*.

The following result, proved by Arnold [1, 2] for $\mathrm{Gl}(n;\mathbb{C})$ acting on $M_{n\times n}(\mathbb{C})$, and generalized by Tannenbaum [18] for a Lie group acting on a complex manifold, provides the relation between the versal deformation of x_0 and the local structure of the orbit and stabilizer of x_0.

THEOREM 2.3.

1. *A deformation $x(\gamma)$ of x_0 is versal if and only if it is transversal to the orbit $\mathcal{O}(x_0)$ at x_0.*
2. *The minimal number of parameters of a versal deformation is equal to the codimension of the orbit of x_0 in $\mathcal{M}$, $\ell = \operatorname{codim}\mathcal{O}(x_0)$.*
3. *If $x(\gamma)$ is a miniversal deformation and values of the mapping $g(\xi)$ are restricted to belong to a smooth submanifold $\mathcal{R} \subset \mathcal{G}$, which is transversal to $\mathcal{S}(x_0)$ at e and has the minimal dimension $\dim\mathcal{R} = \operatorname{codim}\mathcal{S}(x_0)$, then the mappings $\phi(\xi)$ and $g(\xi)$ in representation* (2.15) *are uniquely determined by $z(\xi)$.*

Note that the third assertion of Theorem 2.3 was not explicitly stated in [1, 2, 18] but proved in the proof of the corresponding theorem.

Let us denote by $\{t_1,\dots,t_d\}$, $d = \dim T_{x_0}\mathcal{O}(x_0)$, a basis of the tangent space $T_{x_0}\mathcal{O}(x_0)$; by $\{n_1,\dots,n_\ell\}$, $\ell = \operatorname{codim} T_{x_0}\mathcal{O}(x_0)$, a basis the normal complement $(T_{x_0}\mathcal{O}(x_0))^\perp$; by $\{c_1,\dots,c_\ell\}$ a basis of an arbitrary complementary subspace $(T_{x_0}\mathcal{O}(x_0))^c$ to $T_{x_0}\mathcal{O}(x_0)$; and by $\{r_1,\dots,r_d\}$ a basis of $(T_e\mathcal{S}(x_0))^\perp$. By Corollary 2.2, if we have the basis $\{t_1,\dots,t_d\}$, then the basis $\{r_1,\dots,r_d\}$ can be chosen in the form $\{df^*_{x_0}(t_1),\dots,df^*_{x_0}(t_d)\}$, and, vice versa, if the basis $\{r_1,\dots,r_d\}$ is known, then we can choose the basis $\{t_1,\dots,t_d\}$ in the form $\{df_{x_0}(r_1),\dots,df_{x_0}(r_d)\}$.

COROLLARY 2.4. *The deformation*

$$x(\gamma) = x_0 + \sum_{i=1}^{\ell} c_i\gamma_i \tag{2.16}$$

is a miniversal deformation. The functions $\phi(\xi)$ and $g(\xi)$ in the versal deformation reduction (2.15) *are uniquely determined if the mapping $g(\xi)$ is taken in the form*

$$g(\xi) = e + \sum_{j=1}^{d} r_j\mu_j(\xi), \tag{2.17}$$

where $\mu_j(\xi)$ are smooth functions in $\mathbb{F}$ such that $\mu_j(0) = 0$, $j = 1,\dots,d$.

If we take $c_i = n_i$, $i = 1,\dots,\ell$, in (2.16), then the corresponding miniversal deformation is called *orthogonal*.

If the pencil $x_0 = A_0 - \lambda B_0$ is reduced to the Kronecker canonical form (this is not a restriction because of the homogeneity of the orbit), it is possible to write down explicitly the bases $\{c_1,\dots,c_\ell\}$, $\{n_1,\dots,n_\ell\}$, $\{t_1,\dots,t_d\}$, and $\{r_1,\dots,r_d\}$. Explicit forms of the bases $\{c_1,\dots,c_\ell\}$ and $\{n_1,\dots,n_\ell\}$ were given in [4, 10].

Example 2.2. Let us consider a matrix pencil (2.10). The matrix pencils n_i, t_j and matrix pairs r_j can be obtained from (2.11), (2.12), and (2.13), respectively, by taking $\gamma_i = \mu_j = 1$ and zeros for other variables. Using the explicit form of the tangent space $T_{x_0}\mathcal{O}(x_0)$ given in (2.12), we can choose a basis $\{c_1,\dots,c_\ell\}$, $\ell = 4$, of a complementary subspace $(T_{x_0}\mathcal{O}(x_0))^c$ such that every c_i has exactly one nonzero element. This will give us a *simplest* miniversal deformation, for example,

$$x(\gamma) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \gamma_1 & 0 & \gamma_3 & 0 \\ \gamma_2 & 0 & 0 & 1+\gamma_4 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.18}$$

2.3. Reduction to miniversal deformation. Let us assume that the pencil x_0 and its miniversal deformation $x(\gamma)$ in the form (2.16) are given. To reduce an arbitrary deformation $z(\xi)$ of x_0 to the miniversal deformation, we need to find smooth mappings $\phi(\xi)$ and $g(\xi)$ satisfying relation (2.15). Recall that these mappings are unique if $g(\xi)$ is taken in the form (2.17). Since these mappings are determined in the neighborhood of the origin $\xi = 0$, they can be represented in Taylor series form.

Let $h = (h_1, \dots, h_k)$ be a vector with nonnegative integer components $h_i \in \mathbb{Z}_+$. We will use the conventional notation

$$|h| = h_1 + \cdots + h_k, \quad h! = h_1! \cdots h_k!, \quad C_h^{h'} = \frac{h!}{h'!(h-h')!},$$

$$\xi^h = \xi_1^{h_1} \cdots \xi_k^{h_k}, \quad \phi^{(h)} = \frac{\partial^{|h|}\phi}{\partial\xi^{h_1} \cdots \partial\xi^{h_k}},$$

where derivatives are evaluated at $\xi = 0$; the derivative of zero order denotes the function value at zero, i.e., $\phi^{(0)} = \phi(0)$. Using expression (2.17), we can write the Taylor series for the mappings $\phi(\xi)$ and $g(\xi)$ as

$$\begin{aligned} \phi(\xi) &= \sum_{|h|\le s} \frac{\phi^{(h)}}{h!}\,\xi^h + o(\|\xi\|^s), \\ g(\xi) &= e + \sum_{j=1}^{d} r_j \sum_{|h|\le s} \frac{\mu_j^{(h)}}{h!}\,\xi^h + o(\|\xi\|^s), \end{aligned} \tag{2.19}$$

where $\phi^{(0)} = 0$ and $\mu_j^{(0)} = 0$; $\|\xi\|$ is the norm in the parameter space $\mathbb{F}^k$. Therefore, to find the transformation functions $\phi(\xi)$ and $g(\xi)$, we need to determine the derivatives $\phi^{(h)} = (\phi_1^{(h)}, \dots, \phi_\ell^{(h)})$ and $\mu_1^{(h)}, \dots, \mu_d^{(h)}$. The following theorem provides explicit recurrent formulae for calculation of these derivatives up to an arbitrary order $|h|$.

THEOREM 2.5. *The derivatives $\phi_1^{(h)}, \dots, \phi_\ell^{(h)}$ and $\mu_1^{(h)}, \dots, \mu_d^{(h)}$ determining transformation functions* (2.19)*, which reduce the deformation $z(\xi)$ of x_0 to the miniversal deformation* (2.16)*, satisfy the recurrent formulae*

$$\begin{pmatrix} \phi_1^{(h)} \\ \vdots \\ \phi_\ell^{(h)} \end{pmatrix} = Z^{-1} \begin{pmatrix} \langle s_h, n_1\rangle_1 \\ \vdots \\ \langle s_h, n_\ell\rangle_1 \end{pmatrix}, \tag{2.20}$$

$$\begin{pmatrix} \mu_1^{(h)} \\ \vdots \\ \mu_d^{(h)} \end{pmatrix} = W^{-1} \begin{pmatrix} \langle s_h - \sum_{i=1}^{\ell} c_i\phi_i^{(h)}, t_1\rangle_1 \\ \vdots \\ \langle s_h - \sum_{i=1}^{\ell} c_i\phi_i^{(h)}, t_d\rangle_1 \end{pmatrix}, \tag{2.21}$$

where Z and W are nonsingular $\ell \times \ell$ and $d \times d$ matrices with the elements $z_{ij} = \langle c_j, n_i\rangle_1$ and $w_{ij} = \langle df_{x_0}(r_j), t_i\rangle_1 = \langle r_j, df_{x_0}^*(t_i)\rangle_2$, respectively. The pencil $s_h \in \mathcal{M}$ has the form

$$s_h = z^{(h)} - \sum_{\substack{h'+h''=h \\ |h'|>0,\,|h''|>0}} C_h^{h'}\,\alpha\Big(\sum_{i=1}^{\ell} c_i\phi_i^{(h')}, \sum_{j=1}^{d} r_j\mu_j^{(h'')}, z^{(h')}\Big). \tag{2.22}$$

The mapping $\alpha : \mathcal{M} \times T_e\mathcal{G} \times \mathcal{M} \longrightarrow \mathcal{M}$ *is defined by the expression*

$$\alpha(x, y, z) = (AV - UX) - \lambda(BV - UY), \tag{2.23}$$

where $x = A - \lambda B$, $y = (U, V)$, *and* $z = X - \lambda Y$.

Proof. Using the notation $x = A - \lambda B$, $g = (P, Q)$, and $z = X - \lambda Y$, we can write expression (2.15) in the form

$$X(\xi) - \lambda Y(\xi) = P^{-1}(\xi)\big(A(\phi(\xi)) - \lambda B(\phi(\xi))\big)Q(\xi). \tag{2.24}$$

Multiplying (2.24) by $P(\xi)$ from left and collecting all terms at the left-hand side, we obtain

$$P(\xi)\big(X(\xi) - \lambda Y(\xi)\big) - \big(A(\phi(\xi)) - \lambda B(\phi(\xi))\big)Q(\xi) = 0. \tag{2.25}$$

Taking the derivative of order h of (2.25) and using the Leibniz formula for differentiation of a function product, we get

$$\begin{aligned} \sum_{h'+h''=h} C_h^{h'} \Big[& P^{(h'')}\big(X^{(h')} - \lambda Y^{(h')}\big) \\ & - \big((A(\phi(\xi)))^{(h')} - \lambda (B(\phi(\xi)))^{(h')}\big) Q^{(h'')} \Big] = 0. \end{aligned} \tag{2.26}$$

Using expressions (2.16), (2.17), (2.22), (2.23) in (2.26) and taking into account that $P^{(0)} = I_m$, $Q^{(0)} = I_n$, $A^{(0)} = X^{(0)} = A_0$, $B^{(0)} = Y^{(0)} = B_0$, after permutation of terms we find

$$df_{x_0}\left(\sum_{j=1}^{d} r_j \mu_j^{(h)}\right) = s_h - \sum_{i=1}^{\ell} c_i \phi_i^{(h)}, \tag{2.27}$$

where the linear mapping df_{x_0} is defined in (2.7).

Equality (2.27) represents a system of linear equations with respect to $\ell + d = 2mn$ unknowns $\phi_1^{(h)}, \dots, \phi_\ell^{(h)}$ and $\mu_1^{(h)}, \dots, \mu_d^{(h)}$. The solution of (2.27) exists if and only if its right-hand side belongs to $\operatorname{Im} df_{x_0} = T_{x_0}\mathcal{O}(x_0)$. Hence, the right-hand side has to be orthogonal to every pencil from the basis $\{n_1, \dots, n_\ell\}$ of $(T_{x_0}\mathcal{O}(x_0))^{\perp}$. This condition, written in the matrix form, yields

$$\begin{pmatrix} \langle s_h - \sum_{i=1}^{\ell} c_i \phi_i^{(h)}, n_1 \rangle_1 \\ \vdots \\ \langle s_h - \sum_{i=1}^{\ell} c_i \phi_i^{(h)}, n_\ell \rangle_1 \end{pmatrix} = \begin{pmatrix} \langle s_h, n_1 \rangle_1 \\ \vdots \\ \langle s_h, n_\ell \rangle_1 \end{pmatrix} - Z \begin{pmatrix} \phi_1^{(h)} \\ \vdots \\ \phi_\ell^{(h)} \end{pmatrix} = 0. \tag{2.28}$$

The solution of this system gives expression (2.20) of the theorem.

To determine values of the derivatives $\mu_1^{(h)}, \dots, \mu_d^{(h)}$, we take the scalar product of (2.27) and t_i. For the left-hand side this yields

$$\left\langle df_{x_0}\left(\sum_{j=1}^{d} r_j \mu_j^{(h)}\right), t_i \right\rangle_1 = \sum_{j=1}^{d} \langle df_{x_0}(r_j), t_i \rangle_1 \, \mu_j^{(h)} = \sum_{j=1}^{d} w_{ij} \mu_j^{(h)}. \tag{2.29}$$

Recall that $\langle df_{x_0}(r_j), t_i \rangle_1 = \langle r_j, df_{x_0}^*(t_i) \rangle_2$ by definition (2.8). Taking $i = 1, \dots, d$, we obtain the following system of linear equations:

$$W \begin{pmatrix} \mu_1^{(h)} \\ \vdots \\ \mu_d^{(h)} \end{pmatrix} = \begin{pmatrix} \langle s_h - \sum_{i=1}^{\ell} c_i \phi_i^{(h)}, t_1 \rangle_1 \\ \vdots \\ \langle s_h - \sum_{i=1}^{\ell} c_i \phi_i^{(h)}, t_d \rangle_1 \end{pmatrix}. \tag{2.30}$$

The solution of (2.30) gives expression (2.21) of the theorem. ∎

Note that for evaluation of derivatives $\phi_i^{(h)}$, $\mu_j^{(h)}$, expressions of Theorem 2.5 require only derivatives $\phi_i^{(h')}$, $\mu_j^{(h')}$ of lower orders $|h'| < |h|$ and derivatives $z^{(h')}$ of orders $|h'| < |h|$ and $h' = h$. This makes it possible to use Theorem 2.5 for successive calculation of the derivatives $\phi_i^{(h)}$, $\mu_j^{(h)}$ in order to find the transformation functions $\phi(\xi)$ and $g(\xi)$ in the form of Taylor series (2.19) up to small terms of arbitrary order. Recall that at the initial step of the recurrent procedure we take $\phi_i^{(0)} = 0$ and $\mu_j^{(0)} = 0$.

The matrices Z^{-1} and W^{-1} have to be computed only once in the beginning of the recurrent procedure. The size d of the matrix W is typically close to $2mn$ and can be big. Nevertheless, this matrix is usually very sparse. Moreover, we can avoid difficulties with the inversion by making the matrices Z and W diagonal. For this purpose, we need to choose the bases $\{c_1, \dots, c_\ell\}$, $\{n_1, \dots, n_\ell\}$, $\{t_1, \dots, t_d\}$, and $\{r_1, \dots, r_d\}$ such that $\langle c_j, n_i\rangle_1 = 0$ and $\langle df_{x_0}(r_j), t_i\rangle_1 = 0$ for $i \neq j$.

Note that the orthogonal miniversal deformation, represented by the orthonormal basis $\{n_1, \dots, n_\ell\}$ of $(T_{x_0}\mathcal{O}(x_0))^\perp$, keeps the metric information in the normal direction to the orbit $\mathcal{O}(x_0)$. This deformation is useful for the numerical problem of computation of a Kronecker canonical form [4]. In many applications, a metric based on properties of the underlying system is defined in the parameter space rather than in the whole space of matrix pencils. Computation on the mapping $\gamma = \phi(\xi)$ connecting the parameter spaces allows us to keep the metric information of the original parameter space and transfer this metric into the parameter space of the miniversal deformation. Theorem 2.5 can be used with an arbitrary versal deformation satisfying the requirements of each particular problem.

As noted by Arnold [1, 2], a miniversal deformation can be chosen in a simple form, which makes it convenient for applications. To avoid numerical instability in transformation to the miniversal deformation, the angle between the image of the miniversal deformation $x(\gamma)$ and the tangent space to the orbit $T_{x_0}\mathcal{O}(x_0)$ should not be small, i.e., the transversality condition of Theorem 2.3 should not be affected by numerical uncertainties and round-off.

Example 2.3. Let us consider the following two-parameter deformation $z(\xi)$, $\xi = (\xi_1, \xi_2)$, of matrix pencil (2.10):

$$z(\xi) = \begin{pmatrix} \sin\xi_1 & 1 & 0 & \xi_2^2 \\ 0 & \sin\xi_2 & \xi_2 & \xi_1 \\ \xi_1\xi_2 & \sin\xi_1 & 0 & \cos\xi_2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 & \xi_1\xi_2 \\ 0 & \sin\xi_2 & 1+\xi_2 & \xi_1 \\ \xi_2^2 & 0 & \xi_2 & \cos\xi_1 \end{pmatrix}. \tag{2.31}$$

Using the pencils $c_1, \dots, c_4$, $n_1, \dots, n_4$, $t_1, \dots, t_{20}$ and pairs $r_1, \dots, r_{20}$, constructed in Examples 2.1, 2.2, and applying Theorem 2.5, we find

$$\begin{aligned}
\phi_1(\xi) &= -\xi_1\xi_2 + \xi_2^2 + \xi_1^3 - 2\xi_1^2\xi_2 + \frac{3}{2}\xi_1\xi_2^2 - \frac{5}{2}\xi_2^3 + o(\|\xi\|^3), \\
\phi_2(\xi) &= \xi_1 - \xi_1^2 + \xi_1\xi_2 - \xi_2^2 + \frac{7}{12}\xi_1^3 + \frac{1}{2}\xi_1^2\xi_2 - \frac{1}{2}\xi_1\xi_2^2 + o(\|\xi\|^3), \\
\phi_3(\xi) &= \xi_2 - \xi_2^2 + \xi_1\xi_2^2 + \xi_2^3 + o(\|\xi\|^3), \\
\phi_4(\xi) &= \frac{1}{2}\xi_1^2 - \frac{1}{2}\xi_2^2 + \frac{1}{3}\xi_1^2\xi_2 - \frac{4}{6}\xi_1\xi_2^2 + o(\|\xi\|^3), \\
\mu_1(\xi) &= \xi_1 - \frac{1}{6}\xi_1^3 + o(\|\xi\|^3), \\
\mu_2(\xi) &= -\frac{2}{9}\xi_1^2\xi_2 - \frac{2}{9}\xi_1\xi_2^2 + o(\|\xi\|^3),
\end{aligned}$$

$$
\begin{aligned}
\mu_3(\xi) &= o(\|\xi\|^3),\\
\mu_4(\xi) &= -\frac{1}{3}\xi_1\xi_2 + \frac{2}{3}\xi_2^2 + o(\|\xi\|^3),\\
\mu_5(\xi) &= 2\xi_2 - 2\xi_1^2 - 3\xi_2^2 + \xi_1^2\xi_2 + 2\xi_1\xi_2^2 + \frac{19}{6}\xi_2^3 + o(\|\xi\|^3),\\
\mu_6(\xi) &= \xi_1 - \frac{1}{2}\xi_1\xi_2 - \frac{1}{2}\xi_1^3 + \frac{1}{2}\xi_1^2\xi_2 + \frac{9}{4}\xi_1\xi_2^2 - 2\xi_2^3 + o(\|\xi\|^3),\\
\mu_7(\xi) &= -\frac{1}{2}\xi_1 + \frac{1}{2}\xi_2^2 - \frac{1}{24}\xi_1^3 - \frac{1}{4}\xi_1^2\xi_2 + o(\|\xi\|^3),\\
\mu_8(\xi) &= -\frac{1}{2}\xi_2^2 + o(\|\xi\|^3),\\
\mu_9(\xi) &= -\frac{3}{4}\xi_1 - \frac{1}{16}\xi_1^3 - \frac{3}{8}\xi_1^2\xi_2 + o(\|\xi\|^3),\\
\mu_{10}(\xi) &= -\xi_2 - \xi_2^2 - \frac{1}{4}\xi_1^2\xi_2 - \frac{1}{2}\xi_1\xi_2^2 + o(\|\xi\|^3),\\
\mu_{11}(\xi) &= -\frac{1}{8}\xi_1^2 - \frac{1}{4}\xi_1\xi_2 - \frac{1}{4}\xi_1^2\xi_2 + o(\|\xi\|^3),\\
\mu_{12}(\xi) &= \frac{1}{9}\xi_1^2\xi_2 + \frac{1}{9}\xi_1\xi_2^2 + o(\|\xi\|^3),\\
\mu_{13}(\xi) &= o(\|\xi\|^3),\\
\mu_{14}(\xi) &= -\frac{1}{6}\xi_1\xi_2^2 - \frac{1}{6}\xi_2^3 + o(\|\xi\|^3),\\
\mu_{15}(\xi) &= \frac{2}{3}\xi_1\xi_2 - \frac{1}{3}\xi_2^2 + o(\|\xi\|^3),\\
\mu_{16}(\xi) &= -\xi_2 + \xi_1^2 + \frac{3}{2}\xi_2^2 - \frac{1}{2}\xi_1^2\xi_2 - \frac{3}{2}\xi_1\xi_2^2 - \frac{19}{12}\xi_2^3 + o(\|\xi\|^3),\\
\mu_{17}(\xi) &= \xi_2 - \frac{1}{2}\xi_2^2 + \frac{1}{2}\xi_1\xi_2^2 + \frac{1}{12}\xi_2^3 + o(\|\xi\|^3),\\
\mu_{18}(\xi) &= \frac{1}{2}\xi_2 - \frac{1}{2}\xi_1\xi_2 - \frac{1}{4}\xi_2^2 + \frac{1}{2}\xi_1\xi_2^2 + \frac{1}{8}\xi_2^3 + o(\|\xi\|^3),\\
\mu_{19}(\xi) &= \frac{1}{2}\xi_1^3 - \frac{3}{2}\xi_1\xi_2^2 + \xi_2^3 + o(\|\xi\|^3),\\
\mu_{20}(\xi) &= \xi_2 + \frac{1}{4}\xi_1^2\xi_2 + \frac{1}{2}\xi_1\xi_2^2 + o(\|\xi\|^3).
\end{aligned}
$$

These expressions determine the change of parameters $\gamma = \phi(\xi)$ and equivalence transformation $g(\xi)$ in the reduction of $z(\xi)$ to the miniversal deformation (2.18).

3. Pairs of matrices under the feedback equivalence. In this section we consider the space of pairs of matrices

$$\widetilde{\mathcal{M}} = \{(F,G) \mid F \in M_{m\times m}(\mathbb{F}),\ \ G \in M_{m\times n}(\mathbb{F})\}. \tag{3.1}$$

Each pair $x = (F,G) \in \widetilde{M}$ represents the time-invariant linear dynamical system $\dot\psi = F\psi + G\nu$, $\psi \in \mathbb{F}^m$, with the input vector $\nu \in \mathbb{F}^n$. The change of basis in the state and input spaces and feedback operation in this system induce an equivalence relation in the space $\widetilde{\mathcal{M}}$ as follows: two pairs of matrices $x_1 = (F_1, G_1)$ and $x_2 = (F_2, G_2)$ are called *feedback equivalent* if and only if there exist matrices $P \in \mathrm{Gl}(m;\mathbb{F})$, $R \in \mathrm{Gl}(n;\mathbb{F})$, and $S \in M_{n\times m}(\mathbb{F})$ such that [16]

$$F_2 = P^{-1}(F_1P + G_1S), \quad G_2 = P^{-1}G_1R. \tag{3.2}$$

The feedback equivalence transformation may be seen as the action of the Lie group

$$\widetilde{\mathcal{G}} = \{g = (P, R, S) \mid P \in \mathrm{Gl}(m; \mathbb{F}),\ R \in \mathrm{Gl}(n; \mathbb{F}),\ S \in M_{n\times m}(\mathbb{F})\} \tag{3.3}$$

with the multiplication of elements $g_1, g_2 \in \widetilde{\mathcal{G}}$ determined by the expression

$$g_1 g_2 = (P_1P_2, R_1R_2, S_1P_2 + R_1S_2), \quad g_i = (P_i, R_i, S_i). \tag{3.4}$$

The unit element of the group $\widetilde{\mathcal{G}}$ is $e = (I_m, I_n, 0)$. We will use the short notation $x_2 = g \circ x_1$ for the equivalence relation (3.2). Note that $g_1g_2 \circ x = g_2 \circ (g_1 \circ x)$.

Given a pair of matrices $x = (F, G) \in \widetilde{\mathcal{M}}$ and a triple $g = (P, R, S) \in \widetilde{\mathcal{G}}$, we can associate a matrix pencil $x' \in \mathcal{M}$ of dimension $m \times (m + n)$ and a pair g' from the corresponding Lie group $\mathcal{G}$ in the following manner:

$$x' = (F \ \ G) - \lambda(I_m \ \ 0), \quad g' = \left(P, \begin{pmatrix} P & 0 \\ S & R \end{pmatrix}\right). \tag{3.5}$$

It is easy to see that $x_2 = g \circ x_1$ (the pairs x_1 and x_2 are feedback equivalent) if and only if $x'_2 = g' \circ x'_1$ (the associated matrix pencils x'_1 and x'_2 are strict equivalent) [11]. Hence, $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{G}}$ can be seen as the subspace of $\mathcal{M}$ and subgroup of $\mathcal{G}$, respectively. Note that the subspace $\widetilde{\mathcal{M}} \subset \mathcal{M}$ is not invariant under the action of the Lie group $\mathcal{G}$ defined over the space of matrix pencils.

3.1. Orbit and stabilizer. Let us fix some pair of matrices $x_0 = (F_0, G_0)$ and define the mapping $\widetilde{f}_{x_0}(g) = g \circ x_0$, $g \in \widetilde{\mathcal{G}}$. Then the orbit $\widetilde{\mathcal{O}}(x_0)$ and stabilizer $\widetilde{\mathcal{S}}(x_0)$ of the pair x_0 are defined as follows:

$$\widetilde{\mathcal{O}}(x_0) = \operatorname{Im} \widetilde{f}_{x_0} = \{g \circ x_0 \mid g \in \widetilde{\mathcal{G}}\}, \tag{3.6}$$

$$\widetilde{\mathcal{S}}(x_0) = \operatorname{Ker} (\widetilde{f}_{x_0} - x_0) = \{g \in \widetilde{\mathcal{G}} \mid g \circ x_0 = x_0\}. \tag{3.7}$$

The sets $\widetilde{\mathcal{O}}(x_0)$ and $\widetilde{\mathcal{S}}(x_0)$ are differentiable submanifolds of $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{G}}$, respectively. Note that under relations (3.5) we have $\widetilde{\mathcal{O}}(x_0) \subset \mathcal{O}(x_0)$ and $\widetilde{\mathcal{S}}(x_0) \subset \mathcal{S}(x_0)$.

Since $\widetilde{\mathcal{G}}$ is an open subset of $M_{m\times m}(\mathbb{F}) \times M_{n\times n}(\mathbb{F}) \times M_{n\times m}(\mathbb{F})$, the tangent space $T_e\widetilde{\mathcal{G}}$ to the manifold $\widetilde{\mathcal{G}}$ at the unit element e is

$$T_e\widetilde{\mathcal{G}} = \{(U, V, W) \mid U \in M_{m\times m}(\mathbb{F}),\ V \in M_{n\times n}(\mathbb{F}),\ W \in M_{n\times m}(\mathbb{F})\}. \tag{3.8}$$

Since $\widetilde{\mathcal{M}}$ is a linear space, $T_{x_0}\widetilde{\mathcal{M}} = \widetilde{\mathcal{M}}$. We consider Euclidean scalar products in $\widetilde{\mathcal{M}}$ and $T_e\widetilde{\mathcal{G}}$ having the form

$$\begin{aligned} &\langle x_1, x_2\rangle_1 = \operatorname{trace}(F_1F_2^*) + \operatorname{trace}(G_1G_2^*), \\ &\langle y_1, y_2\rangle_2 = \operatorname{trace}(U_1U_2^*) + \operatorname{trace}(V_1V_2^*) + \operatorname{trace}(W_1W_2^*), \end{aligned} \tag{3.9}$$

where $x_i = (F_i, G_i) \in \widetilde{\mathcal{M}}$, $y_i = (U_i, V_i, W_i) \in T_e\widetilde{\mathcal{G}}$, $i = 1, 2$.

Let $d\widetilde{f}_{x_0} : T_e\widetilde{\mathcal{G}} \longrightarrow \widetilde{\mathcal{M}}$ be the differential of $\widetilde{f}_{x_0}$ at the unit element e. Using (3.2), it can be shown [6] that

$$d\widetilde{f}_{x_0}(y) = (F_0U - UF_0 + G_0W,\ G_0V - UG_0) \in \widetilde{\mathcal{M}}, \tag{3.10}$$

where $y = (U, V, W) \in T_e\widetilde{\mathcal{G}}$. The adjoint linear mapping $d\widetilde{f}^*_{x_0} : \widetilde{\mathcal{M}} \longrightarrow T_e\widetilde{\mathcal{G}}$ is determined by the relation

$$d\widetilde{f}^*_{x_0}(z) = (F_0^*X - XF_0^* - YG_0^*,\ G_0^*Y,\ G_0^*X) \in T_e\widetilde{\mathcal{G}}, \tag{3.11}$$

where $z = (X, Y) \in \widetilde{\mathcal{M}}$.

Analogously to Theorem 2.1, the mappings $d\widetilde{f}_{x_0}$ and $d\widetilde{f}^*_{x_0}$ provide the following description for the tangent spaces $T_{x_0}\widetilde{\mathcal{O}}(x_0)$, $T_e\widetilde{\mathcal{S}}(x_0)$ and their normal complements.

THEOREM 3.1. *The tangent spaces to the orbit and stabilizer of the pair of matrices x_0 and corresponding normal complementary subspaces can be found in the following form:*

1. $T_{x_0}\widetilde{\mathcal{O}}(x_0) = \operatorname{Im} d\widetilde{f}_{x_0} \subset \widetilde{\mathcal{M}}$.
2. $(T_{x_0}\widetilde{\mathcal{O}}(x_0))^\perp = \operatorname{Ker} d\widetilde{f}^*_{x_0} \subset \widetilde{\mathcal{M}}$.
3. $T_e\widetilde{\mathcal{S}}(x_0) = \operatorname{Ker} d\widetilde{f}_{x_0} \subset T_e\widetilde{\mathcal{G}}$.
4. $(T_e\widetilde{\mathcal{S}}(x_0))^\perp = \operatorname{Im} d\widetilde{f}^*_{x_0} \subset T_e\widetilde{\mathcal{G}}$.

Example 3.1. Let $x_0 = (F_0, G_0)$ be a pair of matrices with

$$F_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad G_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{3.12}$$

Then elements $z = (X, Y)$ of the space $(T_{x_0}\widetilde{\mathcal{O}}(x_0))^\perp$ can be found from the equation $d\widetilde{f}^*_{x_0}(z) = 0$ in the form

$$\left(\begin{pmatrix} 0 & 0 & 0 \\ \gamma_1 & \gamma_2 & 0 \\ \gamma_3 & 0 & \gamma_4 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \gamma_3 \end{pmatrix} \right), \tag{3.13}$$

where $\gamma_1, \ldots, \gamma_4 \in \mathbb{F}$ are arbitrary, and $\dim(T_{x_0}\widetilde{\mathcal{O}}(x_0))^\perp = 4$. The elements of $T_{x_0}\widetilde{\mathcal{O}}(x_0)$ have the form

$$\left(\begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ 0 & 0 & \mu_4 \\ \mu_5 & \mu_6 & 0 \end{pmatrix}, \begin{pmatrix} \mu_7 \\ \mu_8 \\ -\mu_5 \end{pmatrix} \right), \tag{3.14}$$

where $\mu_1, \ldots, \mu_8 \in \mathbb{F}$ are arbitrary and $\dim T_{x_0}\widetilde{\mathcal{O}}(x_0) = 8$. Then, by Theorem 3.1, $\dim(T_e\widetilde{\mathcal{S}}(x_0))^\perp = 8$ and elements $y = (U, V, W)$ of $(T_e\widetilde{\mathcal{S}}(x_0))^\perp = d\widetilde{f}^*_{x_0}(T_{x_0}\widetilde{\mathcal{O}}(x_0))$ take the form

$$\left(\begin{pmatrix} -\mu_7 & 0 & -\mu_3 \\ -\mu_8 & 0 & -\mu_4 \\ 2\mu_5 & \mu_6 & 0 \end{pmatrix}, (\mu_7), (\mu_1,\ \mu_2,\ \mu_3) \right). \tag{3.15}$$

Finally, the space $T_e\widetilde{\mathcal{S}}(x_0) = \operatorname{Ker} d\widetilde{f}_{x_0}$ is formed by the triples

$$\left(\begin{pmatrix} \nu_1 & \nu_2 & \nu_3 \\ 0 & \nu_4 & 0 \\ 0 & 0 & \nu_5 \end{pmatrix}, (\nu_1), (0,\ 0,\ \nu_3) \right), \tag{3.16}$$

where $\nu_1, \ldots, \nu_5 \in \mathbb{F}$ are arbitrary and $\dim T_e\widetilde{\mathcal{S}}(x_0) = 5$.

Note that under relation (3.5), the matrix pencil corresponding to pair (3.12) is equivalent to matrix pencil (2.10) considered in Example 2.1. Dimensions of the tangent space to the stabilizer and normal complement of the tangent space to the orbit are the same for the cases of matrix pairs and matrix pencils. But dimensions of the tangent space to the orbit and the normal complement of the tangent space to the stabilizer are smaller in the case of matrix pairs.

3.2. Versal deformation. Let us consider a deformation $x(\gamma)$ of $x_0 \in \widetilde{\mathcal{M}}$ in the form

$$x(\gamma) = x_0 + \sum_{i=1}^{\ell} c_i \gamma_i, \tag{3.17}$$

where $\{c_1, \dots, c_\ell\}$ is a basis of an arbitrary complementary subspace $\left(T_{x_0}\widetilde{\mathcal{O}}(x_0)\right)^c$ to $T_{x_0}\widetilde{\mathcal{O}}(x_0)$; $\ell = \operatorname{codim} T_{x_0}\widetilde{\mathcal{O}}(x_0)$.

Analogously to Corollary 2.4, we have the following.

COROLLARY 3.2. *The deformation* (3.17) *is a miniversal deformation; i.e., any deformation $z(\xi)$, $\xi \in \mathbb{F}^k$, of x_0 can be represented in the neighborhood of the origin $\mathcal{U}_0 \subset \mathbb{F}^k$ in the form*

$$z(\xi) = g(\xi) \circ x(\phi(\xi)), \tag{3.18}$$

where $\phi : \mathcal{U}_0 \longrightarrow \mathbb{F}^\ell$ and $g : \mathcal{U}_0 \longrightarrow \widetilde{\mathcal{G}}$ are smooth mappings such that $\phi(0) = 0$ and $g(0) = e$. The functions $\phi(\xi)$ and $g(\xi)$ are uniquely determined by the deformation $z(\xi)$ if $g(\xi)$ is taken in the form

$$g(\xi) = e + \sum_{j=1}^{d} r_j \mu_j(\xi), \tag{3.19}$$

where $\mu_j(\xi)$ are smooth functions in $\mathbb{F}$ such that $\mu_j(0) = 0$, $j = 1, \dots, d$, and $\{r_1, \dots, r_d\}$ is a basis of $(T_e\widetilde{\mathcal{S}}(x_0))^\perp$.

Recall that if $\{t_1, \dots, t_d\}$ is a basis of $T_{x_0}\widetilde{\mathcal{O}}(x_0)$, then $\{d\widetilde{f}^*_{x_0}(t_1), \dots, d\widetilde{f}^*_{x_0}(t_d)\}$ is a basis of $(T_e\widetilde{\mathcal{S}}(x_0))^\perp$, and, vice versa, if $\{r_1, \dots, r_d\}$ is a basis of $(T_e\widetilde{\mathcal{S}}(x_0))^\perp$, then $\{d\widetilde{f}_{x_0}(r_1), \dots, d\widetilde{f}_{x_0}(r_d)\}$ is a basis of $T_{x_0}\widetilde{\mathcal{O}}(x_0)$.

For pairs of matrices, reduced to the Brunovsky canonical form, explicit expressions for the bases $\{c_1, \dots, c_\ell\}$ and $\{n_1, \dots, n_\ell\}$ may be found in [6].

Example 3.2. Let $x_0 = (F_0, G_0)$ be the pair of matrices considered in Example 3.1. Using explicit form of the tangent space $T_{x_0}\widetilde{\mathcal{O}}(x_0)$ given in (3.14), we can choose a basis $\{c_1, \dots, c_4\}$ of the complementary space $(T_{x_0}\widetilde{\mathcal{O}}(x_0))^c$ such that every c_i has exactly one nonzero element. For example, we can choose the miniversal deformation in the form

$$x(\gamma) = \left(\begin{pmatrix} 0 & 0 & 0 \\ \gamma_1 & \gamma_2 & 0 \\ \gamma_3 & 0 & 1+\gamma_4 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right), \quad \gamma = (\gamma_1, \dots, \gamma_4). \tag{3.20}$$

3.3. Reduction to miniversal deformation. Let x_0 and $x(\gamma)$ be a pair of matrices and its miniversal deformation. In order to reduce a given deformation $z(\xi)$

of x_0 to the miniversal deformation, we need to find the smooth mappings $\phi(\xi)$ and $g(\xi)$ satisfying (3.18). These mappings can be found in Taylor series form:

$$\phi(\xi) = \sum_{|h|\le s} \frac{\phi^{(h)}}{h!}\xi^h + o(\|\xi\|^s),$$

$$g(\xi) = e + \sum_{j=1}^{d} r_j \sum_{|h|\le s} \frac{\mu_j^{(h)}}{h!}\xi^h + o(\|\xi\|^s), \tag{3.21}$$

where $\phi^{(0)} = 0$ and $\mu_j^{(0)} = 0$.

Analogously to Theorem 2.5, we can find explicit recurrent formulae for calculation of the derivatives $\phi^{(h)}$ and $\mu_j^{(h)}$ up to an arbitrary order.

THEOREM 3.3. *The derivatives $\phi_1^{(h)}, \dots, \phi_\ell^{(h)}$ and $\mu_1^{(h)}, \dots, \mu_d^{(h)}$ determining transformation functions* (3.21)*, which reduce the deformation $z(\xi)$ of x_0 to the miniversal deformation* (3.17)*, satisfy the recurrent formulae*

$$\begin{pmatrix} \phi_1^{(h)} \\ \vdots \\ \phi_\ell^{(h)} \end{pmatrix} = Z^{-1} \begin{pmatrix} \langle s_h, n_1\rangle_1 \\ \vdots \\ \langle s_h, n_\ell\rangle_1 \end{pmatrix}, \tag{3.22}$$

$$\begin{pmatrix} \mu_1^{(h)} \\ \vdots \\ \mu_d^{(h)} \end{pmatrix} = W^{-1} \begin{pmatrix} \langle s_h - \sum_{i=1}^{\ell} c_i\phi_i^{(h)}, t_1\rangle_1 \\ \vdots \\ \langle s_h - \sum_{i=1}^{\ell} c_i\phi_i^{(h)}, t_d\rangle_1 \end{pmatrix}, \tag{3.23}$$

where Z and W are nonsingular $\ell \times \ell$ and $d \times d$ matrices with the elements $z_{ij} = \langle c_j, n_i\rangle_1$, $w_{ij} = \langle d\widetilde{f}_{x_0}(r_j), t_i\rangle_1$, respectively. The pair of matrices $s_h \in \widetilde{\mathcal{M}}$ has the form

$$s_h = z^{(h)} - \sum_{\substack{h'+h''=h \\ |h'|>0,\,|h''|>0}} C_h^{h'}\,\widetilde{\alpha}\Big(\sum_{i=1}^{\ell} c_i\phi_i^{(h')}, \sum_{j=1}^{d} r_j\mu_j^{(h'')}, z^{(h')}\Big). \tag{3.24}$$

The mapping $\widetilde{\alpha} : \widetilde{\mathcal{M}} \times T_e\widetilde{\mathcal{G}} \times \widetilde{\mathcal{M}} \longrightarrow \widetilde{\mathcal{M}}$ is defined as follows:

$$\widetilde{\alpha}(x, y, z) = (FU - UX + GW, GV - UY), \tag{3.25}$$

where $x = (F, G)$, $y = (U, V, W)$, and $z = (X, Y)$.

Analogously to the case of matrix pencils, in order to simplify the calculations we can choose the bases $\{c_1, \dots, c_\ell\}$, $\{n_1, \dots, n_\ell\}$, $\{t_1, \dots, t_d\}$, and $\{r_1, \dots, r_d\}$ in such a way that $\langle c_j, n_i\rangle_1 = 0$ and $\langle d\widetilde{f}_{x_0}(r_j), t_i\rangle_1 = 0$ for $i \ne j$, which implies that Z and W are diagonal matrices.

Example 3.3. Let us consider the following two-parameter deformation $z(\xi)$, $\xi = (\xi_1, \xi_2)$, of the pair of matrices $x_0 = (F_0, G_0)$ considered in Example 3.1:

$$z(\xi) = \left(\begin{pmatrix} \xi_1 & \xi_1\xi_2 & \xi_2^3/6 \\ \xi_2 & \xi_1 & \xi_1 + \xi_2 \\ \xi_1^2\xi_2 & \xi_1\xi_2 & 1 \end{pmatrix}, \begin{pmatrix} 1 + \xi_1\xi_2 \\ \xi_1^2 \\ \xi_2^3 \end{pmatrix} \right). \tag{3.26}$$

Using the bases $\{c_1,\dots,c_4\}$, $\{n_1,\dots,n_4\}$, $\{t_1,\dots,t_8\}$, and $\{r_1,\dots,r_8\}$ constructed in Examples 3.1, 3.2 and applying Theorem 3.3, we find

$$
\begin{aligned}
&\phi_1(\xi)=\xi_2+\xi_1\xi_2^2/2+o(\|\xi\|^3),\quad \phi_2(\xi)=\xi_1-\xi_1^2\xi_2-\xi_1\xi_2^2+o(\|\xi\|^3),\\
&\phi_3(\xi)=\xi_1^2\xi_2+\xi_1\xi_2^2+\xi_2^3+o(\|\xi\|^3),\quad \phi_4(\xi)=\xi_1^2\xi_2+\xi_1\xi_2^2+o(\|\xi\|^3),\\
&\mu_1(\xi)=\xi_1-\xi_1^2\xi_2/2+o(\|\xi\|^3),\quad \mu_2(\xi)=\xi_1\xi_2+o(\|\xi\|^3),\\
&\mu_3(\xi)=\xi_2^3/12+o(\|\xi\|^3),\\
&\mu_4(\xi)=\xi_1+\xi_2+\xi_1^2+\xi_1\xi_2+\xi_1^3+\xi_1^2\xi_2+o(\|\xi\|^3),\\
&\mu_5(\xi)=-\xi_2^3/2+o(\|\xi\|^3),\quad \mu_6(\xi)=\xi_1\xi_2+\xi_1^2\xi_2+o(\|\xi\|^3),\\
&\mu_7(\xi)=\xi_1\xi_2/2+o(\|\xi\|^3),\quad \mu_8(\xi)=\xi_1^2+o(\|\xi\|^3).
\end{aligned}
\tag{3.27}
$$

Expressions (3.27) determine the change of parameters $\gamma=\phi(\xi)$ and the equivalence transformation $g(\xi)$ given by (3.19) in the reduction of $z(\xi)$ to the miniversal deformation (3.20).

4. Local analysis of the uncontrollability set for one-input systems. Let us consider a pair of real matrices $z=(F,G)\in\widetilde{\mathcal{M}}$ with $n=1$ and arbitrary m. This pair corresponds to the system of differential equations

$$\dot{\psi}(t)=F\psi(t)+G\nu(t) \tag{4.1}$$

with m-dimensional state vector $\psi\in\mathbb{R}^m$ and one input variable $\nu\in\mathbb{R}$. System (4.1) is called *controllable* if it is possible to construct a control signal $\nu(t)$ that will transfer an initial state to any final state in finite time [16]. The pair $z=(F,G)$ corresponding to such a system is called controllable. The well-known criterion for controllability says that the pair z is controllable if and only if the *controllability matrix* $C=[G,FG,\dots,F^{m-1}G]$ has full rank [16]

$$\operatorname{rank}[G,FG,\dots,F^{m-1}G]=m. \tag{4.2}$$

For one-input systems, i.e., when the matrix G has dimension $m\times 1$, this criterion takes the form

$$\det[G,FG,\dots,F^{m-1}G]\neq 0. \tag{4.3}$$

Let us consider a family of matrix pairs $z(\xi)=(F(\xi),G(\xi))$ with the parameter vector $\xi\in\mathbb{R}^k$. The set of values of the parameter vector ξ such that the pair $z(\xi)$ is uncontrollable is called the *uncontrollability set* and will be denoted by $\mathcal{N}=\{\xi\in\mathbb{R}^k \mid \operatorname{rank}C(\xi)<m\}$. Let us assume that the pair $z(\xi)$ is uncontrollable at some point $\xi_0\in\mathcal{N}$. We are going to analyze the structure of the uncontrollability set in the neighborhood of this point. Due to the complicated entry of elements of the matrices F and G into the controllability matrix, it is difficult to use the controllability condition (4.3) for analytical analysis of the set $\mathcal{N}$. Using reduction of the family $z(\xi)$ to the miniversal deformation, this analysis can be carried out in a more simple and systematic way, as shown below.

The matrix pair $z_0=z(\xi_0)$ can be reduced to the Brunovsky canonical form $\widehat{z}_0=g_0\circ z_0$ by the state feedback transformation $g_0\in\widetilde{\mathcal{G}}$ [11, 16]. Let us consider the

case when the Brunovsky form $\widehat{z}_0$ is as follows:

$$\widehat{z}_0 = \left(\begin{pmatrix} 0 & 1 & & & \\ & 0 & \ddots & & \\ & & \ddots & 1 & \\ & & & 0 & \\ & & & & \sigma_0 \end{pmatrix}, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{pmatrix} \right), \tag{4.4}$$

where $\sigma_0 \in \mathbb{R}$ is an arbitrary number called the *uncontrollable mode* or the *generalized eigenvalue*. In the generic (typical) case, the parameter vectors ξ, corresponding to the matrix pairs $z(\xi)$ having Brunovsky form (4.4), represent typical elements of the uncontrollability set $\mathcal{N}$ and form a codimension 1 smooth submanifold of $\mathbb{R}^k$. Uncontrollable matrix pairs having different Brunovsky structures form submanifolds of higher codimensions [6]. The following proposition gives explicit formulae for the tangent plane to the uncontrollability set $\mathcal{N}$ at ξ_0 and the first approximation of the uncontrollable mode.

PROPOSITION 4.1. *Let $z_0 = z(\xi_0)$, $\xi_0 \in \mathcal{N}$, be a matrix pair having Brunovsky canonical form* (4.4) *with the triple $g_0 = (P_0, R_0, S_0) \in \widetilde{\mathcal{G}}$ providing the feedback equivalence transformation $\widehat{z}_0 = g_0 \circ z_0$. Let us define real vectors $\eta = (\eta_1, \dots, \eta_k)$ and $\eta_\sigma = (\eta_{\sigma 1}, \dots, \eta_{\sigma k})$ with the components*

$$\begin{aligned} \eta_i &= P_0^{-1}(m,:)\left[\frac{\partial F}{\partial \xi_i} \sum_{j=1}^{m-1} \sigma_0^{j-1} P_0(:,j) + \frac{\partial G}{\partial \xi_i}\left(\sum_{j=1}^{m-1} \sigma_0^{j-1} S_0(:,j) + \sigma_0^{m-1} R_0 \right)\right], \\ \eta_{\sigma i} &= P_0^{-1}(m,:)\left(\frac{\partial F}{\partial \xi_i} P_0(:,m) + \frac{\partial G}{\partial \xi_i} S_0(:,m)\right), \quad i = 1, \dots, k, \end{aligned} \tag{4.5}$$

where $P_0^{-1}(m,:)$, $P_0(:,j)$, and $S_0(:,j)$ denote the mth row of P_0^{-1}, the jth column of P_0, and the jth column of S_0, respectively. Then, if $\eta \neq 0$, the uncontrollability set $\mathcal{N}$ is a smooth hypersurface in the vicinity of ξ_0; the vector η is the normal vector to this hypersurface at ξ_0; the tangent plane to $\mathcal{N}$ at ξ_0 is given by the equation

$$(\eta, \xi - \xi_0) = 0, \tag{4.6}$$

where $(\eta, \xi) = \sum_{i=1}^k \eta_i \xi_i$ is a scalar product in $\mathbb{R}^k$; and the first order approximation of the uncontrollable mode on the hypersurface $\mathcal{N}$ is given by the relation

$$\sigma(\xi) = \sigma_0 + (\eta_\sigma, \xi - \xi_0) + o(\|\xi - \xi_0\|). \tag{4.7}$$

Proof. Without loss of generality, we can take $\xi_0 = 0$. Let us consider the family $\widehat{z}(\xi) = g_0 \circ z(\xi)$, which is a deformation of the matrix $\widehat{z}_0 = g_0 \circ z_0$ given by (4.4). The deformation $\widehat{z}(\xi)$ can be reduced to the orthogonal miniversal deformation of $\widehat{z}_0$ having the form [6]

$$x(\gamma) = \left(\begin{pmatrix} 0 & 1 & & & \\ & 0 & \ddots & & \\ & & \ddots & 1 & \\ & & & 0 & \\ \gamma_1 & \sigma_0\gamma_1 & \cdots & \sigma_0^{m-2}\gamma_1 & \sigma_0 + \gamma_2 \end{pmatrix}, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \sigma_0^{m-1}\gamma_1 \end{pmatrix} \right), \tag{4.8}$$

where $\gamma = (\gamma_1, \gamma_2)$. Since the controllability property is invariant under the feedback group action [16], the controllability of the pair $z(\xi)$ is equivalent to the controllability of the pair $x(\phi(\xi))$, where $\gamma = \phi(\xi)$ represents the change of parameters in the reduction of $\widehat{z}(\xi)$ to the miniversal deformation $x(\gamma)$. Applying the criterion of controllability (4.2) to matrix pair (4.8), we conclude that for small γ the pair $x(\gamma)$ is controllable if and only if $\gamma_1 \neq 0$. Hence, the uncontrollability set in the vicinity of ξ_0 is determined by the equation $\gamma_1 = \phi_1(\xi) = 0$. If $\gamma_1 = 0$, then we find the uncontrollable mode $\sigma = \sigma_0 + \gamma_2 = \sigma_0 + \phi_2(\xi)$.

Using formula (3.22) of Theorem 3.3 and taking into account that the matrix Z is diagonal, we find

$$\frac{\partial \phi_1}{\partial \xi_i} = z_{11}^{-1}\Big\langle \frac{\partial \widehat{z}}{\partial \xi_i}, n_1 \Big\rangle_1 = z_{11}^{-1}\Big(\sum_{j=1}^{m-1} \sigma_0^{j-1} \frac{\partial \widehat{F}_{mj}}{\partial \xi_i} + \sigma_0^{m-1} \frac{\partial \widehat{G}_{m1}}{\partial \xi_i} \Big), \tag{4.9}$$

$$z_{11} = \langle n_1, n_1 \rangle_1 = 1 + \sigma_0^2 + \cdots + \sigma_0^{2m-2},$$

where derivatives are taken at ξ_0, the pair n_1 was found from the orthogonal miniversal deformation (4.8) as a coefficient corresponding to γ_1, and $\widehat{F}_{mj}$, $\widehat{G}_{m1}$ denote the (m,j)th and $(m,1)$th elements of the matrices $(\widehat{F}, \widehat{G}) = \widehat{z}$. Using expression $\widehat{z}(\xi) = g_0 \circ z(\xi)$, we obtain

$$\widehat{F}(\xi) = P_0^{-1}(F(\xi)P_0 + G(\xi)S_0), \quad \widehat{G}(\xi) = P_0^{-1}G(\xi)R_0. \tag{4.10}$$

Substitution of (4.10) into (4.9) yields

$$\begin{aligned} \frac{\partial \phi_1}{\partial \xi_i} = z_{11}^{-1} P_0^{-1}(m,:) \Big[& \frac{\partial F}{\partial \xi_i} \sum_{j=1}^{m-1} \sigma_0^{j-1} P_0(:,j) \\ & + \frac{\partial G}{\partial \xi_i} \Big(\sum_{j=1}^{m-1} \sigma_0^{j-1} S_0(:,j) + \sigma_0^{m-1} R_0 \Big) \Big]. \end{aligned} \tag{4.11}$$

Hence, using the notation of (4.5), we find the gradient vector of the function $\phi_1(\xi)$ at ξ_0 in the form

$$\nabla \phi_1 = \left(\frac{\partial \phi_1}{\partial \xi_1}, \ldots, \frac{\partial \phi_1}{\partial \xi_k} \right) = z_{11}^{-1} \eta. \tag{4.12}$$

If $\eta \neq 0$, then $\nabla \phi_1 \neq 0$ and, by the implicit function theorem applied to the equation $\phi_1(\xi) = 0$, we conclude that the uncontrollability set is a smooth hypersurface in the vicinity of ξ_0 with the tangent plane (4.6). The vector η is normal to this surface at ξ_0.

Analogously, we find

$$\begin{aligned} \frac{\partial \phi_2}{\partial \xi_i} &= z_{22}^{-1}\Big\langle \frac{\partial \widehat{z}}{\partial \xi_i}, n_2 \Big\rangle_1 = \frac{\partial \widehat{F}_{mm}}{\partial \xi_i} \\ &= P_0^{-1}(m,:)\Big(\frac{\partial F}{\partial \xi_i} P_0(:,m) + \frac{\partial G}{\partial \xi_i} S_0(:,m) \Big). \end{aligned} \tag{4.13}$$

Hence, using the notation of (4.5), we find the gradient $\nabla \phi_2 = \eta_\sigma$ at ξ_0, which gives approximation (4.7) for the uncontrollable mode $\sigma(\xi) = \sigma_0 + \phi_2(\xi)$. □

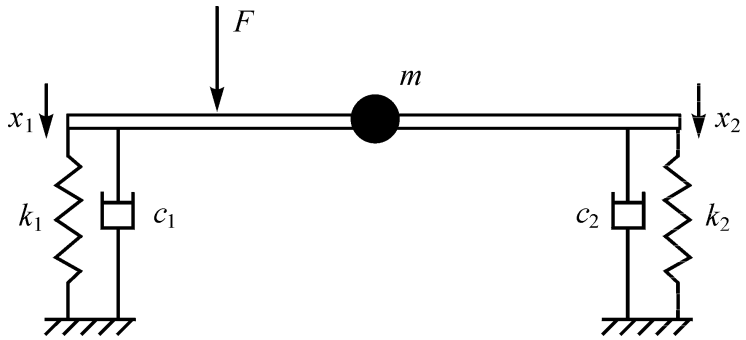


FIG. 2. *Elastic system controlled by a force F.*

Note that Proposition 4.1 provides quantitative local information on the uncontrollability set using only information on the matrix pair $z_0 = z(\xi_0)$ and derivatives of the system matrices $F(\xi)$ and $G(\xi)$ evaluated at the point ξ_0. Using this information we can choose an optimal change of parameters in order to obtain a good-controllable system. Formula for the tangent plane is useful for numerical computation of the uncontrollability set.

A multi-input system is characterized by a vector of real input variables $\nu(t)$ in (4.1). In this case uncontrollable pairs have different Brunovsky forms, and corresponding miniversal deformations are more complicated. The suggested approach can be extended to analysis of the uncontrollability set for a multi-input dynamical system depending on parameters. For this purpose, we need to find the uncontrollability set for that particular versal deformation, and then transfer the result to the original parameter space by means of the mapping $\gamma = \phi(\xi)$ found by Theorem 2.5.

Example 4.1. Let us consider the mechanical system shown in Figure 2. The system consists of a light platform of length L carrying a point mass m in the middle; both ends of the platform are supported on the ground by means of springs with elastic coefficients k_1, k_2 and damping coefficients c_1, c_2. The system is controlled by a force F applied to the platform at the distance $\xi_1 L$ from the left end. We assume that the equilibrium of this system for $F = 0$ corresponds to the horizontal position of the platform. Equations of motion of the system have the form

$$\begin{aligned} &m(\ddot{x}_1 + \ddot{x}_2)/4 + c_1\dot{x}_1 + k_1 x_1 = (1-\xi_1)F, \\ &m(\ddot{x}_1 + \ddot{x}_2)/4 + c_2\dot{x}_2 + k_2 x_2 = \xi_1 F, \end{aligned} \tag{4.14}$$

where x_1 and x_2 are vertical displacements of the left and right ends of the platform, respectively. Taking $m = 1$, $c_1 = c_2 = 1$, $k_1 = \xi_2$, $k_2 = \xi_3$, $F = \nu$ and introducing new state variables $\psi_1 = x_1 + x_2$, $\psi_2 = \dot{\psi}_1$, $\psi_3 = x_2$, after simple manipulations we obtain system (4.1), depending on the vector of parameters $\xi \in \mathbb{R}^3$ with one control variable ν, the state vector $\psi \in \mathbb{R}^3$, and the matrices

$$F(\xi) = \begin{pmatrix} 0 & 1 & 0 \\ -2\xi_2 & -2 & 2(\xi_2 - \xi_3) \\ \xi_2/2 & 1/2 & -(\xi_2 + \xi_3)/2 \end{pmatrix}, \quad G(\xi) = \begin{pmatrix} 0 \\ 2 \\ \xi_1 - 1/2 \end{pmatrix}. \tag{4.15}$$

Let us consider a point $\xi_0 = (1/4,\ 3/2,\ 5/6)$ in the parameter space. At this point

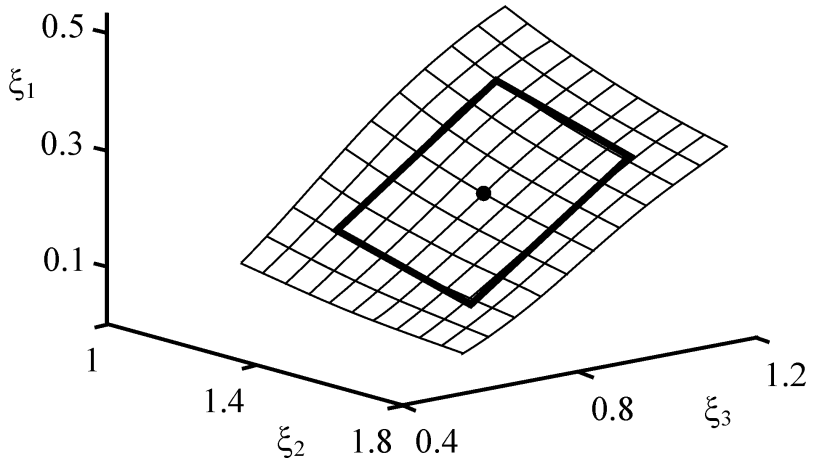


FIG. 3. *Uncontrollability set and its tangent plane.*

the pair of matrices (4.15) takes the form

$$F_0 = \begin{pmatrix} 0 & 1 & 0 \\ -3 & -2 & 4/3 \\ 3/4 & 1/2 & -7/6 \end{pmatrix}, \quad G_0 = \begin{pmatrix} 0 \\ 2 \\ -1/4 \end{pmatrix}. \tag{4.16}$$

It is straightforward to check that the pair (F_0, G_0) is uncontrollable and can be transformed to the Brunovsky form (4.4) with $\sigma_0 = -1$ by the triple $(P_0, R_0, S_0) \in \widetilde{\mathcal{G}}$ of the following form:

$$P_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3/8 & -1/8 & 1 \end{pmatrix}, \quad R_0 = 1/2, \quad S_0 = (5/4,\ 13/12,\ -2/3). \tag{4.17}$$

Using (4.15) and (4.17) in (4.5), we find

$$\eta = (2/3,\ 1/8,\ -3/8), \quad \eta_\sigma = -(2/3,\ 1/4,\ 3/4). \tag{4.18}$$

Hence, by Proposition 4.1, the uncontrollability set is a smooth hypersurface in the vicinity of ξ_0. The tangent plane to this surface at ξ_0 is given by the equation

$$(\eta, \xi - \xi_0) = \frac{2\xi_1}{3} + \frac{\xi_2}{8} - \frac{3\xi_3}{8} - \frac{1}{24} = 0, \tag{4.19}$$

and the perturbation of the uncontrollable mode on this surface has the form

$$\sigma(\xi) = -1 - \frac{2(\xi_1 - 1/4)}{3} - \frac{\xi_2 - 3/2}{4} - \frac{3(\xi_3 - 5/6)}{4} + o(\|\xi - \xi_0\|). \tag{4.20}$$

The plane (4.19) is plotted in Figure 3 (bold rectangular). For comparison, the uncontrollability set found numerically using (4.3) (determinant of the controllability matrix changes the sign when we cross the uncontrollability set) is shown in Figure 3. Numerical computations confirm the analytical results.

5. Conclusion. The general idea of any normal form theory is to transform an object under consideration to a form whose properties are easy to analyze. In this process both the normal form and transformation to it are important. For example, the Jordan normal form of a square matrix determines its spectrum, while knowledge of the transformation to the Jordan form (change of basis) allows us to find explicitly a general solution to the corresponding dynamical system.

In this paper we have solved the second part of the normal form problem (finding the transformation) in the reduction of families of matrix pencils and matrix pairs to the local normal form (miniversal deformation). Information on the transformation (the change of parameters and equivalence transformation) allows the development of the multi-parameter perturbation theory for multi-input linear dynamical systems. In a similar problem for square matrices, advantages of this approach for the perturbation analysis of the spectrum and stability of linear dynamical systems depending on parameters have been illustrated in [3, 12, 14, 15]. In section 4 of this paper it has been shown that the suggested method is useful for the controllability analysis of single-input dynamical systems dependent on parameters.

Acknowledgment. The second author thanks M. I. García-Planas for the hospitality during his staying at the Department of Applied Mathematics I, UPC, Barcelona.

REFERENCES

[1] V. I. Arnold, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.
[2] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1988.
[3] J. V. Burke and M. L. Overton, *Stable perturbations of nonsymmetric matrices*, Linear Algebra Appl., 171 (1992), pp. 249–273.
[4] A. Edelman, E. Elmroth, and B. Kågström, *A geometric approach to perturbation theory of matrices and matrix pencils. Part* I: *Versal deformations*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 653–692.
[5] A. Edelman and Y. Ma, *Staircase failures explained by orthogonal versal forms*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1004–1025.
[6] J. Ferrer, M. I. Garcia-Planas, and F. Puerta, *Brunowsky local form of a holomorphic family of pairs of matrices*, Linear Algebra Appl., 253 (1997), pp. 175–198.
[7] F. R. Gantmacher, *The Theory of Matrices*, Vols. 1, 2, Chelsea, New York, 1959.
[8] M. I. Garcia-Planas and M. D. Magret, *Deformation and stability of triples of matrices*, Linear Algebra Appl., 254 (1997), pp. 159–192.
[9] M. I. Garcia-Planas and M. D. Magret, *Miniversal deformations of linear systems under the full group action*, System Control Lett., 35 (1998), pp. 279–286.
[10] M. I. Garcia-Planas and V. V. Sergeichuk, *Simplest miniversal deformations of matrices, matrix pencils, and contragredient matrix pencils*, Linear Algebra Appl., 302–303 (1999), pp. 45–61.
[11] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, Chelsea, New York, 1977.
[12] A. A. Mailybaev, *Transformation of families of matrices to normal forms and its application to stability theory*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 396–417.
[13] A. A. Mailybaev, *Transformation to versal deformations of matrices*, Linear Algebra Appl., 337 (2001), pp. 87–108.
[14] A. A. Mailybaev, *On stability domains of nonconservative systems under small parametric excitation*, Acta Mechanica, 154 (2002), pp. 11–30.
[15] A. A. Mailybaev and A. P. Seyranian, *On singularities of a boundary of the stability domain*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 106–128.
[16] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Texts Appl. Math. 6, Springer-Verlag, New York, 1990.
[17] L. Stolovitch, *On the computation of a versal family of matrices*, Numer. Algorithms, 4 (1993), pp. 25–46.
[18] A. Tannenbaum, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Math. 845, Springer-Verlag, New York, 1981.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 963–970

AN ELEMENTARY COUNTEREXAMPLE TO THE FINITENESS CONJECTURE*

VINCENT D. BLONDEL†, JACQUES THEYS†, AND ALEXANDER A. VLADIMIROV‡

Abstract. We prove that there exist (infinitely many) values of the real parameters a and b for which the matrices

$$a\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad b\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

have the following property: all infinite periodic products of the two matrices converge to zero, but there exists a nonperiodic product that doesn't. Our proof is self-contained and fairly elementary; it uses only elementary facts from the theory of formal languages and from linear algebra. It is not constructive in that we do not exhibit any explicit values of a and b with the stated property; the problem of finding explicit matrices with this property remains open.

Key words. spectral radius, generalized spectral radius, joint spectral radius, finiteness conjecture, matrix semigroup

AMS subject classifications. 15A18, 15A60, 93D09

PII. S0895479801397846

1. Introduction.

The Lagarias–Wang finiteness conjecture was introduced in 1995 in connection with problems related to spectral radius computation of finite sets of matrices. Let $\rho(A)$ be the spectral radius[1] of the matrix A and let Σ be a finite set of matrices. The generalized spectral radius of Σ is defined by

$$\rho(\Sigma) = \limsup_{k\to+\infty} \max\{\rho(A_1 \cdots A_k)^{1/k} : A_i \in \Sigma,\ i = 1, \ldots, k\}.$$

This quantity was introduced in [7, 8]. The generalized spectral radius is known to coincide (see [1]) with the earlier defined joint spectral radius [13]; we refer to these quantities simply as "spectral radius." The notion of the spectral radius of a set of matrices appears in a wide range of contexts and has led to a number of recent contributions (see, e.g., [2, 3, 6, 8, 11, 15, 16, 17]); a list of over a hundred related contributions is given in [14]. We describe below one particular occurrence in a dynamical system context.

We consider systems of the form $x_{t+1} = A_t x_t$, where Σ is a finite set of matrices, and $A_t \in \Sigma$ for every $t \geq 0$. We do not impose any restrictions on the sequence of matrices A_t. These are exactly the discrete-time linear time-varying systems for

*Received by the editors November 12, 2001; accepted for publication (in revised form) by U. Helmke August 2, 2002; published electronically February 12, 2003. This research was partially supported by the Russian Foundation for Fundamental Research (grants 0001-00571 and 00-15-96116), by INTAS (grant INTAS-265), by NATO (grant CRG-961115), and by the Belgian Prime Minister's Office, Office for Scientific, Technical and Cultural Affairs (Interuniversity Attraction Pole IAP IV/2).

http://www.siam.org/journals/simax/24-4/39784.html

†Division of Applied Mathematics, Université Catholique de Louvain, 4 avenue Georges Lemaitre, B-1348 Louvain-la-Neuve, Belgium (blondel@inma.ucl.ac.be, theys@inma.ucl.ac.be). The second author is a Research Fellow with the Belgian National Research Fund (FNRS).

‡Institute of Information Transmission Problems, Russian Academy of Science, 19 Bol'shoi Karetnyi St., Moscow, 101447, Russia (aav@redline.ru).

[1]The spectral radius of a matrix is equal to the magnitude of its largest eigenvalue.

which the dynamics is taken from a finite set at every time instant. Starting from the initial state x_0, we obtain

$$x_{t+1} = A_t \cdots A_1 A_0 x_0.$$

The spectral radius of Σ is known to characterize how fast x_t can possibly grow with t; see [6, 7]. In particular, the trajectories all converge to the origin if and only if $\rho(\Sigma) < 1$.

We now describe the finiteness conjecture. It is known that

$$\rho(\Sigma) \geq \max\{\rho(A_1 \cdots A_k)^{1/k} : A_i \in \Sigma,\ i = 1, \ldots, k\}$$

for all $k \geq 0$.

Finiteness conjecture. Let Σ be a finite set of matrices. Then there exists some $k \geq 1$ and a matrix $A = A_1 \ldots A_k$ with $A_i \in \Sigma$ such that $\rho(A)^{1/k} = \rho(\Sigma)$.

This conjecture appears in [12]. The problem of determining if the conjecture is true appears under a different guise in [5], where it is attributed to E. S. Pyatnicky.

In terms of the dynamical system interpretation given above, this conjecture can be restated as saying that the convergence to zero of all periodic products of a given finite set of matrices implies the same for all possible products.

The conjecture has recently been proved to be false [4]. The existence of a counterexample is proved in [4] by using iterated function systems, topical maps, and Sturmian sequences. The proof relies in part on a particular fixed point theorem known as Mañé's lemma. In this contribution, we provide an alternative proof. We prove that there are uncountably many values of the real parameter α for which the pair of matrices

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \alpha \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

does not satisfy the finiteness conjecture. Our proof is not constructive in that we do not exhibit any particular value of α for which the corresponding pair of matrices violates the finiteness conjecture. The problem of finding an explicit counterexample and the problem of determining if there exist matrices with rational entries that violate the conjecture remain open questions. As compared to the proof in [4], our proof has the advantage of being self-contained and fairly elementary; it uses only elementary facts from linear algebra.

2. Proof outline. Let us now briefly outline our proof. We define

$$A_0 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

and

$$A_0^\alpha = \frac{1}{\rho_\alpha} A_0, \quad A_1^\alpha = \frac{\alpha}{\rho_\alpha} A_1$$

with $\rho_\alpha = \rho(\{A_0, \alpha A_1\})$. Since $\rho(\lambda\Sigma) = |\lambda|\,\rho(\Sigma)$, the spectral radius of the set $\Sigma_\alpha = \{A_0^\alpha, A_1^\alpha\}$ is equal to one. Let $I = \{0, 1\}$ be a two-letter alphabet and let $I^+ = \{0, 1, 00, 01, 10, 11, 000, \ldots\}$ be the set of finite nonempty words. We will also denote the empty word by $\emptyset$ and use the notation $I^* = I^+ \cup \{\emptyset\}$. To the word $w = w_1 \ldots w_t \in I^+$ we associate the products $A_w = A_{w_1} \ldots A_{w_t}$ and $A_w^\alpha = A_{w_1}^\alpha \ldots A_{w_t}^\alpha$.

A word $w \in I^+$ will be said to be *optimal* for some α if $\rho(A_w^\alpha) = 1$. We use J_w to denote the set of α's for which $w \in I^+$ is optimal. If the finiteness conjecture is true, the union of the sets J_w for $w \in I^+$ covers the real line. We show that this union does not cover the interval $[0, 1]$.

In section 4, we show that if two words $u, v \in I^+$ are essentially equal, then $J_u = J_v$. Two words $u, v \in I^+$ are *essentially equal* if the periodic infinite words $U = uu\ldots$ and $V = vv\ldots$ can be decomposed as $U = xww\ldots$ and $V = yww\ldots$ for some $x, y, w \in I^+$. Words that are not essentially equal are *essentially different*. Obviously, if u and v are essentially different, then so are cyclic permutations of u and v. We show in the same section that the sets J_u and J_v are disjoint if u and v are essentially different. This part of the proof requires some properties of infinite words presented in section 3. The proof is then almost complete. To conclude, we observe in section 5 that the sets $J_w \cap [0, 1]$ are closed subintervals of $[0, 1]$. There are countably many words in I^+, and so $\cup_{w\in I^+}(J_w \cap [0, 1])$ is a countable union of disjoint closed subintervals of $[0, 1]$. Except for a trivial case that we can exclude here, there are always uncountably many points in $[0, 1]$ that do not belong to such a countable union. Each of these points provides a particular counterexample to the finiteness conjecture.

3. Palindromes in infinite words. The *length* of a word $w = w_1 \ldots w_t \in I^*$ is equal to $t \geq 0$ and is denoted by $|w|$. The *mirror image* of w is the word $\tilde{w} = w_t \ldots w_1 \in I^*$. A *palindrome* is a word in I^* that is identical to its mirror image. In particular, the empty word is a palindrome. For $u, v \in I^*$, we write $u > v$ if u is lexicographically larger than v, that is, $u_i = 1$, $v_i = 0$ for some $i \geq 1$ and $u_j = v_j$ for all $j < i$. This is only a partial order since, for example, 101000 and 1010 are not comparable. For an infinite word U, we denote by $F(U)$ the set of all finite factors of U.

LEMMA 3.1. *Let $u, v \in I^+$ be two words that are essentially different. We denote $U = uuu\ldots$ and $V = vvv\ldots$. Then there exists a pair of words $0p0$ and $1p1$ in the set $F(U) \cup F(V)$ such that $p \in I^*$ is a palindrome.*

Proof. Let m and n be the minimal periods of U and V, respectively. The values of m and n are invariant under cyclic permutations of u and v. Let us use induction on $m + n$. The result is obvious for $m + n = 2$ since in this case U and V must be equal to $111\ldots$ and $000\ldots$, and we may then take $p = \emptyset$. Consider now $u, v \in I^+$. If the words 00 and 11 both belong to the set $F(U) \cup F(V)$, then we can set $p = \emptyset$. So assume without loss of generality that 11 does not belong to $F(U) \cup F(V)$. We may also assume that both u and v begin with 0; otherwise, we can take appropriate cyclic permutations of u and v. Then u and v can be factorized in a unique way by factors $0' = 0$ and $1' = 01$.

In the new alphabet $\{0', 1'\}$, the resulting words u' and v' are still essentially different and the minimal periods m' and n' of $U' = u'u'\ldots$ and $V' = v'v'\ldots$ satisfy $m' + n' < m + n$. By induction, there exists a pair of words $0'q'0'$ and $1'q'1'$ in $F(U') \cup F(V')$ such that $q' = \tilde{q'}$. Let q be the word obtained from q' by replacing $0'$ by 0 and $1'$ by 01. From $1'q'1' \in F(U') \cup F(V')$ we get $01q01 \in F(U) \cup F(V)$. Since $0'q'0' \in F(U') \cup F(V')$ we have $0'q'0'0' \in F(U') \cup F(V')$ or $0'q'0'1' \in F(U') \cup F(V')$, and thus $0q00 \in F(U) \cup F(V)$. Define now $p = q0$ and observe that $0p0, 1p1 \in F(U) \cup F(V)$.

Finally, let us show that if q' is a palindrome in $\{0', 1'\}$, then $q0$ is a palindrome in $\{0, 1\}$. We use induction on $|q'|$. For $|q'| = 0, 1$ the statement is obviously true. Suppose that $|q'| \geq 2$. Then $q' = 0's'0'$ or $q' = 1's'1'$ for $s' \in \{0', 1'\}^*$, and s' is a

palindrome in $\{0', 1'\}$. By induction hypothesis, $s0$ is then a palindrome in $\{0, 1\}$. We then have that either $q = 0s00$ or $q = 01s010$, but since $s0$ is a palindrome it follows that p is also a palindrome. ∎

COROLLARY 3.2. *Let $u, v \in I^+$ be two essentially different words and let $U = uuu\ldots$ and $V = vvv\ldots$. Then there exist words $a, b, x, y \in I^+$ satisfying $|x| = |y|$, $x > y$, $\tilde{x} > \tilde{y}$, $x > \tilde{y}$, $\tilde{x} > y$, and a palindrome $p \in I^*$ such that*

$$U = apxpxp\ldots \quad \text{and} \quad V = bpypyp\ldots$$

or one of the words U and V, say U, can be decomposed as

$$U = apxpxp\cdots = bpypyp\ldots.$$

Proof. By Lemma 3.1, there exists a pair of words $0p0$ and $1p1$ in the set $F(U) \cup F(V)$ such that p is a palindrome. Without loss of generality, assume that $1p1$ occurs in U. Then it occurs in U infinitely many times because U is periodic. Let us write

$$U = a'1p1d1p1d\ldots$$

and, analogously,

$$W = b'0p0f0p0f\ldots,$$

where W is either U or V. Without loss of generality we may assume $|d| = |f|$; otherwise, we can always take $d' = d1p1d\ldots1p1d$ instead of d and $f' = f0p0f\ldots0p0f$ instead of f in such a way that $|f'| = |d'|$. It remains to set $a = a'1$, $b = b'0$, $x = 1d1$, and $y = 0f0$. ∎

4. Optimal words are essentially equal. For a given word $w \in I^+$ we define $J_w = \{\alpha : \rho(A_w^\alpha) = 1\}$. Our goal in this section is to prove that J_u and J_v are equal when u and v are essentially equal, and have otherwise empty intersection.

LEMMA 4.1. *Let $u, v \in I^+$ be two words that are essentially equal. Then $J_u = J_v$.*

Proof. Assume $u, v \in I^+$ are essentially equal. Then $U = uu\ldots$ and $V = vv\ldots$ can be written as $U = ss\ldots$ and $V = tt\ldots$ with $|s| = |t|$ and t a cyclic permutation of s. The spectral radius satisfies $\rho(AB) = \rho(BA)$, and so the spectral radius of a product of matrices is invariant under cyclic permutations of the product factors. From this it follows that $\rho(A_s^\alpha) = \rho(A_t^\alpha)$, and hence u is optimal whenever v is. ∎

We need two preliminary lemmas for proving the next result.

LEMMA 4.2. *For any word $w \in I^+$ we have*

$$A_{\tilde{w}} - A_w = k(w)T,$$

where $k(w)$ is an integer and

$$T = A_0A_1 - A_1A_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Moreover, $k(w)$ is positive if and only if $w > \tilde{w}$.

Proof. Let us prove by induction that

$$A_w = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

implies

$$A_{\tilde{w}} = \begin{pmatrix} d & b \\ c & a \end{pmatrix}.$$

Indeed, this is true for $w = 0$ and $w = 1$. Notice also that

$$A_0 \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a+c & b+d \\ c & d \end{pmatrix}$$

and

$$\begin{pmatrix} d & b \\ c & a \end{pmatrix} A_0 = \begin{pmatrix} d & b+d \\ c & a+c \end{pmatrix},$$

and similarly for A_1. From this it follows that $A_{\tilde{w}} - A_w = k(w)T$. The sign relation follows from the fact that

$$A_0 \begin{pmatrix} a & b \\ c & d \end{pmatrix} A_1 - A_1 \begin{pmatrix} d & b \\ c & a \end{pmatrix} A_0 = \begin{pmatrix} a+b+c & 0 \\ 0 & -(a+b+c) \end{pmatrix}. \qquad \square$$

We say that a matrix A *dominates* B if $A \geq B$ componentwise and $\mathrm{tr}A > \mathrm{tr}B$ (tr denotes the trace). The eigenvalues of the 2×2 matrix A are given by $(\mathrm{tr}A \pm \sqrt{(\mathrm{tr}A)^2 - 4\det A})/2$. For all words w, the matrix A_w satisfies $\det(A_w) = 1$ and $\mathrm{tr}(A_w) \geq 2$. We therefore have $\rho(A_u) > \rho(A_v)$ whenever A_u dominates A_v.

LEMMA 4.3. *For any word of the form $w = psq$, where $s > \tilde{s}$ and $q < \tilde{p}$, the matrix $A_{w'}$ with $w' = p\tilde{s}q$ dominates A_w.*

Proof. We have $A_{w'} - A_w = k(s)A_pTA_q$, $k(s) > 0$. The relations $A_iTA_i = T$, $i = 0, 1$, and

$$A_1TA_0 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

finish the proof. $\square$

Let $w = psq$. If $s > \tilde{s}$ and $q < \tilde{p}$, we say that $s \to \tilde{s}$ is a *dominating flip*. We are now ready to prove the main result of this section.

LEMMA 4.4. *Let $u, v \in I^+$ be two words that are essentially different. Then $J_u \cap J_v = \emptyset$.*

Proof. Let $u, v \in I^+$ be two words that are essentially different. We assume without loss of generality that neither $U = uu\ldots$ nor $V = vv\ldots$ is equal to $00\ldots$ or $11\ldots$ because $11\ldots$ is not optimal for any $\alpha \in [0, 1]$ and $00\ldots$ is only optimal for $\alpha = 0$, but no other word is optimal for $\alpha = 0$. In order to prove the result we show that if $\rho(A_u^\alpha) = \rho(A_v^\alpha)$ for some value of α, then there exists a word w satisfying $\rho(A_w^\alpha) > \rho(A_u^\alpha)$.

By Corollary 3.2, there exist words $a, b, x, y \in I^+$ satisfying $|x| = |y|$, $x > y$, $\tilde{x} > \tilde{y}, x > \tilde{y}, \tilde{x} > y$, and a palindrome $p \in I^*$ such that

$$U = apxpxp\ldots \quad \text{and} \quad V = bpypyp\ldots$$

or

$$U = apxpxp\ldots = bpypyp\ldots.$$

Since neither U nor V is equal to $00\ldots$ or $11\ldots$, the matrices A_{xp} and A_{yp} are strictly positive.

Let us consider the word $xpxpxpypypyp$. Setting $s = xpy$, we make a dominating flip in this word and get the word $xpxp\tilde{y}p\tilde{x}pypyp$. Then we set $s = xp\tilde{y}p\tilde{x}py$ and make another dominating flip. As a result, the matrix $A_{xpxpxpypypyp}$ is dominated by the matrix $A_{xp\tilde{y}pxpyp\tilde{x}pyp}$. Analogously, any matrix $A_sA_vA_r$, $v \in I^*$, is dominated by the matrix $A_{s'}A_vA_{r'}$ where $s = xpxpxp$, $r = ypypyp$, $s' = xp\tilde{y}pxp$, and $r' = yp\tilde{x}pyp$. Let us denote the linear operators $A \to A_sAA_r$ and $A \to A_{s'}AA_{r'}$ acting in $\mathbb{R}^4$ as well as their 4×4 matrices by L and L', respectively. It is known that $L = A_r^T \otimes A_s$ and $L' = (A'_r)^T \otimes A'_s$, where $\otimes$ is used to denote the Kronecker (tensor) product (see [10, Lemma 4.3.1]). Both L and L' are strictly positive. The minimal closed convex cone in $\mathbb{R}^4$ containing all matrices A_v, $v \in I^*$, is the cone of all nonnegative 2×2 matrices. Indeed, any nonnegative matrix X with $\det(X) = 0$ can be approximated by matrices of the form βA_w, $\beta > 0$, $w \in I^*$. In particular, this is true for the matrices

$$\begin{pmatrix}1&0\\0&0\end{pmatrix},\quad \begin{pmatrix}0&1\\0&0\end{pmatrix},\quad \begin{pmatrix}0&0\\1&0\end{pmatrix},\quad \begin{pmatrix}0&0\\0&1\end{pmatrix}.$$

Hence $L' \geq L$ elementwise and $L \neq L'$. From the Perron–Frobenius theory (see, for instance, Problem 8.15 in [9]) we get $\rho(L') > \rho(L)$. The spectral radius of a Kronecker product is the product of the spectral radii (see [10, Theorem 4.2.12]), and so

$$\rho(L') = \rho(A_{s'})\rho(A_{r'}) > \rho(A_s)\rho(A_r) = \rho(L).$$

Since the flips performed do not change the average proportion of matrices A_0 and A_1 in the product, we can also write

$$\rho(L^\alpha) = \rho(A_s^\alpha)\rho(A_r^\alpha) \quad \text{and} \quad \rho(L'^\alpha) = \rho(A_{s'}^\alpha)\rho(A_{r'}^\alpha)$$

for each $\alpha > 0$, where L^α and L'^α are defined analogously to L and L'. Suppose that $\rho(A_u^\alpha) = \rho(A_v^\alpha) = 1$. Then $\rho(A_s^\alpha) = \rho(A_r^\alpha) = 1$ and, hence, either $\rho(A_{s'}^\alpha) > 1$ or $\rho(A_{r'}^\alpha) > 1$, which is a contradiction. ∎

5. Finiteness conjecture. We are now ready to prove the main result.

THEOREM 5.1. *There are uncountably many values of the real parameter α for which the pair of matrices*

$$\begin{pmatrix}1&1\\0&1\end{pmatrix},\quad \alpha\begin{pmatrix}1&0\\1&1\end{pmatrix}$$

does not satisfy the finiteness conjecture.

Proof. It is clearly equivalent to prove the statement for the matrices A_0^α and A_1^α. For $\alpha = 0$, all optimal words w are essentially equal to 0. For any other word w, the set $J_w \cap [0,1]$ can be written as

$$J_w \cap [0,1] = \{\alpha \in (0,1] : \rho(A_w^\alpha) = 1\}$$

or, equivalently,

$$J_w \cap [0,1] = \left\{\alpha \in (0,1] : (\rho(A_w)\alpha^{|w|_1})^{\frac{1}{|w|}} = \sup_{v\in I^+}(\rho(A_v)\alpha^{|v|_1})^{\frac{1}{|v|}}\right\}. \tag{5.1}$$

In this expression $|w|_1$ denotes the number of 1's in the word w. Associated to $w \in I^+$ we define the affine function

$$h_w(\beta) = \frac{1}{|w|}(\ln \rho(A_w) + |w|_1 \beta)$$

and let

$$h(\beta) = \sup_{w \in I^+} h_w(\beta).$$

Passing to the logarithmic scale in the expression (5.1), we get

$$J_w \cap [0,1] = \{e^\beta : \beta \in \mathbb{R}, h_w(\beta) = h(\beta)\} \cap [0,1]. \tag{5.2}$$

The functions h_w are affine, h is convex and continuous, and $h(\beta) \geq h_w(\beta)$ for all $w \in I^+$ and $\beta \in \mathbb{R}$. From this it follows that the set $\{\beta \in \mathbb{R} : h_w(\beta) = h(\beta)\}$ is an interval of the real line. This interval is the zero set of a continuous function, and it is therefore closed. From (5.2) we conclude that $J_w \cap [0,1]$ is a closed subinterval of $[0,1]$.

Let us finally show that $[0,1]$ cannot be covered by countably many disjoint closed intervals H_i, $i \geq 1$ (possibly, single points), unless this is a single interval, which is, obviously, not the case here.

We define a function $g(\alpha) : [0,1] \to [0,1]$ as follows. We set $g(0) = 0$, $g(1) = 1$ and then set $g(\alpha) = 1/2$ for all $\alpha \in H_1$. For each subsequent index i, we define $g(\alpha) = g_i = (a_+ + a_-)/2$ for all $\alpha \in H_i$, where a_- is the current highest value of $g(\cdot)$ at the left of H_i and a_+ is the current lowest value of $g(\cdot)$ at the right of H_i.

As a result, the function $g(\cdot)$ is well-defined on $[0,1] \cap (\cup_{i=1,2,\ldots} H_i)$ and non-decreasing. It can be then extended to the whole segment $[0,1]$ by continuity because between any two segments H_i and H_j there exists a segment H_k with $k > i, j$. Since $g(0) = 0$ and $g(1) = 1$, the range of $g(\alpha)$ coincides with $[0,1]$ for $\alpha \in [0,1]$. Therefore, there exist uncountably many values of $\alpha \in [0,1]$ such that $g(\alpha) \neq g_i$, $i = 1, 2, \ldots$. □

REFERENCES

[1] M. Berger and Y. Wang, *Bounded semigroups of matrices,* Linear Algebra Appl., 166 (1992), pp. 21–27.

[2] V. D. Blondel and J. N. Tsitsiklis, *A survey of computational complexity results in systems and control,* Automatica, 36 (2000), pp. 1249–1274.

[3] V. D. Blondel and J. N. Tsitsiklis, *The boundedness of all products of a pair of matrices is undecidable,* Systems Control Lett., 41 (2000), pp. 135–140.

[4] T. Bousch and J. Mairesse, *Asymptotic height optimization for topical IFS, Tetris heaps and the finiteness conjecture,* J. Amer. Math. Soc., 15 (2002), pp. 77–111.

[5] L. Gurvits, *Stability and observability of discrete linear inclusion—Finite automata approach,* in Book of Abstracts, International Symposium on the Mathematical Theory of Networks and Systems, Kobe, Japan, 1991, pp. 166–167.

[6] L. Gurvits, *Stability of discrete linear inclusions,* Linear Algebra Appl., 231 (1995), pp. 47–85.

[7] I. Daubechies and J. C. Lagarias, *Sets of matrices all infinite products of which converge,* Linear Algebra Appl., 161 (1992), pp. 227–263.

[8] I. Daubechies and J. C. Lagarias, *Corrigendum/addendum to "Sets of matrices all infinite products of which converge,"* Linear Algebra Appl., 327 (2001), pp. 69–83.

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

[10] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[11] V. S. Kozyakin, *Algebraic unsolvability of a problem on the absolute stability of desynchronized systems,* Automat. Remote Control, 51 (1990), pp. 754–759.
[12] J. C. Lagarias and Y. Wang, *The finiteness conjecture for the generalized spectral radius of a set of matrices,* Linear Algebra Appl., 214 (1995), pp. 17–42.
[13] G.-C. Rota and W. G. Strang, *A note on the joint spectral radius,* Indag. Math., 22 (1960), pp. 379–381.
[14] G. Strang, *The joint spectral radius,* Commentary by Gilbert Strang on paper number 5, in Collected Works of Gian-Carlo Rota, 2001; available online from http://www-math.mit.edu/~gs.
[15] J. N. Tsitsiklis and V. D. Blondel, *The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate,* Math. Control Signals Systems, 10 (1997), pp. 31–40.
[16] A. A. Vladimirov, L. Elsner, and W.-J. Beyn, *Stability and paracontractivity of discrete linear inclusions,* Linear Algebra Appl., 312 (2000), pp. 125–134.
[17] F. Wirth, *The generalized spectral radius and extremal norms,* Linear Algebra Appl., 342 (2002), pp. 17–40.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 971–989

A SCHUR ALGORITHM FOR COMPUTING MATRIX *P*TH ROOTS*

MATTHEW I. SMITH†

Abstract. Any nonsingular matrix has pth roots. One way to compute matrix pth roots is via a specialized version of Newton's method, but this iteration has poor convergence and stability properties in general. A Schur algorithm for computing a matrix pth root that generalizes methods of Björck and Hammarling [*Linear Algebra Appl.*, 52/53 (1983), pp. 127–140] and Higham [*Linear Algebra Appl.*, 88/89 (1987), pp. 405–430] for the square root is presented. The algorithm forms a Schur decomposition of A and computes a pth root of the (quasi-)triangular factor by a recursion. The backward error associated with the Schur method is examined, and the method is shown to have excellent numerical stability.

Key words. matrix pth root, Schur algorithm, Newton's method, commutativity, stability, real arithmetic, rounding error analysis

AMS subject classification. 65F30

PII. S0895479801392697

1. Introduction.

Given a matrix $A \in \mathbb{C}^{n\times n}$, a matrix X is a pth root of A if

$$X^p = A. \tag{1.1}$$

For the scalar case, $n = 1$, we know that every nonzero complex number has p distinct roots. But for $n > 1$, a matrix pth root may not exist or there may be infinitely many solutions of (1.1). For example, the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

has no square root, while any involutary matrix is a square root of the identity matrix. If the matrix A is nonsingular, it always has a pth root, but for singular matrices existence depends on the structure of the elementary divisors of A corresponding to the zero eigenvalues (see [23, section 8.6], [8, section 8.7]). We will restrict our attention to the roots of nonsingular matrices.

One approach to computing the matrix pth root is to apply Newton's method to the matrix equation (1.1). Hoskins and Walton [14] implement a specialized form of Newton's method based on commutativity assumptions and apply it to symmetric positive definite A. However, this method is not practically viable, as we will show. Björck and Hammarling [1] and Higham [11] offer methods for computing square roots of A that first form a Schur decomposition of A and then use stable and efficient recursive formulae to obtain a square root of the triangular factor. Here we present a generalization of these Schur methods that computes a pth root for arbitrary $p \geq 2$, using only real arithmetic if the matrix A is real.

Applications requiring the matrix pth root arise in system theory in connection with the matrix sector function, defined by $\mathrm{sector}(A) = (A^p)^{-1/p}A$ [19], [2]. Another

*Received by the editors July 20, 2001; accepted for publication (in revised form) by M. Chu July 18, 2002; published electronically February 25, 2003. This research was supported by an Engineering and Physical Sciences Research Council Studentship.
http://www.siam.org/journals/simax/24-4/39269.html
†Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (msmith@ma.man.ac.uk, http://www.ma.man.ac.uk/~msmith/).

application is in the inverse scaling and squaring method for computing the matrix logarithm which can be expressed as $\log A = p \log A^{1/p}$ [16], [3]. Among all pth roots it is usually the principal pth root that is of interest.

DEFINITION 1.1. *Assume that the nonsingular matrix $A \in \mathbb{C}^{n\times n}$ has eigenvalues $\Lambda(A) = \{\lambda_i \mid i = 1\colon n\}$ with $\arg(\lambda_i) \neq \pi$ for all i. Then the principal pth root of A, denoted by $A^{1/p} \in \mathbb{C}^{n\times n}$, is the matrix satisfying*

- $(A^{(1/p)})^p = A$,
- $\arg(\Lambda(A^{1/p}) \in \left(\frac{-\pi}{p}, \frac{\pi}{p}\right)$.

In section 2 we define a function of a matrix. In particular we look at the matrix pth root function and find that in general not all roots of a matrix A are functions of A. This leads to the classification of the solutions of (1.1) into those expressible as polynomials in A and those that are not.

In section 3 we examine Newton's method for solving the matrix pth root problem. Hoskins and Walton [14] show that for a positive definite matrix a specialized version of Newton's method converges to the unique positive definite pth root provided the starting approximation is taken to be A or the identity matrix. We show that for general A this method fails to converge globally and that, when it does converge, it is usually unstable and thus is of little practical interest.

In section 4 we present our Schur method for computing pth roots. The basic step is the calculation of a pth root of a (quasi-)triangular matrix, using entirely real arithmetic if the original matrix is real. We give a rounding error analysis to show that our algorithm is numerically stable.

2. The matrix *p*th root function. For a given function f and $A \in \mathbb{C}^{n\times n}$ Gantmacher [8, Chapter 5] defines $f(A) = p(A)$, where p is a polynomial of minimal degree that interpolates to f on the spectrum of A, that is,

$$p^{(j)}(\lambda_i) = f^{(j)}(\lambda_i), \quad j = 0\colon n_i - 1, \quad i = 1\colon s,$$

where A has s distinct eigenvalues λ_i and n_i is the largest Jordan block in which λ_i appears. We are particularly interested in the function $f(\lambda) = \lambda^{1/p}$, which is clearly defined on the spectrum of nonsingular A. However, $f(\lambda)$ is a multivalued function, giving a choice of p single valued branches for each eigenvalue λ_i. As A has s distinct eigenvalues, we have a total of p^s matrices $f(A)$ when all combinations of branches are accounted for. Hence the matrix pth root function is not uniquely determined until we specify which branch of the pth root function is to be taken in the neighborhood of each eigenvalue λ_i.

We now classify all the pth roots of a nonsingular $A \in \mathbb{C}^{n\times n}$. We require the following result regarding the pth roots of a Jordan block.

THEOREM 2.1. *For $\lambda_k \neq 0$ the Jordan block,*

$$J_k = J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & & 0 \\ & \lambda_k & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k\times m_k}, \tag{2.1}$$

has precisely p upper triangular pth roots

$$f_j(J_k) = \begin{bmatrix} f_j(\lambda_k) & f_j'(\lambda_k) & \cdots & \frac{f_j^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f_j(\lambda_k) & \ddots & \vdots \\ & & \ddots & f_j'(\lambda_k) \\ 0 & & & f_j(\lambda_k) \end{bmatrix}, \quad j = 1\colon p, \tag{2.2}$$

where $f(\lambda) = \lambda^{1/p}$ and the subscript j denotes the branch of the pth root function in the neighborhood of λ_k. The p pth roots are functions of J_k.

Proof. The pth root function $f(\lambda) = \lambda^{1/p}$ is clearly defined on the spectrum of the Jordan block (2.1). Hence the formula (2.2) for the p distinct roots, $f_j(J_k)$, follows directly from the definition of $f(A)$ [8, Chapter 5].

We need to show that these p roots are the only upper triangular pth roots of J_k. Suppose that $X = (x_{\alpha,\beta})$ is an upper triangular pth root of J_k. Equating the (α,α) and $(\alpha,\alpha+1)$ elements in $X^p = J_k$ gives

$$x_{\alpha,\alpha}^p = \lambda_k, \quad 1 \le \alpha \le m_k, \tag{2.3}$$

and

$$x_{\alpha,\alpha+1} \sum_{r=0}^{p-1} x_{\alpha,\alpha}^{p-1-r} x_{\alpha+1,\alpha+1}^r = 1, \quad 1 \le \alpha \le m_k - 1. \tag{2.4}$$

If the eigenvalue λ_k has the polar representation $|\lambda_k|e^{i\theta}$, the p pth roots of (2.3) are

$$x_{\alpha,\alpha} = |\lambda_k|^{1/p} e^{i(\theta+2\pi q)/p}, \quad q = 0\colon p-1.$$

Let the α and $\alpha+1$ diagonal entries of X be

$$x_{\alpha,\alpha} = |\lambda_k|^{1/p} e^{i(\theta+2\pi q_1)/p}, \quad x_{\alpha+1,\alpha+1} = |\lambda_k|^{1/p} e^{i(\theta+2\pi q_2)/p}, \quad q_1, q_2 \in \{0, 1, \ldots, p-1\}.$$

The summation in (2.4) now becomes

$$\begin{aligned} &|\lambda_k|^{(p-1)/p} e^{i\theta(p-1)/p} \sum_{r=0}^{p-1} e^{i2\pi q_1(p-1-r)/p} e^{i(2\pi q_2)/p} \\ &= |\lambda_k|^{(p-1)/p} e^{i\theta(p-1)/p} e^{i2\pi q_1(p-1)/p} \sum_{r=0}^{p-1} e^{i2\pi(q_2-q_1)r/p}. \end{aligned}$$

Equation (2.4) implies that the above sum does not equal zero. In turn, this implies that $\sum_{r=0}^{p-1} e^{i2\pi(q_2-q_1)r/p} \neq 0$. If $x_{\alpha,\alpha}$ and $x_{\alpha+1,\alpha+1}$ are chosen to have the same value, then $q_1 = q_2$, and the summation term becomes

$$\sum_{r=0}^{p-1} e^{i2\pi(q_2-q_1)r/p} = p.$$

If instead the diagonal entries are taken to be roots of λ_k from different branches, then $q_1 \neq q_2$, and the sum becomes

$$\sum_{r=0}^{p-1} e^{i2\pi(q_2-q_1)r/p} = \frac{1 - e^{i2\pi(q_2-q_1)}}{1 - e^{i2\pi(q_2-q_1)/p}} = 0.$$

Hence $q_1 = q_2$. It follows that X has a constant diagonal, and since X can be shown to be uniquely determined by its diagonal elements (see section 4) the result follows. ∎

Theorem 2.1 shows that all roots of a Jordan block, J_k, with constant diagonal entries are functions of J_k and thus, by definition, are polynomials in J_k. However, not all pth roots of a matrix are necessarily functions of the matrix. The pth roots of A that are functions of A are polynomials in A, by definition. Consider, for example, the involutary matrix

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

We have $X^2 = I$, but X is clearly not a polynomial in I. Consequently the identity matrix has square roots that are not functions of the matrix in the sense defined above.

We can classify all the pth roots of a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ into two groups: those that are polynomials in A and those that are not.

THEOREM 2.2. *Let the nonsingular matrix $A \in \mathbb{C}^{n\times n}$ have the Jordan decomposition $A = ZJZ^{-1} = Z\,\mathrm{diag}(J_1, J_2, \ldots, J_r)Z^{-1}$ where each Jordan block $J_i \in \mathbb{C}^{m_i\times m_i}$ and $m_1 + m_2 + \cdots + m_r = n$. Let $s \le r$ be the number of distinct eigenvalues of A. A has precisely p^s pth roots that are functions of A, given by*

$$X_j = Z\,\mathrm{diag}\left(f_{j_1}(J_1), f_{j_2}(J_2), \ldots, f_{j_r}(J_r)\right) Z^{-1}, \quad j = 1\colon p^s, \tag{2.5}$$

corresponding to all possible choices of $j_1, \ldots, j_r$, $j_k \in \{1, 2, \ldots, p\}$, $k = 1\colon r$, subject to the constraint that $j_i = j_k$ whenever $\lambda_i = \lambda_k$.

If $s < r$, A has pth roots which are not functions of A. These pth roots form parameterized families

$$X_j(U) = ZU\,\mathrm{diag}\left(f_{j_1}(J_1), f_{j_2}(J_2), \ldots, f_{j_r}(J_r)\right) U^{-1}Z^{-1}, \quad p^s + 1 \le j \le p^r,$$

where $j_k \in \{1, 2, \ldots, p\}$, $k = 1\colon r$, U is an arbitrary nonsingular matrix that commutes with J, and for each j there exist i and k, depending on j, such that $\lambda_i = \lambda_k$ while $j_i \ne j_k$.

Proof. From the definition of a matrix function there are precisely p^s pth roots of A that are functions of A. We have [8, p. 98ff.]

$$f(A) = f(ZJZ^{-1}) = Zf(J)Z^{-1} = Z\,\mathrm{diag}\left(f(J_k)\right) Z^{-1},$$

and on combining this with Theorem 2.1, it follows that (2.5) gives the p^s pth roots of A that are functions of A.

The second part ensues from [8, pp. 231, 232] and the proof of Higham [11, Theorem 4]. ∎

The essential difference between the roots of A that are functions of A and those that are not is that for all Jordan blocks corresponding to λ_i, the same single valued branch of $\lambda_i^{1/p}$ is chosen. Theorem 2.2 shows that the pth roots of A which are functions of A are isolated pth roots. In contrast, the pth roots that are not functions of A form a finite number of parameterized families. Each family contains infinitely many pth roots sharing the same spectrum.

Note that Theorem 2.2 shows that the principal pth root defined in Definition 1.1 is indeed unique.

3. Newton's method for the matrix pth root. For a general function $F\colon \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, Newton's method for the solution of $F(X) = 0$ (see [6, p. 86], [18, p. 133]) is

$$X_{k+1} = X_k - F'(X_k)^{-1}F(X_k), \quad k = 0, 1, 2, \ldots,$$

where X_0 is given and F' is the Fréchet derivative of F.

Newton's method has been used to compute the positive definite square root of a positive definite matrix A by Higham [10]. The more general problem of determining a matrix pth root is discussed by Hoskins and Walton [14]. Here, for nonsingular $A \in \mathbb{C}^{n\times n}$, we need to solve

$$F(X) \equiv X^p - A = 0. \tag{3.1}$$

Consider the Taylor series for F about X,

$$F(X + H) = F(X) + F'(X)H + O(H^2). \tag{3.2}$$

From the definition of the matrix pth root (3.1) we have

$$\begin{aligned} F(X + H) &= (X + H)^p - A \\ &= F(X) + (X^{p-1}H + X^{p-2}HX + X^{p-3}HX^2 \\ &\qquad + \cdots + XHX^{p-2} + HX^{p-1}) + O(H^2), \end{aligned}$$

and by comparing terms with the Taylor series (3.2), we see that

$$F'(X)H = X^{p-1}H + X^{p-2}HX + \cdots + XHX^{p-2} + HX^{p-1}.$$

Thus, we may write Newton's method for the matrix pth root as, given X_0,

$$\text{solve} \quad \left.\begin{aligned} X_k^{p-1}H_k + X_k^{p-2}H_kX_k + \cdots + H_kX_k^{p-1} &= A - {X_k}^p \\ X_{k+1} &= X_k + H_k \end{aligned}\right\}, \qquad k = 0, 1, 2, \ldots. \tag{3.3}$$

The standard local convergence theorem for Newton's method [6, p. 90] tells us that, provided $\|X - X_0\|$ is sufficiently small and the linear transformation $F'(X)$ is nonsingular, the Newton iteration (3.3) converges quadratically to a pth root X of A.

Newton's method requires us to solve the equation for H_k in (3.3). For $p > 2$ this can be done with the aid of the vec operator, which for $A = [a_1, a_2, \ldots, a_n]$ is defined as $\operatorname{vec}(A) = \left(a_1^T, a_2^T, \ldots, a_n^T\right)^T$, together with the Kronecker product $A \otimes B = (a_{ij}B)$. Applying the vec operator to (3.3) and using the property that $\operatorname{vec}(AXB) = (B^T \otimes A)\operatorname{vec}(X)$ [5, Problem 6.4], we obtain

$$\left((I \otimes X^{p-1}) + (X^T \otimes X^{p-2}) + \cdots + ((X^{p-1})^T \otimes I)\right)\operatorname{vec}(H) = \operatorname{vec}(A - X^p). \tag{3.4}$$

The linear system (3.4) can be solved using any standard method, provided the coefficient matrix is nonsingular. However, (3.4) is an $n^2 \times n^2$ linear system, which makes both storage and computation expensive. A reasonable assumption (which will be justified in Theorem 3.1) to reduce the cost of solving (3.3) is that the commutativity relation

$$X_0H_0 = H_0X_0$$

holds. Then, for example, (3.3) may be written as

$$\left.\begin{array}{cc} \text{solve} & pX_k{}^{p-1}H_k = pH_kX_k{}^{p-1} = A - X_k{}^p \\ & X_{k+1} = X_k + H_k \end{array}\right\}, \qquad k = 0, 1, 2, \ldots.$$

Hence, from the Newton iteration (3.3), we can obtain the two simplified iterations

$$Y_{k+1} = \frac{1}{p}((p-1)Y_k + AY_k^{1-p}) \tag{3.5}$$

and

$$Z_{k+1} = \frac{1}{p}((p-1)Z_k + Z_k^{1-p}A), \tag{3.6}$$

provided that Y_k and Z_k are nonsingular.

3.1. Convergence of Newton's method. In this section we look at the convergence of Newton's method for the matrix pth root. Let us consider the relationship between the Newton iteration (3.3) and the simplified iterations (3.5) and (3.6). Note that the Newton iterates are well defined if and only if, for each k, equation (3.3) has a unique solution, that is, the Fréchet derivative, $F'(X_k)$, is nonsingular.

THEOREM 3.1. *Consider the iterations* (3.3), (3.5), *and* (3.6). *Suppose* $X_0 = Y_0 = Z_0$ *commutes with* A *and that all the Newton iterates* X_k *are well defined. Then*

1. X_k *commutes with* A *for all* k,
2. $X_k = Y_k = Z_k$ *for all* k.

Proof. The proof follows from a suitable modification of the proof of Theorem 1 in [10]. ∎

Hence the Newton iteration (3.3) and its off-shoots (3.5), (3.6) give the same sequence of iterates provided that the initial approximation $X_0 = Y_0 = Z_0$ commutes with A and both X_k and $F'(X_k)$ are nonsingular at each stage. The convergence of this sequence is now examined, concentrating on iteration (3.5) for convenience.

Assume that A is diagonalizable. Then there exists a nonsingular matrix W such that

$$W^{-1}AW = \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{3.7}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. We are now in a position to diagonalize the iteration. If we define

$$D_k = W^{-1}Y_kW, \tag{3.8}$$

then, from (3.5), we have

$$D_{k+1} = \frac{1}{p}\left((p-1)D_k + \Lambda D_k{}^{1-p}\right). \tag{3.9}$$

If the starting matrix D_0 is diagonal, all successive iterates D_k are also diagonal, and so we may analyze the convergence of the diagonalized iterates

$$D_k = \operatorname{diag}(d_i{}^{(k)}).$$

The iteration (3.9) becomes

$$d_i{}^{(k+1)} = \frac{1}{p}\left((p-1)d_i{}^{(k)} + \frac{\lambda_i}{(d_i{}^{(k)})^{(p-1)}}\right), \quad i = 1\colon n,$$

that is, n-uncoupled scalar Newton iterations for the pth root of λ_i. Therefore it suffices to consider the scalar Newton iteration

$$x_{k+1} = \frac{1}{p}\left((p-1)x_k + \frac{a}{x_k^{p-1}}\right) \tag{3.10}$$

for the pth root of a.

For $p = 2$, the Newton iterations (3.3), (3.5), and (3.6) for the matrix square root of A are shown by Higham [10, Theorem 2] to converge quadratically to a square root X of A. From Theorem 2.2 it is clear that the computed square root is a function of A. In particular, for a suitable choice of starting value (e.g., $X_0 = I$ or $X_0 = A$), the Newton iteration converges quadratically to the principal square root of the matrix A. However, for $p > 2$ Newton's method for computing a pth root does not converge in general [19].

The scalar iterations of (3.10) exhibit fractal behavior. Therefore we are interested in finding out for which initial values the iteration (3.10) converges to a particular root. The solution is easy in the case of the square root, but higher order roots present considerable difficulty. The problem in choosing a starting point, x_0, of the Newton iteration is that there exist regions where iterates converge to fixed points or cycles of the function that are not the required roots. A number of people have studied the dynamics of Newton's method applied to a one-parameter family of polynomials, and with the help of numerical experiments and the classical theory of Julia [15] and Fatou [7] were able to describe the behavior of the iterates; see, for example, Curry, Garnett, and Sullivan [4] and Vrscay [22].

To examine the behavior of the Newton iteration (3.10), with $a = 1$, we used MATLAB with a square grid of 160,000 points to generate plots of the attractive basins (the set of points where the iteration converges to a particular root) and their boundary points (the boundary of a basin, B_i, is all points in whose neighborhood, no matter how small, there are points both in B_i and outside B_i) of the iterates, $\{x_k\}$. Each grid point was used as a starting value, x_0, and then shaded gray depending on which root of unity it converged to. Thus the attractive basin associated with each root is assigned a particular shade of gray. The pictures for $p = 2, 3, 4$, and 5 are shown in Figure 3.1.

The plot of the square root case shows that points in the right half plane are iterated to the positive square root of unity and points in the left half plane to -1. The boundary of these two regions is the imaginary axis, which constitutes the set of initial points for which the Newton iteration fails, since points lying on the imaginary axis iterate to points that are purely imaginary.

For $p > 2$ the Newton iterations do not have simple boundaries segmenting their attractive basins. Instead of the plane being bounded into p sectors each $2\pi/p$ wide, the basins of attraction are bounded by petal-type structures. The petals result from the fact that the boundary points of one basin of attraction are actually the boundary points of all the basins. These shared boundary points form a set known as the Julia set. Thus iterations that have more than 2 roots cannot have basin boundaries that are simple connected line segments, and so for $p > 2$, the boundaries of the attractive basins are fractals consisting of totally disconnected point sets. But how do we choose x_0 to achieve convergence to a desired root? We are interested in finding the principal root, so it is natural to start the iteration at a point within the wedge bounded by $\arg = (-\pi/p', \pi/p')$, with $p' > p$. The problem is that we do not know the value of p' due to the size of the Julia set. However, we can see that for any point lying on

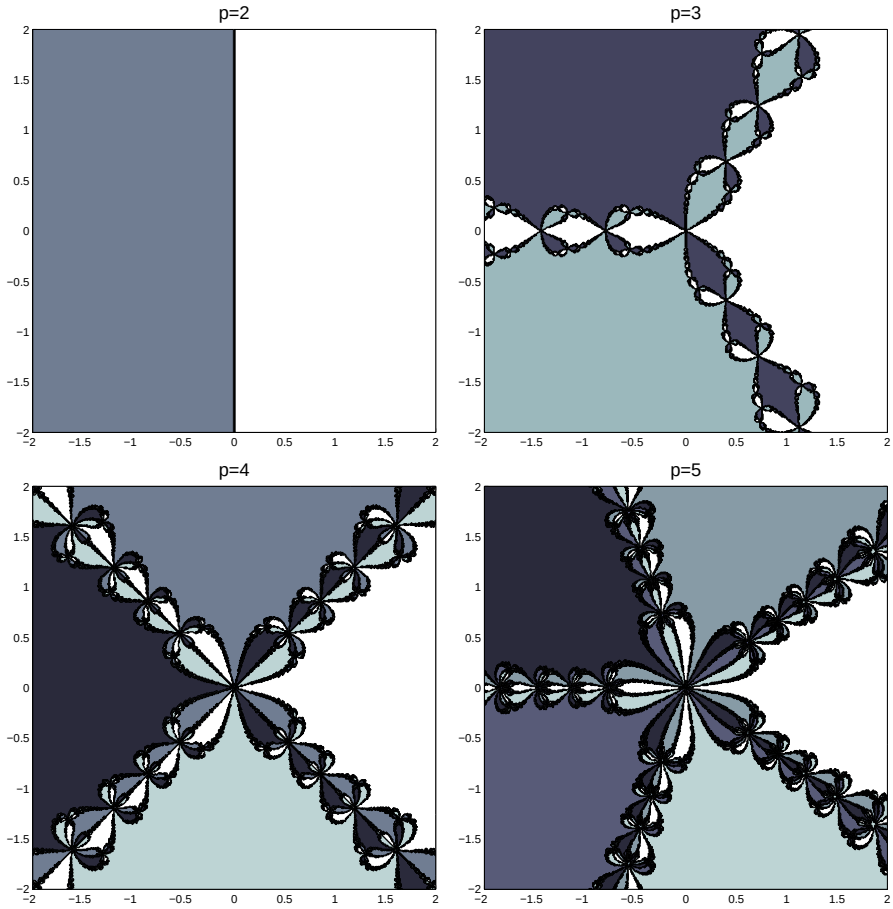


FIG. 3.1. *Fractal behavior of Newton's iteration* (3.10) *for the solution of* $x^p - 1 = 0$.

the nonnegative real axis, Newton's iteration (3.10) will converge to the principal pth root. Hence, for a positive definite matrix, A, the Newton iterations (3.3), (3.5), and (3.6) converge to the unique positive definite pth root of A, provided that the starting matrix is itself positive definite [14].

3.2. Stability analysis. We now consider the stability of Newton's method (3.3) and the two variants (3.5) and (3.6). It is known that Newton's method converges quadratically if started sufficiently close to a solution and, under reasonable assumptions, any errors arising due to floating point arithmetic are damped out in succeeding iterates [21]. But how do perturbations affect the behavior of Newton's method with commutativity assumptions? We will examine iteration (3.5) under the assumptions that the iteration converges in exact arithmetic (e.g., pth root of positive definite A) and A is diagonalizable. Let $\widehat{Y}_k$ denote the kth computed iterate and define

$$\Delta_k = \widehat{Y}_k - Y_k.$$

We make no assumption on the form of Δ_k, since it is intended to model general errors, including rounding errors. Our aim is to analyze how the perturbation Δ_k

propagates, so we assume $\widehat{Y}_{k+1}$ is computed exactly from $\widehat{Y}_k$, to give

$$\begin{aligned}\widehat{Y}_{k+1} &= \frac{1}{p}\left((p-1)\widehat{Y}_k + A\widehat{Y}_k^{1-p}\right)\\ &= \frac{1}{p}\left((p-1)[\Delta_k + Y_k] + A[\Delta_k + Y_k]^{1-p}\right).\end{aligned} \tag{3.11}$$

We need the perturbation result [20, p. 188]

$$(A+E)^{-1} = A^{-1} - A^{-1}EA^{-1} + O(\|E\|^2),$$

which on taking powers gives

$$(A+E)^{1-p} = A^{1-p} - \sum_{r=1}^{p-1} A^{r-p}EA^{-r} + O(\|E\|^2).$$

Equation (3.11) becomes

$$\begin{aligned}\widehat{Y}_{k+1} \;=\; & \frac{1}{p}\left((p-1)[Y_k + \Delta_k] + A\left[Y_k^{1-p} - \sum_{r=1}^{p-1} Y_k^{r-p}\Delta_k Y_k^{-r}\right]\right)\\ & + O(\|\Delta_k\|^2).\end{aligned} \tag{3.12}$$

On subtracting (3.5) from (3.12) we have

$$\Delta_{k+1} = \frac{1}{p}\left((p-1)\Delta_k - A\sum_{r=1}^{p-1} Y_k^{r-p}\Delta_k Y_k^{-r}\right) + O(\|\Delta_k\|^2). \tag{3.13}$$

Using the notation of (3.7) and (3.8), let

$$\widehat{\Delta}_k = Z^{-1}\Delta_k Z$$

and diagonalize (3.13),

$$\widehat{\Delta}_{k+1} = \frac{1}{p}\left((p-1)\widehat{\Delta}_k - \Lambda\sum_{r=1}^{p-1} D_k^{r-p}\widehat{\Delta}_k D_k^{-r}\right) + O(\|\Delta_k\|^2). \tag{3.14}$$

As before, let $D_k = \operatorname{diag}(d_i^{(k)})$ and write $\widehat{\Delta}_k = (\widehat{\delta}_{ij}^{(k)})$, to express (3.14) elementwise as

$$\begin{aligned}\widehat{\delta}_{ij}^{(k+1)} &= \frac{1}{p}\left((p-1)\widehat{\delta}_{ij}^{(k)} - \lambda_i\sum_{r=1}^{p-1}\frac{\widehat{\delta}_{ij}}{\left(d_i^{(k)}\right)^{p-r}\left(d_j^{(k)}\right)^{r}}\right) + O(\|\Delta_k\|^2)\\ &= \pi_{ij}^{(k)}\widehat{\delta}_{ij}^{(k)} + O(\|\Delta_k\|^2), \qquad i,j = 1\colon n,\end{aligned}$$

where

$$\pi_{ij}^{(k)} = \frac{1}{p}\left((p-1) - \lambda_i\sum_{r=1}^{p-1}\frac{1}{\left(d_i^{(k)}\right)^{p-r}\left(d_j^{(k)}\right)^{r}}\right).$$

Since we have assumed that D_k converges to $\Lambda^{1/p}$, we can write

$$d_i^{(k)} = \lambda_i^{1/p} + \epsilon_i^{(k)},$$

where $\epsilon_i^{(k)} \to 0$ as $k \to \infty$. Then

$$\begin{aligned}\pi_{ij}^{(k)} &= \frac{1}{p}\left((p-1) - \lambda_i \sum_{r=1}^{p-1} \frac{1}{\lambda_i^{(p-r)/p}\lambda_j^{r/p}}\right) + O(|\epsilon^{(k)}|) \\ &= \frac{1}{p}\left((p-1) - \sum_{r=1}^{p-1}\left(\frac{\lambda_i}{\lambda_j}\right)^{r/p}\right) + O(|\epsilon^{(k)}|),\end{aligned}$$

where $\epsilon^{(k)} = \max_i |\epsilon_i^{(k)}|$.

For numerical stability of the iteration we require that the error amplification factors $\pi_{ij}^{(k)}$ do not exceed 1 in modulus. Hence we require that

$$\frac{1}{p}\left|(p-1) - \sum_{r=1}^{p-1}\left(\frac{\lambda_i}{\lambda_j}\right)^{r/p}\right| \le 1, \quad i,j = 1\colon n. \tag{3.15}$$

This is a very severe restriction on the matrix A and makes the simplified Newton iteration of little practical use for calculating matrix pth roots. For example, if A is Hermitian positive definite, then in the square root case ($p = 2$) this is equivalent to

$$\kappa_2(A) \le 9,$$

where the condition number $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$. This result was first noted by Laasonen [17] and proved by Higham [10]. For the cube root of a Hermitian positive definite A, (3.15) requires that

$$\kappa_2(A)^{1/3} + \kappa_2(A)^{2/3} \le 5.$$

On solving this quadratic equation, we find that the condition for stability is

$$\kappa_2(A) \le 5.74. \tag{3.16}$$

Clearly as we seek higher order roots the condition for numerical stability becomes more restrictive.

The analysis shows that, depending on the eigenvalues of A, a small perturbation Δ_k in Y_k may cause perturbations of increasing norm in the iterates, resulting in the sequence $\widehat{Y}_k$ diverging from the true sequence Y_k. The loss of stability of the simplified Newton's method is due to the unstable propagation of rounding errors, resulting in a loss of commutativity in the iterates. Hence in simplifying Newton's method, (3.3), to obtain the iterations (3.5) and (3.6), we generally lose the numerical stability of the method.

4. The Schur method. The Newton iterations for computing matrix pth roots considered in section 3 were shown to be of little practical interest due to poor convergence and stability properties. We will overcome these disadvantages by applying a generalization of the direct methods for computing matrix square roots proposed by Björck and Hammarling [1] and Higham [11]. Björck and Hammarling [1] offer a method based on the Schur decomposition and a fast recursion. However, if A is real,

this method may require complex arithmetic even if the desired root we are seeking is itself real. The method of [1] was extended by Higham [11] to compute a real square root of a real matrix using real arithmetic. We will use this technique to derive an algorithm for computing a matrix pth root that uses only real arithmetic if the given matrix is itself real.

To find a pth root X of $A \in \mathbb{R}^{n\times n}$ we first form the real Schur decomposition of A (see [9, p. 341])

$$A = QTQ^T,$$

where T is upper quasi-triangular, each block T_{ii} is either 1×1 or 2×2 with complex conjugate eigenvalues, and Q is real orthogonal. We then find a pth root U of the upper quasi-triangular matrix T, so that

$$U^p = T. \tag{4.1}$$

Finally a pth root X of A is given by

$$X = QUQ^T.$$

Let $R^{(q)}$, $q = 1\colon p-2$, be matrices with the same upper quasi-triangular structure as T such that

$$\begin{array}{rcl} R^{(p-2)} = U^{p-1} & \Rightarrow & UR^{(p-2)} = T, \\ R^{(p-3)} = U^{p-2} & \Rightarrow & UR^{(p-3)} = R^{(p-2)}, \\ \vdots & \vdots & \vdots \\ R^{(2)} = U^3 & \Rightarrow & UR^{(2)} = R^{(3)}, \\ R^{(1)} = U^2 & \Rightarrow & UR^{(1)} = R^{(2)}, \\ R^{(0)} = U & \Rightarrow & UR^{(0)} = R^{(1)}. \end{array} \tag{4.2}$$

Equating (i,j) blocks in the equation $UR^{(p-2)} = T$ we see that, for $i < j$,

$$T_{ij} = \sum_{k=i}^{j} U_{ik}R_{kj}^{(p-2)} = U_{ii}R_{ij}^{(p-2)} + U_{ij}R_{jj}^{(p-2)} + \sum_{k=i+1}^{j-1} U_{ik}R_{kj}^{(p-2)}. \tag{4.3}$$

Similarly for the blocks of the matrices $R^{(q)}$, $q = 1\colon p-2$, in (4.2),

$$R_{ij}^{(q)} = U_{ii}R_{ij}^{(q-1)} + U_{ij}R_{jj}^{(q-1)} + \sum_{k=i+1}^{j-1} U_{ik}R_{kj}^{(q-1)}, \quad \text{where} \quad i < j. \tag{4.4}$$

We are looking to rearrange the expressions of (4.3) and (4.4) in such a way that we can calculate the blocks of the matrices U and $R^{(q)}$, $q = 1\colon p-2$, along one superdiagonal at a time. This can be achieved by first solving (4.1) and (4.2) along the lead diagonal, to give

$$U_{ii} = T_{ii}^{(1/p)}, \quad R_{ii}^{(1)} = U_{ii}^2, \quad \ldots, \quad R_{ii}^{(p-2)} = U_{ii}^{p-1}, \quad 1 \le i \le m.$$

By substituting the expression (4.4) for $R_{ij}^{(q-1)}$ into that of $R_{ij}^{(q)}$, $q = 1\colon p-2$, we are able to find the remaining blocks of the quasi-triangular matrices by moving upwards along the superdiagonals in the order specified by $j-i = 1, 2, \ldots, m-1$. The required form is given in the following lemma.

LEMMA 4.1. *The matrices of* (4.4) *can be expressed as*

$$R_{ij}^{(q)} = \sum_{h=0}^{q} U_{ii}^{q-h} U_{ij} U_{jj}^{h} + \sum_{m=0}^{q-1} U_{ii}^{q-1-m} B_{ij}^{(m)},$$

where

$$B_{ij}^{(m)} = \sum_{k=i+1}^{j-1} U_{ik} R_{kj}^{(m)}.$$

Proof. The proof is by induction. From (4.4) it is clear that the result holds for $q = 1$. Assume that it holds for the first $q-1$ matrices. Then

$$\begin{aligned} R_{ij}^{(q)} &= U_{ii} R_{ij}^{(q-1)} + U_{ij} R_{jj}^{(q-1)} + \sum_{k=i+1}^{j-1} U_{ik} R_{kj}^{(q-1)} \\ &= U_{ii} \left(\sum_{h=0}^{q-1} U_{ii}^{q-1-h} U_{ij} U_{jj}^{h} + \sum_{m=0}^{q-2} U_{ii}^{q-2-m} B_{ij}^{(m)} \right) + U_{ij} R_{jj}^{(q-1)} + B_{ij}^{(q-1)} \\ &= \sum_{h=0}^{q} U_{ii}^{q-h} U_{ij} U_{jj}^{h} + \sum_{m=0}^{q-1} U_{ii}^{q-1-m} B_{ij}^{(m)}, \end{aligned}$$

since $R_{jj}^{(q-1)} = U_{jj}^{q}$. ∎

COROLLARY 4.2. *Equation* (4.3)*, for* $T_{ij}, i < j$*, can be expressed as*

$$T_{ij} = \sum_{h=0}^{p-1} U_{ii}^{p-1-h} U_{ij} U_{jj}^{h} + \sum_{m=0}^{p-2} U_{ii}^{p-2-m} B_{ij}^{(m)}.$$

Proof. Substitute the expression for $R_{ij}^{(p-2)}$ from Lemma 4.1 into (4.3) and collect terms. ∎

We are now in the position to form the matrix pth root U of T, starting with the blocks on the leading diagonal and then moving upwards one superdiagonal at a time. We have

$$U_{ii} = T_{ii}^{(1/p)}, \quad R_{ii}^{(1)} = U_{ii}^{2}, \quad \ldots, \quad R_{ii}^{(p-2)} = U_{ii}^{p-1}, \quad 1 \le i \le m; \tag{4.5}$$

then for $j - i = 1\colon m-1$, we can form

$$T_{ij} = \sum_{h=0}^{p-1} U_{ii}^{p-1-h} U_{ij} U_{jj}^{h} + \sum_{m=0}^{p-2} U_{ii}^{p-2-m} B_{ij}^{(m)}, \quad i < j, \tag{4.6}$$

and

$$R_{ij}^{(q)} = \sum_{h=0}^{q} U_{ii}^{q-h} U_{ij} U_{jj}^{h} + \sum_{m=0}^{q-1} U_{ii}^{q-1-m} B_{ij}^{(m)}, \quad q = 1\colon p-2, \quad i < j. \tag{4.7}$$

We need to solve (4.6) for the blocks U_{ij} of U along one superdiagonal at a time by using only previously computed elements.

ALGORITHM 4.3. *Given an upper triangular quasi-triangular* $T \in \mathbb{R}^{n \times n}$*, this algorithm computes a pth root* U *of the same structure.*

```
Compute U_ii and R_ii^(q), q = 1: p − 2, using (4.5)
for k = 1: n − 1
    for i = 1: n − k
            Solve for U_{i,i+k} in (4.6)
            for q = 1: p − 2
                Compute R_{i,i+k}^(q) from (4.7)
            end
    end
end
```

We can see from (4.5), (4.6), and (4.7) that the matrix pth root U of T is real if and only if each of the blocks U_{ii} is real.

We can compute the principal pth root, U of T, from Algorithm 4.3 provided that each U_{ii} is the principal pth root of the 1×1 or 2×2 matrix T_{ii}. The desired pth roots of a 2×2 matrix can be computed by an extension of the technique of Higham [11, Lemma 2].

Let $T_{ii} \in \mathbb{R}^{2\times 2}$ have complex conjugate eigenvalues $\lambda, \overline{\lambda} = \theta \pm i\mu$, and let

$$W^{-1}T_{ii}W = \operatorname{diag}(\lambda, \overline{\lambda}) = \theta I + i\mu K,$$

where

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

This gives us

$$T_{ii} = \theta I + \mu Z, \tag{4.8}$$

where $Z = iWKW^{-1}$. Since θ and μ are real, it follows that $Z \in \mathbb{R}^{2\times 2}$.

Let $\alpha + i\beta$ be a pth root of $\theta + i\mu$. A pth root of T_{ii} is given by $U_{ii} = WDW^{-1}$, where

$$D = \begin{bmatrix} \alpha + i\beta & 0 \\ 0 & \alpha - i\beta \end{bmatrix}$$

or, alternatively,

$$D = \alpha I + i\beta K.$$

Hence

$$U_{ii} = \alpha I + \beta Z \tag{4.9}$$

is a real pth root of T_{ii} whose complex conjugate eigenvalues $\alpha \pm i\beta$ are the pth roots of the eigenvalues $\theta \pm i\mu$ of T_{ii}.

We now need to compute θ and μ, where $\lambda = \theta + i\mu$ is an eigenvalue of

$$T_{ii} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}.$$

The eigenvalue λ is given by

$$\lambda = \frac{1}{2}\left((t_{11} + t_{22}) + i\sqrt{-(t_{11} - t_{22})^2 - 4t_{12}t_{21}}\right),$$

that is,

$$\theta = \frac{1}{2}\left(t_{11} + t_{22}\right) \quad \text{and} \quad \mu = \frac{1}{2}\sqrt{-\left(t_{11} - t_{22}\right)^2 - 4t_{12}t_{21}}.$$

The next stage requires us to obtain α and β such that $(\alpha + i\beta)^p = \theta + i\mu$. In working out the values α and β it is appropriate to represent λ by its polar coordinates. Namely,

$$(\alpha + i\beta)^p = r\left(\cos\phi + i\sin\phi\right) = re^{i\phi},$$

where $r = \sqrt{\theta^2 + \mu^2}$ and $\phi = \arctan\left(\mu/\theta\right)$. α and β are now easily computed from

$$\alpha = r^{(1/p)}\cos\frac{\phi}{p}, \quad \beta = r^{(1/p)}\sin\frac{\phi}{p}.$$

Finally, the real pth root of T_{ii} is obtained from (4.8) and (4.9):

$$\begin{aligned} U_{ii} &= \alpha I + \beta Z \\ &= \alpha I + \frac{\beta}{\mu}\left(T_{ii} - \theta I\right) \\ &= \begin{bmatrix} \alpha + \frac{\beta}{2\mu}\left(t_{11} - t_{22}\right) & \frac{\beta}{\mu}t_{12} \\ \frac{\beta}{\mu}t_{21} & \alpha - \frac{\beta}{2\mu}\left(t_{11} - t_{22}\right) \end{bmatrix}. \end{aligned}$$

In Algorithm 4.3 we need to solve (4.6), which can be rewritten as

$$\sum_{h=0}^{p-1} U_{ii}^{p-1-h} U_{ij} U_{jj}^{h} = T_{ij} - \sum_{m=0}^{p-2} U_{ii}^{p-2-m} B_{ij}^{(m)}, \quad i < j.$$

Taking the vec of both sides gives

$$\left(\sum_{h=0}^{p-1}\left({U_{jj}}^{hT} \otimes {U_{ii}}^{p-1-h}\right)\right)\operatorname{vec}\left(U_{ij}\right) = \operatorname{vec}\left(T_{ij} - \sum_{m=0}^{p-2} U_{ii}^{p-2-m} B_{ij}^{(m)}\right). \tag{4.10}$$

If U_{ii} is of order y and U_{jj} is of order z, the linear system (4.10) is of order $yz = 1, 2$, or 4 and may be solved using any standard method, provided the coefficient matrix is nonsingular.

THEOREM 4.4. *If $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$ are nonsingular, then the matrix*

$$Y = \sum_{k=0}^{p-1}\left({B^k}^T \otimes A^{p-1-k}\right)$$

is nonsingular, provided that A and $e^{i2\pi q/p}B$, $q = 1\colon p-1$, have no eigenvalues in common.

Proof. Let λ be an eigenvalue of A with corresponding eigenvector u, and let μ be an eigenvalue of B^T with corresponding eigenvector v. For compatible matrices A, B, C, D, we have $(A \otimes B)(C \otimes D) = AC \otimes BD$. Thus

$$\begin{aligned} Y\left(v \otimes u\right) &= \sum_{k=0}^{p-1}\left({B^k}^T v \otimes A^{p-1-k}u\right) \\ &= \sum_{k=0}^{p-1}\left(\mu^k\lambda^{p-1-k}\right)\left(v \otimes u\right). \end{aligned}$$

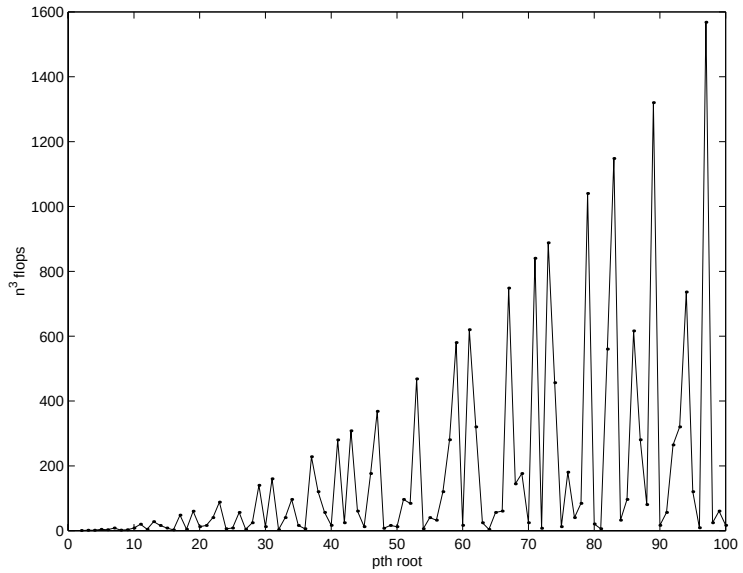


FIG. 4.1. *Operation count associated with computing a pth root of a quasi-triangular matrix by the Schur method when exploiting the factorization of p.*

The matrix Y has eigenvector $v \otimes u$ with associated eigenvalue $\psi = \sum_{k=0}^{p-1} \mu^k \lambda^{p-1-k}$. For Y nonsingular we require $\psi \neq 0$. Clearly for $\lambda = \mu$ we have $\psi = p\lambda^{p-1}$, which is nonzero since A and B are nonsingular. If $\lambda \neq \mu$, then

$$\psi = \frac{\lambda^n - \mu^n}{\lambda - \mu},$$

which is nonzero when $\lambda^n \neq \mu^n$, i.e., $\lambda \neq e^{i2\pi q/p}\mu$, $q = 1\colon p-1$.

It is easy to show that all eigenvalues of Y are of the form ψ. ∎

Therefore, when solving (4.10) we can guarantee the coefficient matrix to be nonsingular by choosing the eigenvalues of U_{ii} and U_{jj} to lie in the same wedge whenever T_{ii} and T_{jj} have eigenvalues in common. As we are usually interested in calculating a root of T that is itself a function of T, the above condition will always be satisfied.

The Schur decomposition can be calculated by numerically stable techniques at a cost of $25n^3$ flops [9, p. 359]. The computation of U as described above requires $p^2n^3/6$ flops and the formation of $X = QUQ^T$ requires $3n^3$ flops. The calculation of U by Algorithm 4.3 requires the formation of $p-2$ intermediary matrices, so for large p the method can be expensive in both computation and storage. By finding the prime factors of p we can form the pth root U by repeatedly applying Algorithm 4.3 over the factors of p. Hence, for highly composite p, we can make considerable computational savings; see Figure 4.1.

Given a matrix A containing real negative eigenvalues, we can find a real odd root of A that is a function of A by using real arithmetic, but this root will not be the principal pth root, as it will have eigenvalues lying in the left half plane. For even p, a real pth root X cannot be computed in real arithmetic since X is real if and only if U_{ii} is real for each i. We now specialize the real Schur method to $A \in \mathbb{C}^{n\times n}$.

Let $A \in \mathbb{C}^{n\times n}$ have the Schur decomposition [9, p. 313], $Q^*AQ = T$. We need

to find a pth root of the strictly upper triangular matrix T. The matrices of (4.2) will also be upper triangular, making (4.5)–(4.7) scalar. This gives us the following recursive formulae for finding a pth root of an upper triangular matrix T.

$$u_{ii} = t_{ii}^{1/p}, \quad r_{ii}^{(1)} = u_{ii}^2, \quad \ldots, \quad r_{ii}^{(p-2)} = u_{ii}^{p-1}.$$

$$\left.\begin{aligned} u_{ij} &= \frac{t_{ij} - \sum_{m=0}^{p-2} u_{ii}^{p-2-m} b_{ij}^{(m)}}{\sum_{h=0}^{p-1} u_{ii}^{p-1-h} u_{jj}^{h}} \\ r_{ij}^{(q)} &= u_{ij} \sum_{h=0}^{q} u_{ii}^{q-h} u_{jj}^{h} + \sum_{m=0}^{q-1} u_{ii}^{q-1-m} b_{ij}^{(m)}, \quad q = 1\colon p-2 \end{aligned}\right\}, \quad i < j, \tag{4.11}$$

where $b_{ij}^{(m)} = \sum_{k=i+1}^{j-1} u_{ik} r_{kj}^{(m)}$. Starting with the leading diagonal, we are able to form the elements of U and $R^{(q)}$ one superdiagonal at a time, as (4.11) uses only previously calculated elements.

4.1. Stability of the Schur method. We consider the numerical stability of the Schur method by examining the rounding error associated with the scalar equations (4.11). We work with the standard model of floating point arithmetic [13, section 2.3]

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta_1) = \frac{x \operatorname{op} y}{1+\delta_2}, \quad |\delta_1|, |\delta_2| \le u, \quad \operatorname{op} = +, -, *, /,$$

where u is the unit roundoff. We define

$$\tilde{\gamma}_k = \frac{cku}{1-cku},$$

where c denotes a small integer constant whose exact value is unimportant. Computed quantities are denoted with a hat. Let

$$|\epsilon_{ij}^{(q)}| = |r_{ij}^{(q)} - fl(r_{ij}^{(q)})|, \quad q = 1\colon p-2.$$

For $q = 1$, equation (4.11) becomes

$$r_{ij}^{(1)} = u_{ij}u_{ii} + u_{ij}u_{jj} + \sum_{k=i+1}^{j-1} u_{ik}u_{kj},$$

which gives

$$\begin{aligned} |\epsilon_{ij}^{(1)}| &\le \gamma_n \left(|\widehat{u}_{ij}||\widehat{u}_{ii}| + |\widehat{u}_{ij}||\widehat{u}_{jj}| + \sum_{k=i+1}^{j-1} |\widehat{u}_{ik}||\widehat{u}_{kj}| \right), \\ &= \gamma_n \sum_{k=i}^{j} |\widehat{u}_{ik}||\widehat{u}_{kj}| = \gamma_n |\widehat{U}|_{ij}^2. \end{aligned} \tag{4.12}$$

The bound of (4.12) is then used in calculating the values $|\epsilon_{ij}^{(q)}|$, $q = 2\colon p-2$, from (4.11) to yield

$$|\epsilon_{ij}^{(q)}| \le \gamma_{qn} |\widehat{U}|_{ij}^{(q+1)}, \quad q = 1\colon p-2. \tag{4.13}$$

We are now in position to examine the rounding errors involved in computing the elements u_{ij} from (4.11). Define

$$|e_{ij}| = |u_{ij} - fl(u_{ij})|,$$

giving

$$|e_{ij}| \leq \gamma_{np}\Bigg(|\widehat{u}_{ij}||\widehat{r}_{ii}^{(p-2)}| + |\widehat{r}_{ii}^{(p-3)}| \sum_{k=i+1}^{j} |\widehat{u}_{ik}||\widehat{u}_{kj}| + |\widehat{r}_{ii}^{(p-4)}| \sum_{k=i+1}^{j} |\widehat{u}_{ik}||\widehat{r}_{kj}^{(1)}| + \cdots + \sum_{k=i+1}^{j} |\widehat{u}_{ik}||\widehat{r}_{kj}^{(p-2)}|\Bigg). \tag{4.14}$$

Finally, using the results of (4.13) in (4.14) renders the bound

$$|e_{ij}| \leq \tilde{\gamma}_{pn}|\widehat{U}|_{ij}^{p},$$

that is

$$|E| \leq c\,p\,n\,u\,|\widehat{U}|^{p}, \tag{4.15}$$

where $E = (e_{ij})$.

By considering

$$\beta(U) = \frac{\|U\|_F^p}{\|T\|_F} \geq 1,$$

we can see that the Schur method is stable, provided $\beta(U)$ is sufficiently small. An analogous result can be found for the backward error of the real Schur algorithm. This generalizes the analysis of [1] and [11] for computing a matrix square root to the problem of the matrix pth root.

If we let X be the exact pth root of A and $\widehat{X}$ be the matrix X rounded to working precision, we see that $\widehat{X}$ satisfies a bound that is essentially the same as (4.15). Hence the bound (4.15) is as good an approximation as we can expect for computing the matrix pth root when working in finite precision arithmetic. However, there exists matrices for which $\beta(U)$ can be large, signaling that the problem in calculating root U is inherently ill conditioned. Therefore it is wise to return the value β whenever implementing Algorithm 4.3.

5. Numerical experiments. All our computations have been done using MATLAB with unit roundoff $u = 2^{-53} \approx 1.1\times10^{-16}$. We use matrices from the Test Matrix Toolbox [12]. For the iterative methods we use A as our starting matrix and report the relative differences

$$\text{reldiff}(X_k) = \frac{\|X_k - X_{k-1}\|_2}{\|X_k\|_2}$$

and the residuals

$$\text{res}(X_k) = \frac{\|X_k^p - A\|_2}{\|A\|_2}.$$

The first example uses the 2×2 symmetric positive definite Lehmer matrix. This is a well-conditioned matrix, with condition number

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = 3.0.$$

TABLE 5.1
Convergence behavior of Newton's method and its variant for the cube root of a Lehmer*(2) matrix.*

	Newton's method (3.3)		Iteration (3.5)	
Iter	reldiff	res	reldiff	res
1	4.09e-01	3.33e-01	4.09e-01	3.33e-01
2	1.45e-01	5.25e-02	1.45e-01	5.25e-02
3	3.30e-02	2.34e-03	3.30e-02	2.34e-03
4	1.62e-03	5.45e-06	1.62e-03	5.45e-06
5	3.78e-06	2.97e-11	3.78e-06	2.97e-11
6	2.06e-11	7.40e-17	2.06e-11	2.47e-16
7	4.85e-17	1.48e-16	2.07e-16	2.64e-16

TABLE 5.2
Results for principal fifth root of minij(5).

	Newton's method (3.3)		Iteration (3.5)	
Iter	reldiff	res	reldiff	res
1	8.55e-01	8.97e+03	9.73e-01	8.97e+03
9	2.84e-01	1.18e+00	2.32e-01	1.18e+00
12	1.33e-01	3.42e-02	1.14e-01	3.42e-02
14	2.87e-02	7.02e-04	2.48e-02	7.02e-04
17	6.65e-09	8.92e-17	5.72e-09	1.23e-10
18	2.14e-16	1.34e-16	4.21e-10	6.82e-10
23	–	–	2.23e-06	3.60e-06
29	–	–	6.53e-02	1.06e-01

The stability condition (3.16) for a cube root of a symmetric positive definite matrix is satisfied by the Lehmer matrix, so the simplified Newton iteration (3.5) is numerically stable. Table 5.1 shows that Newton's method (3.3) and the simplified iteration (3.5) both converge to a positive definite cube root of A after 7 iterations.

For the next example we find the fifth root of the 5×5 minij matrix A, whose elements are given by $A(i,j) = \min(i,j)$. The condition number of the Hermitian positive definite matrix A is $\kappa_2(A) = 45.4552$, which does not satisfy the stability condition (3.15). As A is positive definite, we would expect convergence to the positive definite fifth root, but iteration (3.5) fails to converge due to the unstable propagation of rounding errors which bring loss of commutativity to the iterates. However, the full Newton iteration (3.3) converges to a fifth root after 18 iterations; see Table 5.2.

Let us now consider finding the 4th root via the Schur method of the matrix

$$T = \begin{bmatrix} 1.0000 & -1.0000 & -1.0000 & -1.0000 \\ 0 & 1.3000 & -1.0000 & -1.0000 \\ 0 & 0 & 1.7000 & -1.0000 \\ 0 & 0 & 0 & 2.0000 \end{bmatrix}.$$

We know from Theorem 2.2 that T has $4^4 = 256$ 4th roots that are functions of T and hence upper triangular. These roots yield different β values. For example, the principal root,

$$U = \begin{bmatrix} 1.0000 & -0.2260 & -0.2609 & -0.3058 \\ 0 & 1.0678 & -0.1852 & -0.2125 \\ 0 & 0 & 1.1419 & -0.1578 \\ 0 & 0 & 0 & 1.1892 \end{bmatrix},$$

has $\beta(\widehat{U}) = 6.7854$ and $\mathrm{res}(\widehat{U}) = 2.2288 \times 10^{-16}$. In contrast, the root

$$U = \begin{bmatrix} 1.0000 & 6.8926 & 17.5356 & 37.6656 \\ 0 & -1.0678 & -5.5241 & -18.8185 \\ 0 & 0 & 1.1419 & 7.7702 \\ 0 & 0 & 0 & -1.1892 \end{bmatrix}$$

has $\beta(\widehat{U}) = 1.2526 \times 10^{6}$ and $\mathrm{res}(\widehat{U}) = 1.2324 \times 10^{-14}$. This illustrates that the Schur method returns a pth root of a matrix near A as long as $\beta(U)$ is not too large.

Acknowledgment. I would like to thank Professor Nick Higham for many interesting and helpful discussions concerning matrix roots.

REFERENCES

[1] Å. Björck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.
[2] Ç. K. Koç and B. Bakkaloğlu, *Halley's method for the matrix sector function*, IEEE Trans. Automat. Control, 40 (1995), pp. 944–949.
[3] 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.
[4] J. H. Curry, L. Garnett, and D. Sullivan, *On the iteration of a rational function: Computer experiments with Newton's method*, Comm. Math. Phys., 91 (1983), pp. 267–277.
[5] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
[6] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[7] P. Fatou, *Sur les équations fonctionnelles*, Bull. Soc. Math. France, 47 (1919), pp. 161–271.
[8] F. R. Gantmacher, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.
[9] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[10] N. J. Higham, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–549.
[11] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.
[12] N. J. Higham, *The Test Matrix Toolbox for MATLAB (Version 3.0)*, Numerical Analysis Report 276, Manchester Centre for Computational Mathematics, Manchester, England, 1995.
[13] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[14] W. D. Hoskins and D. J. Walton, *A faster, more stable method for computing the pth roots of positive definite matrices*, Linear Algebra Appl., 26 (1979), pp. 139–163.
[15] G. Julia, *Mémoire sur l'itération des fonctions rationelles*, J. Math. Pure Appl., 8 (1918), pp. 47–245.
[16] C. Kenney and A. J. Laub, *Padé error estimates for the logarithm of a matrix*, Internat. J. Control, 50 (1989), pp. 707–730.
[17] P. Laasonen, *On the iterative solution of the matrix equation $AX^2 - I = 0$*, Math. Tables Aids Comput., 12 (1958), pp. 109–116.
[18] J. M. Ortega, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.
[19] L. S. Shieh, Y. T. Tsay, and C. T. Wang, *Matrix sector functions and their applications to systems theory*, Proc. IEE-D, 131 (1984), pp. 171–181.
[20] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, London, 1973.
[21] F. Tisseur, *Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1038–1057.
[22] E. R. Vrscay, *Julia sets and Mandlebrot-like sets associated with higher order Schröder rational iteration functions: A computer assisted study*, Math. Comp., 46 (1996), pp. 151–169.
[23] J. H. M. Wedderburn, *Lectures on Matrices*, AMS, Providence, RI, 1934.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 990–1002

MATRICES WITH SPECIAL SIGN PATTERNS OF SIGNED GENERALIZED INVERSES*

JIA-YU SHAO†, JIN-LING HE†, AND HAI-YING SHAN†

Abstract. A real matrix A is said to have a *signed generalized inverse* (GI) if the sign pattern of its GI A^+ is uniquely determined by the sign pattern of A. We characterize those sign pattern matrices with a signed GI, and the GI of it is nonnegative, or is positive, or has no zeros.

Key words. matrix, sign, generalized inverse

AMS subject classifications. 15A09, 15A48

PII. S0895479802401485

1. Introduction.

The sign pattern of a real matrix A, denoted by $\mathrm{sgn}A$, is the $(0, 1, -1)$-matrix obtained from A by replacing each entry with its sign. The set of real matrices with the same sign pattern as A is called the *qualitative class of* A and is denoted by $Q(A)$.

A square real matrix is called a *strong sign nonsingular (S^2NS) matrix* if each matrix in $Q(A)$ is nonsingular(i.e., invertible) and the inverses of the matrices in $Q(A)$ all have the same sign pattern.

Let A be an $m \times n$ real matrix. An $n \times m$ real matrix X is the *generalized inverse* (GI) of A (or *Moore–Penrose inverse of A;* see [5]) if X satisfies the following four conditions:

$$AXA = A, \qquad XAX = X, \qquad (AX)^T = AX, \qquad (XA)^T = XA.$$

It is well known that for each matrix A its GI exists and is unique. The GI of A is denoted by A^+. If A is an invertible square matrix, then clearly $A^+ = A^{-1}$.

DEFINITION 1.1. *A real matrix A is said to have a signed GI (or simply say that "A^+ is signed") if $sgnB^+ = sgnA^+$ for each matrix B in $Q(A)$.*

The notion of matrices having signed GI was first introduced in [2] and [6] in the study of the least square sign solvability of linear systems of equations. It is obviously a generalization of the notion of S^2NS matrices, since each S^2NS matrix clearly has a signed GI.

A matrix is said to be *totally nonzero* if it contains no zero entries.

DEFINITION 1.2. *A real matrix A is said to have a nonnegative (or nonpositive, or negative, or positive, or totally nonzero, respectively) signed GI if A has a signed GI and A^+ is nonnegative (or nonpositive, or negative, or positive, or totally nonzero).*

Two $m \times n$ real matrices A and B are said to be permutation equivalent if A can be transformed to B by permuting its rows and columns. It is easy to verify that the property of having a signed GI (or nonnegative, or nonpositive, or negative, or positive, or totally nonzero signed GI) for a matrix is preserved under permutation equivalences.

*Received by the editors January 25, 2002; accepted for publication (in revised form) by H. Woerdeman August 8, 2002; published electronically February 25, 2003. This research was supported by NNSF of China 19831050 and RFDP of China 2000024705.

http://www.siam.org/journals/simax/24-4/40148.html

†Department of Applied Mathematics, Tongji University, Shanghai 200092, China (jyshao@sh163.net, hejinling1@163.com, shan_haiying@sohu.com).

In fact, if P and Q are permutation matrices, then $(PAQ)^+ = Q^T A^+ P^T$. Thus, if A and B are permutation equivalent, then so are A^+ and B^+.

In this paper, we give complete characterizations for the matrices to have nonpositive (or nonnegative), negative (or positive), or totally nonzero signed GIs.

We first consider S^2NS matrices in section 2 and then consider the general cases in sections 3 and 4.

2. S^2NS matrices with special inverse sign patterns. In this section we characterize the S^2NS matrices with nonnegative (respectively, positive or totally nonzero) inverse sign patterns. In order to derive these results, we need to use some graph theoretical concepts and techniques.

A *signed digraph* S is a digraph in which each of its arcs is assigned a sign $+1$ or -1. The *sign* of a subdigraph S_1 of S is defined to be the product of the signs of all the arcs of S_1.

A signed digraph S is called an S^2NS signed digraph if S satisfies the following two conditions:

(1) The sign of each cycle of S is negative.

(2) Each pair of paths in S with the same initial vertex and the same terminal vertex has the same sign.

Let $A = (a_{ij})$ be a square real matrix of order n. The associated digraph $D(A)$ of A is defined to be the digraph with the vertex set $V = \{1, 2, \ldots, n\}$ and arc set $E = \{(i,j) | a_{ij} \neq 0, i \neq j\}$. The associated signed digraph $S(A)$ of A is obtained from $D(A)$ by assigning the sign of a_{ij} to each arc (i,j) in $D(A)$.

The following is a characterization of S^2NS matrices and is proved in [1] and [2].

LEMMA 2.A. *Let A be a square real matrix, all of whose diagonal entries are negative. Then A is an S^2NS matrix if and only if its associated signed digraph $S(A)$ is an S^2NS signed digraph.*

We also need the following result from [2, Lem. 3.2.4, Thm. 3.2.5] (also see [3]).

LEMMA 2.B. *Let A be an S^2NS matrix of order n, all of whose diagonal entries are negative. Let $(A^{-1})_{ij}$ be the (i,j) entry of the inverse matrix A^{-1}. Then we have that*

(1) $(A^{-1})_{ii} < 0$ $(i = 1, \ldots, n)$;

(2) *if $i \neq j$, then $(A^{-1})_{ij} \neq 0$ if and only if there exists a path from vertex i to vertex j in the associated digraph $D(A)$;*

(3) *if $i \neq j$ and $(A^{-1})_{ij} \neq 0$, then $sgn(A^{-1})_{ij} = -\epsilon$, where ϵ is the common sign of all the paths in $S(A)$ from vertex i to vertex j.*

A square matrix A of order n is *fully indecomposable* if A does not contain a nonvacuous zero submatrix whose number of rows and number of columns sum to n. It is a well-known fact (see [4]) that if A is a square matrix of order n, all of whose diagonal entries are nonzero, then its associated digraph $D(A)$ is strongly connected if and only if A is fully indecomposable. From this fact and Lemma 2.B, the following characterization from [2] and [3] follows directly.

THEOREM 2.A. *Let A be an S^2NS matrix. Then A^{-1} is totally nonzero if and only if A is fully indecomposable.*

Now we consider the characterizations of the S^2NS matrices with nonnegative (or nonpositive) inverses.

THEOREM 2.1. *Let A be a square real matrix of order n. Then A is an S^2NS matrix with $A^{-1} \leq 0$ if and only if A is permutation equivalent to a matrix of the following lower triangular form:*

$$\begin{pmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}, \tag{2.1}$$

where $b_{ii} < 0$ $(i = 1, \ldots, n)$ and $b_{ij} \geq 0$ for all $1 \leq j < i \leq n$.

Proof.

Sufficiency. Without loss of generality, we may assume that A equals the matrix in (2.1). Let $S(A)$ be the associated signed digraph of A. Then it is easy to see from (2.1) that $S(A)$ contains no cycle and the sign of each arc of $S(A)$ is positive. It follows from the definition that $S(A)$ is an S^{2}NS signed digraph and thus A is an S^{2}NS matrix.

Now we show that $A^{-1} \leq 0$. From Lemma 2.B we have $(A^{-1})_{ii} < 0$. Also since the sign of each path in $S(A)$ is positive, we have $(A^{-1})_{ij} \leq 0$ for all $i \neq j$ by Lemma 2.B. Thus we have $(A^{-1}) \leq 0$ as desired.

Necessity. By suitably permuting the rows and columns of A, we may assume that all the diagonal entries of A are nonzero. Take a diagonal matrix D with all the diagonal entries in $\{1, -1\}$ such that all the diagonal entries of the S^{2}NS matrix DA are negative. Write $A_1 = DA$. Then all the diagonal entries of the inverse matrix A_1^{-1} are also negative by Lemma 2.B. Now $A_1^{-1}D = A^{-1} \leq 0$, so all the diagonal entries of D are 1. Thus D is the identity matrix, $A = A_1$, and thus all the diagonal entries of A are negative.

By Lemma 2.A, the associated signed digraph $S(A)$ of the S^{2}NS matrix A is an S^{2}NS signed digraph. Also the sign of every path of $S(A)$ is positive by Lemma 2.B and the assumption $A^{-1} \leq 0$. Thus the sign of every arc of $S(A)$ is positive, which implies that each off-diagonal entry of A is nonnegative. Finally, $S(A)$ must be an acyclic digraph (otherwise, the sign of a cycle in $S(A)$ would be positive, contradicting the fact that $S(A)$ is an S^{2}NS signed digraph). Thus A can be transformed into a lower triangular matrix of the form (2.1) by simultaneously permuting its rows and columns, where each diagonal entry $b_{ii} < 0$ and each off-diagonal entry $b_{ij} \geq 0$. ∎

Corollary 2.1 below now follows easily from Theorem 2.1, which claims that the only S^{2}NS matrices with positive inverses are the positive matrices of order one.

COROLLARY 2.1. *There does not exist an S^2NS matrix A of order $n \geq 2$ such that $A^{-1} < 0$ (or $A^{-1} > 0$).*

Proof. Suppose A is an S^{2}NS matrix of order $n \geq 2$ with $A^{-1} < 0$. Then by Theorem 2.1, A is permutation equivalent to a lower triangular matrix. Thus A^{-1} also is a lower triangular matrix and contains some zero entries, which is a contradiction. ∎

3. Matrices with nonnegative and positive signed GIs. In this section we give characterizations of matrices with nonnegative and positive signed GIs. Without loss of generality, we always assume that A is an $m \times n$ matrix with $n \leq m$ (otherwise we may consider A^T instead of A). First we need to introduce some concepts and quote some preliminary results, most of which will be used several times in the proofs of our main results in sections 3 and 4.

DEFINITION 3.1 (see [7]). *Let a, b be two real numbers, and let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ real matrices.*

(1) *We say that b is sign majorized by a, denoted by $b \preceq a$, if $b = 0$ or $sgnb = sgna$.*

(2) *We say that B is sign majorized by A, denoted by $B \preceq A$, if $b_{ij} \preceq a_{ij}$ for each $i = 1, \ldots, m$ and $j = 1, \ldots, n$.*

It is easy to see that $B \preceq A$ if and only if B can be obtained from some $\widetilde{A} \in Q(A)$ by replacing some nonzero entries of $\widetilde{A}$ by zero.

The term rank of a matrix A, denoted by $\rho(A)$, is the maximal cardinality of the sets of nonzero entries of A, no two of which lie on the same row or same column. The matrix A is said to have *"full row (or column) term rank"* if $\rho(A)$ is equal to the number of rows (or columns) of A.

The next three results are from [7] and will be used in the characterizations.

THEOREM 3.A. *Let A and B be two matrices with $B \preceq A$ and $\rho(B) = \rho(A)$. If A has a signed GI, then B also has a signed GI and $B^+ \preceq A^+$.*

LEMMA 3.A. *Let A be an $m \times n$ matrix with $\rho(A) < n \leq m$. Then A is permutation equivalent to a matrix of the form $\begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$, where B has full column term rank and D has full row term rank.*

THEOREM 3.B. *Let $A = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$, where B has full column term rank and D has full row term rank. Then A has a signed GI if and only if A satisfies the following two conditions:*

(1) *Both B and D have signed GIs.*

(2) *$sgn(\widetilde{D}^+\widetilde{C}\widetilde{B}^+) = sgn(D^+CB^+)$ for all $\widetilde{B} \in Q(B), \widetilde{C} \in Q(C)$, and $\widetilde{D} \in Q(D)$.*

Also in this case we have

$$A^+ = \begin{pmatrix} B^+ & 0 \\ -D^+CB^+ & D^+ \end{pmatrix}. \tag{3.1}$$

Now we need to introduce the following three types of matrices: A matrix is of type T1 if it is a column of size at least two with no zero entries, of type T2 if it is an S^2NS matrix, and of type T3 if it has the same zero pattern as the vertex-edge incidence matrix of a tree. We also call a matrix of type T3 a tree matrix. This tree is usually denoted by $T(A)$.

Note that a square nonzero matrix of order one is considered to be of type T2, not of type T1, while a 2×1 matrix with no zero entries can be considered to be of both types T1 and T3.

Obviously, each matrix of type T1 or T2 has a signed GI, and the fact that each matrix of type T3 has a signed GI is proven in [6].

DEFINITION 3.2 (see [8]). *Let A be a matrix of the following block partitioned form:*

$$\begin{pmatrix} A_1 & O & \cdots & O \\ B_{21} & A_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \cdots & A_k \end{pmatrix}. \tag{3.2}$$

Then

(1) *A is a T-type matrix if each diagonal block A_i is a matrix of type* T1, T2, or T3 *$(i = 1, \ldots, k)$.*

(2) *A is a T_{13}-type matrix if each diagonal block A_i is a matrix of type* T1 or T3 *$(i = 1, \ldots, k)$.*

Result (1) of Theorem 3.C below is from [2] and [6], while result (2) of Theorem 3.C is from [8].

THEOREM 3.C. *Let A be an $m \times n$ matrix with no zero rows and $\rho(A) = n$. Suppose that A has a signed GI; then we have that*

(1) *A is permutation equivalent to a T-type matrix.*

(2) *A is permutation equivalent to a T-type matrix of the form $\begin{pmatrix} X & 0 \\ Z & Y \end{pmatrix}$, where X is a T_{13}-type matrix and Y is an S^2NS matrix.*

A matrix A is said to be the direct sum of two matrices B and C if $A = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix}$. It is obvious that if A is the direct sum of B and C, then A^+ is the direct sum of B^+ and C^+.

LEMMA 3.1. *Let A be a blocked partitioned matrix, whose diagonal blocks are denoted by $A_1, \ldots, A_k$ and whose off-diagonal blocks are denoted by $B_{i,j}$ $(1 \le i \le k, 1 \le j \le k$ and $i \ne j)$. Suppose each diagonal block A_i has full column term rank $(i = 1, \ldots, k)$ and A has a nonnegative (or nonpositive) signed GI. Then each A_i also has a nonnegative (or nonpositive) signed GI.*

Proof. Let B be the matrix obtained from A by replacing all the off-diagonal blocks with zeros. Then $B \preceq A$ and $\rho(B) = \rho(A)$. Thus B also has a nonnegative (or nonpositive) signed GI by Theorem 3.A. Since B is a direct sum of $A_1, \ldots, A_k$, it follows that each A_i also has a nonnegative (or nonpositive) signed GI. ∎

LEMMA 3.2. *Let A be an $(n+1) \times n$ T3-type matrix (tree matrix) with $n \ge 2$. Then A^+ contains both positive and negative entries.*

Proof. The proof is by induction on n. If $n = 2$, then we may assume that

$$A = \begin{pmatrix} a & 0 \\ b & c \\ 0 & d \end{pmatrix},$$

where $abcd \ne 0$. By direct computations we can verify that (e.g., see [2, p. 273])

$$sgn(A^+) = sgn \begin{pmatrix} a & b & -bcd \\ -abc & c & d \end{pmatrix}$$

from which we can see that the product of the four entries in the first two columns of A^+ is negative. Thus A^+ contains both positive and negative entries.

Now assume that $n \ge 3$ and proceed by induction on n. By suitable row and column permutations we may assume that

$$A = \begin{pmatrix} a & 0 \cdots 0 \\ \beta & A_1 \end{pmatrix}, \tag{3.3}$$

where a is a nonzero number and A_1 is a tree matrix containing at least two columns. By induction, A_1^+ contains both positive and negative entries. Suppose to the contrary that A^+ does not contain both positive and negative entries, say, A^+ does not contain negative entries. Then A has a nonnegative signed GI (since A is a tree matrix), and thus by Lemma 3.1 we are led to the contradiction that A_1 also has a nonnegative signed GI. ∎

Lemma 3.3 below shows that if a T_{13}-type matrix A has a nonpositive signed GI, then A is a direct sum of several negative columns.

LEMMA 3.3. *Let A be a T_{13}-type matrix as in* (3.2). *Then A has a signed GI with $A^+ \le 0$ if and only if each diagonal block A_i is of type* T1 *with $A_i < 0$ $(i = 1, \ldots, k)$ and each off-diagonal block $B_{ij} = 0$ $(1 \le j < i \le k)$.*

Proof. The sufficiency is obvious and we now prove the necessity. By Lemma 3.1, each diagonal block A_i also has a nonpositive signed GI. Thus by Lemma 3.2, A_i must be of type T1 with $A_i < 0$ $(i = 1, \ldots, k)$.

Now take any off-diagonal block B_{ij} $(1 \le j < i \le k)$. Write $B = \begin{pmatrix} A_j & 0 \\ B_{ij} & A_i \end{pmatrix}$. By Lemma 3.1, we can derive that B has a signed GI and $B^+ \le 0$. Suppose that $B_{ij} \neq 0$. Then since we have already shown that both A_i and A_j are matrices of type T1 (containing at least two rows), there exist a 3×2 submatrix C of B and a 3×2 tree matrix D such that $D \preceq C$. Thus we have $B^+ \le 0 \Longrightarrow C^+ \le 0 \Longrightarrow D^+ \le 0$ by Theorem 3.A, contradicting Lemma 3.2. Hence we have $B_{ij} = 0$. □

LEMMA 3.4. *Let $A = (a_{ij})$ be an $m \times n$ nonnegative matrix with no zero columns and $C = (c_{ij})$ be a $k \times h$ nonnegative matrix with no zero rows. Suppose $B = (b_{ij})$ is an $n \times k$ real matrix containing some positive entry $b_{pq} > 0$. Then there exists some $\widetilde{B} \in Q(B)$ such that $A\widetilde{B}C$ contains some positive entry.*

Proof. Take $\widetilde{B} = (\widetilde{b_{ij}}) \in Q(B)$ such that

$$\widetilde{b_{ij}} = \begin{cases} b_{ij} & \text{if } (i,j) = (p,q), \\ \epsilon b_{ij} & \text{if } (i,j) \neq (p,q), \end{cases}$$

where ϵ is a positive number. Take an index i with $a_{ip} > 0$ and an index j with $c_{qj} > 0$. Then we have

$$(A\widetilde{B}C)_{ij} = \sum_{r=1}^{n} \sum_{s=1}^{k} a_{ir} \widetilde{b_{rs}} c_{sj} = a_{ip} b_{pq} c_{qj} + \epsilon a,$$

where a is a constant independent of ϵ. Thus for ϵ sufficiently small $(A\widetilde{B}C)_{ij}$ is positive. □

LEMMA 3.5. *Let $A = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$, where B has full column term rank and D has full row term rank. Then A has a nonpositive signed GI if and only if*

(1) *both B and D have nonpositive signed GIs, and*

(2) $C \ge 0$.

Proof. The sufficiency part follows from Theorem 3.B. We now prove the necessity part. (1) also follows from Theorem 3.B. For (2), let $X = -D^+ \ge 0$, $Z = -B^+ \ge 0$, and $Y = -C$. Suppose to the contrary that Y contains some positive entry. Then by Lemma 3.4, there exists some $\widetilde{Y} = -\widetilde{C} \in Q(Y)$ (where $\widetilde{C} \in Q(C)$) such that $X\widetilde{Y}Z$ contains some positive entry. Thus $-D^+\widetilde{C}B^+ = X\widetilde{Y}Z$ contains some positive entry. Now let $\widetilde{A} = \begin{pmatrix} B & 0 \\ \widetilde{C} & D \end{pmatrix} \in Q(A)$. Then $\widetilde{A}^+ = \begin{pmatrix} B^+ & 0 \\ -D^+\widetilde{C}B^+ & D^+ \end{pmatrix}$ contains some positive entry, which is a contradiction. So we have $C \ge 0$. □

The next result characterizes those matrices with full column term rank that have nonpositive signed GIs.

THEOREM 3.1. *Let A be an $m \times n$ real matrix with no zero rows and full column term rank $\rho(A) = n$. Then A has a nonpositive signed GI if and only if A is permutation equivalent to a matrix of the form $\begin{pmatrix} X & 0 \\ Z & Y \end{pmatrix}$ satisfying the following:*

(1) *X is a direct sum of negative columns of sizes at least two.*

(2) *Y is a square lower triangular matrix whose diagonal entries are all negative and whose off-diagonal entries are all nonnegative.*

(3) $Z \ge 0$.

Proof. The sufficiency follows directly from Lemma 3.5, Lemma 3.3, and Theorem 2.1. We now prove the necessity. From Theorem 3.C we know that A is permutation

equivalent to a T-type matrix of the form $\begin{pmatrix} X & 0 \\ Z & Y \end{pmatrix}$, where X is a T_{13}-type matrix and Y is an S^{2}NS matrix. Now conditions (1), (2), and (3) follow directly from Lemma 3.5, Lemma 3.3, and Theorem 2.1. □

Notice that the matrix $\begin{pmatrix} X & 0 \\ Z & Y \end{pmatrix}$ in Theorem 3.1 can be written in the following form:

$$\begin{pmatrix} A_1 & 0 & \cdots & O & & & & \\ 0 & A_2 & \cdots & 0 & & & & \\ \vdots & \vdots & \ddots & \vdots & & O & & \\ O & 0 & \cdots & A_r & & & & \\ & & & & b_1 & 0 & \cdots & O \\ & & & & +_0 & b_2 & \cdots & 0 \\ & & +_0 & & \vdots & \vdots & \ddots & \vdots \\ & & & & +_0 & +_0 & \cdots & b_s \end{pmatrix}, \tag{3.4}$$

where $A_1, \ldots, A_r$ are negative columns of sizes at least two, $b_1, \ldots, b_s$ are negative numbers, and "$+_0$" means "nonnegative."

Now the characterization for a general real matrix A to have nonpositive signed GI can be obtained by using Lemma 3.5, Theorem 3.1, and an extended version of Lemma 3.A. In fact, if we allow B or D in Lemma 3.A to be vacuous, then we can extend Lemma 3.A to say that any matrix A is permutation equivalent to a matrix of the form $\begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$, where B has full column term rank and D has full row term rank (we allow the special cases $A = B$ and $A = D$). Thus by using Lemma 3.5 and Theorem 3.1, we have the following characterization.

THEOREM 3.2. *Let A_0 be the submatrix of a matrix A obtained by deleting all the zero rows and zero columns of A. Then A has a nonpositive signed GI if and only if A_0 is permutation equivalent to a matrix of the form $\begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$ such that*

(1) *both B and D^T are of the form* (3.4).

(2) $C \geq 0$.

Proof. Obviously A has a nonpositive signed GI if and only if A_0 has. Then the case where A_0 has full row term rank follows from the extended version of Lemma 3.A and Theorem 3.1, while the case where A_0 does not have full row term rank follows from Lemmas 3.A and 3.5. □

Finally, we characterize those matrices having positive (or negative) signed GIs.

THEOREM 3.3. *Let A be an $m \times n$ real matrix. Then A has a negative signed GI if and only if A is a negative row or a negative column.*

Proof. The sufficiency part is obvious, so we need only prove the necessity part. Without loss of generality, we may assume that $n \leq m$. Also A contains no zero rows since $A^+ < 0$. If $\rho(A) < n \leq m$, then by Lemma 3.A, A is permutation equivalent to a matrix of the form $\begin{pmatrix} B & 0 \\ C & D \end{pmatrix}$ and A^+ contains some zero entries by Theorem 3.B, contradicting $A^+ < 0$. Thus we have $\rho(A) = n$. Then by Theorem 3.1, A is permutation equivalent to a matrix B of the form $\begin{pmatrix} X & 0 \\ Z & Y \end{pmatrix}$ satisfying the three conditions in Theorem 3.1. If both X and Y are nonvacuous, then B^+ also contains some zero entries by Theorem 3.B, contradicting $A^+ < 0$. Thus either $B = X$ or $B = Y$. If $B = X$, then B, and hence A, must be a negative column by condition (1) of Theorem 3.1 and $B^+ < 0$. If $B = Y$, then B, and hence A, must be a square negative matrix of order one by condition (2) of Theorem 3.1 and $B^+ < 0$. □

4. Matrices with totally nonzero signed GIs. In this section we give characterizations of matrices with totally nonzero signed GIs.

Let $[m] = \{1, \dots, m\}$ and $[n] = \{1, \dots, n\}$. Let A be an $m \times n$ matrix. If S is a subset of $[m]$ and T is a subset of $[n]$, then $A[S|T]$ denotes the submatrix of A whose rows have index in S and whose columns have index in T. The complement of T in $[n]$ is denoted by $\overline{T}$. If $T = [n]$, we abbreviate $A[S|T]$ to $A[S| :]$.

For convenience, we use the notation $(A)_{ij}$ to denote the (i, j) entry of the matrix A.

We first introduce three types of matrices called CR-matrices, CC-matrices, and RR-matrices (where C stands for "column" and R stands for "row") in Definitions 4.1–4.3 below.

DEFINITION 4.1. *An $m \times n$ matrix A is called a CR-matrix if for each position (p, q) of A there exists a square submatrix $B = A[T| :]$ of order n with $p \in T$ such that both B and its submatrix $A[T\backslash\{p\}|\overline{\{q\}}]$ have full row term ranks.*

Theorem 4.1 below gives an important relation between CR-matrices and matrices with totally nonzero signed GIs.

THEOREM 4.1. *Let A be an $m \times n$ matrix having a signed GI and $\rho(A) = n \le m$. Then A^+ contains no zero entries if and only if A is a CR-matrix.*

Proof.

Necessity. Suppose A is not a CR-matrix. Then there exists a position (p, q) such that for each $T \subseteq [m]$ with $p \in T$ and $|T| = n$, we have

$$\det A[T| :] \cdot \det A[T\backslash\{p\}|\overline{\{q\}}] = 0.$$

Since A has a signed GI and $\rho(A) = n$, A is an L-matrix by [7, Lem. 2.B]. Now use the following formula for $(A^+)_{qp}$ in [5] and [6]:

$$(A^+)_{qp} = \frac{(-1)^{q+1}}{\det(A^T A)} \sum_{T \subseteq [m], p \in T, |T| = n} (-1)^{inv(p,T)} \det A[T| :] \det A[T\backslash\{p\}|\overline{\{q\}}].$$

We then have $(A^+)_{qp} = 0$, which is a contradiction.

Sufficiency. We first show that $(A^+)_{11} \neq 0$. By the CR-property of A, there exists $T \subseteq [m]$ with $1 \in T$ and $|T| = n$ such that both of the two square matrices $A[T| :]$ (of order n) and $A[T\backslash\{1\}|\overline{\{1\}}]$ (of order $n - 1$) have full row term rank. It follows from [2, Thm. 11.2.10] and from Theorem 3.A that $A[T| :]$ is an S^2NS matrix, since A has a signed GI. Thus $A[T\backslash\{1\}|\overline{\{1\}}]$ is an SNS matrix (since it is a submatrix of the S^2NS matrix $A[T| :]$ of order $n - 1$ with full column term rank), and so $(A[T| :]^{-1})_{11} \neq 0$. For convenience, we assume that $T = [n]$. Let A_T be the matrix obtained from A by replacing all the entries not in the first n rows with zeros, namely,

$$A_T = \begin{pmatrix} A[T| :] \\ 0 \end{pmatrix}.$$

Then $A_T^+ = \begin{pmatrix} A[T| :]^+ & 0 \end{pmatrix} = \begin{pmatrix} A[T| :]^{-1} & 0 \end{pmatrix}$ and thus $(A_T^+)_{11} = (A[T| :]^{-1})_{11} \neq 0$. On the other hand, we have $A_T \preceq A$ and $\rho(A_T) = \rho(A)$, so by Theorem 3.A we have $A_T^+ \preceq A^+$. Thus $(A_T^+)_{11} \neq 0$ implies that $(A^+)_{11} \neq 0$.

By similar arguments we can show that $(A^+)_{ij} \neq 0$ for each $i \in [n]$ and $j \in [m]$. Thus A^+ contains no zero entries. □

Next we want to show that certain special T_{13}-type matrices are CR-matrices. In order to prove this by using induction, we introduce the following notions.

DEFINITION 4.2. *An $m \times n$ matrix A is called a CC-matrix if for each pair of indices p and q in $[n]$ with $p \neq q$ there exists an $(n-1) \times n$ submatrix B of A such that both square submatrices $B[:|\{\bar{p}\}]$ and $B[:|\{\bar{q}\}]$ have full row term ranks.*

DEFINITION 4.3. *An $m \times n$ matrix A is called an RR-matrix if for each pair of indices i and j in $[m]$ with $i \neq j$ there exists an $(n+1) \times n$ submatrix $B = A[T|:]$ of A with $i \in T$ and $j \in T$ such that both square submatrices $A[T\backslash\{i\}|:]$ and $A[T\backslash\{j\}|:]$ have full term ranks.*

It is obvious from the definitions that the properties of being a CR-, CC-, or RR-matrix for a matrix A depend only on the zero pattern of A and are preserved under permutation of rows and columns. Also, if A is a CR- (or CC-, or RR-) matrix and B is obtained from A by replacing some zero entries of A with nonzero elements, then B is also a CR- (or CC-, or RR-) matrix.

A matrix A is called a CC+CR+RR matrix if A is a CC-matrix, a CR-matrix, and also an RR-matrix.

LEMMA 4.1. *Let B be an $m \times n$ CC+CR+RR matrix. Let A be a matrix obtained by adding a nonzero row to B. Then A is also a CC+CR+RR matrix.*

Proof. Without loss of generality, we may assume that

$$A = \begin{pmatrix} B \\ 1 * \cdots * \end{pmatrix}.$$

(1) It is obvious that A is a CC-matrix.

(2) We verify that A is a CR-matrix.

Take any position (p, q) of A.

Case 1. $p \leq m$. Then the desired submatrix can be obtained by using the CR-property of B.

Case 2. $p = m + 1$ and $q = 1$. Since B is a CC-matrix, B contains a square submatrix B_1 of order $n - 1$ with full term rank which does not use the first column of B. Appending the first column and last row of A to B_1 we obtain the desired submatrix.

Case 3. $p = m + 1$ and $q \geq 2$. Using the CC-property of B for the two indices 1 and q, there exists an $(n-1) \times n$ submatrix B_2 of B satisfying the corresponding CC-property for 1 and q. Appending the last row of A to B_2 we obtain the desired submatrix of A satisfying the CR-property for $(m + 1, q)$.

(3) We verify that A is still an RR-matrix.

Let i and j be any two indices in $[m + 1]$ with $i \neq j$.

Case 1. $i \leq m$ and $j \leq m$. Then the desired submatrix can be obtained by using the RR-property of B.

Case 2. One of i and j is $m + 1$. Say $j = m + 1$. Using the CR-property of B, we obtain a submatrix B_1 of order n of B satisfying the CR-property for the $(i, 1)$ position of B. Appending the last row of A to B_1 we obtain the desired submatrix of A satisfying the RR-property for i and $m + 1$. ∎

LEMMA 4.2. *Let A be an $(m + 1) \times (n + 1)$ matrix and $B = A[[m]|[n]]$ be an $m \times n$ submatrix of A as in* (4.1):

$$A = \begin{pmatrix} & & * \\ & B & \vdots \\ & & * \\ & & 1 \\ * \cdots * & & 1 \end{pmatrix}. \tag{4.1}$$

If B is a CC+CR+RR matrix, then A is also a CC+CR+RR matrix.

Proof. (1) We show that A is a CC-matrix.

Let p and q be any indices in $[n+1]$ with $p \neq q$.

Case 1. $p \leq n$ and $q \leq n$. By the CC-property of B there is an $(n-1) \times n$ submatrix B_1 of B satisfying the CC-property for p and q. Appending the last row and last column of A to B_1 we obtain the desired submatrix of A.

Case 2. One of p and q is $n+1$. Say $q = n+1$. By the CR-property of B, there is a submatrix B_1 of order n of B satisfying the CR-property for the (m, p) position of B. Appending the last column of A to B_1 we obtain the desired submatrix of A satisfying the CC-property for p and $n+1$.

(2) We show that A is an RR-matrix.

Let i and j be any two indices in $[m+1]$ with $i \neq j$.

Case 1. $i \leq m$ and $j \leq m$. By the RR-property of B, there is an $(n+1) \times n$ submatrix B_1 of B satisfying the RR-property for i and j. Appending the last row and last column of A to B_1 we obtain the desired $(n+2) \times (n+1)$ submatrix of A satisfying the RR-property for i and j.

Case 2. $i = m$ and $j = m+1$. By the RR-property of B there is a square submatrix B_1 of order n of B with full term rank which does not use the last row of B. Appending the last two rows and the last column of A to B_1 we obtain the desired $(n+2) \times (n+1)$ submatrix of A satisfying the RR-property for m and $m+1$.

Case 3. $i \leq m-1$ and $j = m+1$. By the RR-property of B there is an $(n+1) \times n$ submatrix B_1 of B satisfying the RR-property for i and m. Appending the last row and the last column of A to B_1 we obtain the desired $(n+2) \times (n+1)$ submatrix of A satisfying the RR-property for i and $m+1$.

(3) We show that A is a CR-matrix.

Take any position (p, q) of A.

Case 1. $p \leq m$ and $q \leq n$. By the CR-property of B there is a submatrix B_1 of order n of B satisfying the CR-property for the (p, q) position of B. Appending the last row and the last column of A to B_1 we obtain the desired submatrix (of order $(n+1)$) of A satisfying the CR-property for the (p, q) position of A.

Case 2. $p = m+1$ and $q = n+1$. By the CR-property of B there is a submatrix B_1 of order n of B with full term rank. Appending the last row and the last column of A to B_1 we obtain the desired submatrix of A.

Case 3. $p = m+1$ and $q \leq n$. By the CR-property of B there is a submatrix B_1 of order n of B satisfying the CR-property for the (m, q) position of B. Appending the last row and the last column of A to B_1 we obtain the desired submatrix of A.

Case 4. $p = m$ and $q = n+1$. By the RR-property of B there is a square submatrix B_1 of order n of B with full term rank which does not use the last row of B. Appending the mth row and the last column of A to B_1 we obtain the desired submatrix of A.

Case 5. $p \leq m-1$ and $q = n+1$. By the RR-property of B there is an $(n+1) \times n$ submatrix B_1 of B satisfying the RR-property for p and m. Appending the last column of A to B_1 we obtain the desired submatrix of A. ∎

LEMMA 4.3. *Let*

$$A = \begin{pmatrix} B & Z \\ Y & C \end{pmatrix},$$

where C is a matrix of type T1*, and where Y and Z are not both zero matrices. If B is a CC+CR+RR matrix, then A is also a CC+CR+RR matrix.*

Proof.

Case 1. $Y \neq 0$. We may assume that the first row Y_1 of Y is not a zero row. Then (where Y_2 denotes the second row of Y) B *is* $CC+CR+RR \overset{(Lemma\ 4.1)}{\Longrightarrow}$, $\begin{pmatrix} B \\ Y_1 \end{pmatrix}$ *is* $CC+CR+RR \overset{(Lemma\ 4.2)}{\Longrightarrow}$, $\begin{pmatrix} B & Z \\ Y_1 & 1 \\ Y_2 & 1 \end{pmatrix}$ *is* $CC+CR+RR \overset{(Lemma\ 4.1)}{\Longrightarrow}$, and $\begin{pmatrix} B & Z \\ Y & C \end{pmatrix}$ *is* $CC+CR+RR$.

Case 2. $Z \neq 0$. Then B *is* $CC+CR+RR \overset{(Lemma\ 4.2)}{\Longrightarrow}$, $\begin{pmatrix} B & Z \\ Y_1 & 1 \end{pmatrix}$ *is* $CC+CR+RR \overset{(Lemma\ 4.1)}{\Longrightarrow}$, and $\begin{pmatrix} B & Z \\ Y & C \end{pmatrix}$ *is* $CC+CR+RR$. □

LEMMA 4.4. *Let A be a matrix of type* T1 or T3. *Then A is a CC+CR+RR matrix.*

Proof. The case of type T1 is obvious (actually in this case the properties of being a CC- or CR-matrix hold vacuously). The case of type T3 can be proved by an inductive argument using Lemma 4.2 on the matrix B obtained from A by deleting a row and a column of A corresponding to a pendant vertex and to the pendant edge of the tree $T(A)$. □

DEFINITION 4.4 (see [8]). *Let A be a lower triangular blocked matrix as in the form* (3.2). *The block associated digraph $BD(A)$ is defined to be a digraph with vertex set $V = \{v_1, \ldots, v_k\}$ and arc set $E = \{(v_i, v_j) | B_{ij} \neq 0, i \neq j\}$. The (undirected) block associated graph $BG(A)$ is defined to be the graph obtained from $BD(A)$ by ignoring the directions of all the arcs of $BD(A)$.*

THEOREM 4.2. *Let*

$$A = \begin{pmatrix} A_1 & O & \cdots & O \\ B_{21} & A_2 & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ B_{k1} & B_{k2} & \cdots & A_k \end{pmatrix} \tag{4.2}$$

be an $m \times n$ T_{13}-type $(0,1)$ matrix such that the block associated graph $BG(A)$ is connected. Then A is a CC+CR+RR matrix.

Proof. We prove the result by using induction on m. The case $k = 1$ follows from Lemma 4.4. Thus we may assume that $k \geq 2$.

Since $BG(A)$ is connected, there exists a vertex (say, v_i) in $BG(A)$ such that the graph $BG(A)\backslash\{v_i\}$ is still connected (e.g., we can take v_i to be a pendant vertex in some spanning tree of $BG(A)$). Now we consider the corresponding diagonal block A_i of A according to the following two cases.

Case 1. A_i is of type T3 containing at least two columns. Then we may assume that

$$A_i = \begin{pmatrix} & & 0 \\ & A_i' & \vdots \\ & & 0 \\ & & 1 \\ 0 \cdots 0 & & 1 \end{pmatrix}, \tag{4.3}$$

where A_i' is still a matrix of type T3, and we may also write that

$$(B_{i1},\dots,B_{i,i-1},A_i)=\begin{pmatrix} & & 0\\ Y_1 & A_i' & \vdots\\ & & 0\\ & & 1\\ Y_2 & 0\cdots 0 & 1\end{pmatrix} \tag{4.4}$$

and

$$\begin{pmatrix} A_i\\ B_{i+1,i}\\ \vdots\\ B_{k,i}\end{pmatrix}=\begin{pmatrix} & 0\\ A_i' & \vdots\\ & 0\\ & 1\\ 0\cdots 0 & 1\\ Z_1 & Z_2\end{pmatrix}, \tag{4.5}$$

where Y_2 is a row and Z_2 is a column. We may also assume that Y_1 and Z_1 are not both zero, since the tree $T(A_i)$ contains at least two pendant vertices and at least two pendant edges, and $\binom{Y_1}{Y_2}$ and (Z_1Z_2) are not both zero by the connectedness of $BG(A)$.

Let B be a matrix obtained from A by deleting a row and a column of A corresponding to the last row and last column of A_i. Then B is still a T_{13}-type matrix and $BG(A)$ is still connected. By induction B is a CC+CR+RR matrix. So by Lemma 4.2, A is also a CC+CR+RR matrix.

Case 2. A_i is of type T1.

Let B be the matrix obtained from A by deleting those rows in the ith row block and those columns in the ith column block of A. Then B is a CC+CR+RR matrix by induction. Thus A is also a CC+CR+RR matrix by Lemma 4.3 (here we need the connectedness of $BG(A)$ to ensure that A and B satisfy the assumptions in Lemma 4.3). □

THEOREM 4.3. *Let A be an $m\times n$ real matrix. Then A has a totally nonzero signed GI if and only if A is a matrix of one of the following three types:*

(1) *A is a fully indecomposable S^2NS matrix.*

(2) *A is permutation equivalent to a T_{13}-type matrix X of the form* (4.2) *such that each off-diagonal block B_{ij} contains at most one nonzero entry, and the block associated graph $BG(X)$ is a tree.*

(3) *A^T is a matrix of type* (2).

Proof.

Sufficiency. Type (1) having totally nonzero signed GIs follows from Theorem 2.A, while types (2) and (3) having totally nonzero signed GIs follows from Theorem 4.1, Theorem 4.2, and [8].

Necessity. Without loss of generality, we may assume that $n\le m$.

Case 1. Assume $\rho(A)<n$. Then by Lemma 3.A and Theorem 3.B, A is permutation equivalent to a matrix of the form $\begin{pmatrix} B & 0\\ C & D\end{pmatrix}$ and $A^+=\begin{pmatrix} B^+ & 0\\ -D^+CB^+ & D^+\end{pmatrix}$. So in this case A does not have a totally nonzero signed GI, which is a contradiction.

Case 2. Assume $\rho(A) = n$. Then by Theorem 3.C and [8] (Theorem 4.2), A is permutation equivalent to a T-type matrix B of the form

$$B = \begin{pmatrix} X & 0 \\ Z & Y \end{pmatrix}, \tag{4.6}$$

where X is a T_{13}-type matrix each of whose off-diagonal blocks contains at most one nonzero entry and $BG(X)$ contains no (undirected) cycle, and Y is an S^2NS matrix. Now if both X and Y are nonvacuous, then by Theorem 3.B, we have $B^+ = \begin{pmatrix} X^+ & 0 \\ -Y^+ZX^+ & Y^+ \end{pmatrix}$, so B (and hence A) does not have a totally nonzero signed GI, which is a contradiction. So either $B = Y$ or $B = X$. If $B = Y$, then B (and hence A) is a fully indecomposable S^2NS matrix (i.e., a matrix of type (1)) by Theorem 2.A. If $B = X$, then $BG(X)$ must be a tree. Otherwise, the acyclic graph $BG(X)$ is not connected. Consequently X is permutation equivalent to a direct sum of some two (nonvacuous) matrices, and thus X^+ contains some zero entries, which is a contradiction. So $BG(X)$ is a tree and A is a matrix of type (2). □

REFERENCES

[1] L. Bassett, J. Maybee, and J. Quirk, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 36 (1968), pp. 544–563.

[2] R.A. Brualdi and B.L. Shader, *Matrices of Sign-Solvable Linear Systems*, Cambridge University Press, Cambridge, UK, 1995.

[3] R.A. Brualdi, K.L. Chavey, and B.L. Shader, *Bipartite graphs and inverse sign patterns of strong sign-nonsingular matrices*, J. Combin. Theory Ser. B, 62 (1994), pp. 133–152.

[4] R.A. Brualdi and H.J. Ryser, *Combinatorial Matrix Theory*, Cambridge University Press, New York, 1991.

[5] R.B. Bapat, K.P.S. Bhaskara Rao, and K.M. Prasad, *Generalized inverses over integral domains*, Linear Algebra Appl., 140 (1990), pp. 181–196.

[6] B.L. Shader, *Least squares sign-solvability*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1056–1073.

[7] J.-Y. Shao and H.-Y. Shan, *Matrices with signed generalized inverses*, Linear Algebra Appl., 322 (2001), pp. 105–127.

[8] J.-Y. Shao and H.-Y. Shan, *The solution of a problem on matrices having signed generalized inverses*, Linear Algebra Appl., 345 (2002), pp. 43–70.

SIAM J. MATRIX ANAL. APPL.
Vol. 0, No. 0, pp. 000–000

THE BLOCK NUMERICAL RANGE OF AN $n \times n$ BLOCK OPERATOR MATRIX*

CHRISTIANE TRETTER† AND MARKUS WAGENHOFER†

Abstract. We introduce the new notion of the block numerical range for bounded $n \times n$ block operator matrices. The main results concern spectral inclusion, inclusion between block numerical ranges for refined block decompositions, an estimate of the resolvent in terms of the block numerical range, and block numerical ranges of companion operators.

Key words. block numerical range, numerical range, spectrum, eigenvalue, resolvent, operator polynomial

AMS subject classifications. 47A10, 47A12, 47A75, 15A60

PII. S0895479801394076

1. Introduction.

The classical notion of the numerical range has been generalized in various ways in the literature (see, e.g., [1] and the references therein). Recently, the new concept of the quadratic numerical range of a 2×2 block operator matrix $\mathcal{A}$ in a Hilbert space $\mathcal{H}$ with respect to a decomposition $\mathcal{H} = H_1 \times H_2$ has been introduced in [7] and further studied in [5], [6]. It has been shown that the quadratic numerical range $W^2(\mathcal{A})$ is contained in the numerical range $W(\mathcal{A})$ of $\mathcal{A}$ and, like the numerical range, it contains the spectrum of $\mathcal{A}$ in its closure. Therefore (and because the quadratic numerical range need not be convex), it may give a better localization of the spectrum than the usual numerical range.

In the present paper we introduce the more general notion of the block numerical range of an $n \times n$ block operator matrix $\mathcal{A}$ in a Hilbert space $\mathcal{H}$ with respect to a decomposition $\mathcal{H} = H_1 \times H_2 \times \cdots \times H_n$. If, with respect to this decomposition,

$$\mathcal{A} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nn} \end{pmatrix},$$

then the block numerical range $W_{H_1 \times H_2 \times \cdots \times H_n}(\mathcal{A}) = W^n(\mathcal{A})$ is defined as the set of all $\lambda \in \mathbb{C}$ for which there exist $x_1 \in H_1, \ldots, x_n \in H_n$, $\|x_1\| = \cdots = \|x_n\| = 1$, such that

$$\det \left(\begin{pmatrix} (A_{11}x_1, x_1) & \cdots & (A_{1n}x_n, x_1) \\ \vdots & & \vdots \\ (A_{n1}x_1, x_n) & \cdots & (A_{nn}x_n, x_n) \end{pmatrix} - \lambda I_n \right) = 0,$$

where I_n denotes the identity matrix in $\mathbb{C}^n$. Like the numerical range, the block numerical range is bounded, and it is closed if $\mathcal{H}$ is finite dimensional. However,

*Received by the editors August 22, 2001; accepted for publication (in revised form) by A. C. M. Ran September 17, 2002; published electronically DATE. This research was supported by the Deutsche Forschungsgemeinschaft, DFG, under grant TR 368/4–1 and by the British Engineering and Physical Sciences Research Council, EPSRC, under grant GR/R40753.

http://www.siam.org/journals/simax/x-x/39407.html

†Fachbereich 3 – Mathematik, Universität Bremen, Bibliothekstr. 1, D–28359 Bremen, Germany (ctretter@math.uni-bremen.de, wagenhofer@math.uni-bremen.de).

whereas the numerical range of $\mathcal{A}$ consists of one convex component, the block numerical range of $\mathcal{A}$ may consist of n components (see [9]), and even its components need not be convex.

The paper is organized as follows. In section 2 we are going to show that the block numerical range contains the eigenvalues of the block operator matrix $\mathcal{A}$ and that the spectrum of $\mathcal{A}$ is contained in the closure of the block numerical range:

$$\sigma_p(\mathcal{A}) \subset W^n(\mathcal{A}), \quad \sigma(\mathcal{A}) \subset \overline{W^n(\mathcal{A})}.$$

In section 3 we will prove that if $\tilde{n} \geq n$ and $\mathcal{H} = \widetilde{H}_1 \times \widetilde{H}_2 \times \cdots \times \widetilde{H}_{\tilde{n}}$ is a refinement of the decomposition $\mathcal{H} = H_1 \times H_2 \times \cdots \times H_n$, then $W_{\widetilde{H}_1 \times \widetilde{H}_2 \times \cdots \times \widetilde{H}_{\tilde{n}}}(\mathcal{A}) \subset W_{H_1 \times H_2 \times \cdots \times H_n}(\mathcal{A})$, or, briefly,

$$W^{\tilde{n}}(\mathcal{A}) \subset W^n(\mathcal{A}), \quad \tilde{n} \geq n.$$

This suggests that a subsequent refinement of decompositions of the given space $\mathcal{H}$ leads to an increasingly improving approximation of the spectrum of $\mathcal{A}$ by the respective block numerical ranges. In section 4 we will prove that the resolvent of $\mathcal{A}$ can be estimated by

$$\|(\mathcal{A} - \lambda)^{-1}\| \leq \frac{(\|\mathcal{A}\| + |\lambda|)^{n-1}}{(\operatorname{dist}(\lambda, W^n(\mathcal{A}))^n}, \quad \lambda \notin \overline{W^n(\mathcal{A})},$$

and we will show more detailed estimates in the case when $W^n(\mathcal{A})$ consists of several components. This will enable us to estimate the length of Jordan chains in boundary points of the block numerical range. Finally, we will consider the connection between the numerical range of an operator polynomial and the block numerical range of its companion operator.

2. Definition and spectral inclusion. Let $n \in \mathbb{N}$, let $H_1, \ldots, H_n$ be complex Hilbert spaces, $\mathcal{H} = H_1 \times \cdots \times H_n$, and consider the operator $\mathcal{A} \in L(\mathcal{H})$ given by

$$\mathcal{A} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nn} \end{pmatrix}, \tag{2.1}$$

where $A_{ij} \in L(H_j, H_i)$, $i, j = 1, \ldots, n$. For $x = (x_1, \ldots, x_n) \in H_1 \times \cdots \times H_n$ define $\mathcal{A}_x \in M_n(\mathbb{C})$ (the space of $n \times n$ matrices over $\mathbb{C}$) by

$$\mathcal{A}_x := \begin{pmatrix} (A_{11}x_1, x_1) & \cdots & (A_{1n}x_n, x_1) \\ \vdots & & \vdots \\ (A_{n1}x_1, x_n) & \cdots & (A_{nn}x_n, x_n) \end{pmatrix}, \tag{2.2}$$

that is, $(\mathcal{A}_x)_{ij} := (A_{ij}x_j, x_i)$, $i, j = 1, \ldots, n$. Further, we denote by $\mathcal{S}_{H_1 \times \cdots \times H_n} := \{x = (x_1, \ldots, x_n) \in \mathcal{H} : \|x_1\| = \cdots = \|x_n\| = 1\}$ the product of unit spheres.

Definition 2.1. *The set*

$$W_{H_1 \times \cdots \times H_n}(\mathcal{A}) := \{\lambda \in \mathbb{C} : \exists\, x \in \mathcal{S}_{H_1 \times \cdots \times H_n} \ \det(\mathcal{A}_x - \lambda) = 0\} \tag{2.3}$$

is called the block numerical range *of $\mathcal{A}$ with respect to the block operator representation* (2.1). *If the decomposition of $\mathcal{H}$ is fixed, we also write $W^n(\mathcal{A}) = W_{H_1 \times \cdots \times H_n}(\mathcal{A})$ and $\mathcal{S}^n = \mathcal{S}_{H_1 \times \cdots \times H_n}$, respectively.*

REMARK 2.2. *Since* $\{\lambda \in \mathbb{C} : \det(\mathcal{A}_x - \lambda) = 0\} = \sigma(\mathcal{A}_x) = \sigma_p(\mathcal{A}_x)$ *for all* $x \in \mathcal{S}^n$, *the set* $W^n(\mathcal{A})$ *has the equivalent representation*

$$W^n(\mathcal{A}) = \bigcup_{x \in \mathcal{S}^n} \sigma(\mathcal{A}_x) = \bigcup_{x \in \mathcal{S}^n} \sigma_p(\mathcal{A}_x). \tag{2.4}$$

For $n = 1$, the block numerical range coincides with the usual numerical range; for $n = 2$ it coincides with the quadratic numerical range introduced in [7]; and for an $n \times n$ matrix $W^n(\mathcal{A})$ coincides with the spectrum of $\mathcal{A}$.

If $\mathcal{A}$ is a lower or upper tridiagonal matrix, then $W^n(\mathcal{A}) = W(A_{11}) \cup \cdots \cup W(A_{nn})$. In general, $W^n(\mathcal{A})$ need not be convex; it may consist of at most n components which need not be convex.

As an example, we consider the 3×3 block operator matrix

$$\mathcal{A}_1 = \left(\begin{array}{cc|c|c} 2 & 0 & 1 & 1 \\ 0 & -2 & 1 & 1 \\ \hline \mathrm{i} & \mathrm{i} & 1 & 0 \\ \hline \mathrm{i} & \mathrm{i} & 0 & -1 \end{array}\right)$$

for which the cubic numerical range has 3 components, but none of them is convex (see Figure 1).

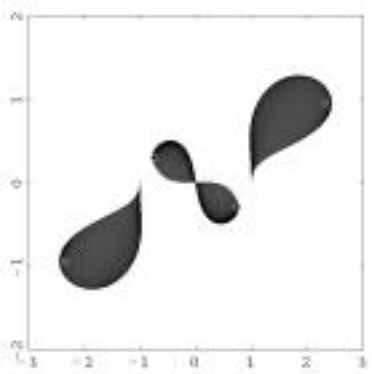


FIG. 1. $W_{\mathbb{C}^2 \times \mathbb{C} \times \mathbb{C}}(\mathcal{A}_1)$.

REMARK 2.3. *The following properties hold:*

(i) *For all* $x \in \mathcal{H}$ *we have* $\|\mathcal{A}_x\| \le \|\mathcal{A}\|$. *Hence*

$$W^n(\mathcal{A}) \subset K_{\|\mathcal{A}\|}(0) := \{\lambda \in \mathbb{C} : |\lambda| \le \|\mathcal{A}\|\}$$

is bounded and compact if $\mathcal{H}$ *is finite dimensional.*

(ii) *Since* $(\mathcal{A}_x)^* = (\mathcal{A}^*)_x$, *we have*

$$W^n(\mathcal{A}^*) = W^n(\mathcal{A})^* := \{\lambda \in \mathbb{C} : \overline{\lambda} \in W^n(\mathcal{A})\}.$$

(iii) *If* $\mathcal{A}$ *is self-adjoint, then* $W^n(\mathcal{A})$ *is real.*

Proof. The assertions (ii) and (iii) are clear. For (i) let $x = (x_1, \ldots, x_n) \in \mathcal{S}^n$, $z = (z_1, \ldots, z_n) \in \mathbb{C}^n$, $\|z\| = 1$, and define $y_j := z_j x_j$, $j = 1, \ldots, n$, $y := (y_1, \ldots, y_n)$. Then $\|y\| = 1$ and

$$\|\mathcal{A}_x z\|^2 = \sum_{i=1}^n \left| \sum_{j=1}^n (A_{ij} x_j, x_i) z_j \right|^2 \le \sum_{i=1}^n \left\| \sum_{j=1}^n A_{ij} y_j \right\|^2 \|x_i\|^2 = \|\mathcal{A} y\|^2 \le \|\mathcal{A}\|^2.$$

The remaining statement follows from (2.4) and $\max|\sigma_p(\mathcal{A}_x)| \le \|\mathcal{A}_x\|$, $x \in \mathcal{S}^n$. ∎

In Figure 1 the eigenvalues of $\mathcal{A}_1$, which are marked by little circles, are obviously contained in $W^3(\mathcal{A}_1)$. In order to prove general spectral inclusion, we need the following lemma.

LEMMA 2.4. *Let $A \in M_n(\mathbb{C})$. If A is invertible, then*

$$\|A^{-1}\| \le \frac{\|A\|^{n-1}}{|\det A|}. \tag{2.5}$$

For all $x \in \mathbb{C}^n$, $\|x\| = 1$, we have

$$\operatorname{dist}(0, \sigma(A)) \le \sqrt[n]{\|A\|^{n-1}\|Ax\|}.$$

Proof. The first estimate has been proved in [2, Lem. 1] (see also [3, Chap. I, eq. (4.12)] and note that $\mathbb{C}^n$ is a unitary space). The second statement is trivial if A is not invertible. Now let A be invertible and let $x \in \mathbb{C}^n$, $\|x\| = 1$. Then $\|Ax\| > 0$ and

$$\|A^{-1}\| \ge \left\| A^{-1}\left(\frac{Ax}{\|Ax\|}\right)\right\| = \frac{\|x\|}{\|Ax\|} = \frac{1}{\|Ax\|}.$$

If $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A, we obtain, using (2.5),

$$(\operatorname{dist}(0, \sigma(A)))^n = \left(\min_{i=1}^{n} |\lambda_i|\right)^n \le |\lambda_1 \cdots \lambda_n| = |\det A| \le \frac{\|A\|^{n-1}}{\|A^{-1}\|} \le \|A\|^{n-1}\|Ax\|. \qquad ∎$$

THEOREM 2.5. *The following inclusions hold:*

$$\sigma_p(\mathcal{A}) \subset W^n(\mathcal{A}), \quad \sigma(\mathcal{A}) \subset \overline{W^n(\mathcal{A})}.$$

Proof. First let $\lambda \in \sigma_p(\mathcal{A})$. Then there exists $x = (x_1, \dots, x_n) \in \mathcal{H}$, $x \ne 0$, such that $\mathcal{A}x - \lambda x = 0$. Write $x_i = \|x_i\|\,\hat{x}_i$ with $\|\hat{x}_i\| = 1$, $i = 1, \dots, n$. Then we have $\hat{x} := (\hat{x}_1, \dots, \hat{x}_n) \in \mathcal{S}^n$, $(x_i, \hat{x}_i) = \|x_i\|$, $i = 1, \dots, n$, and

$$\begin{aligned}
(\mathcal{A}_{\hat{x}} - \lambda)\begin{pmatrix} \|x_1\| \\ \vdots \\ \|x_n\| \end{pmatrix} &= \begin{pmatrix} (A_{11}\hat{x}_1, \hat{x}_1) & \cdots & (A_{1n}\hat{x}_n, \hat{x}_1) \\ \vdots & & \vdots \\ (A_{n1}\hat{x}_1, \hat{x}_n) & \cdots & (A_{nn}\hat{x}_n, \hat{x}_n) \end{pmatrix}\begin{pmatrix} \|x_1\| \\ \vdots \\ \|x_n\| \end{pmatrix} - \begin{pmatrix} \lambda\|x_1\| \\ \vdots \\ \lambda\|x_n\| \end{pmatrix} \\
&= \begin{pmatrix} (A_{11}x_1, \hat{x}_1) + \cdots + (A_{1n}x_n, \hat{x}_1) - \lambda(x_1, \hat{x}_1) \\ \vdots \\ (A_{n1}x_1, \hat{x}_n) + \cdots + (A_{nn}x_n, \hat{x}_n) - \lambda(x_n, \hat{x}_n) \end{pmatrix} \\
&= \begin{pmatrix} \left(\sum_{j=1}^n A_{1j}x_j - \lambda x_1, \hat{x}_1\right) \\ \vdots \\ \left(\sum_{j=1}^n A_{nj}x_j - \lambda x_n, \hat{x}_n\right) \end{pmatrix} = 0.
\end{aligned}$$

Hence $\lambda \in \sigma(\mathcal{A}_{\hat{x}}) \subset W^n(\mathcal{A})$ by (2.4).

For the proof of the second inclusion, let $\lambda \in \sigma(\mathcal{A})$. Then either $\lambda \in \sigma_p(\mathcal{A}^*)^*$ or $\lambda \in \sigma_{app}(\mathcal{A})$ (the approximate point spectrum of $\mathcal{A}$). If $\lambda \in \sigma_p(\mathcal{A}^*)^*$, then $\bar{\lambda} \in \sigma_p(\mathcal{A}^*)$ and hence $\bar{\lambda} \in W^n(\mathcal{A}^*) = W^n(\mathcal{A})^*$ by the first inclusion and Remark 2.3(i). This shows $\lambda \in W^n(\mathcal{A})$.

Now let $\lambda \in \sigma_{app}(\mathcal{A})$. Then there is a sequence $(x^{(\nu)})_{\nu=1}^{\infty} \subset \mathcal{H}$ with $\|x^{(\nu)}\| = 1$ and $\mathcal{A}x^{(\nu)} - \lambda x^{(\nu)} \longrightarrow 0$ for $\nu \to \infty$. Writing $x_i^{(\nu)} = \|x_i^{(\nu)}\|\,\hat{x}_i^{(\nu)}$ with $\|\hat{x}_i^{(\nu)}\| = 1$, $i = 1, \ldots, n$, $\nu = 1, 2, \ldots$, we obtain $\hat{x}^{(\nu)} := (\hat{x}_1^{(\nu)}, \ldots, \hat{x}_n^{(\nu)}) \in \mathcal{S}^n$ and, in a similar way as above,

$$(\mathcal{A}_{\hat{x}^{(\nu)}} - \lambda) \begin{pmatrix} \|x_1^{(\nu)}\| \\ \vdots \\ \|x_n^{(\nu)}\| \end{pmatrix} \longrightarrow 0, \quad \nu \to \infty,$$

whence $\varepsilon_\nu := \big\|(\mathcal{A}_{\hat{x}^{(\nu)}} - \lambda)(\|x_1^{(\nu)}\|, \ldots, \|x_n^{(\nu)}\|)^{\mathrm{t}}\big\| \longrightarrow 0$, $\nu \to \infty$. Here $^{\mathrm{t}}$ denotes the transpose of a matrix or a vector. Since $\big\|(\|x_1^{(\nu)}\|, \ldots, \|x_n^{(\nu)}\|)\big\| = \|x^{(\nu)}\| = 1$, Lemma 2.4 and Remark 2.3(i) imply that

$$\begin{aligned} \operatorname{dist}(\lambda, \sigma(\mathcal{A}_{x^{(\nu)}})) = \operatorname{dist}(0, \sigma(\mathcal{A}_{x^{(\nu)}} - \lambda)) &\le \sqrt[n]{\|\mathcal{A}_{x^{(\nu)}} - \lambda\|^{n-1}\varepsilon_\nu} \\ &\le \sqrt[n]{(\|\mathcal{A}\| + |\lambda|)^{n-1}\varepsilon_\nu} \longrightarrow 0, \quad \nu \to \infty, \end{aligned}$$

and therefore

$$\lambda \in \overline{\bigcup_{\nu\in\mathbb{N}} \sigma(\mathcal{A}_{x^{(\nu)}})} \subset \overline{\bigcup_{x\in\mathcal{S}^n} \sigma(\mathcal{A}_x)} = \overline{W^n(\mathcal{A})}. \qquad \square$$

As an illustration of Theorem 2.5, we consider the two block operator matrices

$$\mathcal{A}_2 = \left(\begin{array}{c|cc|c} 2 & \mathrm{i} & 1 & 0 \\ \hline \mathrm{i} & 2 & 0 & 1 \\ 1 & 0 & -2 & \mathrm{i} \\ \hline 0 & 1 & \mathrm{i} & -2 \end{array}\right), \quad \mathcal{A}_3 = \left(\begin{array}{cc|cc} -2 & -1 & 1 & 0 \\ -1 & -2 & 0 & 1 \\ \hline -2 & -1 & 0 & -3\mathrm{i} \\ -1 & -2 & 3\mathrm{i} & 0 \end{array}\right).$$

Figure 2 shows their eigenvalues marked by little circles and their cubic numerical ranges. Note that the horizontal line in the right picture is part of the cubic numerical range.

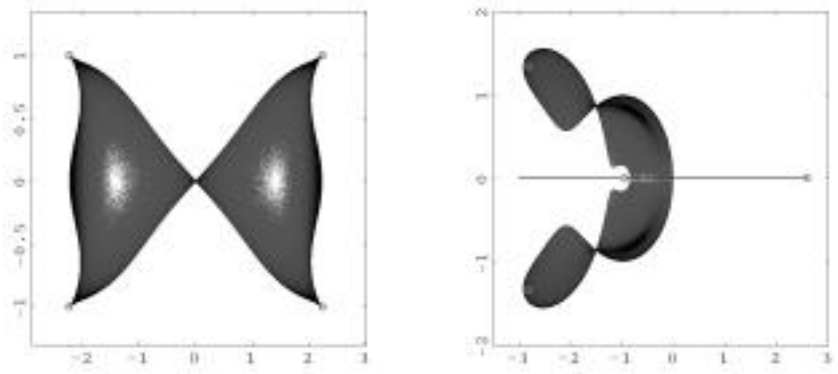


FIG. 2. $W_{\mathbb{C}\times\mathbb{C}^2\times\mathbb{C}}(\mathcal{A}_2)$ *and* $W_{\mathbb{C}\times\mathbb{C}\times\mathbb{C}^2}(\mathcal{A}_3)$.

3. Inclusions between block numerical ranges. In this section we will prove that the block numerical range of a principal submatrix of an $n \times n$ block operator matrix $\mathcal{A}$ is contained in the block numerical range of $\mathcal{A}$ if a certain dimension condition is satisfied. Further we will show that a refinement of the decomposition $H_1 \times \cdots \times H_n$ makes the block numerical range smaller.

THEOREM 3.1. *Let $k \in \mathbb{N}$, $1 \leq k \leq n$, $1 \leq i_1 < \cdots < i_k \leq n$ and let P be the orthogonal projection of $H_1 \times \cdots \times H_n$ onto $H_{i_1} \times \cdots \times H_{i_k}$.*

If there exists an enumeration $i'_1, \ldots, i'_{n-k}$ of the elements of the set $\{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$ such that $\dim H_{i'_j} > n - j$, $j = 1, \ldots, n-k$, then

$$W_{H_{i_1} \times \cdots \times H_{i_k}}(P\mathcal{A}P) \subset W_{H_1 \times \cdots \times H_n}(\mathcal{A}).$$

Proof. For $k = n$ the statement is trivial. For $k = n-1$ there is an $i \in \{1, \ldots, n\}$ such that $\{i_1, \ldots, i_k\} \cup \{i\} = \{1, \ldots, n\}$. If we denote $\mathcal{H}'_i := H_1 \times \cdots \times H_{i-1} \times H_{i+1} \times \cdots \times H_n$ and $\mathcal{A}'_i := P\mathcal{A}P$, then

$$\mathcal{A}'_i = \begin{pmatrix} A_{11} & \cdots & A_{1,i-1} & A_{1,i+1} & \cdots & A_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ A_{i-1,1} & \cdots & A_{i-1,i-1} & A_{i-1,i+1} & \cdots & A_{i-1,n} \\ A_{i+1,1} & \cdots & A_{i+1,i-1} & A_{i+1,i+1} & \cdots & A_{i+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ A_{n1} & \cdots & A_{n,i-1} & A_{n,i+1} & \cdots & A_{nn} \end{pmatrix}.$$

Now let $\lambda \in W_{H_1 \times \cdots \times H_{i-1} \times H_{i+1} \times \cdots \times H_n}(\mathcal{A}'_i)$. Then there exists an element $x' = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathcal{S}_{\mathcal{H}'_i}$ with $\det((\mathcal{A}'_i)_{x'} - \lambda) = 0$. Since

$$\dim \operatorname{span}\{A_{i1}x_1, \ldots, A_{i,i-1}x_{i-1}, A_{i,i+1}x_{i+1}, \ldots, A_{in}x_n\} \leq n - 1 < \dim H_i$$

by assumption, there is an $x_i \in H_i$, $\|x_i\| = 1$, with $(\mathcal{A}_x)_{ij} = (A_{ij}x_j, x_i) = 0$ for $j = 1, \ldots, i-1, i+1, \ldots, n$. Then we have $x := (x_1, \ldots, x_n) \in \mathcal{S}_{\mathcal{H}}$ and

$$\mathcal{A}_x = \begin{pmatrix} (\mathcal{A}_x)_{11} & \cdots & (\mathcal{A}_x)_{1,i-1} & (\mathcal{A}_x)_{1i} & (\mathcal{A}_x)_{1,i+1} & \cdots & (\mathcal{A}_x)_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \\ (\mathcal{A}_x)_{i-1,1} & \cdots & (\mathcal{A}_x)_{i-1,i-1} & (\mathcal{A}_x)_{i-1,i} & (\mathcal{A}_x)_{i-1,i+1} & \cdots & (\mathcal{A}_x)_{i-1,n} \\ 0 & \cdots & 0 & (\mathcal{A}_x)_{ii} & 0 & \cdots & 0 \\ (\mathcal{A}_x)_{i+1,1} & \cdots & (\mathcal{A}_x)_{i+1,i-1} & (\mathcal{A}_x)_{i+1,i} & (\mathcal{A}_x)_{i+1,i+1} & \cdots & (\mathcal{A}_x)_{i+1,n} \\ \vdots & & \vdots & \vdots & \cdots & & \\ (\mathcal{A}_x)_{n1} & \cdots & (\mathcal{A}_x)_{n,i-1} & (\mathcal{A}_x)_{ni} & (\mathcal{A}_x)_{n,i+1} & \cdots & (\mathcal{A}_x)_{nn} \end{pmatrix}.$$

Thus $\det(\mathcal{A}_x - \lambda) = ((\mathcal{A}_x)_{ii} - \lambda) \det((\mathcal{A}'_i)_{x'} - \lambda) = 0$ and hence $\lambda \in W_{H_1 \times \cdots \times H_n}(\mathcal{A})$. The case $k < n-1$ follows by induction. ∎

As a special case of Theorem 3.1 we obtain that under a certain dimension condition, the numerical ranges of the diagonal entries A_{ii} of $\mathcal{A}$ are contained in the block numerical range of $\mathcal{A}$.

COROLLARY 3.2. *Let $i \in \mathbb{N}$. If there exists an enumeration $i'_1, \ldots, i'_{n-1}$ of the elements of the set $\{1, \ldots, i-1, i+1, \ldots, n\}$ with $\dim H_{i'_j} > n - j$, $j = 1, \ldots, n-1$, then*

$$W(A_{ii}) \subset W^n(\mathcal{A}).$$

If $\dim H_i \geq n$ *for all* $i = 1, \ldots, n$, *then* $W(A_{ii}) \subset W^n(\mathcal{A})$ *for all* $i = 1, \ldots, n$.

COROLLARY 3.3. *If* $\dim H_i \geq n$ *for all* $i = 1, \ldots, n$ *and* $W^n(\mathcal{A})$ *consists of* n *components* $K_1, \ldots, K_n \subset \mathbb{C}$, *then there exists a permutation* π *of* $\{1, \ldots, n\}$ *such that* $W(A_{ii}) \subset K_{\pi(i)}$, $i = 1, \ldots, n$.

Proof. Choose an arbitrary $x_1 \in S_{H_1}$. Then, by the dimension condition, we can choose $x_k \in S_{H_k}$, $k = 2, \ldots, n$, recursively so that $x_k \perp \{A_{k1}x_1, \ldots, A_{k,k-1}x_{k-1}\}$. Since $W^n(\mathcal{A})$ consists of n components $K_1, \ldots, K_n \subset \mathbb{C}$, every matrix $\mathcal{A}_x$, $x \in \mathcal{S}_\mathcal{H}$, has exactly one eigenvalue in each component of $W^n(\mathcal{A})$. In particular, if we let $x := (x_1, \ldots, x_n) \in \mathcal{S}_\mathcal{H}$, then

$$\mathcal{A}_x = \begin{pmatrix} (A_{11}x_1, x_1) & \cdots & (A_{1n}x_n, x_1) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (A_{nn}x_n, x_n) \end{pmatrix},$$

and hence there exists a permutation π of $\{1, \ldots, n\}$ such that $(A_{ii}x_i, x_i) \in K_{\pi(i)}$ for $i = 1, \ldots, n$. By Corollary 3.2 we have $W(A_{ii}) \subset W^n(\mathcal{A})$ for all $i = 1, \ldots, n$, and since the numerical range is convex, this implies the assertion. ∎

That the dimension condition in Theorem 3.1 is essential can be seen from the following example. For the matrix $\mathcal{A}_4$ and its principal submatrix $\mathcal{A}_4'$ given by

$$\mathcal{A}_4 = \left(\begin{array}{c|ccc} 1 & 3+\mathrm{i} & 2 & \mathrm{i} \\ \hline 3+\mathrm{i} & 1 & \mathrm{i} & 2 \\ -2 & \mathrm{i} & 1 & 3+\mathrm{i} \\ \hline \mathrm{i} & -2 & 3+\mathrm{i} & 1 \end{array}\right), \quad \mathcal{A}_4' = \left(\begin{array}{cc|c} 1 & \mathrm{i} & 2 \\ \mathrm{i} & 1 & 3+\mathrm{i} \\ \hline -2 & 3+\mathrm{i} & 1 \end{array}\right),$$

Figure 3 shows that the quadratic numerical range of $\mathcal{A}_4'$ is not contained in the cubic numerical range of $\mathcal{A}_4$.

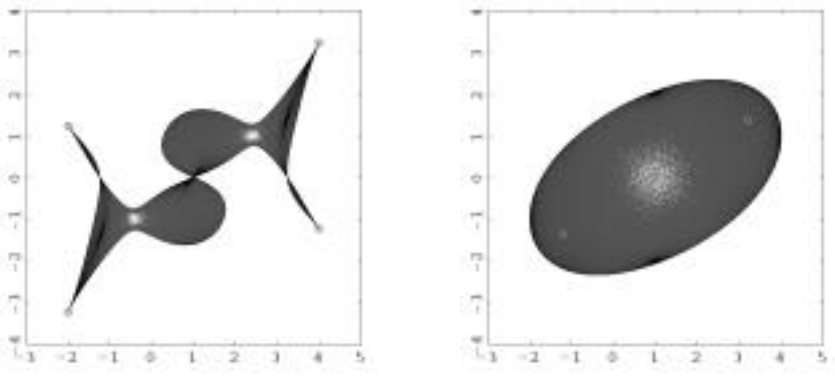


FIG. 3. $W_{\mathbb{C}\times\mathbb{C}^2\times\mathbb{C}}(\mathcal{A}_4)$ *and* $W_{\mathbb{C}^2\times\mathbb{C}}(\mathcal{A}_4')$.

In what follows we are going to consider the behavior of the block numerical range under refinements of the decomposition of $\mathcal{H}$.

DEFINITION 3.4. *Let* $n, \tilde{n} \in \mathbb{N}$ *and let* $\mathcal{H} = H_1 \times \cdots \times H_n = \widetilde{H}_1 \times \cdots \times \widetilde{H}_{\tilde{n}}$ *with Hilbert spaces* $H_1, \ldots, H_n$ *and* $\widetilde{H}_1, \ldots, \widetilde{H}_n$. *Then* $\widetilde{H}_1 \times \cdots \times \widetilde{H}_{\tilde{n}}$ *is called a* refinement of $H_1 \times \cdots \times H_n$ *if* $n \leq \tilde{n}$ *and there are integers* $0 = i_0 < \cdots < i_n = \tilde{n}$ *such that* $H_k = \widetilde{H}_{i_{k-1}+1} \times \cdots \times \widetilde{H}_{i_k}$ *for all* $k = 1, \ldots, n$.

THEOREM 3.5. *If* $\mathcal{H} = \widetilde{H}_1 \times \cdots \times \widetilde{H}_{\tilde{n}}$ *is a refinement of* $\mathcal{H} = H_1 \times \cdots \times H_n$, *then*

$$W_{\widetilde{H}_1 \times \cdots \times \widetilde{H}_{\tilde{n}}}(\mathcal{A}) \subset W_{H_1 \times \cdots \times H_n}(\mathcal{A}),$$

or, briefly,

$$W^{\tilde{n}}(\mathcal{A}) \subset W^n(\mathcal{A}), \quad \tilde{n} \geq n.$$

Proof. It is sufficient to consider the case $\tilde{n} = n+1$. The general case then follows easily by induction. If $\tilde{n} = n+1$, there exists a $k \in \{1, \dots, n\}$ such that the refinement $\mathcal{H} = \widetilde{H}_1 \times \cdots \times \widetilde{H}_{\tilde{n}}$ of $H_1 \times \cdots \times H_n$ is of the form $\mathcal{H} = H_1 \times \cdots \times H_{k-1} \times H_k^1 \times H_k^2 \times H_{k+1} \times \cdots \times H_n$, where $H_k = H_k^1 \times H_k^2$. With respect to this refined decomposition, $\mathcal{A}$ has the representation

$$\mathcal{A} = \begin{pmatrix} A_{11} & \cdots & A_{1k}^1 & A_{1k}^2 & \cdots & A_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ A_{k1}^1 & \cdots & A_{kk}^{11} & A_{kk}^{12} & \cdots & A_{kn}^1 \\ A_{k1}^2 & \cdots & A_{kk}^{21} & A_{kk}^{22} & \cdots & A_{kn}^2 \\ \vdots & & \vdots & \vdots & & \vdots \\ A_{n1} & \cdots & A_{nk}^1 & A_{nk}^2 & \cdots & A_{nn} \end{pmatrix},$$

where for the entries $A_{ij} \in L(H_j, H_i)$, $i, j = 1, 2, \dots, n$, of the representation (2.1) of $\mathcal{A}$ with respect to $\mathcal{H} = H_1 \times \cdots \times H_n$, we have

$$A_{kk} = \begin{pmatrix} A_{kk}^{11} & A_{kk}^{12} \\ A_{kk}^{21} & A_{kk}^{22} \end{pmatrix}, \quad A_{ki} = \begin{pmatrix} A_{ki}^1 \\ A_{ki}^2 \end{pmatrix}, \quad A_{jk} = \begin{pmatrix} A_{jk}^1 & A_{jk}^2 \end{pmatrix}, \quad i, j = 1, \dots, n,$$

with $A_{kk}^{st} \in L(H_k^t, H_k^s)$, $A_{jk}^s \in L(H_k^s, H_j)$, and $A_{ki}^t \in L(H_i, H_k^t)$, $s, t = 1, 2$.

By Theorem 2.5 about the spectral inclusion, we conclude that

$$\begin{aligned} W_{H_1 \times \cdots \times H_k^1 \times H_k^2 \times \cdots \times H_n}(\mathcal{A}) &= \bigcup \{\sigma(\mathcal{A}_x) : x \in \mathcal{S}_{H_1 \times \cdots \times H_k^1 \times H_k^2 \times \cdots \times H_n}\} \\ &\subset \bigcup \{W_{\mathbb{C} \times \cdots \times \mathbb{C}^2 \times \cdots \times \mathbb{C}}(\mathcal{A}_x) : x \in \mathcal{S}_{H_1 \times \cdots \times H_k^1 \times H_k^2 \times \cdots \times H_n}\}. \end{aligned}$$

The theorem is proved if we show that for $x \in \mathcal{S}_{H_1 \times \cdots \times H_k^1 \times H_k^2 \times \cdots \times H_n}$

$$W_{\mathbb{C} \times \cdots \times \mathbb{C}^2 \times \cdots \times \mathbb{C}}(\mathcal{A}_x) \subset \bigcup \{\sigma(\mathcal{A}_y) : y \in \mathcal{S}_{H_1 \times \cdots \times H_n}\} = W_{H_1 \times \cdots \times H_n}(\mathcal{A}).$$

To this end, let $x \in \mathcal{S}_{H_1 \times \cdots \times H_k^1 \times H_k^2 \times \cdots \times H_n}$, $x = (x_1, \dots, x_k^1, x_k^2, \dots, x_n)$ such that $\|x_1\| = \cdots = \|x_k^1\| = \|x_k^2\| = \cdots = \|x_n\| = 1$. Then

$$\mathcal{A}_x = \left(\begin{array}{c|c|cc|c|c} (A_{11}x_1, x_1) & \cdots & (A_{1k}^1 x_k^1, x_1) & (A_{1k}^2 x_k^2, x_1) & \cdots & (A_{1n}x_n, x_1) \\ \hline \vdots & & \vdots & \vdots & & \vdots \\ \hline (A_{k1}^1 x_1, x_k^1) & \cdots & (A_{kk}^{11} x_k^1, x_k^1) & (A_{kk}^{12} x_k^2, x_k^1) & \cdots & (A_{kn}^1 x_n, x_k^1) \\ (A_{k1}^2 x_1, x_k^2) & \cdots & (A_{kk}^{21} x_k^1, x_k^2) & (A_{kk}^{22} x_k^2, x_k^2) & \cdots & (A_{kn}^2 x_n, x_k^2) \\ \hline \vdots & & \vdots & \vdots & & \vdots \\ \hline (A_{n1}x_1, x_n) & \cdots & (A_{nk}^1 x_k^1, x_n) & (A_{nk}^2 x_k^2, x_n) & \cdots & (A_{nn}x_n, x_n) \end{array}\right)$$

$$=: \begin{pmatrix} B_{11} & \cdots & B_{1k} & \cdots & B_{1n} \\ \vdots & & \vdots & & \vdots \\ B_{k1} & \cdots & B_{kk} & \cdots & B_{kn} \\ \vdots & & \vdots & & \vdots \\ B_{n1} & \cdots & B_{nk} & \cdots & B_{nn} \end{pmatrix} = \mathcal{B} \in L(\mathbb{C} \times \cdots \times \mathbb{C}^2 \times \cdots \times \mathbb{C}).$$

If for any $z \in \mathcal{S}_{\mathbb{C}\times\cdots\times\mathbb{C}^2\times\cdots\times\mathbb{C}}$ we find a $y \in \mathcal{S}_{H_1\times\cdots\times H_n}$ with $\mathcal{B}_z = \mathcal{A}_y$, then $\sigma((\mathcal{A}_x)_z) = \sigma(\mathcal{B}_z) = \sigma(\mathcal{A}_y)$, and hence

$$\begin{aligned} W_{\mathbb{C}\times\cdots\times\mathbb{C}^2\times\cdots\times\mathbb{C}}(\mathcal{A}_x) &= \bigcup \{\sigma((\mathcal{A}_x)_z) : z \in \mathcal{S}_{\mathbb{C}\times\cdots\times\mathbb{C}^2\times\cdots\times\mathbb{C}}\} \\ &\subset \bigcup \{\sigma(\mathcal{A}_y) : y \in \mathcal{S}_{H_1\times\cdots\times H_n}\} \end{aligned}$$

as required. In order to show this, let $z = (z_1, \ldots, z_k, \ldots, z_n)$, $z_i \in \mathbb{C}$, $i \neq k$, $z_k = (z_k^1, z_k^2) \in \mathbb{C}^2$, with $|z_1|^2 = \cdots = \|z_k\|^2 = \cdots = |z_n|^2 = 1$. Then

$$\mathcal{B}_z = \begin{pmatrix} (B_{11}z_1, z_1) & \cdots & (B_{1k}z_k, z_1) & \cdots & (B_{1n}z_n, z_1) \\ \vdots & & \vdots & & \vdots \\ (B_{k1}z_1, z_k) & \cdots & (B_{kk}z_k, z_k) & \cdots & (B_{kn}z_n, z_k) \\ \vdots & & \vdots & & \vdots \\ (B_{n1}z_1, z_n) & \cdots & (B_{nk}z_k, z_n) & \cdots & (B_{nn}z_n, z_n) \end{pmatrix}.$$

Set $y_k^i := z_k^i x_k^i$, $i = 1, 2$, and $y = (y_1, \ldots, y_k, \ldots, y_n) := (z_1x_1, \ldots, (y_k^1, y_k^2), \ldots, z_nx_n)$. Then $y_i \in H_i$, $i = 1, \ldots, n$, $\|y_i\| = 1$, $i \neq k$, and $\|y_k\|^2 = \|z_k^1 x_k^1\|^2 + \|z_k^2 x_k^2\|^2 = |z_k^1|^2\|x_k^1\|^2 + |z_k^2|^2\|x_k^2\|^2 = |z_k^1|^2 + |z_k^2|^2 = \|z_k\|^2 = 1$, and hence $y \in \mathcal{S}_{H_1\times\cdots\times H_n}$. With this choice of y, we obtain the desired equality $\mathcal{B}_z = \mathcal{A}_y$. Indeed, e.g., for $i = j = k$,

$$\begin{aligned} (B_z)_{kk} &= \left(\begin{pmatrix} (A_{kk}^{11}x_k^1, x_k^1)z_k^1 + (A_{kk}^{12}x_k^2, x_k^1)z_k^2 \\ (A_{kk}^{21}x_k^1, x_k^2)z_k^1 + (A_{kk}^{22}x_k^2, x_k^2)z_k^2 \end{pmatrix}, \begin{pmatrix} z_k^1 \\ z_k^2 \end{pmatrix} \right) \\ &= \big((A_{kk}^{11}y_k^1, x_k^1) + (A_{kk}^{12}y_k^2, x_k^1)\big)\bar{z}_k^1 + \big((A_{kk}^{21}y_k^1, x_k^2) + (A_{kk}^{22}y_k^2, x_k^2)\big)\bar{z}_k^2 \\ &= (A_{kk}^{11}y_k^1 + A_{kk}^{12}y_k^2, y_k^1) + (A_{kk}^{21}y_k^1 + A_{kk}^{22}y_k^2, y_k^2) \\ &= \left(\begin{pmatrix} A_{kk}^{11}y_k^1 + A_{kk}^{12}y_k^2 \\ A_{kk}^{21}y_k^1 + A_{kk}^{22}y_k^2 \end{pmatrix}, \begin{pmatrix} y_k^1 \\ y_k^2 \end{pmatrix} \right) = (A_{kk}y_k, y_k) = (\mathcal{A}_y)_{kk}. \end{aligned}$$

The proof for the other cases is similar. □

As an example for Theorem 3.5 we consider the 4×4 matrix

$$\mathcal{A}_5 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & -1 & \mathrm{i} & 5\mathrm{i} \\ -1 & -2 & -5\mathrm{i} & \mathrm{i} \end{pmatrix}. \tag{3.1}$$

Its block numerical ranges with respect to the four successively refined decompositions $\mathbb{C}^4 = \mathbb{C}^2 \times \mathbb{C}^2 = \mathbb{C}^2 \times \mathbb{C} \times \mathbb{C} = \mathbb{C} \times \mathbb{C} \times \mathbb{C} \times \mathbb{C}$ (the first one being the numerical range and the last one being the spectrum) are shown in Figure 4 below.

4. Estimate of the resolvent. It is well known that the resolvent of $\mathcal{A}$ can be estimated in terms of the numerical range of $\mathcal{A}$ by

$$\|(\mathcal{A} - \lambda)^{-1}\| \leq \frac{1}{\operatorname{dist}(\lambda, W(\mathcal{A}))}, \quad \lambda \notin \overline{W(\mathcal{A})}.$$

A corresponding result for the block numerical range will be proved in the following.

LEMMA 4.1. *Let $\mathcal{A}_{(\cdot)}$ be uniformly bounded from below; i.e., assume there exists a $\delta > 0$ such that for all $x \in \mathcal{S}^n$*

$$\|\mathcal{A}_x\alpha\| \geq \delta\|\alpha\|, \quad \alpha \in \mathbb{C}^n. \tag{4.1}$$

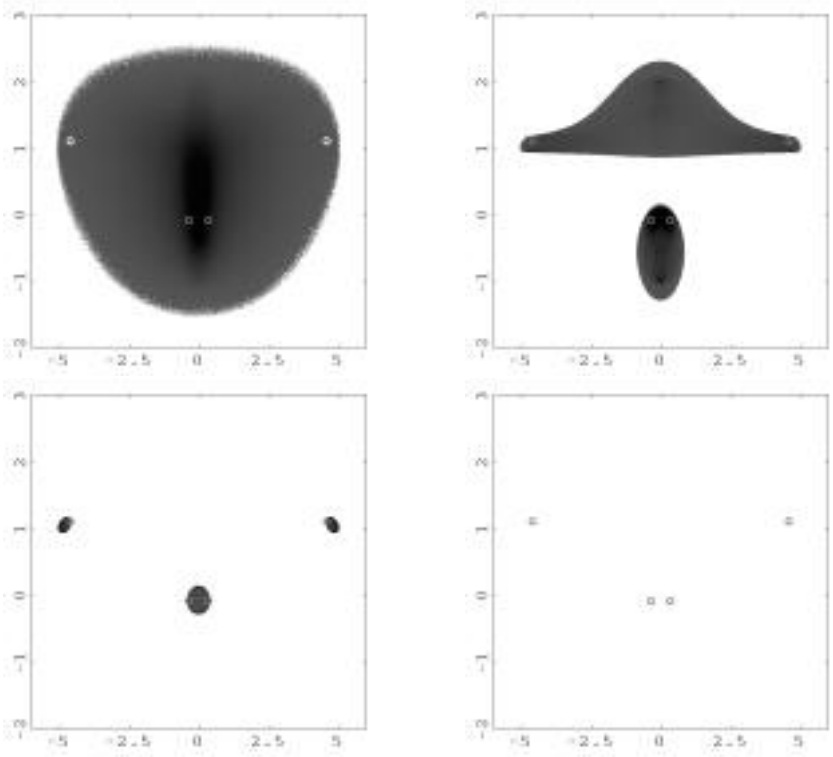


FIG. 4. $W_{\mathbb{C}^4}(\mathcal{A}_5)$, $W_{\mathbb{C}^2\times\mathbb{C}^2}(\mathcal{A}_5)$, $W_{\mathbb{C}^2\times\mathbb{C}\times\mathbb{C}}(\mathcal{A}_5)$, *and* $W_{\mathbb{C}\times\mathbb{C}\times\mathbb{C}\times\mathbb{C}}(\mathcal{A}_5)$.

Then

$$\|\mathcal{A}y\| \geq \delta\|y\|, \quad y \in \mathcal{H}.$$

In particular, if $\mathcal{A}$ *and* $\mathcal{A}_x$ *are invertible with* $\|\mathcal{A}_x^{-1}\| \leq \gamma$ *for all* $x \in \mathcal{S}^n$, *then also* $\|\mathcal{A}^{-1}\| \leq \gamma$.

Proof. Let $y = (y_1, \ldots, y_n) \in \mathcal{H}$ be arbitrary and write $y_i = \|y_i\|\,\hat{y}_i$ with $\|\hat{y}_i\| = 1$, $i = 1, \ldots, n$. Then we have $\hat{y} := (\hat{y}_1, \ldots, \hat{y}_n) \in \mathcal{S}^n$, and by (4.1) we conclude, with $\alpha := (\|y_1\|, \ldots, \|y_n\|) \in \mathbb{C}^n$, that

$$\begin{aligned}
\delta^2\|y\|^2 &= \delta^2(\|y_1\|^2 + \cdots + \|y_n\|^2) = \delta^2\|\alpha\|^2 \leq \|\mathcal{A}_{\hat{y}}\alpha\|^2 \\
&= \left\| \begin{pmatrix} (A_{11}\hat{y}_1, \hat{y}_1)\|y_1\| + \cdots + (A_{1n}\hat{y}_n, \hat{y}_1)\|y_n\| \\ \vdots \\ (A_{n1}\hat{y}_1, \hat{y}_n)\|y_1\| + \cdots + (A_{nn}\hat{y}_n, \hat{y}_n)\|y_n\| \end{pmatrix} \right\|^2 = \sum_{i=1}^n \left| \left(\sum_{j=1}^n A_{ij}y_j, \hat{y}_i \right) \right|^2 \\
&\leq \sum_{i=1}^n \left\| \sum_{j=1}^n A_{ij}y_j \right\|^2 \|\hat{y}_i\|^2 = \sum_{i=1}^n \left\| \sum_{j=1}^n A_{ij}y_j \right\|^2 = \|\mathcal{A}y\|^2. \qquad \square
\end{aligned}$$

THEOREM 4.2. *The resolvent of* $\mathcal{A}$ *admits the estimate*

$$\|(\mathcal{A} - \lambda)^{-1}\| \leq \frac{(\|\mathcal{A}\| + |\lambda|)^{n-1}}{\operatorname{dist}(\lambda, W^n(\mathcal{A}))^n}, \quad \lambda \notin \overline{W^n(\mathcal{A})}. \tag{4.2}$$

More exactly, if $K_1, \dots, K_s$ are the components of $W^n(\mathcal{A})$, then there are integers n_i, $i = 1, \dots, s$, with $\sum_{i=1}^s n_i = n$ such that

$$\|(\mathcal{A} - \lambda)^{-1}\| \leq \frac{(\|\mathcal{A}\| + |\lambda|)^{n-1}}{\prod_{i=1}^s \operatorname{dist}(\lambda, K_i)^{n_i}}, \quad \lambda \notin \overline{W^n(\mathcal{A})}; \tag{4.3}$$

in particular, if $W^n(\mathcal{A})$ consists of n components, then

$$\|(\mathcal{A} - \lambda)^{-1}\| \leq \frac{(\|\mathcal{A}\| + |\lambda|)^{n-1}}{\prod_{i=1}^n \operatorname{dist}(\lambda, K_i)}, \quad \lambda \notin \overline{W^n(\mathcal{A})}.$$

Proof. Let $\lambda \notin \overline{W^n(\mathcal{A})}$. If $K_1, \dots, K_s$ are the components of $W^n(\mathcal{A})$, then there are integers n_i, $i = 1, \dots, s$, with $\sum_{i=1}^s n_i = n$ such that each matrix $\mathcal{A}_x$, $x \in \mathcal{S}^n$, has exactly n_i eigenvalues in K_i for all $i = 1, \dots, s$. Now let $x \in \mathcal{S}^n$ and let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $\mathcal{A}_x$. Then there exists a partition $I_1 \dot\cup \cdots \dot\cup I_s = \{1, \dots, n\}$ so that $\lambda_i \in K_j$ if and only if $i \in I_j$. Then $n_j = \#I_j$, $j = 1, \dots, s$, and

$$\det(\mathcal{A}_x - \lambda) = |\lambda - \lambda_1| \cdots |\lambda - \lambda_n| = \prod_{j=1}^s \prod_{i \in I_j} |\lambda - \lambda_i| \geq \prod_{j=1}^s \operatorname{dist}(\lambda, K_j)^{n_j} > 0$$

for $x \in \mathcal{S}^n$ since $\lambda \notin \overline{W^n(\mathcal{A})}$. In particular, $\mathcal{A}_x - \lambda$ is invertible. Lemma 2.4 then implies that

$$\|(\mathcal{A} - \lambda)_x^{-1}\| \leq \frac{\|(\mathcal{A} - \lambda)_x\|^{n-1}}{|\det(\mathcal{A}_x - \lambda)|} \leq \frac{(\|\mathcal{A}_x\| + |\lambda|)^{n-1}}{\prod_{i=1}^s \operatorname{dist}(\lambda, K_i)^{n_i}} \leq \frac{(\|\mathcal{A}\| + |\lambda|)^{n-1}}{\prod_{i=1}^s \operatorname{dist}(\lambda, K_i)^{n_i}} \tag{4.4}$$

for all $x \in \mathcal{S}^n$. However, $\mathcal{A} - \lambda$ is invertible since $\lambda \notin \overline{W^n(\mathcal{A})}$ and thus $\lambda \in \rho(A)$. Together with (4.4), the second assertion of the theorem follows from Lemma 4.1.

The first and third estimates are immediate consequences of the second inequality. ∎

The estimate of the resolvent in terms of the numerical range implies that in a boundary point of the numerical range there are no associated vectors. In the following we prove the analogue of this statement for the block numerical range.

Definition 4.3. *Let $W \subset \mathbb{C}$. A point $\mu \in \mathbb{C}$ is said to have the* exterior cone property with respect to W *if there exist a closed cone V with positive aperture and vertex μ and an $r > 0$ such that for the closed ball $K_r(\mu) := \{\lambda \in \mathbb{C} : |\lambda - \mu| \leq r\}$,*

$$V \cap K_r(\mu) \cap \overline{W} = \{\mu\}.$$

Note that a point which has the exterior cone property with respect to W lies necessarily on the boundary ∂W of W.

Proposition 4.4. *Let $\lambda_0 \in \partial W^n(\mathcal{A})$ have the exterior cone property. Then the length of the Jordan chains at λ_0 is at most n.*

More exactly, if $W^n(\mathcal{A})$ consists of components $K_1, \dots, K_s$ such that $\overline{K_i}$ are disjoint for $i = 1, \dots, s$, the integers n_i, $i = 1, \dots, s$, are as in Theorem 4.2, and $\lambda_0 \in \overline{K_j}$, then the length of the Jordan chains at λ_0 is at most n_j.

In particular, if $\overline{W^n(\mathcal{A})}$ consists of n components, then $\mathcal{A}$ has no associated vectors in λ_0.

Proof. Again, it is sufficient to prove the second statement. Assume $\lambda_0 \in \overline{K_j}$ has the exterior cone property with respect to $W^n(\mathcal{A})$ and suppose there is a Jordan

chain $\{x_1,\dots,x_{n_j+1}\}$ of $\mathcal{A}$ in λ_0 of length n_j+1. Then there is a $\delta'>0$ and a constant $C>0$ such that for $\lambda\in\rho(\mathcal{A})\cap B_{\delta'}(\lambda_0)$,

$$\|(\mathcal{A}-\lambda)^{-1}x_{n_j+1}\| = \left\|-\sum_{i=1}^{n_j+1}\frac{x_i}{(\lambda-\lambda_0)^{n_j+2-i}}\right\| \geq \frac{C}{|\lambda-\lambda_0|^{n_j+1}}.$$

Now let V be a closed cone with vertex λ_0 and aperture 2α, where $0<2\alpha<\pi$ and $r>0$ are such that $V\cap B_r(\lambda_0)\cap\overline{W^n(\mathcal{A})}=\{\lambda_0\}$. Furthermore, define $d:=\min_{i\neq j}\operatorname{dist}(\lambda_0,K_i)>0$ and set $\delta:=\min\{r,\delta',d\}/2$. Then we have, for any $\lambda\neq\lambda_0$ on the axis of the cone with $|\lambda-\lambda_0|<\delta$,

$$\operatorname{dist}(\lambda,K_j)\geq\operatorname{dist}\big(\lambda,\mathbb{C}\setminus(V\cap B_{2\delta}(\lambda_0))\big)=|\lambda-\lambda_0|\sin\alpha,$$

where we have used $\overline{K_j}\subset\overline{W^n(\mathcal{A})}\subset\mathbb{C}\setminus(V\cap(B_{2\delta}(\lambda_0)\setminus\{\lambda_0\}))$. Hence, with $C':=C(\sin\alpha)^{n_j+1}>0$,

$$\frac{C'}{\operatorname{dist}(\lambda,K_j)^{n_j+1}}\leq\frac{C}{|\lambda-\lambda_0|^{n_j+1}}\leq\|(\mathcal{A}-\lambda)^{-1}x_{n_j+1}\|.$$

On the other hand, by estimate (4.3) in Theorem 4.2,

$$\|(\mathcal{A}-\lambda)^{-1}x_{n_j+1}\|\leq\frac{(\|\mathcal{A}\|+|\lambda|)^{n-1}}{\prod_{i=1}^s\operatorname{dist}(\lambda,K_i)^{n_i}}\|x_{n_j+1}\|\leq\frac{(\|\mathcal{A}\|+|\lambda_0|+\delta)^{n-1}\|x_{n_j+1}\|}{(\prod_{i\neq j}\delta^{n_i})\operatorname{dist}(\lambda,K_j)^{n_j}},$$

and therefore, with $C'':=(\prod_{i\neq j}\delta^{n_i})^{-1}(\|\mathcal{A}\|+|\lambda_0|+\delta)^{n-1}\|x_{n_j+1}\|>0$, we have

$$0<\frac{C'}{C''}\leq\operatorname{dist}(\lambda,K_j)\longrightarrow 0,\quad\lambda\to\lambda_0,$$

which is a contradiction. ∎

5. The block numerical range of a companion operator. In this section we consider a special case of $n\times n$ block operator matrices, namely, companion operators of operator polynomials of degree n, and we study the connection between the block numerical range of a companion operator and the numerical range of the corresponding operator polynomial.

To this end, let $A_i\in L(H)$, $i=0,\dots,n-1$, $A:=(A_0,\dots,A_{n-1})$, and let the operator polynomial P_A be given by

$$P_A(\lambda):=\lambda^nI+\lambda^{n-1}A_{n-1}+\cdots+\lambda A_1+A_0,\quad\lambda\in\mathbb{C}.$$

The numerical range of P_A is defined as (see [8, sect. 26.1])

$$W(P_A):=\{\lambda\in\mathbb{C}:\exists\,x\in H,\ \|x\|=1\ \ (P_A(\lambda)x,x)=0\}.$$

The companion operator of P_A is the $n\times n$ block operator matrix in $\mathcal{H}=H^n$ given by

$$\mathcal{C}^A:=\begin{pmatrix}0&I&&&0\\&0&I&&\\&&\ddots&\ddots&\\&&&0&I\\-A_0&-A_1&\cdots&-A_{n-2}&-A_{n-1}\end{pmatrix}.$$

THEOREM 5.1. *The numerical range of the operator polynomial P_A is contained in the block numerical range of its companion operator $\mathcal{C}^A$:*

$$W(P_A) \subset W^n(\mathcal{C}^A).$$

Proof. Let $\lambda \in W(P_A)$. Then there exists an $x \in H$, $\|x\| = 1$, such that

$$\lambda^n + \lambda^{n-1}(A_{n-1}x, x) + \cdots + \lambda(A_1x, x) + (A_0x, x) = 0.$$

We will show that $\lambda \in \sigma_p(\mathcal{C}^A_{(x,\ldots,x)})$. Since $(x, \ldots, x) \in \mathcal{S}^n$, this will prove the theorem. We have

$$\det\Big(\mathcal{C}^A_{(x,\ldots,x)} - \lambda\Big) = \begin{vmatrix} -\lambda & 1 & & & 0 \\ & -\lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & -\lambda & 1 \\ -(A_0x, x) & -(A_1x, x) & \cdots & -(A_{n-2}x, x) & -(A_{n-1}x, x) - \lambda \end{vmatrix}.$$

If $\lambda = 0$, we have $(A_0x, x) = 0$, and hence the above determinant vanishes as required. If $\lambda \neq 0$, we eliminate the entries above the diagonal successively by adding $1/\lambda$ times the kth column to the $(k+1)$th column for $k = 1, \ldots, n-1$. This shows that

$$\begin{aligned} \det\Big(\mathcal{C}^A_{(x,\ldots,x)} - \lambda\Big) &= (-\lambda)^{n-1}\bigg(-(A_{n-1}x, x) - \lambda - \frac{1}{\lambda}(A_{n-2}x, x) - \cdots - \frac{1}{\lambda^{n-1}}(A_0x, x)\bigg) \\ &= (-1)^n(P_A(\lambda)x, x) = 0. \qquad \square \end{aligned}$$

As an illustration for Theorem 5.1, we consider again the matrix $\mathcal{A}_5$ from (3.1) which is the companion operator of the quadratic polynomial

$$P_5(\lambda) := \lambda^2 I_2 + \lambda \begin{pmatrix} -\mathrm{i} & -5\mathrm{i} \\ 5\mathrm{i} & -\mathrm{i} \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \lambda \in \mathbb{C}.$$

Figure 5 shows the pencil numerical range of P_5 which is contained in the quadratic numerical range of the companion operator $\mathcal{A}_5$ with respect to the decomposition $\mathbb{C}^2 \times \mathbb{C}^2$.

Next we are going to show that if H is finite dimensional, $H = \mathbb{C}^k$, then, up to the origin, the numerical range of P_A coincides with a higher degree block numerical range of its companion operator. To this end, we consider $\mathcal{C}^A$ with respect to a refined decomposition of $\mathcal{H} = H^n = \mathbb{C}^{nk}$.

THEOREM 5.2. *If we consider the companion operator $\mathcal{C}^A$ with respect to the decomposition*

$$\mathbb{C}^{nk} = \overbrace{\mathbb{C} \times \cdots \times \mathbb{C}}^{(n-1)k} \times \mathbb{C}^k, \tag{5.1}$$

then we have

$$W_{\mathbb{C}\times\cdots\times\mathbb{C}\times\mathbb{C}^k}(\mathcal{C}^A) = W^{(n-1)k+1}(\mathcal{C}^A) = \begin{cases} W(P_A), & n = 1, \\ W(P_A) \cup \{0\}, & n \geq 2. \end{cases}$$

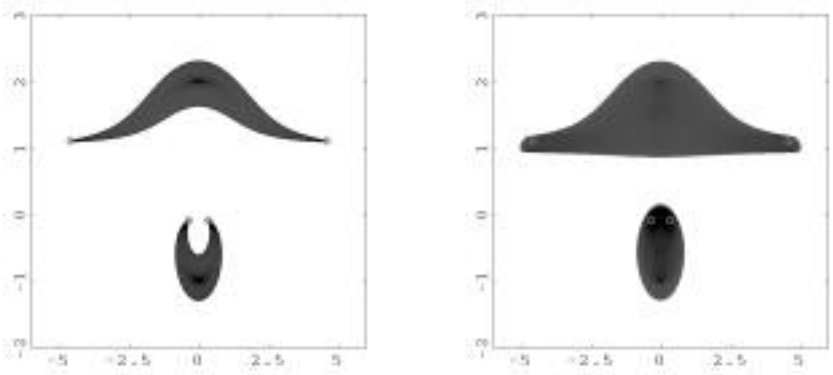


FIG. 5. $W(P_5)$ *and* $W_{\mathbb{C}^2\times\mathbb{C}^2}(\mathcal{A}_5)$.

Proof. For $n = 1$ the assertion is immediate. If $n > 1$, then, with respect to decomposition (5.1), $\mathcal{C}^A$ has the block operator representation

$$\begin{pmatrix}
0 & \cdots & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 & 0_{1,k} \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \cdots & \vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & & 0 & \cdots & 0 & 0_{1,k} \\
0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & & 0 & \cdots & 0 & 0_{1,k} \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \cdots & \vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & & 0 & \cdots & 0 & 0_{1,k} \\
& \vdots & & & \vdots & & & \vdots & & \ddots & & \vdots & & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & & 1 & \cdots & 0 & 0_{1,k} \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \cdots & \vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 1 & 0_{1,k} \\
0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 & e_1 \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \cdots & \vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & & 0 & \cdots & 0 & e_k \\
-A_0^{(1)} & \cdots & -A_0^{(k)} & -A_1^{(1)} & \cdots & -A_1^{(k)} & -A_2^{(1)} & \cdots & -A_2^{(k)} & \cdots & -A_{n-2}^{(1)} & \cdots & -A_{n-2}^{(k)} & -A_{n-1}
\end{pmatrix},$$

where $0_{1,k} = (0 \cdots 0) \in L(\mathbb{C}^k, \mathbb{C})$ is the zero vector, $e_j = (0 \cdots 1 \cdots 0) \in L(\mathbb{C}^k, \mathbb{C})$ is the jth unit vector, $j = 1, \ldots, k$, and

$$A_i^{(j)} = \begin{pmatrix} a_{1j}^{(i)} \\ \vdots \\ a_{kj}^{(i)} \end{pmatrix} \in L(\mathbb{C}, \mathbb{C}^k)$$

is the jth column of $A_i = (a_{st}^{(i)})_{s,t=1}^k$, $i = 1, \ldots, n$, $j = 1, \ldots, k$. Now let

$$x = \big(x_0^{(1)}, \ldots, x_0^{(k)}, \ldots, x_{n-2}^{(1)}, \ldots, x_{n-2}^{(k)}, (\xi_1, \ldots, \xi_k)\big) \in \mathbb{C} \times \cdots \times \mathbb{C} \times \mathbb{C}^k$$

with $|x_0^{(1)}| = \cdots = |x_{n-2}^{(k)}| = \|\xi\| = 1$, where $\xi := (\xi_1, \ldots, \xi_n)$. By similar manipulations of determinants as in the proof of the previous theorem, one can show that

$$\det(\mathcal{C}_x^A - \lambda) = (-1)^n \lambda^{(n-1)(k-1)} (P_A(\lambda)\xi, \xi),$$

which implies that $W^{(n-1)k+1}(\mathcal{C}^A) \setminus \{0\} = W(P_A) \setminus \{0\}$. ∎

As an example for Theorem 5.2 we consider the quadratic operator polynomial (compare [4])

$$P_6(\lambda) := \lambda^2 I_2 + \lambda \begin{pmatrix} 0 & 2.8\mathrm{i} \\ -2.8\mathrm{i} & 0 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \lambda \in \mathbb{C},$$

with its companion operator $\mathcal{A}_6$ decomposed as

$$\mathcal{A}_6 = \left(\begin{array}{c|c|cc} 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline -2 & -1 & 0 & -2.8\mathrm{i} \\ -1 & -2 & 2.8\mathrm{i} & 0 \end{array}\right).$$

In Figure 6 the cubic numerical range of $\mathcal{A}_6$ with respect to this decomposition and the pencil numerical range of P_6 are displayed. Note that the pencil numerical range on the right does not contain the origin, whereas the cubic numerical range on the left does.

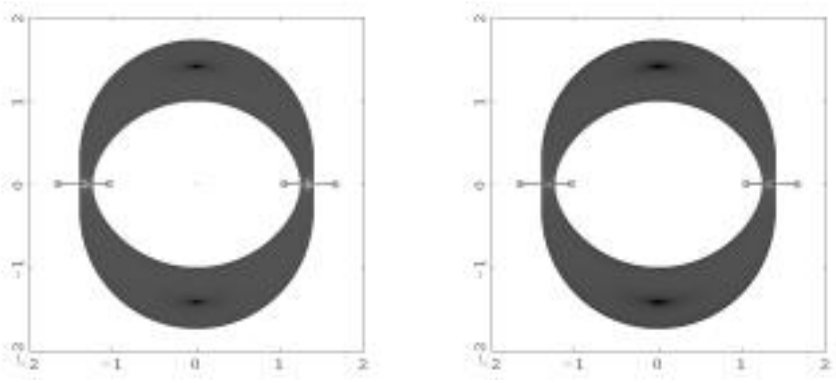


FIG. 6. $W_{\mathbb{C}\times\mathbb{C}\times\mathbb{C}^2}(\mathcal{A}_6)$ *and* $W(P_6)$.

REFERENCES

[1] K. E. GUSTAFSON AND D. K. M. RAO, *Numerical Range*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1997.

[2] T. KATO, *Estimation of iterated matrices, with application to the von Neumann condition*, Numer. Math., 2 (1960), pp. 22–29.

[3] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, 1995.

[4] P. LANCASTER, J. MAROULAS, AND P. ZIZLER, *The numerical range of selfadjoint matrix polynomials*, in Contributions to Operator Theory in Spaces with an Indefinite Metric, Oper. Theory Adv. Appl. 106, Birkhäuser, Basel, 1998, pp. 291–308.

[5] H. LANGER, A. S. MARKUS, V. I. MATSAEV, AND C. TRETTER, *A new concept for block operator matrices: The quadratic numerical range*, Linear Algebra Appl., 330 (2001), pp. 89–112.

[6] H. LANGER, A. S. MARKUS, AND C. TRETTER, *Corners of numerical ranges*, in Recent Advances in Operator Theory, Oper. Theory Adv. Appl. 124, Birkhäuser, Basel, 2001, pp. 385–400.

[7] H. LANGER AND C. TRETTER, *Spectral decomposition of some nonselfadjoint block operator matrices*, J. Oper. Theory, 39 (1998), pp. 339–359.

[8] A. S. MARKUS, *Introduction to the Spectral Theory of Polynomial Operator Pencils*, AMS, Providence, RI, 1988.

[9] M. WAGENHOFER, *Der blocknumerische Wertebereich*, Diploma thesis, Regensburg, Germany, 2000.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1018–1037

BLIND DECONVOLUTION USING A REGULARIZED STRUCTURED TOTAL LEAST NORM ALGORITHM*

ARMIN PRUESSNER† AND DIANNE P. O'LEARY‡

Abstract. Rosen, Park, and Glick proposed the structured total least norm (STLN) algorithm for solving problems in which both the matrix and the right-hand side contain errors. We extend this algorithm for ill-posed problems by adding regularization, and we use the resulting algorithm to solve blind deconvolution problems as encountered in image deblurring when both the image and the blurring function have uncertainty. The resulting regularized structured total least norm (RSTLN) algorithm preserves any affine structure of the matrix and minimizes a discrete ℓ_p-norm measure of the error, where $p = 1$, 2, or ∞. We demonstrate the effectiveness of these algorithms for blind deconvolution.

Key words. least squares, total least squares, total least norm, structured total least norm, minimization, regularization, ill-posed problem, 1-norm, 2-norm, ∞-norm, overdetermined linear system, blind deconvolution, image deblurring, boundary conditions, constrained total least squares

AMS subject classifications. 65F22, 65K10, 90C05

PII. S0895479801395446

1. Introduction and background.

Most image recording devices fail to record the intensity of a given image scene exactly. Each recorded image section (or pixel) describing the corresponding scene has errors in the form of either random noise, or blurring, or both. *Blurring* occurs when the recorded intensity of a given pixel is in effect influenced by the intensity of neighboring sections. Because of these imperfections in recorded images, it is often necessary to apply deblurring techniques to obtain clearer images.

The problem of image deblurring [6, 11] is modeled as an integral equation of the first kind,

$$\int_\Omega a(s,t)x(t)\ dt = b(s) - r(s) = b_c(s), \tag{1.1}$$

where $s, t \in \mathbf{R}^2$ are the spatial coordinates, Ω the domain or (nonzero) support of the image, $x : \mathbf{R}^2 \to \mathbf{R}$ the true image, $a : \mathbf{R}^4 \to \mathbf{R}$ the point spread function, and $r : \mathbf{R}^2 \to \mathbf{R}$ random noise. The function $b(s)$ is the observed, blurred, noisy image, and $b_c(s)$ is the noiseless blurred image.

In particular, if it is assumed that $a(s,t)$ is *spatially invariant*, that is, its effect depends only on the spatial distance between s and t, then the preceding equation represents a convolution integral, where $a(s,t) = a(s-t)$. In this case, $b_c(s)$ is the result of convolving $a(s)$ and $x(s)$.

*Received by the editors September 25, 2001; accepted for publication (in revised form) by P. C. Hansen September 10, 2002; published electronically February 25, 2003. This work was supported in part by the National Science Foundation under grant CCR-97-32022 and by the Office of Naval Research under grant N000140110181.

http://www.siam.org/journals/simax/24-4/39544.html

†Applied Mathematics and Scientific Computation Program, University of Maryland, College Park, MD 20742. Current address: GAMS Development Corporation, 1217 Potomac Street NW, Washington, DC 20007 (armin@gams.com).

‡Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (oleary@cs.umd.edu).

Since recording devices make only a finite number of measurements, the imaging model can be discretized and (1.1) can be written as a matrix equation. The discretized model is

$$Ax = b - r, \tag{1.2}$$

where the matrix A is the discretized counterpart of $a(s,t)$, and x and b also are the discretized versions of the corresponding continuous functions. If the blurring function a is assumed to be spatially invariant, then the matrix A has a special structure: for 1-dimensional signals it is Toeplitz and for 2-dimensional signals it is block Toeplitz with Toeplitz blocks.

If the cause of the blur, and hence a, is not known exactly, then our estimate of A has errors and the problem is known as *blind deconvolution*. In this case the model in (1.2) should be replaced by

$$(A + E)x = b - r, \tag{1.3}$$

a problem of the total least norm variety. If the matrix A has no special structure and the error $\|[E, r]\|_p$ is measured using the Frobenius norm, then the problem can be solved using the total least squares (TLS) method [5]. For image processing problems, the matrix A has a special structure, and it is desirable to enforce the same structure on the error matrix E. Rosen, Park, and Glick [23] developed the structured total least norm (STLN) method to solve such problems.

While STLN is useful for many structured linear problems, the blind deconvolution problem as encountered for image deblurring is generally ill-posed [9]. In particular, the matrix A is often ill-conditioned, resulting in poor recovered images when STLN is applied.

Regularization methods must be implemented in order to stabilize STLN and to obtain useful results. In this paper it is shown how to implement Tikhonov regularization [20, 26] to arrive at the regularized structured total least norm (RSTLN) algorithm. Implementation of Tikhonov regularization for constrained TLS problems had been developed previously [15, 17]. The first of these works predated the work of Rosen, Park, and Glick on the simpler problem. These works, however, focused solely on the 2-norm case. The contributions herein are the extension for $p = 1$ and $p = \infty$ norms and the comparison of methods. In the $p = 1$ and $p = \infty$ cases, the main computational task lies in solving a linear program (LP).

The paper is structured as follows. In the next section the STLN method is introduced and derived. In section 3 the general RSTLN method is introduced and derivations are presented for the $p = 1$, 2, and ∞ cases. Finally, in section 4 we present numerical results, and in section 5 we draw conclusions.

2. Derivation of the STLN method. In order to understand the RSTLN method, a brief derivation of STLN based on [23] is given. For a more thorough derivation, the reader is referred to [23] and [12].

2.1. TLS and STLN. The TLS [5] formulation for solving problems as in (1.3) is to find a matrix E and a vector r such that

$$\|[E,\ r]\|_F \tag{2.1}$$

is minimized, where F denotes the Frobenius norm and $r = b - (A + E)x$ is the residual. If the matrix A has a special structure which the user wants to enforce

on E, then the TLS formulation is not applicable. Instead, the STLN formulation proves useful.

As in [23], assume that the matrix $E \in \mathbf{R}^{m\times n}$ is parameterized by elements of the vector $\alpha \in \mathbf{R}^q$, $q < mn$. Then the residual is a function of α and x. The STLN formulation is to find vectors α and x such that

$$\left\| \begin{array}{c} r(\alpha, x) \\ D\alpha \end{array} \right\|_p \tag{2.2}$$

is minimized, where $p = 1$, 2, or ∞ and D is a diagonal *weighting matrix* through which the size of α is measured. Note that the norm in (2.2) is a norm over the space of structured matrices crossed with vectors in $\mathbf{R}^m$. For $p = 2$, it is the same as the Frobenius norm in (2.1) but, for $p = 1$ and $p = \infty$, it is not equivalent to any matrix norm.

If the elements of E are linear functions of the parameters α, then there exists a matrix X parameterized by x such that

$$X\alpha = Ex. \tag{2.3}$$

For a detailed description on construction of the matrix X, see [23] and [12]. Note that if the matrix E is structured, then so is X.

Now let Δx and ΔE denote small changes in x and E, respectively. Then

$$X\Delta\alpha = (\Delta E)x. \tag{2.4}$$

If we expand $r(\alpha, x)$ in a Taylor series about $[\alpha^T \; x^T]^T$ and ignore second order and higher terms, we have

$$\begin{array}{rcl} r(\alpha + \Delta\alpha, x + \Delta x) & \approx & b - (A + E)x - X\Delta\alpha - (A + E)\Delta x \\ & = & r(\alpha, x) - X\Delta\alpha - (A + E)\Delta x. \end{array} \tag{2.5}$$

Hence, we have a linearization of (2.2),

$$\min_{\Delta\alpha,\Delta x} \left\| \begin{bmatrix} X & A + E \\ D & 0 \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \end{pmatrix} + \begin{pmatrix} -r \\ D\alpha \end{pmatrix} \right\|_p. \tag{2.6}$$

The general idea behind the STLN method is to start with some initial estimates for x and E, solve the minimization problem in (2.6) for $\Delta\alpha$ and Δx, set $x = x + \Delta x$ and $\alpha = \alpha + \Delta\alpha$, and update the residual r and the matrices E and X. The procedure is repeated iteratively until $\|\Delta\alpha\|$ and $\|\Delta x\|$ are below a specified tolerance, at which point the algorithm has converged to a solution. For a detailed description the reader is referred to [23].

3. Derivation of RSTLN. In order to make STLN more robust in the presence of noise (as is encountered in most image deblurring applications), a form of regularization must be introduced. The method of Tikhonov [26] is implemented herein, which prevents the solution x from becoming too large. In particular, (2.2) can be modified to arrive at the RSTLN algorithm. The new problem formulation is to find vectors α and x so that

$$\left\| \begin{array}{c} r(\alpha, x) \\ D\alpha \\ \lambda x \end{array} \right\|_p \tag{3.1}$$

TABLE 3.1

RSTLN Algorithm

1. Set $E = 0_{m\times n}$ and $\alpha = 0_{q\times 1}$.
2. Compute x by $\min_x \|Ax - b\|_p$ (for $p = 2$ this is just least squares).
3. Compute X from x and the residual $r = b - Ax$.
4. For $k = 1, 2, \ldots$ until $\|\Delta x\|, \|\Delta\alpha\| \le \epsilon$ repeat steps 4.1–4.3.
   - 4.1. Solve
   $$\min_{\Delta\alpha,\Delta x} \left\| \begin{bmatrix} X & A+E \\ D & 0 \\ 0 & \lambda I \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \end{pmatrix} + \begin{pmatrix} -r(\alpha, x) \\ D\alpha \\ \lambda x \end{pmatrix} \right\|_p .$$
   - 4.2. Set $x = x + \Delta x$ and $\alpha = \alpha + \Delta\alpha$.
   - 4.3. Construct E from α, and X from x and compute $r = b - (A + E)x$.
5. The recovered image is x and the recovered blurring matrix $(A + E)$.

is minimized, where λ is a positive scalar known as the regularization parameter and $p = 1$, 2, or ∞. More generally, we could replace λx by λLx, where L is an operator chosen to force some desirable property on the solution x. For example, L might be a difference operator if we want a smooth image; see, for example [8, sect. 4.3]. For simplicity, we will write the algorithm for the case $L = I$, although the generalization is straightforward.

Using the relation in (2.5) and similar reasoning as for the STLN method, the linearization of (3.1) results in

$$\min_{\Delta\alpha,\Delta x} \left\| \begin{bmatrix} X & A+E \\ D & 0 \\ 0 & \lambda I \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \end{pmatrix} + \begin{pmatrix} -r \\ D\alpha \\ \lambda x \end{pmatrix} \right\|_p . \tag{3.2}$$

The general RSTLN algorithm (for arbitrary norm p) is listed in Table 3.1. We remark that Tikhonov regularization can be added in the same manner to the structured nonlinear total least norm (SNTLN) method [24], a variant of STLN for structured nonlinear parameter estimation problems. The resulting regularized algorithm is similar to RSTLN and may be the focus of future work.

3.1. RSTLN for $p = 2$. The minimization problem in the RSTLN formulation is equivalent to minimizing the function

$$\phi(\alpha, x) = \frac{1}{2}\|r(\alpha, x)\|_2^2 + \frac{1}{2}\|D\alpha\|_2^2 + \frac{1}{2}\|\lambda x\|_2^2. \tag{3.3}$$

The 2-norm case has the property of differentiability so that Gauss–Newton theory is applicable. Using similar reasoning as in [23] for the STLN method, it follows that step 4.1 is a Gauss–Newton method which approximates the Hessian of $\phi(\alpha, x)$ by the positive definite matrix $M^T M$, where

$$M = \begin{bmatrix} X & A+E \\ D & 0 \\ 0 & \lambda I \end{bmatrix}. \tag{3.4}$$

See also [3].

The least squares normal equations can be solved using the conjugate gradient method, where the Toeplitz (or block Toeplitz with Toeplitz blocks) structure of M is exploited. In particular, the FFT is used for efficient computation of matrix-vector products.

Another, more efficient, approach for $p = 2$ may be to apply the techniques of [14] for the nonregularized STLN to RSTLN. In particular, an algorithm based on the generalized Schur algorithm [16] for solving least squares problems is used which exploits the structure of the matrix

$$\begin{bmatrix} X & A+E \\ D & 0 \end{bmatrix}. \tag{3.5}$$

Since the RSTLN matrix M has a similar structure to this, the method in [14] should be applicable. This may be the focus of future work.

3.2. RSTLN for $p = \infty$. For both the $p = 1$ and $p = \infty$ cases, step 4.1 of the RSTLN algorithm is an LP. To see this, an approach similar to that in [23] is used.

Let us first consider the derivation for $p = \infty$. Suppose the original image in vector form is $x \in \mathbf{R}^{n\times 1}$, that $\alpha \in \mathbf{R}^{q\times 1}$, and that the residual $r \in \mathbf{R}^{m\times 1}$. Then the optimal function value in step 4.1 is $\bar{\sigma}$, where $\bar{\sigma}$ is determined from the LP

$$\begin{array}{lrcccl} \min\limits_{\Delta\alpha,\Delta x,\bar{\sigma}} & \bar{\sigma} & & & & \\ \text{subject to} & -\bar{\sigma}e_m & \le & X\Delta\alpha + (A+E)\Delta x - r & \le & \bar{\sigma}e_m, \\ & -\bar{\sigma}e_q & \le & D\Delta\alpha + D\alpha & \le & \bar{\sigma}e_q, \\ & -\bar{\sigma}e_n & \le & \lambda\Delta x + \lambda x & \le & \bar{\sigma}e_n, \end{array} \tag{3.6}$$

where $e_k \in \mathbf{R}^{k\times 1}$ is a vector of ones.

Using the matrix M we can write the linear programming problem in standard form,

$$\begin{array}{ll} \min\limits_{\Delta\alpha,\Delta x,\bar{\sigma}} & \bar{\sigma} \\ \text{subject to} & \begin{bmatrix} M & -e_{m+n+q} \\ -M & -e_{m+n+q} \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \\ \bar{\sigma} \end{pmatrix} \le \begin{pmatrix} r \\ -D\alpha \\ -\lambda x \\ -r \\ D\alpha \\ \lambda x \end{pmatrix}. \end{array} \tag{3.7}$$

Depending on the method used to solve the LP, it may be useful to consider the dual formulation. Setting $\sigma = -\bar{\sigma}$, it follows that the dual is

$$\begin{array}{ll} \min\limits_{y_i \ge 0} & r^T y_1 - \alpha^T D y_2 - \lambda x^T y_3 - r^T y_4 + \alpha^T D y_5 + \lambda x^T y_6 \\ \text{subject to} & \begin{bmatrix} M^T & -M^T \\ e_{m+n+q}^T & e_{m+n+q}^T \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \end{array} \tag{3.8}$$

where $y_1, y_4 \in \mathbf{R}^{m\times 1}$, $y_2, y_5 \in \mathbf{R}^{q\times 1}$, and $y_3, y_6 \in \mathbf{R}^{n\times 1}$. System (3.8) can be solved using any standard simplex or interior point method.

The reader should note that since the matrix M has a special structure (Toeplitz or block Toeplitz with Toeplitz blocks), any practical implementation of RSTLN for $p = 1$ or $p = \infty$ should exploit this property when solving the LP.

3.3. RSTLN for $p = 1$. The derivation for the $p = 1$ case is similar to the $p = \infty$ case. Again, let $\bar{\sigma}$ be the optimal function value in step 4.1. In particular, assuming x, α, and r are defined as previously, we have that $\bar{\sigma}$ is determined by

$$\begin{aligned} &\min_{\Delta\alpha,\Delta x,\bar{\sigma}} \quad \bar{\sigma} = \sum_{i=1}^{m} \bar{\sigma}_{1_i} + \sum_{i=1}^{q} \bar{\sigma}_{2_i} + \sum_{i=1}^{n} \bar{\sigma}_{3_i} \\ &\text{subject to} \quad \begin{array}{rcccl} -\bar{\sigma}_1 & \le & X\Delta\alpha + (A+E)\Delta x - r & \le & \bar{\sigma}_1, \\ -\bar{\sigma}_2 & \le & D\Delta\alpha + D\alpha & \le & \bar{\sigma}_2, \\ -\bar{\sigma}_3 & \le & \lambda\Delta x + \lambda x & \le & \bar{\sigma}_3, \end{array} \end{aligned} \tag{3.9}$$

where $\bar{\sigma}_1 \in \mathbf{R}^{m\times 1}$, $\bar{\sigma}_2 \in \mathbf{R}^{q\times 1}$, and $\bar{\sigma}_3 \in \mathbf{R}^{n\times 1}$. Using the matrix M we can write the LP as

$$\begin{aligned} &\min_{\Delta\alpha,\Delta x,\bar{\sigma}} \quad \bar{\sigma} = \sum_{i=1}^{m} \bar{\sigma}_{1_i} + \sum_{i=1}^{q} \bar{\sigma}_{2_i} + \sum_{i=1}^{n} \bar{\sigma}_{3_i} \\ &\text{subject to} \quad \begin{bmatrix} M & -I_{m+n+q} \\ -M & -I_{m+n+q} \end{bmatrix} \begin{pmatrix} \Delta\alpha \\ \Delta x \\ \bar{\sigma}_1 \\ \bar{\sigma}_2 \\ \bar{\sigma}_3 \end{pmatrix} \le \begin{pmatrix} r \\ -D\alpha \\ -\lambda x \\ -r \\ D\alpha \\ \lambda x \end{pmatrix}. \end{aligned} \tag{3.10}$$

As for the $p = \infty$ case, the user may want to use the dual formulation. Setting $\sigma = -\bar{\sigma}$, our formulation becomes

$$\begin{aligned} &\min_{y_i \ge 0} \quad r^T y_1 - \alpha^T D y_2 - \lambda x^T y_3 - r^T y_4 + \alpha^T D y_5 + \lambda x^T y_6 \\ &\text{subject to} \quad \begin{bmatrix} M^T & -M^T \\ I_{m+n+q} & I_{m+n+q} \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 0_{m\times 1} \\ 0_{q\times 1} \\ 0_{n\times 1} \\ e_m \\ e_q \\ e_n \end{pmatrix}, \end{aligned} \tag{3.11}$$

where all y_i are as defined previously for the ∞-norm case, and $0_{k\times 1}$ is a vector of zeros.

3.4. Convergence of RSTLN for $p = 1$ or $p = \infty$. As for the STLN problem, the function minimized in (3.1) is nonconvex, so that there is no guarantee that the RSTLN algorithm converges to a global minimum. For the $p = 2$ norm case the Gauss–Newton theory is applicable: a suitable line search method (see, for example, [3]) can be used to guarantee convergence to a local minimizer from any starting point.

For $p = 1$ and $p = \infty$, Gauss–Newton theory is not directly applicable since differentiability is lost. But the essential idea is the same as that for the $p = 2$ norm.

In particular, the solutions $\begin{bmatrix}\Delta\alpha^T & \Delta x^T\end{bmatrix}^T$ to the LPs given in (3.7) and (3.10) can be thought of as descent directions to the function in (3.1) for the respective p-norm. Again, in order to guarantee convergence to a local minimizer from any starting point, a line search method can be implemented.

4. Numerical results. In this section, experiments are presented to show that the RSTLN method deblurs images better than the STLN method. In particular, examples are shown comparing RSTLN and STLN for the $p = 1$, 2, and ∞ norms. We also compare our results with other blind deconvolution algorithms.

4.1. Experimental design.

4.1.1. Numerical issues. All of our code was written in MATLAB to take advantage of its image visualization capabilities.

We constructed our experiments by taking a known image x_{true} (stretched out to a vector by stacking the columns of the image) and a known blurring function A_{true} and using them to construct a blurred image $b_{true} = A_{true}x_{true}$. Then we added *n-bit noise* to the data: the elements of A_{noisy} were equal to those of A_{true} plus noise from a normal distribution with mean zero and standard deviation $\max_{i,j} A_{true}(i,j)/2^n$. Similarly, b_{noisy} was equal to $A_{noisy}x_{true}$ plus noise with standard deviation $\max_i b_{true}(i)/2^n$. Thus, the data perturbation can be measured by

$$\begin{aligned} pert(b) &= \|b_{noisy} - b_{true}\|_2/\|b_{true}\|_2, \\ pert(A) &= \|A_{noisy} - A_{true}\|_F/\|A_{true}\|_F. \end{aligned}$$

To evaluate the algorithms, we took the computed (recovered) image x_{rec} and the computed blurring function $(A+E)_{rec}$ and computed relative errors

$$\begin{aligned} err(x) &= \|x_{rec} - x_{true}\|_2/\|x_{true}\|_2, \\ err(A) &= \|(A+E)_{rec} - A_{true}\|_F/\|A_{true}\|_F, \\ err(b) &= \|b_{rec} - b_{true}\|_2/\|b_{true}\|_2, b_{rec} = (A+E)_{rec}x_{rec}. \end{aligned}$$

4.1.2. Implementation issues for STLN and RSTLN. For the STLN and RSTLN algorithms, a linear problem needs to be solved at each iteration; see step 4.1 of Table 3.1. For the $p = 2$ norm, we used the conjugate gradient least squares method to solve this problem. We set the conjugate gradient termination condition to a relative residual tolerance of 10^{-6} or 1000 iterations. This generally produces satisfactory accuracy to determine the descent direction, but for larger images the maximum number of iterations was sometimes taken.

For the $p = 1$ and $p = \infty$ cases we solved the LP in step 4.1 using the MATLAB function `linprog.m` with the *largescale* model employed. The function uses the LIPSOL [27] algorithm and is based on a primal-dual interior point method. Because of limitations in the MATLAB interface to LIPSOL, we were only able to set our stopping criteria to 10^{-2} to 10^{-3} compared with tolerances of 10^{-6} for the STLN experiments in [23]; a smaller tolerance caused LIPSOL to fail to converge. Even with this difficulty, RSTLN gives better results than STLN. Our current implementation is restricted to fairly small images because of the large number of constraints in the LP. While the constraint matrix M passed into `linprog.m` is sparse, its factorization generally is not. Hence, the LP solver as implemented in MATLAB is very memory intensive and currently restricts our test cases to images no larger than 100×100.

We stop the STLN or RSTLN iterations when the relative change in the recovered image and the recovered A matrix drops below some tolerance *tol*. At times

we stopped the RSTLN method prematurely before reaching the desired tolerance because for a higher number of iterations the reconstructed image was, in fact, deteriorating. This is a common phenomenon in the numerical solution of ill-posed problems shared, for example, by the Lucy–Richardson (LR) algorithm, and the number of iterations can be viewed as an additional regularization parameter [8, Chap. 6], [10]. Initial iterations tend to reconstruct the image while later ones tend to focus on the noise. Problems with low signal-to-noise ratios are particularly prone to such noise amplification; the basic problem is that we do not want to minimize the function in (2.1) but just drive its value down to noise level. Thus, in our experiments, a lower number of RSTLN iterations sometimes yielded a better recovered image than one recovered using more iterations, even if the latter yielded a better function value for (3.1) and satisfied lower tolerances.

The choice of the regularization parameter λ for algorithms such as RSTLN is a well-studied problem (see, for example, [4, 7, 18, 19] and [8, Chap. 7]). Ideally, the choice balances the need to stay close to the original noise-contaminated problem without causing its ill-conditioning to produce unacceptable noise in the solution. In our experiments, we were concerned with the best solution obtainable for any choice of parameter. We set $D = I$ and solved each problem for a wide range of values $\lambda > 0$, choosing the parameter resulting in the smallest value for the 2-norm of the image error. The solution was sometimes quite sensitive to this choice.

4.1.3. Comparison with other blind deconvolution methods. We compare RSTLN with two other blind deconvolution methods: the blind LR method and the APEX/SECB method of Carasso.

The blind LR algorithm is an extension of the well-known original LR method [13, 22] to problems in which the blurring function is unknown. The original iterative method was derived from Bayes' theorem and assumes that the blurred image, the original image, and the point spread function (PSF) are (possibly nonnormalized) probability density functions. The most common and efficient implementation makes use of the FFT to compute convolutions. This implicitly imposes periodic boundary conditions on the image.

The blind version is similar to the original method; each iteration alternately uses several iterations of the nonblind algorithm to estimate a new PSF and then a new image. It is generally more effective for images having many pixels and for images with fewer sharp edges, since convolution tends to smooth edge boundaries [9].

The algorithm can be used without FFTs, but it is computationally much slower and may produce *ringing* (high frequency oscillations) in the recovered image if the image does not have finite support. This ringing arises because the method has a probabilistic basis, and any implementation must conserve energy. Thus, a nonperiodic (for example, zero boundary condition) implementation is useful only for images having support strictly inside the image boundaries. Convolutions involving images that violate this assumption do not conserve energy since data outside of the original image boundary are lost; this lost energy tends to be recovered as ringing. Conservation of energy, image support, and ringing are discussed in more detail in [21]. To reduce the amount of ringing, we experimented with techniques such as *tapering*, implemented in the MATLAB routine `edgetaper.m`, which seek to transform a nonperiodic image to a periodic one by reblurring the edges of the image with a suitable PSF. The reader is referred to [25] for details.

The stopping criterion for MATLAB's blind LR function `deconvblind.m` is based solely on the input number of iterations. The user may specify this total number of

TABLE 4.1

RSTLN errors for $p = 1$, 2, and ∞. We list the errors in the image x, the matrix A, and the residual error $err(b)$ for the unregularized STLN and the RSTLN methods for each of the norms. For the $p = 1$ and $p = 2$ norms the RSTLN recovered image error $err(x)$ is much smaller than for STLN. For $p = \infty$ the image error is near optimal and the error using RSTLN is only slightly smaller than for STLN.

Test Case 1	$err(x)$	$err(A)$	$err(b)$
$p = 2$ STLN	1.19	3.97e−2	1.1e−3
$p = 2$ RSTLN	0.39	4.10e−2	1.1e−3
$p = 1$ STLN	0.97	3.99e−2	1.4e−3
$p = 1$ RSTLN	0.44	4.00e−2	1.1e−2
$p = \infty$ STLN	0.50	4.02e−2	5.5e−1
$p = \infty$ RSTLN	0.45	3.98e−2	4.9e−1

iterations or use the default value of 10. Our non-FFT implementation is similar to the nonblind MATLAB routine `deconvlucy.m` but lets the user specify the total number of iterations and, for each, the number of LR inner iterations to update the image and PSF estimates. We estimate the optimal number of iterations by recovering images using a wide variety of choices and then choosing the image resulting in the smallest 2-norm error. For our comparison test cases, where our goal was to show only general trends in the recovered images, we often used a default of 10 iterations, modifying this number as needed.

Carasso's APEX/SECB method [1] can be applied to the class of PSFs a whose FFT, denoted by $\hat{a}(\xi, \eta)$, is of the form

$$\hat{a}(\xi, \eta) = e^{-\alpha(\xi^2+\eta^2)^\beta}, \tag{4.1}$$

where ξ and η are the respective frequency coordinates. If the blurred image $b = a \otimes x$ is obtained by (periodic) convolution, then in the Fourier domain,

$$\begin{aligned} \hat{b}(\xi, \eta) &= \hat{x}(\xi, \eta) \cdot \hat{a}(\xi, \eta) \\ &= \hat{x}(\xi, \eta) \cdot e^{-\alpha(\xi^2+\eta^2)^\beta}. \end{aligned} \tag{4.2}$$

The idea behind the PSF identification method is to fit the function $\alpha|\xi|^{2\beta}$ to the logarithm of the Fourier transform of the blurred image minus an estimate of the true image; see [1] for details. If the image or the PSF fails to meet necessary requirements, then such a fit will not be possible.

4.2. Results.

Test 1. Our first test consists of a cross of size 21×21. The true PSF is a Gaussian blur with variance 2.5, truncated to a support of size 11×11.

The blurred image was obtained by convolving the original image and PSF, assuming that pixel values outside the image are zero (zero boundary conditions). The original and blurred images are shown in Figures 4.1(A) and (B). 6-bit noise was added to the PSF to obtain the initial PSF estimate. This resulted in $pert(A) = 3.99 \times 10^{-2}$. Furthermore, 11-bit noise was added to the blurred image, resulting in $pert(b) = 1.10 \times 10^{-3}$.

The errors resulting from the STLN and RSTLN methods for the different p-norms are shown in Table 4.1. The corresponding images are shown in Figures 4.1(C)–(H). From the error table we see that the use of RSTLN generally increases the error $err(A)$ in the blurring matrix and the residual error $err(b)$. For the 1- and 2-norms, however,

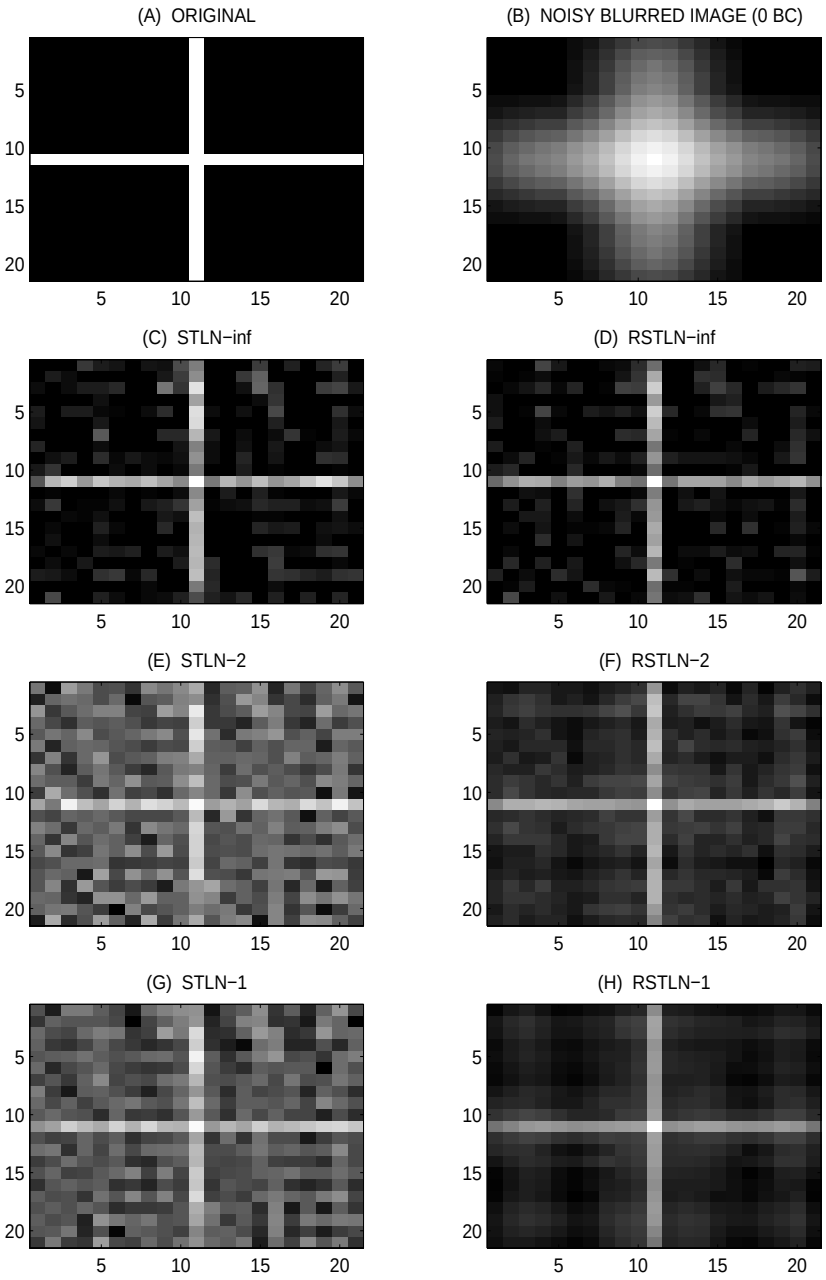


FIG. 4.1. *RSTLN cross (noise, Gaussian blur). Test* 1*, results of STLN and RSTLN methods using* $p = 1$, 2, *and* ∞ *norms. Random noise is present in the blurred image. The blur estimate is the true blur plus the addition of* 6*-bit noise so that* $pert(A) = 3.99 \times 10^{-2}$. 11*-bit noise was added to the blurred image so that* $pert(b) = 1.10 \times 10^{-3}$. (A) *Original image,* 21×21. (B) *Noisy blurred image (zero BC).* (C) *STLN* (∞*-norm) solution with* $tol = 10^{-2}$. *Solution is near optimal:* 13 *iterations.* (D) *RSTLN* (∞*-norm) recovered image with* $tol = 10^{-2}$*, regularization parameter* $\lambda = 0.001$, 12 *iterations.* (E) *STLN* (2*-norm) solution with* $tol = 10^{-3}$, 22 *iterations.* (F) *RSTLN* (2*-norm) recovered image with* $tol = 10^{-3}$, $\lambda = 0.05$, 27 *iterations.* (G) *STLN* (1*-norm) solution with* $tol = 10^{-2}$, 13 *iterations.* (H) *RSTLN* (1*-norm) recovered image with* $tol = 10^{-2}$, $\lambda = 0.5$, 50 *iterations.*

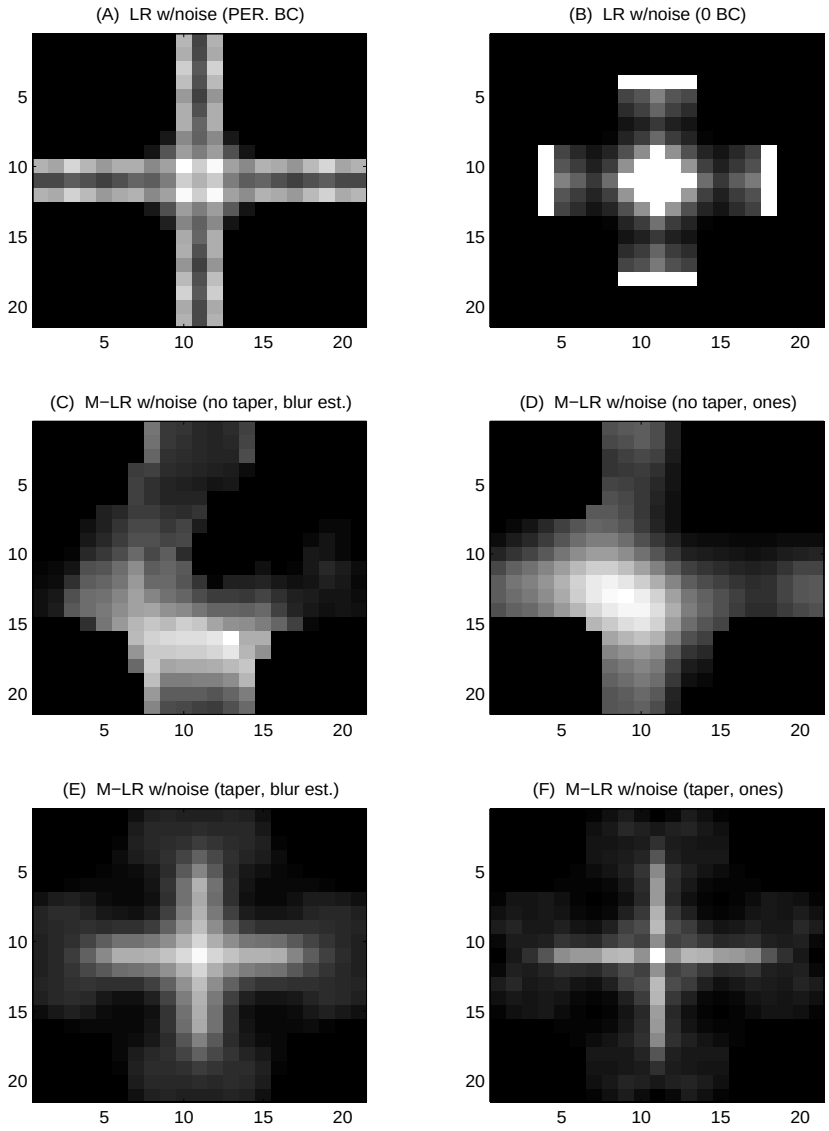


FIG. 4.2. *Test* 1, *LR results.* (A) *Periodic LR implementation using a periodic blurred image,* 20 *LR iterations each with* 10 *iterations.* (B) *Zero boundary condition LR implementation using a zero BC blurred image,* 5 *LR iterations each with* 10 *iterations.* (C) *M-LR result without tapering and using the RSTLN initial PSF estimate,* 10 *iterations.* (D) *M-LR result without tapering and using an* 11×11 *matrix of ones for the initial PSF estimate,* 10 *iterations.* (E) *M-LR result with tapering and using the RSTLN initial PSF estimate,* 50 *iterations.* (F) *M-LR result with tapering and using an* 11×11 *matrix of ones for the initial PSF estimate,* 100 *iterations.*

the error $err(x)$ in the image estimate is considerably lower, so the reconstructed image is improved. For the $p = \infty$ norm, the image obtained from STLN was near optimal, and all RSTLN experiments for nonzero values of the regularization parameter λ resulted in higher image errors.

In Figure 4.2 we present the results of the blind LR method. In (A) we show results obtained by LR in reconstructing images blurred with periodic boundary conditions

TABLE 4.2

RSTLN errors for $p = 2$ for the large cross test case. We list the errors in the image x, the matrix A, and the residual error $err(b)$ for the unregularized STLN and the RSTLN methods for $p = 2$. For the RSTLN ($\lambda = 2.5$) recovered image error $err(x)$ is much smaller than for STLN.

Test Case 2	$err(x)$	$err(A)$	$err(b)$
$p = 2$ STLN	4.2895	4.03e−2	1.03e−2
$p = 2$ RSTLN	0.5885	1.15e+0	9.20e−3

(6-bit noise added) using 20 outer iterations with 10 LR iterations in each. The width of the cross is broadened due to blurring of the edges during the reconstruction.

In Figures 4.2(B)–(F), we present the result of various attempts to reconstruct the image with zero boundary conditions from Figure 4.1. In (B) we show the result obtained by using 5 outer iterations with 10 LR iterations each, computing convolutions using zero padded images. It is clear that the image is distorted, and ringing is observed throughout. The other images are reconstructed using the MATLAB-supplied implementation of blind LR, which we call M-LR. In (C) we show the M-LR result, beginning with the blur estimate as for RSTLN, and stopping after the MATLAB default of 10 iterations. We repeat this experiment in (D) but starting from a flat PSF estimate (a matrix of ones of size 11×11). In both cases only poor reconstructions are obtained. In (E) and (F) we show similar results as in (C) and (D), except that the image is tapered using `edgetaper.m`. The reader is referred to [25] for details. We performed 50 and 100 M-LR iterations, respectively. The reader should note that the algorithm is not able to reconstruct data near the image boundary, although the interior is adequately recovered.

The APEX/SECB method cannot be applied to this image because it is too small to yield enough data points.

Test 2. Our next test consists of a somewhat broader cross image of size 41×41 with a nonzero cross width of 5. The image was blurred with an 11×11 Gaussian. 8-bit noise was added to the blurred images, resulting in $pert(b) = 1.05 \times 10^{-2}$ and 9.8×10^{-3}, respectively. The blur estimate was obtained by adding 6-bit noise to the original blur, resulting in $pert(A) = 3.91 \times 10^{-2}$.

Again, we present results comparing the STLN, RSTLN, LR, and M-LR methods, as well as Carasso's APEX/SECB method. In Figures 4.3(A) and (B) we show the original and blurred images. In (C) we show the STLN 2-norm solution (that is, without any regularization), and in (D) the best RSTLN 2-norm solution with regularization (using $\lambda = 0.75$). (The RSTLN $p = 1$ and $p = \infty$ were not computed due to the expense of solving the linear programming problems.) The resulting STLN and RSTLN errors for the 2-norm are shown in Table 4.2.

For APEX/SECB, the original image in Figure 4.3(A) was blurred using periodic boundary conditions as in (4.2) using parameters $\alpha = 0.075$ and $\beta = 1$. This resulted in a blurred image nearly identical to (B). Again, 8-bit noise was added to the blurred image. In subplot (E) we show the results of using APEX/SECB for PSF identification and subsequent deblurring of the periodic noisy blurred image. The APEX PSF identification procedure resulted in parameter estimates of $\alpha_{est} = 0.0749$ and $\beta_{est} = 0.9756$, which are fairly close to the true parameter values. Unfortunately, this method was unsuccessful for images blurred with zero boundary conditions and noise added. In (F), we show the APEX optimization function for different scalar value image estimates. The nonsmooth family of curves corresponds to the optimization function for different scalar estimates for the unknown frequency-domain

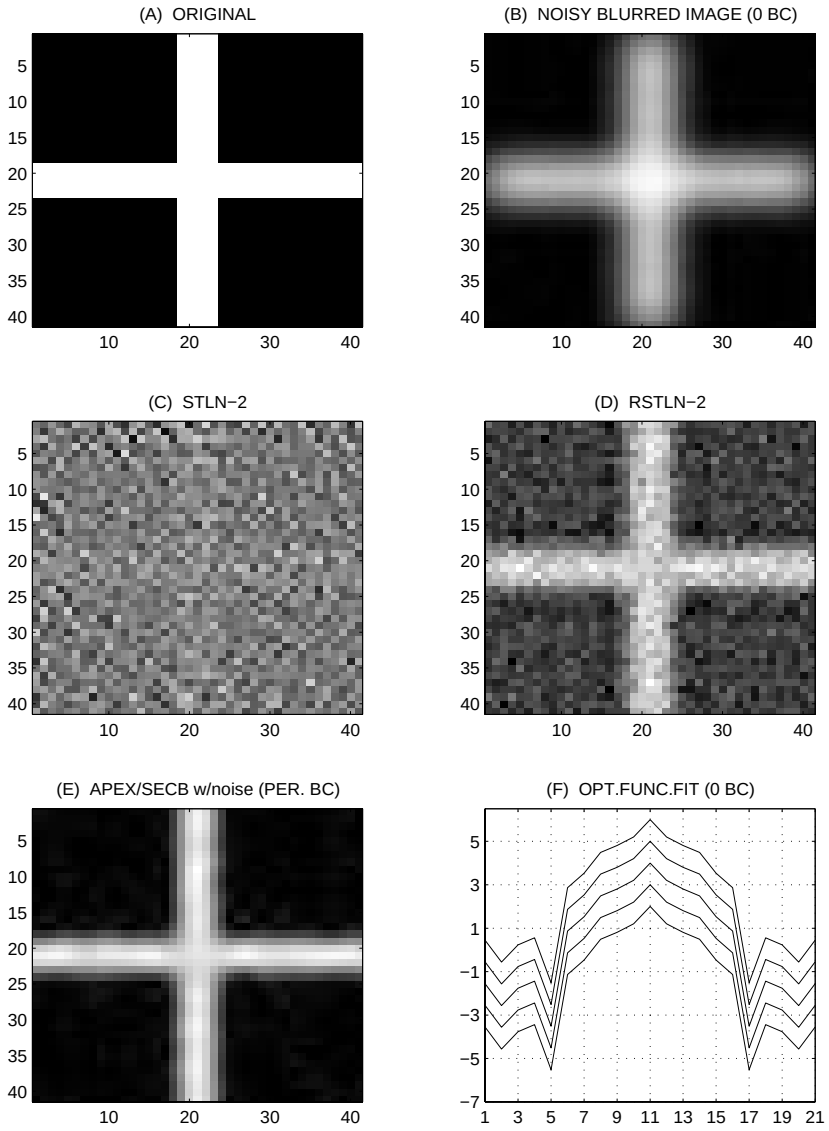


FIG. 4.3. *Test* 2, *RSTLN, and APEX/SECB results. The image was blurred using zero boundary conditions.* 8*-bit noise was added to obtain the image in* (B)*, resulting in* $pert(b) = 1.05 \times 10^{-2}$. *The blur estimate was obtained by adding* 6*-bit noise to the original blur, resulting in* $pert(A) = 3.91 \times 10^{-2}$. (C) *STLN* 2*-norm solution,* $tol = 10^{-3}$*,* 26 *iterations.* (D) *Best RSTLN* 2*-norm solution,* $\lambda = 0.75$*,* $tol = 10^{-3}$*,* 25 *iterations.* (E) *APEX/SECB recovered image using a noisy periodic image. The image was blurred as in* (4.2) *using parameters* $\alpha = 0.075$ *and* $\beta = 1$. *The recovered PSF parameter estimates are* $\alpha_{est} = 0.0749$ *and* $\beta_{est} = 09756$ *using a scalar image component estimate of* $K = 2.2$. (F) *APEX optimization function for a zero BC noisy image. Since the function does not have the proper form* $\alpha|\xi|^{2\beta}$*, no fit can be obtained. In this case no PSF was found.*

image quantity $\log|\hat{x}^*(\xi, 0)|$ if the natural logarithm is applied to the right- and left-hand sides in (4.2) and when a noisy zero boundary condition blurred image is used. The curves do not have the proper form and thus do not permit a curve fit of the form $\alpha|\xi|^{2\beta}$. For this case no proper PSF can be found.

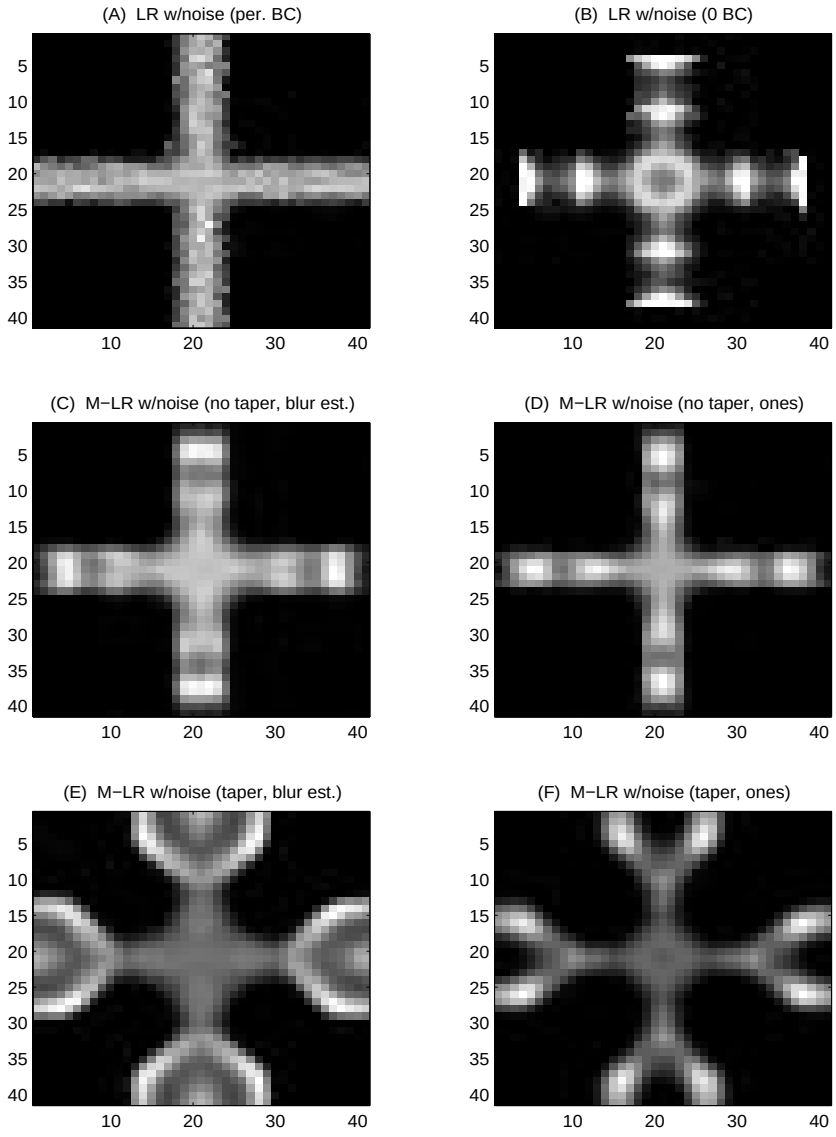


FIG. 4.4. *Test* 2*, LR results.* (A) *Periodic LR implementation using a periodic blurred image,* 50 *LR iterations each with* 10 *iterations.* (B) *Zero boundary LR implementation using a zero BC blurred image,* 50 *LR iterations each with* 10 *iterations.* (C) *M-LR result without tapering and using the RSTLN blur estimate,* 25 *iterations.* (D) *M-LR result without tapering and using an* 11×11 *matrix of ones for the PSF estimate,* 25 *iterations.* (E) *M-LR result with tapering and using the RSTLN blur estimate,* 10 *iterations.* (F) *M-LR result with tapering and using an* 11×11 *matrix of ones for the PSF estimate,* 10 *iterations.*

In Figure 4.4 we present results of the blind LR algorithm. In (A) we see that the algorithm gives a good result for periodic blurs, but the reconstruction for a zero boundary condition exhibits ringing and distortion. These results used 50 outer iterations, each using 10 LR iterations. In (B) we give the result for the zero boundary condition image using the zero boundary implementation. We then apply the M-LR algorithm to a noisy zero boundary blurred image. In (C) and (D) we show results

TABLE 4.3

RSTLN errors for $p = 2$ for the sun test case. We list the errors in the image x, the matrix A, and the residual error $err(b)$ for the unregularized STLN and the RSTLN methods for $p = 2$. For the RSTLN ($\lambda = 75$) recovered image error $err(x)$ is much smaller than for STLN.

Test Case 3	$err(x)$	$err(A)$	$err(b)$
$p = 2$ STLN	20.01	2.47e−2	2.19e−2
$p = 2$ RSTLN	0.9265	3.8483e+0	6.71e−1

using no tapering, 25 iterations, and an initial guess of either the RSTLN blur estimate or a matrix of ones of size 11×11. Both results exhibit ringing due to improper boundary conditions. In (E) and (F) we show M-LR results with tapering, using 10 outer iterations and initial blur estimates as in (C) and (D). The reconstructions are not useful.

Test 3. Our final comparison test consists of an image obtained from the NASA Image Exchange (http://nix.nasa.gov). It shows the corona of the sun and a large solar eruption. We truncated the image to size 99×99 and reduced it to grayscale.

Again, the image was blurred with a Gaussian PSF of size 11×11 in two ways: one assuming zero values for pixels outside the image, and the other assuming a periodic image. 6-bit noise was added to the image after blurring using a zero boundary condition. This resulted in $pert(b) = 2.20 \times 10^{-2}$. For the periodic image no noise was added to the blurred image. The blur estimate was obtained by adding 6-bit noise to the original blur ($pert(A) = 2.46 \times 10^{-2}$).

In Figure 4.5(A) we show the original and in (B) the noisy blurred image using zero boundary conditions. In (C) we show the STLN result using the 2-norm. Due to the high noise level in both the blurred image and the blur estimate, no useful result was obtained. In (D) we show the best result using the RSTLN method with a regularization value of $\lambda = 75$. We remark that in this case the algorithm did not converge to a tolerance of 10^{-2}. Instead, we stopped prematurely after 10 iterations. A larger number of iterations which did achieve the desired tolerance produced an image of lesser quality (see section 4.1.1 on noise amplification).

In Table 4.3 we computed the resulting errors for the STLN and RSTLN methods. Although $err(A)$ and $err(b)$ are increased for RSTLN with respect to STLN, clearly the image error is drastically reduced using the RSTLN method.

For the APEX/SECB method the image was blurred with a Gaussian blur using periodic boundary conditions and parameters $\alpha = 0.01$ and $\beta = 1$ as in (4.2). This resulted in a blurred image very similar to the one in Figure 4.5(B). 6-bit noise was added to the blurred image. Using the APEX PSF identification method, a curve fit to the optimization function was done, resulting in parameter estimates of $\alpha_{est} = 0.0108$ and $\beta_{est} = 1.028$. These are fairly close to the true PSF parameters. In (E) we show the APEX/SECB recovered image using the noisy blurred image with periodic boundary conditions. In (F) we show the function to be fit using the noisy image with zero boundary conditions. We plot the function using different scalar estimates for the original image component in (4.2). None of the functions have the proper form and a suitable curve fit of the form $\alpha|\xi|^{2\beta}$ is not possible. For this case no useful PSF was found.

In Figure 4.6 we show the results from the various LR experiments. In subplot (A) we have the LR result using a periodic image using our own periodic LR implementation. We performed 10 iterations, each with 10 LR iterations. In (B) we show the result using the zero boundary implementation and a zero boundary blurred

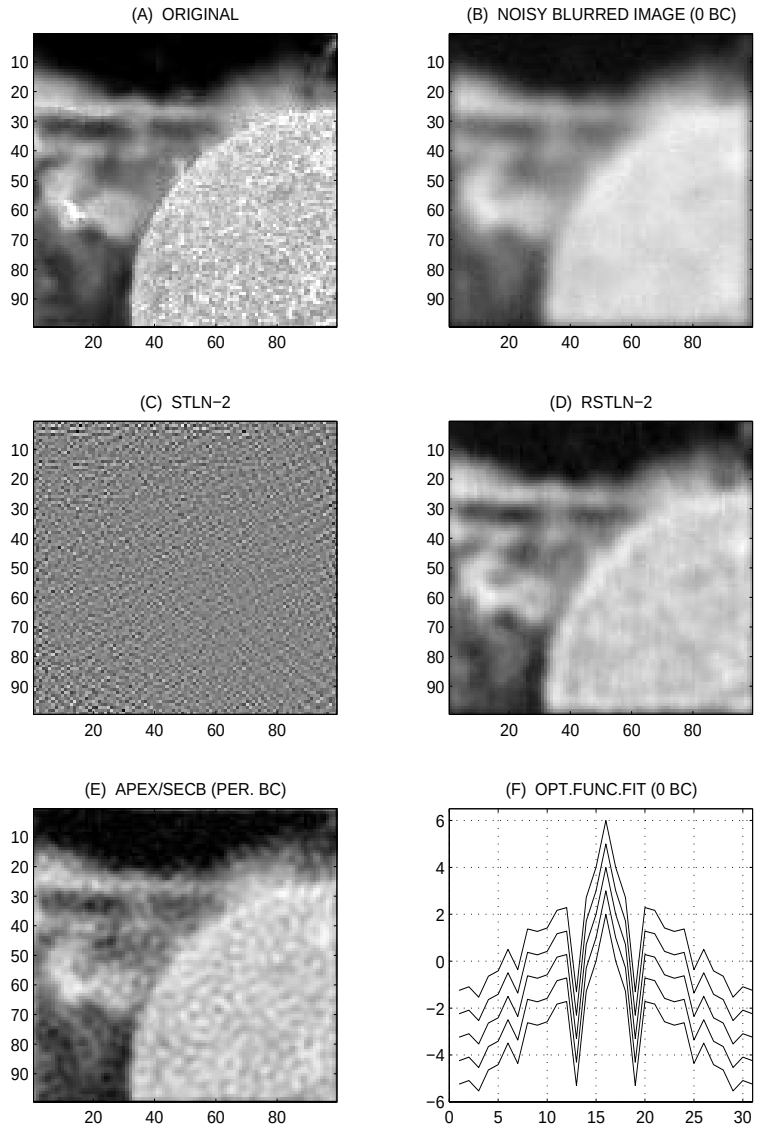


FIG. 4.5. *Test* 3, *RSTLN and APEX/SECB results.* (A) *Original image,* 99×99. (B) *Noisy blurred image (zero BC).* (C) *STLN (2-norm) solution with tol* $= 10^{-2}$, 2 *iterations.* (D) *RSTLN (2-norm) recovered image with initial tol* $= 10^{-2}$ *and regularization* $\lambda = 75$. *The experiment was stopped prematurely after* 10 *iterations. While a larger number of iterations did achieve the desired tolerance, the results were distorted by ringing.* (E) *APEX/SECB recovered image. Image is blurred assuming a periodic image as in* (4.2) *with parameters* $\alpha = 0.01$ *and* $\beta = 1$. (F) *Plot of optimization function if the image is blurred using zero BC. The different plots represent the optimization function for different scalar estimates for the unknown quantity* $\log|\hat{x}^*(\xi, 0)|$, *where* $\hat{x}^*(\xi, \eta)$ *denotes the normalized FFT of the original image* x. *Since none of the curves possess the proper shape, no useful PSF can be found.*

image. We performed 15 outer iterations, each with 10 iterations to estimate the new PSF and image. Severe ringing is present. In (C) and (D) we show the nontapered M-LR results using the RSTLN blur estimate, an 11×11 matrix of ones for the blur

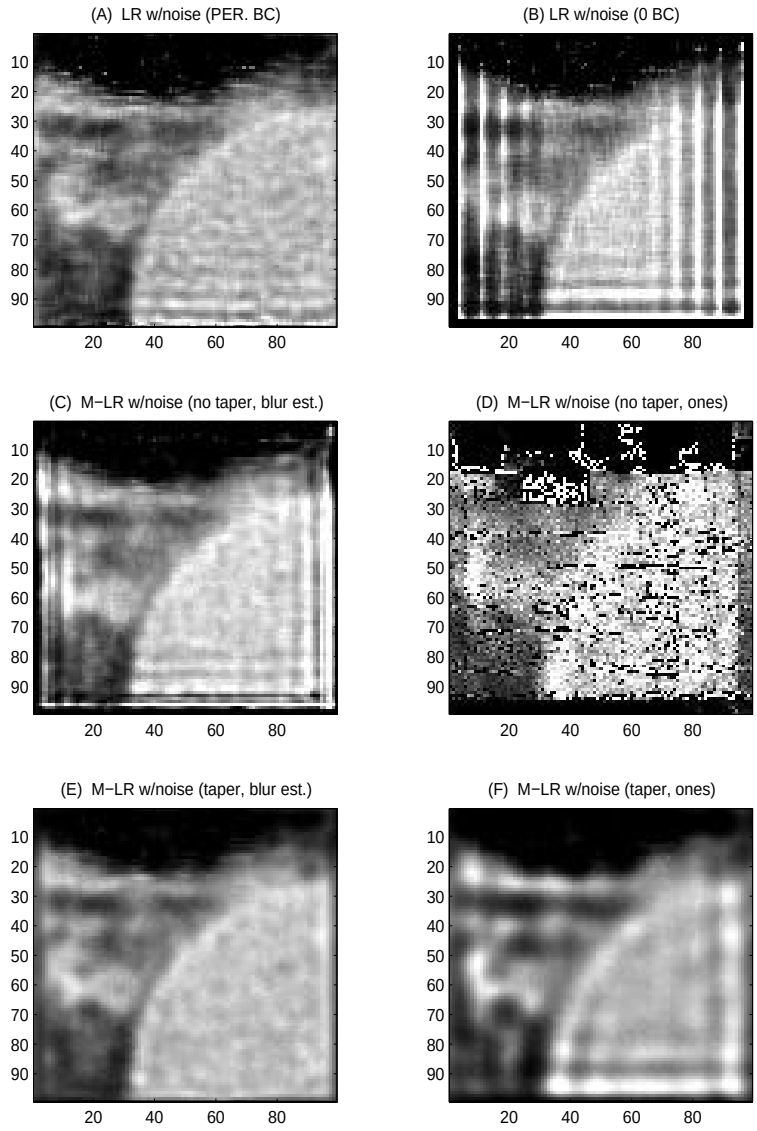


FIG. 4.6. *Test* 3*, LR results.* (A) *Periodic LR implementation using a periodic blurred image,* 10 *LR iterations each with* 10 *iterations.* (B) *Zero boundary LR implementation using a zero BC blurred image,* 15 *LR iterations each with* 10 *iterations.* (C) *M-LR result without tapering and using the RSTLN blur estimate,* 25 *iterations.* (D) *M-LR result without tapering and using an* 11×11 *matrix of ones for the PSF estimate,* 10 *iterations.* (E) *M-LR result with tapering and using the RSTLN blur estimate,* 25 *iterations.* (F) *M-LR result with tapering and using an* 11×11 *matrix of ones for the PSF estimate,* 10 *iterations.*

estimate, and a zero boundary blurred image. 25 outer iterations were performed, with 10 iterations each. For the result in (C), ringing is observed near the image boundary, whereas in (D) the image is severely distorted. Finally, in (E) and (F) we obtained results using M-LR and a tapered noisy blurred image using the two different initial blur estimate types. For the result in (E), 25 iterations were performed which produced reasonable results. The result in (F) was obtained after 10 iterations with less favorable results.

TABLE 4.4
Summary of methods.

Algorithm	Requirements	Comments
Blind LR	• FFT version requires periodic boundary conditions or finite support	• more effective with larger images • ringing if image fails to have finite support
Carasso's method	• periodic boundary conditions or finite support • image must have specific properties [1] • PSF must have specific properties [1]	• more effective with larger images
STLN	• substantial computation for 1- and ∞-norm methods	• sensitive to noise
RSTLN	• substantial computation for 1- and ∞-norm methods	• robust to noise

4.3. Effectiveness of methods. As the experiments indicate, some of the methods presented prove useful only if specific requirements are satisfied. This section summarizes the effectiveness of each of the methods.

The blind LR method using FFTs is useful only if the original image either was blurred using periodic boundary conditions or has finite support. If it does not satisfy either of these conditions, the recovered image often suffers from ringing. It is also observed that the method is sometimes more useful for larger images or if preprocessing techniques such as tapering or flat PSF initial estimates are used (see [25] for details).

Like the blind LR method, Carasso's APEX/SECB method requires periodic boundary conditions or finite support [2]. Furthermore, it can be applied only to the class of PSFs satisfying (4.1) and requires images to belong to a specific class as defined in [1].

In contrast, neither STLN nor RSTLN imposes any restrictions on the image or PSF and both are effective on small images. While STLN is useful for some total least norm problems, the blind deconvolution problem is generally ill-posed, so that small perturbations in the data can cause large changes in the solution. Thus, the RSTLN method proves to be more useful for most blind image deblurring applications where regularization is usually necessary.

If the noise is Gaussian, then least squares theory provides ample justification for choosing the 2-norm in RSTLN rather than the 1-norm or ∞-norm. However, in order to take advantage of this theory, the standard deviations of the two error distributions must be known so that the error terms can be balanced. When this data is unavailable, or when the noise distributions are not Gaussian, then the 1-norm and ∞-norm have no theoretical disadvantages. Our experiments show that the 1-norm in particular provides high-quality reconstructions and is not sensitive to outliers in the data.

A summary of the requirements and effectiveness of each method is given in Table 4.4.

5. Conclusions. We have presented the RSTLN algorithm for blind deconvolution. Like the STLN method, RSTLN preserves any affine structure in the matrix, and the user has the choice of minimizing the error for the 2-norm or for other norms such as the 1- and ∞-norms. The use of norms other than the 2-norm leads to good image recovery, although the cost is substantially higher.

In contrast to other methods, such as that of Carasso's APEX/SECB, the RSTLN method does not depend on having a periodic image. Ringing in the reconstructed images is less of a problem. Therefore, we can apply the RSTLN method for arbitrary boundary conditions, for example, zero (Dirichlet), Neumann (data outside the image boundary is a reflection of the corresponding data inside), or periodic.

Acknowledgments. The authors wish to thank Cleve Moler and Bruce Golden for their help with linear programming and MATLAB, Jon McCoy for spirited discussions on the LR method, and the referees for their insightful comments.

REFERENCES

[1] A. S. Carasso, *Direct blind deconvolution*, SIAM J. Appl. Math., 61 (2001), pp. 1980–2007.
[2] P. J. Davis, *Circulant Matrices*, Wiley, New York, 1979.
[3] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., Vol. 2, Wiley and Sons, New York, 1980.
[4] G. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.
[5] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.
[6] R. C. Gonzalez and P. Wintz, *Digital Image Processing*, Addison-Wesley, Reading, MA, 1977.
[7] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Rev., 34 (1992), pp. 561–580.
[8] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.
[9] P. C. Hansen, *Numerical Aspects of Deconvolution*, Lecture Notes, Department of Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark, 2000.
[10] M. E. Kilmer and D. P. O'Leary, *Choosing regularization parameters in iterative methods for ill-posed problems*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1204–1221.
[11] R. L. Lagendijk and J. Biemond, *Iterative Identification and Restoration of Images*, Kluwer Academic Publishers, Norwell, MA, 1991.
[12] P. Lemmerling, N. Mastronardi, and S. Van Huffel, *Fast algorithm for solving Hankel/Toeplitz structured total least squares problem*, Numer. Algorithms, 21 (2000), pp. 371–392.
[13] L. B. Lucy, *An iterative technique for the rectification of observed distributions*, Astronomical Journal, 79 (1974), pp. 745–754.
[14] N. Mastronardi, P. Lemmerling, and S. Van Huffel, *Fast structured total least squares algorithm for solving the basic deconvolution problem*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 533-553.
[15] V. Mesarović, N. Galatsanos, and A. Katsaggelos, *Regularized constrained total least squares image restoration*, IEEE Trans. Image Process., 4 (1995), pp. 1096–1108.
[16] B. De Moor, *Total least squares for affinely structured matrices and the noisy realization problem*, IEEE Trans. Signal Process., 42 (1994), pp. 3004–3113.
[17] M. K. Ng, R. J. Plemmons, and F. Pimentel, *A new approach to constrained total least squares image restoration*, Linear Algebra Appl., 316 (2000), pp. 237–258.
[18] D. P. O'Leary, *Near-optimal parameters for Tikhonov and other regularization methods*, SIAM J. Sci. Comput., 23 (2001), pp. 1161–1171.
[19] D. P. O'Leary and J. A. Simmons, *A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems*, SIAM J. Sci. Stat. Comput., 2 (1981), pp. 474–489.
[20] D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. Assoc. Comput. Mach., 9 (1962), pp. 84–97.
[21] A. Pruessner, *Blind Deconvolution Using a Regularized Structured Total Least Norm Algorithm*, Master's Thesis, University of Maryland, College Park, MD, 2001.
[22] W. H. Richardson, *Bayesian-based iterative method of image restoration*, J. Opt. Soc. Amer. A, 62 (1972), pp. 55–59.
[23] J. B. Rosen, H. Park, and J. Glick, *Total least norm formulation and solution for structured problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 110–126.
[24] J. B. Rosen, H. Park, and J. Glick, *Structured total least norm for nonlinear problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 14–30.

[25] J. Skilling and S. F. Gull, *Algorithms and applications*, in Proc. 1st Workshop on Maximum Entropy and Bayesian Methods in Inverse Problems, C. R. Smith and W. T. Grandy, Jr., eds., D. Reidel Publishing Company, Boston, 1985, pp. 83–132.

[26] A. N. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 4 (1963), pp. 1035–1038.

[27] Y. Zhang, *Solving Large-Scale Linear Programs by Interior-Point Methods under the MATLAB Environment*, Technical Report, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1997.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1038–1059

THE SOLUTION OF PARAMETRIZED SYMMETRIC LINEAR SYSTEMS*

KARL MEERBERGEN†

Abstract. We compare the Lanczos and MINRES methods for the solution of symmetric linear systems with a linear parameter arising from structural engineering. We show a connection with the Padé-via-Lanczos method. We propose an error estimation for the solution in finite precision arithmetic. The use of reorthogonalization is discussed and practical algorithms are given. The theory is illustrated with numerical examples.

Key words. Lanczos method, MINRES, linear systems

AMS subject classifications. 65F15, 65F50

PII. S0895479800380386

1. Introduction.

This paper analyzes a technique for the efficient solution of

$$L(\alpha)x(\alpha) = f \quad \text{with} \quad L(\alpha) = K - \alpha M \tag{1.1}$$

for different values of α. This problem arises in many engineering applications where (1.1) is the equation that results from the Fourier transform of a second order differential equation. The parameter α is the square of the frequency ω, K is an $n \times n$ (real symmetric) stiffness matrix, and M is an $n \times n$ (real symmetric) positive definite mass matrix. K and M are typically large and sparse. Usually, ω is selected from a given frequency range $[\omega_{\min}, \omega_{\max}]$ in $\mathbf{R}^+$. Hence, we are interested in the computation of $x(\alpha)$ in the interval $[\alpha_{\min}, \alpha_{\max}]$ with $\alpha_{\min} = \omega_{\min}^2$ and $\alpha_{\max} = \omega_{\max}^2$.

The traditional approach to this problem is to discretize $[\alpha_{\min}, \alpha_{\max}]$ into $\{\alpha_1, \ldots, \alpha_m\}$ and factorize $L(\alpha)$ and solve (1.1) by backtransformation for $\alpha = \alpha_1, \ldots, \alpha_m$. This can be quite expensive when m is large. This paper uses more efficient techniques for solving (1.1) that are based on model reduction methods and iterative solvers for linear systems.

An approach that received recent attention in the engineering community is the Padé approximation. (See Kuzuoglu and Mittra [23] and Malhotra and Pinsky [26].) In this approach, $x(\alpha)$ is developed in a vector-Padé series around a central value σ, and the coefficients of the series are evaluated with a recurrence relation. This technique is very general and also can be applied when the right-hand side f depends on α, but the Padé expansion may suffer from ill-conditioning of the basis functions; see, e.g., Feldman and Freund [11] and Skoogh [37].

When the quantity of interest has the form $c^*x(\alpha)$ with $c \in \mathbf{R}^n$, model reduction techniques based on Krylov methods can be employed. Padé-via-Lanczos (PVL) is discussed by Feldman and Freund [11], Bay and Ye [4], Bai and Freund [3], and Grimme, Sorensen, and Van Dooren [19], and multipoint Padé (or rational Lanczos) is discussed by Gallivan, Grimme, and Van Dooren [13, 14] and Grimme [20]. Jaimoukha and Kasenally [22] use a two-sided Arnoldi method. Skoogh [37, 36] analyzes rational Krylov methods. In structural analysis and acoustics the quantity of interest may

*Received by the editors November 3, 2000; accepted for publication (in revised form) by Z. Strakos June 25, 2002; published electronically February 25, 2003.
http://www.siam.org/journals/simax/24-4/38038.html
†Free Field Technologies, 16 place de l'Université, 1348 Louvain-la-Neuve, Belgium (karl.meerbergen@fft.be).

be comprised of another combination of the components of $x(\alpha)$. Very often, tens or even hundreds of components of the solution vector are of interest and, in this case, the model reduction methods are not necessarily the methods of choice. However, we shall see that the Lanczos method studied in this paper is intimately connected to the PVL method [11].

A possible approach consists of the iterative solution of

$$(M^{-1}K - \alpha I)x(\alpha) = M^{-1}b\ . \tag{1.2}$$

Because Krylov subspaces are shift-invariant, the Krylov space for $M^{-1}K - \alpha I$ is independent of α. It is also this property that makes PVL, mentioned above, very efficient. Iterative methods for shifted linear systems were proposed by Datta and Saad [8] and Frommer and Glässner [12]. Krylov methods applied to $M^{-1}K - \alpha I$ generally converge slowly for our applications because the spectrum of $M^{-1}K - \alpha I$ is spread over a large interval on the real axis. This technique may be useful if the multiplication of M^{-1} with a vector is very cheap, e.g., when M is a diagonal matrix, which is the case when a lumped finite element formulation is used. Shifting K and inverting the role of K and M in (1.2) leads to

$$(K - \sigma M)^{-1}(K - \alpha M)x(\alpha) = (K - \sigma M)^{-1}b, \tag{1.3}$$

for which the shift-invariance property of Krylov subspaces also holds, since

$$(K - \sigma M)^{-1}(K - \alpha M) = I + (\sigma - \alpha)(K - \sigma M)^{-1}M.$$

Although the application of $(K - \sigma M)^{-1}$ is, in general, not cheap, the spectrum of $(K-\sigma M)^{-1}(K-\alpha M)$ is more favorable for fast convergence in Krylov methods than $M^{-1}K-\alpha I$, provided σ is well chosen. In the context of iterative methods, $K-\sigma M$ can be regarded as a preconditioner for (1.1). The use of $K-\sigma M$ for preconditioning is widely accepted for computing eigenvalues of large linear eigenvalue problems [10, 2]. Model reduction techniques and iterative methods for parametrized linear systems also use preconditioning by $K - \sigma M$ to improve the quality of the reduced model [14, 35, 34]. The idea of using Krylov subspaces for computing $x(\alpha)$ is known in the engineering community as the Ritz vector technique, discussed by Wilson, Yuan, and Dickens [39], Ibrahimbegovic et al. [21], and Coyette [6].

This work is a comparison of the Lanczos and MINRES methods with M-inner-product, the derivation of error estimates for computations with finite precision arithmetic or the iterative computation of the product $(K - \sigma M)^{-1}$ with a vector, and the use of Krylov methods for parametrized linear systems on problems arising from low frequency behavior of structures. We also discuss implementation aspects. We do not develop advanced strategies for choosing σ. This is still an open problem that deserves some attention.

The paper is organized as follows. In section 2, we present the Lanczos and MINRES methods with M-innerproduct. We discuss the quality of the preconditioner in section 3. The MINRES and Lanczos methods are compared in section 4. We derive error estimates of the approximation computed in sections 5 and 6. We also give practical algorithms and comment on the loss of orthogonality of the Lanczos vectors in section 7. Numerical examples are given in section 8. In section 9, we formulate our main conclusions.

1.1. Notation. The M-innerproduct of two vectors x and y is x^*My, and the induced norm is denoted by $\|x\|_M$. We have that

$$\|x\|_2^2/\|M^{-1}\| \leq \|x\|_M^2 \leq \|x\|_2^2\|M\|.$$

We define the matrix M-norm $\|A\|_M = \max_{\|x\|_M=1} \|Ax\|_M$. Throughout the paper we will frequently use the shift-and-invert transformation

$$A = (K - \sigma M)^{-1} M \ , \tag{1.4}$$

the (matrix) Cayley transformation

$$A(\alpha) \equiv (K - \sigma M)^{-1}(K - \alpha M) \ , \tag{1.5}$$

the preconditioned right-hand side

$$b = (K - \sigma M)^{-1} f, \tag{1.6}$$

and its M-norm

$$\beta = \|b\|_M \ . \tag{1.7}$$

The Cayley transform and its inverse are defined by

$$\theta = c(\lambda) := \frac{\lambda - \alpha}{\lambda - \sigma} \quad \text{and} \quad \lambda = c^{-1}(\theta) := \frac{\alpha - \sigma\theta}{1 - \theta} \ . \tag{1.8}$$

Underscored uppercase letters are used to denote rectangular matrices. For example, $\underline{I}$ denotes the identity matrix with an additional lower row of zeros.

2. Iterative methods. The idea is to use an iterative method for solving (1.1). In order to improve the speed of convergence, we solve the preconditioned linear system

$$A(\alpha)x(\alpha) = b \ , \tag{2.1}$$

where σ is fixed for a large number of α's. Since $A(\alpha)$ is generally no longer symmetric, we cannot apply methods such as Lanczos [25] or MINRES [30, 5, 9] for symmetric indefinite linear systems in their standard form. For the inversion of $(K - \sigma M)$ we assume it is practical to use a direct linear solver. Since σ is fixed for a large number of α's, only a few large-scale sparse factorizations are required, as is also the case for solving the corresponding eigenvalue problem $Ku = \lambda Mu$ by the spectral transformation Lanczos method, which we will discuss further.

2.1. The Lanczos process. First, suppose we want to solve the system

$$Ax = b \ . \tag{2.2}$$

Although A is a nonsymmetric matrix, we can use the (symmetric) Lanczos method, as we now explain.

Since A is self-adjoint with respect to the M-innerproduct, Nour-Omid et al. [28] and Grimes, Lewis, and Simon [18] suggest the M-innerproduct so that the Lanczos method can be used for solving the eigenvalue problem $L(\lambda)u = 0$. Ericsson and Ruhe [10] call the matrix A the spectral transformation, and Saad [32] uses the term shift-and-invert transformation. The process proposed by Lanczos [24] builds a set of $k+1$ basis vectors $V_{k+1} = [v_1, \ldots, v_{k+1}] \in \mathbf{R}^{n\times(k+1)}$, with $V_{k+1}^* M V_{k+1} = I$, of the Krylov space

$$\mathcal{K}_{k+1}(A, b) = \text{span}\{b, Ab, A^2 b, \ldots, A^k b\},$$

and a $(k+1) \times k$ tridiagonal matrix $\underline{T}_k$ satisfying the recurrence relation

$$AV_k - V_{k+1}\underline{T}_k = 0 \ . \tag{2.3}$$

2.2. Lanczos and MINRES. In the Lanczos method [30, 5], we compute the approximate solution of (2.2) as

$$\tilde{x} = V_k \tilde{z} \quad \text{with} \quad \tilde{z} = \beta T_k^{-1} e_1,$$

where T_k is the leading $k \times k$ submatrix of $\underline{T}_k$ and β is as defined in (1.7). The residual is

$$\begin{aligned} r = b - A\tilde{x} &= V_{k+1}(\beta e_1 - \underline{T}_k \tilde{z}) \\ &= -v_{k+1}\beta_k e_k^* \tilde{z} \end{aligned}$$

and $\rho = \|r\|_M = \beta_k |e_k^* \tilde{z}|$, where $\beta_k = e_{k+1}^* \underline{T}_k e_k$. The application of T_k^{-1} usually is performed using a direct tridiagonal linear solver without pivoting based on LU or QR factorization.

With MINRES the approximate solution of (2.2) is found as $\hat{x} = V_k \hat{z}$, where $\hat{z}$ minimizes $\|\beta e_1 - \underline{T}_k z\|_2$. Then $\hat{x}$ minimizes $\|b - Ax\|_M$ among all $x \in \mathcal{K}_k(A, b)$. The minimization problem can be solved by first forming the QR factorization $\underline{Q}R = \underline{T}_k$ with R a $k \times k$ upper triangular matrix and $Q = [\underline{Q} \;\; q]$ $(k+1) \times (k+1)$ orthogonal. Then $\hat{z} = \beta R^{-1} \underline{Q}^* e_1$ and the residual for $\hat{x} = V_k \hat{z}$ is

$$\begin{aligned} r &= b - A\hat{x} \\ &= V_{k+1}(\beta e_1 - \underline{T}_k \hat{z}) \\ &= \beta V_{k+1}(\underline{I} - \underline{Q}\underline{Q}^*) e_1 \\ &= \beta V_{k+1} q q^* e_1 \ . \end{aligned}$$

The residual norm for $\hat{x}$ is $\rho := \|r\|_M = \beta |q^* e_1|$.

2.3. Parametrized linear systems. We now return to the solution of (2.1). Since

$$A(\alpha) = (K - \sigma M)^{-1}(K - \alpha M) = I + (\sigma - \alpha)A,$$

the Krylov space is the same for all $\alpha \neq \sigma$ as for A, so it is sufficient to compute V_k only once. In addition, if the Krylov space is computed for A, i.e., if V_{k+1} and $\underline{T}_k$ satisfy (2.3), then

$$A(\alpha)V_k - V_{k+1}\underline{H}_k(\alpha) = 0 \quad \text{with} \quad \underline{H}_k(\alpha) = \underline{I} + (\sigma - \alpha)\underline{T}_k \ . \tag{2.4}$$

We also define

$$\eta_k = (\sigma - \alpha)\beta_k \ . \tag{2.5}$$

To summarize, the solution of (1.1) is computed as

$$\begin{aligned} \text{Lanczos}: \quad \tilde{z}(\alpha) &= \beta H_k^{-1}(\alpha) e_1, \\ \tilde{x}(\alpha) &= V_k \tilde{z}(\alpha), \\ r(\alpha) &= -v_{k+1}\eta_k e_k^* \tilde{z}(\alpha), \end{aligned} \tag{2.6}$$

$$\begin{aligned} \text{MINRES}: \quad \hat{x}(\alpha) &= \beta V_k R^{-1} \underline{Q}^* e_1, \\ r(\alpha) &= \beta V_{k+1} q q^* e_1, \end{aligned} \tag{2.7}$$

where $\underline{Q}$, R, and q come from the QR factorization $\underline{H}_k(\alpha) = \underline{Q}R$ and $Q = [\underline{Q} \;\; q]$ unitary.

The following algorithm gives a brief overview of the computational steps to be performed to compute $\tilde{x}(\alpha)$ and $\hat{x}(\alpha)$. Practical algorithms are discussed in section 7.

ALGORITHM 2.1 (MINRES/Lanczos(α)).

1. Choose σ and factorize $K - \sigma M$.
   Solve $(K - \sigma M)b = f$ for b.
2. Set initial vector $v_1 = b/\beta$ and $v_{-1} = 0$ and $\beta_0 = 0$.
3. For $j = 1, \ldots, k$ do
   3.1. Solve $(K - \sigma M)w_j = Mv_j$ for w_j.
   3.2. Compute $w_j' = w_j - \beta_{j-1}v_{j-1}$.
   3.3. Compute $\alpha_j = v_j^* M w_j'$.
   3.4. Compute $w_j'' = w_j' - \alpha_j v_j$.
   3.5. Compute $\beta_j = \|w_j''\|_M$.
   3.6. Normalize: $v_{j+1} = w_j''/\beta_j$.
4. For $j = 1, \ldots, m$ do
   4.1. Form $\underline{H}_k(\alpha_j) = \underline{I} + (\sigma - \alpha_j)\underline{T}_k$.
   4.2. Lanczos : solve $H_k(\alpha_j)\tilde{z}(\alpha_j) = \beta e_1$ and compute $\tilde{x}(\alpha_j) = V_k\tilde{z}(\alpha_j)$.
   MINRES : solve $\min_{\hat{z}} \|\underline{H}_k(\alpha_j)\hat{z} - \beta e_1\|_2$ for $\hat{z}$ and compute $\hat{x}(\alpha_j) = V_k\hat{z}(\alpha_j)$.

Steps 3.2–3.6 form an orthogonalization step that makes v_{j+1} M-orthogonal to v_{j-1} and v_j. Because of properties of the Lanczos method, orthogonality is guaranteed to $v_1, \ldots, v_{j-2}$ (in exact arithmetic). The coefficients α_j form the main diagonal elements, and β_j form the off-diagonal elements of the tridiagonal matrix $\underline{T}_k$.

3. Analysis of the preconditioner. An important remaining question is, How good is $K - \sigma M$ as a preconditioner? The preconditioner is good when the eigenvalues of $A(\alpha)$ are clustered away from zero. In the following, we will analyze the spectrum of this matrix.

The matrix $A(\alpha)$ is a generalized Cayley transform, whose spectral properties are analyzed in detail by Garratt, Moore, and Spence [15]. If λ is an eigenvalue of $M^{-1}K$, then $\theta = c(\lambda)$ is an eigenvalue of $A(\alpha)$.

A typical situation in structures and acoustics is shown in Figure 3.1. Usually, the spectrum of $M^{-1}K$ is spread out over the real axis, which makes the iterative solution of (1.2) hard. Since $\lim_{\lambda\to\infty}(\lambda - \alpha)/(\lambda - \sigma) = 1$, most λ's are mapped to θ's close to 1. Clustered eigenvalues very much favor that Krylov methods converge. The eigenvalues $\lambda > \alpha$ are mapped between zero and one. The eigenvalues $\lambda < \sigma$ are mapped between 1 and $+\infty$. The eigenvalues in between σ and α are mapped between $-\infty$ and 0. If $M^{-1}K$ has no eigenvalues between σ and α, $A(\alpha)$ has positive eigenvalues and the convergence of the Lanczos method is similar to that of conjugate gradients. The Cayley transform is ill-conditioned (and may thus perform badly in an iterative method) when it has eigenvalues near zero ($\lambda \approx \alpha$) and some far away from zero ($\lambda \approx \sigma$).

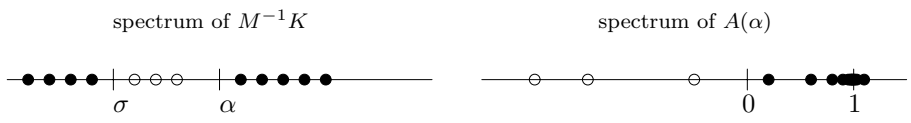


FIG. 3.1. *Mapping properties of the Cayley transform.*

The spectrum of $A(\alpha)$ in Figure 3.1 has a large cluster of eigenvalues near 1 and a few isolated eigenvalues that converge one by one in the Krylov subspace. As in many applications, this is a situation where classical convergence rates are too pessimistic; see, e.g., [38, 17].

4. Characterization.

4.1. Relation to the eigenvalue problem. If $Ku = \lambda Mu$, then $A(\alpha)u = \theta u$ with $\theta = c(\lambda)$. Let $(\tilde{\theta}_j, \tilde{y}_j)$ for $j = 1, \ldots, k$ be the eigenpairs of H_k; if k is large enough, some $\tilde{\lambda}_j = c^{-1}(\tilde{\theta}_j)$ for $j = 1, \ldots, k$ are good approximations to the eigenvalues of $M^{-1}K$ near σ. In this paper, we call $\tilde{\lambda}_j$ a Ritz value of $M^{-1}K$ to distinguish it from the (exact) eigenvalue λ_j.

When the Lanczos method is used, we can write from (2.6) that

$$\tilde{x}(\alpha) = \beta V_k \sum_{j=1}^{k} \frac{\tilde{\lambda}_j - \sigma}{\tilde{\lambda}_j - \alpha} (\tilde{y}_j^* e_1) \tilde{y}_j \ . \tag{4.1}$$

The following lemma illustrates the link between the Ritz values and the residual norm of the Lanczos solution.

LEMMA 4.1. *Let $(\tilde{\theta}_j, \tilde{u}_j)$ with $\tilde{u}_j = V_k \tilde{y}_j$ and $\|\tilde{y}_j\|_2 = 1$ be a Ritz pair of $A(\alpha)$. Let*

$$\rho_j = \|A(\alpha)\tilde{u}_j - \tilde{\theta}_j \tilde{u}_j\|_M$$

be the residual norm of the Ritz pair. Then

$$\begin{aligned} \|r(\alpha)\|_M &:= \|b - A(\alpha)\tilde{x}(\alpha)\|_M \\ &\leq \sum_{j=1}^{k} |\tilde{u}_j^* M b| \cdot |\tilde{\theta}_j^{-1}| \cdot \rho_j \ . \end{aligned} \tag{4.2}$$

Proof. From (4.1) and (2.6), we derive

$$r(\alpha) = -v_{k+1} \beta \eta_k \sum_{j=1}^{k} \tilde{\theta}_j^{-1} (\tilde{y}_j^* e_1)(e_k^* \tilde{y}_j) \ .$$

From (2.4), we have

$$A(\alpha)\tilde{u}_j - \tilde{\theta}_j \tilde{u}_j = v_{k+1} \eta_k e_k^* \tilde{y}_j$$

and we also have that $\beta \tilde{y}_j^* e_1 = \tilde{u}_j^* M b$. This proves the lemma. ∎

This lemma shows the importance of Ritz values and Ritz vectors in the convergence of the iterative method. From (4.2), it follows that if $|\tilde{u}_j^* M b| / \|b\|_M$, $|\tilde{\theta}_j|^{-1}$, or ρ_j for $j = 1, \ldots, k$ are much smaller than 1, then $\|r(\alpha)\|_M \ll \|b\|_M$. In other words, $|\tilde{u}_j^* M b| / \|b\|_M$ is small when $\tilde{u}_j$ has a small component in b. In this case $\tilde{x}(\alpha)$ has a small component in $\tilde{u}_j$. We do not consider this situation any further. Assume that $|\tilde{\theta}_j^{-1}|$ is significantly larger than 1. This happens when $|\alpha - \tilde{\lambda}_j| \ll |\sigma - \tilde{\lambda}_j|$. Then the only possibility of having a small $\|r(\alpha)\|$ is if ρ_j is small. This must be true for all α in the interval of interest. If ρ_j is small, $\tilde{\lambda}_j$ is near an eigenvalue of $M^{-1}K$. This implies that all eigenvalues of $M^{-1}K$ in this interval must be computed fairly accurately. If the Ritz values $\tilde{\lambda}_j$ are good approximations to the eigenvalues λ_j, the peaks in $\|\tilde{x}(\alpha)\|$

correspond well to peaks in $\|x(\alpha)\|$. The error in the Ritz values gives the error in the positions of the peaks. So, in practice, k should at least be equal to the number of eigenvalues in the interval.

With MINRES, we use the harmonic Ritz values and vectors. Paige, Parlett, and van der Vorst [29] call $(\hat{\theta}_j, \hat{u}_j)$ a harmonic Ritz pair for $A(\alpha)$ if

$$\underline{H}_k^* \underline{H}_k \hat{y}_j = \hat{\theta}_j H_k^* \hat{y}_j \tag{4.3}$$

with $\hat{y}_j \neq 0$ and $\hat{u}_j = V_k \hat{y}_j$. We define the harmonic Ritz values of $M^{-1}K$ by the inverse Cayley transform $\hat{\lambda}_j = c^{-1}(\hat{\theta}_j)$; see (1.8). The $\hat{\lambda}_j$'s depend on α. Thus, we have

$$\hat{x}(\alpha) = \beta V_k \sum_{j=1}^{k} \frac{\hat{\lambda}_j(\alpha) - \sigma}{\hat{\lambda}_j(\alpha) - \alpha} (\hat{y}(\alpha)_j^* e_1) \hat{y}_j(\alpha) \ . \tag{4.4}$$

The following theorem and the next section illustrate that MINRES cannot identify a singularity in $x(\alpha)$ as clearly as the Lanczos method. The next theorem considers the situation where $\alpha = \tilde{\lambda}_1$.

THEOREM 4.2. *Assume that $\alpha = \tilde{\lambda}_1$. Then σ is a harmonic Ritz value of $M^{-1}K$. Moreover, all $\hat{\theta}_j$ for $j = 1, \ldots, k$ satisfy*

$$|\hat{\theta}_j| \geq \frac{\sigma_{\min}(\underline{H}_k)}{\kappa(\underline{H}_k)}, \tag{4.5}$$

where $\sigma_{\min}$ denotes the smallest singular value and κ the condition number.

Proof. From (1.8), it is clear that $\tilde{\theta}_1 = 0$. Note that $H_k \tilde{y}_1 = 0$, but $\underline{H}_k^* \underline{H}_k \tilde{y}_1 \neq 0$ so that (4.3) has an infinite eigenvalue. The proof that σ is a harmonic Ritz value follows from $\lim_{\theta \to \infty} c^{-1}(\theta) = \sigma$. Second, from (4.3) and $\|\tilde{y}_j\|_2 = 1$, it follows that

$$|\hat{\theta}_j| \|H_k^* \hat{y}_j\| \geq \sigma_{\min}^2(\underline{H}_k)$$

and

$$|\hat{\theta}_j| \geq \frac{\sigma_{\min}^2(\underline{H}_k)}{\|H_k^* \hat{y}_j\|} \geq \frac{\sigma_{\min}^2(\underline{H}_k)}{\sigma_{\max}(\underline{H}_k)},$$

where $\sigma_{\max}$ denotes the largest singular vector. This proves the theorem. ∎

The theorem shows that the Ritz value $\tilde{\theta}_1 = 0$ corresponds to $\hat{\theta}_1 = \infty$, while there is no other harmonic Ritz value closer to 0 than that given by (4.5). The fact that the harmonic Ritz values depend on α is important. Since $\underline{H}_k$ always has full rank and $\hat{\lambda}_j$ can never be equal to α, $\hat{x}(\alpha)$ is always bounded. This implies that peaks may not always be identified as clearly as with the Lanczos method.

4.2. Minimization properties. MINRES makes the residual norm $\|r(\alpha)\|_M$ minimum, while the Lanczos method makes $r(\alpha)$ M-orthogonal to $\mathcal{K}_k(A(\alpha), b)$. If $A(\alpha)$ has positive eigenvalues, then the Lanczos method is equivalent to conjugate gradients (CG) so that the error $e(\alpha) = x(\alpha) - \tilde{x}(\alpha)$ is minimized with respect to the $MA(\alpha)$ norm, i.e.,

$$e(\alpha)^* MA(\alpha) e(\alpha) = r(\alpha)^* A(\alpha)^{-1} M r(\alpha)$$

is minimized. In general, $MA(\alpha)$ is not a positive definite matrix. MINRES minimizes the error with respect to the $A(\alpha)^* MA(\alpha)$ norm, i.e.,

$$e(\alpha)^* A(\alpha)^* MA(\alpha) e(\alpha) = r(\alpha)^* M r(\alpha).$$

MINRES minimizes the error in the direction of the eigenvectors with a weighting factor equal to the square of the eigenvalues. Let (λ_j, u_j) denote the eigenpairs of $M^{-1}K$. Then

$$\text{MINRES} : \min \quad \sum_{j=1}^{n} \left(\frac{\lambda_j - \alpha}{\lambda_j - \sigma} \right)^2 (u_j^* M e(\alpha))^2 \ ,$$

$$\text{CG (Lanczos)} : \min \quad \sum_{j=1}^{n} \left(\frac{\lambda_j - \alpha}{\lambda_j - \sigma} \right) (u_j^* M e(\alpha))^2 \ .$$

The eigenvalues of $A(\alpha)$ near 1 are not magnified or wiped out. MINRES reduces the error in the direction of the dominant eigenvectors of $A(\alpha)$ more effectively, but it does not easily reduce the error in the direction of the dominant eigenvectors of $A(\alpha)^{-1}$. These eigenvectors make the peaks in $\|x(\alpha)\|$ and are most important for the approximation. The conclusion is that, although MINRES minimizes the residual norm, the Lanczos method may produce more accurate results.

The following analysis shows that when α approaches a Ritz value $\tilde{\lambda}_j$, so that $|\tilde{\theta}_j|$ is small, the MINRES method may have difficulty identifying a clear peak in $\|x(\alpha)\|$. The MINRES method is usually described in terms of harmonic Ritz values. The next theorem shows that if $\alpha = \tilde{\lambda}_j$, the behavior of MINRES can be explained in terms of Ritz values.

THEOREM 4.3. *Let $k > 1$. Let $\hat{z}$ minimize $\|\underline{H}_k z - e_1\|_2$ for $z \in \mathbf{R}^k$. Let $(\tilde{\theta}_i, \tilde{y}_i)$ with $\|\tilde{y}_i\|_2 = 1$ for $i = 1, \ldots, k$ be the eigenpairs of H_k. Let $\rho_i = \eta_k e_k^* \tilde{y}_i$ with η_k defined by (2.5). If $\tilde{\theta}_j = 0$ and $\tilde{\theta}_i \neq 0$ when $i \neq j$, then*

$$\tilde{y}_i^* \hat{z} = \frac{\tilde{y}_i^* e_1}{\tilde{\theta}_i} \quad \text{for} \quad i = 1, \ldots, j-1, \quad j+1, \ldots, k \ , \tag{4.6}$$

$$\tilde{y}_j^* \hat{z} = - \sum_{i=1, i \neq j}^{k} \frac{\rho_i}{\rho_j} (\tilde{y}_i^* \hat{z}) \ . \tag{4.7}$$

Proof. Minimizing $\|\underline{H}_k z - e_1\|$ is equivalent to solving

$$(H_k^2 + \eta_k^2 e_k e_k^*) \hat{z} = \underline{H}_k^* e_1 = H_k^* e_1 \ . \tag{4.8}$$

Define $Y_k = [\tilde{y}_1, \ldots, \tilde{y}_k]$ and $\Theta_k = \operatorname{diag}(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$. Decompose $\hat{z} = Y_k g$ and define $r^* = \eta_k e_k^* Y_k = [\rho_1, \ldots, \rho_k]$. Then (4.8) is equivalent to

$$(\Theta_k^2 + r r^*) g = \Theta_k Y_k^* e_1 \ . \tag{4.9}$$

Since $\tilde{\theta}_j = 0$, the jth row in (4.9) produces $r^* g = 0$, from which (4.7) follows. Row $i \neq j$ produces (4.6). ∎

The next theorem describes the behavior of the Lanczos method and MINRES for α's near $\tilde{\lambda}_j$.

THEOREM 4.4. *Let $k > 1$. Define $\delta = \alpha - \tilde{\lambda}_j \in \mathbf{R}$. Recall the definitions of $\hat{z}$, ρ_i, $\tilde{\theta}_i$, and $\tilde{y}_i$ from Theorem 4.3. Decompose the solution of $\min_z \|\underline{H}_k(\alpha) z - e_1\|$ into*

$$\hat{z}(\alpha) = \hat{z} + \delta d + O(\delta^2); \tag{4.10}$$

then

$$\tilde{y}_i^* d = - \frac{\rho_i}{\rho_j} \frac{\tilde{y}_j^* e_1}{(\sigma - \tilde{\lambda}_j) \tilde{\theta}_i^2} + \frac{\tilde{y}_i^* e_1}{\tilde{\theta}_i^2} \frac{\tilde{\theta}_i - 1}{\sigma - \tilde{\lambda}_j}, \tag{4.11}$$

$$\tilde{y}_j^* d = \frac{\tilde{y}_j^* e_1}{\rho_j^2(\sigma - \tilde{\lambda}_j)} - \sum_{i=1, i\neq j}^{k} \frac{\rho_i}{\rho_j} \tilde{y}_i^* d \ . \tag{4.12}$$

Proof. Denote $\underline{H}_k = \underline{H}_k(\tilde{\lambda}_j)$. Decompose $\underline{H}_k(\alpha) = \underline{H}_k - \delta \underline{T}_k$ and minimize

$$\|(\underline{H}_k - \delta \underline{T}_k)\hat{z}(\alpha) - e_1\|_2 \ . \tag{4.13}$$

The analysis is similar to the proof of Theorem 5.3.1 in [16]. By definition (4.10), $d = (\partial \hat{z}(\alpha)/\partial \alpha)_{|\alpha=\tilde{\lambda}_j}$. The solution of (4.13) satisfies

$$(\underline{H}_k - \delta \underline{T}_k)^*(\underline{H}_k - \delta \underline{T}_k)(\hat{z} + \delta d) = (\underline{H}_k - \delta \underline{T}_k)^* e_1 + O(\delta^2).$$

Differentiation for δ and setting $\delta = 0$ gives

$$\underline{H}_k^* \underline{H}_k d = \underline{H}_k^* \underline{T}_k \hat{z} + \underline{T}_k^*(\underline{H}_k \hat{z} - e_1) \ . \tag{4.14}$$

Using the eigendecomposition of H_k, we have

$$\underline{H}_k = \begin{bmatrix} Y_k \Theta_k \\ r^* \end{bmatrix} Y_k^*, \qquad \underline{T}_k = \frac{1}{\sigma - \tilde{\lambda}_j} \begin{bmatrix} Y_k(\Theta_k - I) \\ r^* \end{bmatrix} Y_k^*,$$

where $r^* = [\rho_1, \ldots, \rho_k]$. Following Theorem 4.3, $\underline{H}_k \hat{z} - e_1 = -\binom{\tilde{y}_j(\tilde{y}_j^* e_1)}{0}$. With $d = Y_k t$ and $\hat{z} = Y_k g$, (4.14) becomes

$$\begin{bmatrix} \Theta_k \\ r^* \end{bmatrix}^* \begin{bmatrix} \Theta_k \\ r^* \end{bmatrix} t = \frac{1}{\sigma - \tilde{\lambda}_j} \begin{bmatrix} \Theta_k \\ r^* \end{bmatrix}^* \begin{bmatrix} \Theta_k - I \\ r^* \end{bmatrix} g - \frac{1}{\sigma - \tilde{\lambda}_j} \begin{bmatrix} \Theta_k - I \\ r^* \end{bmatrix}^* \begin{bmatrix} e_j(\tilde{y_j^*} e_1) \\ 0 \end{bmatrix},$$

$$(\Theta_k^2 + r r^*) t = \frac{1}{\sigma - \tilde{\lambda}_j}(\Theta_k(\Theta_k - I)g - (\Theta_k - I)e_j(\tilde{y}_j^* e_1)),$$

since $r^* g = 0$ (see the proof of Theorem 4.3). The jth equation produces

$$\rho_j r^* t = \frac{\tilde{y}_j^* e_1}{\sigma - \tilde{\lambda}_j},$$

from which we derive (4.12). The ith equation becomes

$$\tilde{\theta}_i^2 e_i^* t = \frac{1}{\sigma - \tilde{\lambda}_j} \tilde{\theta}_i(\tilde{\theta}_i - 1) e_i^* g - \rho_i r^* t,$$

from which (4.11) follows. ∎

From Theorem 4.3, we see that the solution computed by the MINRES method behaves in a very similar way to the Lanczos solution for the components in the direction of the Ritz vectors corresponding to $\tilde{\lambda}_j$ away from α (i.e., large $|\tilde{\theta}_j|$). The $\tilde{\lambda}_j$ near α do not participate in the same way: in the Lanczos method, they give rise to an unbounded solution, while in the MINRES method the ratio of the residual norms ρ_i of the Ritz values of $A(\alpha)$ determines the norm of the solution. Clearly, the more accurate $\tilde{\lambda}_j$ (smaller ρ_j) is, the larger is the component in the direction of $\tilde{y}_j$ and the larger is $\|z(\alpha)\|_2$ for the MINRES method.

From Theorem 4.4, it follows that if ρ_j is small, a small change in α may cause a large change in the MINRES solution. If ρ_j is large, there is a fairly large interval of α's around $\tilde{\lambda}_j$ for which the MINRES solution does not rapidly change and thus cannot produce a clear peak.

4.3. Relation with PVL. The PVL method can be formulated for solving the problem

$$\begin{aligned} A(\alpha)x(\alpha) &= b, \\ y(\alpha) &= c^* M x(\alpha), \end{aligned}$$

where $y(\alpha)$ is the output. The two-sided Lanczos method is used to build the two Krylov subspaces $\mathcal{K}_{k+1}(A(\alpha), b)$ and $\mathcal{K}_{k+1}(A(\alpha), c)$. The response is approximated as

$$\tilde{y}(\alpha) = \beta \|c\|_M e_1^* (I + (\sigma - \alpha)T_k)^{-1} e_1.$$

When we choose $c = b$, both Krylov spaces collapse and the tridiagonal matrix is the one computed by the symmetric Lanczos method. Note that PVL does not solve a least squares problem, so it is closer to the Lanczos method than to MINRES. The PVL method matches $2k-1$ Taylor coefficients of $y(\alpha)$ and $\tilde{y}(\alpha) = c^* M \tilde{x}(\alpha)$ [11].

The following theorem shows the connection between the Taylor series of $x(\alpha)$ and the Lanczos approximation $\tilde{x}(\alpha)$.

THEOREM 4.5. *If we define the Taylor polynomials*

$$\begin{aligned} x(\alpha) &= x_0 + (\alpha - \sigma)x_1 + (\alpha - \sigma)^2 x_2 + \cdots, \\ \tilde{x}(\alpha) &= \tilde{x}_0 + (\alpha - \sigma)\tilde{x}_1 + (\alpha - \sigma)^2 \tilde{x}_2 + \cdots, \end{aligned}$$

then $\tilde{x}_j = x_j$ *for* $j = 0, \ldots, k-1$.

Proof. Equation (2.1) is equivalent to

$$(I + (\sigma - \alpha)A)x(\alpha) = b.$$

Ordering following powers of $(\sigma - \alpha)$ produces $x_0 = b$ and $x_j = Ax_{j-1}$ for $j > 1$. Decompose

$$z(\alpha) = z_0 + (\alpha - \sigma)z_1 + (\alpha - \sigma)^2 z_2 + \cdots.$$

From $(I + (\sigma - \alpha)T_k)z(\alpha) = \beta e_1$ we find that $z_0 = \beta e_1$ and $z_j = T_k z_{j-1}$. By induction on j, we find $z_j = \beta T_k^j e_1$ for $j \geq 0$. The proof follows by noting that $\tilde{x}_j = \beta V_k T_k^j e_1 = A^j v_1 \beta = A^j b$ for $0 \leq j \leq k-1$. ∎

Only k Taylor coefficients of $x(\alpha)$ and $\tilde{x}(\alpha)$ match, but (4.1) is a vector-Padé approximation, where the poles are the Ritz values. Taylor polynomials usually have small convergence intervals, but the poles in the vector-Padé approximation make the approximation converge in the presence of singularities (eigenvalues).

5. Error estimation. Error estimates for PVL are studied by Skoogh [37] and Bai and Ye [4]. In the next section we suggest simple and cheap error bounds inspired by the work of Skoogh [37]. The results in [4] and [37] illustrate that the error is hard to estimate. The estimates are heuristics. Antoulas and Sorensen [1] call this one of the drawbacks of Krylov methods for model reduction. The heuristics we discuss here are thus very rough approximations to the error and are not guaranteed to produce accurate estimates in all situations.

The error of the Lanczos solution, $e(\alpha) = x(\alpha) - \tilde{x}(\alpha)$, satisfies $e(\alpha) = A(\alpha)^{-1} r(\alpha)$ with $r(\alpha) = b - A(\alpha)\tilde{x}(\alpha)$. In this section, we present two bounds. We will compare them by numerical examples in section 8.

The most straightforward bound for $\|e(\alpha)\|_2$ is

$$\|A(\alpha)^{-1}\|_M \sqrt{\|M^{-1}\|_2} \|r(\alpha)\|_M \simeq \max_{j=1,\ldots,k} \left| \frac{\tilde{\lambda}_j - \sigma}{\tilde{\lambda}_j - \alpha} \right| \sqrt{\|M^{-1}\|_2}\, \|r(\alpha)\|_M.$$

The maximum must be taken over the eigenvalues of $Ku = \lambda Mu$. In practice, these are unknown, so we use the Ritz values. Since these must be accurate to guarantee a small $\|e(\alpha)\|$, the estimate of $\|A(\alpha)^{-1}\|_M$ can be rather sharp. The bound may be too large, since $r(\alpha)$ has a small component in the Ritz vectors that make small angles with eigenvectors. The norm $\sqrt{\|M^{-1}\|_2}$ can be estimated as $\max_{i=1,\ldots,k+1} \|v_i\|_2$, so, in practice, we can use the estimate

$$\text{Err}_1 := \max_{j=1,\ldots,k} \left| \frac{\tilde{\lambda}_j - \sigma}{\tilde{\lambda}_j - \alpha} \right| \left(\max_{i=1,\ldots,k+1} \|v_i\|_2 \right) \|r(\alpha)\|_M. \tag{5.1}$$

Let us now try to develop a more accurate estimate. The residual is proportional to v_{k+1}, so we must approximate $\|A(\alpha)^{-1}v_{k+1}\|_M$. Let $w \in \mathbf{R}^k$ be chosen so that $e_k^* w = \beta_k^{-1}$. (If $\beta_k = 0$, then $\tilde{x}(\alpha) = x(\alpha)$ and so the error estimation is no longer required.) Then

$$v_{k+1} = V_{k+1}\underline{T}_k w - V_k T_k w. \tag{5.2}$$

We first estimate

$$\gamma_k := \|A(\alpha)^{-1}V_{k+1}\underline{T}_k w\|_M = \|(K - \alpha M)^{-1}MV_k w\|_M.$$

Consider the eigendecomposition $T_k Y = Y\tilde{\Theta}$, $\tilde{\Lambda} = \sigma I + \tilde{\Theta}^{-1}$ and the matrix of Ritz vectors $U_k = V_k Y$. Then approximate $KU_k \approx MU_k\tilde{\Lambda}$ and

$$(K - \alpha M)^{-1}M \approx U_k(\tilde{\Lambda} - \alpha I)^{-1}U_k^* M.$$

Hence we find

$$\gamma_k \approx \tilde{\gamma}_k(w) := \|(\tilde{\Lambda} - \alpha I)^{-1}Y^* w\|_2.$$

Next, we estimate

$$\|A(\alpha)^{-1}V_k T_k w\|_M = \|(\tilde{\Lambda} - \alpha I)^{-1}(\tilde{\Lambda} - \sigma I)Y^* T_k w\|_2 \simeq \|(\tilde{\Lambda} - \alpha I)^{-1}Y^* w\|_2 = \tilde{\gamma}_k(w).$$

Thus, we get

$$\|A(\alpha)^{-1}v_{k+1}\|_M \le \|A(\alpha)^{-1}V_{k+1}\underline{T}_k w\|_M + \|A(\alpha)^{-1}V_k T_k w\|_M \simeq 2\tilde{\gamma}_k(w)$$

and use

$$\text{Err}_2 := \tilde{\gamma}_k(w) \left(\max_{i=1,\ldots,k+1} \|v_i\|_2 \right) \|r(\alpha)\|_M$$

as an estimate for $\|e(\alpha)\|_2$. The problem now is which w is suitable for our estimate. One possibility is to choose $w = e_k\beta_k^{-1}$. The smallest estimate can be found by choosing w so that $\tilde{\gamma}_k(w)$ is minimal, i.e.,

$$\min_{w:\beta_k e_k^* w = 1} \|(\tilde{\Lambda} - \alpha I)^{-1}Y^* w\|_2.$$

We found this a good choice in our experiments.

6. Accuracy of the inversion of $K-\sigma M$. The quality of the approximation to the solution of (1.1) is bounded by the accuracy of the application of $(K-\sigma M)^{-1}$. In step 3.1 of Algorithm 2.1, let the residual be $s_j = Mv_j - (K-\sigma M)w_j$, and collect all the residuals s_j for $j = 1, \ldots, k$ in S_k. Then the recurrence relation (2.3) becomes

$$(K-\sigma M)^{-1} M V_k - V_{k+1}\underline{T}_k = (K-\sigma M)^{-1} S_k.$$

We can rewrite this relation using a backward error E on $K-\sigma M$ as follows:

$$(K-\sigma M+E)^{-1} M V_k - V_{k+1}\underline{T}_k = 0, \quad \text{where} \quad S_k = -E V_{k+1}\underline{T}_k \ . \tag{6.1}$$

This is the recurrence relation for the perturbed matrix $(K-\sigma M+E)^{-1}M$. V_{k+1} and $\underline{T}_k$ can be considered as computed by the (exact) Lanczos method applied to $(K-\sigma M+E)^{-1}M$. The following lemma shows the impact on the error of the solution of (1.3) by the Lanczos and MINRES methods.

LEMMA 6.1. *Let $\tilde{A}(\alpha) = (K-\sigma M+E)^{-1}(K-\alpha M)$. The error for the Lanczos method $e(\alpha) = x(\alpha) - \tilde{x}(\alpha)$ satisfies*

$$e(\alpha) = -\eta_k \tilde{A}(\alpha)^{-1} v_{k+1} e_k^* \tilde{z}(\alpha) + (K-\alpha M)^{-1} E \tilde{x}(\alpha) \ . \tag{6.2}$$

Proof. After shifting (6.1), we get the relationship

$$\tilde{A}(\alpha) V_k - V_{k+1}(I + (\sigma-\alpha)\underline{T}_k) = -(K-\sigma M+E)^{-1} E V_k.$$

Recall that the solution $x(\alpha)$ is approximated by $\tilde{x}(\alpha) = V_k \tilde{z}(\alpha)$. The residual computed by the Lanczos method is

$$\begin{aligned} r(\alpha) &= b - \tilde{A}(\alpha)\tilde{x} \\ &= -v_{k+1}\eta_k e_k^* \tilde{z}(\alpha) + (K-\sigma M+E)^{-1} E V_k \tilde{z}(\alpha) \ . \end{aligned}$$

The error (6.2) follows from $e(\alpha) = \tilde{A}(\alpha)^{-1} r(\alpha)$. This proves the lemma. ∎

The first term in the right-hand side of (6.2) is what we obtain by using the data V_{k+1} and $\underline{T}_k$ that we obtain from the Lanczos method in Algorithm 2.1, as if $E = 0$. The shift-invariance property is violated when $E \neq 0$. This is where the second term comes from. The lemma says that even if $e_k^* \tilde{z}(\alpha)$ tends to zero for increasing k, $e(\alpha)$ does not necessarily go to zero. It all depends on the impact of E on $\tilde{x}(\alpha)$. The norm of the first term can be estimated by Err_1 or Err_2 from section 5.

In order to estimate the last term in the right-hand side of (6.2), note that

$$\|(K-\alpha M)^{-1}\|_2 \geq \|x(\alpha)\|_2 / \|f\|_2.$$

This can be used to estimate

$$\|(K-\alpha M)^{-1} E \tilde{x}(\alpha)\|_2 \approx \|x(\alpha)\|_2 / \|f\|_2 \, \|E\|_2 \, \|\tilde{x}(\alpha)\|_2 \ . \tag{6.3}$$

The residual term s_j must have a norm small enough to guarantee the desired error level for the solution. If an iterative linear solver is used for step 3.1 in Algorithm 2.1, then we can control this. If a direct linear solver is employed, we usually have that

$$\|s_j\|_2 \leq \|K-\sigma M\| \|w_j\|_2 \epsilon,$$

where ϵ is the machine precision. This corresponds to $\|E\|_1 \approx \|K-\sigma M\|_1 \epsilon$. Replacing $x(\alpha)$ by $\tilde{x}(\alpha)$ and $\|E\|_2$ by $\|K-\sigma M\|_1 \epsilon$ in (6.3), we obtain the estimate

$$\mathrm{Err}_3 := \|\tilde{x}(\alpha)\|_2^2 / \|f\|_2 \|K-\sigma M\|_1 \epsilon \tag{6.4}$$

for (6.3). Adding estimates of the two terms in the right-hand of (6.2) then produces

$$\|e(\alpha)\| \simeq \mathrm{Err}_j + \mathrm{Err}_3 \quad \text{with} \quad j = 1 \text{ or } 2, \tag{6.5}$$

where Err_1 or Err_2 are used to estimate the first term in the right-hand side of (6.2).

7. Implementation issues. The Lanczos vectors v_j computed by Algorithm 2.1 may lose orthogonality. Iterative linear system solvers still converge but may need a few more iterations to obtain the solution with the same accuracy as with orthogonal Lanczos vectors.

A major point requiring attention is the memory usage. The α's are usually available as a discrete set $\alpha_1, \ldots, \alpha_m$, where m can be relatively high, say 20 to 200. In simulation codes the solution $x(\alpha_j)$ is passed on to a postprocessing procedure or stored in a database. In practice, $\tilde{x}(\alpha_1), \ldots, \tilde{x}(\alpha_m)$ cannot be stored simultaneously in core.

We consider two different approaches. Algorithm 7.1 requires the storage of the iteration vectors v_j, $j = 1, \ldots, k+1$, and Algorithm 7.2 requires the storage of all solution vectors. We now discuss the two algorithms.

7.1. Algorithm 7.1: Storing the iteration vectors. The first algorithm stores only the iteration vectors. The idea is that if the number of solution vectors, m, is significantly larger than k, we can reduce the memory cost.

ALGORITHM 7.1.

1. Discretize the interval $[\alpha_{\min}, \alpha_{\max}]$ into $\{\alpha_1, \ldots, \alpha_m\}$.
2. Choose σ and factorize $K - \sigma M$.
3. Let $b = (K - \sigma M)^{-1} f$, $\beta = \|b\|_M$, and $v_1 = b/\beta$.
4. Build the Krylov subspace for a fixed k.
5. For $j = 1, \ldots, m$
   5.1. Compute $\tilde{x}(\alpha_j)$.
   5.2. If the solution does not satisfy the stopping criterion, select a new σ, and go to step 2.

The advantage of this approach is that we do not have to store all $\tilde{x}(\alpha_j)$ but only V_k. Instead we compute $\tilde{x}(\alpha) = \beta V_k H_k(\alpha)^{-1} e_1$. Each new σ requires the additional storage of k iteration vectors.

Since the iteration vectors are stored, we can use reorthogonalization to reduce the number of iterations.

In ACTRAN [7], k is fixed based on the ratio of factorization cost and cost for a linear solve with $K - \sigma M$. The philosophy is that the cost for factorization and building the Krylov subspace must be in balance. A new σ requires a new factorization, but it brings σ closer to the α's for which the solution should be computed, so it improves the convergence speed. The first $\sigma = \alpha_1$. The $\tilde{x}(\alpha_j)$ are computed from $j = 1$ until $j = m$ in this order. When the stopping criterion fails in step 5.2, we select $\sigma = \alpha_j$ and continue with step 2. This is a very simple but robust strategy: even if the Lanczos method never converges for $\alpha \neq \sigma$, the solution for $\alpha_j = \sigma$ is computed since $x(\alpha_j) = b$. So, in the worst case, σ takes the values α_j for $j = 1, \ldots, m$. The choice of σ is still an open question.

7.2. Algorithm 7.2: Storing the solution vectors. The preceding approach requires k to be fixed beforehand. The following algorithm updates the solution on each iteration of the Lanczos method.

ALGORITHM 7.2.

1. Discretize the interval $[\alpha_{\min}, \alpha_{\max}]$ into $\alpha_1, \ldots, \alpha_m$.
2. Initialize the solutions $\tilde{x}_j = 0$ for $j = 1, \ldots, m$.
3. Choose σ and factorize $K - \sigma M$.
4. Let $b = (K - \sigma M)^{-1} f$, $\beta = \|b\|_M$, and $v_1 = b/\beta$.
5. For $k = 1, 2, \ldots$
   5.1. Compute the iteration vector v_{k+1} from v_{k-1} and v_k.

5.2. Update the solutions $\tilde{x}_j$ for $j = 1, \ldots, m$.

5.3. Stop when all solutions satisfy the stopping criterion.

We refer to [35] for a practical implementation of the update of $\tilde{x}_j$ in step 5.2. Note that only v_{k-1} and v_k need to be stored to form v_{k+1} in step 5.1. However, for each α we need to store not only $\tilde{x}$ but also auxiliary vectors that depend on α. In the case of the Lanczos method (CG) we have to store one vector for each α for handling the CG vectors. See [16] and Algorithm 6.16 in Saad [33]. In the case of MINRES, we need two additional vectors for each value of α.

This approach is attractive when $m \approx k$ or when only a few components of $x(\alpha)$ are wanted. In the latter case, we need to store only these components. Unfortunately, this is not the case in many industrial codes that have been developed over many years. Reorthogonalization is impractical because the iteration vectors are not stored simultaneously. An advantage is that k need not be fixed beforehand. The user can, in step 5.3, decide to change σ. For the Lanczos method, this requires a jump to step 4 with $b = v_{k+1}$, which is the direction of the residual vector. For the MINRES method, this is not possible since the direction of the residual depends on α.

The disadvantage of Algorithm 7.2 is that $[\alpha_{\min}, \alpha_{\max}]$ must be discretized beforehand, while Algorithm 7.1 allows $[\alpha_{\min}, \alpha_{\max}]$ to be discretized when the Ritz values are computed. This may allow the use of a finer discretization when there are many peaks in $[\alpha_{\min}, \alpha_{\max}]$ and a coarser discretization if the number of peaks is low. A practical algorithm may combine both approaches so that refinement of $[\alpha_{\min}, \alpha_{\max}]$ in Algorithm 7.1 is easy and the choice of k is flexible, as in Algorithm 7.2.

7.3. Choice of σ. In section 6, we analyzed the influence of the quality of the preconditioner on the error of the approximation. Since

$$\|E\| \sim \|K - \sigma M\|\epsilon \sim \max(\|K\|, |\sigma|\|M\|)\epsilon,$$

the choice of σ may influence the level of accuracy that can be reached. In the low frequency range, $\|K - \sigma M\| \approx \|K\|$. In section 8.4, we show that a σ near an eigenvalue may quickly lead to loss of orthogonality of the Lanczos vectors and thus to a delay of convergence. Strategies for choosing poles have been proposed by Grimme [20] and Grimes, Lewis, and Simon [18]. Although the latter paper deals with the solution of eigenvalue problems, the problem of the choice of pole is very similar to the solution of parametrized linear systems, since the choice of pole determines the eigenvalues that converge in k iterations. It is, in general, not easy to determine which eigenvalues converge first, but as a rule of thumb, it is correct to state that the eigenvalues nearest to σ come first. This makes it natural to choose σ near the α's of interest. With the selection of σ in [18] we compute $x(\alpha)$ in a relatively small interval $[\alpha_i, \alpha_{i+1}]$ and then move this interval using a new σ. The choice of pole is a problem that deserves more attention.

When σ is close to an eigenvalue, T_k has very small and very large eigenvalues in modulus. This may lead to very large and very small elements in $|\underline{T}_k|$. Following section 4.3 in [18], this is a situation where orthogonality may be quickly lost. The ratio

$$\frac{\max_j\{|\alpha_j|, |\beta_j|\}}{\min_j |\beta_j|}$$

is a measure of the growth of the loss of orthogonality.

8. Numerical examples. All our examples arise from structural finite element problems. They were generated and run within the ACTRAN [7] simulation package

for acoustic transmission. The ACTRAN direct linear solver was used for the preconditioner. All figures in this section show the frequency $\omega = \sqrt{\alpha}$ on the horizontal axis and $\log_{10}$ of another quantity on the vertical axis, e.g., one component of $\tilde{x}(\alpha)$, $\|e(\alpha)\|_2$ or $\|r(\alpha)\|_2$. The computations were performed on an HP B2000 workstation running HP-UX10.20.

We have chosen Algorithm 7.1 for our computations since the number of α's is far larger than the number of Lanczos vectors. We have skipped the stopping criterion in step 4.2 so that we have used just one matrix factorization for a fixed σ.

8.1. Description of the test problems.

Aluminum plate. The first test problem illustrates the method for a structural model of an aluminum plate measuring 0.5m×0.5m, with Young modulus $7.0\,10^{10}\mathrm{N/m}^2$, thickness 0.001m, Poisson ratio 0.33, density $2700\mathrm{kg/m}^3$, and no structural damping. The plate was discretized by a grid of 16×16 solid shell elements (HEX08). The plate is subjected to a unit point force in the coordinate $(0.125\mathrm{m}, 0.125\mathrm{m})$. The goal is to compute the amplitude of the vertical displacement $x(\alpha)$ in the same position for the frequency range $[10, 110]$ (Hertz). In order to generate the plots the frequency range was discretized as $\{\omega_1, \ldots, \omega_m\} = \{10 + (j-1), j = 1, \ldots, m\}$ with $m = 101$. We used $\alpha_j = \omega_j^2$ for $j = 1, \ldots, m$. The problem has order $n = 1734$.

Car windscreen. The second test problem illustrates the method on a structural model of a car windscreen. This is a three-dimensional problem discretized with 7564 nodes and 5400 linear hexahedral elements HEX08 (3 layers of 60×30 elements). The material is glass with the following properties: the Young modulus is $7.0\,10^{10}\mathrm{N/m}^2$, the density is $2490\mathrm{kg/m}^3$, and the Poisson ratio is 0.23. The natural damping of glass is not taken into account. The structural boundaries are free (free-free boundary conditions). The plate is subjected to a point force applied on node 1891 (i.e., on a corner).

The discretized problem has dimension $n = 22,692$. The goal is to estimate $x(\alpha)$ with $\alpha = \omega^2$ for $\omega \in [0.5, 200]$. In order to generate the plots the frequency range was discretized as $\{\omega_1, \ldots, \omega_m\} = \{0.5j, j = 1, \ldots, m\}$ with $m = 400$.

8.2. Choice of the pole σ. Before we give the major results for this paper we want to motivate the choice of the pole σ in our calculations.

Car windscreen. From the analysis in section 3, we see that σ should not be selected far away from α. In addition, σ should not be selected close to an eigenvalue because this might also introduce very large eigenvalues in $A(\alpha)$ with a risk of overflow. In this section, we compare the error norm $\|e(\alpha)\|_2$ for $k = 40$ iterations of the Lanczos method with $\sigma = -100$, 0.25, 100, and 384.813. Note that the σ's correspond to the frequencies $10\sqrt{-1}$, 0.5, 10, and 19.617, respectively. The second and last choices are near eigenvalues. The relative errors $\|e(\alpha)\|_2/|e_j^* \tilde{x}(\alpha)|$ with $j = 22,683$ in Figure 8.1 show that the convergence speed is similar for $\sigma = -100$ and $\sigma = 100$ but is slower for $\sigma = 0.25$, which is relatively close to the eigenvalue 0. The choice $\sigma = 384.813$ is best. In the remainder of the paper we use $\sigma = -100$ for the car windscreen problem. We have $\kappa_1(K - \sigma M) \simeq 10^{11}$.

Aluminum plate. For the aluminum plate problem, we use $\sigma = \alpha_1^2 = 100$ since there is no nearby eigenvalue. We choose the pole at the left end of the frequency range. Note the condition number $\kappa_1(K - \sigma M) \simeq 10^{10}$.

8.3. MINRES versus Lanczos.

Aluminum plate. Figure 8.2 compares the errors obtained for MINRES and the Lanczos method with $k = 10$. The results for the MINRES and Lanczos methods are

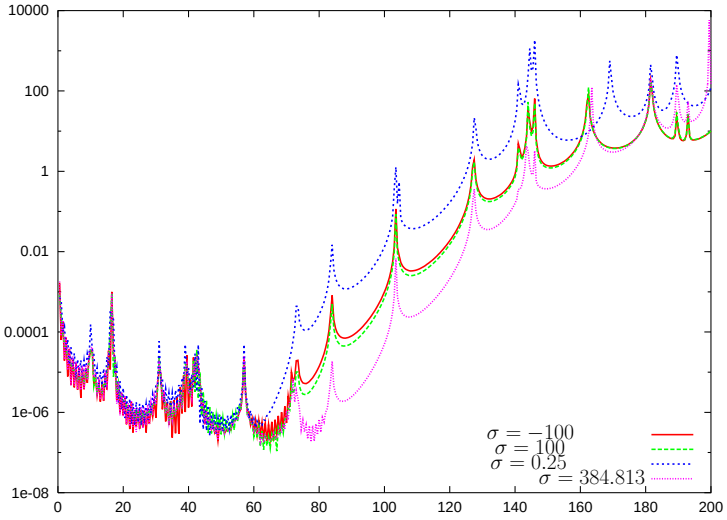


FIG. 8.1. *Car windscreen. Comparison of* $\|e(\alpha)\|_2/e_j^*\tilde{x}(\alpha)|$ *with* $j = 22,683$ *(vertical axis) versus the frequency* $\omega = \sqrt{\alpha}$ *(horizontal axis) for* $\sigma = -100$, $\sigma = 0.25$, $\sigma = 100$, *and* $\sigma = 384.813$.

the same except for α far away from σ where the convergence is slow, in particular for α's near eigenvalues of $M^{-1}K$. In all situations, MINRES minimizes the residual but not the error. Clearly, the Lanczos method is slightly better when α is far away from σ, i.e., where $A(\alpha)$ has small eigenvalues. There is no visible difference between the solution curves.

Let $\alpha = 10325$ ($\omega = 102$). The zero Ritz value of $H_k(\alpha)$ has $|\rho_j| = 7.1\,10^{-40}$. The other $|\rho_i|$ are between $1.9\,10^{-64}$ and $2.7\,10^{-3}$. The residual norm $|\rho_j|$ is fairly small, so, following Theorem 4.3, the solution is bounded by a large value. In addition, following Theorem 4.4, the least squares solution in MINRES can quickly grow large for α's around the Ritz value.

Car windscreen. In Figure 8.2 we compare the MINRES and Lanczos methods for $k = 20$. For this example, the difference between both methods is more pronounced than for the aluminum plate model. MINRES clearly minimizes the residual much more effectively than the Lanczos method. However, the solution computed by the Lanczos method is closer to the exact solution. To illustrate the difference in convergence behavior, we have computed the spectrum of $A(\alpha_{j_1})$ and $A(\alpha_{j_2})$ with $\omega_{j_1} = 20$ and $\omega_{j_2} = 180$. The eigenvalues of $A(\alpha_{j_1})$ are -0.9559, -0.5596, 0.7912, 0.8271, and the other eigenvalues are clustered around 1. The spectrum of $A(\alpha_{j_2})$ is -0.5237301, -0.6288101, -1.9862195, -2.068593, -5.1368919, -5.3885997, -17.614441, -21.483162, -209.58525, -166.91462, 0.0183553, 0.0994259, and the other eigenvalues are spread between 0.1 and 1. The matrix $A(\alpha_{j_2})$ has several eigenvalues near zero, and MINRES has more difficulties finding the solution in the direction of the corresponding eigenvectors. It is precisely these small eigenvalues that make the peaks in the solution. The bottom right-hand picture in Figure 8.2 shows that the peak near 189Hz is missed with MINRES.

Let $\alpha = 33184.8$ near the peak at frequency $\omega \approx 189$. The zero Ritz value of $H_k(\alpha)$ has a residual norm $|\rho_j| \approx 1.2\,10^{-2}$ while the smallest $|\rho_i|$ is $2.3\,10^{-58}$ and the largest is $6.3\,10^{-2}$. The residual norm of the Ritz value is fairly large, so, following Theorem 4.3, the norm of the MINRES solution remains relatively small. Following

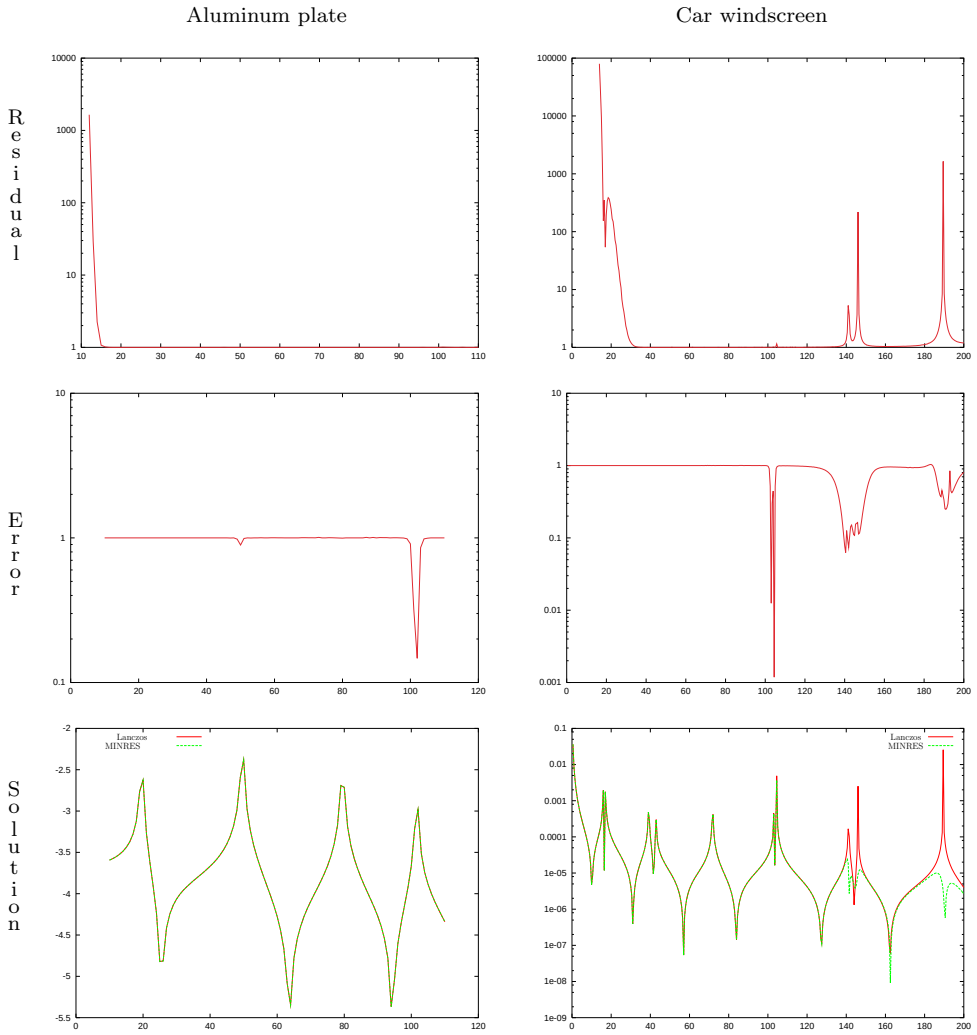


FIG. 8.2. *Aluminum plate/car windscreen. Comparison between the Lanczos method (solid line) and MINRES (dotted line). The horizontal axis shows the frequency* $\omega = \sqrt{\alpha}$. *The top figures show the ratio of the residual norms for the Lanczos method and MINRES as a function of* ω. *The middle figures show the ratio of the error norms in the Lanczos method and MINRES. The bottom figures show the absolute value of one component of the solution. All figures use a logarithmic scale on the vertical axis.*

Theorem 4.4, d is relatively small; i.e., the solution does not rapidly change with α. It should be noted that the situation is not as bad for MINRES as the numerical results would suggest. Since $|\rho_j|$ is relatively large, we conclude from Lemma 4.1 that $\|r(\alpha)\|_M$ is expected to be large too, and so the solution computed by the Lanczos method is not expected to be very accurate. This is confirmed by the error curve for $k = 20$ in Figure 8.7.

8.4. Reorthogonalization. Orthogonality of the Krylov vectors is not required for linear solvers, so the solution should still be computed accurately: the convergence may just be delayed. Since the Lanczos vectors are stored, we do not want to perform more iterations than necessary, since this requires more storage. Therefore we may

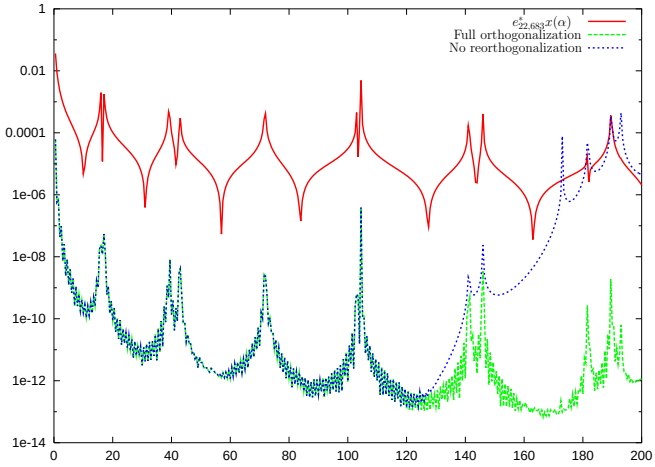


FIG. 8.3. *Car windscreen. Comparison of the error with and without reorthogonalization. The horizontal axis shows the frequency* $\omega = \sqrt{\alpha}$ *while the vertical axis shows the absolute value of the* $22,683$*th component in a logarithmic axis (solid line), the error when no reorthogonalization is used (dotted line), and the error when reorthogonalization is used (gray dots).*

perform partial reorthogonalization [31] as for eigenvalue computations. Since the major cost of the method lies in the backtransformation, the additional cost is low. In all our examples, we have imposed full orthogonality of the Lanczos vectors on each iteration.

Aluminum plate. We have recomputed the result with 20 iteration vectors using no reorthogonalization. We found the same solution, and the error curves did not differ much. Full reorthogonalization on each iteration increased the total computation time of Algorithm 7.1 (i.e., for sparse matrix factorization, back transformations, sparse matrix-vector products, error estimations, etc.) by less than five percent.

Car windscreen. Figure 8.3 compares the results obtained by full and no reorthogonalization for $k = 40$ iterations. As the figure shows, there is a clear difference in the accuracy of both cases for α far away from σ. This is a case where the Lanczos method (and MINRES) converges in more iterations when no reorthogonalization is employed. In this case, the total computation time for Algorithm 7.1 increased by only four percent due to reorthogonalization. This is negligible since, without reorthogonalization, additional iterations are required to obtain the same accuracy and this is far more expensive.

Illustration of a pole near an eigenvalue. We illustrate the loss of orthogonality when σ is close to a Ritz value. We ran $k = 20$ iterations of the Lanczos method for the aluminum plate problem for $\sigma = 100$ ($\omega = 10$) and $\sigma = 384.813$ ($\omega = 19.617$), which is near an eigenvalue. Figure 8.4 shows the errors for both poles with and without reorthogonalization. This is an illustration where loss of orthogonality becomes significant. The conclusion is that if reorthogonalization is not used, it is advisable to choose the pole in between a cluster of eigenvalues.

8.5. Illustration of error estimates.

Aluminum plate. Figure 8.5 shows the solution and the error estimates Err_1 and Err_2 (left) and Err_2 and Err_3 (right) for $k = 10$ Lanczos iterations. From the left panel we can see that the error bounds follow the exact error, but Err_2 is more accurate.

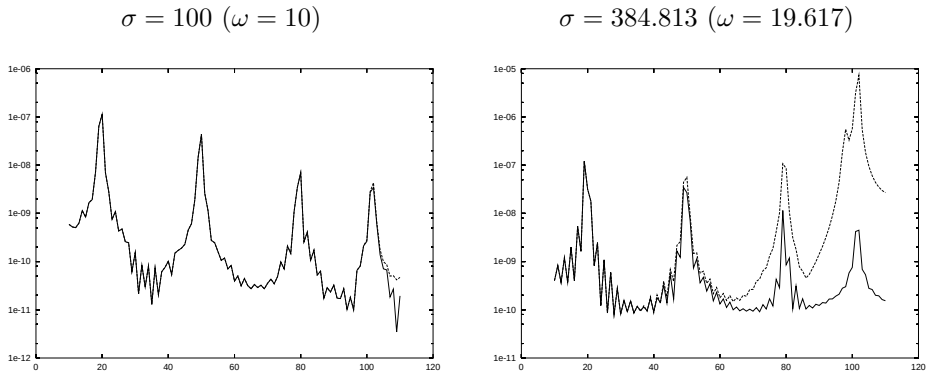


FIG. 8.4. *Aluminum plate. The solid line shows the error as a function of* $\omega = \sqrt{\alpha}$ *when reorthogonalization is used, and the dashed line shows the error when no reorthogonalization is used.*

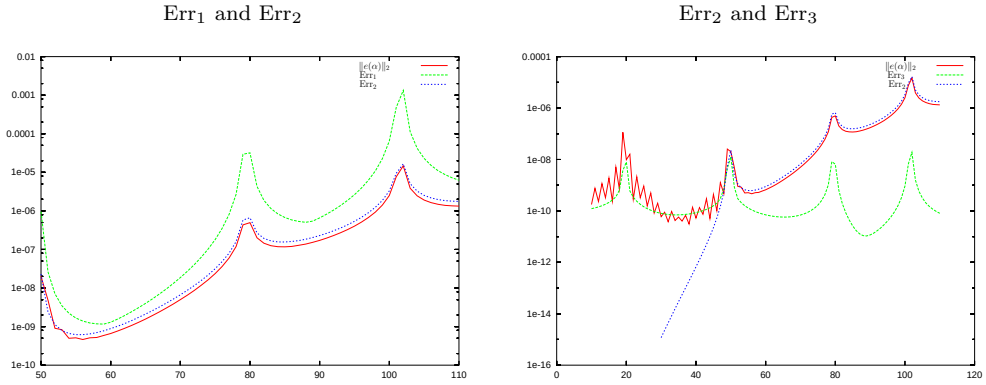


FIG. 8.5. *Aluminum plate. The horizontal axis shows the frequency* $\omega = \sqrt{\alpha}$ *while the vertical axis shows the absolute value of* $\|e(\alpha)\|_2$ *and the error estimates* Err_1 *and* Err_2.

Recall from (6.5) that $\|e(\alpha)\|_2$ is estimated as the sum of two terms. The term Err_3 shows the smallest error that we can obtain in finite precision arithmetic. The term Err_2 is small near σ and is zero when $\alpha = \sigma$. So, it is not a surprise that there is a point where Err_2 and Err_3 cross: when the solution computed by the Lanczos method becomes less accurate than the rounding errors, Err_2 dominates.

Figure 8.6 compares the results for different values of k. The exact $x(\alpha)$ is computed by a direct method whose result is improved by iterative refinement with two iterations of GMRES. To the right, there is an α for which Err_1 and Err_3 cross each other. On the left of this point, we can use Err_2 as error bound. On the right-hand side, we use Err_3 as an error bound. The estimate Err_3 is fairly accurate. Note that for all solutions, the exact solution and the approximation correspond well.

Car windscreen. We plotted the 22,683th component of $\tilde{x}(\alpha)$ together with the error estimates in Figure 8.7. This component corresponds to the vertical displacement on the same corner as the one subjected to the point force. On the left, the error first follows Err_3 and later Err_2. Note that Err_2 becomes larger than the solution for larger values of α. Thus, we have no accurate digits at all. The estimate Err_3 is always smaller than the solution. In the neighborhood of $\omega = 110$ the error Err_3 gets dangerously close to the solution. On the right, Err_2 is always smaller than

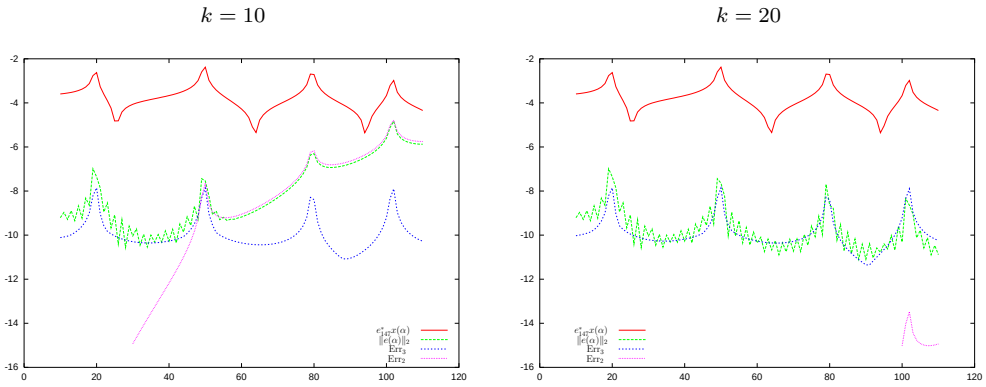


FIG. 8.6. *Aluminum plate. The horizontal axis shows the frequency* $\omega = \sqrt{\alpha}$. *The figure compares* $\|e(\alpha)\|_2$, *the error estimates* Err_2 *and* Err_3, *and component* 147 *of the exact solution.*

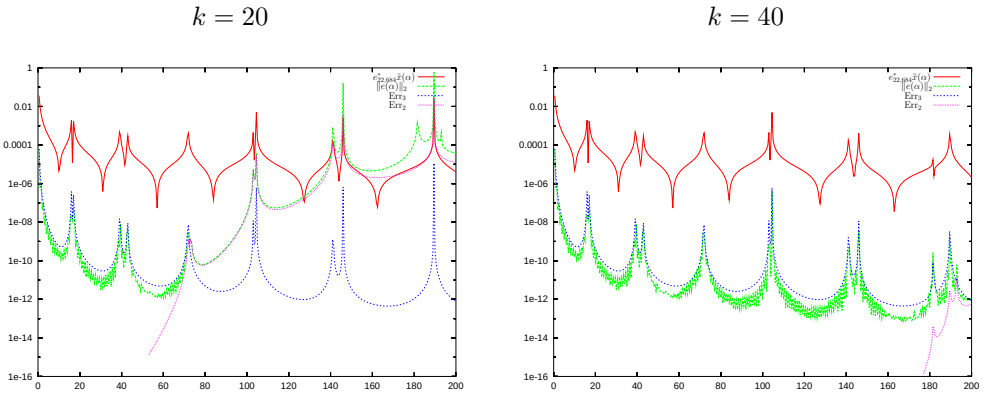


FIG. 8.7. *Car windscreen. Computed solution in unknown* 22,683, *the error norm* $\|e(\alpha)\|_2$, *and the estimates* Err_2 *and* Err_3.

Err_3. The accuracy of the solution is thus determined by Err_3. Note that Err_3 is an overestimate of the error in between two eigenvalues.

9. Conclusions. In this paper, we discussed a preconditioned Lanczos technique for solving the parametrized linear system (1.1) using a few sparse factorizations of $L(\sigma)$. The nice features are that the preconditioner is well suited for values of α near the reference value σ, and application of the preconditioner uses only one sparse matrix solve per iteration.

If the iteration vectors are stored, we suggest using reorthogonalization in the Lanczos method. In general, this reduces the number of iterations (and thus the number of vectors) for obtaining the same accuracy. If the iteration vectors are not stored, the pole should not be chosen near an eigenvalue, as this may quickly lead to loss of orthogonality and a need for significantly more iterations to attain the same accuracy. The choice of σ is not that important as long as σ is near the α's of interest.

The most important conclusion of this paper is that the Lanczos method appears to be more reliable than the MINRES method for estimating the peaks in the solution. The situation is not that bad for MINRES. We noticed that where MINRES did not identify clear peaks, the Lanczos method did not produce an accurate solution either.

From the error estimates we can conclude that the estimation of the rounding errors is reliable and that Err_2 is a better estimate than Err_1 for $\|e(\alpha)\|_2$ when the error is larger than the rounding errors. Since these are estimates, we should always be aware that there can be situations where the estimates are not as accurate as the numerical examples may suggest.

The choice (and change) of σ is not discussed in this paper. The change of σ may be interesting for computing $x(\alpha)$ when α is far away from σ. A technique developed for the eigenvalue problem, called rational Lanczos [27], can be used here.

Acknowledgments. I thank my colleague Thomas Leclercq for providing the test examples, Mickaël Robbé for reading the paper, and Serge Goossens, Paul Van Dooren, Henk van der Vorst, and Jörg Liesen for inspiring discussions. I am grateful to the referees who have encouraged me with many useful suggestions for further improvements of the paper.

REFERENCES

[1] A. C. Antoulas and D. C. Sorensen, *Approximation of Large-Scale Dynamical Systems: An Overview*, Technical Report CAAM TR01-01, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2001.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, PA, 2000.

[3] Z. Bai and R. Freund, *A partial Padé-via-Lanczos method for reduced-order modeling*, Linear Algebra Appl., 332/334 (2001), pp. 141–166.

[4] Z. Bai and Q. Ye, *Error estimation of the Padé approximation of transfer functions via the Lanczos process*, Electron. Trans. Numer. Anal., 7 (1998), pp. 1–17.

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

[6] J.-P. Coyette, *Evaluation of different computational strategies for acoustic finite element modelling*, in Proceedings of the Institute of Acoustics, Vol. 12, 1990, pp. 851–864.

[7] J.-P. Coyette, J.-L. Migeot, T. Leclerq, G. Lielens, K. Meerbergen, and P. Ploumhans, *ACTRAN Rev.* 1.3*, User's Manual*, Free Field Technologies, Louvain-la-Neuve, Belgium, 2001.

[8] B. N. Datta and Y. Saad, *Arnoldi methods for large Sylvester-like observer matrix equations, and an associated algorithm for partial spectrum assignment*, Linear Algebra Appl., 154/156 (1991), pp. 225–244.

[9] J. J. Dongarra, I. S. Duff, D. C. Sorensen, and H. A. van der Vorst, *Numerical Linear Algebra for High-Performance Computers*, SIAM, Philadelphia, PA, 1998.

[10] 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.

[11] P. Feldman and R. W. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Trans. Computer-Aided Design, CAD-14 (1995), pp. 639–649.

[12] A. Frommer and U. Glässner, *Restarted GMRES for shifted linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 15–26.

[13] K. Gallivan, E. Grimme, and P. Van Dooren, *Reduction and simulation of large-scale dynamical systems with Lanczos methods*, in Proceedings of the IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 1994, pp. 443–448.

[14] K. Gallivan, E. Grimme, and P. Van Dooren, *A rational Lanczos algorithm for model reduction*, Numer. Algorithms, 12 (1996), pp. 33–63.

[15] T. Garratt, G. Moore, and A. Spence, *A generalised Cayley transform for the numerical detection of Hopf bifurcations in large systems*, in Contributions in Numerical Mathematics, R. Agarwal, ed., World Scientific, River Edge, NJ, 1993, pp. 177–185.

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

[17] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, PA, 1997.

[18] R. G. Grimes, J. G. Lewis, and H. D. Simon, *A shifted block Lanczos algorithm for solving sparse symmetric generalized eigenproblems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 228–272.

[19] E. Grimme, D. Sorensen, and P. Van Dooren, *Model reduction of state space systems via an implicitly restarted Lanczos method*, Numer. Algorithms, 12 (1996), pp. 1–31.

[20] E. J. Grimme, *Krylov Projection Methods for Model Reduction*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, IL, 1997.

[21] H. C. Ibrahimbegovic, E. L. Chen, E. L. Wilson, and R. L. Taylor, *Ritz method for dynamic analysis of large discrete linear systems with non-proportional damping*, Earthquake Engineering and Structural Dynamics, 19 (1990), pp. 877–889.

[22] I. M. Jaimoukha and E. M. Kasenally, *Implicitly restarted Krylov subspace methods for stable partial realizations*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 633–652.

[23] M. Kuzuoglu and R. Mittra, *Finite element solution of electromagnetic problems over a wide frequency range via the Padé approximation*, Comput. Methods Appl. Mech. Engrg., 169 (1997), pp. 263–277.

[24] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45 (1950), pp. 255–282.

[25] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 33–53.

[26] M. Malhotra and P. M. Pinsky, *Efficient computation of multi-frequency far-field solutions of the Helmholtz equation using Padé approximation*, J. Comput. Acoust., 8 (2000), pp. 223–240.

[27] K. Meerbergen, *The rational Lanczos method for Hermitian eigenvalue problems*, Numer. Linear Algebra Appl., 8 (2001), pp. 33–52.

[28] B. Nour-Omid, B. Parlett, T. Ericsson, and P. Jensen, *How to implement the spectral transformation*, Math. Comp., 48 (1987), pp. 663–673.

[29] C. Paige, B. Parlett, and H. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.

[30] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[31] B. Parlett, *The Symmetric Eigenvalue Problem*, SIAM Classics in Appl. Math. 20, SIAM, Philadelphia, PA, 1998.

[32] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Algorithms and Architectures for Advanced Scientific Computing, Manchester University Press, Manchester, UK; Halsted Press [John Wiley & Sons, Inc.], New York, 1992.

[33] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.

[34] V. Simoncini, *Linear systems with a quadratic parameter and application to structural dynamics*, in Iterative Methods in Scientific Computation II, D. R. Kincaid and A. Elster, eds., IMACS Series in Computational and Applied Mathematics 5, 1999, pp. 451–461.

[35] V. Simoncini and F. Perotti, *On the numerical solution of* $(\lambda^2 A + \lambda B + C)x = b$ *and application to structural dynamics*, SIAM J. Sci. Comput., 23 (2002), pp. 1875–1897.

[36] D. Skoogh, *Krylov Subspace Methods for Linear Systems, Eigenvalues and Model Reduction*, Ph.D. thesis, Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden, 1998.

[37] D. Skoogh, *A Rational Krylov Method for Model Order Reduction*, Technical Report 1998-47, Department of Mathematics, Chalmers University of Technology and the University of Göteborg, Göteborg, Sweden, 1998.

[38] A. Wathen, B. Fischer, and D. Silvester, *The convergence of iterative solution methods for symmetric and indefinite linear systems*, in Numerical Analysis 1997, D. F. Griffiths and G. A. Watson, eds., Pitman Res. Notes in Math. Ser. 380, Longman, Harlow, UK, 1998, pp. 230–243.

[39] E. L. Wilson, M.-W. Yuan, and J. M. Dickens, *Dynamic analysis by direct superposition of Ritz vectors*, Earthquake Engineering and Structural Dynamics, 10 (1982), pp. 813–821.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1060–1070

FAT DIAGONALS AND FOURIER ANALYSIS*

MILAN HLADNIK†, JOHN HOLBROOK‡, AND STEFANO SERRA CAPIZZANO§

Abstract. We use Fourier analysis and Toeplitz matrices to study the effect on matrix norms obtained by various "truncations" of the matrix. Examples include the truncation of a matrix to a central band of diagonals, truncation to the off-diagonal part, and the upper-triangular truncation. Our tools include tensor product constructions, asymptotic eigenvalue distributions, and multilevel Toeplitz matrices.

Key words. matrix norms, truncation, Fourier analysis, Toeplitz matrices, tensor product, eigenvalue distribution

AMS subject classifications. 15A60, 15A45, 15A18, 47B35

PII. S0895479802411742

1. Introduction.

In a recent article, Bhatia describes links between Fourier series and the effect on matrix norms of operations that modify the matrix entries (see [B2000]). For example, he shows that

$$\|\mathcal{T}_3(M)\| \le \left(\frac{1}{3} + \frac{2\sqrt{3}}{\pi}\right)\|M\|, \tag{1}$$

where $\|M\|$ denotes the operator norm of an $n \times n$ matrix M, and $\mathcal{T}_3(M)$ is the tridiagonal part of M (i.e., all m_{ij} with $|i-j| > 1$ are replaced by zeros). The constant in (1) is in fact L_1, where L_k is the kth *Lebesgue constant*,

$$L_k = \int_{-\pi}^{\pi} \left| \sum_{j=-k}^{k} e^{\mathbf{i}j\theta} \right| \frac{d\theta}{2\pi} \tag{2}$$

($\mathbf{i}$ is our notation for the complex root of -1). Bhatia shows, moreover, that this constant is best possible if inequality (1) holds for all matrix sizes n. His argument is based on the exact knowledge of the eigenvalues for $\mathcal{T}_3(1_n)$, where 1_n denotes the $n \times n$ matrix whose entries are all 1. This method of determining the best possible constant may be extended to "fatter" diagonal parts such as $\mathcal{T}_5(M)$ by using information about the asymptotic eigenvalue distributions of Toeplitz matrices (see [S–T2002], [Ty1996]). In fact, in what follows we shall explore several different approaches that put such estimates in a general context and that show they are all asymptotically best possible.

In [B2000] Bhatia also discusses inequalities governing the upper-triangular part of a matrix. He shows that $\|\Delta_U(M)\| \le L_{n+1}\|M\|$, where $\Delta_U(M)$ denotes the upper-triangular part of the $n \times n$ matrix M, and notes that standard estimates for the

*Received by the editors July 17, 2002; accepted for publication (in revised form) by R. Bhatia September 24, 2002; published electronically February 25, 2003. This work was supported in part by NSERC of Canada and by the Ministry of Education, Science, and Sport of Slovenia.
http://www.siam.org/journals/simax/24-4/41174.html

†Department of Mathematics, Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, SI–1000 Ljubljana, Slovenia (milan.hladnik@fmf.uni-lj.si).

‡Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G 2W1, Canada (jholbroo@uoguelph.ca).

§Dipartimento di Chimica, Fisica e Matematica, Università dell'Insubria – sede di Como, via Valleggio, 11, 22100 Como, Italy (stefano.serrac@uninsubria.it).

Lebesgue constants imply that the growth of the norm $\|\Delta_U\|$, as an operator on the space of $n \times n$ matrices, is bounded, asymptotically, by $4\pi^{-2} \log n$. This is close to the exact asymptotic growth, $\pi^{-1} \log n$ (obtained in [A–C–N1992]; see also [Ha1995]). We shall obtain the exact asymptotic growth in section 3 by combining the results of section 2 with a tensor product construction, neatly explaining the factor $4/\pi$ that links the two estimates.

Let us specify some notation. We deal with complex vectors and matrices, and $\|v\|$ denotes the usual Euclidean norm of a vector v in $\mathbb{C}^n$ (complex n-space). The space of $n \times n$ complex matrices is denoted by $M_n(\mathbb{C})$, and for $M \in M_n(\mathbb{C})$ the operator norm $\|M\|$ is the norm of M as a transformation on $\mathbb{C}^n$, i.e.,

$$\|M\| = \max\{\|Mu\| : u \in \mathbb{C}^n,\ \|u\| = 1\}. \tag{3}$$

Given (square integrable) functions f and g on $[-\pi, \pi]$, the inner product (f, g) is defined by

$$(f, g) = \int_{-\pi}^{\pi} f(\theta)\overline{g(\theta)}\frac{d\theta}{2\pi}. \tag{4}$$

The corresponding norm is $\|f\|_2 = \sqrt{(f, f)}$, and we shall also need the essential supremum norm $\|f\|_\infty$ and the L^1-norm

$$\|f\|_1 = \int_{-\pi}^{\pi} |f(\theta)|\frac{d\theta}{2\pi}. \tag{5}$$

For example, the Lebesgue constant L_k defined by (2) is $\|D_k\|_1$, where D_k is the kth *Dirichlet kernel*

$$D_k(\theta) = \sum_{j=-k}^{k} e^{\mathbf{i}j\theta}. \tag{6}$$

Given an integrable function f on $[-\pi, \pi]$, we denote by $\hat{f}(k)$ the kth Fourier coefficient $(f, e^{\mathbf{i}k\cdot})$. For example, $\widehat{D_k}(j) = 1$ when $|j| \le k$, while the other Fourier coefficients of D_k vanish.

The "fat diagonals" and upper-triangular parts of matrices may be expressed in terms of the Fourier coefficients of appropriate functions. For example, the diagonal of width $2k + 1$ satisfies

$$\mathcal{T}_{2k+1}(M) = [\widehat{D_k}(i - j)m_{ij}]. \tag{7}$$

Thus the fact (proved in section 2) that the inequality

$$\|[\hat{f}(i - j)m_{ij}]\| \le \|f\|_1\|M\| \tag{8}$$

cannot be improved (if it is to apply to $M = [m_{ij}] \in M_n(\mathbb{C})$ for all n) broadens considerably the asymptotic results of [B2000]. The corresponding analysis of $\|\Delta_U\|$ is more subtle (see section 3) since the appropriate function f must change with n.

2. Schur norms and asymptotic results. It is instructive to express our results in terms of Schur products and the corresponding *Schur norm* of a matrix. Given $M, X \in M_n(\mathbb{C})$, the Schur (or Hadamard) product $M \circ X$ is defined elementwise as follows:

$$M \circ X = [m_{ij}x_{ij}]. \tag{9}$$

The Schur norm $\|M\|_S$ of M is the norm of Schur multiplication by M as an operator on $M_n(\mathbb{C})$:

$$\|M\|_S = \max\{\|M \circ X\| : X \in M_n(\mathbb{C}),\ \|X\| = 1\}. \tag{10}$$

For example, the norm of $\mathcal{T}_{2k+1}$ (acting on $M_n(\mathbb{C})$) may be viewed as the Schur norm $\|\mathcal{T}_{2k+1}(1_n)\|_S$, where 1_n is our notation for the $n \times n$ matrix with 1 in every position (not to be confused with the identity matrix I_n).

The study of Schur norms has by now covered a lot of ground. Some useful entry points to this territory may be found, for example, in [A–C–N1992], [Be1977], [B–C–D1989], [B–H2000], [C–D–P1994], [Haa1984], [Ha1995], [Hl1999], [Ho2001], [M1993a], and [M1993b]. Here we are concerned with the Schur norms of Toeplitz matrices.

Given an integrable function f we define the corresponding truncated Toeplitz matrix $T_n(f)$ as that element of $M_n(\mathbb{C})$ having the form

$$[\hat{f}(i-j)] \quad (i,j \in \{1,2,\ldots,n\}).$$

A basic idea in [B2000] (also found in a somewhat different form in [B–D–M1983]) yields the following proposition.

PROPOSITION 1. *For any integrable f defined on $[-\pi,\pi]$, and any positive integer n,*

$$\|T_n(f)\|_S \le \|f\|_1. \tag{11}$$

Proof. Let $U(\theta)$ denote the diagonal matrix whose kth diagonal entry is $e^{\mathbf{i}k\theta}$ $(k = 1,2,\ldots,n)$. Since $U(\theta)$ is unitary, $\|U^*(\theta)XU(\theta)\| = 1$ for any $X \in M_n(\mathbb{C})$ such that $\|X\| = 1$. It follows that

$$\left\| \int_{-\pi}^{\pi} f(\theta)U^*(\theta)XU(\theta)\frac{d\theta}{2\pi} \right\| \le \int_{-\pi}^{\pi} |f(\theta)|\,\|U^*(\theta)XU(\theta)\|\frac{d\theta}{2\pi} = \|f\|_1. \tag{12}$$

Since $U^*(\theta)XU(\theta) = [e^{\mathbf{i}(j-i)\theta}x_{ij}]$,

$$\int_{-\pi}^{\pi} f(\theta)U^*(\theta)XU(\theta)\frac{d\theta}{2\pi} = [\hat{f}(i-j)x_{ij}] = T_n(f) \circ X. \tag{13}$$

Thus (12) says that $\|T_n(f) \circ X\| \le \|f\|_1$ whenever $\|X\| = 1$, i.e., that (11) holds. ∎

Equality in (11) may be far from the truth for fixed n and f. For example, taking $f(\theta) = D_n(\theta) - 1$, (11) tells us that the "off-diagonal part" $OD_n(M)$ of an $n \times n$ matrix M satisfies $\|OD_n(X)\| \le \|f\|_1\|X\|$. Now $\|f\|_1$ is close to the Lebesgue constant L_n, so that it grows like $\log n$, whereas the best inequality governing the off-diagonal part is

$$\|OD_n(X)\| \le 2\left(1 - \frac{1}{n}\right)\|X\| \tag{14}$$

(see [B–C–D1989]). We shall take another look at (14) in section 3. Proposition 3, below, shows that (11) becomes more accurate as $n \to \infty$ with f fixed.

Note that the proof of Proposition 1 may be easily adapted to show that

$$|||T_n(f) \circ X||| \le \|f\|_1|||X||| \tag{15}$$

for any *weakly unitarily invariant* matrix norm $|||\cdot|||$ (i.e., a norm such that $|||U^*XU||| = |||X|||$ for any X and unitary U in $M_n(\mathbb{C})$).

Note also that these considerations extend naturally to the case where $f(\theta)/2\pi$ is replaced by a more general Borel measure μ on $(-\pi, \pi]$. We define the corresponding Fourier coefficients by

$$\hat{\mu}(k) = \int_{-\pi}^{\pi} e^{-\mathbf{i}k\theta}\, \mu(d\theta)$$

and interpret $\|\mu\|_1$ as the total variation of μ. A simple modification of the proof of Proposition 1 also yields

$$\|T_n(\mu)\|_S \le \|\mu\|_1.$$

PROPOSITION 2. *For any integrable f defined on $[-\pi, \pi]$, and any positive integer n,*

$$\|T_n(f)\|_S \le \|T_{n+1}(f)\|_S. \tag{16}$$

Proof. Let $X \in M_n(\mathbb{C})$ have $\|X\| = 1$ and be such that $\|T_n(f)\|_S = \|T_n(f) \circ X\|$. Let $\widetilde{X}$ denote the matrix in $M_{n+1}(\mathbb{C})$ obtained by augmenting X with a final column of zeros, then a final row of zeros. Then $\|\widetilde{X}\| = \|X\| = 1$ and

$$\|T_{n+1}(f) \circ \widetilde{X}\| = \|\widetilde{T_n(f) \circ X}\| = \|T_n(f) \circ X\| = \|T_n(f)\|_S. \tag{17}$$ ∎

Again, the proof of Proposition 2 is easily adapted to give the same result for a measure μ in place of f.

The following proposition shows that (11) is asymptotically sharp as $n \to \infty$. The result is inherent in a theorem of Bennett [Be1977] concerning infinite Toeplitz matrices as Schur multipliers. We shall discuss that approach and some related error estimates. At this point, however, we speak only of finite matrices, and we offer a proof based on standard techniques of Fourier analysis.

PROPOSITION 3. *For any integrable f defined on $[-\pi, \pi]$,*

$$\|T_n(f)\|_S \uparrow_n \|f\|_1 \tag{18}$$

as $n \to \infty$.

Proof. In view of Propositions 1 and 2, it simply remains to find matrices $X_n \in M_n(\mathbb{C})$ such that

$$\frac{\|T_n(f) \circ X_n\|}{\|X_n\|} \to \|f\|_1 \tag{19}$$

as $n \to \infty$. We may take $X_n = T_n(g)$, where $g(\theta) = \operatorname{sign} f(-\theta)$, i.e., $g(\theta) = |f(-\theta)|/f(-\theta)$ (if $f(-\theta) = 0$, set $g(\theta) = 0$). To see this, recall the properties of the convolution product $f \star g$ as follows: In this setting the appropriate definition is

$$f \star g(t) = \int_{-\pi}^{\pi} f(t-s)g(s)\frac{ds}{2\pi}, \tag{20}$$

where the values of f are extended 2π-periodically beyond $[-\pi, \pi]$. It is easy to verify, for any integrable f, g, that $\widehat{f \star g}(k) = \hat{f}(k)\hat{g}(k)$. It follows that $T_n(f) \circ T_n(g) = T_n(f \star g)$ so that in (19) we are dealing with the ratio

$$\frac{\|T_n(f \star g)\|}{\|T_n(g)\|}. \tag{21}$$

Recall that for any essentially bounded function φ

$$\|T_n(\varphi)\| \to_n \|\varphi\|_\infty; \tag{22}$$

Equation (22) may be viewed as a version of the well-known formula for infinite Toeplitz matrices, $\|T(\varphi)\| = \|\varphi\|_\infty$ (see, for example, [D1972, Chap. 7]). Alternatively, a direct approach might be based on standard results on the convergence of Fourier series. Given (22), we need only note that $\|g\|_\infty = \|\operatorname{sign} f(-\theta)\|_\infty = 1$ and that

$$\|f \star g\|_\infty \geq |f \star g(0)| = \left| \int_{-\pi}^{\pi} f(-s)g(s)\frac{ds}{2\pi} \right| = \int_{-\pi}^{\pi} |f(-s)|\frac{ds}{2\pi} = \|f\|_1. \tag{23}$$

Actually, we have equality in (23) since, always, $\|f \star g\|_\infty \leq \|f\|_1 \|g\|_\infty$. □

Turning to a version of Proposition 3 based on Bennett's theorem, we also obtain the natural extension to measures. This extension will be useful in section 3.

PROPOSITION 4. *For any Borel measure μ on $(-\pi, \pi]$ we have*

$$\|T_n(\mu)\|_S \uparrow_n \|\mu\|_1 \tag{24}$$

as $n \to \infty$.

Proof. Bennett's theorem from [Be1977] says that the infinite Toeplitz matrix

$$T(\mu) = [\hat{\mu}(i-j)] \quad (i, j = 1, 2, \dots)$$

acts on $\mathcal{B}(\ell^2)$ as a Schur multiplier with $\|T(\mu)\|_S = \|\mu\|_1$. That is,

$$\|\mu\|_1 = \sup\{\|T(\mu) \circ X\| : \|X\| = 1\},$$

where X denotes an infinite matrix representing a bounded operator in $\mathcal{B}(\ell^2)$. Given $\epsilon > 0$ we have X with $\|X\| = 1$ and $\|T(\mu) \circ X\| > \|\mu\|_1 - \epsilon$. If X_n denotes the truncation of X obtained by extracting the $n \times n$ block, where $i, j \in \{1, 2, \dots, n\}$, we have $X_n \in M_n(\mathbb{C})$ and $\|X_n\| \uparrow_n \|X\|$. Thus, for some n,

$$\|T_n(\mu) \circ X_n\| = \|(T(\mu) \circ X)_n\| > \|\mu\|_1 - \epsilon$$

and, since $\|X_n\| \leq \|X\| = 1$, we have $\|T_n(\mu)\|_S > \|\mu\|_1 - \epsilon$. □

3. Triangular truncation and tensor products. The asymptotic result established in Propositions 3 and 4 may be applied to tensor products to evaluate (or estimate) Schur norms for certain matrices of fixed size. Given an $n \times n$ matrix X and an $m \times m$ matrix Y, we may define the tensor product $X \otimes Y$ by means of the block matrix

$$X \otimes Y = [x_{ij}Y]_{i,j=1,2,\dots,m}. \tag{25}$$

We shall use the fact obtained in [Hl1999] that the Schur norm is multiplicative on tensor products. There the argument is based in part on Haagerup's theorem [Haa1984]

$$\|X\|_S = \min\{\|R\|_r \|C\|_c : RC = X\}.$$

Here $X, R, C \in M_n(\mathbb{C})$, $\|R\|_r$ denotes the row-norm of R, i.e., $\max_i \|r_{i\cdot}\|$, and $\|C\|_c$ denotes the column-norm of C, i.e., $\max_j \|c_{\cdot j}\|$. The proof below is perhaps more direct than that of [Hl1999].

PROPOSITION 5. *The Schur norm respects tensor products, i.e.,* $\|X \otimes Y\|_S = \|X\|_S \|Y\|_S$.

Proof. Let R_1, C_1, R_2, C_2 be matrices (of the appropriate dimensions) such that $R_1 C_1 = X$, $\|X\|_S = \|R_1\|_r \|C_1\|_c$, $R_2 C_2 = Y$, and $\|Y\|_S = \|R_2\|_r \|C_2\|_c$. Direct computation shows that $\|R_1 \otimes R_2\|_r = \|R_1\|_r \|R_2\|_r$ and $\|C_1 \otimes C_2\|_c = \|C_1\|_c \|C_2\|_c$. Thus

$$\|X \otimes Y\|_S = \|(R_1 \otimes R_2)(C_1 \otimes C_2)\|_S \leq \|R_1 \otimes R_2\|_r \|C_1 \otimes C_2\|_c$$

$$= \|R_1\|_r \|R_2\|_r \|C_1\|_c \|C_2\|_c = \|X\|_S \|Y\|_S.$$

On the other hand, if U and V are unitaries such that $\|X \circ U\| = \|X\|_S$ and $\|Y \circ V\| = \|Y\|_S$, then $U \otimes V$ is also unitary and

$$\|X \otimes Y\|_S \geq \|(X \otimes Y) \circ (U \otimes V)\| = \|(X \circ U) \otimes (Y \circ V)\| = \|X\|_S \|Y\|_S. \qquad \square$$

As a "warm-up" and first illustration of the tensor product technique, let us return to relation (14), which we may express in terms of Schur norms as follows.

PROPOSITION 6 (see [B–C–D1989]). *Let* M_n *denote the* $n \times n$ *"off-diagonal" matrix, i.e.,* $M_n = OD_n(1_n)$, *where* 1_n *is the* $n \times n$ *matrix of* 1*'s. Then*

$$\|M_n\|_S = 2\left(1 - \frac{1}{n}\right). \tag{26}$$

Remark. This elegant formula may be justified by several, rather different, arguments. The original approach of Bhatia, Choi, and Davis expresses $OD_n(X)$ as a convex combination of unitary conjugates of X. In [Ho2001] the argument depends on an "explicit Haagerup factorization" of M_n. The technique used here is closely related to that found in [Hl1999].

Proof. Note that $1_m \otimes M_n = T_{nm}(\mu_n)$, where μ_n is the measure with periodic Fourier coefficients as follows: $\hat{\mu_n}(k) = 1$ unless n divides k, and $\hat{\mu_n}(jn) = 0$. Now μ_n is easily identified: If δ_θ denotes the δ-measure (point mass) located at θ, and $\theta_n = 2\pi/n$, we have

$$\mu_n = \delta_0 - \frac{1}{n}\sum_{k=0}^{n-1} \delta_{k\theta_n} = \left(1 - \frac{1}{n}\right)\delta_0 + \left(-\frac{1}{n}\right)\sum_{k=1}^{n-1} \delta_{k\theta_n}. \tag{27}$$

Proposition 4 tells us that

$$\|T_{nm}(\mu_n)\|_S \uparrow_m \|\mu_n\|_1, \tag{28}$$

where $\|\mu_n\|_1$ is the total variation of μ_n, namely $(1-\frac{1}{n})+(n-1)\frac{1}{n} = 2(1-\frac{1}{n})$. Invoking Proposition 5, $\|1_m\|_S \|M_n\|_S = \|T_{nm}(\mu_n)\|_S \to_m 2(1-\frac{1}{n})$, and since $\|1_m\|_S = 1$, we must have (26). $\square$

In the same spirit, we may complete the program outlined in section 1 for finding the asymptotic behavior of upper-triangulation, i.e., we relate the precise asymptotic growth of $\|\Delta_U(1_n)\|_S$ to the Lebesgue constants L_n. Let Δ_n denote $\Delta_U(1_n)$ and let

$$f_n(\theta) = \sum_{k=0}^{n-1} e^{-\mathbf{i}k\theta}. \tag{29}$$

Since $e^{\mathbf{i}[n/2]\theta} f_n(\theta)$ differs from the Dirichlet kernel $D_{[n/2]}$ (see (6)) by at most a few terms, we have

$$\|f_n\|_1 \sim L_{[n/2]} \sim 4\pi^{-2} \log[n/2] \sim 4\pi^{-2} \log n \tag{30}$$

in view of standard estimates for Lebesgue constants (2). On the other hand, the truncated Toeplitz matrix $T_{nm}(f_n)$ has a superdiagonal band of 1's, with width n, and we may view this as m blocks of Δ_n on the diagonal together with *lower* triangular blocks on the superdiagonal; more precisely,

$$T_{nm}(f_n) = I_m \otimes \Delta_n + T_m(e^{-\mathbf{i}\theta}) \otimes (1_n - \Delta_n). \tag{31}$$

We shall use the ingredients above for our proof of the precise asymptotic result from [A–C–N1992].

PROPOSITION 7. *Asymptotically, as* $n \to \infty$, $\|\Delta_n\|_S \sim \frac{1}{\pi} \log n$.

Proof. Rewrite (31) as

$$T_{nm}(f_n) = T_m(1 - e^{-\mathbf{i}\theta}) \otimes \Delta_n + T_m(e^{-\mathbf{i}\theta}) \otimes 1_n, \tag{32}$$

and note that the second summand has Schur norm 1. Invoking Proposition 5, we see that $\|T_{nm}(f_n)\|_S$ and $\|T_m(1 - e^{-\mathbf{i}\theta})\|_S \|\Delta_n\|_S$ differ by at most 1. Let $m \to \infty$ and recall Proposition 3 to see that $\|f_n\|_1$ and $\|1 - e^{-\mathbf{i}\theta}\|_1 \|\Delta_n\|_S$ also differ by at most 1. Since $\|1 - e^{-\mathbf{i}\theta}\|_1 = \frac{4}{\pi}$,

$$4\pi^{-2} \log n \sim \frac{4}{\pi} \|\Delta_n\|_S \tag{33}$$

so that $\|\Delta_n\|_S \sim \frac{1}{\pi} \log n$. ∎

4. Alternate approaches and refinements. The operator norm $\|\cdot\|$ is the Schatten p-norm $\|\cdot\|_p$ corresponding to $p = \infty$,

$$\|T\| = \|T\|_\infty = \max_k s_k,$$

where $s_1, s_2, \ldots, s_n$ are the singular values of $T \in M_n(\mathbb{C})$. This norm is dual to the trace-norm $\|T\|_1 = \sum_{k=1}^n s_k$,

$$\|T\|_\infty = \max\{|\langle T, S\rangle_F| : \|S\|_1 = 1\}$$

and

$$\|S\|_1 = \max\{|\langle T, S\rangle_F| : \|T\|_\infty = 1\},$$

where $\langle T, S\rangle_F$ is the Frobenius inner product $\sum_{i,j} t_{i,j} \overline{s_{i,j}}$ (= trace(S^*T)). This duality, along with the relation $\langle T \circ X, S\rangle_F = \langle T, S \circ \overline{X}\rangle_F$, shows that $\|T\|_S = \|T\|_{S,\infty} = \|T\|_{S,1}$, where $\|T\|_{S,p}$ denotes the Schur norm of T with respect to the Schatten p-norm,

$$\|T\|_{S,p} = \max\{\|T \circ X\|_p : \|X\|_p = 1\}.$$

This alternate way of computing $\|T\|_S$ is sometimes helpful. For example, the matrix 1_n represents n times the orthogonal projection onto the span of $(1, 1, \ldots, 1)^*$ so that $\|1_n\|_1 = n$. Thus

$$\frac{\|T\|_1}{n} = \frac{\|T \circ 1_n\|_1}{\|1_n\|_1} \le \|T\|_{S,1} = \|T\|_S,$$

and to prove Proposition 3 it suffices to show that

$$\frac{\|T_n(f)\|_1}{n} \to \|f\|_1 \quad \text{as } n \to \infty. \tag{34}$$

This, in fact, was the approach of Bhatia in the special case $f(\theta) = e^{-\mathbf{i}\theta} + 1 + e^{\mathbf{i}\theta}$ described in our introduction. A venerable result (which some trace back to Cauchy) tells us precisely the eigenvalues of $T_n(f)$ in that case. Such precise results are not available in general, but much is known about the asymptotic distribution of Toeplitz eigenvalues. For example, (34) follows from the distributional results of [S2002] and [S–T2002], yielding another point of view on Proposition 3.

Suppose that f is a fixed trigonometric polynomial,

$$f(\theta) = \sum_{k=-N}^{N} c_k e^{\mathbf{i}k\theta}.$$

Let $g(z) = \sum_{-N}^{N} c_k z^k$ be the corresponding function of $z = e^{\mathbf{i}k\theta}$ and let $C_n(f)$ denote the circulant matrix $g(Z_n)$, where Z_n is the permutation matrix mapping $e_j \to e_{j+1}$ (interpreting e_{n+1} as e_1). For circulant matrices we have the following precise information:

$$\|C_n(f)\|_S = \frac{\|C_n(f)\|_1}{n} = \frac{1}{n}\sum_{k=1}^{n} |f(k\theta_n)|, \tag{35}$$

where, as before, $\theta_n = 2\pi/n$. These relations were established by Mathias in [M1993b] and also may be proved by extending the argument we have given for Proposition 6; this approach has been described in [Hl1999]. Because f' is bounded, the Riemann sum on the right of (35) will approximate $\|f\|_1$ to within $O(1/n)$. Moreover, $C_n(f)$ and $T_n(f)$ differ only in blocks of size N in the NE and SW corners so that $\operatorname{rank}(C_n(f) - T_n(f)) \le 2N$ and $\|C_n(f) - T_n(f)\|_1 = O(1)$. Thus we also have

$$\frac{\|T_n(f)\|_1}{n} - \|f\|_1 = O(1/n).$$

Recalling that, in general,

$$\frac{\|T_n(f)\|_1}{n} \le \|T_n(f)\|_{S,1} = \|T_n(f)\|_S \le \|f\|_1,$$

we obtain one example of a quantitative version of Proposition 3.

PROPOSITION 8. *For any fixed trigonometric polynomial f,*

$$\|T_n(f)\|_S \uparrow_n \|f\|_1 \quad \textit{with} \quad \|f\|_1 - \|T_n(f)\|_S = O(1/n).$$

We remark that such results may be extended to classes of sufficiently smooth functions and that the constants hidden in the "big O" can be computed explicitly. In this respect we mention that if f is real-valued and even, then a better bound for the constants involved can be obtained using the τ algebra (i.e., the one generated by $T_n(2\cos(\theta))$ and introduced by Bini and Capovani [B–C1983]).

5. Multilevel extensions. Given a vector $n = (n_1, \dots, n_d)$ of positive integers n_j, $j = 1, \dots, d$, a matrix X, indexed by n (and having dimension $\hat{n} = n_1 \cdots n_d$), is called a d-level matrix. This means that it can be viewed as an $n_1 \times n_1$ block matrix whose blocks are $(d-1)$-level matrices. When $d = 1$ we have a standard matrix. For $d = 2$ a typical 2-level matrix is the discrete Laplacian L, which can be written as

$$L = I_{n_1} \otimes (2I_{n_2} - J_1(n_2) - J_{-1}(n_2)) + (2I_{n_1} - J_1(n_1) - J_{-1}(n_1)) \otimes I_{n_2},$$

where $J_q(m)$ is the $m \times m$ Toeplitz matrix having 1 on the qth diagonal and 0 otherwise (with $1 - m \le q \le m - 1$). It is interesting to notice that L is also the 2-level Toeplitz matrix generated by $4 - 2\cos(\theta_1) - 2\cos(\theta_2)$: in general, a 2-level Toeplitz matrix X indexed by $n = (n_1, n_2)$ and generated by a 2-variate symbol $f(\theta_1, \theta_2)$ is such that

$$X_{(i_1,j_1),(i_2,j_2)} = \hat{f}(i_1 - j_1, i_2 - j_2)$$

with

$$\hat{f}(s,t) = \frac{1}{4\pi^2} \int_Q f(\theta_1, \theta_2) e^{-\mathbf{i}(s\theta_1 + t\theta_2)}\, d\theta_1\, d\theta_2, \quad Q = (-\pi, \pi)^2.$$

Here the 2-index notation $X_{(i_1,j_1),(i_2,j_2)}$ means that we are selecting the block (i_1, j_1) of size $n_2 \times n_2$, and in this block we are selecting the entry (i_2, j_2).

Some of the results of earlier sections have natural extensions to the multilevel setting. For notational simplicity, let us consider the 2-level case (the d-level case is much the same), showing how multivariate Fourier analysis comes into play.

Let $\mathcal{I} \subset \{1 - n_1, \dots, -1, 0, 1, \dots, n_1 - 1\} \times \{1 - n_2, \dots, -1, 0, 1, \dots, n_2 - 1\}$ be a set of pairs of indices and define $\Delta_{\mathcal{I}}^{(n)} : \mathbf{M}(\hat{n}) \to \mathbf{M}(\hat{n})$ as the operator that maps any complex-valued block matrix of size n_1 with blocks of size n_2 into the sum of its "2-level diagonals" indexed by $\mathcal{I}$. More precisely, we define

$$\Delta_{\mathcal{I}}^{(n)}(X) = \sum_{(q_1,q_2) \in \mathcal{I}} J_{(q_1,q_2)} \circ X,$$

where, as before, $\circ$ denotes the componentwise Schur or Hadamard product and

$$J_{(q_1,q_2)} = J_{q_1} \otimes J_{q_2}.$$

Following the analysis of Bhatia, for any weakly unitarily invariant (w.u.i.) norm $|||\cdot|||$ we have

$$|||\Delta_{\mathcal{I}}^{(n)}(X)||| \le \|f_{\mathcal{I}}\|_{L^1} |||X|||, \tag{36}$$

where $f_{\mathcal{I}} = \sum_{(q_1,q_2) \in \mathcal{I}} e^{-\mathbf{i}(q_1\theta_1 + q_2\theta_2)}$, $(\theta_1, \theta_2) \in Q$.

PROPOSITION 9. *For any set of indices $\mathcal{I}$ independent of $n = (n_1, n_2)$, the quantity $\|f_{\mathcal{I}}\|_{L^1}$ is the best constant satisfying* (36) *for any w.u.i. norm and for any dimension $n = (n_1, n_2) \ge (1, 1)$.*

Proof. Let $E = 1_{\hat{n}}$ ($\hat{n} = n_1 n_2$) be the $\hat{n} \times \hat{n}$ matrix of all ones. Then a simple calculation shows that

$$\Delta_{\mathcal{I}}^{(n)}(E) = T_n(f_{\mathcal{I}}).$$

For any $f \in L^1(Q)$, we know from [S2002], [S–T2002] that

$$\lim_{n \to \infty} \frac{\|T_n(f)\|_1}{\hat{n}} = \|f\|_{L^1}, \quad \frac{\|T_n(f)\|_1}{\hat{n}} \le \|f\|_{L^1}$$

with $\|X\|_1 = \sum_{j=1}^{\hat{n}} s_j(X)$ (the Schatten 1-norm). Therefore we have

$$\|\Delta_{\mathcal{I}}^{(n)}(E)\|_1 = \|T_n(f_{\mathcal{I}})\|_1 = \hat{n} \cdot \|f_{\mathcal{I}}\|_{L^1}(1-\epsilon_n)$$

with $\epsilon_n > 0$ and $\epsilon_n \to 0$ when both n_1 and n_2 go to infinity. Since $\|E\|_1 = \hat{n}$ and $\|\cdot\|_1$ is a w.u.i. norm we deduce

$$\|\Delta_{\mathcal{I}}^{(n)}(E)\|_1 = \|f_{\mathcal{I}}\|_{L^1}\|E\|_1(1-\epsilon_n), \tag{37}$$

which proves the asymptotic sharpness of (36). □

Note that by duality (see section 4) we obtain the same asymptotic results with respect to the operator norm.

In order to give a precise bound on the error term ϵ_n it is enough to use the same reasoning as in the unilevel case: instead of the standard circulant algebra [Da1979] we use the 2-level circulant algebra generated by $I_{n_1} \otimes Z(n_2) + Z(n_1) \otimes I_{n_2}$, where $Z(m) = J_1 + J_{1-m}$. By invoking the same low rank correction trick as in the 1-level case, and by using the precision of bivariate Riemann sums for smooth arguments, the following result is obtained.

PROPOSITION 10. *The quantity ϵ_n in* (37) *satisfies $\epsilon_n = O(n_1^{-1} + n_2^{-1})$.*

In the case where $\mathcal{I}$ is symmetric with respect to 0 (so that $f_{\mathcal{I}}$ is real-valued and even with respect to both variables), a slightly better bound on the constant hidden in the "big O" can be obtained by using the embedding argument in the two-level τ algebra [B–C1983] generated by $I_{n_1} \otimes (J_1(n_2) + J_{-1}(n_2)) + (J_1(n_1) + J_{-1}(n_1)) \otimes I_{n_2}$.

Acknowledgments. The authors thank the organizers and participants of the Linear Algebra Workshop held in 1999 at Bled, Slovenia, where some of these ideas were developed. Chi-Kwong Li's suggestions were especially helpful.

REFERENCES

[A–C–N1992] J. R. ANGELOS, C. C. COWEN, AND S. K. NARAYAN, *Triangular truncation and finding the norm of a Hadamard multiplier*, Linear Algebra Appl., 170 (1992), pp. 117–135.

[Be1977] G. BENNETT, *Schur multipliers*, Duke Math. J., 44 (1977), pp. 603–639.

[B2000] R. BHATIA, *Pinching, trimming, truncating, and averaging of matrices*, Amer. Math. Monthly, 107 (2000), pp. 602–608.

[B–C–D1989] R. BHATIA, M.-D. CHOI, AND CH. DAVIS, *Comparing a matrix to its off-diagonal part*, in The Gohberg Anniversary Collection, Vol. I, Oper. Theory Adv. Appl. 40, Birkhäuser, Basel, 1989, pp. 151–164.

[B–D–M1983] R. BHATIA, CH. DAVIS, AND A. MCINTOSH, *Perturbation of spectral subspaces and solution of linear operator equations*, Linear Algebra Appl., 52/53 (1983), pp. 45–67.

[B–H2000] R. BHATIA AND J. HOLBROOK, *Fréchet derivatives of the power function*, Indiana Univ. Math. J., 49 (2000), pp. 1155–1173.

[B–C1983] D. BINI AND M. CAPOVANI, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52/53 (1983), pp. 99–125.

[C–D–P1994] C. C. COWEN, M. A. DRITSCHEL, AND R. C. PENNEY, *Norms of Hadamard multipliers*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 313–320.

[Da1979] P. DAVIS, *Circulant Matrices*, John Wiley and Sons, New York, 1979.

[D1972] R. G. DOUGLAS, *Banach Algebra Techniques in Operator Theory*, Academic Press, New York, 1972.

[Haa1984] U. HAAGERUP, *Decompositions of Completely Bounded Maps on Operator Algebras*, preprint, 1984.

[Ha1995] D. HADWIN, *Triangular truncation and normal limits of nilpotent operators*, Proc. Amer. Math. Soc., 123 (1995), pp. 1741–1745.

[Hl1999] M. Hladnik, *Schur norms of bicirculant matrices*, Linear Algebra Appl., 286 (1999), pp. 261–272.

[Ho2001] J. Holbrook, *Schur norms and the multivariate von Neumann inequality*, in Recent Advances in Operator Theory and Related Topics, Oper. Theory Adv. Appl. 127, Birkhäuser, Basel, 2001, pp. 375–386.

[M1993a] R. Mathias, *Matrix completions, norms, and Hadamard products*, Proc. Amer. Math. Soc., 117 (1993), pp. 905–918.

[M1993b] R. Mathias, *The Hadamard operator norm of a circulant and applications*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1152–1167.

[S2002] S. Serra Capizzano, *Test functions, growth conditions and Toeplitz matrices*, in Proc. Conf. on Functional Analysis and Approximation Theory (Maratea, Italy, 2000), Rend. Circ. Mat. Palermo (2), Vol. 68, Circ. Mat. Palermo, Palermo, Italy, 2002, pp. 791–795.

[S–T2002] S. Serra Capizzano and P. Tilli, *On unitarily invariant norms of matrix-valued linear positive operators*, J. Inequal. Appl., 7 (2002), pp. 309–330.

[Ty1996] E. Tyrtyshnikov, *A unifying approach on some old and new results on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1071–1109

SPECTRAL ANALYSIS OF THE TRANSITION OPERATOR AND ITS APPLICATIONS TO SMOOTHNESS ANALYSIS OF WAVELETS*

RONG-QING JIA† AND QINGTANG JIANG‡

Abstract. This paper investigates spectral properties of the transition operator associated to a multivariate vector refinement equation and their applications to the study of smoothness of the corresponding refinable vector of functions.

Let $\Phi = (\phi_1, \ldots, \phi_r)^T$ be an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R}^s)$ satisfying $\Phi = \sum_{\alpha \in \mathbb{Z}^s} a(\alpha)\Phi(M\cdot - \alpha)$, where M is an expansive integer matrix. The smoothness of Φ is measured by the Sobolev critical exponent $\lambda(\Phi) := \sup\Big\{\lambda : \int_{\mathbb{R}^s} |\hat{\phi}_j(\xi)|^2(1+|\xi|^\lambda)^2\, d\xi < \infty, 1 \le j \le r\Big\}$. Suppose M is similar to $\mathrm{diag}(\sigma_1, \ldots, \sigma_s)$ with $|\sigma_1| = \cdots = |\sigma_s|$ and $\mathrm{supp}\, a := \{\alpha \in \mathbb{Z}^s : a(\alpha) \ne 0\}$ is finite. For $\mu = (\mu_1, \ldots, \mu_s) \in \mathbb{N}_0^s$, define $\sigma^{-\mu} := \sigma_1^{-\mu_1} \cdots \sigma_s^{-\mu_s}$. Let $A := \sum_{\alpha \in \mathbb{Z}^s} a(\alpha)/|\det M|$ and $b(\alpha) := \sum_{\beta \in \mathbb{Z}^s} \overline{a(\beta)} \otimes a(\alpha + \beta)/|\det M|$, $\alpha \in \mathbb{Z}^s$, where $\otimes$ denotes the (right) Kronecker product. Suppose that the highest total degree of polynomials reproduced by Φ is $k-1$ and $\mathrm{spec}\,(A)$ (the spectrum of A) is $\{\eta_1, \eta_2, \ldots, \eta_r\}$ with $\eta_1 = 1$ and $\eta_j \ne 1, 2 \le j \le r$. Set

$$E_k := \{\eta_j \overline{\sigma^{-\mu}}, \overline{\eta_j}\sigma^{-\mu} : |\mu| < k, j = 2, \ldots, r\} \cup \{\sigma^{-\mu} : |\mu| < 2k\}.$$

The main result of this paper asserts that if Φ is stable, then $\lambda(\Phi) = -(\log_{|\det M|} \rho_k)\, s/2$, where

$$\rho_k := \max\left\{ |\nu| : \nu \in \mathrm{spec}\Big(b(M\alpha - \beta)\Big)_{\alpha,\beta \in K} \setminus E_k \right\},$$

and K is the set $\mathbb{Z}^s \cap \sum_{n=1}^{\infty} M^{-n}(\mathrm{supp}\, b)$. This result is obtained through an extensive use of linear algebra and matrix theory. Three examples are provided to illustrate the general theory.

Key words. refinement equations, wavelets, subdivision operators, transition operators, polynomial reproducibility, spectral analysis, smoothness analysis

AMS subject classifications. 42C40, 39B72, 15A18, 41A25

PII. S0895479801397858

1. Introduction.

The purpose of this paper is to investigate spectral properties of the transition operator associated with a multivariate vector refinement equation and their applications to the study of smoothness of the corresponding refinable vector of functions. This study is important in applications of wavelets to image processing, computer aided geometric design, and numerical solutions to partial differential equations.

Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{R}^s$ the s-dimensional Euclidean space. An element of $\mathbb{R}^s$ is also viewed as an $r \times 1$ vector of real numbers. The inner product of two vectors x and y in $\mathbb{R}^s$ is denoted by $x \cdot y$. The norm of x is $|x| := \sqrt{x \cdot x}$.

Let f be a (Lebesgue) measurable function from $\mathbb{R}^s$ to $\mathbb{C}$, where $\mathbb{C}$ denotes the set of complex numbers. For $1 \le p < \infty$, let

$$\|f\|_p := \left(\int_{\mathbb{R}^s} |f(x)|^p\, dx\right)^{1/p}.$$

*Received by the editors November 9, 2001; accepted for publication (in revised form) by M. Hanke August 8, 2002; published electronically March 13, 2003.

http://www.siam.org/journals/simax/24-4/39785.html

†Department of Mathematical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2G1 (jia@xihu.math.ualberta.ca). This author's research was supported in part by NSERC Canada under grant OGP 121336.

‡Department of Mathematics, West Virginia University, Morgantown, WV 26506. Current address: Department of Mathematics and Computer Science, University of Missouri-St. Louis, St. Louis, MO 63121 (jiang@math.umsl.edu).

For $p = \infty$, let $\|f\|_\infty$ be the essential supremum of $|f|$ on $\mathbb{R}^s$. By $L_p(\mathbb{R}^s)$ we denote the Banach space of all measurable functions f such that $\|f\|_p < \infty$. A function f is said to be integrable if f lies in $L_1(\mathbb{R}^s)$.

The Fourier transform of a function $f \in L_1(\mathbb{R}^s)$ is defined by

$$\hat{f}(\xi) := \int_{\mathbb{R}^s} f(x) e^{-ix\cdot\xi}\, dx, \qquad \xi \in \mathbb{R}^s,$$

where i denotes the imaginary unit. The domain of the Fourier transform can be naturally extended to $L_2(\mathbb{R}^s)$.

Let $\mathbb{N}$ denote the set of positive integers and $\mathbb{N}_0$ the set of nonnegative integers. An s-tuple $\mu = (\mu_1, \ldots, \mu_s) \in \mathbb{N}_0^s$ is called a *multi-index*. The length of μ is $|\mu| := \mu_1 + \cdots + \mu_s$, and the factorial of μ is $\mu! := \mu_1! \cdots \mu_s!$. For $\mu, \nu \in \mathbb{N}_0^s$, $\nu \le \mu$ means $\nu_j \le \mu_j$, $j = 1, \ldots, s$. If $\nu \le \mu$ and $\nu \ne \mu$, we write $\nu < \mu$. For $\nu \le \mu$, we define

$$\binom{\mu}{\nu} := \frac{\mu!}{\nu!(\mu - \nu)!}.$$

For $\mu = (\mu_1, \ldots, \mu_s) \in \mathbb{N}_0^s$ and $x = (x_1, \ldots, x_s) \in \mathbb{R}^s$, define

$$x^\mu := x_1^{\mu_1} \cdots x_s^{\mu_s}.$$

The function $x \mapsto x^\mu$ ($x \in \mathbb{R}^s$) is called a monomial and its (total) degree is $|\mu|$. A polynomial is a linear combination of monomials. The degree of a polynomial $q = \sum_\mu c_\mu x^\mu$ is defined to be $\deg q := \max\{|\mu| : c_\mu \ne 0\}$. For $k \in \mathbb{N}_0$, we use Π_k to denote the linear space of all polynomials of degree at most k.

Let $\mathbb{Z}$ denote the set of integers. By $\ell(\mathbb{Z}^s)$ we denote the linear space of all sequences on $\mathbb{Z}^s$. A sequence a on $\mathbb{Z}^s$ is said to be finitely supported if $a(\alpha) \ne 0$ only for finitely many α. Let $\ell_0(\mathbb{Z}^s)$ denote the linear space of all finitely supported sequences on $\mathbb{Z}^s$. Let $u \in \ell(\mathbb{Z}^s)$. For $1 \le p < \infty$, we define

$$\|u\|_p := \left(\sum_{\alpha \in \mathbb{Z}^s} |u(\alpha)|^p \right)^{1/p}.$$

For $p = \infty$, define $\|u\|_\infty$ to be the supremum of $|u|$ on $\mathbb{Z}$. For $1 \le p \le \infty$, let $\ell_p(\mathbb{Z}^s)$ denote the Banach space of all sequences u for which $\|u\|_p < \infty$.

For positive integers m and n, by $\mathbb{C}^{m\times n}$ we denote the collection of all $m \times n$ complex matrices. The transpose of a matrix A is denoted by A^T. When $n = 1$, $\mathbb{C}^{m\times 1}$ is abbreviated as $\mathbb{C}^m$. The linear span of a set E of vectors is denoted by $\operatorname{span}(E)$.

We use $\ell(\mathbb{Z}^s \to \mathbb{C}^{m\times n})$ to denote the linear space of all sequences of $m\times n$ matrices. Similarly, we use $\ell_0(\mathbb{Z}^s \to \mathbb{C}^{m\times n})$ to denote the linear space of all finitely supported sequences of $m \times n$ matrices. For simplicity, $\ell(\mathbb{Z}^s \to \mathbb{C}^{m\times n})$ and $\ell_0(\mathbb{Z}^s \to \mathbb{C}^{m\times n})$ will be abbreviated as $\ell^{m\times n}(\mathbb{Z}^s)$ and $\ell_0^{m\times n}(\mathbb{Z}^s)$, respectively. When $n = 1$, $\ell^{m\times 1}(\mathbb{Z}^s)$ and $\ell_0^{m\times 1}(\mathbb{Z}^s)$ will be further abbreviated as $\ell^m(\mathbb{Z}^s)$ and $\ell_0^m(\mathbb{Z}^s)$, respectively. For a subset $K \subseteq \mathbb{Z}^s$, $\ell^{m\times n}(K)$ denotes the linear space of those elements $u \in \ell^{m\times n}(\mathbb{Z}^s)$ for which $u(\alpha) = 0$ for all $\alpha \in \mathbb{Z}^s \setminus K$.

The *symbol* of an element $v \in \ell_0(\mathbb{Z}^s)$, denoted $\hat{v}$, is the trigonometric polynomial given by

$$\hat{v}(\xi) := \sum_{\alpha \in \mathbb{Z}^s} v(\alpha) e^{-i\alpha\cdot\xi}, \quad \xi \in \mathbb{R}^s.$$

The symbol of an element in $\ell_0^{m\times n}(\mathbb{Z}^s)$ is defined accordingly.

By $\mathbb{T}(\mathbb{R}^s)$ we denote the set of all trigonometric polynomials on $\mathbb{R}^s$. Accordingly, by $\mathbb{T}^{m\times n}(\mathbb{R}^s)$ we denote the set of all $m\times n$ matrices of trigonometric polynomials on $\mathbb{R}^s$.

The *spectrum* of a square matrix A is denoted by $\operatorname{spec}(A)$ and it is understood to be the multiset of its eigenvalues. In other words, multiplicities of eigenvalues are counted in the spectrum. The multiset of nonzero eigenvalues of a square matrix A is denoted by $\operatorname{spec}'(A)$. By $\rho(A)$ we denote the spectral radius of A. Clearly, if $\operatorname{spec}'(A)$ is not empty,

$$\rho(A) = \max\{|\nu| : \nu \in \operatorname{spec}(A)\} = \max\{|\nu| : \nu \in \operatorname{spec}'(A)\}.$$

Let M be an $s\times s$ integer matrix. We assume that M is *expansive*, i.e., all the eigenvalues of M are greater than 1 in modulus.

An $r\times 1$ vector $\Phi = (\phi_1,\dots,\phi_r)^T$ of compactly supported functions in $L_p(\mathbb{R}^s)$ is said to be M-refinable if Φ satisfies the vector refinement equation

$$\Phi = \sum_{\alpha\in\mathbb{Z}^s} a(\alpha)\Phi(M\cdot - \alpha), \tag{1.1}$$

where $a \in \ell_0^{r\times r}(\mathbb{Z}^s)$. We call a the (refinement) *mask*. Taking Fourier transform of both sides of (1.1), we obtain

$$\hat\Phi(\xi) = A\big((M^T)^{-1}\xi\big)\hat\Phi\big((M^T)^{-1}\xi\big), \quad \xi\in\mathbb{R}^s, \tag{1.2}$$

where

$$A(\xi) := \frac{1}{|\det M|}\sum_{\alpha\in\mathbb{Z}^s} a(\alpha)e^{-i\alpha\cdot\xi}. \tag{1.3}$$

It follows from (1.2) that $\hat\Phi(0) = A(0)\hat\Phi(0)$, where

$$A(0) = \frac{1}{d}\sum_{\alpha\in\mathbb{Z}^s} a(\alpha) \quad\text{and}\quad d := |\det M|. \tag{1.4}$$

Our goal is to determine the smoothness of Φ in the L_2 norm strictly in terms of the mask a. For $\lambda\ge 0$, we denote by $W_2^\lambda(\mathbb{R}^s)$ the Sobolev space of all functions $f\in L_2(\mathbb{R}^s)$ such that

$$\int_{\mathbb{R}^s} |\hat f(\xi)|^2(1+|\xi|^\lambda)^2\,d\xi < \infty.$$

The smoothness of $\Phi = (\phi_1,\dots,\phi_r)^T$ is measured by the critical exponent $\lambda(\Phi)$, which is defined by

$$\lambda(\Phi) := \sup\big\{\lambda : \phi_j \in W_2^\lambda(\mathbb{R}^s) \text{ for all } j = 1,\dots,r\big\}.$$

The smoothness of refinable functions is an important issue in all multiresolution analyses and has a strong impact on applications of wavelets to image processing and geometric modeling, e.g., subdivision schemes.

The smoothness order of refinable functions has been studied extensively. For the scalar case ($r = 1$), a characterization of the critical exponent of a refinable function

in terms of the corresponding mask was given in [12], [45], and [5]. In particular, it was shown that the critical exponent of a refinable function could be calculated in terms of the spectral radius of a transition matrix associated with the mask.

The aforementioned results rely on factorization of the symbol of the mask. In the multivariate case $s > 1$, however, the symbol of the refinement mask is often irreducible. This difficulty was overcome in [21] by considering certain invariant subspaces of the transition operator associated with the mask. Based on the characterization of smoothness of multivariate refinable functions given in [21], a useful algorithm for calculation of the critical exponent was given in [29]. These results are valid when the matrix M is isotropic. In the case when M is anisotropic, smoothness of multivariate refinable functions was investigated in [7].

For the vector case ($r > 1$), smoothness of univariate refinable vectors of functions was studied in [6] and [36] on the basis of a factorization technique. A different approach was employed in [28] to give the optimal smoothness of refinable vectors of functions. Smoothness of multivariate refinable vectors of functions was analyzed in [30] and [31]. Also, see [41] and [26] for a recent study of the Sobolev regularity of refinable functions without the requirement of stability.

The study of smoothness of Φ is related to properties of shift-invariant spaces. Suppose $\Phi = (\phi_1, \dots, \phi_r)^T$ is an $r \times 1$ vector of compactly supported functions in $L_p(\mathbb{R}^s)$. We use $\mathbb{S}(\Phi)$ to denote the shift-invariant space generated from Φ, which is the linear space of functions of the form

$$\sum_{j=1}^{r} \sum_{\alpha \in \mathbb{Z}^s} u_j(\alpha) \phi_j(\cdot - \alpha),$$

where $u_1, \dots, u_r \in \ell(\mathbb{Z}^s)$. The (multi-integer) shifts of $\phi_1, \dots, \phi_r$ are said to be stable if there exist two positive constants C_1 and C_2 such that the inequalities

$$C_1 \left(\sum_{j=1}^{r} \|u_j\|_p \right) \le \left\| \sum_{j=1}^{r} \sum_{\alpha \in \mathbb{Z}^s} u_j(\alpha) \phi_j(\cdot - \alpha) \right\|_p \le C_2 \left(\sum_{j=1}^{r} \|u_j\|_p \right)$$

are valid for all $u_1, \dots, u_r \in \ell_p(\mathbb{R}^s)$. If this is the case, we simply say that Φ is stable. It was proved in [27] and [19] that the shifts of $\phi_1, \dots, \phi_r$ are stable if and only if, for every $\xi \in \mathbb{R}^s$,

$$\operatorname{span} \{ \hat{\Phi}(\xi + 2\pi\beta) : \beta \in \mathbb{Z}^s \} = \mathbb{C}^r.$$

The Kronecker product of two matrices is a useful tool in our study of vector refinement equations. Let us recall some basic properties of the Kronecker product. Suppose $A = (a_{ij})_{1 \le i \le m, 1 \le j \le n}$ and $B = (b_{ij})_{1 \le i \le r, 1 \le j \le s}$ are two matrices. The (right) Kronecker product of A and B, written $A \otimes B$, is defined to be the block matrix

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}.$$

For three matrices A, B, and C of the same type, we have

$$(A + B) \otimes C = (A \otimes C) + (B \otimes C) \quad \text{and} \quad A \otimes (B + C) = (A \otimes B) + (A \otimes C).$$

If A, B, C, D are four matrices such that the products AC and BD are well defined, then

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

Moreover, if $\lambda_1, \ldots, \lambda_r$ are the eigenvalues of an $r \times r$ matrix A and $\mu_1, \ldots, \mu_s$ are the eigenvalues of an $s \times s$ matrix B, then the eigenvalues of the Kronecker product $A \otimes B$ are $\lambda_m \mu_n$, $m = 1, \ldots, r, n = 1, \ldots, s$. See [34, Chap. 12] for a proof of these results.

For a matrix $A = (a_{ij})_{1 \le i \le m, 1 \le j \le n}$, the vector

$$(a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, \ldots, a_{1n}, \ldots, a_{mn})^T$$

is said to be the vec-function of A and is written as $\operatorname{vec} A$. If A, X, and B are three matrices such that the product AXB is well defined, then

$$\operatorname{vec}(AXB) = (B^T \otimes A)\operatorname{vec} X. \tag{1.5}$$

For two functions f, g in $L_2(\mathbb{R}^s)$, $f \odot g$ is defined as follows:

$$f \odot g(x) := \int_{\mathbb{R}^s} f(x+y)\overline{g(y)}\, dy, \qquad x \in \mathbb{R}^s,$$

where $\overline{g(y)}$ stands for the complex conjugate of $g(y)$. In other words, $f \odot g$ is the convolution of f with the function $y \mapsto \overline{g(-y)}$, $y \in \mathbb{R}^s$. It is easily seen that $f \odot g$ lies in $C_0(\mathbb{R}^s)$, the space of continuous functions on $\mathbb{R}$ which vanish at ∞. In particular, $f \odot g$ is uniformly continuous.

Suppose $\Phi = (\phi_1, \ldots, \phi_r)^T$ is an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R}^s)$ satisfying the refinement equation (1.1). Let

$$\Phi \odot \Phi^T := \begin{bmatrix} \phi_1 \odot \phi_1 & \phi_1 \odot \phi_2 & \cdots & \phi_1 \odot \phi_r \\ \phi_2 \odot \phi_1 & \phi_2 \odot \phi_2 & \cdots & \phi_2 \odot \phi_r \\ \vdots & \vdots & \ddots & \vdots \\ \phi_r \odot \phi_1 & \phi_r \odot \phi_2 & \cdots & \phi_r \odot \phi_r \end{bmatrix}.$$

It follows from (1.1) that

$$\Phi \odot \Phi^T = \sum_{\alpha \in \mathbb{Z}^s} \sum_{\beta \in \mathbb{Z}^s} a(\alpha)\Phi(M\cdot - \alpha) \odot \Phi^T(M\cdot - \beta)\overline{a(\beta)}^T.$$

Let $F := \operatorname{vec}(\Phi \odot \Phi^T)$. With the help of (1.5) we obtain

$$F = \sum_{\alpha \in \mathbb{Z}^s} b(\alpha) F(M\cdot - \alpha),$$

where $b \in \ell_0^{r^2 \times r^2}(\mathbb{Z}^s)$ is given by

$$b(\alpha) := \frac{1}{d} \sum_{\beta \in \mathbb{Z}^s} \overline{a(\beta)} \otimes a(\alpha + \beta), \quad \alpha \in \mathbb{Z}^s. \tag{1.6}$$

For a bounded subset H of $\mathbb{R}^s$, the set $\sum_{n=1}^{\infty} M^{-n}H$ is defined as

$$\left\{ \sum_{n=1}^{\infty} M^{-n} h_n \,:\, h_n \in H \ \text{ for } \ n = 1, 2, \ldots \right\}.$$

If H is a compact set, then $\sum_{n=1}^{\infty} M^{-n}H$ is also compact. By $\operatorname{supp} b$ we denote the set $\{\alpha \in \mathbb{Z}^s : b(\alpha) \neq 0\}$. Let

$$K := \left(\sum_{n=1}^{\infty} M^{-n}(\operatorname{supp} b)\right) \cap \mathbb{Z}^s.$$

We assume that M is isotropic, i.e., M is similar to a diagonal matrix $\operatorname{diag}(\sigma_1, \ldots, \sigma_s)$ with $|\sigma_1| = \cdots = |\sigma_s|$. For $\mu = (\mu_1, \ldots, \mu_s) \in \mathbb{Z}^s$, define

$$\sigma^\mu := \sigma_1^{\mu_1} \cdots \sigma_s^{\mu_s}.$$

Suppose $r = 1$ and that Φ is stable. Let k be the largest integer such that $\mathbb{S}(\Phi) \supset \Pi_{k-1}$. It was proved in [29] that $\lambda(\Phi) = -(\log_d \rho_k)\, s/2$, where

$$\rho_k := \max\Big\{ |\nu| : \nu \in \operatorname{spec}\big(b(M\alpha - \beta)\big)_{\alpha,\beta \in K} \setminus \{\sigma^{-\mu} : |\mu| < 2k\} \Big\}.$$

A straightforward generalization of this result to the vector case ($r > 1$) does not work. See section 8 for a counterexample. In fact, in the vector case, a correct formula for $\lambda(\Phi)$ must involve the spectrum of the $r \times r$ matrix $A(0)$ given in (1.4). Suppose $\operatorname{spec}(A(0)) = \{\eta_1, \eta_2, \ldots, \eta_r\}$. We assume that $\eta_1 = 1$ and $\eta_j \neq 1$ for $j = 2, \ldots, r$. The following theorem is the main result of this paper.

THEOREM 1.1. *Let Φ be an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R}^s)$. Suppose Φ satisfies the refinement equation* (1.1) *with mask a. Let k be the largest integer such that $\mathbb{S}(\Phi) \supset \Pi_{k-1}$. Let*

$$E_k := \{\eta_j \overline{\sigma^{-\mu}}, \overline{\eta_j}\sigma^{-\mu} : |\mu| < k, j = 2, \ldots, r\} \cup \{\sigma^{-\mu} : |\mu| < 2k\}.$$

If, in addition, Φ is stable, then

$$\lambda(\Phi) = -\big(\log_d \rho_k\big)\, s/2,$$

where

$$\rho_k := \max\Big\{ |\nu| : \nu \in \operatorname{spec}\big(b(M\alpha - \beta)\big)_{\alpha,\beta \in K} \setminus E_k \Big\}.$$

Here is an outline of the paper. Section 2 is devoted to a study of subdivision and transition operators. The fact that the subdivision operator is the algebraic adjoint of the transition operator will be employed to derive useful spectral properties of these linear operators. In section 3 we will review polynomial reproducibility of refinable vectors of functions and introduce certain invariant subspaces of the subdivision and transition operators, which will be needed in the smoothness analysis of refinable functions. In section 4 we will give a characterization of the smoothness order of a refinable vector Φ of functions in terms of the corresponding mask a. This characterization is difficult to implement. Thus, in section 5, we will give a formula for the critical exponent of Φ in terms of the spectral radius of the transition operator T_b restricted to a certain invariant subspace, where b is obtained from a by (1.6). In order to calculate this spectral radius, we will carefully analyze the relevant invariant subspaces and spectra of the subdivision operator and the transition operator in sections 6 and 7. This analysis enables us to prove Theorem 1.1 and other related results. Finally, in section 8, we will provide three examples to illustrate the general

theory. These examples demonstrate the usefulness of Theorem 1.1 in various applications such as multiwavelets, numerical solutions of partial differential equations, and computer aided geometric design.

In work related to their study of $\sqrt{3}$-subdivision schemes (see [38]), Jiang and Oswald [32] developed Matlab software to calculate $\lambda(\Phi)$ in Theorem 1.1. It can be freely downloaded from http://cm.bell-labs.com/who/poswald/ and from http://www.math.umsl.edu/~jiang. The reader is referred to [32] for explanations of the Matlab routines.

2. Subdivision and transition operators. This section is devoted to a study of the subdivision and transition operators. To each $a \in \ell_0^{r\times r}(\mathbb{Z}^s)$ we associate two linear operators: the subdivision operator S_a and the transition operator T_a. The subdivision operator S_a is the linear operator on $\ell^{1\times r}(\mathbb{Z}^s)$ defined by

$$S_a u(\alpha) := \sum_{\beta\in\mathbb{Z}^s} u(\beta)a(\alpha - M\beta), \quad \alpha \in \mathbb{Z}^s,\ u \in \ell^{1\times r}(\mathbb{Z}^s).$$

The transition operator T_a is the linear operator on $\ell_0^r(\mathbb{Z}^s)$ defined by

$$T_a v(\alpha) := \sum_{\beta\in\mathbb{Z}^s} a(M\alpha - \beta)v(\beta), \quad \alpha \in \mathbb{Z}^s,\ v \in \ell_0^r(\mathbb{Z}^s).$$

See [3] and [10] for some earlier work on these operators.

We introduce a bilinear form on a pair of linear spaces $\ell_0^r(\mathbb{Z}^s)$ and $\ell^{1\times r}(\mathbb{Z}^s)$ as follows:

$$\langle u, v\rangle := \sum_{\alpha\in\mathbb{Z}^s} u(-\alpha)v(\alpha), \quad u \in \ell^{1\times r}(\mathbb{Z}^s),\ v \in \ell_0^r(\mathbb{Z}^s).$$

Then $\ell^{1\times r}(\mathbb{Z}^s)$ is the algebraic dual of $\ell_0^r(\mathbb{Z}^s)$ with respect to this bilinear form. For $u \in \ell^{1\times r}(\mathbb{Z}^s)$ and $v \in \ell_0^r(\mathbb{Z}^s)$, we have

$$\begin{aligned}\langle S_a u, v\rangle &= \sum_{\alpha\in\mathbb{Z}^s} (S_a u)(\alpha)v(-\alpha) = \sum_{\alpha\in\mathbb{Z}^s}\sum_{\beta\in\mathbb{Z}^s} u(\beta)a(\alpha - M\beta)v(-\alpha)\\ &= \sum_{\beta\in\mathbb{Z}^s}\sum_{\alpha\in\mathbb{Z}^s} u(-\beta)a(M\beta - \alpha)v(\alpha) = \sum_{\beta\in\mathbb{Z}^s} u(-\beta)(T_a v)(\beta) = \langle u, T_a v\rangle.\end{aligned}$$

Consequently, S_a is the algebraic adjoint of T_a.

The annihilator of a linear subspace U of $\ell^{1\times r}(\mathbb{Z}^s)$ is defined by

$$U^\perp := \{v \in \ell_0^r(\mathbb{Z}^s) : \langle u, v\rangle = 0\ \forall\, u \in U\}.$$

Similarly, the annihilator of a linear subspace V of $\ell_0^r(\mathbb{Z}^s)$ is defined by

$$V^\perp := \{u \in \ell^{1\times r}(\mathbb{Z}^s) : \langle u, v\rangle = 0\ \forall\, v \in V\}.$$

Clearly, $U \subseteq (U^\perp)^\perp$. If U is a *finite* dimensional subspace of $\ell^{1\times r}(\mathbb{Z}^s)$, then $(U^\perp)^\perp = U$. This comes from the *theorem on linear dependence* (see [33, p. 7]), which states that a linear functional f is a linear combination of a finite set $\{f_1, \dots, f_n\}$ of linear functionals if and only if the null space of f contains the intersection of the null spaces of $f_1, \dots, f_n$. Indeed, an element $u \in \ell^{1\times r}(\mathbb{Z}^s)$ can be viewed as a linear functional on $\ell_0^r(\mathbb{Z}^s)$. Suppose $\{u_1, \dots, u_n\}$ is a basis for U. Then $u \in (U^\perp)^\perp$ means the null space of u contains the intersection of the null spaces of $u_1, \dots, u_n$. Hence, by the theorem on linear dependence, u lies in U.

Moreover, if V is a linear subspace of $\ell_0^r(\mathbb{Z}^s)$, then $(V^\perp)^\perp = V$. In this case, V is not required to be finite dimensional. Clearly, $V \subseteq (V^\perp)^\perp$. The inclusion relation $(V^\perp)^\perp \subseteq V$ can be proved by a version of the Hahn–Banach theorem. Suppose $w \in \ell_0^r(\mathbb{Z}^s) \setminus V$. Let W be the linear span of V and w. Then we can find a linear functional f on W such that f vanishes on V and $f(w) = 1$. This linear functional can be extended to a linear functional on $\ell_0^r(\mathbb{Z}^s)$. Since $\ell^{1\times r}(\mathbb{Z}^s)$ is the algebraic dual of $\ell_0^r(\mathbb{Z}^s)$, this means that there exists some element u in $\ell^{1\times r}(\mathbb{Z}^s)$ such that $u \in V^\perp$ and $\langle u, w\rangle = 1$. Hence, $w \notin (V^\perp)^\perp$. This shows $(V^\perp)^\perp \subseteq V$.

LEMMA 2.1. *Let U be a finite dimensional linear subspace of $\ell^{1\times r}(\mathbb{Z}^s)$, and let $V := U^\perp$. Then U is invariant under the subdivision operator S_a if and only if V is invariant under the transition operator T_a.*

Proof. Suppose U is invariant under S_a. For $v \in V$ we have

$$\langle u, T_a v\rangle = \langle S_a u, v\rangle = 0 \quad \forall\, u \in U.$$

Hence, $T_a v \in U^\perp = V$. This shows that V is invariant under T_a.

Suppose V is invariant under T_a. For $u \in U$ we have

$$\langle S_a u, v\rangle = \langle u, T_a v\rangle = 0 \quad \forall\, v \in V.$$

Hence, $S_a u \in V^\perp = U$. This shows that U is invariant under S_a. ∎

It was proved in [15] that T_a has only finitely many nonzero eigenvalues. The following is an outline of this proof. By $\operatorname{supp} a$ we denote the set $\{\alpha \in \mathbb{Z}^s : a(\alpha) \neq 0\}$. Similarly, for $v \in \ell_0^r(\mathbb{Z}^s)$, $\operatorname{supp} v$ stands for the set $\{\alpha \in \mathbb{Z}^s : v(\alpha) \neq 0\}$. By the definition of T_a we see that $T_a v(\alpha) \neq 0$ if and only if

$$M\alpha - \beta \in \operatorname{supp} a \ \text{ for some } \ \beta \in \operatorname{supp} v.$$

Hence,

$$\operatorname{supp}(T_a v) \subseteq M^{-1}\operatorname{supp} a + M^{-1}\operatorname{supp} v.$$

Applying the above argument repeatedly, we obtain

$$\operatorname{supp}(T_a^n v) \subseteq \sum_{j=1}^{n} M^{-j}\operatorname{supp} a + M^{-n}\operatorname{supp} v. \tag{2.1}$$

Let

$$K := \mathbb{Z}^s \bigcap \left(\sum_{n=1}^{\infty} M^{-n}(\operatorname{supp} a) \right). \tag{2.2}$$

The preceding discussion tells us that $\operatorname{supp} v \subseteq K$ implies $\operatorname{supp}(T_a v) \subseteq K$. Therefore, $\ell^r(K)$ is invariant under T_a. Suppose v is an arbitrary element in $\ell_0^r(\mathbb{Z}^s)$. Comparing (2.1) with (2.2), we see that there exists a positive integer N such that, for $n \geq N$ and each $\alpha \in \operatorname{supp}(T_a^n v)$, the distance from the point α to the set K is less than $1/2$. However, $\alpha \in \mathbb{Z}^s$ and $K \subset \mathbb{Z}^s$, so α lies in K. This shows that $T_a^n v \in \ell^r(K)$ for sufficiently large n.

Now suppose θ is a nonzero eigenvalue of T_a and $T_a v = \theta v$ for some $v \in \ell_0^r(\mathbb{Z}^s)$. For sufficiently large n we have $\theta^n v = T_a^n v \in \ell^r(K)$. It follows that $v \in \ell^r(K)$. Since $\ell^r(K)$ is finite dimensional, T_a has only finitely many nonzero eigenvalues.

The following lemma extends the above results.

LEMMA 2.2. *Let V and W be two invariant subspaces of the transition operator T_a. Suppose W is finite dimensional and $V \cap \ell^r(K) \subseteq W \subseteq V$, where K is the set given in* (2.2). *Then*

$$\operatorname{spec}'\big(T_a|_W\big) = \operatorname{spec}'\big(T_a|_{V\cap\ell^r(K)}\big).$$

Proof. Let $\tilde{T}_a$ denote the quotient linear operator induced by T_a on the quotient space $W/(V \cap \ell^r(K))$. Clearly,

$$\operatorname{spec}\big(T_a|_W\big) = \operatorname{spec}(\tilde{T}_a) \cup \operatorname{spec}\big(T_a|_{V\cap\ell^r(K)}\big).$$

Thus, it suffices to show that all the eigenvalues of $\tilde{T}_a$ are zero. Let θ be an eigenvalue of $\tilde{T}_a$. Then there exists some $v \in W \setminus (V \cap \ell^r(K))$ such that

$$\tilde{T}_a(v + V \cap \ell^r(K)) = \theta(v + V \cap \ell^r(K)).$$

It follows that

$$T_a v - \theta v \in V \cap \ell^r(K).$$

Since $V \cap \ell^r(K)$ is invariant under T_a, for $n \in \mathbb{N}$ we have

$$T_a^n v - \theta^n v = (T_a^{n-1} + \cdots + \theta^{n-1})(T_a v - \theta v) \in V \cap \ell^r(K).$$

For sufficiently large n, $T_a^n v \in \ell^r(K)$. Hence, $\theta^n v \in V \cap \ell^r(K)$ for sufficiently large n. However, $v \notin V \cap \ell^r(K)$. Therefore, $\theta = 0$. The proof is complete. □

Lemma 2.2 tells us that

$$\rho\big(T_a|_W\big) = \rho\big(T_a|_{V\cap\ell^r(K)}\big).$$

This motivates us to define the spectral radius of $T_a|_V$ as $\rho\big(T_a|_{V\cap\ell^r(K)}\big)$.

LEMMA 2.3. *Let U be a finite dimensional invariant subspace of the subdivision operator S_a, and let $V := U^\perp$. Then*

$$\rho(T_a|_V) = \max\big\{|\nu| : \nu \in \operatorname{spec}\big((a(M\alpha-\beta))_{\alpha,\beta\in K}\big) \setminus \operatorname{spec}(S_a|_U)\big\}, \tag{2.3}$$

where $K = \mathbb{Z}^s \cap \sum_{n=1}^\infty M^{-n}(\operatorname{supp} a)$.

Proof. Suppose $\{u_1, \ldots, u_N\}$ is a basis for U. Then there exist $v_1, \ldots, v_N \in \ell_0^r(\mathbb{Z}^s)$ such that

$$\langle u_j, v_m\rangle = \delta_{jm} \quad \text{for } \ j,m = 1,\ldots,N, \tag{2.4}$$

where δ stands for the Kronecker sign. Let G be a bounded subset of $\mathbb{R}^s$ such that

$$G \supseteq \{0\} \cup \operatorname{supp} a \cup \big(\cup_{m=1}^N \operatorname{supp}(M v_m)\big),$$

and let $J := \mathbb{Z}^s \cap \big(\sum_{n=1}^\infty M^{-n} G\big)$. Then $K \subseteq J$ and $v_1, \ldots, v_N \in \ell^r(J)$. Moreover, $\ell^r(J) \cap V$ is an invariant subspace of T_a.

Consider the quotient space $\ell^r(J)/(\ell^r(J) \cap V)$. For $v \in \ell^r(J)$, let $\tilde{v}$ denote the coset $v + \ell^r(J) \cap V$. We claim that $\{\tilde{v}_1, \ldots, \tilde{v}_N\}$ forms a basis for $\ell^r(J)/(\ell^r(J) \cap V)$. Indeed, for $v \in \ell^r(J)$ we have

$$v - \sum_{j=1}^N \langle u_j, v\rangle\, v_j \in \ell^r(J) \cap V.$$

Consequently, $\tilde{v}$ lies in the span of $\{\tilde{v}_1, \ldots, \tilde{v}_N\}$. Furthermore, suppose $\sum_{m=1}^N c_m \tilde{v}_m = 0$. Then $\sum_{m=1}^N c_m v_m \in \ell^r(J) \cap V$. It follows that, for $j = 1, \ldots, N$,

$$c_j = \left\langle u_j, \sum_{m=1}^N c_m v_m \right\rangle = 0.$$

Hence, $\tilde{v}_1, \ldots, \tilde{v}_N$ are linearly independent. This justifies our claim.

Let $\tilde{T}_a$ denote the linear quotient operator induced by T_a on the quotient space $\ell^r(J)/(\ell^r(J) \cap V)$, that is, $\tilde{T}_a$ is defined by $\tilde{T}_a \tilde{v} := \widetilde{T_a v}$. Suppose

$$S_a u_j = \sum_{m=1}^N b_{jm} u_m \quad \text{and} \quad \tilde{T}_a \tilde{v}_j = \sum_{m=1}^N c_{jm} \tilde{v}_m, \quad j = 1, \ldots, N.$$

By (2.4) we have

$$b_{jm} = \langle S_a u_j, v_m \rangle = \langle u_j, T_a v_m \rangle = c_{mj}, \quad j, m = 1, \ldots, N.$$

Therefore,

$$\operatorname{spec}(\tilde{T}_a) = \operatorname{spec}\big(S_a|_U\big).$$

Consequently, we have

$$\operatorname{spec}(T_a|_{\ell^r(J)}) = \operatorname{spec}(\tilde{T}_a) \cup \operatorname{spec}(T_a|_{\ell^r(J)\cap V}) = \operatorname{spec}(S_a|_U) \cup \operatorname{spec}(T_a|_{\ell^r(J)\cap V}).$$

It follows that

$$\operatorname{spec}'(T_a|_{\ell^r(J)}) = \operatorname{spec}'(S_a|_U) \cup \operatorname{spec}'(T_a|_{\ell^r(J)\cap V}).$$

By Lemma 2.2,

$$\operatorname{spec}'(T_a|_{\ell^r(J)}) = \operatorname{spec}'(T_a|_{\ell^r(K)}) \quad \text{and} \quad \operatorname{spec}'(T_a|_{\ell^r(J)\cap V}) = \operatorname{spec}'(T_a|_{\ell^r(K)\cap V}).$$

Hence,

$$\operatorname{spec}'(T_a|_{\ell^r(K)}) = \operatorname{spec}'(S_a|_U) \cup \operatorname{spec}'(T_a|_{\ell^r(K)\cap V}). \tag{2.5}$$

Note that

$$\rho(T_a|_V) = \rho(T_a|_{\ell^r(K)\cap V}) = \max\{|\nu| : \nu \in \operatorname{spec}'(T_a|_{\ell^r(K)\cap V})\}.$$

In light of (2.5) we have

$$\rho(T_a|_V) = \max\{|\nu| : \nu \in \operatorname{spec}'(T_a|_{\ell^r(K)}) \setminus \operatorname{spec}'(S_a|_U)\}.$$

Finally,

$$\rho(T_a|_V) = \max\{|\nu| : \nu \in \operatorname{spec}(T_a|_{\ell^r(K)}) \setminus \operatorname{spec}(S_a|_U)\}.$$

However, $\operatorname{spec}(T_a|_{\ell^r(K)}) = \operatorname{spec}\big((a(M\alpha - \beta))_{\alpha,\beta\in K}\big)$. Taking this into account, we obtain the desired formula (2.3). □

3. Polynomial reproducibility. Let Φ be an $r \times 1$ vector $(\phi_1, \dots, \phi_r)^T$, where $\phi_1, \dots, \phi_r$ are compactly supported integrable functions on $\mathbb{R}^s$. If there exists a (finite) linear combination ψ of shifts of $\phi_1, \dots, \phi_r$ such that

$$\sum_{\alpha \in \mathbb{Z}^s} q(\alpha)\psi(\cdot - \alpha) = q \quad \forall\, q \in \Pi_{k-1}, \tag{3.1}$$

then we say that Φ reproduces all polynomials of degree at most $k-1$. In this section we review results on polynomial reproducibility relevant to our study of smoothness of refinable vectors of functions.

For $j = 1, \dots, s$, let e_j denote the jth column of the $s \times s$ identity matrix. We may view $e_1, \dots, e_s$ as the coordinate unit vectors in $\mathbb{R}^s$. By D_j we denote the partial derivative with respect to the jth coordinate. For a multi-index $\mu = (\mu_1, \dots, \mu_s)$, D^μ stands for the differential operator $D_1^{\mu_1} \cdots D_s^{\mu_s}$.

The conditions in (3.1) are equivalent to the following conditions:

$$D^\mu \hat{\psi}(2\pi\beta) = \delta_{0\mu}\delta_{0\beta} \qquad \forall\, |\mu| < k \text{ and } \beta \in \mathbb{Z}^s.$$

If ψ satisfies the above conditions, then we say that Φ satisfies the Strang–Fix conditions of order k (see [44]). In [8] Dahmen and Micchelli investigated approximation order on the basis of the Strang–Fix conditions.

It is easily seen that Φ satisfies the Strang–Fix conditions of order k if and only if there exists a $1 \times r$ vector y of trigonometric polynomials such that

$$D^\mu (y\hat{\Phi})(2\pi\beta) = \delta_{0\mu}\delta_{0\beta} \qquad \forall\, |\mu| < k \text{ and } \beta \in \mathbb{Z}^s. \tag{3.2}$$

If y satisfies the conditions in (3.2), then we have

$$\frac{x^\mu}{\mu!} = \sum_{\alpha \in \mathbb{Z}^s} u_\mu(\alpha)\Phi(x - \alpha), \quad x \in \mathbb{R}^s,\ |\mu| < k, \tag{3.3}$$

where

$$u_\mu(\alpha) := \sum_{\nu \le \mu} \frac{(-iD)^{\mu-\nu} y(0)}{(\mu - \nu)!} \frac{\alpha^\nu}{\nu!}, \quad \alpha \in \mathbb{Z}^s. \tag{3.4}$$

See the recent survey paper [24] for a proof of this result.

Now suppose Φ satisfies the refinement equation (1.1). Naturally, we wish to find the optimal order of the Strang–Fix conditions satisfied by Φ in terms of the mask. There has been a lot of research done on this problem. See [17] and [39] for the univariate case ($s = 1$) and [1], [2], and [31] for the multivariate case ($s > 1$). The results in these papers can be summarized as follows (see [24]). Suppose $\Phi = (\phi_1, \dots, \phi_r)^T$ satisfies the refinement equation (1.1) with a being its mask. Let $A(\xi)$ ($\xi \in \mathbb{R}^s$) be the $r \times r$ matrix given in (1.3). Let y be a $1 \times r$ vector of trigonometric polynomials, and let $g(\xi) := y(M^T\xi)A(\xi)$, $\xi \in \mathbb{R}^s$. Then (3.2) is valid, provided the following three conditions are satisfied:

(P1) $y(0)\hat{\Phi}(0) = 1$;

(P2) $D^\mu g(2\pi(M^T)^{-1}\omega) = 0$ for all $|\mu| < k$ and $\omega \in \mathbb{Z}^s \setminus (M^T\mathbb{Z}^s)$;

(P3) $D^\mu g(0) = D^\mu y(0)$ for all $|\mu| < k$.

Conversely, if Φ is stable and (3.2) is valid, then conditions (P1), (P2), and (P3) are satisfied.

For the special case $k = 1$, it is known (see, e.g., [24]) that conditions (P2) and (P3) together are equivalent to

$$y_0 \sum_{\beta\in\mathbb{Z}^s} a(\alpha - M\beta) = y_0 \quad \forall\, \alpha \in \mathbb{Z}^s, \tag{3.5}$$

where $y_0 := y(0) \in \mathbb{C}^{1\times r}$. If this is the case, we say that a satisfies the basis sum rule with respect to y_0. For the general case $k \geq 1$, conditions (P2) and (P3) can also be expressed as sum rules involving a. Thus, we say that a satisfies the sum rules of order k with respect to y if y and $g : \xi \mapsto y(M^T\xi)A(\xi)$ $(\xi \in \mathbb{R}^s)$ satisfy conditions (P2) and (P3). If the meaning of y is clear from the context, then the reference to y may be omitted. We always assume that $y(0) \neq 0$.

Let Ω be a complete set of representatives of the distinct cosets of $\mathbb{Z}^s/M^T\mathbb{Z}^s$. We assume $0 \in \Omega$. Clearly, $\#\Omega$ (the number of elements in Ω) is equal to $d := |\det M|$. Note that condition (P2) can be restated as $D^\mu g(2\pi(M^T)^{-1}\omega) = 0$ for all $|\mu| < k$ and $\omega \in \Omega \setminus \{0\}$. For $v \in \ell_0^r(\mathbb{Z}^s)$ and $\alpha \in \mathbb{Z}^s$ we have

$$\begin{aligned}\sum_{\omega\in\Omega} \hat{v}\big((M^T)^{-1}(\xi + 2\pi\omega)\big) &= \sum_{\omega\in\Omega}\sum_{\alpha\in\mathbb{Z}^s} v(\alpha)e^{-i\alpha\cdot((M^T)^{-1}(\xi+2\pi\omega))} \\ &= \sum_{\alpha\in\mathbb{Z}^s} v(\alpha)e^{-iM^{-1}\alpha\cdot\xi} \sum_{\omega\in\Omega} e^{-2\pi iM^{-1}\alpha\cdot\omega}.\end{aligned}$$

With the help of the identity (see, e.g., [20, Lem. 3.2])

$$\sum_{\omega\in\Omega} e^{-2\pi iM^{-1}\alpha\cdot\omega} = \begin{cases} d & \text{if } \alpha \in M\mathbb{Z}^s, \\ 0 & \text{if } \alpha \notin M\mathbb{Z}^s, \end{cases}$$

we obtain

$$\sum_{\omega\in\Omega} \hat{v}\big((M^T)^{-1}(\xi + 2\pi\omega)\big) = d \sum_{\alpha\in\mathbb{Z}^s} v(M\alpha)e^{-i\alpha\cdot\xi}, \quad \xi \in \mathbb{R}^s.$$

The convolution of $u \in \ell^{m\times n}(\mathbb{Z}^s)$ and $v \in \ell_0^n(\mathbb{Z}^s)$ is the element in $\ell^m(\mathbb{Z}^s)$ given by

$$u{*}v(\alpha) := \sum_{\beta\in\mathbb{Z}^s} u(\alpha - \beta)v(\beta), \quad \alpha \in \mathbb{Z}^s.$$

Suppose $v \in \ell_0^r(\mathbb{Z}^s)$. By the definition of the transition operator T_a, we have

$$(T_a v)(\alpha) = (a{*}v)(M\alpha), \quad \alpha \in \mathbb{Z}^s.$$

Hence

$$(T_a v)\hat{}(\xi) = \sum_{\alpha\in\mathbb{Z}^s} (a{*}v)(M\alpha)e^{-i\alpha\cdot\xi} = \frac{1}{d}\sum_{\omega\in\Omega} (a{*}v)\hat{}\big((M^T)^{-1}(\xi + 2\pi\omega)\big), \quad \xi \in \mathbb{R}^s.$$

It follows that

$$(T_a v)\hat{}(\xi) = \sum_{\omega\in\Omega} A\big((M^T)^{-1}(\xi + 2\pi\omega)\big)\, \hat{v}\big((M^T)^{-1}(\xi + 2\pi\omega)\big), \quad \xi \in \mathbb{R}^s. \tag{3.6}$$

LEMMA 3.1. *Let $a \in \ell_0^{r\times r}(\mathbb{Z}^s)$. Suppose a satisfies the sum rules of order k with respect to $y \in \mathbb{T}^{1\times r}(\mathbb{R}^s)$. Then the linear space H_j $(0 \le j < k)$ given by*

$$H_j := \big\{\, v \in \ell_0^r(\mathbb{Z}^s) : D^\mu(y\hat{v})(0) = 0 \;\forall\, |\mu| = j \,\big\}$$

is invariant under the transition operator T_a.

Proof. By (3.6) we have

$$(T_a v)\hat{}\,(M^T\xi) = \sum_{\omega\in\Omega} A\big(\xi + 2\pi(M^T)^{-1}\omega\big)\hat{v}\big(\xi + 2\pi(M^T)^{-1}\omega\big), \quad \xi \in \mathbb{R}^s.$$

It follows that

$$y(M^T\xi)(T_a v)\hat{}\,(M^T\xi) = \sum_{\omega\in\Omega} y(M^T\xi)A\big(\xi+2\pi(M^T)^{-1}\omega\big)\hat{v}\big(\xi+2\pi(M^T)^{-1}\omega\big), \quad \xi \in \mathbb{R}^s.$$

For $\omega \in \Omega \setminus \{0\}$, we have by (P2)

$$D^\mu\big(y(M^T\xi)A(\xi + 2\pi(M^T)^{-1}\omega)\big)|_{\xi=0} = D^\mu g\big(2\pi(M^T)^{-1}\omega\big) = 0 \quad \forall\, |\mu| < k.$$

For $\omega = 0$, we have $D^\mu g(0) = D^\mu y(0)$ for all $|\mu| < k$. Hence,

$$D^\mu\big(y(M^T\xi)A(\xi)\hat{v}(\xi)\big)|_{\xi=0} = D^\mu\big(g(\xi)\hat{v}(\xi)\big)|_{\xi=0} = D^\mu(y\hat{v})(0).$$

However, $v \in H_j$ implies $D^\mu(y\hat{v})(0) = 0$ for all $|\mu| = j$. Therefore,

$$D^\mu\big(y(M^T\xi)(T_a v)\hat{}\,(M^T\xi)\big)|_{\xi=0} = 0 \quad \forall\, |\mu| = j.$$

Let $f(\xi) := y(\xi)(T_a v)\hat{}\,(\xi)$, $\xi \in \mathbb{R}^s$. We use $f\circ M^T$ to denote the composition of f and M^T. The above equation tells us that, for all $|\mu| = j$, $D^\mu(f\circ M^T)(0) = 0$. Clearly, $f = (f\circ M^T)\circ(M^T)^{-1}$. By the chain rule, $D^\mu f$ is a linear combination of $D^\nu(f\circ M^T)$, $|\nu| = j$. Therefore, $D^\mu(y\widehat{(T_a v)})(0) = D^\mu f(0) = 0$ for all $|\mu| = j$, i.e., $T_a v \in H_j$. This shows that H_j is invariant under T_a. ∎

By the Leibniz rule for differentiation we have

$$(-iD)^\mu(y\hat{v})(0) = \sum_{\nu\le\mu} \binom{\mu}{\nu} (-iD)^{\mu-\nu} y(0)(-iD)^\nu \hat{v}(0).$$

However,

$$(-iD)^\nu \hat{v}(\xi) = \sum_{\alpha\in\mathbb{Z}^s} v(\alpha)(-\alpha)^\nu e^{-i\alpha\cdot\xi}, \quad \xi \in \mathbb{R}^s.$$

It follows that

$$(-iD)^\nu \hat{v}(0) = \sum_{\alpha\in\mathbb{Z}^s} v(\alpha)(-\alpha)^\nu.$$

Hence,

$$\frac{(-iD)^\mu(y\hat{v})(0)}{\mu!} = \sum_{\alpha\in\mathbb{Z}^s}\sum_{\nu\le\mu} \frac{(-iD)^{\mu-\nu}y(0)}{(\mu-\nu)!}\,\frac{(-\alpha)^\nu}{\nu!}\,v(\alpha) = \sum_{\alpha\in\mathbb{Z}^s} u_\mu(-\alpha)v(\alpha) = \langle u_\mu, v\rangle,$$

where u_μ is the element in $\ell^{1\times r}(\mathbb{Z}^s)$ as defined in (3.4). Consequently, v lies in H_j if and only if $\langle u_\mu, v\rangle = 0$ for all $|\mu| = j$. In other words, $H_j = G_j^\perp$, where

$$G_j := \operatorname{span}\{u_\mu : |\mu| = j\}.$$

Let $U_k := \operatorname{span}\{u_\mu : |\mu| < k\}$ and

$$V_k := \{v \in \ell_0^r(\mathbb{Z}^s) : D^\mu(y\hat{v})(0) = 0 \ \forall\, |\mu| < k\}. \tag{3.7}$$

Then $V_k = U_k^\perp$. Moreover,

$$U_k = G_0 + G_1 + \cdots + G_{k-1} \quad \text{and} \quad V_k = H_0 \cap H_1 \cap \cdots \cap H_{k-1}.$$

We may write u_μ as $\sum_{\nu\le\mu} y_{\mu-\nu} q_\nu$, where $y_{\mu-\nu} := (-iD)^{\mu-\nu} y(0)/(\mu-\nu)!$ and q_ν is the sequence given by $q_\nu(\alpha) = \alpha^\nu/\nu!$, $\alpha \in \mathbb{Z}^s$. If $y_0 \ne 0$, then the set $\{u_\mu : |\mu| < k\}$ is linearly independent. To justify our claim, let c_μ $(|\mu| < k)$ be complex numbers such that $\sum_{|\mu|<k} c_\mu u_\mu = 0$. It follows that

$$\sum_{|\mu|=k-1} c_\mu y_0 q_\mu + \sum_{|\nu|<k-1} h_\nu q_\nu = 0,$$

where h_ν $(|\nu| < k-1)$ are some elements in $\mathbb{C}^{1\times r}$. Since q_μ $(|\mu| < k)$ are linearly independent, we have $c_\mu y_0 = 0$ for all $|\mu| = k-1$. However, $y_0 \ne 0$. Hence, $c_\mu = 0$ for all $|\mu| = k-1$. By using this argument repeatedly, we see that $c_\mu = 0$ for all $|\mu| = j$, $j = k-1, k-2, \ldots, 0$. This shows that $\{u_\mu : |\mu| < k\}$ is linearly independent. Consequently, $\{u_\mu : |\mu| = j\}$ is a basis for G_j $(j < k)$.

For $\gamma \in \mathbb{Z}^s$, the difference operator ∇_γ on the space $\ell^{m\times n}(\mathbb{Z}^s)$ is defined by

$$\nabla_\gamma u = u - u(\cdot - \gamma), \quad u \in \ell^{m\times n}(\mathbb{Z}^s).$$

Let us consider $\nabla_\gamma u_\mu$. For $\alpha \in \mathbb{Z}^s$ we have

$$u_\mu(\alpha) - u_\mu(\alpha-\gamma) = \sum_{\nu\le\mu} \frac{1}{\nu!}\big[\alpha^\nu - (\alpha-\gamma)^\nu\big] y_{\mu-\nu} = \sum_{\nu\le\mu} \sum_{0<\tau\le\nu} -\frac{1}{\nu!}\binom{\nu}{\tau}(-\gamma)^\tau \alpha^{\nu-\tau} y_{\mu-\nu}.$$

It follows that

$$\begin{aligned} \nabla_\gamma u_\mu(\alpha) &= \sum_{0<\tau\le\mu} -\frac{(-\gamma)^\tau}{\tau!} \sum_{\tau\le\nu\le\mu} \frac{\alpha^{\nu-\tau}}{(\nu-\tau)!} y_{(\mu-\tau)-(\nu-\tau)} \\ &= \sum_{0<\tau\le\mu} -\frac{(-\gamma)^\tau}{\tau!} u_{\mu-\tau}(\alpha). \end{aligned} \tag{3.8}$$

Consequently, $\nabla_\gamma u_\mu \in \operatorname{span}\{u_\nu : \nu < \mu\}$.

For $\mu \in \mathbb{N}_0^s$, recall that q_μ is the sequence given by $q_\mu(\alpha) = \alpha^\mu/\mu!$, $\alpha \in \mathbb{Z}^s$. When $\mu \in \mathbb{Z}^s \setminus \mathbb{N}_0^s$, we agree that $q_\mu = 0$. With this convention, we may interpret $D_j q_\mu$ as $q_{\mu-e_j}$. For $\gamma = (\gamma_1, \ldots, \gamma_s) \in \mathbb{Z}^s$, let $D_\gamma := \gamma_1 D_1 + \cdots + \gamma_s D_s$. Then it follows from (3.8) that

$$\nabla_\gamma u_\mu - D_\gamma u_\mu \in \operatorname{span}\{u_\nu : |\nu| \le |\mu| - 2\}.$$

Let Γ be a finite multiset of elements in $\mathbb{Z}^s$. If $\#\Gamma \ge |\mu|$, then the above relation yields

$$\left(\prod_{\gamma\in\Gamma} \nabla_\gamma\right) u_\mu = \left(\prod_{\gamma\in\Gamma} D_\gamma\right) u_\mu. \tag{3.9}$$

Moreover, both sides of the above equation vanish when $\#\Gamma > |\mu|$.

For $j = 1, \dots, s$, the difference operator ∇_{e_j} is abbreviated as ∇_j. For a multi-index $\tau = (\tau_1, \dots, \tau_s) \in \mathbb{N}_0^s$, the difference operator ∇^τ is defined as $\nabla_1^{\tau_1} \cdots \nabla_s^{\tau_s}$. As a consequence of (3.9) we have

$$\nabla^\tau u_\mu = D^\tau u_\mu = \delta_{\tau\mu} u_0 \quad \text{for } |\tau| \ge |\mu|. \tag{3.10}$$

Furthermore, it follows from (3.9) that

$$\nabla_{Me_1}^{\tau_1} \cdots \nabla_{Me_s}^{\tau_s} u_\mu = D_{Me_1}^{\tau_1} \cdots D_{Me_s}^{\tau_s} u_\mu \quad \text{for } |\tau| \ge |\mu|. \tag{3.11}$$

Suppose $Me_n = m_{n1}e_1 + \cdots + m_{ns}e_s$ with suitable coefficients m_{nj}, $n, j = 1, \dots, s$. Then for $|\tau| = j$ we have

$$D_{Me_1}^{\tau_1} \cdots D_{Me_s}^{\tau_s} = \prod_{n=1}^{s} \big(m_{n1}D_1 + \cdots + m_{ns}D_s\big)^{\tau_n} =: \sum_{|\nu|=j} b_{\tau\nu} D^\nu.$$

Since $\operatorname{spec}(M) = \{\sigma_1, \dots, \sigma_s\}$, the spectrum of the matrix $(b_{\tau\nu})_{|\tau|=j,|\nu|=j}$ is $\{\sigma^\mu : |\mu| = j\}$ (see [2, Lem. 4.2]). In light of (3.10), (3.11) yields

$$\nabla_{Me_1}^{\tau_1} \cdots \nabla_{Me_s}^{\tau_s} u_\mu = \sum_{|\nu|=j} b_{\tau\nu} D^\nu u_\mu = b_{\tau\mu} u_0 \quad \text{for } |\tau| \ge |\mu|, \tag{3.12}$$

where $b_{\tau\mu}$ is understood to be 0 if $|\tau| > |\mu|$.

LEMMA 3.2. *Under the conditions in Lemma* 3.1*, the linear space* G_j $(j < k)$ *is invariant under the subdivision operator* S_a*. If, in addition,* $y(0) \neq 0$*, then*

$$\operatorname{spec}(S_a|_{G_j}) = \{\sigma^{-\mu} : |\mu| = j\}.$$

Proof. Note that $y_0 = y(0)$ and $u_0 = y_0 q_0$, where $q_0(\alpha) = 1$ for all $\alpha \in \mathbb{Z}^s$. Since a satisfies the basic sum rule with respect to y_0, (3.5) is valid. Hence, for $\alpha \in \mathbb{Z}^s$ we have

$$S_a u_0(\alpha) = \sum_{\beta \in \mathbb{Z}^s} u_0(\beta) a(\alpha - M\beta) = y_0 \sum_{\beta \in \mathbb{Z}^s} a(\alpha - M\beta) = y_0 = u_0(\alpha).$$

This shows $S_a u_0 = u_0$.

Since $G_j^\perp = H_j$ and H_j is invariant under T_a, the linear space G_j is invariant under S_a by Lemma 2.1. Thus, there exist complex numbers $c_{\mu\nu}$ such that

$$S_a u_\mu = \sum_{|\nu|=j} c_{\mu\nu} u_\nu, \quad |\mu| = j.$$

Let C denote the matrix $(c_{\mu\nu})_{|\mu|=j,|\nu|=j}$. Then $\operatorname{spec}(S_a|_{G_j}) = \operatorname{spec}(C)$.

For $\gamma \in \mathbb{Z}^s$, it can be easily verified that

$$S_a(\nabla_\gamma u_\mu) = \nabla_{M\gamma}(S_a u_\mu).$$

Consequently, for $\tau = (\tau_1, \dots, \tau_s) \in \mathbb{N}_0^s$ we have

$$S_a(\nabla^\tau u_\mu) = \nabla_{Me_1}^{\tau_1} \cdots \nabla_{Me_s}^{\tau_s} (S_a u_\mu) = \sum_{|\nu|=j} c_{\mu\nu} \big(\nabla_{Me_1}^{\tau_1} \cdots \nabla_{Me_s}^{\tau_s}\big) u_\nu.$$

In light of (3.9) and (3.12), it follows that

$$\delta_{\tau\mu} u_0 = \sum_{|\nu|=j} c_{\mu\nu} b_{\tau\nu} u_0.$$

Hence, $C = (B^T)^{-1}$, where B denotes the matrix $(b_{\tau\nu})_{|\tau|=j,|\nu|=j}$, but the spectrum of B is $\{\sigma^\mu : |\mu| = j\}$. Therefore, $\operatorname{spec}(C) = \{\sigma^{-\mu} : |\mu| = j\}$. This completes the proof. □

Recall that U_k is the direct sum of $G_0, \ldots, G_{k-1}$ and $V_k = U_k^\perp$. Hence, we have the following result.

LEMMA 3.3. *Under the conditions in Lemma* 3.1*, V_k is invariant under the transition operator T_a, and U_k is invariant under the subdivision operator S_a. If, in addition, $y(0) \neq 0$, then*

$$\operatorname{spec}(S_a|_{U_k}) = \{\sigma^{-\mu} : |\mu| < k\}.$$

4. Characterization of smoothness. In this section we give a characterization for the smoothness of a refinable vector of functions in terms of the corresponding mask.

Sobolev spaces are related to Lipschitz spaces, which are defined on the basis of the modulus of smoothness. The *modulus of continuity* of a function f in $L_p(\mathbb{R}^s)$ is defined by

$$\omega(f,h)_p := \sup_{|t|\le h} \big\|\nabla_t f\big\|_p, \qquad h \ge 0,$$

where $\nabla_t f := f - f(\cdot - t)$. Let k be a positive integer. The kth *modulus of smoothness* of $f \in L_p(\mathbb{R}^s)$ is defined by

$$\omega_k(f,h)_p := \sup_{|t|\le h} \big\|\nabla_t^k f\big\|_p, \qquad h \ge 0.$$

For $1 \le p \le \infty$ and $0 < \lambda \le 1$, the Lipschitz space $\operatorname{Lip}(\lambda, L_p(\mathbb{R}^s))$ consists of all functions $f \in L_p(\mathbb{R}^s)$ for which

$$\omega(f,h)_p \le C\, h^\lambda \qquad \forall\, h > 0,$$

where C is a positive constant independent of h. For $\lambda > 0$ we write $\lambda = m + \eta$, where m is an integer and $0 < \eta \le 1$. The Lipschitz space $\operatorname{Lip}(\lambda, L_p(\mathbb{R}^s))$ consists of those functions $f \in L_p(\mathbb{R}^s)$ for which $D^\mu f \in \operatorname{Lip}(\eta, L_p(\mathbb{R}^s))$ for all multi-indices μ with $|\mu| = m$. For $\lambda > 0$, let k be an integer greater than λ. The generalized Lipschitz space $\operatorname{Lip}^*(\lambda, L_p(\mathbb{R}^s))$ consists of those functions $f \in L_p(\mathbb{R}^s)$ for which

$$\omega_k(f,h)_p \le C\, h^\lambda \qquad \forall\, h > 0,$$

where C is a positive constant independent of h. If $\lambda > 0$ is not an integer, then

$$\operatorname{Lip}(\lambda, L_p(\mathbb{R}^s)) = \operatorname{Lip}^*(\lambda, L_p(\mathbb{R}^s)), \qquad 1 \le p \le \infty.$$

See [11, Chap. 2] for a discussion about Lipschitz spaces.

It is well known that, for $\lambda > \varepsilon > 0$, the inclusion relations

$$\operatorname{Lip}(\lambda, L_2(\mathbb{R}^s)) \subseteq \operatorname{Lip}^*(\lambda, L_2(\mathbb{R}^s)) \subseteq \operatorname{Lip}(\lambda - \varepsilon, L_2(\mathbb{R}^s))$$

and

$$W_2^\lambda(\mathbb{R}^s) \subseteq \mathrm{Lip}(\lambda, L_2(\mathbb{R}^s)) \subseteq W_2^{\lambda-\varepsilon}(\mathbb{R}^s)$$

hold true. See [43, Chap. V] for these facts. Therefore, we have

$$\lambda(f) = \sup\{\lambda : f \in \mathrm{Lip}(\lambda, L_2(\mathbb{R}^s))\} = \sup\{\lambda : f \in \mathrm{Lip}^*(\lambda, L_2(\mathbb{R}^s))\}.$$

The inner product of two functions $f, g \in L_2(\mathbb{R}^s)$ is defined as

$$\langle f, g\rangle := \int_{\mathbb{R}^s} f(x)\,\overline{g(x)}\,dx.$$

This definition still makes sense if f is a compactly supported function in $L_2(\mathbb{R}^s)$ and g is a polynomial on $\mathbb{R}^s$.

By $(L_p(\mathbb{R}^s))^r$ we denote the linear space of all $r \times 1$ vectors $F = (f_1, \dots, f_r)^T$ such that $f_1, \dots, f_r \in L_p(\mathbb{R}^s)$. This space is equipped with the norm given by

$$\|F\|_p := \left(\sum_{j=1}^r \|f_j\|_p^p\right)^{1/p}, \quad F = (f_1, \dots, f_r)^T \in (L_p(\mathbb{R}^s))^r.$$

Suppose $u \in \ell^{m\times n}(\mathbb{Z}^s)$ and $u(\alpha) = (u_{jk}(\alpha))_{1\le j\le m, 1\le k\le n}$ for $\alpha \in \mathbb{Z}^s$. We define

$$\|u\|_p := \left(\sum_{\alpha\in\mathbb{Z}^s} \sum_{1\le j\le m} \sum_{1\le k\le n} \big|u_{jk}(\alpha)\big|^p\right)^{1/p}, \quad 1 \le p \le \infty.$$

Let Φ be an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R}^s)$. Suppose Φ satisfies the refinement equation (1.1). We claim that

$$\Phi = \sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)\Phi(M^n\cdot - \alpha), \tag{4.1}$$

where the sequences a_n are given by $a_1 = a$ and, for $n = 2, 3, \dots,$

$$a_n(\alpha) = \sum_{\beta\in\mathbb{Z}^s} a_{n-1}(\beta)a(\alpha - M\beta), \quad \alpha \in \mathbb{Z}^s. \tag{4.2}$$

This can be proved by induction on n. Indeed, (4.1) is valid for $n = 1$. Suppose (4.1) holds true for $n-1$. Then we have

$$\Phi = \sum_{\beta\in\mathbb{Z}^s} a_{n-1}(\beta)\Phi(M^{n-1}\cdot - \beta) = \sum_{\beta\in\mathbb{Z}^s} a_{n-1}(\beta) \sum_{\alpha\in\mathbb{Z}^s} a(\alpha)\Phi(M^n\cdot - M\beta - \alpha).$$

It follows that

$$\Phi = \sum_{\alpha\in\mathbb{Z}^s} \left(\sum_{\beta\in\mathbb{Z}^s} a_{n-1}(\beta)a(\alpha - M\beta)\right)\Phi(M^n\cdot - \alpha) = \sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)\Phi(M^n\cdot - \alpha).$$

This completes the induction procedure.

Let $\Phi = (\phi_1, \dots, \phi_r)^T$ be an $r \times 1$ vector of compactly supported functions in $L_2(\mathbb{R}^s)$ satisfying the refinement equation (1.1) with a being the mask. Recall that

$d = |\det M|$. Suppose a satisfies the sum rules of order k with respect to $y \in \mathbb{T}^{1\times r}(\mathbb{R}^s)$ satisfying (3.2). Let

$$V_k := \big\{ v \in \ell_0^r(\mathbb{Z}^s) : D^\mu(y\hat{v})(0) = 0 \,\forall\, |\mu| < k \big\}.$$

THEOREM 4.1. *If for every $v \in V_k$ there exists a positive constant C_v independent of n such that*

$$\|a_n * v\|_2 \le C_v d^{(1/2-\lambda/s)n} \quad \forall\, n \in \mathbb{N}, \tag{4.3}$$

then $\Phi \in (Lip^*(\lambda, L_2(\mathbb{R}^s)))^r$. Conversely, if $\Phi \in (Lip(\lambda, L_2(\mathbb{R}^s)))^r$, and if Φ is stable, then (4.3) is valid for $v \in V_k$ and $k > \lambda$.

Proof. Recall that $e_1, \ldots, e_s$ are the coordinate unit vectors in $\mathbb{R}^s$. If there exists a constant C such that

$$\big\|\nabla^k_{M^{-n}e_j}\Phi\big\|_2 \le C d^{(-\lambda/s)n} \quad \forall\, n \in \mathbb{N} \quad \text{and} \quad j = 1, \ldots, s, \tag{4.4}$$

then [21, Thm. 2.1] tells us that Φ lies in $(\mathrm{Lip}^*(\lambda, L_2(\mathbb{R}^s)))^r$.

It follows from (4.1) that

$$\nabla_{M^{-n}e_j}\Phi = \sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)\big[\Phi(M^n\cdot - \alpha) - \Phi(M^n\cdot - \alpha - e_j)\big] = \sum_{\alpha\in\mathbb{Z}^s} \nabla_j a_n(\alpha)\Phi(M^n\cdot - \alpha).$$

Applying the difference operator $\nabla_{M^{-n}e_j}$ to (4.1) repeatedly, we obtain

$$\nabla^k_{M^{-n}e_j}\Phi = \sum_{\alpha\in\mathbb{Z}^s} \nabla^k_j a_n(\alpha)\Phi(M^n\cdot - \alpha).$$

Since Φ is compactly supported, it follows that

$$\big\|\nabla^k_{M^{-n}e_j}\Phi\big\|_2 \le C d^{-n/2}\big\|\nabla^k_j a_n\big\|_2,$$

where C is a constant independent of n. For $m = 1, \ldots, r$, let v_m be the element in $\ell_0^r(\mathbb{Z}^s)$ such that $v_m(\alpha) = 0$ for all $\alpha \in \mathbb{Z}^s \setminus \{0\}$ and $v_m(0)$ is the mth column of the $r \times r$ identity matrix. We have

$$\big\|\nabla^k_j a_n\big\|_2 \le \sum_{m=1}^r \big\|(\nabla^k_j a_n) * v_m\big\|_2 = \sum_{m=1}^r \big\|a_n * (\nabla^k_j v_m)\big\|_2.$$

We observe that $(\nabla^k_j v_m)\hat{}\,(\xi) = (1 - e^{-i\xi_j})^k \hat{v}_m(\xi)$ for $\xi = (\xi_1, \ldots, \xi_s) \in \mathbb{R}^s$. Hence, for $|\mu| < k$, $D^\mu\big(y(\nabla^k_j v_m)\hat{}\,\big)(0) = 0$ with y as in (3.2). In other words, $\nabla^k_j v_m \in V_k$, $m = 1, \ldots, r$.

If (4.3) is valid, then

$$\big\|a_n * (\nabla^k_j v_m)\big\|_2 \le C_m d^{(1/2-\lambda/s)n} \quad \forall\, n \in \mathbb{N},$$

where C_m is a constant independent of n. Combining the above estimates, we obtain the desired estimate (4.4). Therefore, $\Phi \in (\mathrm{Lip}^*(\lambda, L_2(\mathbb{R}^s)))^r$.

Now suppose $\Phi = (\phi_1, \ldots, \phi_r)^T \in (\mathrm{Lip}(\lambda, L_2(\mathbb{R}^s)))^r$ and Φ is stable. We wish to show that (4.3) is true. For this purpose, we shall use approximation schemes induced by quasi-projection operators (see [35] and [23]).

For $\nu \in \mathbb{N}_0^s$, let q_ν be the monomial given by $q_\nu(x) := x^\nu/\nu!$, $x \in \mathbb{R}^s$. Recall that $y_\nu = (-iD)^\nu y(0)/\nu!$. Each y_ν is a $1 \times r$ vector $(y_{\nu 1}, \dots, y_{\nu r})$. There exist real-valued compactly supported functions $g_1, \dots, g_r$ in $L_2(\mathbb{R}^s)$ such that

$$\langle q_\nu, g_j \rangle = y_{\nu j} \quad \forall\, |\nu| < k \text{ and } j = 1, \dots, r.$$

For $|\mu| < k$ and $\alpha \in \mathbb{Z}^s$ we have

$$\begin{aligned}\langle q_\mu, g_j(\cdot - \alpha)\rangle &= \langle q_\mu(\cdot + \alpha), g_j\rangle = \sum_{\nu \le \mu} \int_{\mathbb{R}^s} \frac{1}{\mu!}\binom{\mu}{\nu} x^{\mu-\nu}\alpha^\nu g_j(x)\, dx \\ &= \sum_{\nu \le \mu} \frac{\alpha^\nu}{\nu!} \int_{\mathbb{R}^s} \frac{x^{\mu-\nu}}{(\mu-\nu)!}\, g_j(x)\, dx = \sum_{\nu \le \mu} \frac{\alpha^\nu}{\nu!} y_{\mu-\nu, j}.\end{aligned}$$

Let P_Φ be the quasi-projection operator given by

$$P_\Phi f := \sum_{\alpha \in \mathbb{Z}^s} \sum_{j=1}^r \langle f, g_j(\cdot - \alpha)\rangle \phi_j(\cdot - \alpha), \quad f \in L_2(\mathbb{R}^s).$$

For $|\mu| < k$ we have

$$P_\Phi q_\mu = \sum_{\alpha \in \mathbb{Z}^s} \sum_{j=1}^r \langle q_\mu, g_j(\cdot - \alpha)\rangle \phi_j(\cdot - \alpha) = \sum_{\alpha \in \mathbb{Z}^s} \sum_{\nu \le \mu} \frac{\alpha^\nu}{\nu!} y_{\mu-\nu} \Phi(\cdot - \alpha) = q_\mu,$$

where (3.3) has been used to derive the last equality. Thus, P_Φ reproduces all polynomials of degree at most $k-1$, i.e., $P_\Phi q = q$ for all $q \in \Pi_{k-1}$. Consequently, for $f \in \mathrm{Lip}(\lambda, L_2(\mathbb{R}^s))$ $(0 < \lambda < k)$ we have

$$\left\| f - \sum_{\alpha \in \mathbb{Z}^s} \sum_{j=1}^r \langle f, d^n g_j(M^n \cdot - \alpha)\rangle \phi_j(M^n \cdot - \alpha) \right\|_2 \le C(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N}, \tag{4.5}$$

where C is a constant independent of n (see [23]).

Let v be an element in V_k, and let

$$H(x) := \sum_{\alpha \in \mathbb{Z}^s} v(\alpha) h(x - \alpha), \quad x \in \mathbb{R}^s,$$

where h is a compactly supported continuous function on $\mathbb{R}^s$ such that the shifts of h are stable, and $D^\mu \hat{h}(2\pi\beta) = 0$ for all $|\mu| < k$ and $\beta \in \mathbb{Z}^s \setminus \{0\}$. By our choice of H, we have

$$D^\mu(y\hat{H})(2\pi\beta) = D^\mu(y\hat{v}\hat{h})(2\pi\beta) = 0 \quad \forall\, |\mu| < k \text{ and } \beta \in \mathbb{Z}^s.$$

Let $\Psi := \Phi + H$. Taking (3.2) into account, we obtain

$$D^\mu(y\hat{\Psi})(2\pi\beta) = D^\mu(y\hat{\Phi})(2\pi\beta) = \delta_{0\mu}\delta_{0\beta} \quad \forall\, |\mu| < k \text{ and } \beta \in \mathbb{Z}^s.$$

Suppose $\Psi = (\psi_1, \dots, \psi_r)^T$. Let P_Ψ be the quasi-projection operator given by

$$P_\Psi f := \sum_{\alpha \in \mathbb{Z}^s} \sum_{j=1}^r \langle f, g_j(\cdot - \alpha)\rangle \psi_j(\cdot - \alpha), \quad f \in L_2(\mathbb{R}^s).$$

Then P_Ψ also reproduces all polynomials of degree at most $k-1$.

For $n = 1, 2, \ldots$, let c_n be the sequence of $r \times r$ matrices given by

$$c_n(\alpha) := \big(\langle \phi_j, d^n g_m(M^n \cdot - \alpha)\rangle\big)_{1\le j,m\le r}, \quad \alpha \in \mathbb{Z}^s.$$

Suppose $\Phi \in (\mathrm{Lip}(\lambda, L_2(\mathbb{R}^s)))^r$ and $0 < \lambda < k$. Since $P_\Phi q = P_\Psi q = q$ for all $q \in \Pi_{k-1}$, the estimate in (4.5) tells us that there exists a positive constant C_1 such that

$$\Big\|\Phi - \sum_{\alpha\in\mathbb{Z}^s} c_n(\alpha)\Phi(M^n\cdot - \alpha)\Big\|_2 \le C_1(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N} \tag{4.6}$$

and

$$\Big\|\Phi - \sum_{\alpha\in\mathbb{Z}^s} c_n(\alpha)\Psi(M^n\cdot - \alpha)\Big\|_2 \le C_1(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N}. \tag{4.7}$$

It follows from (4.1) and (4.6) that

$$\Big\|\sum_{\alpha\in\mathbb{Z}^s} (a_n - c_n)(\alpha)\Phi(M^n\cdot - \alpha)\Big\|_2 \le C_1(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N}.$$

Since Φ is stable, we deduce from the above estimate that

$$\|a_n - c_n\|_2 \le C_2 d^{(1/2-\lambda/s)n} \quad \forall\, n \in \mathbb{N},$$

where C_2 is a constant independent of n. This in connection with (4.7) gives

$$\Big\|\Phi - \sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)\Psi(M^n\cdot - \alpha)\Big\|_2 \le C_3(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N},$$

where C_3 is a constant independent of n. However, $\Psi = \Phi + H$. So the above inequality together with (4.1) yields

$$\Big\|\sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)H(M^n\cdot - \alpha)\Big\|_2 \le C_3(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N}.$$

However,

$$\begin{aligned}\sum_{\alpha\in\mathbb{Z}^s} a_n(\alpha)H(M^n\cdot - \alpha) &= \sum_{\alpha\in\mathbb{Z}^s}\sum_{\beta\in\mathbb{Z}^s} a_n(\alpha)v(\beta)h(M^n\cdot - \alpha - \beta)\\ &= \sum_{\gamma\in\mathbb{Z}^s} (a_n * v)(\gamma)h(M^n\cdot - \gamma).\end{aligned}$$

Consequently,

$$\Big\|\sum_{\gamma\in\mathbb{Z}^s} (a_n * v)(\gamma)h(M^n\cdot - \gamma)\Big\|_2 \le C_3(d^{-1/s})^{\lambda n} \quad \forall\, n \in \mathbb{N}.$$

Since the shifts of h are stable, there exists a constant C_v such that (4.3) holds true. □

5. Spectral radius. To apply the results in the previous section to smoothness analysis of refinable vectors of functions, we need to evaluate the limit

$$\lim_{n\to\infty} \|a_n * v\|_2^{1/n}.$$

In this section we shall show that this limit can be evaluated as the spectral radius of a certain (finite) matrix. Some ideas in [13], [22], and [25] will be employed in our discussion.

For $u, v \in \ell_0^r(\mathbb{Z}^s)$, we define $u \odot v^T$ as follows:

$$u \odot v^T(\alpha) := \sum_{\beta\in\mathbb{Z}^s} u(\alpha+\beta)\overline{v(\beta)}^T, \quad \alpha \in \mathbb{Z}^s.$$

Let $u_n := a_n * u$ and $v_n := a_n * v$, where a_n $(n = 1, 2, \ldots)$ are the sequences given in (4.2). Moreover, let $w := \mathrm{vec}(u \odot v^T)$ and $w_n := \mathrm{vec}(u_n \odot v_n^T)$. For $\alpha \in \mathbb{Z}^s$, we have

$$u_n \odot v_n^T(\alpha) = \sum_{\beta\in\mathbb{Z}^s} u_n(\alpha+\beta)\overline{v_n(\beta)}^T = \sum_{\beta\in\mathbb{Z}^s}\sum_{\gamma\in\mathbb{Z}^s}\sum_{\eta\in\mathbb{Z}^s} a_n(\alpha+\beta-\gamma)u(\gamma)\overline{v(\eta)}^T\overline{a_n(\beta-\eta)}^T.$$

It follows by (1.5) that

$$w_n(\alpha) = \sum_{\gamma\in\mathbb{Z}^s}\left(\sum_{\beta\in\mathbb{Z}^s}\overline{a_n(\beta)}\otimes a_n(\alpha+\beta-\gamma)\right)\left(\sum_{\eta\in\mathbb{Z}^s}\mathrm{vec}\left(u(\gamma+\eta)\overline{v(\eta)}^T\right)\right).$$

Let b_n $(n = 1, 2, \ldots)$ be the sequences given by

$$b_n(\alpha) := \frac{1}{d^n}\sum_{\beta\in\mathbb{Z}^s}\overline{a_n(\beta)}\otimes a_n(\alpha+\beta), \quad \alpha\in\mathbb{Z}^s. \tag{5.1}$$

Consequently,

$$\mathrm{vec}\left((a_n*u)\odot(a_n*v)^T\right) = d^n b_n * \left(\mathrm{vec}\,(u\odot v^T)\right). \tag{5.2}$$

Clearly, b_1 is the same as the sequence b given in (1.6). Furthermore, for $n > 1$, it follows from (5.1) and (4.2) that

$$\begin{aligned} d^n b_n(\alpha) &= \sum_{\beta\in\mathbb{Z}^s}\sum_{\eta\in\mathbb{Z}^s}\sum_{\gamma\in\mathbb{Z}^s}\left(\overline{a_{n-1}(\eta)a(\beta-M\eta)}\right)\otimes\left(a_{n-1}(\gamma)a(\alpha+\beta-M\gamma)\right)\\ &= \sum_{\gamma\in\mathbb{Z}^s}\left(\sum_{\eta\in\mathbb{Z}^s}\overline{a_{n-1}(\eta)}\otimes a_{n-1}(\eta+\gamma)\right)\left(\sum_{\beta\in\mathbb{Z}^s}\overline{a(\beta)}\otimes a(\alpha+\beta-M\gamma)\right). \end{aligned}$$

It follows that

$$b_n(\alpha) = \sum_{\gamma\in\mathbb{Z}^s} b_{n-1}(\gamma)b(\alpha-M\gamma), \quad \alpha\in\mathbb{Z}^s. \tag{5.3}$$

THEOREM 5.1. *Let $a \in \ell_0^{r\times r}(\mathbb{Z}^s)$, and let a_n $(n = 1, 2, \ldots)$ be given as in* (4.2). *Then for $v \in \ell_0^r(\mathbb{Z}^s)$,*

$$\lim_{n\to\infty}\|a_n*v\|_2^{1/n} = \sqrt{d\rho(T_b|_W)},$$

where b is the sequence given in (1.6) *and W is the minimal invariant subspace of the transition operator T_b generated by $w := vec(v \odot v^T)$.*

Proof. We first establish the following identity for $w \in \ell_0^{r^2}(\mathbb{Z}^s)$:

$$T_b^n w(\alpha) = \sum_{\beta \in \mathbb{Z}^s} b_n(M^n\alpha - \beta) w(\beta), \quad \alpha \in \mathbb{Z}^s. \tag{5.4}$$

This will be proved by induction on n. By the definition of the transition operator T_b, (5.4) is true for $n = 1$. Suppose $n > 1$ and that (5.4) is valid for $n-1$. For $\alpha \in \mathbb{Z}^s$ we have

$$\begin{aligned} T_b^n w(\alpha) &= \sum_{\beta \in \mathbb{Z}^s} b_{n-1}(M^{n-1}\alpha - \beta)(T_b w)(\beta) \\ &= \sum_{\beta \in \mathbb{Z}^s} \sum_{\gamma \in \mathbb{Z}^s} b_{n-1}(M^{n-1}\alpha - \beta) b(M\beta - \gamma) w(\gamma) \\ &= \sum_{\gamma \in \mathbb{Z}^s} \left[\sum_{\beta \in \mathbb{Z}^s} b_{n-1}(\beta) b(M^n\alpha - \gamma - M\beta) \right] w(\gamma) \\ &= \sum_{\gamma \in \mathbb{Z}^s} b_n(M^n\alpha - \gamma) w(\gamma), \end{aligned}$$

where (5.3) has been used to derive the last equality. This completes the induction procedure.

Let v be an element in $\ell_0^r(\mathbb{Z}^s)$ and let $w := \mathrm{vec}(v \odot v^T)$. For $n \in \mathbb{N}$, let $v_n := a_n * v$ and $w_n := \mathrm{vec}\,(v_n \odot v_n^T)$. Then $w_n = d^n b_n * w$ by (5.2). This, together with (5.4), yields

$$d^n T_b^n w(\alpha) = d^n b_n * w(M^n\alpha) = w_n(M^n\alpha), \quad \alpha \in \mathbb{Z}^s.$$

Since $w_n = \mathrm{vec}\,(v_n \odot v_n^T)$, we have

$$d^n \|T_b^n w\|_\infty \le \|w_n\|_\infty \le \|v_n\|_2^2.$$

On the other hand,

$$d^n T_b^n w(0) = w_n(0) = \mathrm{vec}\left(\sum_{\beta \in \mathbb{Z}^s} v_n(\beta) \overline{v_n(\beta)}^T \right).$$

Consequently,

$$\|v_n\|_2^2 \le r d^n \|T_b^n w\|_\infty \le r \|v_n\|_2^2.$$

Therefore,

$$\lim_{n\to\infty} \|a_n * v\|_2^{2/n} = \lim_{n\to\infty} \|v_n\|_2^{2/n} = d \lim_{n\to\infty} \|T_b^n w\|_\infty^{1/n} = d\rho(T_b|_W),$$

where W is the minimal invariant subspace of T_b generated by w. ∎

Now suppose a satisfies the sum rules of order k with respect to $y \in \mathbb{T}^{1\times r}(\mathbb{R}^s)$. Let

$$W_k := \mathrm{span}\{\mathrm{vec}\,(u \odot v^T) : u, v \in V_k\},$$

where V_k is the linear space given in (3.7). By Lemma 3.3, V_k is invariant under the transition operator T_a. We claim that W_k is invariant under the transition operator T_b. Suppose $w = \operatorname{vec}(u \odot v^T)$, where $u, v \in V_k$. By (5.2) we have

$$T_b w(\alpha) = b*w(M\alpha) = \frac{1}{d} \operatorname{vec}\big((a*u) \odot (a*v)^T\big)(M\alpha), \quad \alpha \in \mathbb{Z}^s.$$

Let E be a complete set of representatives of the distinct cosets of $\mathbb{Z}^s/M\mathbb{Z}^s$. Then we have

$$\begin{aligned}\big((a*u) \odot (a*v)^T\big)(M\alpha) &= \sum_{\beta\in\mathbb{Z}^s} (a*u)(M\alpha+\beta)\overline{(a*v)(\beta)}^T \\ &= \sum_{\eta\in E}\sum_{\gamma\in\mathbb{Z}^s} (a*u)(M\alpha+M\gamma+\eta)\overline{(a*v)(M\gamma+\eta)}^T.\end{aligned}$$

However,

$$(a*u)(M\alpha+M\gamma+\eta) = T_a\big(u(\cdot+\eta)\big)(\alpha+\gamma), \quad \alpha \in \mathbb{Z}^s.$$

We observe that V_k is shift-invariant, i.e., $u \in V_k$ implies $u(\cdot+\eta) \in V_k$ for $\eta \in \mathbb{Z}^s$. Since V_k is invariant under T_a, we see that $u_\eta := T_a(u(\cdot+\eta))$ lies in V_k. Similarly, $v_\eta := T_a(v(\cdot+\eta))$ lies in V_k. Consequently,

$$\big((a*u) \odot (a*v)^T\big)(M\alpha) = \sum_{\eta\in E}\sum_{\gamma\in\mathbb{Z}^s} u_\eta(\alpha+\gamma)\,\overline{v_\eta(\gamma)}^T, \quad \alpha \in \mathbb{Z}^s.$$

Therefore,

$$T_b w = \frac{1}{d}\sum_{\eta\in E} \operatorname{vec}(u_\eta \odot v_\eta^T) \in W_k.$$

This shows that W_k is invariant under T_b.

Let us consider the special case $k = 1$. Suppose a satisfies the basic sum rule with respect to $y_0 \neq 0$. In this case, it is easily seen that

$$V_1 = \left\{ v \in \ell_0^r(\mathbb{Z}^s) : y_0 \sum_{\alpha\in\mathbb{Z}^s} v(\alpha) = 0 \right\}$$

and

$$W_1 = \left\{ w \in \ell_0^{r^2}(\mathbb{Z}^s) : (\overline{y_0} \otimes y_0) \sum_{\alpha\in\mathbb{Z}^s} w(\alpha) = 0 \right\}.$$

It was shown in [25] and [4] that the cascade algorithm associated with mask a converges in the L_2 norm if $\lim_{n\to\infty} \|a_n * v\|_2 = 0$ for each $v \in V_1$. Conversely, suppose $\Phi \in (L_2(\mathbb{R}^s))^r$ is a compactly supported solution to the refinement equation (1.1) and Φ is stable. Then the proof of Theorem 4.1 tells us that $\lim_{n\to\infty} \|a_n * v\|_2 = 0$ for each $v \in V_1$. Thus, we have the following result.

THEOREM 5.2. *Let $b \in \ell_0^{r^2}(\mathbb{Z}^s)$ be defined as in* (1.6). *If a satisfies the basic sum rule, and if $\rho(T_b|_{W_1}) < 1$, then there exists a compactly supported solution $\Phi \in (L_2(\mathbb{R}^s))^r$ to the refinement equation* (1.1) *with a being the mask. Conversely, if*

$\Phi \in (L_2(\mathbb{R}^s))^r$ is a compactly supported solution to the refinement equation (1.1) *with a being the mask, and if Φ is stable, then a satisfies the basic sum rule and $\rho(T_b|_{W_1}) < 1$.*

We conclude this section with the following characterization of the critical exponent of Φ in terms of the mask.

THEOREM 5.3. *Let Φ be a $1\times r$ vector of compactly supported functions in $L_2(\mathbb{R}^s)$ satisfying the refinement equation* (1.1). *Suppose the mask a satisfies the sum rules of order k and the matrix M is isotropic. Then*

$$\lambda(\Phi) \ge -\big(\log_d \rho\big(T_b|_{W_k}\big)\big)\, s/2.$$

The equality holds true in the above relation if, in addition, Φ is stable and k is the largest integer such that $\mathbb{S}(\Phi) \supset \Pi_{k-1}$.

Proof. Let $v \in V_k$. Then $w := \mathrm{vec}(v \odot v^T)$ lies in W_k. By Theorem 5.1 we have

$$\lim_{n\to\infty} \|a_n * v\|_2^{2/n} \le d\rho(T_b|_{W_k}).$$

Write ρ_k for $\rho(T_b|_{W_k})$. For $\varepsilon > 0$, there exists a positive constant C such that

$$\|a_n * v\|_2 \le C d^{n/2} (\rho_k + \varepsilon)^{n/2} \quad \forall\, n \in \mathbb{N}.$$

Let

$$\lambda_\varepsilon := -(\log_d(\rho_k + \varepsilon))s/2.$$

Then the above inequality can be rewritten as

$$\|a_n * v\|_2 \le C d^{(1/2 - \lambda_\varepsilon/s)n} \quad \forall\, n \in \mathbb{N}.$$

By Theorem 4.1, Φ lies in $(\mathrm{Lip}(\lambda_\varepsilon, L_2(\mathbb{R}^s)))^r$. Hence,

$$\lambda(\Phi) \ge \lambda_\varepsilon = -(\log_d(\rho_k + \varepsilon))s/2.$$

However, $\varepsilon > 0$ could be arbitrarily small. Therefore, we obtain

$$\lambda(\Phi) \ge -(\log_d \rho_k)s/2.$$

Now suppose Φ is stable and k is the largest integer such that $\mathbb{S}(\Phi) \supset \Pi_{k-1}$. We must have $\lambda(\Phi) \le k$, for otherwise $\lambda(\Phi) > k$ would imply $\mathbb{S}(\Phi) \supset \Pi_k$ (see [40] and [4]). Since Φ is stable and $\mathbb{S}(\Phi) \supset \Pi_{k-1}$, the corresponding mask a satisfies the sum rules of order k with respect to some $y \in \mathbb{T}^{1\times r}(\mathbb{R}^s)$. Let $\lambda_\varepsilon := \lambda(\Phi) - \varepsilon$, where $0 < \varepsilon < \lambda(\Phi)$. Then Φ lies in $(\mathrm{Lip}(\lambda_\varepsilon, L_2(\mathbb{R}^s)))^r$ and $k > \lambda_\varepsilon$. Note that $\rho(T_b|_{W_k}) = \rho(T_b|_{W_k \cap \ell^{r^2}(K)})$, where K is the set $\mathbb{Z}^s \cap \sum_{n=1}^\infty M^{-n}(\mathrm{supp} b)$. Since $W_k \cap \ell^{r^2}(K)$ is finite dimensional, we can find $u_j, v_j \in V_k$, $j = 1, \ldots, N$, such that

$$W_k \cap \ell^{r^2}(K) \subseteq \mathrm{span}\big\{\mathrm{vec}(u_j \odot v_j^T) : j = 1, \ldots, N\big\}.$$

Let $w_j := \mathrm{vec}(u_j \odot v_j^T) : j = 1, \ldots, N$. We have

$$\rho_k = \rho(T_b|_{W_k \cap \ell^{r^2}(K)}) \le \max_{1\le j\le N} \Big\{ \lim_{n\to\infty} \|T_b^n w_j\|_\infty^{1/n} \Big\}.$$

By (5.2) we have

$$d^n \|b_n * w_j\|_\infty \le \|a_n * u_j\|_2 \, \|a_n * v_j\|_2.$$

Thus, from the proof of Theorem 5.1 we obtain

$$\lim_{n\to\infty} \|T_b^n w_j\|_\infty^{1/n} \le d^{-1} \Big(\lim_{n\to\infty} \|a_n * u_j\|_2^{1/n} \Big) \Big(\lim_{n\to\infty} \|a_n * v_j\|_2^{1/n} \Big).$$

Since $\Phi \in (\mathrm{Lip}(\lambda_\varepsilon, L_2(\mathbb{R}^s)))^r$ with $\lambda_\varepsilon < k$, and since Φ is stable, by Theorem 4.1 we have

$$\lim_{n\to\infty} \|a_n * u_j\|_2^{1/n} \le d^{1/2-\lambda_\varepsilon/s} \quad \text{and} \quad \lim_{n\to\infty} \|a_n * v_j\|_2^{1/n} \le d^{1/2-\lambda_\varepsilon/s}.$$

Therefore,

$$\rho_k \le d^{-1} d^{1/2-\lambda_\varepsilon/s} d^{1/2-\lambda_\varepsilon/s} = d^{-2\lambda_\varepsilon/s}.$$

It follows that

$$\lambda(\Phi) - \varepsilon = \lambda_\varepsilon \le -(\log_d \rho_k) s/2.$$

However, $\varepsilon > 0$ could be arbitrarily small. We conclude that $\lambda(\Phi) \le -(\log_d \rho_k) s/2$. This completes the proof. □

6. Invariant subspaces. In the previous section, we reduced calculation of the critical exponent of a refinable vector of functions to the spectral radius of the transition operator T_b restricted to W_k. The purpose of this section is to find a basis for $W_k^\perp$. In this way, we will be able to apply Lemma 2.3 to calculate $\rho(T_b|_{W_k})$.

Let $y \in \mathbb{T}^{1\times r}(\mathbb{R}^s)$. Recall that

$$V_k = \{v \in \ell_0^r(\mathbb{Z}^s) : D^\mu(y\hat{v})(0) = 0 \ \forall\, |\mu| < k\}$$

and

$$W_k = \mathrm{span}\{\mathrm{vec}(u \odot v^T) : u \in V_k, v \in V_k\}. \tag{6.1}$$

For $\xi \in \mathbb{R}^s$, we have

$$\begin{aligned}(u \odot v^T)\hat{}\,(\xi) &= \sum_{\alpha\in\mathbb{Z}^s} (u \odot v^T)(\alpha) e^{-i\alpha\cdot\xi} = \sum_{\alpha\in\mathbb{Z}^s} \sum_{\beta\in\mathbb{Z}^s} u(\alpha+\beta)\overline{v(\beta)}^T e^{-i(\alpha+\beta)\cdot\xi} e^{i\beta\cdot\xi} \\ &= \sum_{\beta\in\mathbb{Z}^s} \left(\sum_{\alpha\in\mathbb{Z}^s} u(\alpha+\beta) e^{-i(\alpha+\beta)\cdot\xi} \right) \overline{v(\beta)e^{-i\beta\cdot\xi}}^T = \hat{u}(\xi)\,\overline{\hat{v}(\xi)}^T.\end{aligned}$$

Let us first consider the special case $r = 1$ and $y = 1$. In this case, we claim that

$$W_k = \{w \in \ell_0(\mathbb{Z}^s) : D^\mu \hat{w}(0) = 0 \ \forall\, |\mu| < 2k\}.$$

Indeed, if $u, v \in V_k$, then $D^\nu \hat{u}(0) = D^\nu \hat{v}(0) = 0$ for all $|\nu| < k$. Hence, $D^\mu(\hat{u}\overline{\hat{v}})(0) = 0$ for all $|\mu| < 2k$. Conversely, suppose $w \in \ell_0(\mathbb{Z}^s)$ and that $h := \hat{w}$ satisfies $D^\mu h(0) = 0$ for all $|\mu| < 2k$. The following lemma tells us $w \in W_k$.

LEMMA 6.1. *Let h be a trigonometric polynomial on $\mathbb{R}^s$ such that $D^\mu h(0) = 0$ for all $|\mu| < 2k$. Then*

$$h \in \mathrm{span}\,\{g_1\overline{g_2} : g_1, g_2 \in \mathbb{T}(\mathbb{R}^s),\ D^\nu g_1(0) = D^\nu g_2(0) = 0 \ \forall\, |\nu| < k\}.$$

Proof. For $\beta \in \mathbb{Z}^s$ we use δ_β to denote the sequence on $\mathbb{Z}^s$ given by $\delta_\beta(\alpha) = 0$ for $\alpha \in \mathbb{Z}^s \setminus \{\beta\}$ and $\delta_\beta(\beta) = 1$. Let

$$V := \operatorname{span}\{\nabla^\mu \delta_\beta : |\mu| = 2k, \beta \in \mathbb{Z}^s\},$$

and let $U := V^\perp$. Suppose u is a polynomial sequence of degree at most $2k-1$. For $|\mu| = 2k$ we have

$$\langle u, \nabla^\mu \delta_\beta \rangle = \sum_{\alpha \in \mathbb{Z}^s} u(\alpha) \nabla^\mu \delta_\beta(-\alpha) = \sum_{\alpha \in \mathbb{Z}^s} \nabla^\mu u(\alpha) \delta_\beta(-\alpha) = 0.$$

Hence, u lies in U. Conversely, if $u \in U$, then $\langle u, \nabla^\mu \delta_\beta \rangle = 0$ for all $|\mu| = 2k$ and $\beta \in \mathbb{Z}^s$. It follows that $\langle \nabla^\mu u, \delta_\beta \rangle = 0$ for all $\beta \in \mathbb{Z}^s$. Therefore, $\nabla^\mu u = 0$ for all $|\mu| = 2k$. This shows that u is a polynomial sequence of degree at most $2k-1$.

Suppose $h(\xi) = \sum_{\alpha \in \mathbb{Z}^s} v(\alpha) e^{-i\alpha \cdot \xi}$, where $v \in \ell_0(\mathbb{Z}^s)$. If $D^\mu h(0) = 0$ for all $|\mu| < 2k$, then

$$\sum_{\alpha \in \mathbb{Z}^s} (-i\alpha)^\mu v(\alpha) = 0 \quad \forall\, |\mu| < 2k.$$

Consequently, $\langle u, v \rangle = 0$ for every polynomial sequence u of degree at most $2k-1$. This shows $v \in U^\perp = (V^\perp)^\perp = V$. Thus, $h = \hat{v}$ lies in $\operatorname{span}\{(\nabla^\mu \delta_\beta)\hat{} : |\mu| = 2k, \beta \in \mathbb{Z}^s\}$. The symbol of $\nabla^\mu \delta_\beta$ is

$$(1 - e^{-i\xi_1})^{\mu_1} \cdots (1 - e^{-i\xi_s})^{\mu_s} e^{-i\beta \cdot \xi}, \quad \xi = (\xi_1, \ldots, \xi_s) \in \mathbb{R}^s.$$

For $|\mu| = 2k$, this expression can be written as $g_1(\xi)\overline{g_2(\xi)}$, where g_1 and g_2 are trigonometric polynomials satisfying $D^\nu g_1(0) = D^\nu g_2(0) = 0$ for all $|\nu| < k$. ∎

The following lemma extends Lemma 6.1 to the general case.

LEMMA 6.2. *Suppose $y = (y_1, \ldots, y_r) \in \mathbb{T}^{1\times r}(\mathbb{R}^s)$ and $y(0) \neq 0$. Let*

$$G := \big\{ g \in \mathbb{T}^r(\mathbb{R}^s) : D^\nu(yg)(0) = 0 \quad \forall\, |\nu| < k \big\},$$

and let H be the set of those $r \times r$ matrices h of trigonometrical polynomials for which $D^\nu(y\,h)(0) = D^\nu(h\,\overline{y}^T)(0) = 0$ for all $|\nu| < k$ and $D^\mu(y\,h\,\overline{y}^T)(0) = 0$ for all $k \le |\mu| < 2k$. Then

$$H = \operatorname{span}\{g_1 \overline{g_2}^T : g_1, g_2 \in G\}.$$

Proof. We observe that both G and H are linear spaces. If $g_1, g_2 \in G$, then $h := g_1 \overline{g_2}^T$ satisfies

$$D^\nu(yh)(0) = D^\nu(y g_1 \overline{g_2}^T)(0) = 0 \quad \forall\, |\nu| < k$$

and

$$D^\nu(h\overline{y}^T)(0) = D^\nu(g_1 \overline{g_2}^T \overline{y}^T)(0) = D^\nu(y g_2 \overline{g_1}^T)(0) = 0 \quad \forall\, |\nu| < k.$$

Moreover,

$$D^\mu(yh\overline{y}^T)(0) = D^\mu(y g_1 \overline{g_2}^T \overline{y}^T)(0) = D^\mu\big((y g_1)(\overline{y g_2})^T\big)(0) = 0 \quad \forall |\mu| < 2k.$$

Hence, $h = g_1 \overline{g_2}^T \in H$ for all $g_1, g_2 \in G$.

Conversely, suppose $h = (h_{mn})_{1\le m,n\le r} \in H$. Then $D^\nu(yh)(0) = D^\nu(h\overline{y}^T)(0) = 0$ for all $|\nu| < k$. Consequently, $D^\nu(y\,(h_{1m},\dots,h_{rm})^T)(0) = D^\nu((h_{m1},\dots,h_{mr})\,\overline{y}^T)(0) = 0$ for each $m = 1,\dots,r$ and all $|\nu| < k$. Hence, $(h_{1m},\dots,h_{rm})^T \in G$ and $\overline{(h_{m1},\dots,h_{mr})}^T \in G$. Without loss of any generality, we may assume that $y_1(0) \neq 0$. Thus, for $m = 2,\dots,r$ we can find $u_m \in \mathbb{T}(\mathbb{R}^s)$ such that

$$D^\nu(y_1u_m + y_m)(0) = 0 \quad \forall\, |\nu| < k.$$

For $m = 2,\dots,r$, consider the vector $(u_m, 0,\dots,0,1,0,\dots,0)^T$, where 1 is in the mth position. In light of our choice of u_m, we have $(u_m, 0,\dots,0,1,0,\dots,0)^T \in G$. Let

$$h' := h - \sum_{m=2}^{r} (u_m, 0,\dots,0,1,0,\dots,0)^T (h_{m1},\dots,h_{mr}).$$

Recall that $\overline{(h_{m1},\dots,h_{mr})}^T \in G$. Therefore, h' lies in H. Moreover, for $m = 2,\dots,r$, the mth row of h' vanishes. Suppose the first row of h' is $(h'_{11}, h'_{12},\dots,h'_{1r})$. Since $h' \in H$, we have $(h'_{1m}, 0,\dots,0)^T \in G$ for $m = 1,\dots,r$. Let

$$h'' := h' - \sum_{m=2}^{r} (h'_{1m}, 0,\dots,0)^T \overline{(u_m, 0,\dots,0,1,0,\dots,0)}.$$

Then $h'' \in H$. All the entries except the $(1,1)$-entry of the matrix h'' are zero. Let h''_{11} be the $(1,1)$-entry of h''. Since $h'' \in H$, we have $D^\nu(y_1h''_{11})(0) = 0$ for all $|\nu| < k$. Moreover, $D^\mu(|y_1|^2h''_{11})(0) = D^\mu(y_1h''_{11}\overline{y_1})(0) = 0$ for $k \le |\mu| < 2k$, but $y_1(0) \neq 0$. Hence, it follows that $D^\mu(h''_{11})(0) = 0$ for all $|\mu| < 2k$. By Lemma 6.1,

$$h''_{11} \in \operatorname{span}\{f_1\overline{f_2} : f_1, f_2 \in \mathbb{T}(\mathbb{R}^s), D^\nu f_1(0) = D^\nu f_2(0) = 0\ \forall\, |\nu| < k\}.$$

If $f_1, f_2 \in \mathbb{T}(\mathbb{R}^s)$ satisfy $D^\nu f_1(0) = D^\nu f_2(0) = 0$ for all $|\nu| < k$, then $g_1 := (f_1, 0,\dots,0)^T$ and $g_2 := (f_2, 0,\dots,0)^T$ belong to G. This shows that

$$h'' \in \operatorname{span}\{g_1\overline{g_2}^T : g_1, g_2 \in G\}.$$

Therefore, h itself lies in $\operatorname{span}\{g_1\overline{g_2}^T : g_1, g_2 \in G\}$. □

Since $(u \odot v^T)\hat{} = \hat{u}\bar{\hat{v}}^T$, we have

$$\operatorname{span}\{(u \odot v^T)\hat{} : u \in V_k, v \in V_k\} = \operatorname{span}\{\hat{u}\bar{\hat{v}}^T : u \in V_k, v \in V_k\}.$$

By Lemma 6.2, $w \in W_k$ if and only if $\hat{w} = \operatorname{vec}(h)$ for some h satisfying the following conditions:

$$D^\mu(yh)(0) = D^\mu(h\overline{y}^T)(0) = 0\ \forall\, |\mu| < k \quad \text{and} \quad D^\mu(yh\overline{y}^T)(0) = 0\ \forall\, k \le |\mu| < 2k.$$

Let $\{t_1,\dots,t_r\}$ be a basis for $\mathbb{C}^{1\times r}$. It is easily seen that

$$D^\mu(yh)(0) = 0 \iff D^\mu(yht_m^T) = 0 \quad \forall\, m = 1,\dots,r.$$

Similarly,

$$D^\mu(h\overline{y}^T)(0) = 0 \iff D^\mu(t_mh\overline{y}^T) = 0 \quad \forall\, m = 1,\dots,r.$$

We observe that

$$\mathrm{vec}(yht_m^T) = (t_m \otimes y)\mathrm{vec}(h),\ \mathrm{vec}(t_m h\overline{y}^T) = (\overline{y} \otimes t_m)\mathrm{vec}(h),$$
$$\text{and } \mathrm{vec}(yh\overline{y}^T) = (\overline{y} \otimes y)\mathrm{vec}(h).$$

Therefore, $u \in W_k$ if and only if

$$D^\mu\big((t_m \otimes y)\hat{w}\big)(0) = D^\mu\big((\overline{y} \otimes t_m)\hat{w}\big)(0) = 0 \quad \forall\, |\mu| < k$$

and

$$D^\mu\big((\overline{y} \otimes y)\hat{w}\big)(0) = 0 \quad \forall\, k \le |\mu| < 2k.$$

By the Leibniz rule for differentiation we have

$$\frac{(-iD)^\mu\big((t_m \otimes y)\hat{w}\big)(0)}{\mu!} = \sum_{\nu \le \mu} \frac{(-iD)^{\mu-\nu}(t_m \otimes y)(0)}{(\mu-\nu)!}\, \frac{(-iD)^\nu \hat{w}(0)}{\nu!}.$$

However, $(-iD)^\nu \hat{w}(0) = \sum_{\alpha \in \mathbb{Z}^s} (-\alpha)^\nu w(\alpha)$. Hence,

$$\frac{(-iD)^\mu\big((t_m \otimes y)\hat{w}\big)(0)}{\mu!} = \sum_{\alpha \in \mathbb{Z}^s} (t_m \otimes u_\mu)(-\alpha) w(\alpha) = \langle t_m \otimes u_\mu, w\rangle,$$

where u_μ ($|\mu| < k$) is given by

$$u_\mu := \sum_{\nu \le \mu} \frac{(-iD)^{\mu-\nu} y(0)}{(\mu-\nu)!}\, q_\nu,$$

and $q_\nu(\alpha) = \alpha^\nu/\nu!$, $\alpha \in \mathbb{Z}^s$. Thus,

$$D^\mu\big((t_m \otimes y)\hat{w}\big)(0) = 0 \Longleftrightarrow \langle t_m \otimes u_\mu, w\rangle = 0.$$

Similarly,

$$D^\mu\big((\overline{y} \otimes t_m)\hat{w}\big)(0) = 0 \Longleftrightarrow \langle u'_\mu \otimes t_m, w\rangle = 0,$$

where u'_μ is given by $u'_\mu(\alpha) = \overline{u_\mu(-\alpha)}$, $\alpha \in \mathbb{Z}^s$. Finally, for $|\mu| \le 2k$, let

$$\tilde{u}_\mu := \sum_{\nu \le \mu} \frac{(-iD)^{\mu-\nu}(\overline{y} \otimes y)(0)}{(\mu-\nu)!}\, q_\nu. \tag{6.2}$$

Then

$$D^\mu\big((\overline{y} \otimes y)\hat{w}\big)(0) = 0 \Longleftrightarrow \langle \tilde{u}_\mu, w\rangle = 0.$$

The above discussions are summarized in the following lemma.

LEMMA 6.3. *Suppose y is a $1 \times r$ vector of trigonometric polynomials on $\mathbb{R}^s$ such that $y(0) \neq 0$. Let $\{t_1, \ldots, t_r\}$ be a basis for $\mathbb{C}^{1\times r}$. If W_k is the linear space defined in* (6.1)*, then $W_k = U_k^\perp$, where*

$$U_k := \mathrm{span}\{t_m \otimes u_\mu, u'_\mu \otimes t_m : |\mu| < k \text{ and } m = 1, \ldots, r\} + \mathrm{span}\{\tilde{u}_\mu : k \le |\mu| < 2k\}.$$

Since W_k is invariant under the transition operator T_b, U_k is invariant under the subdivision operator S_b, by Lemma 2.1.

In the above lemma, $\{t_1, \dots, t_r\}$ could be any basis for $\mathbb{C}^{1\times r}$. However, a particular choice of bases will facilitate our study. Recall that $A(0) = \sum_{\alpha\in\mathbb{Z}^s} a(\alpha)/d$. Suppose

$$\operatorname{spec}(A(0)) = \{\eta_1, \eta_2, \dots, \eta_r\},$$

where $\eta_1 = 1$ and $\eta_j \neq 1$ for $j = 2, \dots, r$. We choose a basis $\{t_1, t_2, \dots, t_r\}$ for $\mathbb{C}^{1\times r}$ such that $t_1 A(0) = t_1$ and

$$t_m A(0) \in \operatorname{span}\{t_2, \dots, t_r\}, \quad m = 2, \dots, r.$$

Suppose

$$t_m A(0) = \sum_{n=1}^{r} \eta_{mn} t_n, \quad m = 1, \dots, r.$$

Then $\eta_{11} = 1$ and $\eta_{m1} = \eta_{1m} = 0$ for $m = 2, \dots, r$.

7. Spectral analysis. In this section we will establish Theorem 1.1 and other related results. For this purpose we shall first find the spectrum of the subdivision operator S_b restricted to U_k.

Let y be a $1 \times r$ vector of trigonometric polynomials on $\mathbb{R}^s$ such that $y(0) \neq 0$ and $y(0)A(0) = y(0)$. We choose a basis $\{t_1, t_2, \dots, t_r\}$ for $\mathbb{C}^{1\times r}$ such that $t_1 = y(0)$ and $t_m A(0) \in \operatorname{span}\{t_2, \dots, t_r\}$, $m = 2, \dots, r$. Recall that $q_\nu(\alpha) = \alpha^\nu/\nu!$, $\alpha \in \mathbb{Z}^s$, and

$$u_\mu = \sum_{\nu\le\mu} y_{\mu-\nu} q_\nu, \quad |\mu| < k,$$

where $y_{\mu-\nu} = (-iD)^{\mu-\nu} y(0)/(\mu-\nu)!$. In particular, $y_0 = y(0) = t_1$. Moreover,

$$u'_\mu = \sum_{\nu\le\mu} \overline{y_{\mu-\nu}} (-1)^{|\nu|} q_\nu.$$

For $j = 1, \dots, k$, let

$$U'_j := \operatorname{span}\{t_m \otimes u_\mu, u'_\mu \otimes t_m : |\mu| < j \text{ and } m = 1, \dots, r\}.$$

LEMMA 7.1. *The set*

$$\{\overline{t_m} \otimes u_\mu : |\mu| < k, m = 1, \dots, r\} \cup \{u'_\mu \otimes t_m : |\mu| < k, m = 2, \dots, r\} \tag{7.1}$$

forms a basis for U'_k.

Proof. For $|\mu| = 0$, we have

$$u'_0 \otimes y_0 = \overline{y_0} q_0 \otimes y_0 = \overline{y_0} \otimes y_0 q_0 = \overline{y_0} \otimes u_0.$$

For $|\mu| > 0$ we have

$$\begin{aligned} &u'_\mu \otimes y_0 - (-1)^{|\mu|} \overline{y_0} \otimes u_\mu \\ &= \sum_{\nu\le\mu} (-1)^{|\nu|} \overline{y_{\mu-\nu}} \otimes y_0 q_\nu - (-1)^{|\mu|} \sum_{\nu\le\mu} \overline{y_0} q_\nu \otimes y_{\mu-\nu} \\ &= \sum_{\nu<\mu} (-1)^{|\nu|} \overline{y_{\mu-\nu}} \otimes y_0 q_\nu - (-1)^{|\mu|} \sum_{\nu<\mu} \overline{y_0} q_\nu \otimes y_{\mu-\nu}. \end{aligned}$$

Note that

$$y_0 q_\nu = u_\nu - \sum_{\tau<\nu} y_{\nu-\tau} q_\tau \quad \text{and} \quad (-1)^{|\nu|}\overline{y_0} q_\nu = u'_\nu - \sum_{\tau<\nu} \overline{y_{\nu-\tau}}(-1)^{|\tau|} q_\tau.$$

Hence,

$$u'_\mu \otimes y_0 - (-1)^{|\mu|}\overline{y_0} \otimes u_\mu = \sum_{\nu<\mu}(-1)^{|\nu|}\overline{y_{\mu-\nu}} \otimes u_\nu - \sum_{\nu<\mu}(-1)^{|\mu-\nu|} u'_\nu \otimes y_{\mu-\nu} + J,$$

where

$$J := \sum_{\nu<\mu}\sum_{\tau<\nu}\Big[-(-1)^{|\nu|}\overline{y_{\mu-\nu}} \otimes y_{\nu-\tau} + (-1)^{|\mu-\nu+\tau|}\overline{y_{\nu-\tau}} \otimes y_{\mu-\nu}\Big] q_\tau.$$

It follows that

$$J = \sum_{\tau<\mu}\sum_{\tau<\nu<\mu}\Big[-(-1)^{|\nu|}\overline{y_{\mu-\nu}} \otimes y_{\nu-\tau} + (-1)^{|\mu-\nu+\tau|}\overline{y_{\nu-\tau}} \otimes y_{\mu-\nu}\Big] q_\tau.$$

Replacing ν by $\mu-\nu+\tau$ in the first part of the above inner sum, we obtain

$$\sum_{\tau<\nu<\mu} -(-1)^{|\nu|}\overline{y_{\mu-\nu}} \otimes y_{\nu-\tau} = \sum_{\tau<\nu<\mu} -(-1)^{|\mu-\nu+\tau|}\overline{y_{\nu-\tau}} \otimes y_{\mu-\nu}.$$

This shows $J = 0$. Therefore,

$$u'_\mu \otimes y_0 - (-1)^{|\mu|}\overline{y_0} \otimes u_\mu = \sum_{\nu<\mu}\Big[(-1)^{|\nu|}\overline{y_{\mu-\nu}} \otimes u_\nu - (-1)^{|\mu-\nu|} u'_\nu \otimes y_{\mu-\nu}\Big]. \tag{7.2}$$

In light of (7.2) we see that the set in (7.1) spans U'_k. Actually, this set is linearly independent. To justify our claim, we first make the following observation. Suppose $t_1, \dots, t_r$ are linearly independent $1\times r$ vectors and $w_1, \dots, w_r$ are $1\times r$ vectors. Then

$$w_1 \otimes t_1 + \cdots + w_r \otimes t_r = 0 \Longrightarrow w_1 = 0, \dots, w_r = 0. \tag{7.3}$$

Indeed, there exist $r \times 1$ vectors v_n $(n = 1, \dots, r)$ such that

$$t_m v_n = \delta_{mn}, \quad m, n = 1, \dots, r,$$

since $t_1, \dots, t_r$ are linearly independent. Let I be the $r \times r$ identity matrix. Then

$$(w_1 \otimes t_1 + \cdots + w_r \otimes t_r)(I \otimes v_n) = 0.$$

However, $(w_m \otimes t_m)(I \otimes v_n) = (w_m I) \otimes (t_m v_n) = w_m \delta_{mn}$. Hence, $w_n = 0$ for $n = 1, \dots, r$. This verifies (7.3).

Suppose $c_{j\mu}$ $(|\mu| < k,\ j = 1, \dots, r)$ and $c'_{j\mu}$ $(|\mu| < k,\ j = 2, \dots, r)$ are complex numbers such that

$$\sum_{|\mu|<k}\left[\sum_{j=1}^{r} c_{j\mu}\overline{t_j} \otimes u_\mu + \sum_{j=2}^{r} c'_{j\mu} u'_\mu \otimes t_j\right] = 0.$$

We wish to show that all $c_{j\mu} = 0$ and $c'_{j\mu} = 0$. In terms of the expressions of u_μ and u'_μ, we have

$$\sum_{|\mu|<k}\sum_{j=1}^{r}\sum_{\nu\le\mu} c_{j\mu}\overline{t_j} \otimes y_{\mu-\nu} q_\nu + \sum_{|\mu|<k}\sum_{j=2}^{r}\sum_{\nu\le\mu} c'_{j\mu}(-1)^{\nu}\overline{y_{\mu-\nu}} \otimes t_j q_\nu = 0.$$

As sequences on $\mathbb{Z}^s$, q_ν ($|\nu| < k$) are linearly independent. In the above sums, consider those terms involving q_ν with $|\nu| = k-1$. Then we have

$$\sum_{|\mu|=k-1} \left[\sum_{j=1}^{r} c_{j\mu} \overline{t_j} \otimes y_0 + \sum_{j=2}^{r} c'_{j\mu} (-1)^\mu \overline{y_0} \otimes t_j \right] q_\mu = 0.$$

It follows that

$$\left(\sum_{j=1}^{r} c_{j\mu} \overline{t_j} \right) \otimes t_1 + \sum_{j=2}^{r} c'_{j\mu} (-1)^\mu \overline{y_0} \otimes t_j = 0.$$

Since $t_1, t_2, \dots, t_r$ are linearly independent, by (7.3) we have

$$\sum_{j=1}^{r} c_{j\mu} \overline{t_j} = 0 \quad \text{and} \quad c'_{j\mu} (-1)^\mu \overline{y_0} = 0, \ j = 2, \dots, r.$$

Consequently, $c_{j\mu} = 0$ for all $|\mu| = k-1$ and $j = 1, \dots, r$, and $c'_{j\mu} = 0$ for all $|\mu| = k-1$ and $j = 2, \dots, r$. By using this argument repeatedly, we see that all $c_{j\mu} = 0$ and $c'_{j\mu} = 0$. Therefore, the set in (7.1) is linearly independent, so it forms a basis for U'_k. □

Recall that $\operatorname{spec}(M) = \{\sigma_1, \dots, \sigma_s\}$ and $\sigma^\mu = \sigma_1^{\mu_1} \cdots \sigma_s^{\mu_s}$ for $\mu = (\mu_1, \dots, \mu_s) \in \mathbb{Z}^s$.

LEMMA 7.2. *The spectrum of the subdivision operator S_b restricted to U'_k is*

$$\{\overline{\eta_m} \sigma^{-\mu} : m = 1, \dots, r, |\mu| < k\} \cup \{\eta_m \overline{\sigma^{-\mu}} : m = 2, \dots, r, |\mu| < k\}.$$

Proof. Suppose $|\mu| = j < k$. For $m = 1, \dots, r$ and $\alpha \in \mathbb{Z}^s$, we have

$$\begin{aligned} S_b(\overline{t_m} \otimes u_\mu)(\alpha) &= \sum_{\gamma \in \mathbb{Z}^s} (\overline{t_m} \otimes u_\mu)(\gamma) b(\alpha - M\gamma) \\ &= \frac{1}{d} \sum_{\gamma \in \mathbb{Z}^s} \sum_{\beta \in \mathbb{Z}^s} \big(\overline{t_m} \otimes u_\mu(\gamma)\big) \big(\overline{a(\beta)} \otimes a(\alpha - M\gamma + \beta)\big) \\ &= \frac{1}{d} \sum_{\beta \in \mathbb{Z}^s} (\overline{t_m a(\beta)}) \otimes \big((S_a u_\mu)(\alpha + \beta)\big). \end{aligned}$$

By Lemma 3.2, there are complex numbers $c_{\mu\nu}$ such that

$$S_a u_\mu = \sum_{|\nu|=j} c_{\mu\nu} u_\nu, \quad |\mu| = j.$$

Moreover, the spectrum of the matrix $(c_{\mu\nu})_{|\mu|=j, |\nu|=j}$ is $\{\sigma^{-\mu} : |\mu| = j\}$. For $|\nu| = j$, (3.8) tells us that

$$u_\nu(\alpha + \beta) - u_\nu(\alpha) = \sum_{|\tau|<j} h_{\nu\tau}(\beta) u_\tau(\alpha), \quad \alpha, \beta \in \mathbb{Z}^s,$$

where $h_{\nu\tau} \in \ell(\mathbb{Z}^s)$. Thus, for $\alpha, \beta \in \mathbb{Z}^s$ we have

$$(S_a u_\mu)(\alpha + \beta) = \sum_{|\nu|=j} c_{\mu\nu} u_\nu(\alpha + \beta) = \sum_{|\nu|=j} c_{\mu\nu} u_\nu(\alpha) + \sum_{|\nu|=j} \sum_{|\tau|<j} c_{\mu\nu} h_{\nu\tau}(\beta) u_\tau(\alpha).$$

Hence, there exists some element $w_{m\mu} \in U'_j$ such that

$$S_b(\overline{t_m} \otimes u_\mu) = \frac{1}{d} \sum_{\beta \in \mathbb{Z}^s} (\overline{t_m a(\beta)}) \otimes \left(\sum_{|\nu|=j} c_{\mu\nu} u_\nu \right) + w_{m\mu}.$$

However,

$$\frac{1}{d} \sum_{\beta \in \mathbb{Z}^s} (\overline{t_m a(\beta)}) = \overline{t_m A(0)} = \sum_{n=1}^{r} \overline{\eta_{mn} t_n}.$$

Therefore, for $m \in \{1, \dots, r\}$ and $|\mu| = j$ we have

$$S_b(\overline{t_m} \otimes u_\mu) = \sum_{n=1}^{r} \sum_{|\nu|=j} \big(\overline{\eta_{mn}} c_{\mu\nu}\big)(\overline{t_n} \otimes u_\nu) + w_{m\mu}. \tag{7.4}$$

Let Δ_j denote the index set $\{(m, \mu) : m = 1, \dots, r, |\mu| = j\}$. With an appropriate ordering, the matrix

$$\big(\overline{\eta_{mn}}\, c_{\mu\nu}\big)_{(m,\mu)\in\Delta_j, (n,\nu)\in\Delta_j}$$

can be viewed as the Kronecker product of the matrices $(\overline{\eta_{mn}})_{1\le m,n\le r}$ and $(c_{\mu\nu})_{|\mu|=j,|\nu|=j}$. Hence, its spectrum is

$$\{\overline{\eta_m}\, \sigma^{-\mu} : m = 1, \dots, r,\ |\mu| = j\}.$$

An analogous argument shows that, for $|\mu| = j$ and $m \in \{2, \dots, r\}$,

$$S_b(u'_\mu \otimes t_m) = \sum_{n=2}^{r} \sum_{|\nu|=j} (\eta_{mn}\, \overline{c_{\mu\nu}})(u'_\nu \otimes t_n) + w'_{m\mu}, \tag{7.5}$$

where $w'_{m\mu} \in U'_j$. Note that the spectrum of the matrix $(\eta_{mn})_{2\le m,n\le r}$ is $\{\eta_2, \dots, \eta_r\}$.

For $j = 1, \dots, k$, let $\tilde{S}_b^{(j)}$ denote the quotient linear operator induced by S_b on the quotient space U'_j/U'_{j-1}. Then (7.4) and (7.5) tell us that

$$\operatorname{spec}(\tilde{S}_b^{(j)}) = \{\overline{\eta_m}\sigma^{-\mu} : m = 1, \dots, r, |\mu| = j-1\} \cup \{\eta_m \overline{\sigma^{-\mu}} : m = 2, \dots, r, |\mu| = j-1\}.$$

Since

$$\operatorname{spec}(S_b|_{U'_k}) = \cup_{j=1}^{k} \operatorname{spec}(\tilde{S}_b^{(j)}),$$

the proof of the lemma is complete. □

By Lemma 6.3, we have $U_k = U'_k + \operatorname{span}\{\tilde{u}_\mu : k \le |\mu| < 2k\}$, where $\tilde{u}_\mu$ ($|\mu| < 2k$) are given by (6.2). As was done in section 3, it can be easily proved that U_k is the direct sum of U'_k and $\operatorname{span}\{\tilde{u}_\mu : k \le |\mu| < 2k\}$. Also, the set $\{\tilde{u}_\mu : k \le |\mu| < 2k\}$ is linearly independent. For $j = k, k+1, \dots, 2k$, let

$$U''_j := U'_k + \operatorname{span}\{\tilde{u}_\mu : k \le |\mu| < j\}.$$

In particular, $U''_k = U'_k$ and $U''_{2k} = U_k$.

LEMMA 7.3. *The spectrum of the subdivision operator S_b restricted to U_k is*

$$\{\overline{\eta_m}\sigma^{-\mu}, \eta_m\overline{\sigma^{-\mu}} : m = 2, \dots, r, |\mu| < k\} \cup \{\sigma^{-\mu} : |\mu| < 2k\}. \tag{7.6}$$

Proof. Suppose $|\mu| = j \in \{k, \dots, 2k-1\}$. Since U_k is invariant under S_b, there exist complex numbers $c_{\mu\nu}$ $(k \le |\nu| < 2k)$ and an element $w_\mu \in U_k'$ such that

$$S_b\tilde{u}_\mu = \sum_{k\le|\nu|<2k} c_{\mu\nu}\tilde{u}_\nu + w_\mu.$$

Since $S_b(\nabla_\gamma \tilde{u}_\mu) = \nabla_{M\gamma}(S_b\tilde{u}_\mu)$ for $\gamma \in \mathbb{Z}^s$, it follows that

$$S_b(\nabla^\tau \tilde{u}_\mu) = \sum_{k\le|\nu|<2k} c_{\mu\nu}(\nabla_{Me_1}^{\tau_1}\cdots\nabla_{Me_s}^{\tau_s})\tilde{u}_\nu + (\nabla_{Me_1}^{\tau_1}\cdots\nabla_{Me_s}^{\tau_s})w_\mu, \quad \tau \in \mathbb{N}_0^s.$$

We claim that $c_{\mu\nu} = 0$ for $|\nu| > j$. If this is not the case, then $N := \max\{|\nu| : c_{\mu\nu} \ne 0\} > j$. For $|\tau| = N$, we have $\nabla^\tau \tilde{u}_\mu = 0$ and $(\nabla_{Me_1}^{\tau_1}\cdots\nabla_{Me_s}^{\tau_s})w_\mu = 0$. Moreover, by (3.12) we have

$$(\nabla_{Me_1}^{\tau_1}\cdots\nabla_{Me_s}^{\tau_s})\tilde{u}_\nu = b_{\tau\nu}\tilde{u}_0 \quad \text{for } |\tau| = |\nu| = N,$$

where the matrix $(b_{\tau\nu})_{|\tau|=N,|\nu|=N}$ has $\{\sigma^\mu : |\mu| = N\}$ as its spectrum. Consequently,

$$\sum_{|\nu|=N} c_{\mu\nu}b_{\tau\nu} = 0 \quad \forall\, |\tau| = N. \tag{7.7}$$

Since the matrix $(b_{\tau\nu})_{|\tau|=N,|\nu|=N}$ is invertible, we obtain $c_{\mu\nu} = 0$ for all $|\nu| = N$. This contradiction justifies our claim. Therefore,

$$S_b\tilde{u}_\mu = \sum_{|\nu|=j} c_{\mu\nu}\tilde{u}_\nu + w_\mu', \tag{7.8}$$

where $w_\mu' \in U_j''$. For $|\tau| = j$, we deduce from (7.8) that

$$\delta_{\mu\tau}\tilde{u}_0 = S_b(\nabla^\tau\tilde{u}_\mu) = \sum_{|\nu|=j} c_{\mu\nu}(\nabla_{Me_1}^{\tau_1}\cdots\nabla_{Me_s}^{\tau_s})\tilde{u}_\nu = \sum_{|\nu|=j} c_{\mu\nu}b_{\tau\nu}\tilde{u}_0.$$

Hence, the spectrum of the matrix $(c_{\mu\nu})_{|\mu|=j,|\nu|=j}$ is $\{\sigma^{-\mu} : |\mu| = j\}$.

For $j = k+1, \dots, 2k$, let $\tilde{S}_b^{(j)}$ denote the quotient linear operator induced by S_b on the quotient space U_j''/U_{j-1}''. Then (7.8) tells us that

$$\operatorname{spec}(\tilde{S}_b^{(j)}) = \{\sigma^{-\mu} : |\mu| = j-1\}.$$

Since

$$\operatorname{spec}(S_b|_{U_k}) = \operatorname{spec}(S_b|_{U_{2k}''}) = \operatorname{spec}(S_b|_{U_k'}) \cup \Big(\cup_{j=k+1}^{2k}\operatorname{spec}(\tilde{S}_b^{(j)})\Big),$$

we conclude that the set in (7.6) is indeed the spectrum of S_b restricted to U_k. ∎

By Lemmas 7.3 and 2.3 we have the following formula:

$$\rho(T_b|_{W_k}) = \max\Big\{\, |\nu| : \nu \in \operatorname{spec}\big(b(M\alpha-\beta)\big)_{\alpha,\beta\in K} \setminus E_k \,\Big\},$$

where

$$E_k := \{\eta_j\overline{\sigma^{-\mu}}, \overline{\eta_j}\sigma^{-\mu} : |\mu| < k, j = 2, \dots, r\} \cup \{\sigma^{-\mu} : |\mu| < 2k\}.$$

This, together with Theorem 5.3, verifies Theorem 1.1.

Let B be the matrix $\big(b(M\alpha - \beta)\big)_{\alpha,\beta\in K}$. We say that B satisfies condition E if 1 is a simple eigenvalue of B and other eigenvalues of B are less than 1 in modulus. Suppose a satisfies the basic sum rule. Then W_1 is invariant under T_b and

$$\rho(T_b|_{W_1}) = \max\Big\{ |\nu| : \nu \in \operatorname{spec}(B) \setminus \{1, \eta_2, \dots, \eta_r, \overline{\eta_2}, \dots, \overline{\eta_r}\} \Big\}.$$

Thus, if B satisfies condition E, then $\rho(T_b|_{W_1}) < 1$, and hence the refinement equation (1.1) has a compactly supported solution $\Phi \in (L_2(\mathbb{R}^s))^r$ by Theorem 5.2. Conversely, if $\Phi \in (L_2(\mathbb{R}^s))^r$ is a compactly supported solution to the refinement equation (1.1), and if Φ is stable, then W_1 is invariant under T_b and $\rho(T_b|_{W_1}) < 1$. But, in this case, $|\eta_j| < 1$ for $j = 2, \dots, r$ (see [9] and [4]). Therefore, the matrix B satisfies condition E. This result was established in [42] for the case when the matrix M is 2 times the $s \times s$ identity matrix.

8. Examples. In this section we give three examples to illustrate the general theory. Our first example, taken from [14], is concerned with orthogonal multiwavelets.

Example 8.1. Let $r = 2$, $s = 1$, and $M = (2)$. Suppose $a \in \ell_0^2(\mathbb{Z})$ is supported on $\{0, 1, 2, 3\}$. Moreover,

$$a(0) = \frac{1}{10}\begin{bmatrix} 6 & 8\sqrt{2} \\ \frac{-1}{\sqrt{2}} & -3 \end{bmatrix}, \qquad a(1) = \frac{1}{10}\begin{bmatrix} 6 & 0 \\ \frac{9}{\sqrt{2}} & 10 \end{bmatrix},$$

$$a(2) = \frac{1}{10}\begin{bmatrix} 0 & 0 \\ \frac{9}{\sqrt{2}} & -3 \end{bmatrix}, \qquad a(3) = \frac{1}{10}\begin{bmatrix} 0 & 0 \\ \frac{-1}{\sqrt{2}} & 0 \end{bmatrix}.$$

We have

$$A(0) = [a(0) + a(1) + a(2) + a(3)]/2 = \frac{1}{10}\begin{bmatrix} 6 & 4\sqrt{2} \\ 4\sqrt{2} & 2 \end{bmatrix}.$$

The eigenvalues of $A(0)$ are $\eta_1 = 1$ and $\eta_2 = -1/5$. It can be easily verified that a satisfies the sum rules of order 2, but a does not satisfy the sum rules of order 3. Let b be the element in $\ell_0^4(\mathbb{Z})$ given by

$$b(\alpha) = \sum_{\beta\in\mathbb{Z}} a(\beta) \otimes a(\alpha + \beta)/2, \quad \alpha \in \mathbb{Z}.$$

Then b is supported on $\mathbb{Z}^2 \cap [-3, 3]$. Let B be the 28×28 matrix $(b(2\alpha - \beta))_{-3\le\alpha,\beta\le 3}$. The nonzero eigenvalues of B are

$$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{10}, -\frac{1}{10}, -\frac{1}{20}, -\frac{1}{20}, -\frac{1}{20}, -\frac{1}{20}, -\frac{1}{25}, -\frac{1}{50}, -\frac{1}{50}.$$

Thus, there exists a unique compactly supported solution $\Phi = (\phi_1, \phi_2)^T \in (L_2(\mathbb{R}))^2$ to the refinement equation

$$\Phi = \sum_{\alpha=0}^{3} a(\alpha)\Phi(2\cdot - \alpha)$$

subject to the condition $[\sqrt{2},1]\hat{\Phi}(0)=1$. The shifts of ϕ_1 and ϕ_2 are orthogonal (see [14]). Consequently, Φ is stable. Hence, we may apply Theorem 1.1 to obtain

$$\lambda(\Phi) = -(\log_2 \rho_2)/2,$$

where $\rho_2 = \max\{|\nu| : \nu \in \operatorname{spec}(B) \setminus E_2\}$ and

$$E_2 = \{1, 1/2, 1/4, 1/8, -1/5, -1/5, -1/10, -1/10\}.$$

Therefore, $\rho_2 = 1/8$ and $\lambda(\Phi) = -(\log_2 \rho_2)/2 = 3/2$. Note that

$$\max\big\{|\nu| : \nu \in \operatorname{spec}(B) \setminus \{(1/2)^\mu : \mu < 4\}\big\} = 1/5.$$

However, we have $\lambda(\Phi) = 3/2 > -(\log_2 1/5)/2$. □

Our second example is motivated by the study given in [37] on norm bounds for iterated transfer operators related to numerical solutions to partial differential equations.

Example 8.2. Let $r=2$, $s=2$, and $M = 2I_2$, where I_2 denotes the 2×2 identity matrix. Suppose $a \in \ell_0^2(\mathbb{Z}^2)$ is supported on $(\mathbb{Z}^2 \cap [0,5]^2) \setminus \{(4,0),(5,0),(4,1),(5,1),(0,5)\}$. Moreover, $a(0,0), a(1,0), a(2,0)$ are given by

$$\frac{1}{8}\begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix};$$

$a(0,1), a(1,1), a(2,1)$ are given by

$$\frac{1}{8}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 0 \\ 5 & 1 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix};$$

$a(0,2), a(1,2), a(2,2), a(3,2), a(4,2)$ are given by

$$\frac{1}{8}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 1 & -1 \\ 5 & 8 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 8 & 1 \\ 1 & 8 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix};$$

$a(0,3), a(1,3), a(2,3), a(3,3), a(4,3)$ are given by

$$\frac{1}{8}\begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 8 & 5 \\ -1 & 1 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 1 & 5 \\ 0 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix};$$

and $a(1,4), a(2,4), a(3,4), a(4,4)$ are given by

$$\frac{1}{8}\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \frac{1}{8}\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}.$$

We have

$$A(0) = \frac{1}{4}\sum_{\alpha\in\mathbb{Z}^2} a(\alpha) = \frac{1}{8}\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}.$$

The eigenvalues of $A(0)$ are $\eta_1 = 1$ and $\eta_2 = 1/4$. Moreover, $[1,1]A(0) = [1,1]$. It can be verified that the optimal order of sum rules satisfied by a is $k=2$. Let b be the element in $\ell_0^4(\mathbb{Z}^2)$ given by

$$b(\alpha) = \sum_{\beta\in\mathbb{Z}^2} a(\beta)\otimes a(\alpha+\beta)/4, \quad \alpha \in \mathbb{Z}^2.$$

Then b is supported in $[-5,5]^2$. Let B be the 484×484 matrix $(b(2\alpha-\beta))_{\alpha,\beta\in[-5,5]^2}$. The leading eigenvalues of B are

$$1,\ 1/2,\ 1/2,\ 1/4,\ 1/4,\ 1/4,\ 1/4,\ 1/4, 0.13129521, 0.13060779, \ldots.$$

Thus, there exists a unique compactly supported solution $\Phi\in(L_2(\mathbb{R}^2))^2$ to the refinement equation

$$\Phi=\sum_{\alpha\in\mathbb{Z}^2} a(\alpha)\Phi(2\cdot-\alpha)$$

subject to the condition $[1,1]\hat{\Phi}(0)=1$. By using the method in [18] we can show that Φ is stable. Hence, we may apply Theorem 1.1 to obtain

$$\lambda(\Phi)=-\log_4\rho_2,$$

where $\rho_2=\max\{|\nu|:\nu\in\operatorname{spec}(B)\setminus E_2\}$ and

$$E_2=\left\{\frac{1}{4},\frac{1}{4},\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8}\right\}\cup\left\{1,\frac{1}{2},\frac{1}{2},\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8},\frac{1}{8}\right\}.$$

Therefore, $\rho_2\approx 0.13129521$ and $\lambda(\Phi)=-\log_4\rho_2\approx 1.46436842$. □

Our third example is a refinable vector of functions with Hermite interpolation properties (see [16]). Such refinable functions are useful in computer aided geometric design.

Example 8.3. Let $r=3$, $s=2$, and

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

Clearly, the eigenvalues of M are $\sigma_1=1+i$ and $\sigma_2=1-i$, where i denotes the imaginary unit. Suppose $a\in\ell_0^3(\mathbb{Z}^2)$ is supported on $\{(0,0),(1,0),(0,1),(-1,0),(0,-1)\}$. Moreover, $a(0,0)$, $a(1,0)$, and $a(0,1)$ are given by

$$\begin{bmatrix}1 & 0 & 0\\ 0 & 1/2 & 1/2\\ 0 & -1/2 & 1/2\end{bmatrix},\quad \begin{bmatrix}1/4 & -3/4 & 0\\ 1/16 & -1/8 & 0\\ -1/16 & 1/8 & 0\end{bmatrix},\quad \begin{bmatrix}1/4 & 0 & -3/4\\ 1/16 & 0 & -1/8\\ 1/16 & 0 & -1/8\end{bmatrix},$$

and $a(-1,0)$, $a(0,-1)$ are given by

$$\begin{bmatrix}1/4 & 3/4 & 0\\ -1/16 & -1/8 & 0\\ 1/16 & 1/8 & 0\end{bmatrix},\quad \begin{bmatrix}1/4 & 0 & 3/4\\ -1/16 & 0 & -1/8\\ -1/16 & 0 & -1/8\end{bmatrix}.$$

We have

$$A(0)=\frac{1}{2}\sum_{\alpha\in\mathbb{Z}^2} a(\alpha)=\begin{bmatrix}1 & 0 & 0\\ 0 & 1/8 & 1/8\\ 0 & -1/8 & 1/8\end{bmatrix}.$$

The eigenvalues of $A(0)$ are $\eta_1=1$, $\eta_2=(1+i)/8$, and $\eta_3=(1-i)/8$. Moreover, $[1,0,0]A(0)=[1,0,0]$. It can be verified that the optimal order of sum rules satisfied by a is $k=4$ (see [16]). Let b be the element in $\ell_0^9(\mathbb{Z}^2)$ given by

$$b(\alpha)=\sum_{\beta\in\mathbb{Z}^2} a(\beta)\otimes a(\alpha+\beta)/2,\quad \alpha\in\mathbb{Z}^2.$$

Then b is supported on the set

$$\{(\alpha_1,\alpha_2)\in\mathbb{Z}^2 : -2\le\alpha_1-\alpha_2\le 2,\ -2\le\alpha_1+\alpha_2\le 2\}.$$

We observe that

$$\begin{aligned}K &:= \mathbb{Z}^2\cap\left(\sum_{n=1}^{\infty}M^{-n}(\operatorname{supp} b)\right)\\ &= \{(\alpha_1,\alpha_2)\in\mathbb{Z}^2 : |\alpha_1|\le 6,\ |\alpha_2|\le 6,\ |\alpha_1-\alpha_2|\le 8,\ |\alpha_1+\alpha_2|\le 8\}.\end{aligned}$$

The set K has exactly 129 points. Let B be the 1161×1161 matrix $(b(2\alpha-\beta))_{\alpha,\beta\in K}$. The first 27 eigenvalues of B (in terms of their absolute values) are

$$1,\ (1+i)/2,\ (1-i)/2,\ 1/2,\ i/2,\ -i/2,\ (1+i)/4,\ (1-i)/4,-(1+i)/4,\ (-1+i)/4,$$

$$1/4,\ -1/4,\ -1/4,\ i/4,\ -i/4,\ (1+i)/8,\ (1+i)/8,\ (1+i)/8,\ (1-i)/8,\ (1-i)/8,\ (1-i)/8,$$

$$-(1+i)/8,\ -(1+i)/8,\ (-1+i)/8,\ (-1+i)/8,\ 0.149024,\ 0.148796.$$

Thus, there exists a unique compactly supported solution $\Phi\in(L_2(\mathbb{R}^2))^2$ to the refinement equation

$$\Phi=\sum_{\alpha\in\mathbb{Z}^2}a(\alpha)\Phi(M\cdot-\alpha)$$

subject to the condition $[1,0,0]\hat{\Phi}(0)=1$. It is known that Φ is stable (see [16]). Hence, we may apply Theorem 1.1 to obtain

$$\lambda(\Phi)=-\log_2\rho_4,$$

where $\rho_4=\max\{|\nu| : \nu\in\operatorname{spec}(B)\setminus E_4\}$ and

$$E_4=\{\eta_2\overline{\sigma^{-\mu}},\overline{\eta_2}\sigma^{-\mu},\eta_3\overline{\sigma^{-\mu}},\overline{\eta_3}\sigma^{-\mu} : |\mu|<4\}\cup\{\sigma^{-\mu} : |\mu|<8\}.$$

We see that

$$\rho_4=\max\{|\nu| : \nu\in\operatorname{spec}(B)\setminus E_4\}\approx 0.149024.$$

Therefore, $\lambda(\Phi)=-\log_2\rho_4\approx 2.746387$. □

REFERENCES

[1] C. de Boor, R. A. DeVore, and A. Ron, *Approximation orders of FSI spaces in $L_2(\mathbb{R}^d)$*, Constr. Approx., 14 (1998), pp. 631–652.

[2] C. Cabrelli, C. Heil, and U. Molter, *Accuracy of lattice translates of several multidimensional refinable functions*, J. Approx. Theory, 95 (1998), pp. 5–52.

[3] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc., 93 (1991), no. 453.

[4] D. R. Chen, R.-Q. Jia, and S. D. Riemenschneider, *Convergence of vector subdivision schemes in Sobolev spaces*, Appl. Comput. Harmon. Anal., 12 (2002), pp. 128–149.

[5] A. Cohen and I. Daubechies, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591.

[6] A. Cohen, I. Daubechies, and G. Plonka, *Regularity of refinable function vectors*, J. Fourier Anal. Appl., 3 (1997), pp. 295–324.

[7] A. Cohen, K. Gröchenig, and L. F. Villemoes, *Regularity of multivariate refinable functions*, Constr. Approx., 15 (1999), pp. 241–255.
[8] W. Dahmen and C. A. Micchelli, *On the approximation order from certain multivariate spline spaces*, J. Austral. Math. Soc. Ser. B, 26 (1984), pp. 233–246.
[9] W. Dahmen and C. A. Micchelli, *Biorthogonal wavelet expansions*, Constr. Approx., 13 (1997), pp. 293–328.
[10] I. Daubechies and J. C. Lagarias, *Two-scale difference equations:* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.
[11] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.
[12] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.
[13] T. N. T. Goodman, C. A. Micchelli, and J. D. Ward, *Spectral radius formulas for subdivision operators*, in Recent Advances in Wavelet Analysis, L. L. Schumaker and G. Webb, eds., Academic Press, New York, 1994, pp. 335–360.
[14] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, *Fractal functions and wavelet expansions based on several complex variables*, J. Approx. Theory, 78 (1994), pp. 373–401.
[15] B. Han and R.-Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.
[16] B. Han, T. P.-Y. Yu, and B. Piper, *Multivariate Refinable Hermite Interpolants*, preprint, 2002.
[17] C. Heil, G. Strang, and V. Strela, *Approximation by translates of refinable functions*, Numer. Math., 73 (1996), pp. 75–94.
[18] T. A. Hogan and R.-Q. Jia, *Dependence relations among the shifts of a multivariate refinable functions*, Constr. Approx., 17 (2001), pp. 19–37.
[19] R.-Q. Jia, *Stability of the shifts of a finite number of functions*, J. Approx. Theory, 95 (1998), pp. 194–202.
[20] R.-Q. Jia, *Approximation properties of multivariate wavelets*, Math. Comp., 67 (1998), pp. 647–665.
[21] R.-Q. Jia, *Characterization of smoothness of multivariate refinable functions in Sobolev spaces*, Trans. Amer. Math. Soc., 351 (1999), pp. 4089–4112.
[22] R.-Q. Jia, *Convergence of vector subdivision schemes and construction of biorthogonal multiple wavelets*, in Advances in Wavelets, K.-S. Lau, ed., Springer-Verlag, Singapore, 1999, pp. 199–227.
[23] R.-Q. Jia, *Approximation with Scaled Shift-Invariant Spaces by Means of Quasi-Projection Operators*, manuscript.
[24] R.-Q. Jia and Q. T. Jiang, *Approximation power of refinable vectors of functions*, in Wavelet Analysis and Applications, D. Deng, D. Huang, R.-Q. Jia, W. Lin, and J. Wang, eds., AMS/IP Stud. Adv. Math. 25, AMS, Providence, RI, 2002, pp. 155–178.
[25] R.-Q. Jia, Q. T. Jiang, and Z. W. Shen, *Convergence of cascade algorithms associated with nonhomogeneous refinement equations*, Proc. Amer. Math. Soc., 129 (2000), pp. 415–427.
[26] R.-Q. Jia, K. S. Lau, and D. X. Zhou, *L_p solutions of refinement equations*, J. Fourier Anal. Appl., 7 (2001), pp. 143–167.
[27] R.-Q. Jia and C. A. Micchelli, *On linear independence of integer translates of a finite number of functions*, Proc. Edinburgh Math. Soc., 36 (1992), pp. 69-85.
[28] R.-Q. Jia, S. D. Riemenschneider, and D.-X. Zhou, *Smoothness of multiple refinable functions and multiple wavelets*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 1–28.
[29] R.-Q. Jia and S. R. Zhang, *Spectral properties of the transition operator associated to a multivariate refinement equation*, Linear Algebra Appl., 292 (1999), pp. 155–178.
[30] Q. Jiang, *On the regularity of matrix refinable functions*, SIAM J. Math. Anal., 29 (1998), pp. 1157–1176.
[31] Q. T. Jiang, *Multivariate matrix refinable functions with arbitrary matrix dilation*, Trans. Amer. Math. Soc., 351 (1999), pp. 2407–2438.
[32] Q. T. Jiang and P. Oswald, *On the analysis of $\sqrt{3}$-subdivision*, J. Comput. Appl. Math., to appear.
[33] J. L. Kelley and I. Namioka, *Linear Topological Spaces*, Springer-Verlag, New York, 1963.
[34] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, Orlando, FL, 1985.
[35] J. Lei, R.-Q. Jia, and E. W. Cheney, *Approximation from shift-invariant spaces by integral operators*, SIAM J. Math. Anal., 28 (1997), pp. 481–498.
[36] C. A. Micchelli and T. Sauer, *Regularity of multiwavelets*, Adv. Comput. Math., 7 (1997), pp. 455–545.

[37] P. Oswald, *On Norm Bounds for Iterated Intergrid Transfer Operators*, Arbeitsberichte der GMD 1079, GMD, St. Augustin, June 1997.
[38] P. Oswald and P. Schröder, *Composite Primal/Dual $\sqrt{3}$-subdivision Schemes*, preprint, 2002.
[39] G. Plonka, *Approximation order provided by refinable function vectors*, Constr. Approx., 13 (1997), pp. 221–244.
[40] A. Ron, *Smooth refinable functions provide good approximation orders*, SIAM J. Math. Anal., 28 (1997), pp. 731–748.
[41] A. Ron and Z. W. Shen, *The Sobolev regularity of refinable functions*, J. Approx. Theory, 106 (2000), pp. 185–225.
[42] Z. W. Shen, *Refinable function vectors*, SIAM J. Math. Anal., 29 (1998), pp. 235–250.
[43] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[44] G. Strang and G. Fix, *A Fourier analysis of the finite-element variational method*, in Constructive Aspects of Functional Analysis, (C.I.M.E., 1973), G. Geymonat, ed., Edizioni Cremonese, Rome, pp. 795–840.
[45] L. F. Villemoes, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 1433–1460.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1110–1135

ON DISCRETE MAXIMUM PRINCIPLES FOR LINEAR EQUATION SYSTEMS AND MONOTONICITY OF DIFFERENCE SCHEMES*

V. S. BORISOV†

Abstract. Discrete maximum principles for linear equation systems are discussed. A novel maximum principle for linear difference schemes is established. Sufficient conditions for an arbitrary linear scheme to satisfy the maximum principle are provided. Easily verifiable sufficient as well as necessary and sufficient conditions of nonsingularity for a diagonally dominant matrix, be it reducible or irreducible, are derived. Necessary and sufficient conditions for the validity of the maximum principle for explicit difference schemes are developed. The notion of submonotonicity for linear difference schemes as well as the notions of linear monotonicity and linear submonotonicity for nonlinear difference schemes are introduced and associated criteria developed. The developed approaches are demonstrated by examples of known linear and nonlinear difference schemes associated, in general, with the numerical analysis of systems of partial differential equations.

Key words. difference schemes, variational difference schemes, monotonicity, submonotonicity, grid connectedness, discrete maximum principle, stability

AMS subject classifications. 35B50, 34A30, 15A09, 39A70

PII. S0895479802409687

1. Introduction.

We consider a linear difference scheme written in the form

$$\mathbf{B} \cdot \mathbf{y} = \mathbf{q}, \quad \mathbf{y} \in Y, \ \mathbf{q} \in Q, \tag{1.1}$$

where $\mathbf{B}$ is a rectangular matrix, and Y and Q denote the linear vector spaces with the dimensionalities $N_Y = \dim(Y)$ and $N_Q = \dim(Q)$, respectively. In what follows it is assumed that $\mathbf{q}$ in (1.1) belongs to the image of Y under $\mathbf{B}$, i.e.,

$$\mathbf{q} \in \mathbf{B}(Y) \subseteq Q. \tag{1.2}$$

If $\mathbf{B}$, as well as $\mathbf{y}$ and $\mathbf{q}$ in (1.1), is partitioned, i.e., (1.1) can be written as

$$\sum_{j=1}^{M} \mathbf{B}_i^j \cdot \mathbf{y}_j = \mathbf{q}_i, \quad i = 1, 2, \ldots, K, \tag{1.3}$$

then (1.1) will be referred to as a vector difference scheme or, otherwise, as a scalar one. The null element in any linear space, as well as the number zero, will be denoted by the same symbol 0. The empty set will be denoted by the symbol $\emptyset$. If $\theta \subseteq \Theta$, then the symbol $\overline{\theta}$ denotes the complement of the subset θ with respect to the set Θ. The abbreviation "iff" will be used for "if and only if" (="if" in the definitions). The natural partial ordering is introduced for the vectors of a real space, i.e., $\mathbf{x} \equiv \{x_1, \ldots, x_N\}^T \leq \{y_1, \ldots, y_N\}^T \equiv \mathbf{y}$ iff $x_i \leq y_i$, $i = 1, 2, \ldots, N$.

Discrete maximum principles (or simply, *maximum principles*) are of importance for numerical analysis in mathematical modeling of physical, chemical, biological, hydrogeological, soil, economic, and technological, processes, among others (see, e.g.,

*Received by the editors June 16, 2002; accepted for publication (in revised form) by A. Wathen November 4, 2002; published electronically March 13, 2003.
http://www.siam.org/journals/simax/24-4/40968.html
†Environmental Hydrology and Microbiology, J. Blaustein Institute for Desert Research, Ben-Gurion University, Sede Boker Campus, 84990, Israel (viatslav@bgumail.bgu.ac.il).

[8], [19], [23], [25], [29], [33]). The maximum principle for difference schemes was established in [16] and has since been developed primarily for the spectrum of scalar difference schemes, e.g., the boundary maximum principle, the region maximum principle, the maximum principle for inverse column entries, and the maximum principle for the absolute values (see, e.g., [2], [4], [8], [15], [19], [23], [25], [29], [31], [33], and references therein). Usually, we find that most linear difference schemes being investigated are considered as special cases of the general formulation

$$b_i^j y_j = q_i, \quad i, j = 1, 2, \ldots, M, \tag{1.4}$$

where $b_i^j \in \mathbb{K}$ is an element of the square matrix $\mathbf{B} \equiv \{b_i^j\}$. Here and in what follows, repeating superscript indexes together with subscript indexes denote summation, and $\mathbb{K}$ denotes the field of either real ($\mathbb{R}$) or complex ($\mathbb{C}$) numbers. It is a well-known fact that a difference scheme approximating a differential equation is stable as well as, in general, it does not produce spurious [6] oscillations if a maximum principle holds for this scheme (see, e.g., [2], [14], [23], [24], [29], [32]). Such schemes are often termed monotone (see, e.g., [2], [13], [25], [27], [32]). Hereinafter, a scheme will be referred to as monotone iff the scheme satisfies a maximum principle.

In studies of the monotonicity of difference schemes, two approaches can be distinguished. The first one is purely algebraic [4], [29], [33, pp. 46-53], whereas the second makes further use of geometrical and topological elements [19], [25], [33, p. 44]. To show the interplay between these approaches let us consider a difference scheme written in the canonical form [25]; i.e., the equation associated with y_i is written as

$$c_i y_i = \sum_{j \in P_i} c_i^j y_j + q_i, \quad i = 1, 2, \ldots, M, \tag{1.5}$$

where i, j are grid nodes, c_i, c_i^j, y_i, $q_i \in \mathbb{R}$, and $P_i \subseteq \Omega \diagdown \{i\}$ denotes the punctured neighborhood (or simply, neighborhood) of the node i belonging to the grid $\Omega \equiv \{1\ , 2, \ldots, M\}$; i.e., P_i is the stencil [19] of the node $i \in \Omega$ excluding the i node. It is assumed [19], [25] that $P_i = \{j \in \Omega \mid c_i^j \neq 0\}$. Notice that we designate the grid nodes by their indices only. Such an approach is convenient for specifying completely the grid nodes within the framework of our investigation. A grid node is referred to as a boundary one [25] if at this node the grid function, y_i, is equal to a prescribed value. In such a case we write

$$y_i = f_i. \tag{1.6}$$

Equality (1.6) can be viewed as (1.5) for which $c_i = 1$, P_i becomes an empty set, and $q_i = f_i$. A node will be referred to as being interior if its neighborhood is not an empty set. Denoting by Ω_* the subset of interior nodes and by γ the subset of boundary nodes, we have $\Omega = \Omega_* + \gamma$. The grid Ω is referred to as stencil-connected [19] if there is a possibility of passing from any interior node $i \in \Omega_*$ to an arbitrary node $j \in \Omega$ using a sequence of neighborhoods; i.e., there exist nodes $i_1,\ i_2, \ldots, i_k$ such that $i_1 \in P_i,\ i_2 \in P_{i_1}, \ldots, i_k \in P_{i_{k-1}},\ j \in P_{i_k}$.

Let the grid Ω be stencil-connected with respect to the scheme (1.5), and let

$$c_i > 0,\ c_i^j > 0\ \forall j \in P_i,\ c_i - \sum_{j \in P_i} c_i^j \geq 0,\ q_i = 0 \quad \forall i \in \Omega_*; \tag{1.7}$$

then the scheme (1.5) satisfies the boundary maximum principle [4], [25], [33], [19].

Clearly the scheme (1.5) can be viewed as (1.4) for which $b_i^j = c_i$ if $j = i$, and $b_i^j = -c_i^j$ if $j \neq i$. Assuming that $M' < j \leq M$ for $j \in \gamma$, $\mathbf{y}' = \{y_1, y_2, \ldots, y_{M'}\}^T$, $\mathbf{y}'' = \{y_{M'+1}, \ldots, y_M\}^T$, $\mathbf{q}' = \{q_1, q_2, \ldots, q_{M'}\}^T = 0$, and $\mathbf{q}'' = \{q_{M'+1}, \ldots, q_M\}^T$, we can rewrite (1.5) in the form

$$\mathbf{B} \cdot \mathbf{y} = \begin{Bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ 0 & \mathbf{I} \end{Bmatrix} \begin{Bmatrix} \mathbf{y}' \\ \mathbf{y}'' \end{Bmatrix} = \begin{Bmatrix} 0 \\ \mathbf{q}'' \end{Bmatrix} = \mathbf{q}. \tag{1.8}$$

Hereinafter, $\mathbf{I}$ denotes the identity matrix.

Let us note that the directed graph [12], [33] associated with the matrix $\mathbf{B}$ of (1.8) consists of M distinct vertices that can be viewed as the nodes of the grid Ω. Then (in view of Definition 2.3, Theorem 2.4, and Lemma 3.19 in [33]) we can conclude that the matrix $\mathbf{B}$ of (1.8) is a nonsingular M-matrix and the matrix $\mathbf{B}_{11}$ of (1.8) is irreducible. Let us recall that a matrix is irreducible iff its associated graph is strongly connected [33]. Hence, the matrix $\mathbf{B}_{11}$ of (1.8) is irreducible iff the grid Ω is stencil-connected.

It is necessary to stress that stencil-connectedness is inherent in a rather specific part of difference schemes. In particular, such grid connectedness reflects features associated with the numerical solution of an elliptic partial differential equation (PDE) [25]. The schemes employed in the numerical solution of hyperbolic PDEs do not, in general, possess stencil-connectedness. Hence, it is quite important to develop a maximum principle for (1.8) subject to the condition that $\mathbf{B}_{11}$ is reducible.

By assuming that $\mathbf{B}_{11}$ in (1.8) is a nonsingular matrix, Smelov [29] proved that the maximum principle is valid even though $\mathbf{B}_{11}$ may be reducible. A similar situation takes place with regard to, in general, any sort of maximum principles; i.e., the matrices associated with difference schemes are assumed to be irreducible, or nonsingular, or both for validity of a maximum principle (see, e.g., [4], [19], [25], [29], [31], [33]). From the practical standpoint, nonsingularity of a matrix is still far from the desirable criterion for validity of a maximum principle, and hence there is a need for more easily verifiable criteria which do not involve determinants.

In addition to the restrictions associated with the grid connectedness, one restricts the coefficients in (1.5) to satisfy the inequalities in (1.7) for a maximum principle to hold. However, these restrictions could be too demanding [13] and thus many attempts to circumvent them were reported (see, e.g., [4], [15], [31], [33], and references therein), which are mainly concerned with operators of monotone kind [5, p. 350]. Of considerable importance for investigation of difference schemes associated with such operators was Ciarlet's work [4], particularly his Theorem 1, concerning the necessary and sufficient conditions for the validity of the discrete maximum principle [4, p. 341]. This theorem claims, in fact, that the discrete maximum principle cannot be valid for the scheme (1.8) if $\mathbf{B}$ in (1.8) is not of monotone kind. However, the discrete maximum principle [4, p. 341] does not imply in general that $\mathbf{B}$ in (1.8) is of monotone kind, as is clear from the following counterexample:

$$\mathbf{B} \cdot \mathbf{y} = \begin{Bmatrix} 1 & 0.5 & -1 \\ -1 & 1 & -0.5 \\ 0 & 0 & 1 \end{Bmatrix} \begin{Bmatrix} 0.5q \\ q \\ q \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ q \end{Bmatrix} = \mathbf{q}. \tag{1.9}$$

Obviously $\mathbf{B}$ in (1.9) is not of monotone kind, and yet the discrete maximum principle [4, p. 341] holds for (1.9).

Since a matrix is of monotone kind iff all elements of its inverse matrix are nonnegative [5], we are, in general, facing the above-mentioned problem of nonsingularity.

Because of this, more easily verifiable criteria are used in practice even if such criteria are only sufficient. Except in a few special cases [15], these simple and easily verifiable in practice criteria (see, e.g., [4], [5], [19], [25], [29], [33]) contain the inequalities similar to those in (1.7).

The necessary and sufficient conditions for the maximum principle for inverse column entries to hold for (1.4) have also been considered in [33, p. 54] (see also Theorem 3 and Corollary in [31, pp. 153–154]). It is assumed that in (1.4), $q_i = 0$ for all $i \in \Omega_0 \subset \Omega$. If $\mathbf{q} \geq 0, \neq 0$ in (1.4), then the maximum principle for inverse column entries [33] states

$$\mathbf{y} \geq 0, \quad \max_{i\in\Omega} y_i = \max_{i\in\overline{\Omega}_0} y_i. \tag{1.10}$$

The theorem is formulated as follows [33], [31]. Let $\mathbf{B}$ in (1.4) be a nonsingular M-matrix. Then $\mathbf{B}$ satisfies the maximum principle for inverse column entries iff $\mathbf{B}\cdot\mathbf{e} \geq 0, \neq 0$, where $\mathbf{e}$ is the vector having all components equal to unit. Notice that the condition $\mathbf{B}\cdot\mathbf{e} \geq 0, \neq 0$ is not necessary for the validity of the maximum principle provided Ω_0 is a prescribed set. As a counterexample let us consider the following equation system:

$$\mathbf{B}\cdot\mathbf{y} = \begin{Bmatrix} 1 & -2 & 0 \\ -0.25 & 1 & -0.25 \\ 0 & 0 & 1 \end{Bmatrix} \begin{Bmatrix} q \\ 0.5q \\ q \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ q \end{Bmatrix} = \mathbf{q}. \tag{1.11}$$

It is easy to see that $\mathbf{B}$ in (1.11) is a nonsingular M-matrix, which satisfies the maximum principle. However, the condition $\mathbf{B}\cdot\mathbf{e} \geq 0, \neq 0$ is violated, namely, $\mathbf{B}\cdot\mathbf{e} = \{-1, 0, 1\}^T$.

Quite apparently, when $\mathbf{B}$ in (1.4) is an arbitrary matrix, the question of monotonicity of the scheme (1.4) is, in general, more involved.

Stoyan [31] attempted to extend the maximum principle to the (1.4) scalar scheme when $\mathbf{B}$ is not necessarily of monotone kind. It is assumed that in (1.4) $q_i = 0$ for all $i \in \Omega_0 \subseteq \Omega$. The matrix $\mathbf{B}$ in (1.4) is said to satisfy the maximum principle for the absolute values [31] if

$$\overline{\Omega}_0 = \emptyset \implies \mathbf{y} = 0, \tag{1.12}$$

$$\overline{\Omega}_0 \neq \emptyset \implies \max_{i\in\Omega_0} |y_i| < \max_{j\in\overline{\Omega}_0} |y_j|. \tag{1.13}$$

A theorem was suggested [31] claiming that the maximum principle is valid for the absolute values iff $\mathbf{B}$ in (1.4) is a strictly row diagonally dominant [33, p. 8] H-matrix. However, the claim [31] of a sufficient condition for the validity of the maximum principle is proven only in the trivial case when $\mathbf{q} = 0$ in (1.4). The proof of the case $\mathbf{q} \neq 0$ is, in fact, absent, since the validity of (1.13) is assumed to be obvious with the understanding that $\mathbf{B}$ in (1.4) is nonsingular. Let us note that the maximum principle for the absolute values does not imply in general that the matrix of the scheme is strictly row diagonally dominant, provided that Ω_0 is a prescribed subset of the set Ω, as is clear from the following counterexample:

$$\mathbf{B}\cdot\mathbf{y} = \begin{Bmatrix} 2 & 1 \\ 3 & 1 \end{Bmatrix} \begin{Bmatrix} q \\ -2q \end{Bmatrix} = \begin{Bmatrix} 0 \\ q \end{Bmatrix} = \mathbf{q}. \tag{1.14}$$

It is easy to see that **B** in (1.14) is not a strictly row diagonally dominant matrix, and yet it satisfies the maximum principle for the absolute values.

Since a homogeneous system of linear equations always possesses the trivial solution, (1.12) is equivalent to $\overline{\Omega}_0 = \emptyset \Longleftrightarrow \mathbf{y} = 0$. Thus, (1.12) implies that (1.4) has a unique solution [17, p. 101]. Hence, the demand that **B** in (1.4) must be invertible is embedded into the definition of the maximum principle for the absolute values. Once again, we obtain the above-mentioned problem of nonsingularity. The well-known sufficient criterion [12], [33] ensures that an irreducibly diagonally dominant matrix is nonsingular. With this criterion, namely, under the requirement of irreducibility, we have a broad spectrum of schemes, which are of great importance in practice and which cannot be tested for monotonicity.

So, it is quite important to establish an easily verifiable criterion of nonsingularity for a diagonally dominant matrix, be it reducible or irreducible. The sought-after criterion will be established in section 2.

Let us note that the more general case when **B** in (1.1) is a rectangular matrix is not uncommon in practice [3], [12], [28], and hence such schemes should also be tested for monotonicity. An example of such an investigation for the rectangular matrix **B** in (1.1) when $N_Y > N_Q$ can be found in [29].

We assign $N_Y - N_Q$ unknowns in (1.1) as disposable [18], using equality (1.6). In such a case equality (1.6) can be seen as introducing specific designations for the disposable unknowns; (1.6) can be viewed as (1.5) as well. Thus, we obtain a scheme in the form (1.8), where $M = N_Y$ and $M' = N_Q$. Using such an approach we reduce the investigation of the monotonicity of the scheme with a rectangular matrix ($N_Y > N_Q$) to the scheme with a square one. Thus, we will consider overdetermined equation systems, namely, the scheme (1.1) where $N_Y \leq N_Q$. In such a case the canonical form of the difference scheme would be written in such a way that there exist several neighborhoods (or at least one) for each of the grid nodes.

Difference schemes appearing in practice (see, e.g., [6], [13], [21], [23], [24]) are often formulated in the vector form (1.3). We find that these schemes are converted into the scalar form (1.4) in order to test for monotonicity (see, e.g., [21]). Such an approach can facilitate the investigation to some extent, yet it may yield too restrictive conditions for the scheme monotonicity, as shown in section 4. Because of this we will use the approach which has long been exploited in the theory of PDE systems (see, e.g., [10]), where the uniqueness theorems are proved with the help of the maximum principle which holds for the Euclidean norm of the solutions to the PDE system. Thus, a maximum principle, as applied to a vector difference scheme, can be formulated in terms of a vector norm for the vectors associated with the grid nodes. Moreover, the maximum principle must be formulated in such a way as to imply a unique solution to the difference scheme rather than the reverse.

The idea of using the above approach for a discrete version of a PDE system is obvious and has been considered in the literature. For instance, Ladyzhenskaya [11] discussed the possibility (positive and negative aspects) of using a discrete maximum principle for the analysis of difference schemes approximating PDE systems. Furthermore, Samarskiy and Nikolaev [26] considered the conditions for stability of an algorithm (à la Thomas) for solving a block tridiagonal system of equations. These conditions are, in fact, sufficient for such a system of equations to satisfy a discrete maximum principle with respect to a vector norm. We will emphasize this fact in section 4. We should also mention Stoyan's paper [30], where the author dealt with the difference scheme for weakly coupled systems of PDEs; i.e., the equations are coupled

only through source terms. Moreover, the equations can differ (from one another) only by the source terms. Using the sufficient conditions for a scalar difference scheme to satisfy the classic boundary maximum principle [25] and, due to several additional assumptions, Stoyan proved the maximum norm stability of the scheme. A close look at the proof shows that he obtained an additional interesting result, namely, that the Euclidean norm of the vectors of the discrete solution takes on its maximum on the boundary. It is this auxiliary result that implies the maximum norm stability.

In sections 2 and 3 we will establish a novel maximum principle for, in general, overdetermined vector difference schemes. The sufficient as well as necessary and sufficient conditions for the monotonicity of difference schemes will be developed in section 2. In section 3 we will provide the concept and criteria of submonotonicity for, in general, nonlinear difference schemes. In section 4 we will discuss the interrelationship between different criteria of monotonicity and submonotonicity applied to, in general, vector schemes.

2. Monotonicity. We will consider, in general, vector difference schemes, and hence we will assume that $\mathbf{B}$, $\mathbf{y}$, and $\mathbf{q}$ in (1.1) can be partitioned. Let $\mathbf{y}$ in (1.1) be represented in the form

$$\mathbf{y} = \left\{\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_M^T\right\}^T, \quad 1 \le M < \infty, \tag{2.1}$$

where $\mathbf{y}_i \in L$, $i = 1, 2, \ldots, M$, and L is a linear vector space with $N = \dim(L)$. We will refer to the set of subscript indexes in (2.1) as a set of grid nodes Ω, i.e., $\Omega = \{i \mid i = 1, 2, \ldots, M\}$. Then the equation system (1.3) can be seen as a vector difference scheme for the vector-valued grid function $\mathbf{y}_i$ determined on the grid Ω. We will, in general, assume that (1.3) is overdetermined, i.e., $M \le K$. Let M_i ($M_i \ge 1$ for all $i \in \Omega$) denote the number of vector equations associated with $\mathbf{y}_i$, i.e., with the node i. It is evident that $\sum_i M_i = K$. For the sake of convenience, we take into consideration that there exist several equations associated with $\mathbf{y}_i$, and we introduce new designations for the matrix-valued coefficients in (1.3). The equations associated with $\mathbf{y}_i$ will be written in the canonical form

$$\mathbf{B}_{m_i,i} \cdot \mathbf{y}_i = \mathbf{B}^j_{m_i,i} \cdot \mathbf{y}_j + \mathbf{q}_{m_i,i}, \quad m_i = 1, 2, \ldots, M_i, \quad i, j = 1, 2, \ldots, M, \tag{2.2}$$

where $\mathbf{B}_{m_i,i} \in L^2$ denotes a prescribed nonsingular matrix, $\mathbf{B}^j_{m_i,i} \in L^2$ denotes a prescribed matrix ($\mathbf{B}^j_{m_i,i} = 0$ if $j = i$), and $\mathbf{q}_{m_i,i} \in L$ denotes a prescribed vector. Let $P_{m_i,i}$ denote a neighborhood of the node $i \in \Omega$, i.e., $P_{m_i,i} = \{j \mid \mathbf{B}^j_{m_i,i} \ne 0\}$. In view of (2.2), there exist several, namely M_i, neighborhoods of the node $i \in \Omega$. The one-to-one correspondence between the neighborhoods $P_{m_i,i}$ of the node $i \in \Omega$ on the one hand and the vectors $\mathbf{q}_{m_i,i}$, the matrices $\mathbf{B}_{m_i,i}$, and the collections $\left\{\mathbf{B}^1_{m_i,i}, \mathbf{B}^2_{m_i,i}, \ldots, \mathbf{B}^M_{m_i,i}\right\}$ of matrices on the other hand is obvious.

A grid node will be referred to as a boundary one if at this node $\mathbf{y}_i$ is equal to a prescribed vector. In such a case we write

$$\mathbf{y}_i = \mathbf{g}_i. \tag{2.3}$$

Equality (2.3) can be viewed as the equation (2.2) for which $\mathbf{B}_{m_i,i}$ becomes the identity matrix, $P_{m_i,i}$ becomes the empty set, and $\mathbf{q}_{m_i,i} = \mathbf{g}_i$. Thus, if $i \in \Omega$ is a boundary node, then $\mathbf{q}_{m_i,i}$ for all m_i ($1 \le m_i \le M_i$) must be, in view of (1.2), equal to the same prescribed vector $\mathbf{g}_i$. Because of this fact we will assume that $M_i = 1$ for a boundary

node. A node will be referred to as being interior if there exists at least one nonempty neighborhood for this node.

As $\mathbf{B}_{m_i,i}$ is nonsingular, (2.2) can be rewritten in the form

$$\mathbf{y}_i = \sum_{j\in P_{m_i,i}} \mathbf{A}^j_{m_i,i}\cdot \mathbf{y}_j + \mathbf{f}_{m_i,i},\ m_i = 1,2,\dots,M_i,\quad i = 1,2,\dots,M, \tag{2.4}$$

where $\mathbf{A}^j_{m_i,i} = (\mathbf{B}_{m_i,i})^{-1}\cdot \mathbf{B}^j_{m_i,i}$, $\mathbf{f}_{m_i,i} = (\mathbf{B}_{m_i,i})^{-1}\cdot \mathbf{q}_{m_i,i}$. For brevity's sake we shall omit the index m_i in (2.4). In such an event, the scheme (2.4) will be written as

$$\mathbf{y}_i = \sum_{j\in P_i} \mathbf{A}^j_i\cdot \mathbf{y}_j + \mathbf{f}_i \quad \forall P_i\in\Pi_i,\quad \forall i\in\Omega, \tag{2.5}$$

where $\Pi_i = \{P_{m_i,i} \mid m_i = 1,2,\dots,M_i\}$, and where $\mathbf{f}_i \in L$ and $\mathbf{A}^j_i \in L^2$, respectively, denote the vector and the matrix corresponding to the neighborhood P_i. If $M_i = 1$ for all $i\in\Omega$, then, bearing in mind that $\mathbf{A}^j_i = 0$ iff $j\notin P_i$, we can write (2.5) as

$$\mathbf{y}_i = \mathbf{A}^j_i\cdot \mathbf{y}_j + \mathbf{f}_i,\quad i,j = 1,2,\dots,M. \tag{2.6}$$

DEFINITION 2.1. *A grid node $i\in\Omega$ will be referred to as being connected to a grid node $j\in\Omega$ if either $j = i$ or there is a directed path from i to j, i.e., there exists a sequence of grid nodes $i_1,\ i_2,\dots,i_k$ and neighborhoods $P_i,\ P_{i_1},\dots,P_{i_{k-1}},\ P_{i_k}$ such that $i_1\in P_i,\ i_2\in P_{i_1},\dots,i_k\in P_{i_{k-1}},\ j\in P_{i_k}$. If $i\in\Omega$ is connected to $j\in\Omega$, then we write $i\rightarrowtail j$. A grid node $i\in\Omega$ will be referred to as being connected to a subset $\theta\subseteq\Omega$ if there exists at least one node $j\in\theta$ such that $i\rightarrowtail j$. In this case we write $i\rightarrowtail\theta$. A subset $\varphi\subseteq\Omega$ will be referred to as being connected to a subset $\theta\subseteq\Omega$ if every grid node $i\in\varphi$ is connected to θ. This will be written $\varphi\rightarrowtail\theta$.*

2.1. Linear system of inequalities with rectangular matrix. In this subsection we consider a linear system of inequalities corresponding to a scalar difference scheme. We assume that the system can be written in the form

$$\Lambda_i(y_1,\dots,y_i,\dots,y_M) \equiv y_i - \sum_{j\in P_i} a^j_i y_j \le f_i \quad \forall P_i\in\Pi_i,\ \forall i\in\Omega, \tag{2.7}$$

where $y_i,\ f_i,\ a^j_i\in\mathbb{R}$, Ω is the set of grid nodes, and Π_i is the set of neighborhoods of a node $i\in\Omega$. For brevity, we will write $\Lambda(y_i)$ instead of $\Lambda_i(y_1,\dots,y_i,\dots,y_M)$.

LEMMA 2.2. *Let $\Omega_0\subseteq\Omega_\beta\subseteq\Omega$, $\Omega_0\neq\emptyset$, and*

$$\beta_i \equiv \sum_{j\in P_i} a^j_i \le 1,\qquad a^j_i > 0\ \forall P_i\in\Pi_i,\ \forall i\in\Omega_\beta. \tag{2.8}$$

Let $f_i = 0$ for all $P_i\in\Pi_i$, for all $i\in\Omega_0$. If every grid node $i\in\Omega_0$ is connected to either the complement $\overline{\Omega}_0\equiv\Omega\setminus\Omega_0$ or a node $l\in\Omega_0$ such that $\beta_l < 1$ for at least one neighborhood P_l, then the maximum positive value of y_i at the nodes belonging to Ω_0 cannot be greater than the maximum positive value of y_i at the nodes belonging to $\overline{\Omega}_0$.

Proof. Assume the contrary. Then we have $m\in\Omega_0$ such that $y_m > 0$ and

$$y_m = \mu = \max_{i\in\Omega_0} y_i > \max_{i\in\overline{\Omega}_0} y_i > 0. \tag{2.9}$$

Hence, in view of (2.9) we can argue that $\mu = \max_{i\in\Omega} y_i$. As $y_m \geq y_j$ for all $j \in P_m$, and in view of (2.8), we obtain for every neighborhood of $m \in \Omega_0$ that

$$\Lambda(y_m) = (1-\beta_m)\, y_m + \sum_{j\in P_m} a_m^j (y_m - y_j) \geq (1-\beta_m)\, y_m \geq 0. \tag{2.10}$$

In the case when $\beta_m < 1$ for at least one neighborhood $P_m \in \Pi_m$, we consider the condition $\Lambda(y_m) \leq 0$ and obtain the contradiction with the last inequality in (2.10), which proves Lemma 2.2. If, however, $\beta_m = 1$ for each neighborhood $P_m \in \Pi_m$, we conclude that

$$\Lambda(y_m) = 0; \quad y_j = y_m = \mu \quad \forall\, j \in P_m,\ \forall P_m \in \Pi_m. \tag{2.11}$$

In the case when there exist P_m and $j \in P_m$ such that $j \in \overline{\Omega}_0$ we obtain, in view of (2.11), that $\mu = y_j \leq \max_{i\in\overline{\Omega}_0} y_i$. Thus, we obtain the contradiction with (2.9), which proves Lemma 2.2. The only remaining alternative to be considered is when $\beta_m = 1$ for all P_m, viz., every node $j \in P_m$ for all P_m, belongs to Ω_0. Since the node $m \in \Omega_0$ is connected to the node l such that either $l \in \Omega_0$ and $\beta_l < 1$ for at least one neighborhood P_l or $l \in \overline{\Omega}_0$, there exist nodes i_1, i_2, ..., i_k and neighborhoods $P_m,\ P_{i_1}, \ldots, P_{i_{k-1}},\ P_{i_k}$ such that

$$i_1 \in P_m, \quad i_2 \in P_{i_1},\ \ldots,\ i_k \in P_{i_{k-1}}, \quad l \in P_{i_k}. \tag{2.12}$$

By virtue of (2.11) we conclude that $y_{i_1} = \mu$. Using similar arguments it can be shown that $y_{i_2} = \mu,\ \ldots,\ y_{i_k} = \mu,\ y_l = \mu$. In the case when $l \in \Omega_0$ and $\beta_l < 1$, we consider the condition $\Lambda(y_l) \leq 0$ and obtain the contradiction with the inequality similar to the last one in (2.10). In the case when $l \in \overline{\Omega}_0$ we obtain $\mu = y_l \leq \max_{i\in\overline{\Omega}_0} y_i$, which is in contradiction with (2.9). These contradictions manifest the proof of Lemma 2.2. □

2.2. Schemes composed of arbitrary linear operators. In this subsection we consider the vector schemes written in the canonical form (2.5). Based on (2.5), we can construct the operators

$$\mathbf{A}_i = \left\{ \mathbf{A}_i^1\ \mathbf{A}_i^2\ \ldots\ \mathbf{A}_i^j\ \ldots\ \mathbf{A}_i^M \right\} \quad \forall P_i \in \Pi_i, \quad \forall i \in \Omega. \tag{2.13}$$

Then, by virtue of (2.13) and (2.1), the scheme (2.5) can be written in the form

$$\mathbf{y}_i = \mathbf{A}_i \cdot \mathbf{y} + \mathbf{f}_i \quad \forall P_i \in \Pi_i, \quad \forall i \in \Omega. \tag{2.14}$$

Adopting the Chebyshev norm for $\mathbf{y}$ of (2.14) in the form

$$\|\mathbf{y}\|_C = \max_{i\in\Omega} \|\mathbf{y}_i\|, \tag{2.15}$$

where $\|\mathbf{y}_i\|$ is a prescribed vector norm defined on the vector space L, we can obtain a matrix norm, $\|\mathbf{A}_i\|_{\hat{C}}$, which is compatible [12] with $\|\mathbf{y}_i\|$ and $\|\mathbf{y}\|_C$, i.e.,

$$\|\mathbf{A}_i \cdot \mathbf{y}\| \leq \|\mathbf{A}_i\|_{\hat{C}}\, \|\mathbf{y}\|_C. \tag{2.16}$$

By way of example let us consider the norm

$$\|\mathbf{A}_i\|_{\hat{C}} = \sum_{j\in P_i} \left\| \mathbf{A}_i^j \right\| \quad \forall P_i \in \Pi_i, \quad \forall i \in \Omega, \tag{2.17}$$

where the matrix norm $\|\mathbf{A}_i^j\|$ is induced [12] by the prescribed vector norm $\|\mathbf{y}_j\|$. The norm $\|\mathbf{A}_i\|_{\hat{C}}$ of (2.17) is compatible with $\|\mathbf{y}_i\|$ and $\|\mathbf{y}\|_C$. Actually, by virtue of (2.13) and (2.1), we obtain

$$\|\mathbf{A}_i \cdot \mathbf{y}\| = \left\|\mathbf{A}_i^j \cdot \mathbf{y}_j\right\| \le \left\|\mathbf{A}_i^j\right\| \|\mathbf{y}_j\| \le \left(\sum_{j\in P_i} \left\|\mathbf{A}_i^j\right\|\right) \max_{j\in\Omega} \|\mathbf{y}_j\|. \tag{2.18}$$

Thus, in view of (2.18), (2.17), and (2.15), inequality (2.16) is valid for $\|\mathbf{A}_i\|_{\hat{C}}$ of (2.17).

DEFINITION 2.3. *A vector scheme that can be written in the canonical form* (2.5) *is said to satisfy a row-contraction criterion* (RC-criterion) *if the following conditions hold:*

$$\beta_i \equiv \|\mathbf{A}_i\|_{\hat{C}} \le 1 \quad \forall P_i \in \Pi_i \quad \forall i \in \Omega; \tag{2.19}$$

$$\omega \equiv \{i \in \Omega \mid \exists P_i \in \Pi_i \text{ such that } \beta_i < 1\} \ne \emptyset. \tag{2.20}$$

An example of the scheme satisfying the RC-criterion is the scalar scheme (1.4), where $\mathbf{B}$ is a weakly row diagonally dominant matrix [33, p. 8].

The subset $\omega \subseteq \Omega$ of (2.20) contains only those grid nodes at which there exist the *strict row-contractions* (*SRC*, i.e., $\beta_i < 1$) with respect to the norm $\|\mathbf{A}_i\|_{\hat{C}}$. This subset will be important in the subsequent discussion, and it will be referred to as the *SRC-subset*.

DEFINITION 2.4. *Consider a linear scheme that can be written in the canonical form* (2.5). *Let the SRC-subset* $\omega \ne \emptyset$, *and let* $\mathbf{f}_i = 0$ *for all* $P_i \in \Pi_i$, *for all* $i \in \overline{\omega}$. *The scheme is said to satisfy the maximum principle with respect to the vector norm if*

$$\max_{i\in\Omega} \|\mathbf{y}_i\| = \max_{i\in\omega} \|\mathbf{y}_i\|. \tag{2.21}$$

THEOREM 2.5. *If a linear difference scheme satisfies the maximum principle in Definition* 2.4, *then the scheme possesses a unique solution.*

Proof. The equation system (2.5) possesses a unique solution iff the associated system of homogeneous equations implies the trivial solution $\mathbf{y}_i = 0$ for all $i \in \Omega$ [17, p. 101]. Let us assume that $\mathbf{f}_i \equiv 0$ in (2.5) (i.e., $\mathbf{f}_i = 0$ for all $P_i \in \Pi_i$, for all $i \in \Omega$) implies

$$\mu = \max_{i\in\Omega} \|\mathbf{y}_i\| > 0. \tag{2.22}$$

Then, in view of (2.21), there exists $m \in \omega$ such that

$$\|\mathbf{y}_m\| = \mu. \tag{2.23}$$

Inasmuch as $m \in \omega$, there is $P_m \in \Pi_m$ such that $\|\mathbf{A}_m\|_{\hat{C}} < 1$. Hence, by virtue of (2.14) and (2.16), we obtain

$$\|\mathbf{y}_m\| = \|\mathbf{A}_m \cdot \mathbf{y}\| \le \|\mathbf{A}_m\|_{\hat{C}} \|\mathbf{y}\|_C < \|\mathbf{y}\|_C = \max_{i\in\Omega} \|\mathbf{y}_i\|. \tag{2.24}$$

In view of (2.24), (2.22), and (2.23) we obtain the contradiction $\|\mathbf{y}_m\| < \|\mathbf{y}_m\|$, which finishes the proof of Theorem 2.5. ∎

THEOREM 2.6. *Let a linear scheme satisfy the RC-criterion* (2.19), (2.20) *with respect to the norm* (2.17). *If* $\Omega \rightarrowtail \omega$, *then the maximum principle from Definition* 2.4 *holds for this scheme.*

Proof. If $\overline{\omega} = \emptyset$, i.e., $\omega = \Omega$, then the validity of (2.21) is obvious. Consider the case when $\overline{\omega} \neq \emptyset$. By virtue of (2.5), we obtain

$$\|\mathbf{y}_i\| \leq \sum_{j \in P_i} \left\| \mathbf{A}_i^j \right\| \|\mathbf{y}_j\| + \|\mathbf{f}_i\| \quad \forall P_i \in \Pi_i, \quad \forall i \in \Omega. \tag{2.25}$$

Assume that $\mathbf{f}_i = 0$ for all $P_i \in \Pi_i$, for all $i \in \overline{\omega}$; then we obtain, in view of (2.25), that

$$\Lambda(\|\mathbf{y}_i\|) \equiv \|\mathbf{y}_i\| - \sum_{j \in P_i} \left\| \mathbf{A}_i^j \right\| \|\mathbf{y}_j\| \leq 0 \quad \forall P_i \in \Pi_i, \quad \forall i \in \overline{\omega}. \tag{2.26}$$

The conditions of Lemma 2.2 are, in view of (2.19), (2.20), and (2.26), fulfilled under $\Omega_0 = \overline{\omega}$, $\Omega_\beta = \Omega$, $a_i^j = \|\mathbf{A}_i^j\|$, and $y_i = \|\mathbf{y}_i\|$. Thus, if $\max_{i \in \overline{\omega}} \|\mathbf{y}_i\| > 0$, then we obtain, in view of Lemma 2.2, that $\max_{i \in \overline{\omega}} \|\mathbf{y}_i\| \leq \max_{i \in \omega} \|\mathbf{y}_i\|$ and, consequently, the validity of (2.21). If $\max_{i \in \overline{\omega}} \|\mathbf{y}_i\| = 0$, then the validity of (2.21) is obvious inasmuch as $\max_{i \in \omega} \|\mathbf{y}_i\| \geq 0$. □

Let us now consider the question of nonsingularity of a weakly row diagonally dominant matrix [33, p. 8]. It is apparent that all diagonal entries of such a matrix are nonzero; otherwise the entries of the associated row are all equal to zero, and hence the matrix is singular. Thus, without loss of generality, we suppose that all diagonal entries of $\mathbf{B}$ in (1.4) are equal to unit. Let

$$\omega \equiv \left\{ i \in \Omega \mid \sum_{j \neq i} \left| b_i^j \right| < 1 \right\} \neq \emptyset. \tag{2.27}$$

With the above assumptions we can now state the following result.

THEOREM 2.7. (i) *The weakly row diagonally dominant matrix* $\mathbf{B}$ *in* (1.4) *will be nonsingular if* $\Omega \rightarrowtail \omega$. (ii) *Let the weakly row diagonally dominant matrix* $\mathbf{B}$ *in* (1.4) *be real and the off-diagonal elements each be nonpositive; then* $\mathbf{B}$ *will be nonsingular iff* $\Omega \rightarrowtail \omega$.

Proof. (i) If $\Omega \rightarrowtail \omega$ then, in view of Theorems 2.6 and 2.5, the equation system (1.4) possesses a unique solution. Hence, the matrix $\mathbf{B}$ in (1.4) is nonsingular. (ii) The sufficiency is already proven above. To prove the necessity we assume the contrary. Let $\mathbf{B}$ in (1.4) be nonsingular, and suppose there exists at least one node $i \in \Omega$ which is not connected with ω.

Since $\mathbf{B}$ in (1.4) is nonsingular, the scheme possesses a unique solution. Hence, the associated homogeneous equation system possesses the unique solution

$$y_j = 0 \quad \forall j \in \Omega. \tag{2.28}$$

On the other hand, considering that the node i is not connected to ω, we obtain the subset $\Omega_i = \{ j \in \Omega \mid i \rightarrowtail j \}$ such that $\Omega_i \cap \omega = \emptyset$. That is, we obtain the lower order equation system

$$y_j + \sum_{k \neq j} b_j^k y_k = 0, \quad j, k \in \Omega_i, \tag{2.29}$$

where $\sum_{k\neq j}\left|b_j^k\right|=1$ for all $j\in\Omega_i$. Consequently,

$$\sum_{k\neq j} b_j^k=-1 \quad \forall j\in\Omega_i, \tag{2.30}$$

as the off-diagonal elements of $\mathbf{B}$ in (1.4) are all nonpositive. By virtue of (2.30) we can see that an arbitrary constant $a\neq 0$ is the solution to (2.29). So, in view of (2.28), we obtain the contradiction $0=y_j=a\neq 0$ for all $j\in\Omega_i$ which proves the necessity. □

A similar condition of nonsingularity can be formulated based on the columns of $\mathbf{B}$ by considering $\mathbf{B}^T$ instead of $\mathbf{B}$ in (1.4).

Let us note that the proposed criterion ($\Omega\rightarrowtail\omega$) of nonsingularity of a weakly (row or column) diagonally dominant matrix is even easier to verify than the irreducibility of the matrix. Actually, the condition $\Omega\rightarrowtail\omega$ is fulfilled, when $i\rightarrowtail\omega$ for all $i\in\Omega$, and hence the computational work is proportional to the number of grid nodes M, whereas for verification of irreducibility one has to prove that the directed graph is strongly connected, and thus the computational work is proportional to M^2.

Let us now demonstrate that if the scheme (2.5) satisfies the RC-criterion (2.19), (2.20), then the maximum response takes place at nodes $i\in\Omega$ (i.e., in such y_i), where the vectors $\mathbf{f}_i$ are nonzero.

PROPOSITION 2.8. *Consider a linear scheme that can be written in the form* (2.5). *Let the scheme satisfy the RC-criterion with respect to the norm* (2.17), *and let* $\Omega\rightarrowtail\omega$. *If* $\mathbf{f}_i=0$ *for all* $P_i\in\Pi_i$ *for all* $i\in\Omega_0\subset\Omega$, *then*

$$\max_{i\in\Omega_0}\|\mathbf{y}_i\|\leq\max_{i\in\overline{\Omega}_0}\|\mathbf{y}_i\|. \tag{2.31}$$

Proof. The proof is trivial. In perfect analogy to the derivation of (2.25), we obtain here that $\Lambda(\|\mathbf{y}_i\|)\leq 0$ for all $P_i\in\Pi_i$ for all $i\in\Omega_0$. In such a case the conditions of Lemma 2.2 are fulfilled when $\Omega_\beta=\Omega$. Thus, if $\max_{i\in\Omega_0}\|\mathbf{y}_i\|>0$, then we obtain, in view of Lemma 2.2, the validity of (2.31). If $\max_{i\in\Omega_0}\|\mathbf{y}_i\|=0$, then the validity of (2.31) is obvious inasmuch as $\max_{i\in\overline{\Omega}_0}\|\mathbf{y}_i\|\geq 0$. □

REMARK 2.9. *The maximum principle (Definition* 2.4*), Theorems* 2.5 *and* 2.6, *and Proposition* 2.8 *will be valid even if* $\mathbf{A}_i^j$ *in* (2.5) *is an arbitrary linear bounded operator defined on a linear infinite-dimensional vector space.*

Several more points need to be made in this subsection. When testing a scheme for monotonicity, we use the directed paths (Definition 2.1) and the SRC-subsets as the fundamental objects. The construction of these objects in the case of overdetermined schemes has the following peculiarity. A grid node is included in the SRC-subset ω (2.20) if there exists strict row-contraction for at least one neighborhood of this node. Notice that there is no assumption relative to the occurrence of strict row-contraction for each of the neighborhoods. A similar situation takes place when constructing the directed paths, insofar as for each of the sequence nodes one may take one appropriate neighborhood at a time.

Given an overdetermined equation system we can construct a reduced scheme with the same unknown quantities by eliminating several equations associated with the nodes having more than one neighborhood. It is obvious that the overdetermined scheme can be monotone while the reduced one cannot. Alternatively, a scheme that does not satisfy the maximum principle can be transformed into monotone by adding several equations.

2.3. Schemes composed of normal matrices. Hereinafter, for simplicity's sake, we will consider difference schemes that can be written in the canonical form (2.6), i.e., the schemes with square matrices, if not stated otherwise. The extension of the results to the overdetermined schemes is almost obvious. Hereinafter, commuting operators (matrices) also will be referred to as permutable.

In this subsection we consider the scheme (2.6) with the operators $\mathbf{A}_i^j$ in (2.6) depending on pairwise permutable normal operators (matrices) $\mathbf{D}^r$, $r = 1,\ 2, \ldots, R$:

$$\mathbf{A}_i^j = \varphi_i^j\left(\mathbf{D}^1, \mathbf{D}^2, \ldots, \mathbf{D}^R\right), \quad R \geq 1, \quad i, j = 1, 2, \ldots, M. \tag{2.32}$$

Let $s\left(\mathbf{D}^r\right) \in \mathbb{C}$ denote the spectrum of the operator $\mathbf{D}^r$, and let $S = s\left(\mathbf{D}^1\right) \times s\left(\mathbf{D}^2\right) \times \cdots \times s\left(\mathbf{D}^R\right)$ denote the Cartesian product of the spectra. Let us assume that each of the functions $\varphi_i^j\left(\lambda_1, \lambda_2, \ldots, \lambda_R\right)$ can be represented by a convergent Laurent series at each point $\mathbf{\Lambda} \equiv \{\lambda_1, \ldots, \lambda_R\} \in S$. Here, we will use the following matrix norm of $\mathbf{A}_i$ in (2.13):

$$\|\mathbf{A}_i\|_{\hat{C}} = \max_{\Lambda \in S} \sum_{j \in P_i} \left|\varphi_i^j\left(\lambda_1, \lambda_2, \ldots, \lambda_R\right)\right|. \tag{2.33}$$

THEOREM 2.10. *Consider a difference scheme that can be written in the canonical form* (2.6). *Let the scheme satisfy the RC-criterion* (2.19), (2.20) *with respect to the norm* (2.33), *and let* $\varphi_i^j\left(\lambda_1, \ldots,\ \lambda_R\right) \neq 0$ *for all* $\Lambda \in S$, *for all* $i, j \in \Omega$. *If* $\Omega \rightarrowtail \omega$, *then the maximum principle formulated by Definition* 2.4 *holds for this scheme.*

Proof. As $\mathbf{D}^r$ belongs to the set of pairwise permutable normal operators, the matrices of the set are simultaneously unitary similar to diagonal matrices [18]; i.e., there exists a unitary matrix $\mathbf{U}$ such that

$$\mathbf{U}^{-1} \cdot \mathbf{D}^r \cdot \mathbf{U} = \{\lambda_n^r \delta_{nm}\}, \quad m, n = 1, 2, \ldots, N,\ r = 1, 2, \ldots R, \tag{2.34}$$

where λ_n^r denotes the nth eigenvalue of $\mathbf{D}^r$. Hereinafter, δ_{nm} denotes the Kronecker delta. Let us rewrite (2.6) in the form (2.14) and let

$$\mathbf{x}_i = \mathbf{U}^{-1} \cdot \mathbf{y}_i \quad \forall i \in \Omega. \tag{2.35}$$

Then we obtain

$$\mathbf{x}_i = \mathbf{W}_i \cdot \mathbf{x} + \mathbf{F}_i \quad \forall i \in \Omega, \quad \mathbf{x} = \left\{\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_M^T\right\}^T, \tag{2.36}$$

where $\mathbf{F}_i = \mathbf{U}^{-1} \cdot \mathbf{f}_i$, $\mathbf{W}_i = \left\{\mathbf{W}_i^1\ \mathbf{W}_i^2\ \ldots\ \mathbf{W}_i^M\right\}$,

$$\mathbf{W}_i^j = \mathbf{U}^{-1} \cdot \mathbf{A}_i^j \cdot \mathbf{U} \quad \forall i, j \in \Omega. \tag{2.37}$$

As $\varphi_i^j\left(\lambda_1, \lambda_2, \ldots, \lambda_R\right)$ can be expanded into a convergent Laurent series, every block $\mathbf{W}_i^j$ in (2.37) can be written in the form

$$\mathbf{W}_i^j = \varphi_i^j\left(\mathbf{U}^{-1} \cdot \mathbf{D}^1 \cdot \mathbf{U}, \mathbf{U}^{-1} \cdot \mathbf{D}^2 \cdot \mathbf{U}, \ldots, \mathbf{U}^{-1} \cdot \mathbf{D}^R \cdot \mathbf{U}\right). \tag{2.38}$$

In view of (2.36) and (2.34), we write $\mathbf{W}_i^j$ in the diagonal form

$$\mathbf{W}_i^j = \left\{W_{in}^j \delta_{nm}\right\}, \quad m, n = 1, 2, \ldots, N, \tag{2.39}$$

where

$$W_{in}^{j} = \varphi_i^j\left(\lambda_n^1, \lambda_n^2, \ldots, \lambda_n^R\right). \tag{2.40}$$

Let x_{in} and F_{in}, $n = 1, 2, \ldots, N$, designate, respectively, the components of the vectors $\mathbf{x}_i$ and $\mathbf{F}_i$, i.e., $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iN}\}^T$, $\mathbf{F}_i = \{F_{i1}, F_{i2}, \ldots, F_{iN}\}^T$. The equation system (2.36) can be rewritten, by virtue of (2.39), in the scalar form

$$x_{in} = W_{in}^{j} x_{jn} + F_{in}, \quad n = 1, 2, \ldots, N,\ j \in P_i,\ i \in \Omega. \tag{2.41}$$

By virtue of (2.39) and (2.40), we can write

$$\max_{n=1,2,\ldots,N} \sum_{j \in P_i} \left|W_{in}^{j}\right| \leq \max_{\Lambda \in S} \sum_{j \in P_i} \left|\varphi_i^j\left(\lambda_1, \lambda_2, \ldots, \lambda_R\right)\right| \ \ \forall i \in \Omega. \tag{2.42}$$

From (2.42), (2.33), (2.19), and (2.20) we obtain

$$\sum_{j \in P_i} \left|W_{in}^{j}\right| \leq 1, \quad n = 1, 2, \ldots, N,\ i \in \Omega, \tag{2.43}$$

$$\sum_{j \in P_i} \left|W_{in}^{j}\right| < 1, \quad n = 1, 2, \ldots, N,\ i \in \omega. \tag{2.44}$$

Let us denote a node of the scheme (2.41) by a twofold index (i, n). The following designations are also used: $\Omega^* = \{(i, n) \mid i \in \Omega\}$, $\omega^* = \{(i, n) \mid i \in \omega\}$, and

$$\Phi^* = \left\{(i, n) \mid \sum_{j \in P_i} \left|W_{in}^{j}\right| < 1\right\}. \tag{2.45}$$

Inasmuch as $\Omega \rightarrowtail \omega$ and $\varphi_i^j\left(\lambda_1, \ldots, \lambda_R\right) \neq 0$ for all $\Lambda \in S$, for all $i, j \in \Omega$ (i.e., $W_{in}^{j} \neq 0$), we can see that for any $i \in \Omega$ there exists $j \in \omega$ such that $(i, n) \rightarrowtail (j, n)$. That is, any node of the scheme (2.41) is connected with the subset where the strict row-contraction of (2.44) is valid. Thus, the scheme (2.41) satisfies the conditions of Theorem 2.6. Because of (2.44) we can write $\omega^* \subseteq \Phi^*$, and hence

$$\max_{(i,n) \in \Omega^*} |x_{in}| = \max_{(i,n) \in \Phi^*} |x_{in}| \geq \max_{(i,n) \in \omega^*} |x_{in}|. \tag{2.46}$$

If $\mathbf{f}_i = 0$ for all $i \in \overline{\omega}$, then $F_{in} = 0$ for all $(i, n) \in \overline{\omega}^* = \Omega^* \diagdown \omega^*$. By virtue of Proposition 2.8, we obtain

$$\max_{(i,n) \in \overline{\omega}^*} |x_{in}| \leq \max_{(i,n) \in \omega^*} |x_{in}|. \tag{2.47}$$

Taking into account that $\Phi^* \diagdown \omega^* \subseteq \overline{\omega}^*$, we obtain from (2.46), (2.47) the equality

$$\max_{(i,n) \in \Phi^*} |x_{in}| = \max_{(i,n) \in \omega^*} |x_{in}|. \tag{2.48}$$

Let us note that

$$\max_{(i,n) \in \Omega^*} |x_{in}| = \max_{i \in \Omega} \|\mathbf{x}_i\|_\infty, \quad \max_{(i,n) \in \omega^*} |x_{in}| = \max_{i \in \omega} \|\mathbf{x}_i\|_\infty. \tag{2.49}$$

Hence, in view of (2.48), (2.49), and (2.46), we can write

$$\max_{i\in\Omega}\|\mathbf{x}_i\|_\infty = \max_{i\in\omega}\|\mathbf{x}_i\|_\infty . \tag{2.50}$$

Considering (2.42), we can see that $\|\mathbf{A}_i\|_{\hat{C}}$ in (2.33) is compatible with $\|\mathbf{x}_i\|_\infty$ and $\|\mathbf{x}\|_C = \max_{(i,n)\in\Omega^*}|x_{in}| = \max_{i\in\Omega}\|\mathbf{x}_i\|_\infty$ in (2.36). Then, in view of (2.35), we obtain that the matrix norm $\|\mathbf{A}_i\|_{\hat{C}}$ in (2.33) is compatible with $\left\|\mathbf{U}^{-1}\cdot\mathbf{y}_i\right\|_\infty$ and $\|\mathbf{y}\|_C = \max_{i\in\Omega}\left\|\mathbf{U}^{-1}\cdot\mathbf{y}_i\right\|_\infty$. We note that (2.50) actually manifests the proof of Theorem 2.10, for it can be written in the form (2.21) with $\|\mathbf{y}_i\| = \left\|\mathbf{U}^{-1}\cdot\mathbf{y}_i\right\|_\infty$. □

REMARK 2.11. *Let us note that Theorem* 2.10 *will be valid even if there exist nodes* $i\in\Omega$ *and* $j\in P_i$ *such that* $\varphi_i^j(\lambda_n^1,\lambda_n^2,\dots,\lambda_n^R)$ *is equal to zero for several* n. *In such a case, however, the inequality* (2.44) *must be valid for these* i *and* n, *or the node* (i,n) *must be connected to* Φ^*.

2.4. Explicit schemes. Thus far we have dealt with the establishment of sufficient conditions to ensure the monotonicity of a scheme. In this subsection we will also consider necessary conditions for explicit schemes. Explicit schemes are defined as having neighborhoods of all interior nodes that contain (see (2.3)) boundary nodes only (otherwise, the schemes will be called implicit).

Let the scheme (2.6) be explicit, and let $\Omega_* = \{1,2,\dots n\}\neq\emptyset$ and $\gamma = \{n+1,\dots,M\}\neq\emptyset$ denote, respectively, the subsets of interior nodes and boundary nodes. For the sake of convenience, the following designations are also used: $\mathbf{z}_i = \mathbf{y}_i$ for $i\in\Omega_*$ $(\equiv\overline{\gamma})$, $\mathbf{x}_i = \mathbf{y}_i$ and $\mathbf{g}_i = \mathbf{f}_i$ for $i\in\gamma$, and $\mathbf{x} = \left\{\mathbf{x}_{n+1}^T,\dots,\mathbf{x}_M^T\right\}^T$. Then (2.6) can be written in the specific form

$$\mathbf{z}_i = \mathbf{A}_i^j\cdot\mathbf{x}_j + \mathbf{f}_i,\quad i\in\overline{\gamma},\ j\in\gamma, \tag{2.51}$$

$$\mathbf{x}_i = \mathbf{g}_i,\quad i\in\gamma. \tag{2.52}$$

Rewriting (2.51) in the form (2.14) with $\mathbf{A}_i = \left\{\mathbf{A}_i^{n+1}\ \mathbf{A}_i^{n+2}\ \dots\ \mathbf{A}_i^M\right\}$, we obtain

$$\mathbf{z}_i = \mathbf{A}_i\cdot\mathbf{x} + \mathbf{f}_i,\quad i\in\overline{\gamma}. \tag{2.53}$$

Let $\mathbf{z}_i$ and $\mathbf{x}_j$ in (2.51) be members of a vector space L, and let h_i, $i=1,2,\dots,n$, be vector norms on L. Adopting Chebyshev norms for $\mathbf{x}$ of (2.53) in the form

$$\|\mathbf{x}\|_{C_i} = \max_{j\in\gamma} h_i(\mathbf{x}_j),\quad i=1,2,\dots,n, \tag{2.54}$$

we obtain the matrix norms, $\|\mathbf{A}_i\|_{C_i}$, induced by $h_i(\mathbf{z}_i)$ and $\|\mathbf{x}\|_{C_i}$:

$$\|\mathbf{A}_i\|_{C_i} = \max_{\|\mathbf{x}\|_{C_i}=1} h_i(\mathbf{A}_i\cdot\mathbf{x}),\quad i=1,2,\dots,n. \tag{2.55}$$

Notice that, in the case of an explicit scheme, a node $i\in\overline{\gamma}$ is not connected to any other $j\in\overline{\gamma}$ since the neighborhood of each interior node contains only boundary nodes. Using the same reasoning we obtain that $\overline{\gamma}\rightarrowtail\gamma$. Furthermore, inasmuch as (2.52) can be also viewed as (2.53) with $\mathbf{A}_i = 0$ for all $i\in\gamma$, we obtain that $\|\mathbf{A}_i\|_{C_i} = 0$ for all $i\in\gamma$; i.e., every boundary node always belongs to the SRC-subset. On the basis of this, for an explicit scheme we define a maximum principle (with respect to the set of vector norms) in the following form.

DEFINITION 2.12. *Consider an explicit scheme that can be written in the form* (2.51), (2.52). *Let* $\mathbf{f}_i = 0$ *for all* $i \in \overline{\gamma}$. *The scheme is said to satisfy the maximum principle if*

$$h_i(\mathbf{z}_i) \le \max_{j\in\gamma} h_i(\mathbf{x}_j) \quad \forall i \in \overline{\gamma}. \tag{2.56}$$

REMARK 2.13. *If there exist numbers* $\alpha_i > 0$, $i = 1, 2, \ldots, n$, *and a vector norm,* $\| \ \|$, *on* L *such that for* h_i *of* (2.56) *we have*

$$\|\mathbf{u}\| = \alpha_i h_i(\mathbf{u}) \quad \forall \mathbf{u} \in L \ \forall i \in \overline{\gamma}, \tag{2.57}$$

then, as is easy to see, the maximum principle formulated by Definition 2.12 *can be represented in the form similar to that of the boundary maximum principle. All one has to do is supersede the condition of* (2.56) *by*

$$\max_{i\in\overline{\gamma}} \|\mathbf{z}_i\| \le \max_{i\in\gamma} \|\mathbf{x}_i\|. \tag{2.58}$$

THEOREM 2.14. *Consider an explicit scheme that can be written in the form* (2.51), (2.52). *The maximum principle formulated by Definition* 2.12 *holds for this scheme iff the matrix norm,* $\|\mathbf{A}_i\|_{C_i}$, *of* (2.55) *satisfies the following condition:*

$$\|\mathbf{A}_i\|_{C_i} \le 1 \quad \forall i \in \overline{\gamma}. \tag{2.59}$$

Proof. Assume that (2.59) holds. If $\mathbf{f}_i = 0$ for all $i \in \overline{\gamma}$, then we obtain from (2.53) that

$$h_i(\mathbf{z}_i) = h_i(\mathbf{A}_i \cdot \mathbf{x}) \le \|\mathbf{A}_i\|_{C_i} \|\mathbf{x}\|_{C_i} \le \max_{j\in\gamma} h_i(\mathbf{x}_j) \quad \forall i \in \overline{\gamma}. \tag{2.60}$$

Hence the sufficiency is proven.

Conversely, assume that there exists an explicit scheme that satisfies the maximum principle formulated by Definition 2.12 and yet does not satisfy (2.59), i.e., $\exists i \in \overline{\gamma}$ such that

$$\|\mathbf{A}_i\|_{C_i} > 1. \tag{2.61}$$

Since the scheme is explicit, the vector $\mathbf{x}_j$ for all $j \in \gamma$ can be chosen arbitrarily. Hence, the vector $\mathbf{x}$ in (2.53) also can be chosen arbitrarily. Let us consider the subset of vectors such that $\|\mathbf{x}\|_{C_i} = 1$. For each of the vectors belonging to this subset we have

$$h_i(\mathbf{A}_i \cdot \mathbf{x}) \le \|\mathbf{A}_i\|_{C_i} \|\mathbf{x}\|_{C_i} = \|\mathbf{A}_i\|_{C_i}. \tag{2.62}$$

As the norm, $h_i(\mathbf{A}_i \cdot \mathbf{x})$, is a continuous function of $\mathbf{x}$ on the closed and bounded subset $\|\mathbf{x}\|_{C_i} = 1$, there exists an $\widetilde{\mathbf{x}} = \left\{\widetilde{\mathbf{x}}_{n+1}^T, \ldots, \widetilde{\mathbf{x}}_M^T\right\}^T$ for which the maximum of the left-hand side (LHS) in (2.62) is attained [12], i.e.,

$$\|\mathbf{A}_i\|_{C_i} = \max_{\|\mathbf{x}\|_{C_i}=1} h_i(\mathbf{A}_i \cdot \mathbf{x}) = h_i(\mathbf{A}_i \cdot \widetilde{\mathbf{x}}). \tag{2.63}$$

Assume that $\mathbf{f}_i = 0$ for all $i \in \overline{\gamma}$ and $\mathbf{x} = \widetilde{\mathbf{x}}$; then (2.53) takes the form $\widetilde{\mathbf{z}}_i = \mathbf{A}_i \cdot \widetilde{\mathbf{x}}$, $i \in \overline{\gamma}$. In view of (2.63), (2.56), and (2.54) we conclude that

$$\|\mathbf{A}_i\|_{C_i} = h_i(\mathbf{A}_i \cdot \widetilde{\mathbf{x}}) = h_i(\widetilde{\mathbf{z}}_i) \le \max_{j\in\gamma} h_i(\widetilde{\mathbf{x}}_j) = \|\widetilde{\mathbf{x}}\|_{C_i} = 1. \tag{2.64}$$

In view of (2.61) and (2.64) we obtain the contradiction

$$1 < \|\mathbf{A}_i\|_{C_i} \leq 1, \tag{2.65}$$

which manifests the proof of the necessity. □

COROLLARY 2.15. *Consider an explicit scalar scheme that can be written in the form*

$$z_i = a_i^j x_j + f_i, \quad i \in \overline{\gamma},\ j \in \gamma, \tag{2.66}$$

$$x_i = g_i, \quad i \in \gamma, \tag{2.67}$$

where $a_i^j \in \mathbb{K}$ *is an element of a rectangular matrix* $\{a_i^j\}$. *The scheme* (2.66), (2.67) *will be monotone iff*

$$\sum_{j\in\gamma} \left|a_i^j\right| \leq 1 \quad \forall i \in \overline{\gamma}. \tag{2.68}$$

Consider the special case of the scheme (2.51), (2.52); namely, $\mathbf{A}_i^j$ in (2.51) depends on a normal matrix $\mathbf{D}_i$:

$$\mathbf{A}_i^j = \varphi_i^j(\mathbf{D}_i) \quad \forall i \in \overline{\gamma},\ \forall j \in \gamma. \tag{2.69}$$

Let $S_i = s(\mathbf{D}_i) \in \mathbb{C}$ denote the spectrum of the matrix $\mathbf{D}_i$, and let $\varphi_i^j(\lambda)$ be represented by a convergent Laurent series at each point $\lambda \in S_i$.

THEOREM 2.16. *Consider an explicit scheme that can be written in the form* (2.51), (2.52) *with* $\mathbf{z}_i,\ \mathbf{x}_j \in \mathbb{K}^N$. *Let* (2.69) *be valid. Then the maximum principle formulated by Definition* 2.12 *holds for this scheme iff*

$$\max_{\lambda\in S_i} \sum_{j\in\gamma} \left|\varphi_i^j(\lambda)\right| \leq 1 \quad \forall i \in \overline{\gamma}. \tag{2.70}$$

Proof. Since a normal matrix is unitary similar to a diagonal one [18], there exists a unitary matrix $\mathbf{U}_i$ such that

$$\mathbf{U}_i^{-1} \cdot \mathbf{D}_i \cdot \mathbf{U}_i = \{\lambda_{in}\delta_{nm}\}, \quad m,n = 1,2,\ldots,N, \quad i \in \overline{\gamma}, \tag{2.71}$$

where λ_{in} denotes the nth eigenvalue of $\mathbf{D}_i$. The following notation is used:

$$\overline{\mathbf{z}}_i = \mathbf{U}_i^{-1} \cdot \mathbf{z}_i,\ \ \overline{\mathbf{x}}_{i,j} = \mathbf{U}_i^{-1} \cdot \mathbf{x}_j,\ \ \mathbf{W}_i^j = \mathbf{U}_i^{-1} \cdot \mathbf{A}_i^j \cdot \mathbf{U}_i,\ \ \overline{\mathbf{f}}_i = \mathbf{U}_i^{-1} \cdot \mathbf{f}_i. \tag{2.72}$$

By virtue of (2.72), we can write (2.53) in the form

$$\overline{\mathbf{z}}_i = \mathbf{W}_i \cdot \overline{\mathbf{x}}_i + \overline{\mathbf{f}}_i, \quad \overline{\mathbf{x}}_i = \left\{\overline{\mathbf{x}}_{i,n+1}^T, \ldots, \overline{\mathbf{x}}_{i,M}^T\right\}^T, \quad i \in \overline{\gamma}, \tag{2.73}$$

where $\mathbf{W}_i = \left\{\mathbf{W}_i^{n+1}\ \mathbf{W}_i^{n+2}\ \ldots\ \mathbf{W}_i^M\right\}$. As $\varphi_i^j(\lambda)$ can be represented by a convergent Laurent series at each point $\lambda \in S_i$, $\mathbf{W}_i^j$ can be written in the form

$$\mathbf{W}_i^j = \varphi_i^j\left(\mathbf{U}_i^{-1} \cdot \mathbf{D}_i \cdot \mathbf{U}_i\right) = \left\{W_{in}^j \delta_{nm}\right\}, \quad m,n = 1,2,\ldots,N, \tag{2.74}$$

where $W_{in}^j = \varphi_i^j(\lambda_{in})$. In view of (2.74), we obtain that

$$\max_{n=1,\dots,N} \sum_{j\in\gamma} \left|W_{in}^j\right| = \max_{\lambda\in S_i} \sum_{j\in\gamma} \left|\varphi_i^j(\lambda)\right|, \quad i\in\overline{\gamma}. \tag{2.75}$$

We conclude from (2.75) that

$$\max_{\lambda\in S_i} \sum_{j\in\gamma} \left|\varphi_i^j(\lambda)\right| = \|\mathbf{W}_i\|_\infty, \quad i\in\overline{\gamma}, \tag{2.76}$$

where $\|\mathbf{W}_i\|_\infty$ is induced by the vector norms $\|\overline{\mathbf{z}}_i\|_\infty = \|\mathbf{U}_i^{-1}\cdot\mathbf{z}_i\|_\infty$ and $\|\overline{\mathbf{x}}_i\|_\infty = \max_{j\in\gamma}\|\mathbf{U}_i^{-1}\cdot\mathbf{x}_j\|_\infty$. Thus, we can see from (2.76) that the LHS in the inequality of (2.70) is the matrix norm, $\|\mathbf{A}_i\|_{C_i}$, induced by $h_i(\mathbf{z}_i)$ and $\|\mathbf{x}\|_{C_i} = \max_{j\in\gamma} h_i(\mathbf{x}_j)$ with $h_i(\mathbf{v}) \equiv \|\mathbf{U}_i^{-1}\cdot\mathbf{v}\|_\infty$ for all $\mathbf{v}\in\mathbb{K}^N$. This, in view of Theorem 2.14, establishes Theorem 2.16. □

3. Submonotonicity. Let us consider the scheme (2.51), (2.52). If we introduce the additional notation $\mathbf{z} = \left\{\mathbf{z}_1^T,\dots,\mathbf{z}_n^T\right\}^T$, $\mathbf{f} = \left\{\mathbf{f}_1^T,\dots,\mathbf{f}_n^T\right\}^T$, $\mathbf{g} = \left\{\mathbf{g}_{n+1}^T,\dots,\mathbf{g}_M^T\right\}^T$, and

$$\mathbf{A} = \left\{\mathbf{A}_i^j\right\}, \quad i\in\overline{\gamma},\ j\in\gamma, \tag{3.1}$$

then the scheme (2.51), (2.52) can be represented in the form

$$\mathbf{z} = \mathbf{A}\cdot\mathbf{x} + \mathbf{f}, \quad \mathbf{x} = \mathbf{g}. \tag{3.2}$$

Notice that (3.2) can be seen as a two-node explicit scheme, where the first node can be associated with $\mathbf{z}\in Z\equiv L^n$ and the second one with $\mathbf{x}\in X\equiv L^{M-n}$; i.e., we can write for this specific scheme that $\overline{\gamma} = \{1\}$, $\gamma = \{2\}$. In such a case the maximum principle can be defined as follows.

DEFINITION 3.1. *Consider an explicit scheme that can be written in the form* (3.2). *Let h_x and h_z be vector norms on the vectors spaces X and Z, respectively, and let $\mathbf{f} = 0$. The scheme is said to satisfy the maximum principle if*

$$h_z(\mathbf{z}) \le h_x(\mathbf{x}). \tag{3.3}$$

PROPOSITION 3.2. *The maximum principle formulated by Definition* 3.1 *holds for* (3.2) *iff*

$$\|\mathbf{A}\| \le 1, \tag{3.4}$$

where $\|\mathbf{A}\|$ is induced by the prescribed vector norms h_x and h_z.

Proof. Actually, considering $\mathbf{f} = 0$ in (3.2), we have

$$h_z(\mathbf{z}) = h_z(\mathbf{A}\cdot\mathbf{x}) \le \|\mathbf{A}\|\, h_x(\mathbf{x}). \tag{3.5}$$

By virtue of (3.4) and (3.5), we obtain (3.3). The proof of the necessity is similar to that of Theorem 2.14. □

Thus, (3.4) is the necessary and sufficient condition for monotonicity of the scheme (2.51), (2.52) with respect to the norms h_x and h_z. Let us note that in practice we usually obtain an explicit scheme in the form (2.51), (2.52), where the nodes have a physical meaning, e.g., the points of an Euclidean space. Representing the scheme in

the artificial form (3.2), we usually lose this meaning. In addition, it is reasonable to investigate the monotonicity of an original scheme with respect to a norm (or a set of norms) of the original vectors. On this basis the scheme (2.51), (2.52) will be referred to as submonotone if (3.4) is fulfilled for this scheme. Thus, every monotone scheme is also submonotone.

REMARK 3.3. *When investigating the monotonicity (or submonotonicity) of a linear scheme via* (3.3)*, it is desirable to take into account the prescribed norm,* $\| \ \|$*, of the original vectors* $\mathbf{x}_j$*,* $\mathbf{z}_i$ *of* (2.51)*. It can be done if* $h_z(\mathbf{z}) = \varphi(\|\mathbf{z}_1\|, \dots, \|\mathbf{z}_n\|)$ *and* $h_x(\mathbf{x}) = \psi(\|\mathbf{x}_{n+1}\|, \dots, \|\mathbf{x}_M\|)$*, where* φ *and* ψ *are the proper functions.*

Consider the special case of the scheme (2.51), (2.52); namely, $\mathbf{A}_i^j$ in (2.51) depends on a matrix $\mathbf{D}^j$ belonging to a set of pairwise permutable normal operators:

$$\mathbf{A}_i^j = \varphi_i^j\left(\mathbf{D}^j\right) \quad \forall i \in \overline{\gamma},\ \forall j \in \gamma. \tag{3.6}$$

Let $S^j = s\left(\mathbf{D}^j\right) \in \mathbb{C}$ denote the spectrum of the matrix $\mathbf{D}^j$, and let $\varphi_i^j(\lambda)$ be represented by a convergent Laurent series at each point $\lambda \in S^j$.

THEOREM 3.4. *Consider an explicit scheme that can be written in the form* (2.51), (2.52)*. Let* (3.6) *be valid. Then the scheme will be submonotone iff*

$$\max_{\lambda \in S^j} \sum_{i \in \overline{\gamma}} \left|\varphi_i^j(\lambda)\right| \le 1 \quad \forall j \in \gamma. \tag{3.7}$$

Proof. As $\mathbf{D}^j$ belongs to the set of pairwise permutable normal operators, the matrices of the set are simultaneously unitary similar to diagonal matrices [18]; i.e., there exists a unitary matrix $\mathbf{U}$ such that

$$\mathbf{U}^{-1} \cdot \mathbf{D}^j \cdot \mathbf{U} = \left\{\lambda_n^j \delta_{nm}\right\}, \quad m, n = 1, 2, \dots, N,\ j \in \gamma, \tag{3.8}$$

where λ_n^j denotes the nth eigenvalue of $\mathbf{D}^j$. The following notation is used:

$$\overline{\mathbf{z}}_i = \mathbf{U}^{-1} \cdot \mathbf{z}_i,\ \ \overline{\mathbf{x}}_j = \mathbf{U}^{-1} \cdot \mathbf{x}_j,\ \ \mathbf{W}_i^j = \mathbf{U}^{-1} \cdot \mathbf{A}_i^j \cdot \mathbf{U},\ \ \overline{\mathbf{f}}_i = \mathbf{U}^{-1} \cdot \mathbf{f}_i. \tag{3.9}$$

By virtue of (3.9), we can write (2.51) in the form

$$\overline{\mathbf{z}}_i = \mathbf{W}_i^j \cdot \overline{\mathbf{x}}_j + \overline{\mathbf{f}}_i \quad \forall i \in \overline{\gamma}. \tag{3.10}$$

Since $\varphi_i^j(\lambda)$ can be represented by a convergent Laurent series at each point $\lambda \in S^j$, $\mathbf{W}_i^j$ can be written as

$$\mathbf{W}_i^j = \varphi_i^j\left(\mathbf{U}^{-1} \cdot \mathbf{D}^j \cdot \mathbf{U}\right) = \left\{W_{in}^j \delta_{nm}\right\}, \quad m, n = 1, 2, \dots, N, \tag{3.11}$$

where $W_{in}^j = \varphi_i^j\left(\lambda_n^j\right)$. In view of (3.11), we obtain that

$$\max_{j \in \gamma} \left(\max_{n=1,\dots,N} \sum_{i \in \overline{\gamma}} \left|W_{in}^j\right| \right) = \max_{j \in \gamma} \left(\max_{\lambda \in S^j} \sum_{i \in \overline{\gamma}} \left|\varphi_i^j(\lambda)\right| \right). \tag{3.12}$$

We conclude from (3.12) that

$$\max_{j \in \gamma} \left(\max_{\lambda \in S^j} \sum_{i \in \overline{\gamma}} \left|\varphi_i^j(\lambda)\right| \right) = \|\mathbf{W}\|_1, \quad \mathbf{W} = \left\{\mathbf{W}_i^j\right\},\ i \in \overline{\gamma},\ j \in \gamma, \tag{3.13}$$

where $\|\mathbf{W}\|_1$ is induced by $\| \left\{\overline{\mathbf{z}}_1^T, \ldots, \overline{\mathbf{z}}_n^T\right\}^T \|_1 = \sum_{i\in\overline{\gamma}} \left\|\mathbf{U}^{-1}\cdot\mathbf{z}_i\right\|_1$ and $\|\{\overline{\mathbf{x}}_{n+1}^T, \ldots, \overline{\mathbf{x}}_M^T\}^T\|_1 = \sum_{j\in\gamma}\left\|\mathbf{U}^{-1}\cdot\mathbf{x}_j\right\|_1$. Thus, we can see from (3.13) that if (3.7) is valid, then (3.4) will be valid for the norm $\|\mathbf{A}\| = \|\{\mathbf{U}^{-1}\cdot\mathbf{A}_i^j\cdot\mathbf{U}\}\|_1$ induced by the vector norms $h_z(\mathbf{z}) = \sum_{i\in\overline{\gamma}}\left\|\mathbf{U}^{-1}\cdot\mathbf{z}_i\right\|_1$ and $h_x(\mathbf{x}) = \sum_{j\in\gamma}\left\|\mathbf{U}^{-1}\cdot\mathbf{x}_j\right\|_1$. This establishes Theorem 3.4. □

Let us now introduce the notion of submonotonicity for, in general, implicit schemes which can be written in the canonical form (2.6). Based on (2.6) we can construct the operator

$$\mathbf{A} = \left\{\mathbf{A}_i^j\right\}, \quad i,j\in\Omega = \{1,2,\ldots,M\}, \tag{3.14}$$

where $\mathbf{A}_i^j = 0$ if $i = j$. Then, by virtue of (3.14) and (2.1), the scheme (2.6) can be written as

$$\mathbf{y} = \mathbf{A}\cdot\mathbf{y} + \mathbf{f}, \quad \mathbf{f} = \left\{\mathbf{f}_1^T, \ldots, \mathbf{f}_M^T\right\}^T. \tag{3.15}$$

In such a case the scheme (2.6) will be referred to as submonotone if $\|\mathbf{A}\| \le 1$, where the matrix norm for $\mathbf{A}$ of (3.15) is induced by a prescribed vector norm $\|\mathbf{y}\|$.

Let us now consider nonlinear schemes that can be written in the form

$$\mathbf{H}_i(\mathbf{y}_1, \ldots, \mathbf{y}_M; \mathbf{q}_1, \ldots, \mathbf{q}_R) = 0, \quad i\in\Omega = \{1,2,\ldots,M\}, \tag{3.16}$$

where $\mathbf{y}_j \in \mathbb{K}^N$, $j = 1,\ldots,M$, denote the sought-after vectors, $\mathbf{q}_r \in \mathbb{K}^{N_r}$, $r = 1,\ldots,R$, denote the prescribed vectors, N and N_r denote the dimensionalities of the corresponding vector spaces, and $\mathbf{H}_i$ is a differentiable (with respect to $\mathbf{y}_j$ and $\mathbf{q}_r$) vector-valued function with the range belonging to $\mathbb{K}^N$. We will assume that there exists $\left(\partial\mathbf{H}_i/\partial\mathbf{y}_i\right)^{-1}$ for all $i\in\Omega$ over the whole domain of $\mathbf{H}_i$. Consider the variational scheme of (3.16),

$$\sum_{j=1}^{M}\frac{\partial\mathbf{H}_i}{\partial\mathbf{y}_j}\cdot\delta\mathbf{y}_j + \sum_{r=1}^{R}\frac{\partial\mathbf{H}_i}{\partial\mathbf{q}_r}\cdot\delta\mathbf{q}_r = 0, \quad i\in\Omega, \tag{3.17}$$

where $\delta\mathbf{y}_j$ and $\delta\mathbf{q}_r$ denote the variations of $\mathbf{y}_j$ and $\mathbf{q}_r$, respectively. The following notation is used:

$$\mathbf{A}_i^j = -\left(\frac{\partial\mathbf{H}_i}{\partial\mathbf{y}_i}\right)^{-1}\cdot\frac{\partial\mathbf{H}_i}{\partial\mathbf{y}_j}, \quad \mathbf{C}_i^r = -\left(\frac{\partial\mathbf{H}_i}{\partial\mathbf{y}_i}\right)^{-1}\cdot\frac{\partial\mathbf{H}_i}{\partial\mathbf{q}_r}, \tag{3.18}$$

$$\delta\mathbf{f}_i = \mathbf{C}_i^r\cdot\delta\mathbf{q}_r, \quad r = 1,\ldots,R, \quad i,j\in\Omega, \quad i\ne j. \tag{3.19}$$

By virtue of (3.18), (3.19), we rewrite (3.17) in the form akin to that of (2.6):

$$\delta\mathbf{y}_i = \mathbf{A}_i^j\cdot\delta\mathbf{y}_j + \delta\mathbf{f}_i, \quad i,j\in\Omega, \tag{3.20}$$

where $\mathbf{A}_i^j = 0$ if $i = j$.

Let us note that any linear scheme coincides with its variational scheme up to notation, and hence all the results which are valid for a linear scheme also will be valid for its variational scheme and vice-versa. In view of (3.20) we can test linear schemes for monotonicity, addressing also nonlinear schemes, as variational schemes (that can emanate from linear or nonlinear equation systems) are always linear in

terms of $\delta\mathbf{y}_i$. An analogous approach has long been exploited for investigation of the stability of motion, and we use here the ideas of Lyapunov (1892) (see, e.g., [1], [9], [24], and references therein).

DEFINITION 3.5. *A nonlinear scheme that can be written in the form* (3.16) *will be referred to as being linearly monotone or linearly submonotone if its variational scheme* (3.20) *is monotone or submonotone, respectively. For simplicity the linearly monotone and linearly submonotone schemes will also be referred to as submonotone.*

4. Exemplification and discussion. In what follows we will exemplify the use and interrelationship of the various criteria developed thus far as well as the interplay between several notions of monotonicity.

4.1. Examples. By way of illustration let us consider the following overdetermined equation system [28, p. 141] in the form (1.1), where $\mathbf{y} = \{y_1,\ y_2\}^T$, $\mathbf{q} = \{q_1, \dots, q_5\}^T$, $\mathbf{B} = \{\mathbf{B}_1\ \ \mathbf{B}_2\}$, $\mathbf{B}_1 = \{2,\ 1,\ 3,\ 4,\ 1\}^T$, $\mathbf{B}_2 = \{3,\ 1,\ 5,\ 1,\ -2\}^T$. In view of (1.2), $\mathbf{q}$ belongs to the linear span of the columns of $\mathbf{B}$: $\mathbf{q} \in \mathbf{B}\,(Y) = \alpha_1\mathbf{B}_1 + \alpha_2\mathbf{B}_2$, $\alpha_1,\ \alpha_2 \in \mathbb{R}$. This scheme [28, p. 141] can be written in the canonical form (2.4):

$$y_1 = -y_2 + q_2, \quad y_1 = -\frac{1}{4}y_2 + \frac{1}{4}q_4, \tag{4.1}$$

$$y_2 = -\frac{2}{3}y_1 + \frac{1}{3}q_1, \quad y_2 = -\frac{3}{5}y_1 + \frac{1}{5}q_3, \quad y_2 = \frac{1}{2}y_1 - \frac{1}{2}q_5. \tag{4.2}$$

Thus, the nonzero coefficients of the scheme will be $A^2_{1,1} = -1$, $A^2_{2,1} = -\frac{1}{4}$; $A^1_{1,2} = -\frac{2}{3}$, $A^1_{2,2} = -\frac{3}{5}$, $A^1_{3,2} = \frac{1}{2}$, and hence

$$\left|A^2_{1,1}\right| = 1; \quad \left|A^2_{2,1}\right|,\ \left|A^1_{1,2}\right|,\ \left|A^1_{2,2}\right|,\ \left|A^1_{3,2}\right| < 1. \tag{4.3}$$

We conclude from (4.1), (4.2), and (4.3) that the set of grid nodes $\Omega = \{1, 2\}$, and the SRC-subset $\omega = \Omega$. Hence, in view of Theorem 2.6, the maximum principle (Definition 2.4) holds for the scheme of [28, p. 141]. Let us note that the scheme will be monotone even if the second equation of (4.1) is erased. In the case of the SRC-subset $\omega = \{2\}$, however, the first grid node is connected to ω ($1 \rightarrowtail 2$) in view of the first equation of (4.1). In view of Theorem 2.5, the scheme of [28, p. 141] possesses a unique solution, viz., $y_1 = 3q_2 - q_1$, $y_2 = q_1 - 2q_2$.

Let us consider a block tridiagonal system of equations [26] that can be written in the form

$$-\mathbf{C}_0\cdot\mathbf{y}_0 + \mathbf{B}_0\cdot\mathbf{y}_1 = \mathbf{q}_0, \quad \mathbf{A}_M\cdot\mathbf{y}_{M-1} - \mathbf{C}_M\cdot\mathbf{y}_M = \mathbf{q}_M, \tag{4.4}$$

$$\mathbf{A}_i\cdot\mathbf{y}_{i-1} - \mathbf{C}_i\cdot\mathbf{y}_i + \mathbf{B}_i\cdot\mathbf{y}_{i+1} = \mathbf{q}_i, \quad 1 \le i \le M-1, \tag{4.5}$$

where $\mathbf{y}_i,\ \mathbf{q}_i \in L \equiv \mathbb{K}^N$; $\mathbf{A}_i,\ \mathbf{C}_i,\ \mathbf{B}_i \in L^2$, $i \in \Omega \equiv \{0, 1, \dots, M\}$. It is also assumed that $\mathbf{C}_i$ for all $i \in \Omega$, is nonsingular. Let us assume that for each of the nodes we have

$$\left\|\mathbf{C}_0^{-1}\cdot\mathbf{B}_0\right\| \le 1, \quad \left\|\mathbf{C}_M^{-1}\cdot\mathbf{A}_M\right\| \le 1, \tag{4.6}$$

$$\left\|\mathbf{C}_i^{-1}\cdot\mathbf{A}_i\right\| + \left\|\mathbf{C}_i^{-1}\cdot\mathbf{B}_i\right\| \le 1, \quad 1 \le i \le M-1, \tag{4.7}$$

and let there exist at least one strict inequality in (4.6)–(4.7), i.e., the SRC-subset $\omega \neq \emptyset$, and, last, let $\Omega \rightarrowtail \omega$. Then, on the strength of Theorem 2.6 the scheme (4.4)–(4.5) will be monotone.

It is interesting to note that the sufficient conditions for stability of the algorithm (similar to the Thomas's) for solving the scheme (4.4)–(4.5) [26] are similar to the conditions of monotonicity obtained by application of Theorem 2.6 to the system. The conditions of monotonicity are even weaker than the conditions of stability obtained in [26]. The distinctions between these conditions amount to the following. Samarskiy and Nikolaev [26] assumed for (4.5) that

$$\mathbf{A}_i \neq 0, \quad \mathbf{B}_i \neq 0, \quad 1 \leq i \leq M-1, \tag{4.8}$$

instead of the connectedness, i.e., $\Omega \rightarrowtail \omega$. It is easy to see that (4.8) implies $\Omega \rightarrowtail \omega$ but not vice-versa.

Let us consider a transient problem for a vector-valued function $\mathbf{V}(x,t) \in \mathbb{R}^N$,

$$\frac{\partial \mathbf{V}}{\partial t} = \frac{\partial}{\partial x}\left(\mathbf{D}\cdot\frac{\partial \mathbf{V}}{\partial x}\right), \qquad -L < x < L,\ t > 0, \tag{4.9}$$

$$\mathbf{V}(x,0) = \mathbf{V}_0(x), \ \ \mathbf{V}|_{x=\pm L} = 0, \tag{4.10}$$

where $\mathbf{V}_0(x)$ denotes a prescribed vector-valued function, and $\mathbf{D} = \{D_{ij}\}$ is a symmetric and positive definite matrix with constant elements; i.e., we can write

$$\delta \leq \sigma(\mathbf{D}) \leq \|\mathbf{D}\|_2, \quad \delta > 0. \tag{4.11}$$

Let us assign the grids $\Omega_\tau = \{k \mid k = 0\,,\,1,\,\ldots\}$, $\Omega_x = \{i = 0, \pm 1,\,\ldots\}$, $\Omega = \Omega_\tau \times \Omega_x$. Let h (= const) and τ (= const) denote the spatial interval and the time increment, respectively, and let $x_i \equiv ih$, $t_k \equiv k\tau$; $\mathbf{U}_i \equiv \mathbf{V}(x_i, t_{k+1})$, $\check{\mathbf{U}}_i \equiv \mathbf{V}(x_i, t_k)$. We make use of the well-known notation [8], [19], [23], [24], [25] for the approximations of $\mathbf{V}(x,t)$ and $\mathbf{V}_0(x)$. A possible difference scheme for (4.9), (4.10) is

$$\frac{\mathbf{U}_i - \check{\mathbf{U}}_i}{\tau} = \mathbf{D}\cdot\left[\sigma\frac{\mathbf{U}_{i+1} - 2\mathbf{U}_i + \mathbf{U}_{i-1}}{h^2} + (1-\sigma)\,\frac{\check{\mathbf{U}}_{i+1} - 2\check{\mathbf{U}}_i + \check{\mathbf{U}}_{i-1}}{h^2}\right], \tag{4.12}$$

where $\sigma \in \mathbb{R}$ is a parameter such that $0 \leq \sigma \leq 1$. Rewriting (4.12) in the canonical form (2.6), we obtain

$$\mathbf{U}_i = \mathbf{T}\cdot\left[\nu\sigma\mathbf{D}\cdot(\mathbf{U}_{i-1} + \mathbf{U}_{i+1}) + \mathbf{P}\cdot\check{\mathbf{U}}_i + \nu(1-\sigma)\,\mathbf{D}\cdot\left(\check{\mathbf{U}}_{i-1} + \check{\mathbf{U}}_{i+1}\right)\right], \tag{4.13}$$

where $\mathbf{T} = (\mathbf{I} + 2\nu\sigma\mathbf{D})^{-1}$, $\mathbf{P} = \mathbf{I} - 2\nu(1-\sigma)\,\mathbf{D}$, $\nu = \tau/h^2$. Let us note that the neighborhood of any interior node (a node at the current time-level) of the two time-level scheme (4.13) contains the nodes (or at least one) belonging to the previous time-level, i.e., the boundary nodes, insofar as the value of the grid function at the previous time-level is assumed to be known. As the boundary nodes belong to the SRC-subset ω, we can write for (4.12) that $\Omega \rightarrowtail \omega$; i.e., the scheme (4.12) provides connectedness.

At first, consider the scheme (4.13) in the scalar form. By virtue of Corollary 2.15, we obtain that the explicit ($\sigma = 0$) scheme (4.13) will be monotone iff

$$2\nu|D_{ii}| + |1 - 2\nu D_{ii}| + 4\nu\sum_{j\neq i}|D_{ij}| \leq 1, \quad i = 1,\ldots,N. \tag{4.14}$$

In view of (4.14), the explicit ($\sigma = 0$) scheme (4.13) can be monotone iff $\mathbf{D}$ becomes a diagonal matrix, which is the case of the uncoupled system of equations (4.9). Thus, we conclude that the scalar form of a vector scheme is not always suitable for investigation of monotonicity. Such an approach can lead to very stiff restrictions. However, in view of Theorem 2.16, we note that the explicit ($\sigma = 0$) scheme (4.13) will be monotone iff

$$\max_{\delta \le \lambda \le \|\mathbf{D}\|_2} \left(|1 - 2\nu\lambda| + |2\nu\lambda|\right) \le 1. \tag{4.15}$$

Hence we have, by virtue of (4.15), a necessary and sufficient condition for monotonicity of the scheme (4.13) under $\sigma = 0$:

$$\tau \le \frac{h^2}{2\,\|\mathbf{D}\|_2}. \tag{4.16}$$

Thus, the validity of the maximum principle for a scheme depends on the vector norm taken for investigation of the scheme on the monotonicity. If a vector scheme is converted into the scalar form (1.4), then, in fact, we investigate the monotonicity with respect to the cubic vector norm, $\|\mathbf{y}\|_\infty$. Such an approach, as we can see from the above, may yield too restrictive conditions for the scheme to satisfy the maximum principle.

In view of Theorem 2.10, we obtain the condition of monotonicity for the implicit ($0 < \sigma \le 1$) scheme (4.13),

$$\max_{\delta \le \lambda \le \|\mathbf{D}\|_2} |1 + 2\nu\sigma|^{-1} \left(2\,|\nu\sigma\lambda| + |1 - 2\nu\,(1-\sigma)\,\lambda| + 2\,|\nu\,(1-\sigma)\,\lambda|\right) \le 1, \tag{4.17}$$

yielding the following sufficient conditions for the validity of the maximum principle (Definition 2.4):

$$\tau \le \frac{h^2}{2\,\|\mathbf{D}\|_2\,(1-\sigma)}, \quad 0 < \sigma \le 1. \tag{4.18}$$

We can, however, obtain necessary and sufficient conditions for the monotonicity of the implicit ($0 < \sigma \le 1$) scheme (4.13). For this purpose we transform the implicit scheme (4.13) into an explicit form of this scheme. For the sake of simplicity assume that $L \longrightarrow \infty$ in (4.9), (4.10). The results of transforming (4.13) into the explicit form, for a scalar scheme, is cited in [8]. In the case of (4.13) as a vector scheme, we obtain

$$\mathbf{U}_i = \mathbf{A}_0 \cdot \check{\mathbf{U}}_i + \sum_{j=1}^{\infty} \mathbf{A}_j \cdot \left(\check{\mathbf{U}}_{i-j} + \check{\mathbf{U}}_{i+j}\right), \tag{4.19}$$

where

$$\mathbf{A}_0 = \mathbf{I} - 4\tau\mathbf{G}^{-1} \cdot (h\mathbf{I} + \mathbf{G})^{-1} \cdot \mathbf{D}, \quad \mathbf{A}_1 = 4\tau h\mathbf{G}^{-1} \cdot (h\mathbf{I} + \mathbf{G})^{-2} \cdot \mathbf{D}, \tag{4.20}$$

$$\mathbf{A}_j = \mathbf{q} \cdot \mathbf{A}_{j-1}, \quad j \ge 2, \tag{4.21}$$

$$\mathbf{q} = 4\sigma\tau\,(h\mathbf{I} + \mathbf{G})^{-2} \cdot \mathbf{D}, \quad \mathbf{G} \equiv \left(h^2\mathbf{I} + 4\sigma\tau\mathbf{D}\right)^{0.5}. \tag{4.22}$$

Since all operators in (4.19) depend on the single symmetric matrix $\mathbf{D}$, in view of Theorem 2.16, the scheme (4.19) as well as (4.12) will be monotone iff

$$\tau \leq \frac{(2-\sigma)\,h^2}{4\,\|\mathbf{D}\|_2\,(1-\sigma)^2}. \tag{4.23}$$

It is instructive to compare (4.23) with the necessary and sufficient condition for the so-called L_2-stability of the scheme (4.12). By applying the Fourier transform method [23] to the scheme (4.12) we obtain

$$\mathbf{u}(\xi) - \breve{\mathbf{u}}(\xi) = 4\nu \sin^2(\xi h/2)\,\mathbf{D}\cdot\left[\sigma\mathbf{u}(\xi) + (1-\sigma)\,\breve{\mathbf{u}}(\xi)\right], \tag{4.24}$$

where $\mathbf{u}(\xi)$ and $\breve{\mathbf{u}}(\xi)$ denote the Fourier transform of $\mathbf{U}(x)$ and $\breve{\mathbf{U}}(x)$, respectively, $\nu = \tau/h^2$. After elementary transformations we obtain from (4.24) that $\mathbf{u}(\xi) = \mathbf{S}\cdot\breve{\mathbf{u}}(\xi)$, where the amplification matrix $\mathbf{S}$ is given by

$$\mathbf{S}(\xi) = \left[\mathbf{I} + 4\nu\sigma\sin^2(\xi h/2)\,\mathbf{D}\right]^{-1}\cdot\left[\mathbf{I} - 4\nu(1-\sigma)\sin^2(\xi h/2)\,\mathbf{D}\right]. \tag{4.25}$$

The scheme (4.12) will be stable iff $\|\mathbf{S}\|_2 \leq 1$ [23], [24], from which we obtain, in view of [22, Theorem IV.1], that

$$\|\mathbf{S}\|_2 = \max_{\delta\leq\lambda\leq\|\mathbf{D}\|_2}\left|\frac{h^2 - 4\lambda\tau(1-\sigma)\sin^2(\xi h/2)}{h^2 + 4\lambda\tau\sigma\sin^2(\xi h/2)}\right| \leq 1\ \forall\xi\in(-\infty,+\infty), \tag{4.26}$$

yielding the necessary and sufficient condition for the stability of (4.12):

$$\tau \leq \begin{cases} +\infty & \text{if } 0.5\leq\sigma\leq 1 \\ \frac{h^2}{2(1-2\sigma)\|\mathbf{D}\|_2} & \text{if } 0\leq\sigma<0.5. \end{cases} \tag{4.27}$$

Since in many cases a scheme can be recommended for the practical implementation if the scheme is stable and provides consistent approximation, let us, for example, consider the scheme (4.12) under $\sigma = 0.5$, i.e., a version of the Crank–Nicholson scheme [23]. In view of (4.27) this scheme is unconditionally stable and provides second order temporal and spatial accuracy [23, p. 189]. However, by virtue of (4.23), the (4.12) scheme will be monotone (subject to $\sigma = 0.5$) iff

$$\tau \leq \frac{3h^2}{2\,\|\mathbf{D}\|_2}, \tag{4.28}$$

which is forcing strict limitation on the time step. If the inequality of (4.28) is violated, then the scheme could produce spurious oscillations.

Let us demonstrate an implementation of the submonotonicity notion for a vector nonlinear scheme. Consider a vector nonlinear equation written in the form

$$\frac{\partial\mathbf{V}}{\partial t} + \frac{\partial\mathbf{F}(\mathbf{V})}{\partial x} + \frac{\partial\mathbf{F}(\mathbf{V})}{\partial y} = 0, \quad \mathbf{V}(x,y)\in\mathbb{R}^N, \tag{4.29}$$

and the explicit upwind difference scheme approximating it,

$$\frac{\mathbf{U}_{i,j} - \breve{\mathbf{U}}_{i,j}}{\tau} + \frac{\mathbf{F}\left(\breve{\mathbf{U}}_{i,j}\right) - \mathbf{F}\left(\breve{\mathbf{U}}_{i-1,j}\right)}{h_x} + \frac{\mathbf{F}\left(\breve{\mathbf{U}}_{i,j}\right) - \mathbf{F}\left(\breve{\mathbf{U}}_{i,j-1}\right)}{h_y} = 0. \tag{4.30}$$

Let $\mathbf{A}_{i,j}$ $[\equiv (\partial\mathbf{F}/\partial\mathbf{V})|_{\mathbf{V}=\breve{\mathbf{U}}_{i,j}}]$ be normal pairwise permutable matrices.

The variational scheme of (4.30) can be written in the form

$$\delta \mathbf{U}_{i,j} = [\mathbf{I} - (\gamma_x + \gamma_y)\, \mathbf{A}_{i,j}] \cdot \delta \breve{\mathbf{U}}_{i,j}$$

$$+ \gamma_x \mathbf{A}_{i-1,j} \cdot \delta \breve{\mathbf{U}}_{i-1,j} + \gamma_y \mathbf{A}_{i,j-1} \cdot \delta \breve{\mathbf{U}}_{i,j-1}, \quad \gamma_x = \tau/h_x,\ \gamma_y = \tau/h_y. \tag{4.31}$$

Applying Theorem 2.16 to (4.31) we obtain that the scheme (4.29) will be submonotone iff

$$\max_{\lambda \in s(\mathbf{A}_{i,j})} [|1 - (\gamma_x + \gamma_y)\, \lambda| + |\gamma_x \lambda| + |\gamma_y \lambda|] \le 1 \quad \forall i, j, \tag{4.32}$$

where $s(\mathbf{A}_{i,j})$ denotes the spectrum of the matrix $\mathbf{A}_{i,j}$. In view of (4.32), we conclude that the scheme (4.29) will be submonotone iff the eigenvalues of the matrices $\mathbf{A}_{i,j}$ are all real nonnegative numbers, and

$$\tau \le \frac{h_x h_y}{(h_x + h_y) \max\limits_{i,j} \|\mathbf{A}_{i,j}\|_2}. \tag{4.33}$$

Let us note that if the matrices $\mathbf{A}_{i,j}$ are not permutable, then these conditions ((4.33) and nonnegativity of the eigenvalues of $\mathbf{A}_{i,j}$) will be necessary only for submonotonicity of the scheme (4.29).

4.2. Interconnection between several notions of monotonicity. As the total variation diminishing (TVD) idea [6], [7], [20] was developed to produce monotone schemes, i.e., the schemes free of spurious oscillations, we will address the interrelation between monotonicity, submonotonicity, and the TVD notions. To start with, let us assume that the explicit scheme (2.51) is scalar and obeys the TVD notion, i.e., $\mathbf{z} = \mathbf{A} \cdot \mathbf{x}$ $(z_i = a_i^j x_j,\ i, j = 1, \ldots, m)$ and $TV(\mathbf{z}) \le TV(\mathbf{x})$ $(\mathbf{x}, \mathbf{z} \in L)$. Let us note that the seminorm TV is associated with quotient (factor) spaces. We will investigate the characteristics of TVD using projections onto norm spaces. Let L/κ denote the quotient space of a vector space L modulo κ, where $\kappa = \{\mathbf{x} \,|\, TV(\mathbf{x}) = 0\}$ is the null-space of the seminorm TV. The projection $L \to L/\kappa$, i.e., $\mathbf{x} \to [\mathbf{x}] \equiv \widetilde{\mathbf{x}}$, $(\mathbf{x} \in L, [\mathbf{x}] \in L/\kappa)$, can be done by $\widetilde{x}_i = x_{i+1} - x_i,\ i = 1, \ldots, m-1$. By virtue of this projection we obtain the associate scheme $\widetilde{\mathbf{z}} = \widetilde{\mathbf{A}} \cdot \widetilde{\mathbf{x}}$ $(\widetilde{z}_i = \widetilde{a}_i^j \widetilde{x}_j,\ i, j = 1, \ldots, m-1)$, where $\widetilde{a}_i^j = \sum_{k=1}^{j} (a_i^k - a_{i+1}^k)$. Having $\widetilde{\mathbf{x}} = [\mathbf{x}]$ and $TV(\mathbf{x}) = \|\widetilde{\mathbf{x}}\|_1 \equiv \sum_i |\widetilde{x}_i|$, the matrix norm of $\widetilde{\mathbf{A}}$ subject to the vector norm $\|\widetilde{\mathbf{x}}\|_1$ fulfills the condition $\|\widetilde{\mathbf{A}}\|_1 \le 1$ insofar as $\|\widetilde{\mathbf{z}}\|_1 = TV(\mathbf{z}) \le TV(\mathbf{x}) = \|\widetilde{\mathbf{x}}\|_1$. Hence, we conclude that the associate scheme is submonotone on the basis that the original explicit scheme is TVD. So, the TVD notion could be associated with the notion of submonotonicity rather than monotonicity (cf. [27]).

It remains to show that the novel maximum principle (see Definitions 2.4, 2.12, and 3.1) is a consequence of each of the previous versions.

The boundary maximum principle can be formulated as follows [4], [19], [25], [33]. The equation system (1.8) satisfies the boundary maximum principle if the following inequalities are valid for its solution:

$$\min_{j \in \gamma} y_j \le y_i \le \max_{j \in \gamma} y_j \quad \forall i \in \overline{\gamma}, \tag{4.34}$$

where $\gamma\ (= [M', M])$ denotes the subset of boundary nodes. Let ω be the SRC-subset; then $\gamma \subseteq \omega$ and hence

$$\min_{j \in \omega} y_j \le \min_{j \in \gamma} y_j, \quad \max_{j \in \gamma} y_j \le \max_{j \in \omega} y_j. \tag{4.35}$$

By virtue of (4.34) and (4.35), we obtain

$$-\max_{j\in\omega}|y_j|\le\min_{j\in\omega}y_j\le y_i\le\max_{j\in\omega}y_j\le\max_{j\in\omega}|y_j|\quad\forall i\in\overline{\gamma} \tag{4.36}$$

or

$$\max_{i\in\overline{\gamma}}|y_i|\le\max_{j\in\omega}|y_j| \tag{4.37}$$

and, since $\overline{\omega}\subseteq\overline{\gamma}$, we can write

$$\max_{i\in\overline{\omega}}|y_i|\le\max_{j\in\omega}|y_j|\,. \tag{4.38}$$

In view of (4.38), the equality (2.21) is valid, and hence the boundary maximum principle implies the maximum principle formulated by Definition 2.4.

Let the maximum principle for inverse column entries [33] be valid. Assume that $\overline{\Omega}_0$ in (1.10) coincides with the SRC-subset ω. In such a case (1.10) implies (2.21), and hence the novel maximum principle (Definition 2.4) holds. The assertion that the novel maximum principle is a consequence of the maximum principle for the absolute values [31] can be proven in perfect analogy to the previous one.

Let us assume that the regional maximum principle [33] holds for the equation system

$$\mathbf{D}\cdot\mathbf{z}=\mathbf{B}\cdot\mathbf{v}+\tau\mathbf{C}\cdot\mathbf{w},\quad\tau>0; \tag{4.39}$$

i.e., we have the estimates

$$\min_{k\in\Omega}v_k+\tau\min_{k\in\Omega}w_k\le z_i\le\max_{k\in\Omega}v_k+\tau\max_{k\in\Omega}w_k\quad\forall i\in\Omega, \tag{4.40}$$

where $\Omega=\{1,2,\ldots,M\}$. Let us note that (4.40) implies

$$|z_i|\le\max_{k\in\Omega}|v_k|+\tau\max_{k\in\Omega}|w_k|\quad\forall i\in\Omega, \tag{4.41}$$

which can be written in the form

$$\|\mathbf{z}\|_\infty\le\|\mathbf{v}\|_\infty+\tau\|\mathbf{w}\|_\infty\,. \tag{4.42}$$

Notice that the condition $\mathbf{v}=\mathbf{w}=0$ implies, by virtue of (4.40), the unique solution $\mathbf{z}=0$ to (4.39), and hence $\mathbf{D}$ in (4.39) is nonsingular. Then, we can rewrite (4.39) as

$$\mathbf{z}=\mathbf{A}\cdot\mathbf{x},\quad\mathbf{A}=\left\{\mathbf{D}^{-1}\cdot\mathbf{B}\quad\mathbf{D}^{-1}\cdot\mathbf{C}\right\},\ \mathbf{x}=\left\{\mathbf{v}^T,\tau\mathbf{w}^T\right\}^T. \tag{4.43}$$

Let us assume that the vector norm of $\mathbf{x}$ in (4.43) is $h_x(\mathbf{x})=\|\mathbf{v}\|_\infty+\tau\|\mathbf{w}\|_\infty$. In such a case, in view of (4.42), we obtain (3.3), and hence the maximum principle (Definition 3.1) holds for (4.39) with respect to $h_z(\mathbf{z})\equiv\|\mathbf{z}\|_\infty$ and $h_x(\mathbf{x})$.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics,* 2nd ed., Benjamin/Cummings Publishing Company, Reading, MA, 1978.

[2] V. V. Akimenko, *The maximum principle and nonlinear monotone schemes for parabolic equations*, Comput. Math. Math. Phys., 39 (1999), pp. 805–816.

[3] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover Publications, New York, 1991.

[4] P. G. Ciarlet, *Discrete maximum principle for finite-difference operators*, Aequationes Math., 4 (1970), pp. 338–352.

[5] L. Collatz, *Functional Analysis and Numerical Mathematics*, Academic Press, New York, 1966.

[6] V. G. Ganzha and E. V. Vorozhtsov, *Numerical Solutions for Partial Differential Equations,* CRC Press, New York, 1996.

[7] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[8] N. N. Kalitkin, *Numerical Methods*, Nauka, Moscow, 1978 (in Russian).

[9] N. N. Krasovskii, *Stability of Motion*, Stanford University Press, Stanford, CA, 1963.

[10] O. A. Ladyzhenskaya and N.N. Ural'tseva, *Linear and Quasi-Linear Elliptic Equations*, Academic Press, New York, 1968.

[11] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics,* Appl. Math. Sci. 49, Springer-Verlag, New York, 1985.

[12] P. Lancaster, *Theory of Matrices*, Academic Press, New York, 1969.

[13] L. Leboucher, *Monotone scheme and boundary conditions for finite volume simulation of magnetohydrodynamic internal flows at high Hartman number*, J. Comput. Phys., 150 (1999), pp. 181–198.

[14] M. Lobo and A. F. Emery, *The discrete maximum principle in finite-element thermal radiation analysis*, Numerical Heat Transfer, 24 (1993), pp. 209–227.

[15] J. Lorenz and W. Mackens, *Toeplitz matrices with totally nonnegative inverses*, Linear Algebra Appl., 24 (1979), pp. 133–141.

[16] L. A. Lyusternik, *Über einige Anwendungen der direkten Methoden in Variationsrechnung*, Mat. Sb., 33 (1926), pp. 173–202.

[17] A. N. Michel and C. J. Herget, *Applied Algebra and Functional Analysis*, Dover Publications, New York, 1993

[18] L. Mirsky, *An Introduction to Linear Algebra*, Dover Publications, New York, 1990.

[19] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, 1996.

[20] V. V. Ostapenko, *On the strong monotonicity of nonlinear difference schemes*, Comput. Math. Math. Phys., 38 (1998), pp. 1119–1133.

[21] C. V. Pao, *Numerical analysis of coupled systems of nonlinear parabolic equations*, SIAM J. Numer. Anal., 36 (1999), pp. 393–416.

[22] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

[23] R. D. Richtmyer and K.W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed., Wiley-Interscience, New York, 1967.

[24] A. A. Samarskiy and A.V. Gulin, *Stability of Finite Difference Schemes*, Nauka, Moscow, 1973 (in Russian).

[25] A. A. Samarskiy, *Theory of Finite Difference Schemes*, Nauka, Moscow, 1977 (in Russian).

[26] A. A. Samarskiy and E. S. Nikolaev, *Methods for Solving Difference Equations*, Nauka, Moscow, 1978 (in Russian).

[27] A. A. Samarskii and P. N. Vabishchevich, *Nonlinear monotone schemes for the transport equation*, Dokl. Math., 361 (1998), pp. 15–17.

[28] S. R. Searle, *Matrix Algebra for the Biological Sciences,* John Wiley & Sons, New York, 1966.

[29] V. V. Smelov, *Algebraic aspect of the discrete maximum principle*, Russ. J. Numer. Anal. Math. Modelling, 16 (2001), pp. 175–190.

[30] G. Stoyan, *On monotone difference schemes for weakly coupled system of partial differential equations*, in Computational Mathematics, Banach Center Pub. 13, A. Wakulicz, ed., PWN-Polish Scientific Publishers, Warsaw, 1984, pp. 33–43.

[31] G. Stoyan, *On maximum principle for monotone matrices*, Linear Algebra Appl., 78 (1986), pp. 147–161.

[32] D. Thangaraj and A. Nathan, *A rotated monotone difference scheme for the two-dimensional anisotropic drift-diffusion equation*, J. Comput. Phys., 145 (1998), pp. 445–461.

[33] G. Windish, *M-matrices in Numerical Analysis*, BSB Teubner, Leipzig, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1136–1148

A NEW CLASS OF INVERSE M-MATRICES OF TREE-LIKE TYPE*

SERVET MARTÍNEZ†, JAIME SAN MARTÍN†, AND XIAO-DONG ZHANG‡

Abstract. In this paper, we use weighted dyadic trees to introduce a new class of nonnegative matrices whose inverses are column diagonally dominant M-matrices.

Key words. nonnegative matrix, inverse M-matrix, weighted dyadic tree

AMS subject classifications. 15A09, 05C50, 15A57

PII. S0895479801396816

1. Introduction.

It is a longstanding and difficult problem to characterize all nonnegative matrices whose inverses are M-matrices, although inverses of all nonsingular M-matrices are always nonnegative matrices. In 1977, Willoughby [16] called the problem of finding or characterizing nonnegative matrices whose inverses are M-matrices the *inverse M-matrix problem*. Johnson [7], Fiedler, Johnson, and Markham [6], and Fiedler [4] devoted much effort to general properties of inverse M-matrices. For definitions, references, and background on M-matrices and the inverse M-matrix problem, the reader is referred to Berman and Plemmons [1] and Johnson [7]. However, until now there have been just a few known classes of inverse M-matrices. The oldest class of symmetric inverse M-matrices is the class of positive type D matrices defined by Markham [8]. In 1994, Martínez, Michon, and San Martín introduced a strictly symmetric ultrametric matrix $A = (a_{ij})$ whose entries satisfy

$$a_{ij} \geq \min\{a_{ik}, a_{kj}\} \quad \text{for all} \quad i, j, k,$$

$$a_{ii} > \max_{j \neq i} a_{ij} \quad \text{for all} \quad i$$

and proved that inverses of strictly symmetric ultrametric matrices are row and column diagonally dominant M-matrices (see [9] and also [13]). Later, nonsymmetric ultrametric matrices were independently introduced by McDonald et al. [11] and Nabben and Varga [14], i.e., nested block form and generalized ultrametric matrices. After a suitable permutation, every generalized ultrametric matrix can be put into nested block form, which contains type D matrices. Recently, Fiedler [5] introduced a new class of inverse M-matrices. Furthermore, Nabben [12] was motivated by Fiedler's result and introduced a new class of inverse M-matrices.

We have been motivated by the results in [3], [5], [10], [11], [14], and [12] to introduce in section 2 a new class of nonnegative matrices by using weighted dyadic trees. We state the following condition under which our main result holds: their

*Received by the editors October 18, 2001; accepted for publication (in revised form) by R. Nabben September 23, 2002; published electronically March 13, 2003. This research was supported by FONDAP in Applied Mathematics.

http://www.siam.org/journals/simax/24-4/39681.html

†Departamento de Ingeniería Matemática y Centro de Modelamiento Matemático, UMR 2071 CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170-3 Correo 3, Santiago, Chile (smartine@dim.uchile.cl, jsanmart@dim.uchile.cl).

‡Centro de Modelamiento Matemático, UMR 2071 CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170-3 Correo 3, Santiago, Chile. Current address: Shanghai Jiao Tong University, Shanghai, China (xiaodong@sjtu.edu.cn).

inverses are column diagonally dominant M-matrices. In section 3, some preliminary properties and lemmas are presented. In particular, it is shown that these weighted tree matrices admit a representation that we call the quasi-nested block form. The proof of the main result is supplied in section 4. Finally, in section 5, we study the class of all the permutations, which leads to the matrix being presented in a quasi-nested block form.

2. Definitions and main result. Let $T = (V, E)$ be a tree on n vertices and edge set E. Sometimes we also write $V = V(T)$, $E = E(T)$. For any two vertices s and t, there is a unique path $geod(s, t)$ from vertex s to vertex t. In particular $geod(s, s) = \{s\}$. Let vertex $r \in V$ be a root of the tree T. We may define a partial order relation "$\preceq$" on T: $s \preceq t$ if and only if $s \in geod(r, t)$. Moreover, for $s, t \in V$, $s \wedge t = \sup\{v : v \in geod(r, s) \cap geod(r, t)\}$ denotes the closest common ancestor of s and t. Thus $\mathbf{s}(t) = \{v \in V : t \preceq v, (t, v) \in E\}$ is the set of successors of t, and $I = \{i \in T : \mathbf{s}(i) = \emptyset\}$ is the set of leaves of the tree T. A tree is called *dyadic* if the cardinality of set $\mathbf{s}(t)$ is $|\mathbf{s}(t)| = 2$ for $t \notin I$. For vertex $t \notin I$ of a dyadic tree T, its successors are signed and denoted by t^- and t^+ (the signs $-$ or $+$ of the successors are fixed). In addition, since vertex $t \in T$ and the set $L(t) = \{i \in I : t \in geod(r, i)\}$ are in one-to-one correspondence relations, we may identify $L(t)$ with t. Thus, the root r is identified with I. The distinction between the roles of $L \in V$ and $L \subseteq I$ will be clear in the context when we use them. We usually say "element L" when referring to $L \in V$ and "set L" to mean $L \subseteq I$.

For $L \in T$, we denote by $T_L = (V_L, E_L)$ the dyadic subtree rooted by L, that is, $V_L = \{v \in V : L \preceq v\}$, $E_L = E \cap (V_L \times V_L)$. Its leaves are the elements of L. For $v \in V_L$, its signed successors in T_L coincide with its signed successors in T.

For a dyadic tree T, its set I of leaves can be totally ordered as follows: $i \leq j$ if $i \in t^-$, $j \in t^+$, where $t = i \wedge j$. We denote by $P^\phi : I \to \{1, \ldots, n\}$ the permutation which assigns i to its rank in the total order and we call it the *canonical permutation*.

DEFINITION 2.1. *A matrix $U = (u_{ij} : i, j \in I)$ is called a $\mathcal{W}$ matrix if there exists a dyadic tree $T = (V, E)$ with set I of leaves and nonnegative vectors $\overrightarrow{\alpha} = (\alpha_i : i \in V)$, $\overrightarrow{\beta} = (\beta_i : i \in V)$ satisfying that*

(i) *$\alpha_i = \beta_i > 0$ for $i \in I$;*

(ii) *$0 \leq \alpha_i \leq 1$ and $0 \leq \beta_i \leq 1$ for $i \in V \setminus I$;*

(iii) *β is $\preceq$-increasing in $V \setminus I$, that is, $s \preceq t \in V \setminus I$ implies $\beta_s \leq \beta_t$;*

(iv) *$u_{ij} = \alpha_i \Pi_{(l,l^-) \in geod(t,i)} \alpha_l$ if $(i, j) \in (t^-, t^+)$, and $u_{ij} = \beta_t \alpha_i \Pi_{(l,l^-) \in geod(t,i)} \alpha_l$ if $(i, j) \in (t^+, t^-)$, where $t = i \wedge j$;*

(v) *$u_{ii} = \alpha_i$ for $i \in I$.*

The matrix U is said to be supported by the dyadic tree T and defined by $\overrightarrow{\alpha}$, $\overrightarrow{\beta}$ on T.

For $J, K \subseteq I$, denote $U_{JK} = (u_{ij} : i \in J, j \in K)$. It is easy to see that if U is a $\mathcal{W}$ matrix supported by $T = (V, E)$ and $L \in V$, then U_{LL} is also a $\mathcal{W}$ matrix supported by T_L and defined by the restricted vectors $\overrightarrow{\alpha}|_{V_L}$ and $\overrightarrow{\beta}|_{V_L}$ on V_L.

The main result of this paper is the following.

THEOREM 2.2. *Let U be a $\mathcal{W}$ matrix. If U does not contain a row of zeros and no two columns in U are the same, then U is nonsingular and its inverse is a column diagonally dominant M-matrix.*

3. Preliminaries and lemmas. In this section, we first present an equivalent condition for $U \in \mathcal{W}$.

DEFINITION 3.1. *Let $C = (c_{ij})$ be a nonnegative matrix of order n with positive main diagonal elements. We define inductively as follows what it means for C to be in quasi-nested block form:*

(i) *If $n = 1$, then C is in quasi-nested block form.*

(ii) *If $n > 1$, and quasi-nested block form has been defined for all $k \times k$ nonnegative matrices with $k < n$, then C is in quasi-nested block form if*

$$C = \begin{pmatrix} C_{11} & b_{12} b_J e_K{}^T \\ b_{21} b_K e_J{}^T & C_{22} \end{pmatrix},$$

where C_{11} and C_{22} are $n_1 \times n_1$ and $n_2 \times n_2$ square matrices in quasi-nested block form with $n_1 \geq 1$, $n_2 \geq 1$, $n = n_1 + n_2$; b_J and b_K are the last columns of C_{11} and C_{22}, respectively; e is a vector of all ones with suitable dimension; $0 \leq b_{12} \leq 1$, $0 \leq b_{21} \leq 1$; and $c_{ij} \geq c_{ik}$ for all $k \geq j \geq i$, $c_{ij} \geq c_{ik}$ for all $i \geq j \geq k$.

THEOREM 3.2. *U is a $\mathcal{W}$ matrix if and only if there exists a permutation matrix P such that PUP^T is a matrix in quasi-nested block form. Moreover, P can be taken to be the matrix associated with the canonical permutation P^ϕ.*

Proof. Necessity. We prove the assertion by induction on n, the dimension of U. It is clear for $n = 1, 2$. Assume that the assertion holds for less than n. Let us consider the total order $\leq$ on I defined by the dyadic tree T supporting U. The successors of the root I are denoted by $J = I^-$ and $K = I^+$. Then there exists a permutation matrix P such that

$$PUP^T = \begin{pmatrix} U_{JJ} & U_{JK} \\ U_{KJ} & U_{KK} \end{pmatrix},$$

where the matrices U_{JJ} and U_{KK} are $\mathcal{W}$ matrices. We denote by n_1 and n_2 the orders of U_{JJ} and U_{KK}, respectively. Clearly $n_1 > 0$, $n_2 > 0$, and $n_1 + n_2 = n$. Hence by the induction hypothesis, there exist permutation matrices Q_J and Q_K such that $Q_J U_{JJ} Q_J^T = C_{11}$ and $Q_K U_{KK} Q_K^T = C_{22}$ are matrices in quasi-nested block form. Moreover, Q_J and Q_K can be taken to be the matrices associated with permutations $Q_J^{\phi_1}$ and $Q_K^{\phi_2}$, respectively.

Let $P_1 = \operatorname{diag}(Q_J, Q_K)P$. Then

$$P_1 U P_1^T = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} := C.$$

For $i \leq n_1 < j$, since $i \wedge j = I$ and $i \in I^-$, we get $c_{ij} = \alpha_I(\alpha_i \prod_{(l,l^-) \in geod(I^-,i)} \alpha_l) = \alpha_I c_{i,n_1}$. Hence $C_{12} = \alpha_I b_J e_K{}^T$, where b_J is the last column of C_{11}. By a similar argument, we may show that $C_{21} = \beta_I b_K e_J^T$, where b_K is the last column of C_{22}.

Let $i \leq j \leq k$. If $i \leq j \leq k \leq n_1$ or $n_1 < i \leq j \leq k$, then by the induction hypothesis, $c_{ij} \geq c_{ik}$; if $i \leq j \leq n_1 < k$, also by the induction hypothesis we get $c_{ij} \geq c_{i,n_1} \geq c_{i,n_1}\alpha_I = c_{ik}$; and in the case $i \leq n_1 < j \leq k$, we find directly $c_{ij} = c_{ik}$. Let $i \geq j \geq k$. If $i \geq j \geq k \geq n_1$ or $n_1 > i \geq j \geq k$, then by the induction hypothesis, $c_{ij} \geq c_{ik}$; if $i > n_1 \geq j \geq k$, then $c_{ij} = c_{ik}$; and if $i \geq j > n_1 \geq k$, then $i \wedge k = I$, $i \wedge j = t$, and

$$c_{ij} = \alpha_i \beta_t \Pi_{(l,l^-) \in geod(t,i)} \alpha_l \text{ and } c_{ik} = \alpha_i \beta_I \Pi_{(l,l^-) \in geod(t,i)} \alpha_l \Pi_{(l,l^-) \in geod(I,t)} \alpha_l$$

since $0 \leq \alpha_l \leq 1$ and $\beta_I \leq \beta_t$, we have $c_{ij} \geq c_{ik}$. Hence C is a matrix in quasi-nested block form. Moreover, with this construction, an induction argument shows that the final P_1 will correspond to the canonical permutation P^ϕ.

Sufficiency. We proceed as before by induction on the size of the matrix. For $n = 2$,

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix},$$

where $c_{12} \le c_{11}$ and $c_{21} \le c_{22}$. Let T be a dyadic tree with tree elements $V = \{I, I^-, I^+\}$, $\alpha_{I^-} = \beta_{I^-} = c_{11}$, $\alpha_{I^+} = \beta_{I^+} = c_{22}$, $\alpha_I = \frac{c_{12}}{c_{11}}$, and $\beta_I = \frac{c_{21}}{c_{22}}$. The matrix U with support tree T is just C. Hence the assertion holds for $n = 2$. Assume that the assertion holds when the dimension is less than n. By the definition of matrix C in quasi-nested block form,

$$C = \begin{pmatrix} C_{11} & b_{12}c_J e_K^T \\ b_{21}c_K e_J^T & C_{22} \end{pmatrix},$$

where C_{ii} is a matrix of order n_i in quasi-nested block form for $i = 1, 2$ and both c_J and c_K are the last columns of C_{11} and C_{22}, respectively. By the induction hypothesis, there exist two dyadic trees T_1 and T_2 with roots J and K and $\overrightarrow{\alpha} = (\alpha_t : t \in V(T_i))$, $\overrightarrow{\beta} = (\beta_t : t \in V(T_i))$ for $i = 1, 2$. Now we define a new tree T obtained from $T_1 \cup T_2$ by adding a new root vertex I associated with $\alpha_I = b_{12}$ and $\beta_I = b_{21}$ and two edges (I, J) and (I, K), where $J = I^-$ and $K = I^+$. Then the matrix associated with T has the following form:

$$U = \begin{pmatrix} U_{JJ} & U_{12} \\ U_{21} & U_{KK} \end{pmatrix} = \begin{pmatrix} C_{11} & U_{12} \\ U_{21} & C_{22} \end{pmatrix}.$$

For $i \le n_1 < j$, $u_{ij} = \alpha_i \Pi_{(l,l^-) \in geod(I,i)} \alpha_l = \alpha_I \alpha_i \Pi_{(l,l^-) \in geod(I^-,i)} \alpha_l = \alpha_I u_{i,n_1}$. Hence $U_{12} = b_{12}c_J e_K^T$, where c_J is the last column of $U_{JJ} = C_{11}$. Similarly, $U_{21} = b_{21}c_K e_J^T$, where c_K is the last column of $U_{KK} = C_{22}$. Therefore $U = C$ and C is a $\mathcal{W}$ matrix. Since the permutation matrix P corresponds to renumbering of the vertices, PCP^T is still a $\mathcal{W}$ matrix. □

LEMMA 3.3. *Let $U = (u_{ij} : i, j \in I)$ be a $\mathcal{W}$ matrix associated with tree T in quasi-nested block form and $\overrightarrow{\alpha}, \overrightarrow{\beta}$. If $0 \le \delta \le \beta_I$ and $\delta < 1$, then $\widetilde{U} = U - \delta b_I e^T$ is still a $\mathcal{W}$ matrix associated with T and $\widetilde{\alpha_I} = \frac{(1-\delta)\alpha_I}{1-\delta\alpha_I}, \widetilde{\beta_I} = \frac{\beta_I - \delta}{1-\delta}$, where b_I is the last column of U.*

Proof. We assume $I = \{1, \ldots, n\}$ is totally ordered by the tree T. We proceed on n, the dimension of matrix U. If U is a 2×2 matrix with the root I of the tree T and the set $\{1, 2\}$ of leaves, then we assume $1 = I^-$, $2 = I^+$. Hence

$$U = \begin{pmatrix} \alpha_1 & \alpha_1\alpha_I \\ \beta_I\alpha_2 & \alpha_2 \end{pmatrix},$$

where $0 \le \alpha_I, \beta_I \le 1$. Then

$$\widetilde{U} = \begin{pmatrix} (1-\delta\alpha_I)\alpha_1 & (1-\delta)\alpha_I\alpha_1 \\ (\beta_I - \delta)\alpha_2 & (1-\delta)\alpha_2 \end{pmatrix}.$$

We take the same tree T with vectors $\overrightarrow{\widetilde{\alpha}}, \overrightarrow{\beta}$ given by $\widetilde{\alpha}_1 = (1-\delta\alpha_I)\alpha_1$, $\widetilde{\alpha_2} = (1-\delta)\alpha_2$, and

$$\widetilde{\alpha_I} = \frac{(1-\delta)\alpha_I}{1-\delta\alpha_I}, \quad \widetilde{\beta_I} = \frac{\beta_I - \delta}{1-\delta}.$$

It is clear that $0 \le \widetilde{\alpha}_I, \widetilde{\beta}_I \le 1$ and that $\widetilde{U}$ is just the matrix defined by vectors $\overrightarrow{\widetilde{\alpha}}$, $\overrightarrow{\widetilde{\beta}}$ on the tree T. Hence the assertion holds for $n = 2$. Assume that the assertion holds when the dimension of a matrix is less than n. Let U be an $n \times n$ matrix. By Theorem 3.2, we may assume that

$$U = \begin{pmatrix} U_{JJ} & \alpha_I b_J e_K^T \\ \beta_I b_K e_J^T & U_{KK} \end{pmatrix}$$

is associated with tree T, U_{JJ} with subtree T_1, and tree U_{KK} with subtree T_2. Then $b_I = \binom{\alpha_I b_J}{b_K}$ and

$$\widetilde{U} = U - \delta b_I e^T = \begin{pmatrix} U_{JJ} - \delta\alpha_I b_J e_J^T & (1-\delta)\alpha_I b_J e_K^T \\ (\beta_I - \delta) b_K e_J^T & U_{KK} - \delta b_K e_K^T \end{pmatrix} := \begin{pmatrix} \overline{U}_{JJ} & \overline{U}_{12} \\ \overline{U}_{21} & \overline{U}_{KK} \end{pmatrix},$$

where b_J and b_K are the last columns of U_{JJ} and U_{KK}, respectively. Since $\overrightarrow{\beta}$ is increasing and $0 \le \alpha_I \le 1$, we have $0 \le \delta\alpha_I \le \beta_I\alpha_I \le \beta_J$ and $\delta\alpha_I < 1$. Hence by the induction hypothesis, $U_{JJ} - \delta\alpha_I b_J e_J^T = \overline{U}_{JJ}$ is a $\mathcal{W}$ matrix defined by vectors $(\widetilde{\alpha}_t : t \in V(T_1))$ and $(\widetilde{\beta}_t : t \in V(T_1))$ on the subtree T_1. Moreover,

$$\widetilde{\alpha}_J = \frac{(1-\delta\alpha_I)\alpha_J}{1-\delta\alpha_I\alpha_J}, \quad \widetilde{\beta}_J = \frac{\beta_J - \delta\alpha_I}{1-\delta\alpha_I}.$$

By a similar argument, $U_{KK} - \delta b_K e_K^T = \overline{U}_K$ is a $\mathcal{W}$ matrix associated with subtree T_2 and vectors $(\widetilde{\alpha}_t : t \in V(T_2))$ and $(\widetilde{\beta}_t : t \in V(T_2))$. Moreover,

$$\widetilde{\alpha}_K = \frac{(1-\delta)\alpha_K}{1-\delta\alpha_K}, \quad \widetilde{\beta}_K = \frac{\beta_K - \delta}{1-\delta}.$$

Define $\widetilde{\alpha}_I = \frac{(1-\delta)\alpha_I}{1-\delta\alpha_I}$, $\widetilde{\beta}_I = \frac{\beta_I - \delta}{1-\delta}$. We have $0 \le \widetilde{\alpha}_I, \widetilde{\beta}_I \le 1$ and

$$\widetilde{\beta}_I = \frac{\beta_I - \delta}{1-\delta} \le \frac{\beta_K - \delta}{1-\delta} = \widetilde{\beta}_K,$$

$$\widetilde{\beta}_I = \frac{\beta_I - \delta}{1-\delta} \le \frac{\beta_I - \delta\alpha_I}{1-\delta\alpha_I} \le \frac{\beta_J - \delta\alpha_I}{1-\delta\alpha_I} = \widetilde{\beta}_J.$$

Then the matrix X associated with the tree T and vectors $(\widetilde{\alpha}_t : t \in V(T))$, $(\widetilde{\beta}_t : t \in V(T))$ is just $\widetilde{U}$. In fact, $0 \le \widetilde{\alpha}_t, \widetilde{\beta}_t \le 1$ for $t \in V \setminus I$ and $\widetilde{\beta}$ is increasing in $V \setminus I$. For $i, j \in I^- = J$ or $i, j \in I^+ = K$, $X_{ij} = (\overline{U}_{JJ})_{ij} = \widetilde{U}_{ij}$ or $X_{ij} = (\overline{U}_{KK})_{ij} = \widetilde{U}_{ij}$; for $i \in J$, $j \in K$, and $|J| = n_1$, $X_{ij} = \widetilde{\alpha}_i \Pi_{(l,l^-) \in geod(I,i)} \widetilde{\alpha}_l = \widetilde{\alpha}_I X_{i,n_1} = \widetilde{\alpha}_I (\overline{U}_{JJ})_{i,n_1} = (1-\delta)\alpha_I (U_{JJ})_{i,n_1} = (\widetilde{U})_{ij}$; for $i \in K$, $j \in J$, $X_{ij} = \widetilde{\alpha}_i \widetilde{\beta}_I \Pi_{(l,l^-) \in geod(I,i)} \widetilde{\alpha}_l = \widetilde{\beta}_I \widetilde{\alpha}_i \Pi_{(l,l^-) \in geod(s,i)} \widetilde{\alpha}_l = \widetilde{\beta}_I X_{in} = (\widetilde{U})_{ij}$, where $i \wedge n = s$, since each edge from vertex I to vertex s is (t, t^+). This completes our proof. □

4. Proof of Theorem 2.2.

LEMMA 4.1. *Let U be a $\mathcal{W}$ matrix defined by vectors $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$ on tree T. Then U does not contain a row of zeros and no two columns in U are the same if and only if $\alpha_t\beta_t < 1$ for $t \in V(T) \setminus I$ and $\alpha_i > 0$ for $i \in I$, where I is the set of leaves of T.*

Proof. Necessity. We use the induction on the size of matrix U. It is clear that the assertion holds for $|I| = 1, 2$. Since U does not contain a row of zeros, $U_{ii} = \alpha_i > 0$

for $i \in I$. Let $J = I^-$ and $K = I^+$. It is easy to see that no two columns in U_{JJ} and U_{KK} are the same. By the induction hypothesis, it suffices to verify that $\alpha_I\beta_I < 1$. Assume that

$$U = \begin{pmatrix} U_{JJ} & \alpha_I b_J {e_K}^T \\ \beta_I b_K {e_J}^T & U_{KK} \end{pmatrix}.$$

If $\alpha_I\beta_I = 1$, then $\alpha_I = \beta_I = 1$. Hence the $|I^-|$th and nth columns are the same, which is a contradiction. Thus $\alpha_I\beta_I < 1$.

Conversely, since $\alpha_i > 0$ it is clear that the assertion holds for $n = 1, 2$. We may assume that

$$U = \begin{pmatrix} U_{JJ} & \alpha_I b_J e_K^T \\ \beta_I b_K e_J^T & U_{KK} \end{pmatrix},$$

where U_{JJ} is an $n_1 \times n_1$ matrix. By the induction hypothesis, no two columns in U_{JJ} and U_{KK} are the same. Suppose that the ith and jth columns in U are the same with $i < j$. Then $i \le n_1 < j$ and $U_{ii} = U_{ij}$, $U_{ji} = U_{jj}$. On the other hand, $U_{ij} = \alpha_I U_{i,n_1} \le U_{ii}$ and $U_{ji} = \beta_I U_{jn} \le U_{jn} \le U_{jj}$. Hence $\alpha_I\beta_I = 1$, a contradiction. Therefore no two columns in U are the same. □

Now we may present the proof of Theorem 2.2.

Proof of Theorem 2.2. We use induction with respect to the size of the matrix U. For $n = 2$, it is easy to see that $\det(U) = (1 - \alpha_I\beta_I)\alpha_1\alpha_2 > 0$ and

$$U^{-1} = \begin{pmatrix} \alpha_1 & \alpha_1\alpha_I \\ \beta_I\alpha_2 & \alpha_2 \end{pmatrix}^{-1} = \frac{1}{\det(U)} \begin{pmatrix} \alpha_2 & -\alpha_1\alpha_I \\ -\beta_I\alpha_2 & \alpha_1 \end{pmatrix}.$$

Hence U^{-1} is a column diagonally dominant M-matrix. Assume that the assertion holds for less than n. For n, by Theorem 3.2, we may assume that

$$U = \begin{pmatrix} U_{JJ} & \alpha_I b_J e_K^T \\ \beta_I b_K e_J^T & U_{KK} \end{pmatrix}.$$

By Lemma 4.1, U_{JJ} and U_{KK} do not contain a row of zeros and no two columns in U_{JJ} and U_{KK} are the same. By the induction hypothesis, U_{JJ} and U_{KK} are nonsingular. Further, U_{JJ}^{-1} and U_{KK}^{-1} are column diagonally dominant M-matrices. So $\mu_J^t = e^T U_{JJ}^{-1} \ge 0$ and $\mu_K^t = e^T U_{KK}^{-1} \ge 0$. By $\alpha_I\beta_I < 1$ and the Sherman–Morrison formula (see [11]), we have

$$U^{-1} = \begin{pmatrix} U_{JJ}^{-1} + \frac{\alpha_I\beta_I}{1-\alpha_I\beta_I}\varepsilon_J\mu_J^T & -\frac{\alpha_I}{1-\alpha_I\beta_I}\varepsilon_J\mu_K^T \\ -\frac{\beta_I}{1-\alpha_I\beta_I}\varepsilon_K\mu_J^T & U_{KK}^{-1} + \frac{\alpha_I\beta_I}{1-\alpha_I\beta_I}\varepsilon_K\mu_K^T \end{pmatrix} := \begin{pmatrix} C & D \\ E & F \end{pmatrix},$$

where $\varepsilon_J = (0, \ldots, 0, 1)^T$ and $\varepsilon_K = (0, \ldots, 0, 1)^T$. It is easy to see that $D \le 0$ and $E \le 0$. Since $\alpha_I\beta_I \le \beta_J$ and $\alpha_I\beta_I < 1$, by Lemma 3.3, $U_{JJ} - \alpha_I\beta_I b_J e_J^T$ is still a $\mathcal{W}$ matrix. In addition,

$$C = U_{JJ}^{-1} + \frac{\alpha_I\beta_I}{1-\alpha_I\beta_I}\varepsilon_J\mu_J^T = (U_{JJ} - \alpha_I\beta_I b_J e^T)^{-1}$$

is nonsingular. By the induction hypothesis, C is a column diagonally dominant M-matrix. By a similar argument, we may prove that F is a column diagonally

dominant M-matrix. Therefore U^{-1} is an M-matrix. Moreover,

$$e_J^T C + e_K^T E = e_J^T U_{JJ}^{-1} + \frac{\alpha_I\beta_I}{1-\alpha_I\beta_I} e_J^T \varepsilon_J \mu_J^T + \frac{-\beta_I}{1-\alpha_I\beta_I} e_K^T \varepsilon_K \mu_J^T = \frac{1-\beta_I}{1-\alpha_I\beta_I}\mu_J^T \geq 0,$$

$$e_J^T D + e_K^T F = \frac{-\alpha_I}{1-\alpha_I\beta_I} e_J^T \varepsilon_J \mu_K^T + e_K^T U_{KK}^{-1} + \frac{\alpha_I\beta_I}{1-\alpha_I\beta_I} e_K^T \varepsilon_K \mu_K^T = \frac{1-\alpha_I}{1-\alpha_I\beta_I}\mu_K^T \geq 0.$$

Hence U^{-1} is a column diagonally dominant M-matrix. □

Remark 4.2. Neumann in [15] conjectured that the Hadamard product $A \circ A$ is an inverse M-matrix if A is an inverse M-matrix. Clearly, this conjecture is true for $A \in \mathcal{W}$ since $A \circ A \in \mathcal{W}$ (moreover for any $n \geq 1$, $A^{\circ n} \in \mathcal{W}$).

Example 4.3. Let T be a dyadic tree with $\overrightarrow{\alpha}$, $\overrightarrow{\beta}$ defined by Figure 1.

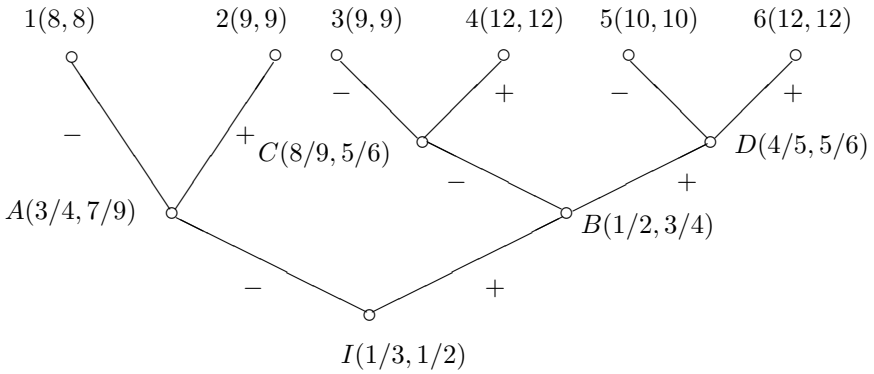


FIG. 1.

Then the matrix U, associated with tree T, and the inverse of U are

$$U = \begin{pmatrix} 8 & 6 & 2 & 2 & 2 & 2 \\ 7 & 9 & 3 & 3 & 3 & 3 \\ 2 & 2 & 9 & 8 & 4 & 4 \\ 3 & 3 & 10 & 12 & 6 & 6 \\ 4 & 4 & 6 & 6 & 10 & 8 \\ 6 & 6 & 9 & 9 & 10 & 12 \end{pmatrix}$$

and

$$U^{-1} = \begin{pmatrix} 0.3000 & -0.2000 & -0.0000 & -0.0000 & -0.0000 & -0.0000 \\ -0.2200 & 0.2800 & -0.0114 & -0.0057 & -0.0160 & -0.0160 \\ -0.0000 & -0.0000 & 0.4286 & -0.2857 & -0.0000 & -0.0000 \\ -0.0000 & -0.0000 & -0.3143 & 0.3429 & -0.0400 & -0.0400 \\ -0.0000 & -0.0000 & -0.0000 & -0.0000 & 0.3000 & -0.2000 \\ -0.0400 & -0.0400 & -0.0800 & -0.0400 & -0.2120 & 0.2880 \end{pmatrix},$$

which is a column diagonally dominant M-matrix.

Remark 4.4. Nabben in [12] described a class of inverse M-matrices whose nested block form is similar to GUMs (generally ultrametric matrices) with the major change being that in the $(2,1)$-block the ee^T was replaced by ce^T, where b corresponds to the

last column of the $(2,2)$-block. From Theorems 3.2 and 2.2, the quasi-nested block form in $\mathcal{W}$ is also similar to GUMs with the major changes being that the $(1,2)$-block was replaced by be^T and the $(2,1)$-block was replaced by ce^T, where b and c are the last columns of the $(1,1)$-block and $(2,2)$-block, respectively. Hence it is natural that the following two questions were proposed.

QUESTION 4.5. *Is it possible to use be^T in the off diagonal blocks, where b is any column of the corresponding diagonal block? Are there any other vectors that will work?*

QUESTION 4.6. *Is it possible to use be^T and ee^T alternately in the nested block form, or must one use one or the other only?*

The following two examples illustrate that the above questions are answered in a negative way.

Example 4.7. Let A be

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad A_{11} = \begin{pmatrix} 8 & 8 \times \frac{1}{8} \\ 10 \times 1 & 10 \end{pmatrix}, \quad A_{12} = \frac{1}{2} \begin{pmatrix} 8 \\ 10 \times 1 \end{pmatrix} (1\ 1),$$

$$A_{21} = \frac{1}{2} \begin{pmatrix} 10 \times \frac{1}{2} \\ 9 \end{pmatrix} (1\ 1), \quad A_{22} = \begin{pmatrix} 10 & 10 \times \frac{1}{2} \\ 9 \times \frac{2}{3} & 9 \end{pmatrix}.$$

But

$$A^{-1} = \begin{pmatrix} 0.1429 & 0.0190 & -0.0333 & -0.0556 \\ -0.1429 & 0.1143 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1500 & -0.0833 \\ 0.0000 & 0.0667 & -0.0833 & 0.1944 \end{pmatrix}$$

is not an M-matrix. Hence in general, we cannot use be^T in the off diagonal blocks for b not being the last column of the corresponding block.

Example 4.8. Let B be

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \quad B_{11} = \begin{pmatrix} 10 & 5 & \frac{2}{5} \times 5 \\ 6 & 10 & \frac{2}{5} \times 10 \\ \frac{1}{2} \times 10 & \frac{1}{2} \times 10 & 10 \end{pmatrix},$$

$$B_{12} = ee^T, \quad B_{21} = 2e^T e, \quad B_{22} = 5,$$

a 4×4 matrix. But

$$B^{-1} = \begin{pmatrix} 0.1439 & -0.0701 & 0.0023 & -0.0152 \\ -0.0708 & 0.1615 & -0.0487 & -0.0084 \\ -0.0350 & -0.0438 & 0.1264 & -0.0095 \\ -0.0152 & -0.0192 & -0.0320 & 0.2133 \end{pmatrix}$$

is not an M-matrix. Hence in general, we cannot use be^T and ee^T alternately in the nested block form.

5. Combinatorial aspects of a $\mathcal{W}$ matrix in quasi-nested block form. In section 3, we have proved that each $\mathcal{W}$ matrix can be put into quasi-nested block form after a suitable permutation. In this section, we try to describe the set of permutations preserving a $\mathcal{W}$ matrix in quasi-nested block form, which is related to the behavior of a sub-Markov chain. The reader is referred to [2] and [3].

We assume that U is a $\mathcal{W}$ matrix in quasi-nested block form with supporting tree T and vectors $\overrightarrow{\alpha}, \overrightarrow{\beta}$, where $I = \{1, 2, \ldots, n\}$. The root of tree T is I and its successors are $I^- = J$ and $I^+ = K$. We also denote $|J| = m$ and write $U[i_1, \ldots, i_t]$ for the principal submatrix of U whose rows and columns are indexed by $1 \le i_1 < i_2 < \cdots < i_t \le n$.

Let $U[i_1, i_2, i_3, i_4]$ be the principal submatrix of U. It is easy to see that $U[i_1, i_2, i_3, i_4]$ is not a $\mathcal{W}$ matrix, in general. But we can obtain a $\mathcal{W}$ in quasi-nested form from $U[i_1, i_2, i_3, i_4]$ by changing the diagonal entries of $U[i_1, i_2, i_3, i_4]$. In fact, without loss of generality, we may assume that $i_1 \wedge i_2 \wedge i_3 \wedge i_4 = P$, $i_1 \wedge i_2 \wedge i_3 = M$, $i_1 \wedge i_2 = N$; $i_1, i_2, i_3 \in P^-$, $i_4 \in P^+$; $i_1, i_2 \in M^-$, $i_3 \in M^+$; $i_1 \in N^-$, $i_2 \in N^+$ (for the other cases, we may show the same result by a similar argument). Let $\gamma_1 = \alpha_{i_1} \prod_{(l,l^-) \in geod(N^-, i_1)} \alpha_l$, $\gamma_2 = \alpha_{i_2} \prod_{(l,l^-) \in geod(N, i_2)} \alpha_l$, $\gamma_3 = \alpha_{i_3} \prod_{(l,l^-) \in geod(M, i_3)} \alpha_l$, and $\gamma_4 = \alpha_{i_4} \prod_{(l,l^-) \in geod(P, i_4)} \alpha_l$; $\gamma_P = \prod_{(l,l^-) \in geod(M,P)} \alpha_l$, $\gamma_M = \prod_{(l,l^-) \in geod(M,N)} \alpha_l$, $\gamma_N = \alpha_N$; $\delta_P = \beta_P$, $\delta_M = \beta_M$, $\delta_N = \beta_N$. Then

$$V_1 = \begin{pmatrix} \gamma_1 & \gamma_1\gamma_N & \gamma_1\gamma_N\gamma_M & \gamma_1\gamma_N\gamma_M\gamma_P \\ \delta_N\gamma_2 & \gamma_2 & \gamma_2\gamma_M & \gamma_2\gamma_M\gamma_P \\ \delta_M\gamma_3 & \delta_M\gamma_3 & \gamma_3 & \gamma_3\gamma_P \\ \delta_P\gamma_4 & \delta_P\gamma_4 & \delta_P\gamma_4 & \gamma_4 \end{pmatrix}$$

is a $\mathcal{W}$ matrix in quasi-nested block form. Hence we may choose a support tree T_1 for V_1 such that the partial order relationship in T_1 is consistent with the partial order relationship in T. Moreover, if $\gamma_t = 1$ or $\delta_t = 1$ for $t \in T_1$, then for the corresponding t in T, we have $\alpha_t = 1$ or $\beta_t = 1$. Hence V_1 is called the *induced $\mathcal{W}$ matrix in quasi-nested block form* from $U[i_1, i_2, i_3, i_4]$. For the principal submatrix $U[i_1, i_2, i_3]$ of U, there is a similar result.

In the rest of this section, we assume U is nonsingular. Hence by Lemma 4.1, $\alpha_t\beta_t < 1$ for any $t \in T \setminus I$. Moreover, we shall also assume that $\varphi : I \mapsto I$ is a permutation such that $U^\varphi := (U_{\varphi(i),\varphi(j)})$ is a $\mathcal{W}$ matrix in quasi-nested block form with support tree T^φ and vectors $\overrightarrow{\alpha^\varphi}, \overrightarrow{\beta^\varphi}$. Let $U^\varphi[i_1, i_2, i_3, i_4]$ be the principal submatrix of U^φ with $1 \le i_1 < i_2 < i_3 < i_4 \le n$. Then there exists a 4×4 permutation matrix P_1 corresponding to rearranging $\varphi^{-1}(i_1), \varphi^{-1}(i_2), \varphi^{-1}(i_3), \varphi^{-1}(i_4)$ in their natural order such that $P_1 U^\varphi[i_1, i_2, i_3, i_4] P_1^T$ is the principal submatrix of U whose rows and columns are indexed by $j_1 < j_2 < j_3 < j_4$, where j_1, j_2, j_3, j_4 are obtained by rearranging $\varphi^{-1}(i_1), \varphi^{-1}(i_2), \varphi^{-1}(i_3), \varphi^{-1}(i_4)$ into their natural order. Hence we have the induced $\mathcal{W}$ matrix V_1 in quasi-nested block form from $U[j_1, j_2, j_3, j_4]$ associated with tree T_1 and $\overrightarrow{\gamma}, \overrightarrow{\delta}$. Moreover, the partial order relationship of $\{\varphi^{-1}(i_1), \varphi^{-1}(i_2), \varphi^{-1}(i_3), \varphi^{-1}(i_4)\}$ in the support tree T_1 is consistent with the partial order relationship of $\{\varphi^{-1}(i_1), \varphi^{-1}(i_2), \varphi^{-1}(i_3), \varphi^{-1}(i_4)\}$ in the support tree T. Therefore, for any $t \in V(T_1)$, $\gamma_t = 1$ ($\delta_t = 1$) implies $\alpha_t = 1$ ($\beta_t = 1$). Moreover, $P_1^T V_1 P_1 := V$ is the induced $\mathcal{W}$ matrix in quasi-nested block form from $U^\varphi[i_1, i_2, i_3, i_4]$.

LEMMA 5.1. *Let $|J| = m$ and $|K| \ge 2$. If there exist $1 \le f < g \le n$ such that $\varphi(f) = n$ and $\varphi(g) = m + 1$, then $\varphi(J) = J$ and $\varphi(K) = K$.*

Proof. We first prove the following claim: There does not exist $f < i < g$ such

that $\varphi(i) := p \le m$.

Assume there exists $f < i < g$ such that $\varphi(i) = p \le m$. Clearly, $p \in I^-$ and $(m+1) \wedge n = K$. Then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^\varphi[f, i, g]$ is

$$V = \begin{pmatrix} \gamma_n & \gamma_n\delta_I & \gamma_n\delta_K \\ \gamma_p\gamma_I & \gamma_p & \gamma_p\gamma_I \\ \gamma_{m+1}\gamma_K & \gamma_{m+1}\gamma_K\delta_I & \gamma_{m+1} \end{pmatrix}.$$

If $f, i \in (f \wedge^\varphi i \wedge^\varphi g)^-$, then $\gamma_I = \delta_I = 1$. Hence $\alpha_I = \beta_I = 1$, a contradiction. If $i, g \in (f \wedge^\varphi i \wedge^\varphi g)^+$, then $\gamma_K = \delta_K = 1$. Hence $\alpha_K = \beta_K = 1$, a contradiction.

By a similar argument, we may prove that there does not exist $i > g$ such that $\varphi(i) = p \le m$. Now let $\varphi(h) > m+1$ and $\varphi(i) \le m$ for $i = 1, \dots, h-1$, where $h \le f$. By a similar argument as used in the proof of the claim, there does not exist $i > h$ such that $\varphi(i) \le m$. Therefore $\varphi(J) = J$ and $\varphi(K) = K$. □

LEMMA 5.2. *Let $|J| = m$ and $|K| \ge 2$. If there exists $1 \le f < g \le n$ such that $\varphi(f) = n$ and $\varphi(g) = m+1$, then $\varphi(i) = i$ for $i \in J$.*

Proof. By Lemma 5.1, $\varphi(J) = J$ and $\varphi(K) = K$. If there exists $1 \le i < j \le m$ such that $\varphi(i) := p > \varphi(j) := q$, then the induced $\mathcal{W}$ matrix of order 4 in quasi-nested block form from $U^\varphi[i, j, f, g]$ is

$$V = \begin{pmatrix} \gamma_p & \gamma_p\delta_L & \gamma_p\gamma_I & \gamma_p\gamma_I \\ \gamma_q\gamma_L & \gamma_q & \gamma_q\gamma_L\gamma_I & \gamma_q\gamma_L\gamma_I \\ \gamma_n\delta_I & \gamma_n\delta_I & \gamma_n & \gamma_n\delta_K \\ \gamma_{m+1}\gamma_K\delta_I & \gamma_{m+1}\gamma_K\delta_I & \gamma_{m+1}\gamma_K & \gamma_{m+1} \end{pmatrix},$$

where $p \wedge q = L$, since $p, q \in I^-$ and $m+1, n \in I^+$. If $j, f, g \in (i \wedge^\varphi j \wedge^\varphi f \wedge^\varphi g)^+$, then $\gamma_K\delta_I\delta_K = \delta_I$, which implies $\gamma_K = \delta_K = 1$. Thus $\alpha_K = \beta_K = 1$, a contradiction. If $i, j \in (i \wedge^\varphi j \wedge^\varphi f \wedge^\varphi g)^-$ and $f, g \in (i \wedge^\varphi j \wedge^\varphi f \wedge^\varphi g)^+$, or $i, j, f \in (i \wedge^\varphi j \wedge^\varphi f \wedge^\varphi g)^-$, then by a similar argument it is easy to see that $\gamma_K = \delta_K = 1$ or $\gamma_I = \delta_I = 1$. Both are contradictions. Hence $\varphi(i) = i$ for $i \in J$. □

COROLLARY 5.3. *If $\alpha_t < 1$, $\beta_t < 1$ for all $t \in V \setminus I$ and $|K| \ge 2$, then there does not exist $f < g$ such that $\varphi(f) = n$ and $\varphi(g) = m+1$.*

Proof. Suppose that there exists $f < g$ such that $\varphi(f) = n$ and $\varphi(g) = m+1$. By Lemma 5.2, $\varphi(i) = i$ for any $i \in J$. Moreover, $f > m$. Hence the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^\varphi[1, f, g]$ is

$$V = \begin{pmatrix} \gamma_1 & \gamma_1\gamma_I & \gamma_1\gamma_I \\ \gamma_n\delta_I & \gamma_n & \gamma_n\delta_K \\ \gamma_{m+1}\gamma_K\delta_I & \gamma_{m+1}\gamma_K & \gamma_{m+1} \end{pmatrix}.$$

If $1, f \in (1 \wedge^\varphi f \wedge^\varphi g)^-$, then $\delta_K = 1$. If $f, g \in (1 \wedge^\varphi f \wedge^\varphi g)^+$, then $\delta_I = 1$, a contradiction. Hence the assertion holds. □

LEMMA 5.4. *Let $\alpha_t < 1$, $\beta_t < 1$ for all $t \in V \setminus I$ and $|K| \ge 2$. If there exists $1 \le f < g \le n$ such that $\varphi(f) = m+1$ and $\varphi(g) = n$, then there does not exist $f < i < g$ such that $\varphi(i) = p \le m$.*

Proof. Suppose that there exists $f < i < g$ such that $\varphi(i) = p \le m$. Then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^\varphi[f, i, g]$ is

$$V = \begin{pmatrix} \gamma_{m+1} & \gamma_{m+1}\gamma_K\delta_I & \gamma_{m+1}\gamma_K \\ \gamma_p\gamma_I & \gamma_p & \gamma_p\gamma_I \\ \gamma_n\delta_K & \gamma_n\delta_I & \gamma_n \end{pmatrix}.$$

By the definition of $\mathcal{W}$ in quasi-nested block form, it is easy to see that $\delta_I = 1$, a contradiction. Hence the assertion holds. ∎

LEMMA 5.5. *Let $\alpha_t < 1$, $\beta_t < 1$ for all $t \in V \setminus I$ and $|K| \geq 2$. If there exists $1 \leq f < g \leq n$ such that $\varphi(f) = m+1$ and $\varphi(g) = n$, then $\varphi(i) \leq m$ for all $i < f$ and $i > g$.*

Proof. We consider the following two cases.

Case 1. Suppose that there exists $i < f$ such that $\varphi(i) = p > m+1$.

If $p, n \in ((m+1) \wedge p \wedge n)^+$, then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^{\varphi}[i, f, g]$ is

$$V = \begin{pmatrix} \gamma_p & \gamma_p\gamma_L\delta_K & \gamma_p\gamma_L \\ \gamma_{m+1}\gamma_K & \gamma_{m+1} & \gamma_{m+1}\gamma_K \\ \gamma_n\delta_L & \gamma_n\delta_K & \gamma_n \end{pmatrix},$$

where $p \wedge n := L$. By the definition of $\mathcal{W}$ in quasi-nested block form, it is easy to see that $\delta_K = 1$. Hence $\beta_K = 1$ and it is a contradiction.

If $m+1, p \in ((m+1) \wedge p \wedge n)^-$, then denote it by $(m+1) \wedge p := M$, and by a similar argument we have $\delta_M = 1$. Hence $\beta_M = 1$ and it is a contradiction.

Case 2. Suppose that there exists $i > g$ such that $\varphi(i) = p > m+1$. By a similar argument as used in the proof of Case 1, it is a contradiction. ∎

LEMMA 5.6. *Let $\alpha_t < 1$, $\beta_t < 1$ for all $t \in V \setminus I$ and $|K| \geq 2$. If there exists $1 \leq f < g \leq n$ such that $\varphi(f) = m+1$ and $\varphi(g) = n$, then there does not exist a pair (i, j) such that $i < f$, $j > g$ and $\varphi(i) \leq m$, $\varphi(j) \leq m$.*

Proof. Suppose that there exist $i < f$ and $j > g$ such that $\varphi(i) := p \leq m$ and $\varphi(j) := q \leq m$. If $p < q$, then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^{\varphi}[i, j, g]$ is

$$V = \begin{pmatrix} \gamma_p & \gamma_p\gamma_L\gamma_I & \gamma_p\gamma_L \\ \gamma_n\delta_I & \gamma_n & \gamma_n\delta_I \\ \gamma_q\delta_L & \gamma_q\gamma_I & \gamma_q \end{pmatrix},$$

where $p \wedge q = L$. By the definition of $\mathcal{W}$ in quasi-nested block form, it is easy to see that $\gamma_I = 1$. Hence $\alpha_I = 1$ and it is a contradiction.

If $p > q$, it is a contradiction by a similar argument. Hence the assertion holds. ∎

LEMMA 5.7. *Let $\alpha_t < 1, \beta_t < 1$ for all $t \in V \setminus I$ and $|K| \geq 2$. If there exists $1 \leq f < g \leq n$ such that $\varphi(f) = m+1$ and $\varphi(g) = n$, then either $\varphi(i) = i$ for all $i \in I$ or $\varphi(i) = m + i \pmod{n}$ for all $i \in I$ and $\alpha_I \leq \min\{\beta_J, \beta_K\}$.*

Proof. By Lemmas 5.4 and 5.5, we have $\varphi(i) \leq m$ for all $i < f$ and $i > g$ and $\varphi(i) > m+1$ for $f < i < g$. Hence we need only consider the following two cases.

Case 1. There exists $1 \leq h < f$ such that $\varphi(h) \leq m$. Then by Lemma 5.6, there does not exist $i > f$ such that $\varphi(i) \leq m$. Further, for $1 \leq i < j < f$, $\varphi(i) := p < \varphi(j) := q$. In fact, if $p > q$, then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^{\varphi}[i, j, g]$ is

$$V = \begin{pmatrix} \gamma_p & \gamma_p\delta_L & \gamma_p\gamma_I \\ \gamma_q\gamma_L & \gamma_q & \gamma_q\gamma_L\gamma_I \\ \gamma_n\delta_I & \gamma_n\delta_I & \gamma_n \end{pmatrix},$$

where $p \wedge q = L$. By the definition of $\mathcal{W}$ in quasi-nested block form, it is easy to see that $\gamma_I = 1$ or $\gamma_L = 1$. Hence $\alpha_L = 1$ or $\alpha_I = 1$. Both are contradictions.

Hence $\varphi(i) = i$ for $i = 1, \ldots, m$. Moreover, it is easy to show that $\varphi(i) < \varphi(j)$ for all $m < i < j \le n$. Therefore $\varphi(i) = i$ for $i = 1, \ldots, n$.

Case 2. There exists $h > g$ such that $\varphi(h) \le m$. Then $\varphi(i) \ge m + 1$ for all $i < g$ and $\varphi(i) \le m$ for any $i > g$ by Lemma 5.6. Furthermore, it is easy to show that $\varphi(i) < \varphi(j)$ for all $g < i < j$, and $\varphi(i) < \varphi(j)$ for all $1 \le i < j \le g$. Hence $\varphi(i) = m + i \pmod n$ for all $i \in I$. Moreover, since U^φ is a $\mathcal{W}$ matrix in quasi-nested block form, then $\alpha_I \le \min\{\beta_J, \beta_K\}$, and the proof is completed. ∎

LEMMA 5.8. *Let $\alpha_t < 1$, $\beta_t < 1$ for all $t \in V \setminus I$. If $|K| = 1$, then φ is the identity permutation, or $\varphi(1) = m + 1$ and $\varphi(i) = i - 1$ for all $i = 2, \ldots, m + 1$ with $\alpha_I \le \beta_J$.*

Proof. Since $|K| = 1$, $n = m + 1$. Let $f \in I$ such that $\varphi(f) = m + 1$. We consider the following three cases.

Case 1. $f = 1$. Then for any $1 < i < j$, $\varphi(i) < \varphi(j)$. In fact, if $\varphi(i) := p > \varphi(j) := q$, then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^\varphi[1, i, j]$ is

$$V = \begin{pmatrix} \gamma_{m+1} & \gamma_{m+1}\delta_I & \gamma_{m+1}\delta_I \\ \gamma_p\gamma_I & \gamma_p & \gamma_p\delta_L \\ \gamma_q\gamma_L\gamma_I & \gamma_q\gamma_L & \gamma_q \end{pmatrix},$$

where $i \wedge j = L$. It is easy to see that $\gamma_L = 1$, which yields $\alpha_L = 1$, a contradiction. Hence $\varphi(1) = m + 1$ and $\varphi(i) = i - 1$ for $i = 2, \ldots, m + 1$. Moreover, $\alpha_I \le \beta_J$, since U^φ is a $\mathcal{W}$ matrix in quasi-nested block form.

Case 2. $1 < f < m + 1$. Then there exists $i < f < j$ such that $\varphi(i) := p$, $\varphi(j) := q \le m$. Without loss of generality, we may assume that $p > q$. Then the induced $\mathcal{W}$ matrix of order 3 in quasi-nested block form from $U^\varphi[i, f, j]$ is

$$V = \begin{pmatrix} \gamma_p & \gamma_p\gamma_I & \gamma_p\delta_L \\ \gamma_{m+1}\delta_I & \gamma_{m+1} & \gamma_{m+1}\delta_I \\ \gamma_q\gamma_L & \gamma_q\gamma_L\gamma_I & \gamma_q \end{pmatrix},$$

where $i \wedge j = L$. It is easy to see that $\gamma_I = 1$, which implies that $\alpha_I = 1$, a contradiction.

Case 3. $f = m + 1$. By an argument similar to the proof of Case 1, it is easy to see that φ is the identity permutation. ∎

Now we present the main result of this section.

THEOREM 5.9. *Let U be a $\mathcal{W}$ matrix of order n in quasi-nested block form with support tree T and defined by $\overrightarrow{\alpha}$, $\overrightarrow{\beta}$ on T. The root of the support tree is $I = \{1, 2, \ldots, n\}$, and $I^- = J$, $I^+ = K$. Denote $|J| = m$. If $\alpha_t < 1$, $\beta_t < 1$ for all $t \in V \setminus I$ and φ is a permutation on I, then $U^\varphi := (U_{\varphi(i),\varphi(j)})$ is a $\mathcal{W}$ matrix in quasi-nested block form if and only if φ is the identity permutation on I or $\alpha_I \le \min\{\beta_J, \beta_K\}$ with $\varphi(i) = m + i \pmod n$ for $i = 1, \ldots, n$.*

Proof. If $U^\varphi := (U_{\varphi(i),\varphi(j)})$ is a $\mathcal{W}$ matrix in quasi-nested block form, it follows from Corollary 5.3 and Lemmas 5.7 and 5.8 that the assertion holds. Conversely, it is easy to show that the assertion holds by the definition of a $\mathcal{W}$ matrix in quasi-nested block form. ∎

Remark 5.10. Theorem 5.9 does not hold in general, as we will see in the following example, if we cancel the conditions $\alpha_t < 1$, $\beta_t < 1$.

Example 5.11. Let U be a $\mathcal{W}$ matrix of order 6 as follows:

$$U = \begin{pmatrix} \alpha_1 & \alpha_1\alpha_J & \alpha_1\alpha_J & \alpha_1\alpha_J & \alpha_1\alpha_J & \alpha_1\alpha_J\alpha_I \\ \alpha_2\alpha_M\beta_J & \alpha_2 & \alpha_2\alpha_M & \alpha_2\alpha_M & \alpha_2\alpha_M & \alpha_2\alpha_M\alpha_I \\ \alpha_3\alpha_L\alpha_N\beta_J & \alpha_3\alpha_L\alpha_N\beta_M & \alpha_3 & \alpha_3\alpha_L & \alpha_3\alpha_L\alpha_N & \alpha_3\alpha_L\alpha_N\alpha_I \\ \alpha_4\alpha_N\beta_J & \alpha_4\alpha_N\beta_M & \alpha_4\beta_L & \alpha_4 & \alpha_4\alpha_N & \alpha_4\alpha_L\alpha_I \\ \alpha_5\beta_J & \alpha_5\beta_M & \alpha_5\beta_N & \alpha_5\beta_N & \alpha_5 & \alpha_5\alpha_I \\ \alpha_6\beta_I & \alpha_6\beta_I & \alpha_6\beta_I & \alpha_6\beta_I & \alpha_6\beta_I & \alpha_6 \end{pmatrix}.$$

If $\alpha_I = 1$ and $\beta_J = \beta_M = \beta_N = \beta_L = 1$, then U^φ is a $\mathcal{W}$ matrix in quasi-nested block form for $\varphi(1) = 6$, $\varphi(2) = 2$, $\varphi(3) = 1$, $\varphi(4) = 5$, $\varphi(5) = 3$, $\varphi(6) = 4$.

Acknowledgments. The authors would like to thank the referees for many helpful suggestions and for proposing Questions 4.5 and 4.6, which resulted in an improvement of the revised paper.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.
[2] C. Dellacherie, S. Martínez, and J. San Martín, *Ultrametric matrices and induced Markov chains*, Adv. in Appl. Math., 17 (1996), pp. 169–183.
[3] C. Dellacherie, S. Martínez, and J. San Martín, *Description of the sub-Markov kernel associated to generalized ultrametric matrices: An algorithmic approach*, Linear Algebra Appl., 318 (2000), pp. 1–21.
[4] M. Fiedler, *Some characterizations of symmetric inverse M-matrices*, Linear Algebra Appl., 275/276 (1998), pp. 179–187.
[5] M. Fiedler, *Special ultrametric matrices and graphs*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 106–113.
[6] M. Fiedler, C. R. Johnson, and T. L. Markham, *Notes on inverse M-matrices*, Linear Algebra Appl., 91 (1987), pp. 75–81.
[7] C. R. Johnson, *Inverse M-matrices*, Linear Algebra Appl., 47 (1982), pp. 159–216.
[8] T. L. Markham, *Nonnegative matrices whose inverses are M-matrices*, Proc. Amer. Math. Soc., 36 (1972), pp. 326–330.
[9] S. Martínez, G. Michon, and J. San Martín, *Inverse of strictly ultrametric matrices are of Stieltjes type*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 98–106.
[10] J. J. McDonald, R. Nabben, M. Neumann, H. Schneider, and M. J. Tsatsomeros, *Inverse tridiagonal Z-matrices*, Linear Multilinear Algebra, 45 (1998), pp. 75–97.
[11] J. J. McDonald, M. Neumann, H. Schneider, and M. J. Tsatsomeros, *Inverse M-matrix inequalities and generalized ultrametric matrices*, Linear Algebra Appl., 220 (1995), pp. 321–341.
[12] R. Nabben, *A class of inverse M-matrices*, Electron. J. Linear Algebra, 7 (2000), pp. 53–58.
[13] R. Nabben and R. S. Varga, *A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieltjes matrix*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 107–113.
[14] R. Nabben and R. S. Varga, *Generalized ultrametric matrices—a class of inverse M-matrices*, Linear Algebra Appl., 220 (1995), pp. 365–390.
[15] M. Neumann, *A conjecture concerning the Hadamard product of inverse of M-matrices*, Linear Algebra Appl., 285 (1995), pp. 277–290.
[16] R. A. Willoughby, *The inverse M-matrix problem*, Linear Algebra Appl., 18 (1977), pp. 75–94.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1149–1160

THE EIGENVALUES AND EIGENSPACES OF SOME DISCRETE DIV- AND CURL-RELATED OPERATORS*

ERIC T. CHUNG† AND JUN ZOU‡

Abstract. The eigenvalues and eigenspaces of some discrete div- and curl-related operators are investigated. The discrete operators give some good discrete analogues of the continuous counterparts and play an important role in developing finite volume schemes for solving div-curl equations and electromagnetic systems. Knowledge of the eigenvalues and eigenspaces is very useful in the numerical analysis of finite volume methods for electromagetic systems in *nonhomogeneous* media.

Key words. eigenvalues, eigenspaces, discrete div-operator, discrete curl-operator

AMS subject classifications. 65F15, 78-08, 65N25

PII. S0895479801382483

1. Introduction.

The aim of this paper is to find the explicit formulae for the complete eigenvalues and eigenspaces of some discrete div- and curl-related operators. These operators play a very important role in the finite volume approximation of the div-curl equations [7], [9] as well as of Maxwell's equations [3], [8]. Knowledge of the eigenvalues and eigenspaces is very useful in the numerical analysis of a newly developed finite volume method for electromagetic systems in *nonhomogeneous* media [3], [6].

We will mainly investigate three discrete operators: the discrete divergence, curl-curl, and Laplacian operators. We will see that these three discrete operators satisfy a relation that resembles the continuous counterpart. The main difficulty for the spectral analysis lies in the fact that all three components of a vector-valued function in $\mathbb{R}^3$ contribute to each component of the curl-curl operator, while this is not the case for the discrete Laplacian operator. Hence, the standard treatment for finding the eigenvalues and eigenspaces of a Laplacian operator does not work for the curl-curl operator. We will present a new approach for finding the complete eigenvalues and eigenspaces of the discrete curl-curl operator. As we will see, the spectra of the discrete curl-curl operator and the discrete Laplacian operator are similar, but their eigenspaces are different.

The paper is organized as follows. In section 2, we give the definitions of the discrete curl, divergence, and Laplacian operators. In section 3, we show an interesting relation among the three operators and study the complete eigenvalues and eigenspaces of the discrete operators. In section 4, we present some applications of the discrete operators and their eigenvalues and eigenspaces.

2. Discrete differential operators.

We consider a nonuniform triangulation, called the primal mesh, of the unit cube $\Omega = [0,1]^3$ by a set of small rectangular

*Received by the editors October 16, 2001; accepted for publication (in revised form) by M. Hanke October 14, 2002; published electronically March 13, 2003.
http://www.siam.org/journals/simax/24-4/38248.html
†Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095-1555 (tschung@math.ucla.edu).
‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (zou@math.cuhk.edu.hk). The work of this author was supported by Hong Kong RGC grants CUHK4292/00P and CUHK4048/02P.

subdomains, called primal elements.[1] We denote by N_i the number of primal elements in the ith axis direction ($i = 1, 2, 3$). The faces, edges, and nodes of each primal element are called the primal faces, edges, and nodes, respectively. Then, we construct the dual mesh by connecting all the centers of primal elements; this gives another nonuniform triangulation of the domain Ω. Each rectangular subdomain in the dual mesh is called a dual element. Dual faces, edges, and nodes are named as in the primal mesh. Later on, by an interior primal edge (face) we mean a primal edge (face) not completely lying on the boundary of Ω. Moreover, we denote by σ_i the ith primal edge and by σ_j' the jth dual edge. Here, we always use a primed form of a primal quantity to represent a dual quantity. For example, by κ_i, κ_j', τ_r, and τ_s' we mean the ith primal face, jth dual face, rth primal element, and sth dual element, respectively.

The above primal and dual meshes have an important internal relation: each interior primal face (edge) is perpendicular to and in one-to-one correspondence with a dual edge (face), and each interior primal node (element) is in one-to-one correspondence with a dual element (node). Now we assign each edge (both primal and dual) a direction in the way that each edge points to the positive axis direction and assign each primal (dual) face a direction such that it has the same direction as the corresponding dual (primal) edge.

Let E, F, and T be the numbers of interior primal edges, faces, and nodes, respectively. Then by the aforementioned internal relation, we know E, F, and T are also the numbers of dual faces, edges, and elements, respectively, and

$$E = \sum_{i=1}^{3} N_i(N_{i+1} - 1)(N_{i+2} - 1), \quad F = \sum_{i=1}^{3}(N_i - 1)N_{i+1}N_{i+2},$$
$$T = (N_1 - 1)(N_2 - 1)(N_3 - 1).$$

Here and in the subsequent sections we will use the convention that $N_i = N_{i-3}$ for $i > 3$.

For a primal edge $\sigma_j \in \partial\kappa_i$, we say it is oriented positively along $\partial\kappa_i$ if its direction agrees with the direction of $\partial\kappa_i$ formed by the right-hand rule with the thumb pointing in the direction of κ_i. Otherwise, we say σ_j is oriented negatively along $\partial\kappa_i$. In light of the Stokes theorem,

$$\int_{\kappa_i} (\nabla \times \mathbf{u}) \cdot \mathbf{n}\, d\sigma = \int_{\partial\kappa_i} \mathbf{u} \cdot \mathbf{t}\, dl, \tag{2.1}$$

where $\mathbf{u}$ is a vector-valued function in $\mathbb{R}^3$, we define a discrete curl matrix G by

$$(G)_{ij} := \begin{cases} 1 & \text{if } \sigma_j \text{ is oriented positively along } \partial\kappa_i, \\ -1 & \text{if } \sigma_j \text{ is oriented negatively along } \partial\kappa_i, \\ 0 & \text{if } \sigma_j \text{ does not meet } \partial\kappa_i. \end{cases}$$

Clearly G is an $F \times E$ matrix, and $\operatorname{rank}(G) = E - T$ (cf. [9]). One of the goals of this paper is to find all the eigenvalues and eigenvectors of the $E \times E$ matrix G^TG, which is of rank $E - T$. G^TG is symmetric positive semidefinite, so all its eigenvalues are nonnegative. Since the null space of G^TG has dimension T, zero is an eigenvalue of G^TG with multiplicity T. In other words, we need only to find all the remaining $E - T$ positive eigenvalues of G^TG.

[1] The results and techniques of this paper are directly applicable to treating the more general case, for instance, where the domain Ω is a union of some rectangular domains.

For a dual face $\kappa_j' \in \partial\tau_i'$, we say it is oriented positively along $\partial\tau_i'$ if its direction is pointing toward the outside of τ_i'. Otherwise, we say κ_j' is oriented negatively along $\partial\tau_i'$. Initiated by the divergence theorem,

$$\int_{\tau_i'} \nabla \cdot \mathbf{u}\, dx = \int_{\partial\tau_i'} \mathbf{u} \cdot \mathbf{n}\, d\sigma, \tag{2.2}$$

where $\mathbf{u}$ is a vector-valued function in $\mathbb{R}^3$, we define a discrete divergence matrix B by

$$(B)_{ij} := \begin{cases} 1 & \text{if } \kappa_j' \text{ is oriented positively along } \partial\tau_i', \\ -1 & \text{if } \kappa_j' \text{ is oriented negatively along } \partial\tau_i', \\ 0 & \text{if } \kappa_j' \text{ does not meet } \partial\tau_i'. \end{cases}$$

Then B is a $T \times E$ matrix. It is known [9] that the rank of B is T, and that $BG^T = 0$, which is a discrete analogue of $\nabla \cdot (\nabla \times \mathbf{u}) = 0$.

Consider an interior primal edge σ_i. We say σ_k is adjacent to σ_i if both σ_i and σ_k lie on the same primal face or if their intersection is a single point. Clearly, for any interior primal edge having no intersection with $\partial\Omega$, it must have 6 adjacent primal edges. But for any interior primal edge having a nonempty intersection with $\partial\Omega$, the edge has only 5 adjacent primal edges. With these definitions in mind, we define an $E \times E$ discrete Laplacian matrix A in the following way:

1. If σ_i is an interior primal edge having no intersection with $\partial\Omega$, then

$$(A)_{ij} := \begin{cases} 6 & \text{if } j = i, \\ -1 & \text{if } \sigma_j \text{ is adjacent to } \sigma_i, \\ 0 & \text{otherwise.} \end{cases}$$

2. If σ_i is an interior primal edge having a nonempty intersection with $\partial\Omega$, then

$$(A)_{ij} := \begin{cases} 5 & \text{if } j = i, \\ -1 & \text{if } \sigma_j \text{ is adjacent to } \sigma_i, \\ 0 & \text{otherwise.} \end{cases}$$

Note that this discrete Laplacian is different from the standard discrete Laplacian resulting from the discretization of the Laplacian operator by the second order central difference scheme. Instead, BB^T is closer to the standard discrete Laplacian; see Theorem 3.4.

3. Eigenvalues and eigenspaces. This section will be devoted to our main results. For any $\mathbf{f} \in \mathbb{R}^E$, we will interpret its ith component f_i as its value on the ith interior primal edge σ_i, as well as its value on the ith dual face κ_i'. We will often write $\mathbf{f} = (\mathbf{u}^T, \mathbf{v}^T, \mathbf{w}^T)^T$, where $\mathbf{u}$ ($\mathbf{v}$ and $\mathbf{w}$, respectively) is a vector in $\mathbb{R}^{\frac{E}{3}}$ and each component of $\mathbf{u}$ corresponds to an interior primal edge parallel to the x-axis (y-axis and z-axis, respectively).

Now, we are ready to present our first result, which is a discrete version of the well-known relation

$$\nabla \times \nabla \times \mathbf{u} = \nabla(\nabla \cdot \mathbf{u}) - \nabla^2 \mathbf{u}. \tag{3.1}$$

THEOREM 3.1. *For the discrete curl, divergence, and Laplacian operators G, B, and A,*

$$G^T G = -B^T B + A. \tag{3.2}$$

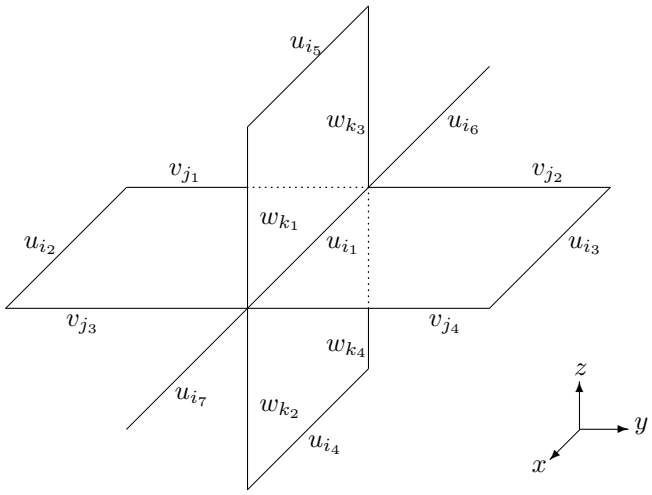


FIG. 3.1. *An interior primal edge having no intersection with $\partial\Omega$ and its adjacent edges.*

Proof. It suffices to show that for any vector $\mathbf{f} = (\mathbf{u}^T, \mathbf{v}^T, \mathbf{w}^T)^T$, we have

$$G^T G\mathbf{f} = -B^T B\mathbf{f} + A\mathbf{f}.$$

First we consider an interior primal edge σ_i having no intersection with $\partial\Omega$; see Figure 3.1.

Here, u_{i_1} denotes a component of $\mathbf{u}$ corresponding to σ_i, which, without loss of generality, is assumed to be parallel to the x-axis. u_{i_s}, $s = 2, 3, \ldots, 7$, denote components of $\mathbf{u}$ corresponding to all adjacent edges of σ_i. Similarly, v_{j_r} and w_{k_r}, $r = 1, 2, 3, 4$, are components of $\mathbf{v}$ and $\mathbf{w}$, respectively, corresponding to the primal edges parallel to the y-axis and z-axis. By the definitions of G, B, and A and direct computations, we know the ith components of $A\mathbf{f}$, $B^T B\mathbf{f}$, and $G^T G\mathbf{f}$ corresponding to σ_i are, respectively, given by

$$\begin{aligned}
(A\mathbf{f})_i &= 6u_{i_1} - u_{i_2} - u_{i_3} - u_{i_4} - u_{i_5} - u_{i_6} - u_{i_7},\\
(B^T B\mathbf{f})_i &= (u_{i_1} - u_{i_6} + v_{j_2} - v_{j_1} + w_{k_3} - w_{k_4}) - (u_{i_7} - u_{i_1} + v_{j_4} - v_{j_3} + w_{k_1} - w_{k_2}),\\
(G^T G\mathbf{f})_i &= (4u_{i_1} - u_{i_2} - u_{i_3} - u_{i_4} - u_{i_5}) + (v_{j_1} - v_{j_2} - v_{j_3} + v_{j_4})\\
&\qquad + (w_{k_1} - w_{k_2} - w_{k_3} + w_{k_4}).
\end{aligned}$$

This implies

$$(G^T G\mathbf{f})_i = -(B^T B\mathbf{f})_i + (A\mathbf{f})_i.$$

Now we consider an interior primal edge σ_i having a single-point intersection with $\partial\Omega$. See Figure 3.2 below, where P is the single-point intersection of σ_i with $\partial\Omega$.

If one of the primal edges corresponding to the component u_{i_s}, $s = 2, 3, 4, 5$, lies on $\partial\Omega$, then we take u_{i_s} to be zero since $\mathbf{f}$ does not contain any boundary component by definition. Then, the ith components of $A\mathbf{f}$, $B^T B\mathbf{f}$, and $G^T G\mathbf{f}$ are, respectively, given by

$$\begin{aligned}
(A\mathbf{f})_i &= 5u_{i_1} - u_{i_2} - u_{i_3} - u_{i_4} - u_{i_5} - u_{i_6},\\
(B^T B\mathbf{f})_i &= u_{i_1} - u_{i_6} + v_{j_2} - v_{j_1} + w_{k_1} - w_{k_2},\\
(G^T G\mathbf{f})_i &= (4u_{i_1} - u_{i_2} - u_{i_3} - u_{i_4} - u_{i_5}) + (v_{j_1} - v_{j_2}) + (w_{k_2} - w_{k_1}).
\end{aligned}$$

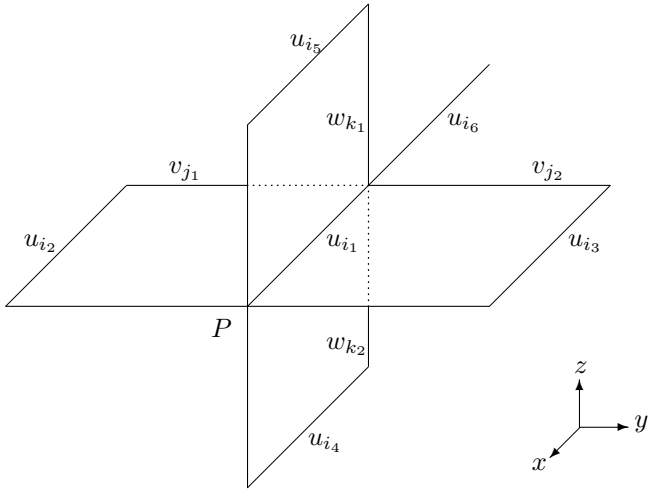


FIG. 3.2. *An interior primal edge having a single-point intersection with $\partial\Omega$ and its adjacent edges.*

It is easy to check that

$$(G^TG\mathbf{f})_i = -(B^TB\mathbf{f})_i + (A\mathbf{f})_i.$$

We complete the proof of Theorem 3.1 by noting that any interior primal edge has either an empty intersection or a single-point intersection with $\partial\Omega$. ∎

Before studying the eigenvalues of G^TG, we will first work on the eigenvalues of A. For convenience, we will let $h_1 = 1/N_1$, $h_2 = 1/N_2$, and $h_3 = 1/N_3$. Note that the original triangulation is nonuniform, so h_1, h_2, and h_3 are not the actual nonequidistant mesh sizes along the x-, y-, and z-axis. However, the definitions of the matrices G, B, and A are independent of the mesh sizes, so we can always assume that the meshes are uniform along each axis and h_1, h_2, and h_3 are the mesh sizes along the x-, y-, and z-axis, respectively. Let k, m, and l be three integers such that $1 \le k \le N_1 - 1$, $1 \le m \le N_2 - 1$, and $1 \le l \le N_3 - 1$. Then for any fixed k, m, and l, we define

$$\begin{aligned}
\lambda^1_{ml} &= 4\sin^2\left(\frac{m\pi h_2}{2}\right) + 4\sin^2\left(\frac{l\pi h_3}{2}\right),\\
\lambda^2_{kl} &= 4\sin^2\left(\frac{k\pi h_1}{2}\right) + 4\sin^2\left(\frac{l\pi h_3}{2}\right),\\
\lambda^3_{km} &= 4\sin^2\left(\frac{k\pi h_1}{2}\right) + 4\sin^2\left(\frac{m\pi h_2}{2}\right),\\
\beta_{kml} &= 4\sin^2\left(\frac{k\pi h_1}{2}\right) + 4\sin^2\left(\frac{m\pi h_2}{2}\right) + 4\sin^2\left(\frac{l\pi h_3}{2}\right).
\end{aligned}$$

For any fixed k, m, and l, we define $\mathbf{f}^1_{ml} = (\mathbf{u}_1^T, \mathbf{v}_1^T, \mathbf{w}_1^T)^T \in \mathbb{R}^E$ to be a vector with only components corresponding to the interior primal edges parallel to the x-axis, i.e., $\mathbf{v}_1 = \mathbf{w}_1 = \mathbf{0}$, and the components of $\mathbf{u}_1$ are given by

$$(\mathbf{u}_1)_j = \sin(ym\pi h_2)\sin(zl\pi h_3), \tag{3.3}$$

where $(\mathbf{u}_1)_j$ is the component of $\mathbf{u}_1$ corresponding to the primal edge σ_j which is parallel to the x-axis, and yh_2 and zh_3 are the y-coordinate and z-coordinate of the

primal edge σ_j, respectively (with y and z being some positive integers). Similarly, we define $\mathbf{f}^2_{kl} = (\mathbf{u}_2^T, \mathbf{v}_2^T, \mathbf{w}_2^T)^T \in \mathbb{R}^E$ and $\mathbf{f}^3_{km} = (\mathbf{u}_3^T, \mathbf{v}_3^T, \mathbf{w}_3^T)^T \in \mathbb{R}^E$ to be two vectors with only the components corresponding to the interior primal edges parallel to the y-axis and z-axis, respectively. Clearly, $\mathbf{f}^1_{ml}$, $\mathbf{f}^2_{kl}$, and $\mathbf{f}^3_{km}$ are linearly independent for any fixed k, m, and l.

Furthermore, for fixed k, m, and l, we define the vector $\mathbf{g}^1_{kml} = (\tilde{\mathbf{u}}_1^T, \tilde{\mathbf{v}}_1^T, \tilde{\mathbf{w}}_1^T)^T \in \mathbb{R}^E$ $(i = 1, 2, 3)$ to be the same as $\mathbf{f}^1_{ml}$, but replace $(\mathbf{u}_1)_j$ in (3.3) by

$$(\tilde{\mathbf{u}}_1)_j = \cos\left(\left(x + \frac{1}{2}\right) k\pi h_1\right) \sin(ym\pi h_2) \sin(zl\pi h_3), \tag{3.4}$$

where $(x + \frac{1}{2})h_1$ is the x-coordinate of the midpoint of the edge σ_j. $\mathbf{g}^2_{kml}$ and $\mathbf{g}^3_{kml}$ are defined similarly. Clearly, $\mathbf{g}^i_{kml}$, $i = 1, 2, 3$, are linearly independent for any fixed k, m, and l.

The following theorem gives the complete spectrum and eigenvectors of A.

THEOREM 3.2. *For k, m, and l satisfying $1 \le k \le N_1 - 1$, $1 \le m \le N_2 - 1$, $1 \le l \le N_3 - 1$, we have*

$$A\mathbf{f}^1_{ml} = \lambda^1_{ml}\mathbf{f}^1_{ml}, \quad A\mathbf{f}^2_{kl} = \lambda^2_{kl}\mathbf{f}^2_{kl}, \quad A\mathbf{f}^3_{km} = \lambda^3_{km}\mathbf{f}^3_{km}; \quad A\mathbf{g}^i_{kml} = \beta_{kml}\mathbf{g}^i_{kml},\ i = 1, 2, 3. \tag{3.5}$$

Proof. We start with the proof of the first relation in (3.5). We first consider an interior primal edge σ_j having no intersection with $\partial\Omega$. If σ_j is parallel to the x-axis and has y-coordinate yh_2 and z-coordinate zh_3, then by the definition of A, we have

$$\begin{aligned}(A\mathbf{f}^1_{ml})_j = {} & \Big\{6\sin(ym\pi h_2) - \sin((y-1)m\pi h_2) - \sin((y+1)m\pi h_2)\Big\}\sin(zl\pi h_3)\\ & - \sin(ym\pi h_2)\Big\{\sin((z-1)l\pi h_3) - \sin((z+1)l\pi h_3)\Big\} - 2\sin(ym\pi h_2)\sin(zl\pi h_3).\end{aligned}$$

A direct computation yields

$$(A\mathbf{f}^1_{ml})_j = 4\left\{\sin^2\left(\frac{m\pi h_2}{2}\right) + \sin^2\left(\frac{l\pi h_3}{2}\right)\right\}\sin(ym\pi h_2)\sin(zl\pi h_3).$$

This shows $(A\mathbf{f}^1_{ml})_j = \lambda^1_{ml}(\mathbf{f}^1_{ml})_j$. Now, if σ_j is an interior primal edge having empty intersection with $\partial\Omega$ and is parallel to the y- or z-axis, then $(\mathbf{f}^1_{ml})_j = (\mathbf{v}_1)_j = 0$ or $(\mathbf{f}^1_{ml})_j = (\mathbf{w}_1)_j = 0$ by definition. This implies $(A\mathbf{f}^1_{ml})_j = 0 = \lambda^1_{ml}(\mathbf{f}^1_{ml})_j$.

Next we consider an interior primal edge σ_j having a single-point intersection with $\partial\Omega$. If σ_j is parallel to the x-axis and has y-coordinate yh_2 and z-coordinate zh_3, then by the definition of A, we have

$$\begin{aligned}(A\mathbf{f}^1_{ml})_j = {} & \Big\{5\sin(ym\pi h_2) - \sin((y-1)m\pi h_2) - \sin((y+1)m\pi h_2)\Big\}\sin(zl\pi h_3)\\ & - \sin(ym\pi h_2)\Big\{\sin((z-1)l\pi h_3) - \sin((z+1)l\pi h_3)\Big\} - \sin(ym\pi h_2)\sin(zl\pi h_3).\end{aligned}$$

A direct computation yields

$$(A\mathbf{f}^1_{ml})_j = 4\left\{\sin^2\left(\frac{m\pi h_2}{2}\right) + \sin^2\left(\frac{l\pi h_3}{2}\right)\right\}\sin(ym\pi h_2)\sin(zl\pi h_3).$$

Therefore we have $(A\mathbf{f}^1_{ml})_j = \lambda^1_{ml}(\mathbf{f}^1_{ml})_j$. The same argument can be applied to prove the second and third relations in (3.5) for the case that σ_j is an interior primal edge

having an empty or a nonempty intersection with $\partial\Omega$ and is parallel to either the y- or z-axis.

We now prove the fourth relation in (3.5). First, consider an interior primal edge σ_j having empty intersection with $\partial\Omega$. If σ_j is parallel to the x-axis and has y-coordinate yh_2 and z-coordinate zh_3, with the x-coordinate of the midpoint of σ_j being $(x+\frac{1}{2})h_1$ for some integer x, then by the definition of A, we have

$$\begin{aligned}(A\mathbf{g}^1_{kml})_j = \cos\left(\left(x+\frac{1}{2}\right)k\pi h_1\right)\Big\{&6\sin(ym\pi h_2)\sin(zl\pi h_3)\\ &-\sin((y-1)m\pi h_2)\sin(zl\pi h_3)-\sin((y+1)m\pi h_2)\sin(zl\pi h_3)\\ &-\sin(ym\pi h_2)\sin((z-1)l\pi h_3)-\sin(ym\pi h_2)\sin((z+1)l\pi h_3)\Big\}\\ &-\left\{\cos\left(\left(x-\frac{1}{2}\right)k\pi h_1\right)+\cos\left(\left(x+\frac{3}{2}\right)k\pi h_1\right)\right\}\sin(ym\pi h_2)\sin(zl\pi h_3),\end{aligned}$$

which, by a direct computation, can be written as

$$(A\mathbf{g}^1_{kml})_j=\beta_{kml}\cos\left(\left(x+\frac{1}{2}\right)k\pi h_1\right)\sin(ym\pi h_2)\sin(zl\pi h_3)=\beta_{kml}(\mathbf{g}^1_{kml})_j.$$

For an interior primal edge σ_j having empty intersection with $\partial\Omega$ and being parallel to the y- or z-axis, we know $(\mathbf{g}^1_{kml})_j=(\tilde{\mathbf{v}}_1)_j=0$ or $(\mathbf{g}^1_{kml})_j=(\tilde{\mathbf{w}}_1)_j=0$ by definition. Therefore

$$(A\mathbf{g}^1_{kml})_j=0=\beta_{kml}(\mathbf{g}^1_{kml})_j.$$

Now, for an interior primal edge σ_j having a single-point intersection with $\partial\Omega$, assume σ_j is parallel to the x-axis and has y-coordinate yh_2 and z-coordinate zh_3, and the x-coordinate of the midpoint of σ_j is $(x+\frac{1}{2})h_1$ for $x=0$ or $x=N_1-1$. Then, by the definition of A, we have

$$\begin{aligned}(A\mathbf{g}^1_{kml})_j = \cos\left(\left(x+\frac{1}{2}\right)k\pi h_1\right)\Big\{&5\sin(ym\pi h_2)\sin(zl\pi h_3)\\ &-\sin((y-1)m\pi h_2)\sin(zl\pi h_3)-\sin((y+1)m\pi h_2)\sin(zl\pi h_3)\\ &-\sin(ym\pi h_2)\sin((z-1)l\pi h_3)-\sin(ym\pi h_2)\sin((z+1)l\pi h_3)\Big\}\\ &-\cos\left(\left(x+\frac{1}{2}\pm 1\right)k\pi h_1\right)\sin(ym\pi h_2)\sin(zl\pi h_3),\end{aligned}$$

where ±1 is taken for $x=0$ and $x=N_1-1$, respectively. Using the fact that $\cos((x+\frac{1}{2})k\pi h_1)=\cos((x-\frac{1}{2})k\pi h_1)$ for $x=0$ and $\cos((x+\frac{1}{2})k\pi h_1)=\cos((x+\frac{3}{2})k\pi h_1)$ for $x=N_1-1$, the above relation can be written as

$$\begin{aligned}(A\mathbf{g}^1_{kml})_j = \cos\left(\left(x+\frac{1}{2}\right)k\pi h_1\right)\Big\{&6\sin(ym\pi h_2)\sin(zl\pi h_3)\\ &-\sin((y-1)m\pi h_2)\sin(zl\pi h_3)-\sin((y+1)m\pi h_2)\sin(zl\pi h_3)\\ &-\sin(ym\pi h_2)\sin((z-1)l\pi h_3)-\sin(ym\pi h_2)\sin((z+1)l\pi h_3)\Big\}\\ &-\cos\left(\left(x-\frac{1}{2}\right)k\pi h_1\right)\sin(ym\pi h_2)\sin(zl\pi h_3)\\ &-\cos\left(\left(x+\frac{3}{2}\right)k\pi h_1\right)\sin(ym\pi h_2)\sin(zl\pi h_3).\end{aligned}$$

This immediately leads to

$$(3.6)\quad (A\mathbf{g}^1_{kml})_j = \beta_{kml}\cos\left(\left(x+\frac{1}{2}\right)k\pi h_1\right)\sin(ym\pi h_2)\sin(zl\pi h_3) = \beta_{kml}(\mathbf{g}^1_{kml})_j.$$

The same argument can be applied to prove (3.6) for the components of $\mathbf{g}^1_{kml}$ corresponding to the primal edges parallel to the y- or z-axis and to prove the last relation in (3.5) with $i = 2, 3$. ∎

The following theorem gives the complete spectrum and eigenvectors of the discrete curl-curl operator G^TG.

THEOREM 3.3. *For each triplet of integers $\{k, m, l\}$ satisfying $1 \le k \le N_1 - 1$, $1 \le m \le N_2 - 1$, $1 \le l \le N_3 - 1$, we have*

$$(3.7)\qquad G^TG\mathbf{f}^1_{ml} = \lambda^1_{ml}\mathbf{f}^1_{ml},\quad G^TG\mathbf{f}^2_{kl} = \lambda^2_{kl}\mathbf{f}^2_{kl},\quad G^TG\mathbf{f}^3_{km} = \lambda^3_{km}\mathbf{f}^3_{km}.$$

Moreover, there exist two linearly independent vectors $\mathbf{p}^1_{kml}$ and $\mathbf{p}^2_{kml}$ in $\mathbb{R}^E$ such that

$$(3.8)\qquad G^TG\mathbf{p}^i_{kml} = \beta_{kml}\mathbf{p}^i_{kml},\quad i = 1, 2.$$

Proof. It is important to notice by the definitions of the matrix B and the vector $\mathbf{f}^1_{ml}$ that, for each dual element τ'_j, $(B\mathbf{f}^1_{ml})_j = 0$. This with (3.2) and (3.5) implies

$$G^TG\mathbf{f}^1_{ml} = -B^TB\mathbf{f}^1_{ml} + A\mathbf{f}^1_{ml} = \lambda^1_{ml}\mathbf{f}^1_{ml}.$$

A similar argument can be applied to show the last two relations in (3.7).

We now prove (3.8). For any fixed integers k, m, and l, we define

$$V_{kml} := \operatorname{span}\{\mathbf{g}^1_{kml}, \mathbf{g}^2_{kml}, \mathbf{g}^3_{kml}\}.$$

Consider any $\mathbf{g} = \alpha_1\mathbf{g}^1_{kml} + \alpha_2\mathbf{g}^2_{kml} + \alpha_3\mathbf{g}^3_{kml} \in V_{kml}$, $\alpha_i \in \mathbb{R}$, $i = 1, 2, 3$. We are going to find all α_i $(i = 1, 2, 3)$ such that $B\mathbf{g} = \mathbf{0}$. For any dual element τ'_j, assume its two primal edges parallel to the x-axis and having nonempty intersection with $\partial\tau'_j$ have x-coordinate $(x - \frac{1}{2})h_1$ and $(x + \frac{1}{2})h_1$, respectively. Clearly, they have the same y- and z-coordinates, namely, yh_2 and zh_3, respectively, for some suitable integers x, y, and z. Then, by a direct computation,

$$\begin{aligned}(B\mathbf{g}^1_{kml})_j &= \left\{\cos\left(\left(x+\frac{1}{2}\right)k\pi h_1\right) - \cos\left(\left(x-\frac{1}{2}\right)k\pi h_1\right)\right\}\sin(ym\pi h_2)\sin(zl\pi h_3)\\ &= -2\sin\left(\frac{k\pi h_1}{2}\right)\sin(xk\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3).\end{aligned}$$

Applying the same argument, we have

$$\begin{aligned}(B\mathbf{g}^2_{kml})_j &= \sin(xk\pi h_1)\left\{\cos\left(\left(y+\frac{1}{2}\right)m\pi h_2\right) - \cos\left(\left(y-\frac{1}{2}\right)m\pi h_2\right)\right\}\sin(zl\pi h_3)\\ &= -2\sin\left(\frac{m\pi h_2}{2}\right)\sin(xk\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3),\\ (B\mathbf{g}^3_{kml})_j &= \sin(xk\pi h_1)\sin(ym\pi h_2)\left\{\cos\left(\left(z+\frac{1}{2}\right)l\pi h_3\right) - \cos\left(\left(z-\frac{1}{2}\right)l\pi h_3\right)\right\}\\ &= -2\sin\left(\frac{l\pi h_3}{2}\right)\sin(xk\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3).\end{aligned}$$

Hence, $(B\mathbf{g})_j = 0$ if and only if

$$\alpha_1 \sin\left(\frac{k\pi h_1}{2}\right) + \alpha_2 \sin\left(\frac{m\pi h_2}{2}\right) + \alpha_3 \sin\left(\frac{l\pi h_3}{2}\right) = 0. \tag{3.9}$$

Notice that (3.9) is a condition that is independent of the choice of the dual element τ_j'. Hence, for any $\alpha_i \in \mathbb{R}$ $(i = 1, 2, 3)$ satisfying (3.9), we have $B\mathbf{g} = \mathbf{0}$. Let $\alpha_1 = s$ and $\alpha_2 = t$ for $s, t \in \mathbb{R}$. We then obtain from (3.9) that

$$\alpha_3 = -\frac{s\sin(\frac{k\pi}{2N_1}) + t\sin(\frac{m\pi}{2N_2})}{\sin(\frac{l\pi}{2N_3})}.$$

Then, we can express $\mathbf{g}$ as

$$\mathbf{g} = s\left(\mathbf{g}^1_{kml} - \frac{\sin(\frac{k\pi}{2N_1})}{\sin(\frac{l\pi}{2N_3})}\mathbf{g}^3_{kml}\right) + t\left(\mathbf{g}^2_{kml} - \frac{\sin(\frac{m\pi}{2N_2})}{\sin(\frac{l\pi}{2N_3})}\mathbf{g}^3_{kml}\right).$$

Define

$$\mathbf{p}^1_{kml} := \mathbf{g}^1_{kml} - \frac{\sin(\frac{k\pi}{2N_1})}{\sin(\frac{l\pi}{2N_3})}\mathbf{g}^3_{kml} \quad \text{and} \quad \mathbf{p}^2_{kml} := \mathbf{g}^2_{kml} - \frac{\sin(\frac{m\pi}{2N_2})}{\sin(\frac{l\pi}{2N_3})}\mathbf{g}^3_{kml}.$$

Clearly, we have $B\mathbf{p}^1_{kml} = B\mathbf{p}^2_{kml} = \mathbf{0}$. Thus by (3.2) and (3.5), we have

$$G^TG\mathbf{p}^i_{kml} = -B^TB\mathbf{p}^i_{kml} + A\mathbf{p}^i_{kml} = \beta_{kml}\mathbf{p}^i_{kml}, \quad i = 1, 2. \qquad \square$$

We remark that Theorem 3.3 gives all the positive eigenvalues of G^TG since the vectors $\mathbf{f}^1_{ml}$, $\mathbf{f}^2_{kl}$, $\mathbf{f}^3_{km}$, and $\mathbf{p}^j_{kml}$ form a complete basis for $\mathbb{R}^{E-T}$. Notice that the smallest positive eigenvalue of G^TG is $8\sin^2(\frac{\pi h}{2})$ which varies as $O(h^2)$ for sufficiently large N. This conclusion is important in the convergence analysis of the finite volume method proposed in [3] for Maxwell's equations with discontinuous physical coefficients.

Recall that B is a discrete divergence matrix, and so B^T represents a discrete gradient matrix. Hence, the matrix BB^T is some sort of scalar discrete Laplacian matrix by the fact that $\nabla \cdot \nabla v = \Delta v$ for any real-valued function v. We have the following.

THEOREM 3.4. *For any fixed integers k, m, and l satisfying $1 \le k \le N_1 - 1$, $1 \le m \le N_2 - 1$, $1 \le l \le N_3 - 1$, there exists a vector $\mathbf{q}_{kml} \in \mathbb{R}^T$ such that*

$$BB^T\mathbf{q}_{kml} = \beta_{kml}\mathbf{q}_{kml}. \tag{3.10}$$

Proof. For any dual element τ_i', we define

$$(\mathbf{q}_{kml})_i := \sin(xk\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3),$$

where xh_1, yh_2, and zh_3 are the x-, y-, and z-coordinates of the corresponding interior primal node. Now, by a direct computation, we have

$$\begin{aligned}
&(BB^T\mathbf{q}_{kml})_i \\
&= 6\sin(xk\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3) \\
&- \sin((x-1)k\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3) - \sin((x+1)k\pi h_1)\sin(ym\pi h_2)\sin(zl\pi h_3) \\
&- \sin(xk\pi h_1)\sin((y-1)m\pi h_2)\sin(zl\pi h_3) - \sin(xk\pi h_1)\sin((y+1)m\pi h_2)\sin(zl\pi h_3) \\
&- \sin(xk\pi h_1)\sin(ym\pi h_2)\sin((z-1)l\pi h_3) - \sin(xk\pi h_1)\sin(ym\pi h_2)\sin((z+1)l\pi h_3) \\
&= \beta_{kml}(\mathbf{q}_{kml})_i. \qquad \square
\end{aligned}$$

We remark here that the vectors $\{\mathbf{q}_{kml}\}$ in Theorem 3.4 are linearly independent, so they form the complete eigensystem of the matrix BB^T.

4. Some applications. In this section, we describe some roles of the discrete div and curl operators and some important results directly derived using the results on eigenvalues and eigenvectors of section 3. Detailed proofs of the results below are given in [4]. We remark that the constant K, or K with subscripts, below is the generic constant independent of mesh sizes, etc.

Let $(\cdot,\cdot)$ be the standard Euclidean inner product with norm $\|\cdot\|_2$. We first recall two mesh and physical parameter dependent inner products introduced in [3], [6],

(4.1)
$$(\mathbf{u},\mathbf{v})_W := (S\mathbf{u}, D'\mathbf{v}) \quad \forall\, \mathbf{u},\mathbf{v} \in \mathbb{R}^F; \qquad (\mathbf{u},\mathbf{v})_{W'} := (S'\mathbf{u}, D\mathbf{v}) \quad \forall\, \mathbf{u},\mathbf{v} \in \mathbb{R}^E,$$

where $S = \mathrm{diag}(s_i)$, $D' = \mathrm{diag}(h'_j)$, $S' = \mathrm{diag}(s'_i)$, and $D = \mathrm{diag}(h_j)$ are all diagonal matrices. s_i and h_j are, respectively, the area of the face κ_i and the length of the edge σ_j, and similar definitions hold for s'_i and h'_j. Then we introduce two discrete circulation matrices $\mathcal{C}$ and $\mathcal{C}'$. Following formula (2.1), we define for each interior primal and dual face κ_i and κ'_i,

$$(\mathcal{C}\mathbf{u})_{\kappa_i} := \sum_{\sigma_j \in \partial\kappa_i} u_j \tilde{h}_j\,, \qquad (\mathcal{C}'\mathbf{u})_{\kappa'_i} := \sum_{\sigma'_j \in \partial\kappa'_i} u_j \tilde{h}'_j\,, \tag{4.2}$$

where $\tilde{h}_j$ is the signed length of h_j [3], [6]; similar meanings hold for $\tilde{h}'_j$ and for $\tilde{s}_j$ and $\tilde{s}'_j$ below.

Further, we introduce two discrete flux matrices $\mathcal{D}$ and $\mathcal{D}'$. Following the divergence theorem (2.2), we define, for each primal and dual element τ_i and τ'_i,

$$(\mathcal{D}\mathbf{u})_i := \sum_{\kappa_j \in \partial\tau_i} u_j \tilde{s}_j\,, \qquad (\mathcal{D}'\mathbf{u})_i := \sum_{\kappa'_j \in \partial\tau'_i} u_j \tilde{s}'_j\,. \tag{4.3}$$

These discrete matrices have the useful relations (cf. [3], [6], [9])

$$\mathcal{C} = GD\,, \qquad \mathcal{C}' = G^T D'\,, \qquad \mathcal{D}' = BS'\,. \tag{4.4}$$

The relations indicate that it is the matrices $\mathcal{D}'$ and $\mathcal{C}'$, not the matrices B and G^T, that directly simulate the divergence and curl operators in the general nonuniform grids.

Discrete Sobolev inequalities. Consider two Sobolev spaces

$$H_0(\mathbf{curl}, \mathrm{div}0; \Omega) = \Big\{ \mathbf{u} \in H(\mathbf{curl}; \Omega);\ \ \nabla\cdot\mathbf{u} = 0 \ \text{ in } \ \Omega\,,\ \mathbf{u}\times\mathbf{n} = 0 \text{ on } \partial\Omega \Big\}\,,$$
$$H_0(\mathbf{curl}0, \mathrm{div}; \Omega) = \Big\{ \mathbf{u} \in H(\mathrm{div}; \Omega);\ \ \nabla\times\mathbf{u} = 0 \ \text{ in } \ \Omega\,,\ \mathbf{u}\times\mathbf{n} = 0 \text{ on } \partial\Omega \Big\}\,.$$

The Sobolev inequalities

$$\|\mathbf{u}\|_{L^2(\Omega)} \le K\,\|\nabla\times\mathbf{u}\|_{L^2(\Omega)} \quad \forall\, \mathbf{u} \in H_0(\mathbf{curl}; \mathrm{div}0; \Omega)\,, \tag{4.5}$$
$$\|\mathbf{u}\|_{L^2(\Omega)} \le K\,\|\nabla\cdot\mathbf{u}\|_{L^2(\Omega)} \quad \forall\, \mathbf{u} \in H_0(\mathrm{div}; \mathbf{curl}0; \Omega) \tag{4.6}$$

are essential to the mathematical analysis of Maxwell's equations [5], [6]. Accordingly, the discrete versions of these two inequalities are important in the convergence analysis

of the numerical methods for Maxwell's equations. Corresponding to (4.5), we have [4]

$$\|\mathbf{u}\|_W \le K\,\|\mathbf{u}\|_{\mathcal{C}'} \quad \forall\,\mathbf{u}\in\{\mathbf{v}\in\mathbb{R}^F\,;\ \mathcal{D}\mathbf{v}=0\}\,, \tag{4.7}$$

$$\|\mathbf{u}\|_{W'} \le K\,\|\mathbf{u}\|_{\mathcal{C}} \quad \forall\,\mathbf{u}\in\{\mathbf{v}\in\mathbb{R}^E\,;\ \mathcal{D}'\mathbf{v}=0\}, \tag{4.8}$$

where $\|\cdot\|_W$ and $\|\cdot\|_{W'}$ are the discrete L^2-norms induced from two inner products in (4.2), while $\|\cdot\|_{\mathcal{C}'}$ and $\|\cdot\|_{\mathcal{C}}$ are two different discrete $H(\mathbf{curl};\Omega)$-norms, one based on the dual circulation matrix $\mathcal{C}'$ and the other based on the primal circulation matrix $\mathcal{C}$,

$$\|\mathbf{u}\|^2_{\mathcal{C}'} = (S'^{-1}\mathcal{C}'\mathbf{u}, D\mathcal{C}'\mathbf{u})\,, \qquad \|\mathbf{u}\|^2_{\mathcal{C}} = (S^{-1}\mathcal{C}\mathbf{u}, D'\mathcal{C}\mathbf{u})\,.$$

Similarly, we can establish the discrete versions of (4.6),

$$\|\mathbf{u}\|_{W'} \le K\,\|\mathbf{u}\|_{\mathcal{D}'} \quad \forall\,\mathbf{u}\in\{\mathbf{v}\in\mathbb{R}^E\,;\ \mathcal{C}\,\mathbf{v}=0\}, \tag{4.9}$$

$$\|\mathbf{u}\|_{W} \le K\,\|\mathbf{u}\|_{\mathcal{D}} \quad \forall\,\mathbf{u}\in\{\mathbf{v}\in\mathbb{R}^F\,;\ \mathcal{C}'\,\mathbf{v}=0\}, \tag{4.10}$$

where $\|\cdot\|_{\mathcal{D}'}$ and $\|\cdot\|_{\mathcal{D}}$ are two different discrete $H(\mathrm{div};\Omega)$-norms, one based on the dual flux matrix $\mathcal{D}'$ and the other based on the primal flux matrix $\mathcal{D}$,

$$\|u\|^2_{\mathcal{D}'} = (V'^{-1}\mathcal{D}'u, \mathcal{D}'u)\,, \quad \|u\|^2_{\mathcal{D}} = (V^{-1}\mathcal{D}u, \mathcal{D}u),$$

where $V' = \mathrm{diag}\,(A'_i)$ and $V = \mathrm{diag}\,(A_i)$, with A'_i and A_i being the volume of the dual element τ'_i and the primal element τ_i, respectively.

Solution of the div-curl equations. Following the discussion in [9], the finite volume discretization of the div-curl equations

$$\mathrm{div}\,\mathbf{u} = f\,, \qquad \mathbf{curl}\,\mathbf{u} = \mathbf{g}\,, \qquad \mathbf{u}\times\mathbf{n}|_\Gamma = 0$$

results in the system of linear algebraic equations of the form

$$V'^{-1}\mathcal{D}'\mathbf{u} = \bar{f}\,, \qquad S^{-1}\mathcal{C}\,\mathbf{u} = \bar{\mathbf{g}}\,. \tag{4.11}$$

System (4.11) is a nonsymmetric and indefinite rectangular system. One way to solve this equation is to solve its least-squares system

$$\Big(\mathcal{D}'^T V'^{-2}\mathcal{D}' + \mathcal{C}^T S^{-2}\mathcal{C}\Big)\mathbf{u} = \mathcal{D}'^T V'^{-1}\bar{f} + \mathcal{C}^T S^{-1}\bar{\mathbf{g}}\,. \tag{4.12}$$

Let $\mathcal{A}$ be the coefficient matrix in (4.12). Then we can derive the following estimate for any $\mathbf{v}\in\mathbb{R}^E$ by using the results of section 3 (see [4] for details):

$$K_0(\mathbf{v},\mathbf{v}) \le (\mathcal{A}\mathbf{v},\mathbf{v}) \le K_1\,h^{-2}\,(\mathbf{v},\mathbf{v})\,. \tag{4.13}$$

By conducting more careful analyses in the derivation, one may derive more explicit bounds of K_0 and K_1 in terms of the physical coefficients, etc. Clearly, (4.13) gives an estimate of order $O(h^{-2})$ of the condition number of the coefficient matrix in (4.12). Also, this inequality provides us with estimates on the smallest and largest eigenvalues of $\mathcal{A}$, which are useful in the convergence analysis of iterative solvers for (4.12).

As a final remark, we mention that there are other, different approaches for numerical solutions of div-curl and Maxwell's equations; see [1], [2], and the references therein. The approaches are based on the so-called de Rham finite element spaces, and the resulting discrete schemes also fulfill (3.1) and the relation $\nabla\cdot(\nabla\times\mathbf{u}) = 0$.

Acknowledgments. The authors would like to thank the referees whose constructive suggestions and comments improved the presentation of the work greatly.

REFERENCES

[1] M. J. Bluck, S. P. Walker, and R. Ordovas, *Time domain finite element methods for EM modelling*, in Proc. Euro. Congr. Comput. Methods Appl. Sci. Engrg., ECCOMAS Computational Fluid Dynamics Conference, Swansea, UK, 2001, pp. 201–214.

[2] D. Boffi, *A note on the discrete compactness property and the de Rham complex*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[3] E. Chung and J. Zou, *A finite volume method for Maxwell's equations with discontinuous physical coefficients*, Int. J. Appl. Math., 7 (2001), pp. 201–223.

[4] E. Chung and J. Zou, *The Eigenvalues and Eigenspaces of Some Discrete* Div- *and* Curl-*Related Operators,* Technical Report 2002-24 (264), Department of Mathematics, The Chinese University of Hong Kong, 2002.

[5] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[6] E. T. Chung, Q. Du, and J. Zou, *Convergence analysis of a finite volume method for Maxwell's equations in nonhomogeneous media*, SIAM J. Numer. Anal., 41 (2003), pp. 37–63.

[7] R. A. Nicolaides, *Direct discretization of planar* div-curl *problems*, SIAM J. Numer. Anal., 29 (1992), pp. 32–56.

[8] R. A. Nicolaides and D. Q. Wang, *Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions*, Math. Comp., 67 (1998), pp. 947–963.

[9] R. A. Nicolaides and X. Wu, *Covolume solutions of three-dimensional* div-curl *equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2195–2203.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1161–1166

COMMENTS ON A SHIFTED CYCLIC REDUCTION ALGORITHM FOR QUASI-BIRTH-DEATH PROBLEMS*

CHUN-HUA GUO†

Abstract. A shifted cyclic reduction algorithm has been proposed by He, Meini, and Rhee [*SIAM J. Matrix Anal. Appl.*, 23 (2001), pp. 673–691] for finding the stochastic matrix G associated with discrete-time quasi-birth-death (QBD) processes. We point out that the algorithm has quadratic convergence even for null recurrent QBDs. We also note that the approximations (to the matrix G) obtained by their algorithm are always stochastic when they are nonnegative.

Key words. matrix equations, minimal nonnegative solution, Markov chains, cyclic reduction, iterative methods, convergence rate

AMS subject classifications. 15A24, 15A51, 60J10, 60K25, 65H05

PII. S0895479802407901

1. Introduction.

A discrete-time quasi-birth-death (QBD) process is a Markov chain with state space $\{(i,j)\,|\,i \geq 0, 1 \leq j \leq m\}$, which has a transition probability matrix of the form

$$P = \begin{pmatrix} C_0 & C_1 & 0 & 0 & \cdots \\ A_0 & A_1 & A_2 & 0 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where C_0, C_1, A_0, A_1, and A_2 are $m \times m$ nonnegative matrices such that P is stochastic. In particular, $(A_0 + A_1 + A_2)e = e$, where e is the column vector with all components equal to one. The matrix P is also assumed to be irreducible.

We assume that $A = A_0 + A_1 + A_2$ is irreducible. Thus, there exists a unique vector $\alpha > 0$ with $\alpha^T e = 1$ and $\alpha^T A = \alpha^T$. The vector α is called the stationary probability vector of A. The QBD is positive recurrent if $\alpha^T A_0 e > \alpha^T A_2 e$ and null recurrent if $\alpha^T A_0 e = \alpha^T A_2 e$.

The minimal nonnegative solution G of the matrix equation

$$G = A_0 + A_1 G + A_2 G^2 \tag{1.1}$$

plays an important role in the study of the QBD process (see [8]). We will also need the equation

$$F = A_2 + A_1 F + A_0 F^2, \tag{1.2}$$

and we let F be its minimal nonnegative solution. It is well known (see [8], for example) that if the QBD is positive recurrent, then G is stochastic and F is substochastic with spectral radius $\rho(F) < 1$; if the QBD is null recurrent, then G and F are both stochastic.

*Received by the editors May 19, 2002; accepted for publication (in revised form) by I. C. F. Ipsen November 2, 2002; published electronically March 13, 2003. This work was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/simax/24-4/40790.html
†Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (chguo@math.uregina.ca).

Recently, a shift technique has been introduced in [6] to a cyclic reduction (CR) algorithm (see [3]) for finding the matrix G in the positive recurrent case, assuming that the only eigenvalue of G on the unit circle is the simple eigenvalue 1. We will make some comments on that interesting paper.

2. Comments. The shift technique introduced in [6] is $H = G - eu^T$, where $u > 0$ and $u^T e = 1$. Then the eigenvalues of H are those of G except that in H the eigenvalue 1 of G is replaced by 0, and H is a solution of the new equation

$$H = B_0 + B_1 H + B_2 H^2, \tag{2.1}$$

where

$$B_0 = A_0(I - eu^T), \quad B_1 = A_1 + A_2 eu^T, \quad B_2 = A_2. \tag{2.2}$$

The shifted CR algorithm is obtained in [6] by applying the CR algorithm to (2.1). For positive recurrent QBDs, it is shown in [6] that the convergence of the shifted CR algorithm is quadratic and faster than that of the CR algorithm, provided that no breakdown occurs. Here we point out that the same is true for null recurrent QBDs. This is a very important feature of the shift technique. Without using the shift technique, all previous methods for finding the matrix G have only linear or sublinear convergence for null recurrent QBDs. For example, the convergence of the Latouche–Ramaswami (LR) algorithm [7] is linear with rate 1/2 for null recurrent QBDs (see [5]). Since the CR algorithm and the LR algorithm are closely related (see [2]), the convergence of the CR algorithm is also linear with rate 1/2 for null recurrent QBDs. Once we have shown that the shift technique recovers quadratic convergence for the CR algorithm in the null recurrent case, then the same will be true for the LR algorithm.

Some work is needed to justify our claim about the shifted CR algorithm for null recurrent QBDs.

Let

$$A(\lambda) = -A_0 + (I - A_1)\lambda - A_2\lambda^2$$

be the matrix polynomial corresponding to (1.1), and let

$$B(\lambda) = -B_0 + (I - B_1)\lambda - B_2\lambda^2$$

be the matrix polynomial associated with (2.1). We first point out that there is a simple proof for the following generalization of Theorem 3.1 in [6].

LEMMA 2.1. *The zeros of* $\det(B(\lambda))$ *are obtained from the zeros of* $\det(A(\lambda))$ *by replacing one zero* 1 *with* 0.

Proof. Since

$$A(\lambda) = (I - A_1 - A_2 G - \lambda A_2)(\lambda I - G),$$

$$B(\lambda) = (I - B_1 - B_2 H - \lambda B_2)(\lambda I - H),$$

and

$$\begin{aligned} I - B_1 - B_2 H - \lambda B_2 &= I - (A_1 + A_2 eu^T) - A_2(G - eu^T) - \lambda A_2 \\ &= I - A_1 - A_2 G - \lambda A_2, \end{aligned}$$

the assertion follows immediately. ∎

Note that $\det(A(\lambda))$ has two zeros 1 for null recurrent QBDs, as seen from the following special case of Theorem 4 in [4].

LEMMA 2.2. *Assume that* $\det(A(\lambda)) \neq 0$ *if* $|\lambda| = 1, \lambda \neq 1$. *Then*

(1) *if the QBD is positive recurrent, then* $\det(A(\lambda))$ *has* $m-1$ *zeros inside the unit circle, one zero* 1, *and* m *zeros outside the unit circle (zeros at infinity are added if the degree of* $\det(A(\lambda))$ *is less than* $2m$*);*

(2) *if the QBD is null recurrent, then* $\det(A(\lambda))$ *has* $m-1$ *zeros inside the unit circle, two zeros* 1, *and* $m-1$ *zeros outside the unit circle (zeros at infinity are added if the degree of* $\det(A(\lambda))$ *is less than* $2m$*).*

We note that the assumption in Lemma 2.2 is equivalent to our earlier assumption that the only eigenvalue of G on the unit circle is the simple eigenvalue 1 (see [4]).

COROLLARY 2.3. *If the QBD is positive recurrent, then* $\det(B(\lambda))$ *has* m *zeros inside the unit circle and no zeros on the unit circle; if the QBD is null recurrent, then* $\det(B(\lambda))$ *has* m *zeros inside the unit circle, one (simple) zero* 1 *on the unit circle, and* $m-1$ *zeros outside the unit circle.*

When the QBD is positive recurrent, $u^T F e < 1$ and $I - eu^T F$ is nonsingular (see [6]). The following result plays a crucial role in [6] for the convergence analysis of the shifted CR algorithm.

LEMMA 2.4 (see [6]). *When the QBD is positive recurrent,*

$$K = (I - eu^T F)F(I - eu^T F)^{-1} \tag{2.3}$$

is a solution of

$$K = B_2 + B_1 K + B_0 K^2. \tag{2.4}$$

When the QBD is null recurrent, we have $Fe = e$. Thus, $(I - eu^T F)e = 0$ and $I - eu^T F$ is singular. The question then arises of whether the norm of the matrix K in (2.3) will become arbitrarily large when the QBD becomes nearly null recurrent. As noted in [6], there is a K-dependent operator norm $\|\cdot\|_K$ such that $\|K\|_K = \rho(K) = \rho(F) < 1$. However, the norm $\|\cdot\|_K$ would be drastically different from practically useful norms, such as $\|\cdot\|_\infty$, as the QBD becomes nearly null recurrent, if $\|K\|_\infty$ couldn't be bounded independent of the nearness to null recurrence. We have the following positive result in this regard. This result also will be the basis for proving quadratic convergence of the shifted CR algorithm in the null recurrent case.

LEMMA 2.5. *If the QBD is positive recurrent, then for the matrix* K *in* (2.3)

$$\|K\|_\infty < 3 + \frac{2}{\min_{1 \le i \le m} u_i},$$

where u_i *is the* i*th component of* u. *In particular,* $\|K\|_\infty < 3 + 2m$ *if* $u = \frac{1}{m}e$.

Proof. By the Sherman–Morrison–Woodbury formula,

$$(I - eu^T F)^{-1} = I + \frac{1}{1 - u^T Fe} eu^T F = I + \frac{1}{u^T(e - Fe)} eu^T F.$$

Thus,

$$K = (I - eu^T F)F + \frac{1}{u^T(e - Fe)}(I - eu^T F)Feu^T F.$$

Note that

$$\begin{aligned}(I - eu^T F)Feu^T F &= (I - eu^T F)eu^T F - (I - eu^T F)(e - Fe)u^T F \\ &= eu^T(e - Fe)u^T F - (I - eu^T F)(e - Fe)u^T F.\end{aligned}$$

Therefore,

$$K = (I - eu^T F)F + eu^T F - \frac{1}{u^T(e - Fe)}(I - eu^T F)(e - Fe)u^T F.$$

It follows that

$$\|K\|_\infty \le \|I - eu^T F\|_\infty \|F\|_\infty + \|eu^T F\|_\infty + \|I - eu^T F\|_\infty \|u^T F\|_\infty \frac{\|e - Fe\|_\infty}{u^T(e - Fe)}.$$

Since $Fe \le e$, $u^T Fe < 1$, and $eu^T Fe < e$, we have $\|F\|_\infty \le 1$, $\|u^T F\|_\infty < 1$, and $\|eu^T F\|_\infty < 1$. Thus, $\|I - eu^T F\|_\infty < 2$ and

$$\|K\|_\infty < 3 + \frac{2}{\min_{1 \le i \le m} u_i}.$$

This completes the proof. □

For the null recurrent case, the role of Lemma 2.4 will be assumed by the following result.

THEOREM 2.6. *If the QBD is null recurrent, then* (2.4) *has a solution* K *having one eigenvalue* 1 *and* $m - 1$ *eigenvalues inside the unit circle.*

Proof. Since the QBD is irreducible, $A_2 \ne 0$. Suppose that $A_2(i, j)$, the (i, j) element of A_2, is positive. For any ϵ with $0 < \epsilon < A_2(i, j)$, define

$$A_0(\epsilon) = A_0, \quad A_1(\epsilon) = A_1 + \epsilon E_{ij}, \quad A_2(\epsilon) = A_2 - \epsilon E_{ij},$$

where E_{ij} is the matrix with one in the (i, j) position and zeros elsewhere. Since $\alpha^T A_0(\epsilon)e > \alpha^T A_2(\epsilon)e$, where α is the stationary probability vector of $A = A_0 + A_1 + A_2 = A_0(\epsilon) + A_1(\epsilon) + A_2(\epsilon)$, the QBD corresponding to $(A_0(\epsilon), A_1(\epsilon), A_2(\epsilon))$ is positive recurrent. We now define

$$B_0(\epsilon) = A_0(\epsilon)(I - eu^T), \quad B_1(\epsilon) = A_1(\epsilon) + A_2(\epsilon)eu^T, \quad B_2(\epsilon) = A_2(\epsilon)$$

and let F_ϵ be the minimal nonnegative solution of

$$F = A_2(\epsilon) + A_1(\epsilon)F + A_0(\epsilon)F^2.$$

Thus, $\rho(F_\epsilon) < 1$. Moreover, $K_\epsilon = (I - eu^T F_\epsilon)F_\epsilon(I - eu^T F_\epsilon)^{-1}$ is a solution of

$$K = B_2(\epsilon) + B_1(\epsilon)K + B_0(\epsilon)K^2$$

by Lemma 2.4. Let the sequence $\{\epsilon_n\}$ be such that $0 < \epsilon_n < A_2(i, j)$ and $\lim \epsilon_n = 0$. Since the sequence $\{K_{\epsilon_n}\}$ is bounded by Lemma 2.5, it has a limit point K. It is clear that this matrix K is a solution of (2.4). Since $\rho(K_{\epsilon_n}) < 1$, we have $\rho(K) \le 1$. Since the zeros of $\det(\hat{B}(\lambda))$, where

$$\hat{B}(\lambda) = -B_2 + (I - B_1)\lambda - B_0\lambda^2,$$

are the reciprocals of the zeros of $\det(B(\lambda))$, and the eigenvalues of K are part of the zeros of $\det(\hat{B}(\lambda))$, we know from Corollary 2.3 that K has $m - 1$ eigenvalues inside the unit circle and one eigenvalue 1. □

The shifted CR algorithm generates a sequence $\widehat{B}_1^{(n)}$ (if no breakdown occurs), and approximations $\tilde{H}_n$ to the matrix H are obtained by $\tilde{H}_n = (I - \widehat{B}_1^{(n)})^{-1}B_0$ (see

[6]). Approximations $\tilde{G}_n$ to the matrix G can be obtained using $\tilde{G}_n = \tilde{H}_n + eu^T$. It is noted in [6] that we also have $\tilde{G}_n = (I - \widehat{B}_1^{(n)})^{-1} A_0$.

For the null recurrent case, the spectral properties of the matrix K in Theorem 2.6 are crucial to show the quadratic convergence of the sequence $\{\tilde{G}_n\}$. Once Theorem 2.6 is proved, quadratic convergence follows from known results.

Let K be the solution of (2.4) given by Lemma 2.4 for the positive recurrent case and given by Theorem 2.6 for the null recurrent case. We have from the discussions in [6] or from Theorem 16 and Remark 17 of [1] that

$$\limsup_{n\to\infty} \sqrt[2^n]{\|\tilde{G}_n - G\|_\infty} \le \rho(K)\rho(H) = \rho(F)\rho(H) < 1. \tag{2.5}$$

In particular, $\tilde{G}_n$ converges to G quadratically for both positive recurrent and null recurrent QBDs. If we apply the CR algorithm directly to (1.1), the approximations G_n for G are such that

$$\limsup_{n\to\infty} \sqrt[2^n]{\|G_n - G\|_\infty} \le \rho(F)\rho(G) \le 1. \tag{2.6}$$

Thus, the convergence of $\{G_n\}$ is slower than that of $\{\tilde{G}_n\}$. One good thing about the sequence $\{G_n\}$ is that it is monotonically increasing to G (see [3]). Thus, $\|G_n e - e\|_\infty = \|(G_n - G)e\|_\infty = \|G_n - G\|_\infty$. So, the actual error $\|G_n - G\|_\infty$ can be obtained easily even though G is not known. For the sequence $\{\tilde{G}_n\}$, the actual error $\|\tilde{G}_n - G\|_\infty$ cannot be obtained in this way. In fact, since $B_0 e = A_0(I - eu^T)e = 0$, we have $\tilde{H}_n e = 0$ and $\tilde{G}_n e = e$ for each $n \ge 0$. Therefore, the matrices $\tilde{G}_n$ are stochastic when they are nonnegative, and we always have $\|\tilde{G}_n e - e\|_\infty = 0$ (in exact arithmetic) no matter how large $\|\tilde{G}_n - G\|_\infty$ is. Nevertheless, computing the values $\|\tilde{G}_n e - e\|_\infty$ in the presence of rounding errors is still of interest. If these values are close to the machine epsilon, we could reasonably assume that the effect of rounding errors on the algorithm is minor. On the other hand, we would have to use the residual error to measure the accuracy of the approximation $\tilde{G}_n$.

We define functions $\mathcal{F}_A$, $\mathcal{F}_B : \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$ by

$$\mathcal{F}_A(X) = X - A_0 - A_1 X - A_2 X^2, \quad \mathcal{F}_B(X) = X - B_0 - B_1 X - B_2 X^2.$$

In [6], the accuracy of the approximations $\tilde{G}_n$ and G_n is compared using the residual errors $\|\mathcal{F}_A(\tilde{G}_n)\|_\infty$ and $\|\mathcal{F}_A(G_n)\|_\infty$. The reported values for $\|\mathcal{F}_A(\tilde{G}_n)\|_\infty$ and $\|\mathcal{F}_A(G_n)\|_\infty$ are roughly of the same magnitude. We note that this does not mean that $\tilde{G}_n$ and G_n have roughly the same accuracy. In fact, $\tilde{G}_n$ is typically much more accurate than G_n when the QBD is null recurrent or nearly null recurrent. The reason for this is the following. When the QBD is null recurrent, the Fréchet derivative of $\mathcal{F}_A$ at the solution G is a singular map. Thus, in general, $\|G_n - G\|_\infty$ is not of the order of $\|\mathcal{F}_A(G_n)\|_\infty = \|\mathcal{F}_A(G_n) - \mathcal{F}_A(G)\|_\infty$. On the other hand, the Fréchet derivative of $\mathcal{F}_B$ at the solution H is a nonsingular map. Using $\tilde{H}_n e = 0$, it is easy to show that $\mathcal{F}_A(\tilde{G}_n) = \mathcal{F}_B(\tilde{H}_n)$. Thus,

$$\|\tilde{G}_n - G\|_\infty = \|\tilde{H}_n - H\|_\infty = O(\|\mathcal{F}_B(\tilde{H}_n) - \mathcal{F}_B(H)\|_\infty) = O(\|\mathcal{F}_A(\tilde{G}_n)\|_\infty).$$

Acknowledgment. The author thanks the referees for their helpful comments.

REFERENCES

[1] D. A. Bini, L. Gemignani, and B. Meini, *Computations with infinite Toeplitz matrices and polynomials*, Linear Algebra Appl., 343/344 (2002), pp. 21–61.

[2] D. A. Bini, G. Latouche, and B. Meini, *Solving matrix polynomial equations arising in queueing problems*, Linear Algebra Appl., 340 (2002), pp. 225–244.

[3] D. Bini and B. Meini, *On the solution of a nonlinear matrix equation arising in queueing problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 906–926.

[4] H. R. Gail, S. L. Hantler, and B. A. Taylor, *Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains*, Adv. in Appl. Probab., 28 (1996), pp. 114–165.

[5] C.-H. Guo, *Convergence analysis of the Latouche–Ramaswami algorithm for null recurrent quasi-birth-death processes*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 744–760.

[6] C. He, B. Meini, and N. H. Rhee, *A shifted cyclic reduction algorithm for quasi-birth-death problems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 673–691.

[7] G. Latouche and V. Ramaswami, *A logarithmic reduction algorithm for quasi-birth-death processes*, J. Appl. Probab., 30 (1993), pp. 650–674.

[8] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, PA, ASA, Alexandria, VA, 1999.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1167–1174

A BOUND FOR THE INVERSE OF A LOWER TRIANGULAR TOEPLITZ MATRIX*

ANTONIA VECCHIO†

Abstract. The expression of a bound of the uniform norm of infinite lower triangular Toeplitz matrices with nonnegative entries is found. All the results are obtained by studying the behavior of the resolvent kernel and of the fundamental matrix of the recurrence relation, which generates the sequence of the entries of the considered matrix.

Key words. inverse matrix, Toeplitz matrix, recurrence formula

AMS subject classifications. 15A09, 61A45, 39A10

PII. S0895479801396762

1. Introduction. We consider the class of lower triangular matrices belonging to $R^{(n+1)\times(n+1)}$,

$$A_n = \begin{pmatrix} a_0 & & & & \\ a_1 & a_0 & & & \\ a_2 & a_1 & a_0 & & \\ . & . & . & . & . \\ a_n & ... & . & a_1 & a_0 \end{pmatrix}, \tag{1.1}$$

the entries of which depend only on the difference between the row and the column numbers. This type of matrix, also called isoclinal or semicirculant, can be viewed as the $n \times n$ truncations of an infinite lower triangular Toeplitz matrix. We refer to [1, 2, 9, 11] for a variety of theorems on Toeplitz matrices and their inverse.

Lower triangular matrices of type (1.1) arise naturally from the application of numerical methods to ordinary differential equations (see, for example, [15]) and to convolution-type integral and integrodifferential equations [13, Chap. 7], [4, Chap. 3], [3, 14, 16]. Moreover, these matrices arise in the matrix representation of formal power series [12] and they are related to probability theory [8, p. 329]. We are concerned with the problem of determining whether the inverse of (1.1) is bounded independently of its dimension.

There exist some results which (directly or indirectly) give information on the inverse of this particular kind of Toeplitz matrix (see [13, Thm. 10.1, p. 173], [6], and [14, Thm. 4.1]) and all of them require the summability of the series of the matrix entries, i.e.,

$$\sum_{n=0}^{\infty} |a_n| < \infty, \tag{1.2}$$

and/or some conditions on the series $a(x) = \sum_{n=0}^{\infty} a_n x^n$ [13, 14]. Our aim is to obtain an explicit bound for $\|A_n^{-1}\|_\infty$, where A is defined by (1.1), without requiring (1.2). Such a result is contained in section 2, where we prove that the uniform norm of

*Received by the editors October 22, 2001; accepted for publication (in revised form) by L. Reichel October 2, 2002; published electronically March 13, 2003.
http://www.siam.org/journals/simax/24-4/39676.html
†Instituto per Applicazioni del Calcolo "M. Picone," Sez. di Napoli, CNR, Via P. Castellino, 111, 80131 Napoli, Italy (vecchio@iam.na.cnr.it).

the inverse of a lower triangular Toeplitz matrix with nonnegative entries is bounded uniformly in n. The conditions we require on the sequence of entries $\{a_n\}_{n\geq 0}$ are related to the sign of its first difference.

All the results are obtained by studying the behavior of the resolvent kernel of the recurrence relation generating the sequence of the entries of A_n^{-1}. As we shall see later, such a relation is nothing but an explicit unbounded order difference equation of convolution type. Several researchers, including the author, have treated such equations in many papers with the goal of studying the stability of numerical methods for Volterra integral and integrodifferential equations; see, for example, [5, 17, 20] and the references therein. Of course, the results presented in this paper also can be exploited in the above mentioned contexts as illustrated in the last section.

2. The inverse matrix. It is well known that the inverse of the lower triangular Toeplitz matrix (1.1) exists if and only if $a_0 \neq 0$. We will assume that this condition holds throughout this paper. If we let $B = A^{-1}$, then B also is a lower triangular Toeplitz matrix,

$$B_n = \begin{pmatrix} b_0 & & & & \\ b_1 & b_0 & & & \\ b_2 & b_1 & b_0 & & \\ . & . & . & . & . \\ b_n & ... & . & b_1 & b_0 \end{pmatrix} \in R^{(n+1)\times(n+1)}, \tag{2.1}$$

where b_i can be obtained by the recurrence formula [13, p. 172]

$$b_0 = \frac{1}{a_0}, \quad b_n = -\frac{1}{a_0}\sum_{l=0}^{n-1} a_{n-l}b_l, \quad n \geq 1. \tag{2.2}$$

As we already mentioned in the introduction, this equation is a linear difference equation of unbounded order. We recall some concepts and results from the theory of such equations.

2.1. Resolvent kernel and fundamental matrix of unbounded difference equations. The solution x_n of

$$x_n = \sum_{l=0}^{n-1} \alpha_{n-l}x_l, \quad n \geq 1, \quad x_0 \text{ given}, \tag{2.3}$$

can be written as [7, 17]

$$x_n = r_n x_0, \quad n \geq 1, \tag{2.4}$$

where r_n is called the resolvent kernel of (2.3) and satisfies

$$r_n = \sum_{l=0}^{n-1} \alpha_{n-l}r_l, \quad n \geq 1, \quad r_0 = 1. \tag{2.5}$$

The sequence $\{r_n\}$ is related to another sequence $\{u_n\}$, known as the fundamental matrix of (2.3), by the following equality [20]:

$$r_n = u_n - u_{n-1}, \quad n \geq 1, \tag{2.6}$$

with

$$u_n = 1 + \sum_{l=0}^{n} \alpha_{n-l} u_l, \quad n \geq 0, \alpha_0 = 0. \tag{2.7}$$

In Theorem 2.1 of [20] the following result is included.

LEMMA 2.1. *Assume that*

(1) $\alpha_n \leq 0, \quad n \geq 0,$
(2) $\alpha_1 + 1 \geq 0,$
(3) $\Delta\alpha_n = \alpha_{n+1} - \alpha_n \geq 0, \quad n \geq 1.$

Then

$$0 \leq u_n \leq 1.$$

2.2. Main results. Taking into account the mentioned results, the entries of B^{-1} given in (2.2) can be expressed as

$$b_n = r_n b_0, \quad n \geq 1, \quad b_0 = \frac{1}{a_0}, \tag{2.8}$$

where r_n satisfies

$$r_n = -\sum_{l=0}^{n-1} \frac{a_{n-l}}{a_0} r_l, \quad r_0 = 1, \tag{2.9}$$

and u_n satisfies

$$u_n = 1 - \sum_{l=0}^{n-1} \frac{a_{n-l}}{a_0} u_l, \quad n \geq 0. \tag{2.10}$$

THEOREM 2.2. *Assume that*

(i) $a_n > 0, \quad n \geq 0,$
(ii) $\Delta a_n \leq 0, \quad n \geq 0,$
(iii) $\inf a_n = a > 0.$

Then

$$\|A_n^{-1}\|_\infty \leq \frac{2}{a} + \frac{1}{a_0}. \tag{2.11}$$

Proof. Consider the sequence $\{u_n\}$ given in (2.10). Hypotheses (i) and (ii) ensure that Lemma 2.1 holds, and hence

$$0 \leq u_n \leq 1. \tag{2.12}$$

Taking into account (iii), we have

$$a \sum_{l=1}^{n} u_{n-l} \leq \sum_{l=1}^{n} a_l u_{n-l} = a_0 \sum_{l=1}^{n} \frac{a_l}{a_0} u_{n-l}$$

and, in view of (2.10),

$$a \sum_{l=0}^{n-1} u_l \leq a_0 (1 - u_n)$$

which, because of (2.12), leads to

$$\sum_{l=0}^{n-1} u_l \leq \frac{a_0}{a}.$$

From (2.6) there results

$$\sum_{n=1}^{\infty} |r_n| \leq 2 \sum_{n=0}^{\infty} |u_n| \leq 2\frac{a_0}{a}, \tag{2.13}$$

and (2.8) implies

$$\|A_n^{-1}\|_\infty = \sum_{l=0}^{n} |b_l| \leq |b_0| \left(1 + \sum_{l=1}^{n} |r_l|\right). \tag{2.14}$$

The desired result follows from here and (2.13). □

Example. The matrix A given by

$$a_0 = 1, \quad a_n = \frac{1}{2}\left(\frac{1}{n} + 1\right), \quad n \geq 0,$$

satisfies the hypotheses of Theorem 2.2 with

$$a = \lim_{n\to\infty} a_n = \frac{1}{2},$$

and hence its inverse is bounded by

$$\|A_n^{-1}\|_\infty \leq 5.$$

Observe that condition (1.2) required in [6, 14] is not satisfied, whereas since

$$\lim_{n\to\infty} \frac{a_n}{a_{n+1}} = 1,$$

the condition

$$\sum_{n=0}^{\infty} a_n z^n < \infty, \quad |z| < 1,$$

is verified. Nevertheless, Theorem 10.1 of [13] cannot be applied since the hypothesis

$$\frac{a_{n+1}}{a_n} \geq \frac{a_n}{a_{n-1}} \tag{2.15}$$

is not fulfilled for $n = 1$.

To check whether the bound (2.11) is tight we have numerically computed $\|A_n^{-1}\|_\infty$ for an increasing value of n and have obtained $\lim_{n\to\infty} \|A_n^{-1}\|_\infty < 2.55$. From here and other numerical examples we conjecture that the quantity $\frac{2}{a} + \frac{1}{a_0}$, appearing in (2.11), is less than double the true value of $\|A_n^{-1}\|_\infty$. This is probably a consequence of (2.6), (2.12), and the first inequality of (2.13).

Observe that the summability of the series $\sum_{n=0}^{\infty} u_n$ is a crucial step in the proof of Theorem 2.2. Now we show that hypothesis (iii) of this theorem is necessary for getting this property, provided that (i) and (ii) hold.

THEOREM 2.3. *Assume that*

(i) $a_n \geq 0$,

(ii) $\Delta a_n \leq 0, \quad n \geq 0$,

(iii) $\inf a_n = 0$.

Then the series $\sum_{n=0}^{\infty} u_n$ *is not convergent.*

Proof. As in the previous theorem, (i) and (ii) ensure (2.12). Assume

$$\sum_{n=0}^{\infty} u_n < \infty. \tag{2.16}$$

From (2.10) we get

$$1 - u_n = \sum_{l=0}^{n} \beta_l u_{n-l} \tag{2.17}$$

with $\beta_0 = 0, \beta_n = a_n/a_0, n \geq 1$. The right-hand side can be considered as the nth term of a sequence $\{\eta_n\}$ obtained by the convolution of sequences $\{\beta_n\}$ and $\{u_n\}$ which, respectively, satisfy

$$\lim_{n\to\infty} \beta_n = 0, \quad \sum_{n=0}^{\infty} |u_n| < \infty.$$

As it can easily be seen, this implies $\lim_{n\to\infty} \eta_n = 0$ or, equivalently, $\lim_{n\to\infty} u_n = 1$ which contradicts (2.16). Thus the desired result is proved by contradiction. □

3. An application. In this section we want to show how the main result of section 2.2 can be useful in the stability analysis of linear methods for solving Volterra integral equations (VIEs). In recent years we proved many results on this subject [16, 19, 18] but most of them consider low order methods and/or nonconvolution Volterra equations. Only recently in [20] we proved the boundedness of the global error of direct quadrature (DQ) methods, up to order three, for solving second kind VIEs [20, sect. 3] of the type

$$y(t) = g(t) + \int_0^t k(t,s)y(s)ds, \quad t \in [0,T], \quad y, g, k \in R. \tag{3.1}$$

The mentioned results regard methods of order higher than the previous ones, but they still must be applied only to nonconvolution kernels $k(t,s)$, which satisfy

$$\int_0^{\infty} |k(t,t)|dt < \infty$$

and do not cover the case of integral equations appearing in a large variety of applied problems, ranging from population dynamics to renewal theory [10], that is, VIEs with a convolution kernel ($k(t,s) = k(t-s)$).

Using Theorem 2.2, the stability result obtained in [20] can be extended to the case of DQ methods applied to VIEs with convolution nonsummable kernels.

According to the notation used in [20], the error E_n (i.e., the difference between the analytical and numerical solutions) due to the application of a DQ method to

$$y(t) = g(t) + \int_0^t k(t-s)y(s)ds, \quad t \in [0,T], \quad y, g, k \in R,$$

satisfies

$$A_n E_n = \Gamma_n, \tag{3.2}$$

where A_n is defined in (1.1) with

$$a_0 = 1 - hw_0k_0, \quad a_n = -hw_nk_n, \quad n \geq 1, \tag{3.3}$$

$$E_n = [e_0, \ldots, e_n]^T, \quad \Gamma_n = [\gamma_0, \ldots, \gamma_n]^T \tag{3.4}$$

with

$$|\gamma_n| \leq \gamma, \quad n \geq 0. \tag{3.5}$$

Here w_n are the entries of the following infinite matrices identifying, respectively, the backward Euler (BE), trapezoidal (TR), and third order (ρ, σ) reducible methods:

BE method. Order 1.

$$W = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ 1 & 1 & 1 & 1 & \\ . & . & . & . & \end{pmatrix}. \tag{3.6}$$

TR method. Order 2.

$$W = \begin{pmatrix} \frac{1}{2} & & & & \\ 1 & \frac{1}{2} & & & \\ 1 & 1 & \frac{1}{2} & & \\ 1 & 1 & 1 & \frac{1}{2} & \\ . & . & . & . & \end{pmatrix}. \tag{3.7}$$

(ρ, σ) reducible method of order 3.

$$W = \frac{1}{12} \begin{pmatrix} 5 & & & & \\ 13 & 5 & & & \\ 12 & 13 & 5 & & \\ 12 & 12 & 13 & 5 & \\ . & . & . & . & \end{pmatrix}. \tag{3.8}$$

Let c_1 and c_2 assume the following values for the different methods (3.6), (3.7), (3.8):

	c_1	c_2
BE	∞	$1 - hk_0$
TR	2	$1 - \frac{h}{2}k_0$
(ρ, σ)	3/2	$1 - \frac{5}{12}hk_0$

THEOREM 3.1. *Assume that*
(i) $k(t) \le 0, \quad t \ge 0,$
(ii) $k'(t) \ge 0, \quad t \ge 0,$
(iii) $\lim_{t\to\infty} k(t) = \tilde{k} < 0,$
(iv) $h|k(0)| < c_1.$
Then the global error of the methods under consideration satisfies

$$\|E_n\|_\infty \le -\frac{2}{h\tilde{k}} + \frac{1}{c_2}.$$

Proof. We give the proof only in the case of the TR method. The BE and (ρ, σ) methods can be handled analogously.

From (3.2) there results

$$\|E_n\|_\infty \le \|A_n^{-1}\|_\infty \|\Gamma_n\|_\infty \tag{3.9}$$

with

$$a_n = \begin{cases} 1 - \frac{h}{2}k_0, & n = 0, \\ -hk_n, & n \ge 1. \end{cases} \tag{3.10}$$

From (3.10) and Theorem 3.1(i) we immediately have that $a_n \ge 0, n \ge 0$. Moreover, (ii) and (iii) ensure that $a_{n+1} \le a_n, n \ge 1$, and $\inf_n a_n = -h\tilde{k} > 0$. From (iv) we find $\frac{h}{2}|k(0)| < 1$, and thus

$$hk_1 - \frac{h}{2}k_0 \ge \frac{h}{2}k_0 > -1.$$

This ensures $a_1 - a_0 \le 0$. The desired result follows by the application of Theorem 2.2, (3.9), and (3.5). ∎

4. Concluding remarks. Starting from some results on the behavior of the resolvent kernel of an unbounded order difference equation, we obtain an explicit bound of the inverse of a lower triangular Toeplitz matrix. Only simple conditions on the matrix are required. As an example of application we show how the main theorem can be useful in the study of numerical stability of some linear methods for VIEs. Namely, we bound the global error of DQ methods for solving VIEs with a convolution kernel by requiring certain simple conditions on the kernel of the considered equation and a restriction on the stepsize of the numerical methods.

REFERENCES

[1] A. BÖTTCHER AND S. GRUDSKY, *Toeplitz Matrices, Asymptotic Linear Algebra and Functional Analysis*, Birkhäuser, Basel, 2000.
[2] A. BÖTTCHER AND B. SILBERMANN, *Introduction to Large Truncated Toeplitz Matrices,* Springer, New York, 1999.
[3] H. BRUNNER, M. R. CRISCI, V. B. E. RUSSO, AND A. VECCHIO, *A family of methods for Abel integral equations of the second order*, J. Comput. Appl. Math., 34 (1991), pp. 211–219.
[4] H. BRUNNER AND P. J. VAN DER HOUWEN, *The Numerical Solution of Volterra Equations*, CWI Monographs 3, North-Holland, Amsterdam, 1986.
[5] M. R. CRISCI, V. B. KOLMANOVSKII, E. RUSSO, AND A. VECCHIO, *Stability of discrete Volterra equations of Hammerstein type*, J. Differ. Equations Appl., 6 (2000), pp. 127–145.
[6] P. P. B. EGGERMONT, *A new analysis of the trapezoidal-discretization method for the numerical solution of Abel-type integral equations*, J. Integral Equations, 3 (1981), pp. 317–332.
[7] S. N. ELAYDI, *An Introduction to Difference Equations*, Springer, New York, 1996.

[8] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley & Sons, New York, 1950.
[9] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, 2nd ed., Chelsea, New York, 1984.
[10] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.
[11] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhäuser, Basel, 1984.
[12] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 1, John Wiley & Sons, New York, 1974.
[13] P. Linz, *Analytical and Numerical Methods for Volterra Equations*, SIAM, Philadelphia, 1985.
[14] Ch. Lubich, *On the stability of linear multistep methods for Volterra convolution equations*, IMA J. Numer. Anal., 3 (1983), pp. 439–465.
[15] M. Meneguitte, *A note on bordered isoclinal matrices*, J. Comput. Appl. Math., 17 (1987), pp. 375–377.
[16] A. Vecchio, *Stability results on some direct quadrature methods for Volterra integro-differential equations*, Dynam. Systems Appl., 7 (1998), pp. 501–518.
[17] A. Vecchio, *On the resolvent kernel of Volterra discrete equations*, Funct. Differ. Equations, 6 (1999), pp. 191–201.
[18] A. Vecchio, *Stability of backward differentiation formulas for Volterra integro-differential equations*, J. Comput. Appl. Math., 115 (2000), pp. 565–576.
[19] A. Vecchio, *Stability of some linear and nonlinear methods for Volterra integral equations*, J. Integral Equations Appl., 12 (2000), pp. 449–465.
[20] A. Vecchio, *Volterra discrete equations: Summability of the fundamental matrix*, Numer. Math., 89 (2001), pp. 783–794.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1175–1185

LYAPUNOV EXPONENTS OF SYSTEMS EVOLVING ON QUADRATIC GROUPS*

LUCA DIECI† AND LUCIANO LOPEZ‡

Abstract. In this paper we show some symmetry properties of Lyapunov exponents of a dynamical system when the linearized problem evolves on a quadratic group, $X^T H X = H$, with H orthogonal. It is well understood that in this case the exponents are symmetric with respect to the origin. Here, we give lower bounds on the number of Lyapunov exponents which are 0 and show that some Lyapunov exponents may have even multiplicity.

Key words. singular values, Lyapunov exponents, regular systems, Lorentz group, Minkowski group, quadratic group

AMS subject classifications. 65F, 65L

PII. S0895479801391011

1. Introduction.

Lyapunov exponents are commonly used to explore stability properties of dynamical systems; e.g., see the collection of works in [3, 4, 15] and the many references there. Given the n-dimensional system of differential equations defined for $t \geq 0$,

$$\dot{x} = f(x), \quad x(0) = x_0, \tag{1.1}$$

the Lyapunov exponents are a characterization of the asymptotic properties of the solution $\phi^t x_0$ via analysis of the linearized problem: $dX/dt = f_x(\phi^t x_0)X$. More generally, we may consider the linear time varying system

$$\dot{x} = A(t)x, \quad A : \mathbb{R}^+ \to \mathbb{R}^{n\times n}. \tag{1.2}$$

By Φ we will indicate the *principal matrix solution* of (1.2), that is, $\dot{\Phi} = A(t)\Phi$, $\Phi(0) = I$, and by X any other fundamental matrix solution (that is, $X(t) = \Phi(t)X(0)$, $X(0)$ invertible). We assume that A is bounded and continuous.

Formally, the Lyapunov exponents associated to (1.2) may be defined as follows (e.g., see [1, 5] and cf. [14]). Let X be a fundamental matrix solution of (1.2), and let $\{e_i\}$ be the standard basis of $\mathbb{R}^n$. Define the numbers $\lambda_i(X)$, $i = 1, \ldots, n$, as (in this paper, the norm is always the 2-norm)

$$\lambda_i(X) = \limsup_{t\to\infty} \frac{1}{t} \log \|X(t)e_i\|. \tag{1.3}$$

When the sum of the $\lambda_i(X)$ is minimized over all initial conditions $X(0)$, the corresponding fundamental solution X is called *normal* and the numbers $\lambda_i(X)$, hereafter simply λ_i, $i = 1, \ldots, n$, are called (upper) *Lyapunov exponents* of the system. In general (see [1]), the Lyapunov exponents satisfy

$$\sum_{i=1}^{n} \lambda_i \geq \limsup_{t\to\infty} \frac{1}{t} \int_0^t \operatorname{trace}\big(A(s)\big) ds. \tag{1.4}$$

*Received by the editors June 15, 2001; accepted for publication (in revised form) by U. Helmke November 20, 2002; published electronically May 6, 2003. This work was supported in part by NSF grant DMS-9973266 and CNR grant 99.01704.CT01.

http://www.siam.org/journals/simax/24-4/39101.html

†School of Mathematics, Georgia Tech, Atlanta, GA 30332 (dieci@math.gatech.edu).

‡Dipartimento di Matematica, University of Bari, I-70125, Bari, Italy (lopezl@dm.uniba.it).

The normal fundamental matrix solution X, or just the system (1.2), is said to be *regular* if the time average of the trace in (1.4) has a finite limit and equality holds in (1.4). If X is regular, then the lim sups in (1.3) can be replaced by ordinary limits. Suppose that (1.2) is regular. Clearly, there are at most n distinct Lyapunov exponents. We will call *Lyapunov spectrum* the collection of all Lyapunov exponents of the system, counted with their multiplicity, and indicate it with $\mathrm{Sp}(X)$.

It is well known that $\mathrm{Sp}(X)$ is unchanged under an orthogonal (time varying) transformation of X. That is, if $R = Q^T X$, Q an orthogonal function, then $\mathrm{Sp}(R) = \mathrm{Sp}(X)$. This fact is often used in computational works (e.g., see [5, 6]), whereby the orthogonal change of variable is used to triangularize X, and thus one brings the coefficient matrix A in (1.2) to upper triangular form, say B. Then (see [1]), regularity implies that the Lyapunov exponents are given by

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \int_0^t B_{ii}(s)ds, \quad i = 1, \ldots, n. \tag{1.5}$$

To infer regularity of a given particular system is not easy. It is therefore important that regularity is a prevalent condition in a certain measure theoretic sense. Furthermore, since (1.2) typically arises from linearization of (1.1), the dependency of $\mathrm{Sp}(X)$ on the initial condition x_0 of (1.1) must also be assessed. These issues are at the heart of the theory of Oseledec. We refer to [5, 8, 12, 14] for details; here we highlight only some of the points from these works.

Suppose that ϕ^t, the flow of (1.1), is a flow on a smooth compact manifold M and let μ be an *invariant probability measure* on M (that is, $\mu(\phi^t A) = \mu(A)$ for all Borel sets A in M). The invariant measure μ is called *ergodic* if every set invariant under ϕ^t has measure 0 or 1. Let Φ_{x_0} be the principal matrix solution associated to the linearization of (1.1) along $\phi^t x_0$. We will write $\mathrm{Sp}\big(\Phi_{x_0}\big)$ for the Lyapunov spectrum (since it generally depends on x_0).

THEOREM 1.1. *Under the above assumptions, there is a subset M_0 of M, invariant under ϕ^t, and of measure* 1*, such that for any $x_0 \in M_0$ the following hold:*

(i) Φ_{x_0} *is regular.*

(ii) *The following limit exists:*[1]

$$\lim_{t\to\infty} \frac{1}{t} \log\big(\Phi_{x_0}^T(t)\Phi_{x_0}(t)\big)^{1/2}. \tag{1.6}$$

(iii) $\mathrm{Sp}\big(\Phi_{x_0}\big)$ *is given by the eigenvalues of the symmetric matrix defined by* (1.6).

(iv) *If μ is ergodic, then* $\mathrm{Sp}\big(\Phi_{x_0}\big)$ *is independent of $x_0 \in M_0$.*

From (1.6), we see that $\mathrm{Sp}\big(\Phi_{x_0}\big)$ is given by the limits of the time averages of the logarithms of the singular values of the principal matrix solution $\Phi_{x_0}(t)$. We refer to [9, 10] for numerical approximation of $\mathrm{Sp}\big(\Phi_{x_0}\big)$ exploiting this point of view. However, regardless of whether one adopts (1.5) or (1.6) as the basis of an algorithm to approximate the Lyapunov exponents, it must be appreciated that either one of (1.5) or (1.6) can be specialized to target the p most dominant Lyapunov exponents, e.g., all the positive Lyapunov exponents of a system.[2] This is convenient, since one may know beforehand that the Lyapunov spectrum enjoys some symmetries. Inarguably, the most important symmetry of the spectrum is the one with respect

[1] For all t, $\log(\Phi_{x_0}^T(t)\Phi_{x_0}(t))^{1/2}$ is the unique symmetric logarithm of the unique symmetric positive definite square root.

[2] Of relevance to approximate the *entropy*; see [8].

to the origin. This property is well known in the symplectic case (see [5, 8, 13]). In this work, we give some results on symmetries of Lyapunov exponents associated to fundamental matrix solutions evolving on other quadratic groups, namely, those for which $X^T(t)HX(t) = H$, with $H^TH = I$, for all t. To be precise, in this case, we will be able to give lower bounds on the number of singular values of X, which are identically 1 for all t, by looking at the distribution of eigenvalues of the matrix H defining the quadratic group. We will further give some bounds on the number of singular values of X which have even multiplicity. These facts, coupled with Theorem 1.1, will translate into bounds on the Lyapunov exponents of $\mathrm{Sp}(\Phi_{x_0})$. As a result, one may end up having to approximate only a few Lyapunov exponents in order to recover the entire Lyapunov spectrum. In particular, our results will apply to the case of the Lorentz and Minkowski groups. Maxwell's equations are the most famous example of a system satisfying invariance under the Lorentz group, and in this case only one Lyapunov exponent will need to be approximated; for this, and other examples of systems invariant under the Lorentz and Minkowski groups, see [2].

2. How many Lyapunov exponents are zero? The following result is essentially given by Gupalo, Kaganovich, and Cohen in [11].

THEOREM 2.1. *Let X be a fundamental matrix solution of* (1.2)*, and suppose that, for all t, $X(t)$ verifies*

$$\text{(2.1)} \qquad \text{(a)} \quad X^T(t)HX(t) = H \qquad \textit{and} \qquad \text{(b)} \quad X(t)HX^T(t) = H,$$

where $H \in \mathbb{R}^{n\times n}$ is nonsingular. Then the function A in (1.2) *satisfies for all t*

$$\text{(2.2)} \qquad \text{(a)} \quad A^T(t)H + HA(t) = 0 \qquad \textit{and} \qquad \text{(b)} \quad A(t)H + HA^T(t) = 0.$$

Further, the logarithms of the singular values of $X(t)$ are symmetric with respect to the origin for all t. Finally, under the assumptions and with the notation of Theorem 1.1*, i.e., if Φ_{x_0}, $x_0 \in M_0$, satisfies* (2.1)*, then*

$$\text{(2.3)} \qquad \mathrm{Sp}(\Phi_{x_0}) \qquad \textit{is symmetric with respect to the origin.}$$

In this paper, we are interested in exploring further symmetries of Lyapunov exponents. From (2.3) in Theorem 2.1, if the dimension n is an odd number, then obviously there must be at least one Lyapunov exponent equal to 0. But, in general, can we anticipate how many Lyapunov exponents are guaranteed to be 0?

To make some progress, we will assume that H in Theorem 2.1 is orthogonal:

$$\text{(2.4)} \qquad X^T(t)HX(t) \;=\; H \text{ for all } t, \quad H^TH = HH^T = I;$$

that is, a fundamental matrix solution X evolves on the quadratic group defined by the orthogonal matrix H. In the case of (2.4), either (a) or (b) in (2.1) and (2.2) is redundant. Indeed, from (2.1)(a) we have $X^T(t)HX(t) = H \Leftrightarrow H^TX^T(t)H = X^{-1}(t) \Leftrightarrow X(t)H^TX^T(t) = H^T \Leftrightarrow X(t)HX^T(t) = H$, and similarly for (2.2).

With this special choice of H orthogonal, we will next show some properties of the singular values of X. These properties, coupled with (1.6), will then be used to obtain bounds on the number of Lyapunov exponents which are zero, and will further tell if some of them have even multiplicity.

Example 2.2. Naturally, the orthogonal group is included in (2.4) if $H = I_n$; in this case, all Lyapunov exponents are 0. Further, the symplectic group is also included if $H = J$ with

$$\text{(2.5)} \qquad J = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}.$$

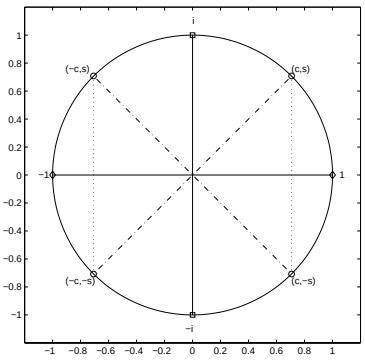


FIG. 2.1.

In this case, a priori one should not expect any of the Lyapunov exponents to be 0. Included in (2.4) is also the Minkowski group (i.e., the *relativity* group), where $H = D$ with

$$D = \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix} \tag{2.6}$$

and $n_1 + n_2 = n$. The particular case $n_1 = 3$, $n_2 = 1$ is the Lorentz group.

Before proceeding, let us simplify the problem. Let U be an orthogonal matrix giving the real Schur form of H, grouping the eigenvalues of H on the unit circle as follows:

$$K := U^T H U = \begin{bmatrix} D & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & J \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} I_{n_1} & 0 \\ 0 & -I_{n_2} \end{bmatrix} & 0 & 0 \\ 0 & C & 0 \\ 0 & 0 & \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix} \end{bmatrix}, \tag{2.7}$$

where C comprises the eigenvalues of H that are different from ± 1 and $\pm i$:

$$\begin{aligned} &C = \operatorname{diag}(C_1, \ldots, C_p), \;\; C_j = \begin{bmatrix} Q_j \otimes I_{n_1(j)} & 0 \\ 0 & -Q_j \otimes I_{n_2(j)} \end{bmatrix}, \\ &Q_j = \begin{bmatrix} c_j & s_j \\ -s_j & c_j \end{bmatrix}, \;\; c_j^2 + s_j^2 = 1,\; c_j \neq 0,\; s_j \neq 0,\; j = 1, \ldots, p. \end{aligned} \tag{2.8}$$

In other words, we have blocked the eigenvalues of H by grouping together the eigenvalues $\cos(\phi_j) \pm i \sin(\phi_j)$ and those out of phase by π: $\cos(\phi_j + \pi) \pm i \sin(\phi_j + \pi)$, and we have ordered them so that the angles are increasing from 0 to $\pi/2$; see Figure 2.1. Naturally, for every complex conjugate pair of eigenvalues, $e^{\pm i\phi}$, there need not be a complex conjugate pair out of phase by π with it or vice versa. That is, in (2.8), $n_1(j)$ or $n_2(j)$ may be 0.

Now, if X is a fundamental matrix solution of (1.2) satisfying (2.4), then the matrix function $R = U^T X U$ satisfies

$$R^T(t) K R(t) \;=\; K \quad \text{for all } t \tag{2.9}$$

with K as in (2.7). Since the singular values of X and R are the same, we can assume to have the simplified form of orthogonal matrices as in (2.7). In this case, we can simplify the form of R satisfying (2.9).

LEMMA 2.3. *Let $R \in \mathbb{R}^{n\times n}$ be any matrix satisfying $R^T K R = K$, with K given in* (2.7). *Then, R has the block structure*

$$R = \begin{bmatrix} W & 0 & 0 \\ 0 & Z & 0 \\ 0 & 0 & S \end{bmatrix}, \tag{2.10}$$

where the partitioning is that inherited by the form of K.

Proof. Write R in block form

$$R = \begin{bmatrix} R_{11} & \begin{bmatrix} R_{12} & R_{13} \end{bmatrix} \\ \begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} & \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix} \end{bmatrix}.$$

Now, use the relations $R^T K R = K$ and $R^T K^T R = K^T$. In particular, from the respective $(2,2)$ blocks, we have

$$\begin{bmatrix} R_{12}^T \\ R_{13}^T \end{bmatrix} D \begin{bmatrix} R_{12} & R_{13} \end{bmatrix} + \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix} = \begin{bmatrix} C & 0 \\ 0 & J \end{bmatrix},$$
$$\begin{bmatrix} R_{12}^T \\ R_{13}^T \end{bmatrix} D \begin{bmatrix} R_{12} & R_{13} \end{bmatrix} + \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} C^T & 0 \\ 0 & -J \end{bmatrix} \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix} = \begin{bmatrix} C^T & 0 \\ 0 & -J \end{bmatrix},$$

from which

$$\begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} (C-C^T)/2 & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix} = \begin{bmatrix} (C-C^T)/2 & 0 \\ 0 & J \end{bmatrix}.$$

Therefore, $\left[\begin{smallmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{smallmatrix}\right]$ must be invertible. Now, from the $(2,1)$ blocks, we get

$$\begin{bmatrix} R_{12}^T \\ R_{13}^T \end{bmatrix} D R_{11} + \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$
$$\begin{bmatrix} R_{12}^T \\ R_{13}^T \end{bmatrix} D R_{11} + \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} C^T & 0 \\ 0 & -J \end{bmatrix} \begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{2.11}$$

from which it follows that

$$\begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} (C-C^T)/2 & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} R_{21} \\ R_{31} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and hence $R_{21} = 0$ and $R_{31} = 0$. At this point, the relation for the $(1,1)$ block gives $R_{11}^T D R_{11} = D$, from which it follows that R_{11} must be invertible. Writing out the relations for the $(1,2)$ blocks, in a way similar to the above, it follows that $R_{12} = 0$ and $R_{13} = 0$. With this, adding the two relations satisfied by the $(2,2)$ blocks, one gets

$$\begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix}^T \begin{bmatrix} (C+C^T)/2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{bmatrix} = \begin{bmatrix} (C+C^T)/2 & 0 \\ 0 & 0 \end{bmatrix}.$$

From this, it follows that R_{22} is invertible and $R_{23} = 0$, $R_{32} = 0$, and hence necessarily that R_{33} is invertible. ∎

Because of Lemma 2.3, we can restrict our attention to simpler cases of fundamental matrix solutions W, Z, and S, where for all t,

$$W^T(t)DW(t) = D, \quad D \text{ in (2.6)}, \tag{2.12}$$

$$Z^T(t)CZ(t) = C, \quad C \text{ in (2.8)}, \tag{2.13}$$

$$S^T(t)JS(t) = J, \quad J \text{ in (2.5)}. \tag{2.14}$$

A goal of ours is to give lower bounds on the number of singular values of fundamental matrix solutions X satisfying (2.4) that are 1 for all t. Our arguments will use the difference in multiplicities of the eigenvalues of H which are out of phase by π with one another. If this difference is 0, the lower bound is 0. For this reason, we will focus attention on (2.12) and (2.13) only, that is, on the W-part and Z-part of the system.

LEMMA 2.4. *Let $W \in \mathbb{R}^{n\times n}$ be any matrix satisfying $W^TDW = D$, with D given in* (2.6)*, and $n = n_1 + n_2$. Let $\nu_0(W)$ be the number of singular values of W which are equal to* 1*. Then, we have*

$$\nu_0(W) \geq |n_1 - n_2|.$$

Proof. Let W be partitioned similarly to D, that is $W = \left[\begin{smallmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{smallmatrix}\right]$, where $W_{ii} \in \mathbb{R}^{n_i\times n_i}$, $i = 1, 2$, and $W_{12} \in \mathbb{R}^{n_1\times n_2}$, $W_{21} \in \mathbb{R}^{n_2\times n_1}$. Let $Y = \left[\begin{smallmatrix} 0 & W_{12} \\ W_{21} & 0 \end{smallmatrix}\right]$. Then, since Y has at most $2\min(n_1, n_2)$ linearly independent columns, we see that $\dim\big(\ker(Y)\big) \geq n - 2\min(n_1, n_2) = |n_1 - n_2|$. Next, observe that, since $W^TDW = D$ and $D^2 = I$, we have

$$\begin{aligned} \ker(W^TW - I) &= \ker(W^TW - DW^TDW) = \ker\big((W^T - DW^TD)W\big) \\ &= \ker\left(\begin{bmatrix} 0 & 2W_{21}^T \\ 2W_{12}^T & 0 \end{bmatrix} W\right). \end{aligned}$$

Thus, we have $\dim\big(\ker(W^TW - I)\big) \geq |n_1 - n_2|$. ∎

Next, we show some results concerning the Z-part of the system.

LEMMA 2.5. *Let $Z \in \mathbb{R}^{n\times n}$ be any matrix satisfying Z^TCZ, where C is given in* (2.8) *with $n = 2\sum_{j=1}^{p}[n_1(j) + n_2(j)]$. Then, Z is a block diagonal matrix*

$$Z = \operatorname{diag}(Z_1, \ldots, Z_p), \quad Z_j^TC_jZ_j = C_j, \quad j = 1, \ldots, p. \tag{2.15}$$

Moreover, for $j = 1, \ldots, p$, Z_j satisfy

$$Z_j^TD_jZ_j = D_j, \quad D_j = \begin{bmatrix} I_2\otimes I_{n_1(j)} & 0 \\ 0 & -I_2\otimes I_{n_2(j)} \end{bmatrix}, \tag{2.16}$$

and

$$Z_j^T\hat{J}_jZ_j = \hat{J}_j,\ \hat{J}_j = \begin{bmatrix} J_{n_1(j)} & 0 \\ 0 & -J_{n_2(j)} \end{bmatrix},\ J_{n_k(j)} = \begin{bmatrix} 0 & I_{n_k(j)} \\ -I_{n_k(j)} & 0 \end{bmatrix},\ k = 1, 2. \tag{2.17}$$

Proof. Since $Z^TCZ = C$ and $ZCZ^T = C$, one also has $Z^TC^TZ = C^T$ and $ZC^TZ^T = C^T$. Adding these relations pairwise, we obtain

$$Z^TNZ = N, \quad ZNZ^T = N, \quad \text{where} \quad N = (C + C^T)/2. \tag{2.18}$$

Given the form of C in (2.8), the matrix N has the form

$$N \;=\; \mathrm{diag}(c_1 D_1, \ldots, c_p D_p),\; D_j = \begin{bmatrix} I_2 \otimes I_{n_1(j)} & 0 \\ 0 & -I_2 \otimes I_{n_2(j)} \end{bmatrix},\; j = 1, \ldots, p,$$

and $c_j = \cos(\phi_j)$, $0 < \phi_1 < \cdots < \phi_p < \pi/2$. Now, from (2.18), one has

$$\text{(a)} \quad Z^{-T} = NZN^{-1} \qquad \text{and} \qquad \text{(b)} \quad Z^{-T} = N^{-1}ZN.$$

Write Z in block form, and equate the (i,j)th blocks of (a) and (b):

$$\frac{c_i}{c_j} D_i Z_{ij} D_j \;=\; \frac{c_j}{c_i} D_i Z_{ij} D_j;$$

thus, we must have $(c_i^2 - c_j^2) Z_{ij} = 0$. For $i \neq j$, this implies $Z_{ij} = 0$. Hence, Z must be block diagonal, and (2.15) holds. The form (2.16) is obtained at once from $Z_j^T (C_j + C_j^T) Z_j = (C_j + C_j^T)$, while (2.17) is obtained from $Z_j^T (C_j - C_j^T) Z_j = (C_j - C_j^T)$.[3] ∎

LEMMA 2.6. *With the notation of* Lemma 2.5*, we have*

$$\nu_0(Z) = \sum_{j=1}^{p} \nu_0(Z_j),$$

where $\nu_0(Z)$ and $\nu_0(Z_j)$ denote the number of singular values of Z and Z_j that are 1*, and where*

$$\nu_0(Z_j) \geq 2|n_1(j) - n_2(j)|,\; j = 1, \ldots, p\ .$$

Further, the singular values of each Z_j have even multiplicity.

Proof. The statement on $\nu_0(Z) = \sum_{j=1}^{p} \nu_0(Z_j)$ is clear from (2.15). The fact that $\nu_0(Z_j) \geq 2|n_1(j) - n_2(j)|$ is now a consequence of Lemma 2.4 and (2.16).

Now, for given j, suppose that $Z_j^T Z_j x = \frac{1}{\lambda} x$, $\|x\| = 1$. Then, we have at once

$$D_j Z_j^T Z_j x \;=\; \frac{1}{\lambda} D_j x \qquad \text{and} \qquad \hat{J}_j Z_j^T Z_j x \;=\; \frac{1}{\lambda} \hat{J}_j x.$$

Now, since $Z_j^T D_j Z_j = D_j$ and $Z_j^T \hat{J}_j Z_j = \hat{J}_j$, one also has $Z_j D_j Z_j^T = D_j$ and $Z_j \hat{J}_j Z_j^T = \hat{J}_j$. Thus, we get

$$(Z_j^T Z_j)^{-1}(D_j x) = \frac{1}{\lambda}(D_j x) \qquad \text{and} \qquad (Z_j^T Z_j)^{-1}(\hat{J}_j x) = \frac{1}{\lambda}(\hat{J}_j x).$$

Therefore, since the eigenvalues of $Z_j^T Z_j$ arise as $\{\lambda, 1/\lambda\}$, and the eigenvectors $D_j x$ and $\hat{J}_j x$ are orthogonal unit vectors, we conclude that each eigenvalue of $Z_j^T Z_j$ has even multiplicity. ∎

Remark 2.7. Suppose that W is a fundamental matrix solution of (1.2) satisfying (2.12) for all t. Since the eigenvalues of the continuous function $W^T W$ can be labeled as continuous functions of t, we can label the singular values of W in such a way that they are continuous functions of t and at least $|n_1 - n_2|$ of them are identically 1 for all t. Likewise, let Z be a fundamental matrix solution of (1.2) satisfying (2.13) for all t. Since the eigenvalues of the functions $Z_j^T Z_j$, $j = 1, \ldots, p$, can be labeled as

[3] Equations (2.16) and (2.17) are equivalent to $Z_j^T C_j Z_j = C_j$.

continuous functions of t, the singular values of Z can be labeled so that they are continuous functions of t, at least $2\sum_{j=1}^{p}|n_1(j)-n_2(j)|$ are identically 1 for all t, and they have even multiplicity for any t.

Remark 2.8. As far as the S-part of the differential system is concerned, that is, when S satisfies (2.14), a priori we cannot be certain that any of its singular values will be identically 1 or that they will possess even multiplicity.

Finally, let $\nu_0(X)$ be the number of singular values of a fundamental matrix solution X satisfying (2.4) which are identically 1 for all t. By putting together the results obtained in this section, we see that a lower bound on $\nu_0(X)$ can be obtained by looking at the distribution of eigenvalues of H on the unit circle. In the case in which the assumptions leading to (1.6) hold, this will give us a lower bound showing how many Lyapunov exponents will be 0. We summarize these considerations in the following theorem, which holds as a consequence of the previous results.

THEOREM 2.9. *Let X be a fundamental matrix solution of* (1.2) *satisfying* (2.4). *Let orthogonal U give the ordered Schur form of H as in* (2.7) *and* (2.8), *with* $n = n_1+n_2+2m+2\sum_{j=1}^{p}[n_1(j)+n_2(j)]$. *With the understanding that some of the indices below may be* 0, *H has*

(1) *n_1 eigenvalues equal to* 1, *and n_2 eigenvalues equal to* -1;
(2) *$2n_1(j)$ eigenvalues equal to $e^{\pm i\phi_j}$, and $2n_2(j)$ eigenvalues equal to $e^{\pm i(\phi_j+\pi)}$, for $j = 1,\dots,p$, and $0 < \phi_1 < \cdots < \phi_p < \pi/2$.*
(3) *$2m$ eigenvalues equal to $\pm i$.*

Then, for $\nu_0(X)$, we have

$$\nu_0(X) \geq |n_1 - n_2| + 2\sum_{j=1}^{p}|n_1(j) - n_2(j)|. \tag{2.19}$$

Moreover, consider the subproblem associated to the eigenvalues $e^{\pm i\phi_j}$, $e^{\pm i(\phi_j+\pi)}$ of (2)*; that is, consider Z in* (2.15). *Then, X has at least as many nonsimple singular values as Z.*

Finally, under the assumptions and with the notation of Theorem 1.1, *for $x_0 \in M_0$, $\mathrm{Sp}(\Phi_{x_0})$ is symmetric with respect to the origin and has at least $\big[|n_1 - n_2| + 2\sum_{j=1}^{p}|n_1(j) - n_2(j)|\big]$ Lyapunov exponents equal to* 0. *Also, $\mathrm{Sp}(\Phi_{x_0})$ contains at least as many repeated Lyapunov exponents as the number of distinct singular values of the Z-part of $U^T\Phi_{x_0}U$, all of which have even multiplicity.*

Proof. The only things to justify are the statements about $\mathrm{Sp}(\Phi_{x_0})$. With previous notation, for all t we must have (see (2.10), (2.12), (2.13), (2.14))

$$U^T\Phi_{x_0}(t)U = \begin{bmatrix} W_{x_0}(t) & 0 & 0 \\ 0 & Z_{x_0}(t) & 0 \\ 0 & 0 & S_{x_0}(t) \end{bmatrix},$$

so that, in particular,

$$\begin{aligned} &U^T \lim_{t\to\infty}\frac{1}{t}\log\big(\Phi_{x_0}^T(t)\Phi_{x_0}(t)\big)^{1/2}\,U \\ &= \lim_{t\to\infty}\frac{1}{t}\begin{bmatrix} \log\big(W_{x_0}^T(t)W_{x_0}(t)\big)^{1/2} & 0 & 0 \\ 0 & \log\big(Z_{x_0}^T(t)Z_{x_0}(t)\big)^{1/2} & 0 \\ 0 & 0 & \log\big(S_{x_0}^T(t)S_{x_0}(t)\big)^{1/2} \end{bmatrix}. \end{aligned}$$

Thus, the symmetry with respect to the origin and the bound on the number of 0 Lyapunov exponents is a consequence of the fact that the singular values of the function Φ_{x_0} can be chosen as continuous functions of t and of previous results. The statement on the multiplicity relative to the Lyapunov exponents associated to Z_{x_0} is also a consequence of continuity of the singular values and of the fact that the limit matrix above is symmetric and hence diagonalizable. In fact, if we let $\Psi_{x_0} = \lim_{t\to\infty} \frac{1}{t} \log(Z_{x_0}^T(t) Z_{x_0}(t))^{1/2}$, then continuity of the eigenvalues of $Z_{x_0}^T(t) Z_{x_0}(t)$ precludes having any of the eigenvalues of Ψ_{x_0} with odd multiplicity. □

Remark 2.10. As we remarked in point (i) of Theorem 1.1, in general $\mathrm{Sp}(\Phi_{x_0})$ depends on $x_0 \in M_0$ and on the invariant measure μ (as M_0 does). The lower bounds given in Theorem 2.9, instead, hold for all x_0 (and μ). The situation is similar to (2.3) in Theorem 2.1, whereby the symmetry of the Lyapunov spectrum with respect to the origin holds regardless of x_0. In order to further infer that $\mathrm{Sp}(\Phi_{x_0})$ does not depend on $x_0 \in M_0$, we would need condition (iv) in Theorem 1.1 to hold.

Remark 2.11. An extension of our results (cf. [7, 11]) is obtained by replacing (2.4) with

$$X^T(t) H X(t) = e^{at} H \text{ for all } t,\ H^T H = I. \tag{2.20}$$

It is a simple verification that one arrives at (2.20) upon considering the shifted system $\dot{x} = (A(t) + a/2\ I)x$ instead of (1.2). In this case, one has $A^T(t)H + HA(t) = aH$ instead of (2.2)(a). Now $\mathrm{Sp}(\Phi_{x_0})$ will be shifted by $a/2$.

3. Examples. The numerical results below have been obtained using the so-called *continuous QR method* (see [6]). That is, we use the technique leading to (1.5) as follows:

- Q is approximated by the classic Runge–Kutta scheme of order 4 to integrate the equation for Q, and the solution is orthogonalized after each step;
- the Lyapunov exponents are approximated from (1.5) using the composite trapezoidal rule.

For the problems below, we fix the interval of integration to $[0, 10^4]$, take initial condition to the identity, and perform integration with a constant step size $h = 1/10$. These examples are purposely built starting from a periodic coefficient matrix, to which we add a term which goes to 0 as $t \to \infty$, so that $\mathrm{Sp}(\Phi)$ reduces to the set of Floquet exponents of the periodic problem. On one hand, this allows us to compute $\mathrm{Sp}(\Phi)$ by other means and to check the accuracy of the obtained answers. On the other hand, we remark that when we attempted a direct time integration for the full monodromy matrix on these problems, we obtained very inaccurate approximations of the Floquet exponents (only the largest one was accurate).

Example 3.1. This is a system evolving on the Lorentz group. We have

$$A(t) = \begin{bmatrix} 0 & \cos(t) & -1 & \frac{1}{1+t} \\ -\cos(t) & 0 & \frac{3}{1+t^2} & 5 \\ 1 & -\frac{3}{1+t^2} & 0 & -\sin(t) \\ \frac{1}{1+t} & 5 & -\sin(t) & 0 \end{bmatrix}, \qquad t \geq 0.$$

The Lyapunov exponents are $\{5, 0, 0, -5\}$. Approximating all four Lyapunov exponents, we obtain (at six digits)

$$\lambda_1 = 4.99959,\ \lambda_2 = 0.000353,\ \lambda_3 = 0.00000277230,\ \lambda_4 = -4.99994.$$

Approximating only the dominant Lyapunov exponent, by integrating just for the first column of Q, we get $\lambda_1 = 4.99958$. This second computation takes 20% of the time required by the first one.

Example 3.2. Here we consider a problem whose fundamental matrix solution $Z(t)$ satisfies $Z^T(t)CZ(t) = C$ with $C = \begin{bmatrix} Q\otimes I_2 & 0 \\ 0 & -Q \end{bmatrix}$, $Q = \left[\begin{smallmatrix} c & s \\ -s & c \end{smallmatrix}\right]$, and $c = \cos(\phi)$, $s = \sin(\phi)$, $0 < \phi < \frac{\pi}{2}$. We take the following coefficient matrix:

$$A(t) = \begin{bmatrix} 0 & 2 & -1 & \frac{1}{1+t} & 1 & 2 \\ -2 & 0 & \frac{1}{1+t} & 5 & \cos(t) & 4 \\ 1 & -\frac{1}{1+t} & 0 & 2 & -2 & 1 \\ -\frac{1}{1+t} & -5 & -2 & 0 & -4 & \cos(t) \\ 1 & \cos(t) & -2 & -4 & 0 & \sin(t) \\ 2 & 4 & 1 & \cos(t) & -\sin(t) & 0 \end{bmatrix}, \qquad t \geq 0.$$

We expect two zero and two possibly nonzero Lyapunov exponents, symmetric with respect to the origin, each of multiplicity 2. In other words, only one Lyapunov exponent really needs to be computed. In fact, the two nonzero Lyapunov exponents for this problem are (at four digits) $\{\pm 3.027\}$. Approximating all six Lyapunov exponents, we get

$$\lambda_1 = 3.028,\ \lambda_2 = 3.028,\ \lambda_3 = 0.5347 \times 10^{-4},$$

$$\lambda_4 = 0.4374 \times 10^{-3},\ \lambda_5 = -3.028,\ \lambda_6 = -3.028.$$

Directly approximating only the dominant Lyapunov exponent, we get $\lambda_1 = 3.027$, and this second computation takes 12.5% of the time required to approximate all Lyapunov exponents.

Acknowledgments. This work was prompted by a visit of the first author to the Department of Mathematics, University of Bari. Both authors are greatly indebted to the referees for improvements to the original version of this work.

REFERENCES

[1] L. Ya. Adrianova, *Introduction to Linear Systems of Differential Equations*, Translations of Mathematical Monographs 146, AMS, Providence, RI, 1995.
[2] J. L. Anderson, *Principles of Relativity Physics*, Academic Press, New York, 1967.
[3] L. Arnold and V. Wihstutz, eds., *Lyapunov Exponents. Proceedings of a Workshop Held in Bremen*, 1984, Lecture Notes in Math. 1186, Springer-Verlag, Berlin, 1986.
[4] L. Arnold, H. Crauel, and J. P. Eckmann, eds., *Lyapunov Exponents. Proceedings of the Second Conference Held in Oberwolfach, May* 28*–June* 2, 1990, Lecture Notes in Math. 1486, Springer-Verlag, Berlin, 1991.
[5] G. Benettin, G. Galgani, L. Giorgilli, and J. M. Strelcyn, *Lyapunov exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part* I: *Theory; Part* II: *Numerical applications*, Meccanica, 15 (1980), pp. 21–30.
[6] L. Dieci, R. D. Russell, and E. S. Van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal., 34 (1997), pp. 402–423.
[7] U. Dressler, *Symmetry property of the Lyapunov spectra of a class of dissipative dynamical systems with viscous damping*, Phys. Rev. A, 38 (1988), pp. 2103–2109.
[8] J. P. Eckmann and D. Ruelle, *Ergodic theory of chaos and strange attractors*, Rev. Modern Phys., 57 (1985), pp. 617–656.
[9] K. Geist, U. Parlitz, and W. Lauterborn, *Comparison of different methods for computing Lyapunov exponents*, Prog. Theoret. Phys., 83 (1990), pp. 875–893.
[10] J. M. Greene and J.-S. Kim, *The calculation of Lyapunov spectra*, Phys. D, 24 (1987), pp. 213–225.
[11] D. Gupalo, A. S. Kaganovich, and E. G. D. Cohen, *Symmetry of Lyapunov spectrum*, J. Statist. Phys., 74 (1994), pp. 1145–1159.

[12] R. A. Johnson, K. J. Palmer, and G. R. Sell, *Ergodic properties of linear dynamical systems*, SIAM J. Math. Anal., 18 (1987), pp. 1–33.
[13] B. Leimkuhler and E. Van Vleck, *Orthosymplectic integration of linear Hamiltonian systems*, Numer. Math., 77 (1997), pp. 269–282.
[14] V. I. Oseledec, *A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc., 19 (1968), p. 197.
[15] E. Ott, T. Sauer, and J. Yorke, *Coping with Chaos*, Wiley, New York, 1994.

SIAM J. MATRIX ANAL. APPL.
Vol. 24, No. 4, pp. 1186–1196

A UNIFIED THEORY OF CONDITIONING FOR LINEAR LEAST SQUARES AND TIKHONOV REGULARIZATION SOLUTIONS*

A. N. MALYSHEV†

Abstract. We develop a unified perturbation theory for the unconstrained linear least squares problem, least squares with linear equality constraints, and least squares with quadratic inequality constraint and Tikhonov regularization solution. The computable condition numbers are exact with respect to the Frobenius norm. For the 2-norm, the computed bounds may differ from the exact condition numbers only by a factor less than $\sqrt{2}$.

Key words. condition number, constrained least squares, Tikhonov regularization

AMS subject classification. 65F35

PII. S0895479801389564

1. Introduction. The theory of conditioning (perturbation or sensitivity analysis) for least squares problems is rather complicated and "not especially easy to read" [11]. Various constraints require different analytical considerations, and the perturbation analysis for each least squares problem is usually long and cumbersome. However, our study was motivated by the observation that most results scattered throughout the literature present only upper bounds on exact condition numbers. Thus, the authors of [11] remark that [4] and [5] "... give exact condition numbers [for the unconstrained linear least squares problem] with respect to the Frobenius norm. For the 2-norm, ... as far as we are aware, exact results are not known."

In the present paper we assess the conditioning theory for several least squares problems and derive computable formulas for their exact condition numbers, thus filling the gap mentioned in the above remark. Our approach also covers Tikhonov regularization solutions.

As should be expected, our algebraic equations for infinitely small variations are essentially similar to those in previous publications [1, 3, 4, 5, 6, 7, 8, 9]. But, in contrast to these earlier studies, we derive final perturbation bounds more accurately. This became possible after extraction of a principal ingredient common to a range of least squares problems. This ingredient is the optimization problem

$$\sup_{\delta A} \frac{\|X(\delta A)x + Y(\delta A)^T y\|_2}{\|\delta A\|_F}, \tag{1.1}$$

where δA, X, Y are matrices and x, y are vectors. Problem (1.1) turns out to have a computable exact solution, which is an "almost exact" solution to the problem

$$\sup_{\delta A} \frac{\|X(\delta A)x + Y(\delta A)^T y\|_2}{\|\delta A\|_2}. \tag{1.2}$$

More precisely, the solution of (1.1) may differ from the solution of (1.2) by a factor less than $\sqrt{2}$ because one of the optimal δA is a matrix of rank 2. It is the introduction

*Received by the editors May 21, 2001; accepted for publication (in revised form) by P. Hansen December 10, 2002; published electronically May 6, 2003.
http://www.siam.org/journals/simax/24-4/38956.html
†Department of Informatics, University of Bergen, PB. 7800, N-5020 Bergen, Norway (sasha@ii.uib.no).

of (1.1) and its exact solution into the conditioning analysis that represents the main contribution of this paper.

Let us consider a standard least squares problem, the unconstrained linear least squares problem

$$\min_{x\in R^n} \|Ax-b\|_2, \tag{1.3}$$

where $A \in R^{m\times n}$ is of full column rank and $b \in R^m$. Its unique solution is A^+b, where A^+ denotes the Moore–Penrose pseudoinverse of A. We also assume that $x+\delta x$ is a unique solution to the perturbed problem

$$\min_{y\in R^n} \|(A+\delta A)y-(b+\delta b)\|_2,$$

where δA and δb are infinitesimal perturbations of A and b. Like the authors of [11], we are interested in the following condition numbers, which were introduced systematically in [10]:

$$\kappa_{b\mapsto x} = \lim_{\delta\to 0} \sup_{\|\delta b\|_2\le\delta} \left(\frac{\|\delta x\|_2}{\|x\|_2} \Big/ \frac{\|\delta b\|_2}{\|b\|_2} \right), \quad \kappa_{A\mapsto x} = \lim_{\delta\to 0} \sup_{\|\delta A\|_2\le\delta} \left(\frac{\|\delta x\|_2}{\|x\|_2} \Big/ \frac{\|\delta A\|_2}{\|A\|_2} \right).$$

To write out δx in terms of δb and δA, subtract the normal equations $A^TAx = A^Tb$ from their perturbed version $(A+\delta A)^T(A+\delta A)(x+\delta x) = (A+\delta A)^T(b+\delta b)$ and drop second order terms. Then transform the result $A^TA\delta x + A^T(\delta A)x + (\delta A)^TAx = A^T\delta b + (\delta A)^Tb$ into

$$\begin{aligned} \delta x &= (A^TA)^{-1}\left[A^T\delta b - A^T(\delta A)x + (\delta A)^T(b-Ax)\right] \\ &= A^+\delta b - A^+(\delta A)x + A^+(A^+)^T(\delta A)^Tr, \end{aligned} \tag{1.4}$$

where $r = b-Ax$ denotes the residual. Here most authors conclude the analysis with the straightforward estimate

$$\frac{\|\delta x\|_2}{\|x\|_2} \le \frac{\|A^+\|_2\|b\|_2}{\|x\|_2}\frac{\|\delta b\|_2}{\|b\|_2} + \|A\|_2\left(\|A^+\|_2 + \|A^+\|_2^2\frac{\|r\|_2}{\|x\|_2}\right)\frac{\|\delta A\|_2}{\|A\|_2}, \tag{1.5}$$

which yields the upper bounds

$$\kappa_{b\mapsto x} \le \|A^+\|_2\frac{\|b\|_2}{\|x\|_2}, \tag{1.6}$$

$$\kappa_{A\mapsto x} \le \|A\|_2\|A^+\|_2\left(1+\|A^+\|_2\frac{\|r\|_2}{\|x\|_2}\right). \tag{1.7}$$

These bounds are commonly accepted as condition numbers, and any discussion about their sharpness is usually avoided.

The value of $\kappa_{b\mapsto x}$ coincides with the upper bound in (1.6), which is a typical situation with perturbations in b. In what follows we often discuss the conditioning with respect to perturbations in A only.

A computable formula for $\kappa_{A\mapsto x}$ does not seem to exist. Fortunately, it was discovered that the parameter

$$\kappa^F_{A\mapsto x} = \lim_{\delta\to 0} \sup_{\|\delta A\|_F\le\delta} \left(\frac{\|\delta x\|_2}{\|x\|_2} \Big/ \frac{\|\delta A\|_F}{\|A\|_2} \right)$$

was better for computation; namely, the following formula holds:

$$\kappa^F_{A\mapsto x} = \|A\|_2\|A^+\|_2\sqrt{1+\left(\|A^+\|_2\frac{\|r\|_2}{\|x\|_2}\right)^2}. \tag{1.8}$$

This formula probably first appeared in [4] and then was slightly generalized in [5]. Since $\kappa^F_{A\mapsto x} \le \kappa_{A\mapsto x}$, the upper bound in (1.7) is a tight approximation to $\kappa_{A\mapsto x}$ and can be used as a strict condition number.

While the techniques from [4, 5] are essentially restricted to the unconstrained linear least squares problem, our unified approach allows one to derive computable expressions of $\kappa^F_{A\mapsto x}$ for a range of least squares problems that includes the unconstrained problem. To demonstrate the latter, let us return to the algebraic equation (1.4) and rewrite it as a sum of linear operators $\delta x = l_b(\delta b) + l(\delta A)$, where $l_b(v) = A^+v$ and $l(V) = -A^+Vx + A^+(A^+)^TV^Tr$. Then the condition number $\kappa_{b\mapsto x}$ simply equals $\|l_b\|_2\|b\|_2/\|x\|_2 = \|A^+\|_2\|b\|_2/\|x\|_2$.

The condition number $\kappa^F_{A\mapsto x}$ is equal to $\sup_V(\|l(V)\|_2/\|V\|_F)\|A\|_2/\|x\|_2$ by definition. By formula (2.3) with $X = -A^+$, $Y = A^+(A^+)^T$, and $y = r$ we obtain

$$\begin{aligned}\sup_V\frac{\|l(V)\|_2}{\|V\|_F} &= \left\|\left[-\|x\|_2A^+,\ \ \|r\|_2A^+(A^+)^T\right]\begin{pmatrix} I-\frac{rr^T}{\|r\|_2^2} & \frac{rx^T}{\|r\|_2\|x\|_2}\\ 0 & I\end{pmatrix}\right\|_2\\ &= \left\|\left[-\|x\|_2A^+,\ \ \|r\|_2A^+(A^+)^T\right]\right\|_2 \quad \text{(because } A^+r=0\text{)}\\ &= \|x\|_2\|A^+\|_2\sqrt{1+\left(\|A^+\|_2\frac{\|r\|_2}{\|x\|_2}\right)^2}.\end{aligned}$$

Hence formula (1.8) is proved.

2. A general framework for the study of the least squares sensitivity to matrix perturbations. Given matrices $V = [V_1|V_2|\cdots|V_n] \in R^{m\times n}$, $X \in R^{n\times m}$, $Y = [Y_1|Y_2|\cdots|Y_n] \in R^{n\times n}$ and vectors $x \in R^n$, $y \in R^m$, we introduce a linear operator

$$l(V) = XVx + YV^Ty. \tag{2.1}$$

The operator $l(V)$ admits the matrix representation

$$l(V) = \sum_{i=1}^n\left(XV_ix_i + Y_iV_i^Ty\right) = \sum_{i=1}^n(x_iXV_i + Y_iy^TV_i) = (x^T\otimes X + Y\otimes y^T)\begin{pmatrix}V_1\\ \vdots\\ V_n\end{pmatrix},$$

where $\otimes$ denotes the Kronecker product of matrices. The matrix $L = x^T\otimes X + Y\otimes y^T$ satisfies the identities

$$\begin{aligned}LL^T &= \sum_{i=1}^n\left(x_iX + Y_iy^T\right)\left(X^Tx_i + yY_i^T\right)\\ &= \|x\|_2^2XX^T + \|y\|_2^2YY^T + (Xy)(Yx)^T + (Yx)(Xy)^T\\ &= (\|x\|_2X,\ \|y\|_2Y)\begin{pmatrix} I & \frac{yx^T}{\|y\|_2\|x\|_2}\\ \frac{xy^T}{\|x\|_2\|y\|_2} & I\end{pmatrix}\begin{pmatrix}\|x\|_2X^T\\ \|y\|_2Y^T\end{pmatrix}\\ &= (\|x\|_2X,\ \|y\|_2Y)\begin{pmatrix} I-\frac{yy^T}{\|y\|_2^2} & \frac{yx^T}{\|y\|_2\|x\|_2}\\ 0 & I\end{pmatrix}\begin{pmatrix} I-\frac{yy^T}{\|y\|_2^2} & 0\\ \frac{xy^T}{\|x\|_2\|y\|_2} & I\end{pmatrix}\begin{pmatrix}\|x\|_2X^T\\ \|y\|_2Y^T\end{pmatrix}.\end{aligned}$$

Since $\sup_V \|l(V)\|_2/\|V\|_F = \|L\|_2$, we have

$$\sup_V \frac{\|l(V)\|_2}{\|V\|_F} = \left\| \|x\|_2^2 XX^T + \|y\|_2^2 YY^T + (Xy)(Yx)^T + (Yx)(Xy)^T \right\|_2^{1/2} \tag{2.2}$$

or

$$\sup_V \frac{\|l(V)\|_2}{\|V\|_F} = \left\| [\|x\|_2 X, \ \|y\|_2 Y] \begin{pmatrix} I - \frac{yy^T}{\|y\|_2^2} & \frac{yx^T}{\|y\|_2\|x\|_2} \\ 0 & I \end{pmatrix} \right\|_2 . \tag{2.3}$$

LEMMA 2.1. *We have*

$$\sup_V \frac{\|l(V)\|_2}{\|V\|_F} \le \sup_V \frac{\|l(V)\|_2}{\|V\|_2} \le \sqrt{2} \sup_V \frac{\|l(V)\|_2}{\|V\|_F}. \tag{2.4}$$

Proof. The lower bound is trivial because $\|V\|_2 \le \|V\|_F$. A similar converse estimate $\|V\|_F \le \sqrt{\min(m,n)}\|V\|_2$ is too rough. Suppose that $\sup_V \frac{\|l(V)\|_2}{\|V\|_2}$ is attained at a matrix V_{opt}. To reveal the structure of V_{opt}, let us choose orthonormal bases in R^n and R^m in which $x = \|x\|_2 e_1$ and $y = \|y\|_2 e_1$. Denoting the first column of V in these bases by v and the first row of V by u^T, we arrive at the identity $l(V) = \|x\|_2 Xv + \|y\|_2 Yu$. Thus, V_{opt} minimizes the 2-norm on the set of matrices with first column v and first row u^T. Such a problem in a more general setting was studied in [2]. Our case is easily treated, assuming that the vectors v and u have all zero components except the first two, i.e., $v = (\alpha_{11}, \alpha_{21}, 0, \ldots, 0)^T$ and $u = (\alpha_{11}, \alpha_{12}, 0, \ldots, 0)^T$. This assumption is trivially satisfied with the help of suitable orthogonal transformations in R^n and R^m, which do not touch the first components. Since the matrix V_{opt} is not unique, we fix a special construction of a matrix with minimum 2-norm for given v and u. The upper left corner of this matrix is the 2×2 matrix $\left(\begin{smallmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{smallmatrix}\right)$, while the rest of V_{opt} consists of zero elements. If $|\alpha_{21}| \ge |\alpha_{12}|$, we put $\alpha_{22} = -\alpha_{11}\alpha_{12}/\alpha_{21}$, else $\alpha_{22} = -\alpha_{11}\alpha_{21}/\alpha_{12}$. The choice of α_{22} guarantees the identity

$$\left\| \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \right\|_2 = \max \left\{ \left\| \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \end{bmatrix} \right\|_2, \left\| \begin{bmatrix} \alpha_{11} & \alpha_{12} \end{bmatrix} \right\|_2 \right\},$$

which proves the correctness of our construction. Since

$$\left\| \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \right\|_F \le \sqrt{2} \left\| \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \right\|_2 ,$$

V_{opt} with the above defined structure satisfies the inequality $\|V_{opt}\|_F \le \sqrt{2}\|V_{opt}\|_2$, whence $\sqrt{2}\frac{\|l(V_{opt})\|_2}{\|V_{opt}\|_F} \ge \sup_V \frac{\|l(V)\|_2}{\|V\|_2}$. ∎

As a result, the tight bounds $\kappa^F_{A\mapsto x} \le \kappa_{A\mapsto x} \le \sqrt{2}\kappa^F_{A\mapsto x}$ allow one to use the computable $\sqrt{2}\kappa^F_{A\mapsto x}$ instead of $\kappa_{A\mapsto x}$ for the least squares problem.

Remarks. Let us look more closely at (2.3). The matrix

$$P = \begin{pmatrix} I - \frac{yy^T}{\|y\|_2^2} & \frac{yx^T}{\|y\|_2\|x\|_2} \\ 0 & I \end{pmatrix}$$

is a projector of rank $m + n - 1$, i.e., $P^2 = P$. Moreover, $\|P\|_2 = \sqrt{2}$, which implies that $\sup_V \frac{\|l(V)\|_2}{\|V\|_F} \le \sqrt{2}\|[\|x\|_2 X, \ \|y\|_2 Y]\|_2$.

The equality

$$[\|x\|_2 X,\ \|y\|_2 Y]\begin{pmatrix} I-\frac{yy^T}{\|y\|_2^2} & \frac{yx^T}{\|y\|_2\|x\|_2} \\ 0 & I \end{pmatrix} = [\|x\|_2 X,\ \|y\|_2 Y] + \|x\|_2 X\frac{y}{\|y\|_2}\left[-\frac{y^T}{\|y\|_2},\ \frac{x^T}{\|x\|_2}\right]$$

shows that this matrix is a rank 1 correction to the matrix $[\|x\|_2 X,\ \|y\|_2 Y]$. By the Courant–Fisher variational principle [1],

$$\left\|[\|x\|_2 X,\ \|y\|_2 Y] + \|x\|_2 X\frac{y}{\|y\|_2}\left[-\frac{y^T}{\|y\|_2},\ \frac{x^T}{\|x\|_2}\right]\right\|_2 \geq \sigma_2([\|x\|_2 X,\ \|y\|_2 Y]).$$

We denote by $\sigma_i(M)$ the singular values of a matrix M in decreasing order so that $\sigma_1(M) = \sigma_{\max}(M) = \|M\|_2$. Thus,

$$\sigma_2([\|x\|_2 X,\ \|y\|_2 Y]) \leq \sup_V \frac{\|l(V)\|_2}{\|V\|_F} \leq \sqrt{2}\sigma_1([\|x\|_2 X,\ \|y\|_2 Y]). \tag{2.5}$$

This observation implies that the exact formula (2.3) can be successfully replaced by the straightforward estimate $\sup_V \frac{\|l(V)\|_2}{\|V\|_F} \leq \sqrt{2}\sigma_1([\|x\|_2 X,\ \|y\|_2 Y])$ or $\sup_V \frac{\|l(V)\|_2}{\|V\|_2} \leq \|x\|_2\|X\|_2 + \|y\|_2\|Y\|_2$ when $\sigma_2([\|x\|_2 X,\ \|y\|_2 Y])$ differs little from $\sigma_1([\|x\|_2 X,\ \|y\|_2 Y])$. This last conclusion is applied in the next section to some examples of the Tikhonov regularization solution.

3. Condition numbers for the Tikhonov regularization solution. If a matrix A is ill-conditioned or of deficient rank, then the vector

$$x_\lambda = (A^TA + \lambda I)^{-1}A^Tb$$

for some fixed $\lambda > 0$ is called the Tikhonov regularization solution to (1.3); cf. [1] and [6].

Eliminating $(A^TA + \lambda I)x_\lambda = A^Tb$ from its perturbed counterpart

$$\left[(A+\delta A)^T(A+\delta A) + \lambda I\right](x_\lambda + \delta x) = (A+\delta A)^T(b+\delta b),$$

and omitting second order terms, we obtain the following expression for δx:

$$\delta x = (A^TA + \lambda I)^{-1}\left[A^T\delta b - A^T(\delta A)x_\lambda + (\delta A)^T(b - Ax_\lambda)\right].$$

A straightforward perturbation bound for the Tikhonov solution x_λ is

$$\frac{\|\delta x\|_2}{\|x_\lambda\|_2} \leq \frac{\|(A^TA+\lambda I)^{-1}A^T\|_2\|b\|_2}{\|x_\lambda\|_2}\frac{\|\delta b\|_2}{\|b\|_2} + \|A\|_2\left(\|(A^TA+\lambda I)^{-1}A^T\|_2 + \|(A^TA+\lambda I)^{-1}\|_2\frac{\|r\|_2}{\|x_\lambda\|_2}\right)\frac{\|\delta A\|_2}{\|A\|_2},$$

where $r = b - Ax_\lambda$, whence

$$\kappa_{A\mapsto x} = \lim_{\delta\to 0}\sup_{\|\delta b\|_2\leq\delta}\left(\frac{\|\delta x\|_2}{\|x_\lambda\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2}\right) \leq \|A\|_2\left(\|(A^TA+\lambda I)^{-1}A^T\|_2 + \|(A^TA+\lambda I)^{-1}\|_2\frac{\|r\|_2}{\|x_\lambda\|_2}\right). \tag{3.1}$$

It is evident that

$$\kappa_{b\mapsto x} = \lim_{\delta\to 0} \sup_{\|\delta b\|_2\le\delta} \left(\frac{\|\delta x\|_2}{\|x_\lambda\|_2} \Big/ \frac{\|\delta b\|_2}{\|b\|_2} \right) = \frac{\|(A^TA+\lambda I)^{-1}A^T\|_2\|b\|_2}{\|x_\lambda\|_2}.$$

To derive exact condition numbers with respect to perturbations of A, let us apply (2.3) with $X = -(A^TA+\lambda I)^{-1}A^T$, $Y = (A^TA+\lambda I)^{-1}$, $y = r = b - Ax_\lambda$ as follows:

$$\begin{aligned}\kappa^F_{A\mapsto x} &= \lim_{\delta\to 0} \sup_{\|\delta A\|_F\le\delta} \left(\frac{\|\delta x\|_2}{\|x_\lambda\|_2} \Big/ \frac{\|\delta A\|_F}{\|A\|_2} \right) \\ &= \|A\|_2 \left\| (A^TA+\lambda I)^{-1} \left[-A^T, \ \frac{\|r\|_2}{\|x_\lambda\|_2} I \right] \begin{pmatrix} I - \frac{rr^T}{\|r\|_2^2} & \frac{rx_\lambda^T}{\|r\|_2\|x_\lambda\|_2} \\ 0 & I \end{pmatrix} \right\|_2 . \end{aligned} \tag{3.2}$$

3.1. An example. It is very tempting to compare the straightforward perturbation bound (3.1) to the exact condition number (3.2). Consider a diagonal matrix $A = \operatorname{diag}(1,\epsilon)$ with positive $\epsilon \ll 1$ and choose $x_\lambda = (0,1)^T$. Then $r = (0,\lambda/\epsilon)^T$,

$$\|(A^TA+\lambda I)^{-1}A^T\|_2 + \|(A^TA+\lambda I)^{-1}\|_2 \frac{\|r\|_2}{\|x_\lambda\|_2} = \max\left\{ \frac{1}{\lambda+1}, \frac{\epsilon}{\lambda+\epsilon^2} \right\} + \frac{\lambda}{\epsilon(\lambda+\epsilon^2)},$$

and

$$(A^TA+\lambda I)^{-1} \left[-A^T, \ \frac{\|r\|_2}{\|x_\lambda\|_2} I \right] \begin{pmatrix} I - \frac{rr^T}{\|r\|_2^2} & \frac{rx_\lambda^T}{\|r\|_2\|x_\lambda\|_2} \\ 0 & I \end{pmatrix}$$

$$= \begin{bmatrix} -\frac{1}{\lambda+1} & 0 & \frac{\lambda}{\epsilon(\lambda+1)} & 0 \\ 0 & 0 & 0 & \frac{\lambda-\epsilon^2}{\epsilon(\lambda+\epsilon^2)} \end{bmatrix}.$$

Let us now choose $\lambda = \epsilon^2$. The exact condition number for the Tikhonov solution x_λ will be

$$\kappa^F_{A\mapsto x} = \max\left\{ \frac{\sqrt{1+(\lambda/\epsilon)^2}}{\lambda+1}, \frac{\lambda-\epsilon^2}{\epsilon(\lambda+\epsilon^2)} \right\} = \frac{1}{\sqrt{1+\epsilon^2}} \approx 1.$$

At the same time, the upper bound in (3.1) equals $\epsilon/(\lambda+\epsilon^2) + (\lambda/\epsilon)/(\lambda+\epsilon^2) = \epsilon^{-1}$, which clearly demonstrates that the straightforward upper bounds can be large overestimations.

3.2. Remarks. The Tikhonov solution case is tractable using the singular values of A, $\sigma_1 \ge \cdots \ge \sigma_n \ge 0$. It is easy to calculate that the singular values of the matrix $M = (A^TA+\lambda I)^{-1}[-A^T, \ \frac{\|r\|_2}{\|x_\lambda\|_2} I]$ equal the numbers

$$\frac{\sqrt{\sigma_i^2 + \|r\|_2^2/\|x_\lambda\|_2^2}}{\sigma_i^2+\lambda}, \qquad i = 1,\dots,n.$$

The derivative of the function $f(\sigma) = \sqrt{\sigma^2 + \|r\|_2^2/\|x_\lambda\|_2^2}/(\sigma^2+\lambda)$ is

$$f'(\sigma) = \frac{x[(\lambda - 2\|r\|_2^2/\|x_\lambda\|_2^2) - \sigma^2]}{(\sigma^2+\lambda)^2\sqrt{\sigma^2 + \|r\|_2^2/\|x_\lambda\|_2^2}}.$$

If $\lambda \le 2\|r\|_2^2/\|x_\lambda\|_2^2$, then $f'(\sigma) \le 0$ for $\sigma \ge 0$. Therefore,

$$\sigma_1(M) = \sigma_{\max}(M) = \frac{\sqrt{\sigma_n^2 + \|r\|_2^2/\|x_\lambda\|_2^2}}{\sigma_n^2 + \lambda}, \qquad \sigma_2(M) = \frac{\sqrt{\sigma_{n-1}^2 + \|r\|_2^2/\|x_\lambda\|_2^2}}{\sigma_{n-1}^2 + \lambda}.$$

Assume that $\sigma_n \le \sigma_{n-1} \ll \sqrt{\lambda}$. Then $\sigma_1(M)$ and $\sigma_2(M)$ are close to each other and are approximately equal to $\frac{\|r\|_2^2}{\lambda\|x_\lambda\|_2^2}$. Taking into account the remarks of section 2, we can conclude that in this case $\kappa^F_{A\mapsto x} \approx \frac{\|A\|_2\|r\|_2^2}{\lambda\|x_\lambda\|_2^2}$.

Another interesting case is when $\|r\|_2$ is small so that $2\|r\|_2^2/\|x_\lambda\|_2^2 \ll \lambda$. The maximum value of $f(\sigma)$ is attained at $\sigma = \sqrt{\lambda}$ and equals $\frac{1}{2\sqrt{\lambda}}$. If there exists a pair of singular values of A sufficiently close to λ, then $\sigma_1(M)$ and $\sigma_2(M)$ are close to $\frac{1}{2\sqrt{\lambda}}$ and $\kappa^F_{A\mapsto x} \approx \frac{\|A\|_2}{2\sqrt{\lambda}}$.

4. Condition numbers for the least squares problem with quadratic inequality constraint. When a matrix A is ill-conditioned or of deficient rank, then problem LSQI (least squares with quadratic inequality constraint) also can be useful. The standard form of LSQI [1] reads

$$\min_x \|Ax - b\|_2 \quad \text{subject to } \|x\|_2 \le \gamma, \tag{4.1}$$

where $\gamma > 0$ is some constant. If $\|A^+b\|_2 \le \gamma$, then A^+b is the unique solution of (4.1); i.e., LSQI reduces to the unconstrained linear least squares problem. Therefore, throughout this section we assume that $\|A^+b\|_2 > \gamma$. The unique solution of (4.1) is then given by

$$x = (A^TA + \lambda I)^{-1}A^Tb, \tag{4.2}$$

where the parameter $\lambda > 0$ is such that the constraint $\|x\|_2 = \gamma$ holds.

A perturbed version of the normal equations $(A^TA + \lambda I)x = A^Tb$ is

$$\left[(A+\delta A)^T(A+\delta A) + (\lambda+\delta\lambda)I\right](x+\delta x) = (A+\delta A)^T(b+\delta b), \tag{4.3}$$

where δA and δb are infinitesimal perturbations of A and b. The constraint $\|x\|_2 = \gamma$ implies $x^T\delta x = 0$. Eliminating $(A^TA + \lambda I)x = A^Tb$ from (4.3) and omitting second order terms, we derive the equation

$$\delta x = (A^TA + \lambda I)^{-1}\left[A^T\delta b - A^T(\delta A)x + (\delta A)^Tr\right] - \delta\lambda(A^TA + \lambda I)^{-1}x,$$

where $r = b - Ax$ is the residual. Multiplying the above equation by x^T from the left and recalling the condition $x^T\delta x = 0$, we obtain the formula

$$\delta\lambda = \frac{x^T(A^TA + \lambda I)^{-1}}{x^T(A^TA + \lambda I)^{-1}x}\left[A^T\delta b - A^T(\delta A)x + (\delta A)^Tr\right].$$

With the help of this formula we derive the final algebraic equation for δx,

$$\delta x = H\left[A^T\delta b - A^T(\delta A)x + (\delta A)^Tr\right], \tag{4.4}$$

where

$$H = (A^TA + \lambda I)^{-1} - \frac{(A^TA + \lambda I)^{-1}xx^T(A^TA + \lambda I)^{-1}}{x^T(A^TA + \lambda I)^{-1}x} \tag{4.5}$$

is a symmetric matrix.

The straightforward perturbation estimate appears as follows:

$$\frac{\|\delta x\|_2}{\|x\|_2} \le \frac{\|HA^T\|_2\|b\|_2}{\|x\|_2}\frac{\|\delta b\|_2}{\|b\|_2} + \|A\|_2\left(\|HA^T\|_2 + \|H\|_2\frac{\|r\|_2}{\|x\|_2}\right)\frac{\|\delta A\|_2}{\|A\|_2}. \tag{4.6}$$

It is trivially verified that

$$\kappa_{b\mapsto x} = \lim_{\delta\to 0}\sup_{\|\delta b\|_2\le\delta}\left(\frac{\|\delta x\|_2}{\|x\|_2}\Big/\frac{\|\delta b\|_2}{\|b\|_2}\right) = \frac{\|HA^T\|_2\|b\|_2}{\|x\|_2}.$$

Applying (2.3) with $X = -HA^T$, $Y = H$, $y = r = b - Ax$, we get the exact condition number

$$\begin{aligned}\kappa^F_{A\mapsto x} &= \lim_{\delta\to 0}\sup_{\|\delta A\|_2\le\delta}\left(\frac{\|\delta x\|_2}{\|x\|_2}\Big/\frac{\|\delta A\|_2}{\|A\|_2}\right)\\ &= \|A\|_2\left\|H\left[-A^T,\ \frac{\|r\|_2}{\|x\|_2}I\right]\begin{pmatrix} I - \frac{rr^T}{\|r\|_2^2} & \frac{rx^T}{\|r\|_2\|x\|_2}\\ 0 & I\end{pmatrix}\right\|_2.\end{aligned} \tag{4.7}$$

5. General LSQI formulation. The general LSQI problem reads

$$\min_x \|Ax - b\|_2 \quad \text{subject to } \|Cx - d\|_2 \le \gamma. \tag{5.1}$$

As in [1], we denote by $x_{A,C}$ solutions of the problem

$$\min_{x\in S}\|Cx - d\|_2, \qquad S = \{x \in R^n \mid \|Ax - b\|_2 = \min\}$$

and assume that $\|Cx_{A,C} - d\|_2 > \gamma$. Under this assumption the unique solution to (5.1) satisfies the generalized normal equations

$$(A^TA + \lambda C^TC)x = A^Tb + \lambda C^Td, \tag{5.2}$$

where the parameter λ is determined by the secular equation $\|Cx - d\|_2 = \gamma$.

A perturbed version of the generalized normal equations is

$$\begin{aligned}&[(A+\delta A)^T(A+\delta A) + (\lambda+\delta\lambda)(C+\delta C)^T(C+\delta C)](x+\delta x)\\ &\qquad = (A+\delta A)^T(b+\delta b) + (\lambda+\delta\lambda)(C+\delta C)^T(d+\delta d).\end{aligned}$$

Eliminating (5.2) from this version and dropping second order terms, we derive the following expression for δx in terms of δA, δb, δC, and δd:

$$\delta x = (A^TA + \lambda C^TC)^{-1}F - \delta\lambda(A^TA + \lambda C^TC)^{-1}C^T(Cx - d), \tag{5.3}$$

where

$$\begin{aligned}F &= A^T\delta b + \lambda C^T\delta d + (\delta A)^T(b - Ax) - A^T(\delta A)x + \lambda(\delta C)^T(d - Cx) - \lambda C^T(\delta C)x\\ &= \begin{pmatrix} A\\ \lambda C\end{pmatrix}^T\begin{pmatrix}\delta b\\ \delta d\end{pmatrix} - \begin{pmatrix} A\\ \lambda C\end{pmatrix}^T\begin{pmatrix}\delta A\\ \delta C\end{pmatrix}x + \begin{pmatrix}\delta A\\ \delta C\end{pmatrix}^T\begin{bmatrix} b - Ax\\ \lambda(d - Cx)\end{bmatrix}.\end{aligned}$$

The constraint $\|Cx-d\|_2=\gamma$ implies $(Cx-d)^TC\delta x=0$, whence

$$\delta\lambda=\frac{(Cx-d)^TC(A^TA+\lambda C^TC)^{-1}}{(Cx-d)^TC(A^TA+\lambda C^TC)^{-1}C^T(Cx-d)}F.$$

Substituting the above formula of $\delta\lambda$ into (5.3), we obtain the algebraic relation

$$\delta x=\mathcal{H}F$$

with the symmetric matrix

$$\mathcal{H}=(A^TA+\lambda C^TC)^{-1}-\frac{(A^TA+\lambda C^TC)^{-1}C^T(Cx-d)(Cx-d)^TC(A^TA+\lambda C^TC)^{-1}}{(Cx-d)^TC(A^TA+\lambda C^TC)^{-1}C^T(Cx-d)}.$$

Applying (2.3) with $X=-\mathcal{H}\binom{A}{\lambda C}^T$, $Y=\mathcal{H}$, and $y=\left[\begin{smallmatrix}b-Ax\\ \lambda(d-Cx)\end{smallmatrix}\right]$, we get the exact condition number

$$\begin{aligned}\kappa^F_{(A,C)\mapsto x}&=\lim_{\delta\to0}\sup_{\|\binom{\delta A}{\delta C}\|_F\le\delta}\left(\frac{\|\delta x\|_2}{\|x\|_2}\Big/\frac{\left\|\binom{\delta A}{\delta C}\right\|_F}{\left\|\binom{A}{C}\right\|_2}\right)\\&=\left\|\begin{pmatrix}A\\C\end{pmatrix}\right\|_2\left\|\mathcal{H}\left[-\begin{pmatrix}A\\\lambda C\end{pmatrix}^T,\ \frac{\|y\|_2}{\|x\|_2}I\right]\begin{pmatrix}I-\frac{yy^T}{\|y\|_2^2}&\frac{yx^T}{\|y\|_2\|x\|_2}\\0&I\end{pmatrix}\right\|_2.\end{aligned}$$

There is a possibility of deriving similar condition numbers in weighted norms like $\|\binom{\alpha A}{\beta C}\|$, where, e.g., $\alpha=1$ and $\beta=\mu=\sqrt{\lambda}$. In the latter case,

$$F=\begin{pmatrix}A\\\mu C\end{pmatrix}^T\begin{pmatrix}\delta b\\\mu\delta d\end{pmatrix}-\begin{pmatrix}A\\\mu C\end{pmatrix}^T\begin{pmatrix}\delta A\\\mu\delta C\end{pmatrix}x+\begin{pmatrix}\delta A\\\mu\delta C\end{pmatrix}^T\begin{bmatrix}b-Ax\\\mu(d-Cx)\end{bmatrix}$$

and

$$\begin{aligned}\kappa^F_{(A,\mu C)\mapsto x}&=\lim_{\delta\to0}\sup_{\|\binom{\delta A}{\mu\delta C}\|_F\le\delta}\left(\frac{\|\delta x\|_2}{\|x\|_2}\Big/\frac{\left\|\binom{\delta A}{\mu\delta C}\right\|_F}{\left\|\binom{A}{\mu C}\right\|_2}\right)\\&=\left\|\begin{pmatrix}A\\\mu C\end{pmatrix}\right\|_2\left\|\left[-\mathcal{H}\begin{pmatrix}A\\\mu C\end{pmatrix}^T,\ \mathcal{H}\frac{\left\|\left[\begin{smallmatrix}b-Ax\\\mu(d-Cx)\end{smallmatrix}\right]\right\|_2}{\|x\|_2}\right]\right\|_2.\end{aligned}$$

For Tikhonov solutions of the form $x_\lambda=(A^TA+\lambda C^TC)^{-1}(A^Tb+\lambda C^Td)$, simply replace $\mathcal{H}$ with $(A^TA+\lambda C^TC)^{-1}$ in all the above formulas.

6. Least squares with linear equality constraints. In this section we study the following problem of least squares with equality constraints [1]: Given matrices $A\in R^{m\times n}$ and $B\in R^{p\times n}$, find a vector $x\in R^n$, which solves

$$\min_x\|Ax-b\|_2\quad\text{subject to } Bx=d. \tag{6.1}$$

The rank conditions $\operatorname{rank}(B)=p$ and $\operatorname{rank}\binom{A}{B}=n$ guarantee existence of the unique solution $x=Q_2(AQ_2)^+b+[I-Q_2(AQ_2)^+A]B^+d$, where the columns of Q_2 form an

orthogonal basis for the nullspace of B, null(B). Standard theorems about Lagrange multipliers provide the augmented system defining the unique solution x as follows:

$$\begin{pmatrix} A^TA & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} A^Tb \\ d \end{pmatrix}. \tag{6.2}$$

The parameter λ equals $(AB^+)^T(b - Ax)$.

If δA and δB are perturbations of A and B, respectively, then the perturbed augmented system reads

$$\begin{bmatrix} (A+\delta A)^T(A+\delta A) & (B+\delta B)^T \\ B+\delta B & 0 \end{bmatrix} \begin{pmatrix} x+\delta x \\ \lambda+\delta\lambda \end{pmatrix} = \begin{bmatrix} (A+\delta A)^T(b+\delta b) \\ d+\delta d \end{bmatrix}.$$

After some algebra, we obtain the linear system

$$\begin{pmatrix} A^TA & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \delta x \\ \delta\lambda \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{6.3}$$

with

$$\begin{aligned} f_1 &= A^T\delta b - A^T(\delta A)x + (\delta A)^T(b - Ax) - (\delta B)^T(AB^+)^T(b - Ax), \\ f_2 &= \delta d - (\delta B)x. \end{aligned}$$

It follows from the second row of (6.3) that $\delta x = B^+f_2 + Q_2\xi$. From the first row of (6.3) we have $Q_2^TA^TAQ_2\xi = Q_2^Tf_1 - Q_2^TA^TAB^+f_2$. The matrix AQ_2 is of full rank, and therefore $\xi = (Q_2^TA^TAQ_2)^{-1}Q_2^Tf_1 - (Q_2^TA^TAQ_2)^{-1}Q_2^TA^TAB^+f_2$. Finally,

$$\delta x = \left[I - Q_2(Q_2^TA^TAQ_2)^{-1}Q_2^TA^TA\right] B^+f_2 + Q_2(Q_2^TA^TAQ_2)^{-1}Q_2^Tf_1.$$

Let us denote

$$K = Q_2(AQ_2)^+ = (AP_0)^+,$$

where $P_0 = I - B^+B = Q_2Q_2^T$ is the orthogonal projector onto null(B). Since $\delta x = KK^Tf_1 + (I - KA)B^+f_2$, the perturbation δx equals

$$\begin{aligned} \delta x = {} & KK^TA^T\delta b + (I - KA)B^+\delta d - (I - KA)B^+(\delta B)x \\ & + KK^T\left[-A^T(\delta A)x + (\delta A)^T(b - Ax) - (\delta B)^T(AB^+)^T(b - Ax)\right]. \end{aligned}$$

Taking into account the identity $KK^TA^T = K$, we derive the representation

$$\begin{aligned} \delta x = {} & \begin{bmatrix} K & (I - KA)B^+ \end{bmatrix} \begin{pmatrix} \delta b \\ \delta d \end{pmatrix} \\ & - \begin{bmatrix} K & (I - KA)B^+ \end{bmatrix} \begin{pmatrix} \delta A \\ \delta B \end{pmatrix} x + KK^T \begin{pmatrix} \delta A \\ \delta B \end{pmatrix}^T \begin{bmatrix} b - Ax \\ -(AB^+)^T(b - Ax) \end{bmatrix}, \end{aligned}$$

which coincides with the representation of Corollary 3.6 from [3].

Applying formula (2.3) with

$$X = -[K\ (I - KA)B^+], \quad Y = KK^T, \quad y = \begin{bmatrix} b - Ax \\ -(AB^+)^T(b - Ax) \end{bmatrix},$$

we get the condition number

$$\kappa^F_{(A,B)\mapsto x} = \lim_{\delta\to 0} \sup_{\left\|\binom{\delta A}{\delta B}\right\|_F \le \delta} \left(\frac{\|\delta x\|_2}{\|x\|_2} \Bigg/ \frac{\left\|\binom{\delta A}{\delta B}\right\|_F}{\left\|\binom{A}{B}\right\|_2} \right)$$

$$= \left\|\begin{pmatrix} A \\ B \end{pmatrix}\right\|_2 \left\| \begin{bmatrix} -K, & -(I-KA)B^+, & KK^T \frac{\|y\|_2}{\|x\|_2} \end{bmatrix} \begin{pmatrix} I - \frac{yy^T}{y^Ty} & \frac{yx^T}{\|y\|_2\|x\|_2} \\ 0 & I \end{pmatrix} \right\|_2 .$$

Note that it is possible to derive similar condition numbers in weighted norms.

7. Conclusion. In this paper we derived new computable formulas for exact normwise condition numbers for several least squares solutions and Tikhonov regularization solutions. All the results were obtained by means of a general construction described in section 2. Further work is needed for interpretation of the exact condition numbers in terms of other characteristics such as, e.g., singular values. The "straightforward" perturbation bounds often appear simpler and are easier to use for interpretation, but we emphasize that they are not always sharp.

REFERENCES

[1] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
[2] C. Davis, W.M. Kahan, and H.F. Weinberger, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 19 (1982), pp. 445–469.
[3] L. Eldén, *Perturbation theory for the least squares problem with linear equality constraints*, SIAM J. Numer. Anal., 17 (1980), pp. 338–350.
[4] A.J. Geurts, *A contribution to the theory of condition*, Numer. Math., 39 (1982), pp. 85–96.
[5] S. Gratton, *On the condition number of linear least squares problems in a weighted Frobenius norm*, BIT, 36 (1996), pp. 523–530.
[6] M. Gulliksson and P.-Å. Wedin, *Perturbation theory for generalized and constrained linear least squares*, Numer. Linear Algebra Appl., 7 (2000), pp. 181–195.
[7] M.E. Gulliksson, P.-Å. Wedin, and Y. Wei, *Perturbation identities for regularized Tikhonov inverses and weighted pseudoinverses*, BIT, 40 (2000), pp. 513–523.
[8] P.C. Hansen, *Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 503–518.
[9] A. Laratta and F. Zironi, *Computation of Lagrange multipliers for linear least squares problems with equality constraints*, Computing, 67 (2001), pp. 335–350.
[10] J.R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.
[11] L.N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.