SPECIAL ISSUE ON ACCURATE SOLUTION OF EIGENVALUE PROBLEMS

The occasion for this special issue is the Sixth International Workshop on Accurate Solution of Eigenvalue Problems, which took place at The Pennsylvania State University from May 22–25, 2006. This special issue provides an outlet for papers from the workshop and recognizes advances in the numerical solution of eigenvalue and related problems. This is the second such special issue published in the [illegible]; the first was published in number 4 of volume 28 in connection with the Fifth International Workshop on Accurate Solution of Eigenvalue Problems, which took place in Hagen, Germany, from June 29 to July 1, 2004.

The twelve papers in the current issue are concerned with a variety of aspects that arise in the computation of eigenvalues and invariant subspaces: perturbation bounds and sensitivity, accuracy and convergence behavior of algorithms, exploitation of structure in matrices, and particular engineering applications.

Thanks go to SIMAX Editor-in-Chief, Henk van der Vorst; guest editors Jesse Barlow, Froilán Dopico, and Zlatko Drmač, who put great effort into the careful and timely review of papers; and Mitch Chernoff, Cherie Trebisky, and other members of the SIAM staff who worked hard to publish this special issue.

[illegible]
[illegible]

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 1–10

THE ARNOLDI EIGENVALUE ITERATION WITH EXACT SHIFTS CAN FAIL[*]

MARK EMBREE[†]

Abstract. The restarted Arnoldi algorithm, implemented in the ARPACK software library and MATLAB's `eigs` command, is among the most common means of computing select eigenvalues and eigenvectors of a large, sparse matrix. To assist convergence, a starting vector is repeatedly refined via the application of automatically constructed polynomial filters whose roots are known as "exact shifts." Though Sorensen proved the success of this procedure under mild hypotheses for Hermitian matrices, a convergence proof for the non-Hermitian case has remained elusive. The present note describes a class of examples for which the algorithm fails in the strongest possible sense; that is, the polynomial filter used to restart the iteration deflates the eigenspace one is attempting to compute.

Key words. implicitly restarted Arnoldi algorithm, Krylov–Schur algorithm, eigenvalues, exact shifts, ARPACK, `eigs`

AMS subject classifications. 15A18, 15A60, 65F15

DOI. 10.1137/060669097

1. Setting.

Large-scale matrix eigenvalue problems typically derive from applications in which one seeks only some small subset of the spectrum, such as the largest magnitude or rightmost eigenvalues for stability analysis of dynamical systems. Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, one can approximate such eigenvalues by projecting $\mathbf{A}$ onto an appropriate low-dimensional subspace and then solving a small eigenvalue problem with a dense method such as the QR algorithm. The Arnoldi method [1, 13] orthogonally projects $\mathbf{A}$ onto the Krylov subspace

$$\mathcal{K}_k(\mathbf{A},\mathbf{v}) := \operatorname{span}\{\mathbf{v}, \mathbf{A}\mathbf{v}, \ldots, \mathbf{A}^{k-1}\mathbf{v}\}$$

generated by the starting vector $\mathbf{v} \in \mathbb{C}^n$. This subspace—the span of iterates of the power method—often produces eigenvalue estimates that fall on the "periphery" of the spectrum [16], though the quality of such approximations depends on the distribution of the eigenvalues of $\mathbf{A}$, the angles between associated eigenvectors, and the starting vector $\mathbf{v}$. Provided that $\dim \mathcal{K}_{k+1}(\mathbf{A},\mathbf{v}) = k+1$, by the end of the kth iteration the Arnoldi algorithm has produced orthonormal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}$ such that

$$\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_j\} = \mathcal{K}_j(\mathbf{A},\mathbf{v})$$

for all $j = 1, \ldots, k+1$. Accumulating $\mathbf{v}_1, \ldots, \mathbf{v}_k$ into the columns of $\mathbf{V}_k \in \mathbb{C}^{n\times k}$, the algebra affecting the construction of the basis can be summarized in the Arnoldi factorization

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + h_{k+1,k}\mathbf{v}_{k+1}\mathbf{e}_k^*, \tag{1}$$

where $\mathbf{e}_k$ denotes the last column of the $k \times k$ identity matrix and $\mathbf{H}_k = \mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$ is an upper Hessenberg matrix. The eigenvalues of $\mathbf{H}_k$, the [illegible] [illegible], approximate

[*] Received by the editors September 5, 2006; accepted for publication by F. Dopico November 2, 2006; published electronically February 27, 2009. Supported by Department of Energy grant DE-FG03-02ER25531 and National Science Foundation grant DMS-CAREER-0449973.
http://www.siam.org/journals/simax/31-1/66909.html

[†] Department of Computational and Applied Mathematics, Rice University, 6100 Main Street, MS 134, Houston, TX 77005–1892 (embree@rice.edu).

eigenvalues of $\mathbf{A}$ in the sense that if $\mathbf{H}_k\mathbf{u} = \theta\mathbf{u}$, then

$$\|\mathbf{A}(\mathbf{V}_k\mathbf{u}) - \theta(\mathbf{V}_k\mathbf{u})\| \leq |h_{k+1,k}||\mathbf{e}_k^*\mathbf{u}|.$$

These Ritz values must fall within the [illegible] (or [illegible]) of $\mathbf{A}$,

$$W(\mathbf{A}) := \{z \in \mathbb{C} : z = \mathbf{x}^*\mathbf{A}\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{C}^n \text{ with } \|\mathbf{x}\| = 1\}, \tag{2}$$

since $\theta = (\mathbf{V}_k\mathbf{u})^*\mathbf{A}(\mathbf{V}_k\mathbf{u})$. The numerical range of a normal matrix is the convex hull of its spectrum [5]; for nonnormal matrices, the numerical range may contain points far from any eigenvalue (see, e.g., [17, Ch. 17]), a fact critical to the examples we shall present. Given the Arnoldi algorithm's broad utility, its nontrivial convergence behavior has attracted considerable attention [3, 4, 6, 7, 8, 13].

The computational cost of enlarging the Krylov subspace in the Arnoldi algorithm grows with the subspace dimension, k, and for large problems storage of the basis vectors alone becomes burdensome. Unfortunately, in many cases practical values of k fail to produce sufficiently accurate eigenvalue estimates. A simple solution is to [illegible] the iteration, using information culled from the factorization (1) to refine the starting vector $\mathbf{v}$ in a manner that enriches components in the direction of desired eigenvectors while damping unwanted eigenvectors. This process is most commonly implemented through a [illegible]: the starting vector for the new Arnoldi factorization takes the form

$$\mathbf{v}_+ = \psi(\mathbf{A})\mathbf{v}$$

for some polynomial ψ, generally of low degree. Strategies for constructing ψ vary. Saad advocated a choice designed to produce $\mathbf{v}_+$ as a linear combination of the desired Ritz vectors [13]. Later methods chose ψ to be small over an adaptively identified region of the complex plane containing only the unwanted eigenvalues, but such procedures proved difficult to reliably automate. Sorensen proposed a less elaborate strategy: take ψ to be a polynomial whose roots match the undesired Ritz values [14]. More precisely, suppose we seek m eigenvalues and have already built an Arnoldi factorization of dimension $k = m+p$ for some $p > 0$. Sort the eigenvalues of $\mathbf{H}_k$ into the m possessing the desired trait (e.g., those rightmost or largest in magnitude), labeled $\theta_1, \ldots, \theta_m$, and the remaining p values $\theta_{m+1}, \ldots, \theta_{m+p}$, which we implicitly assume are coarse approximations to eigenvalues we do not seek. Set $\psi(z) = \prod_{j=1}^{p}(z-\theta_{m+j})$, so that

$$\mathbf{v}_+ = \prod_{j=1}^{p}(\mathbf{A} - \theta_{m+j}\mathbf{I})\mathbf{v}.$$

These roots of ψ, the undesired Ritz values, are called [illegible]. Such shifts form an essential part of the ARPACK library [9], and their efficacy is responsible in no small part for the popularity enjoyed by that software and MATLAB's subordinate `eigs` command. (In the same paper that advocated exact shifts, Sorensen also proposed a robust method of [illegible] restarting the Arnoldi algorithm using the roots of an arbitrary filter polynomial [14]. An alternative implementation, Stewart's Krylov–Schur algorithm, requires the use of exact shifts [15].)

Numerous computational examples demonstrate the success of exact shifts; see, e.g., [3, 4, 9]. A convergence proof would require that one make precise the notion that the p Ritz values used as shifts approximate the unwanted eigenvalues. For Hermitian

$\mathbf{A}$ an appeal to the interlacing theorem [11] suffices. Labeling the eigenvalues of $\mathbf{A}$ as $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, there can be at most one Ritz value in the interval (λ_1, λ_2), two Ritz values in (λ_2, λ_3), and so on. This interlacing forms the basis of Sorensen's proof that for a Hermitian matrix the restarted Arnoldi algorithm converges to extreme eigenvalues.[1] In the years since the introduction of exact shifts, a convergence proof for non-Hermitian matrices has remained elusive. Several results ensure convergence provided the exact shifts satisfy an appropriate distribution [3, 4, 8], but conditions guaranteeing that exact shifts exhibit such behavior have not yet been established. Indeed, few fine results about the Ritz values of non-Hermitian matrices are known.

The purpose of this note is to provide a counterexample to the conjecture that the restarted Arnoldi algorithm with exact shifts must converge under hypotheses resembling those sufficient for the Hermitian case. Our examples illustrate the failure of the algorithm in the strongest sense, in that the filter polynomial ψ exactly deflates a sought-after (perfectly conditioned) eigenvalue.

Throughout we assume that all computations are performed in exact arithmetic and in particular that eigenvalues of the upper Hessenberg matrix $\mathbf{H}_k$ are determined exactly.

2. Simple example. We begin with a small example that clearly demonstrates how restarting with exact shifts can produce catastrophic results. Suppose we seek the largest magnitude (and rightmost) eigenvalue $\lambda = 1$ and associated eigenvector $\mathbf{u}_1 = [1\ 0\ 0\ 0]^T$ of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We denote the bottom-right 3×3 submatrix of $\mathbf{A}$ by

$$\mathbf{D} = \begin{bmatrix} 0 & 6 & -2 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}.$$

The desired eigenvalue $\lambda = 1$ of $\mathbf{A}$ is far from the rest of the spectrum, $\sigma(\mathbf{D}) = \{0\}$, and is also fairly well separated [16, p. 46]:

$$\operatorname{sep}(1, \mathbf{D}) = \|(\mathbf{I} - \mathbf{D})^{-1}\|^{-1} = 0.0837\ldots.$$

In the language of pseudospectra [17], this implies $1 \notin \sigma_\varepsilon(\mathbf{D})$ for all $\varepsilon < 0.0837\ldots$, where

$$\begin{aligned} \sigma_\varepsilon(\mathbf{D}) &= \{z \in \mathbb{C} : z \in \sigma(\mathbf{D} + \mathbf{E}) \text{ for some } \mathbf{E} \in \mathbb{C}^{n \times n} \text{ with } \|\mathbf{E}\| < \varepsilon\} \\ &= \{z \in \mathbb{C} : \|(z\mathbf{I} - \mathbf{D})^{-1}\| > \varepsilon^{-1}\}. \end{aligned} \tag{3}$$

Furthermore, the eigenvector $\mathbf{u}_1$ is orthogonal to the complementary invariant subspace associated with the zero eigenvalue; that is, $\lambda = 1$ is perfectly conditioned [2].

[1] Sorensen's theorem [14, Thm. 5.9] imposes several reasonable caveats: the starting vector cannot be deficient in any of the sought-after eigenvectors, and the iteration must not come arbitrarily close to "lucky breakdown"; i.e., all the subdiagonal entries in $\mathbf{H}_k$ must be bounded away from zero, so none of the intermediate Krylov subspaces fall too close to an invariant subspace.

Apply the restarted Arnoldi method to $\mathbf{A}$ with the starting vector $\mathbf{v} = [1\ 1\ 1\ 1]^T$. We seek the $m = 1$ eigenvalue and will build the Krylov subspace out to dimension $k = m + p = 2$ before restarting. Two steps of the Arnoldi algorithm produce the orthonormal basis vectors

$$\mathbf{v}_1 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \mathbf{v}_2 = \frac{\sqrt{35}}{70}\begin{bmatrix} -3 \\ 9 \\ 1 \\ -7 \end{bmatrix}$$

and upper Hessenberg matrix

$$\mathbf{H}_2 = [\mathbf{v}_1\ \mathbf{v}_2]^*\mathbf{A}[\mathbf{v}_1\ \mathbf{v}_2] = \begin{bmatrix} 7/4 & 3\sqrt{35}/140 \\ \sqrt{35}/4 & 5/4 \end{bmatrix}.$$

The characteristic polynomial for $\mathbf{H}_2$ is

$$\begin{aligned} \det(\lambda\mathbf{I} - \mathbf{H}_2) &= (\lambda - 7/4)(\lambda - 5/4) - \frac{3 \cdot 35}{4 \cdot 140} \\ &= (\lambda - 1)(\lambda - 2), \end{aligned}$$

and hence the Ritz values are $\theta_1 = 2$ and $\theta_2 = 1$ (ordered by decreasing magnitude). The smaller magnitude (and leftmost) of these is chosen as the exact shift; hence the filter polynomial $\psi(z) = z - 1$ produces the new starting vector

$$\mathbf{v}_+ = \psi(\mathbf{A})\mathbf{v} = \begin{bmatrix} 0 \\ 3 \\ 1 \\ -1 \end{bmatrix}.$$

Note that $\mathbf{v}_+$ has no component in the desired eigenvector $\mathbf{u}_1$—that is, $\mathbf{v}_+$ is contained in the invariant subspace associated with the zero eigenvalue, and hence the eigenvalue $\lambda = 1$ will exert no influence upon further Arnoldi iterations.

One might suspect that this starting vector has been specially engineered to yield this behavior. While this is true of the last three components of $\mathbf{v}$, it is curious that the phenomenon persists even when $\mathbf{v}$ is arbitrarily enriched in the eigenvector we wish to compute. If we set $\mathbf{v} = [\alpha\ 1\ 1\ 1]^T$ for any $\alpha \in \mathbb{R}$, then two steps of the Arnoldi algorithm yield the upper Hessenberg matrix

$$\mathbf{H}_2 = \begin{bmatrix} \dfrac{6 + \alpha^2}{3 + \alpha^2} & \dfrac{3\alpha^3}{(3 + \alpha^2)(24 + 11\alpha^2)^{1/2}} \\[2ex] \dfrac{(24 + 11\alpha^2)^{1/2}}{3 + \alpha^2} & \dfrac{3 + 2\alpha^2}{3 + \alpha^2} \end{bmatrix}$$

with characteristic polynomial

$$\det(\lambda\mathbf{I} - \mathbf{H}_2) = \lambda^2 - 3\lambda + 2 = (\lambda - 1)(\lambda - 2).$$

The Ritz values $\theta_1 = 2$ and $\theta_2 = 1$ are independent of the bias of the starting vector $\mathbf{v}$ toward the desired eigenvector. As demonstrated in the next section, this behavior is an instance of a broader phenomenon that will facilitate the design of larger examples.

3. General construction. While the orthogonal basis for the Krylov subspace and the resulting Arnoldi factorization (1) have computational advantages, other bases for $\mathcal{K}_k(\mathbf{A}, \mathbf{v})$ are often better suited to analysis. Among the choices available (see [15, Thm. 2.2]), we shall use a decomposition introduced by Ruhe [12, sect. 2]. Suppose $\dim \mathcal{K}_k(\mathbf{A}, \mathbf{v}) = k$. Arrange the Krylov basis vectors $\mathbf{v}$, $\mathbf{A}\mathbf{v}$, $\ldots$, $\mathbf{A}^{k-1}\mathbf{v}$ into the columns of

$$\mathbf{K}_k = \left[\begin{array}{cccc} \mathbf{v} & \mathbf{A}\mathbf{v} & \cdots & \mathbf{A}^{k-1}\mathbf{v} \end{array}\right],$$

and define the companion matrix

$$\mathbf{C}_k = \begin{bmatrix} & & & c_1 \\ 1 & & & c_2 \\ & \ddots & & \vdots \\ & & 1 & c_k \end{bmatrix} \in \mathbb{C}^{k\times k}$$

for constants $c_1, \ldots, c_k$; unspecified entries equal zero. A direct calculation reveals

$$\mathbf{A}\mathbf{K}_k - \mathbf{K}_k\mathbf{C}_k = \mathbf{r}\mathbf{e}_k^* \tag{4}$$

for the vector

$$\mathbf{r} := \mathbf{A}^k\mathbf{v} - \sum_{j=1}^{k} c_j \mathbf{A}^{j-1}\mathbf{v} \ \in \ \mathcal{K}_{k+1}(\mathbf{A}, \mathbf{v}).$$

Now since

$$\operatorname{Ran}\mathbf{K}_j = \operatorname{Ran}\mathbf{V}_j = \mathcal{K}_j(\mathbf{A}, \mathbf{v}), \quad j = 1, \ldots, k,$$

there exists some invertible $\mathbf{R}_k \in \mathbb{C}^{k\times k}$ such that

$$\mathbf{K}_k = \mathbf{V}_k\mathbf{R}_k.$$

Recalling that $\mathbf{H}_k = \mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$, pre- and postmultiply (4) by $\mathbf{V}_k^*$ and $\mathbf{R}_k^{-1}$ to obtain

$$\mathbf{H}_k - \mathbf{R}_k\mathbf{C}_k\mathbf{R}_k^{-1} = \mathbf{V}_k^*\mathbf{r}\mathbf{e}_k^*\mathbf{R}_k^{-1}.$$

If $c_1, \ldots, c_k$ are chosen such that $\mathbf{r} \perp \mathcal{K}_k(\mathbf{A}, \mathbf{v}) = \operatorname{Ran}\mathbf{V}_k$, we conclude that

$$\mathbf{H}_k = \mathbf{R}_k\mathbf{C}_k\mathbf{R}_k^{-1};$$

i.e., the companion matrix $\mathbf{C}_k$ and the upper Hessenberg matrix $\mathbf{H}_k$ are similar and thus have the same eigenvalues. This development facilitates the following result.

THEOREM 3.1. [illegible] $\theta_1, \ldots, \theta_k$ [illegible] $k \le \ell$ [illegible] $\mathbf{D} \in \mathbb{C}^{\ell\times\ell}$ [illegible] $\mathbf{w} \in \mathbb{C}^{\ell}$ [illegible] $\dim \mathcal{K}_k(\mathbf{D}, \mathbf{w}) = k$ [illegible] $\mathbf{T} \in \mathbb{C}^{m\times m}$ $m \le k$ [illegible] $\sigma(\mathbf{T}) \subseteq \{\theta_1, \ldots, \theta_k\}$ ([illegible]) [illegible] $\mathbf{x} \in \mathbb{C}^m$ k [illegible]

$$\mathbf{A} = \begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} \mathbf{x} \\ \mathbf{w} \end{bmatrix}$$

[illegible] $\theta_1, \ldots, \theta_k$

[illegible]. To write a Krylov factorization for $\mathcal{K}_k(\mathbf{D},\mathbf{w})$, we define

$$\widehat{\mathbf{K}}_k = \left[\begin{array}{cccc} \mathbf{w} & \mathbf{D}\mathbf{w} & \cdots & \mathbf{D}^{k-1}\mathbf{w} \end{array}\right], \qquad \mathbf{C}_k = \left[\begin{array}{cccc} & & & c_1 \\ 1 & & & c_2 \\ & \ddots & & \vdots \\ & & 1 & c_k \end{array}\right] \in \mathbb{C}^{k\times k},$$

with coefficients $c_1,\ldots,c_k$ chosen, e.g., via the Gram–Schmidt process, to ensure

$$\widehat{\mathbf{r}} := \mathbf{D}^k\mathbf{w} - \sum_{j=1}^{k} c_j \mathbf{D}^{j-1}\mathbf{w}$$

is orthogonal to $\mathcal{K}_k(\mathbf{D},\mathbf{w})$. This choice gives $\sigma(\mathbf{C}_k) = \{\theta_1,\ldots,\theta_k\}$. Now with

$$\mathbf{K}_k = \left[\begin{array}{cccc} \mathbf{v} & \mathbf{A}\mathbf{v} & \cdots & \mathbf{A}^{k-1}\mathbf{v} \end{array}\right],$$

observe that

$$\mathbf{A}\mathbf{K}_k - \mathbf{K}_k\mathbf{C}_k = \mathbf{r}\mathbf{e}_k^*,$$

where

$$\mathbf{r} := \mathbf{A}^k\mathbf{v} - \sum_{j=1}^{k} c_j\mathbf{A}^{j-1}\mathbf{v} = \left[\begin{array}{c} \mathbf{T}^k\mathbf{x} - \sum_{j=1}^{k} c_j\mathbf{T}^{j-1}\mathbf{x} \\ \mathbf{D}^k\mathbf{w} - \sum_{j=1}^{k} c_j\mathbf{D}^{j-1}\mathbf{w} \end{array}\right].$$

Since $\sigma(\mathbf{T}) \subseteq \{\theta_1,\ldots,\theta_k\}$ (respecting multiplicity), the polynomial

$$z^k - \sum_{j=1}^{k} c_j z^{j-1} = \prod_{j=1}^{k}(z-\theta_j)$$

annihilates $\mathbf{T}$, leaving

$$\mathbf{r} = \left[\begin{array}{c} \mathbf{0} \\ \widehat{\mathbf{r}} \end{array}\right].$$

It follows that $\mathbf{r}$ is orthogonal to $\mathcal{K}_k(\mathbf{A},\mathbf{v})$, and hence the Ritz values for $\mathbf{A}$ and $\mathbf{v}$ drawn from this subspace must be $\theta_1,\ldots,\theta_k$. □

As Ritz values are invariant to unitary similarity transformations, the matrix $\mathbf{A}$ and $\mathbf{v}$ in the theorem could be replaced by $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ and $\mathbf{Q}^*\mathbf{v}$ for any unitary matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$. Of fundamental importance is the fact that the invariant subspace associated with $\sigma(\mathbf{T})$ be orthogonal to the one associated with $\sigma(\mathbf{D})$. The matrix $\mathbf{T}$ need not be normal nor even diagonalizable.

This result (which relates to the example of maximal Krylov subspaces for derogatory matrices given in [3, p. 1081]) suggests a procedure for manufacturing examples like the one presented in section 2. The following approach presumes that one seeks the largest magnitude eigenvalues of a matrix (as is typical, e.g., after performing a shift-invert or Cayley transformation of the original matrix [10]).

1. Find a matrix $\mathbf{D}$ and starting vector $\mathbf{w}$ such that the Ritz values produced by $k \geq 2m$ steps of the Arnoldi process, when ordered by decreasing magnitude, satisfy

(5) $$|\theta_m| > |\theta_{m+1}| \geq \cdots \geq |\theta_{2m}| > \max_{\lambda\in\sigma(\mathbf{D})} |\lambda|.$$

2. Construct a matrix $\mathbf{T} \in \mathbb{C}^{m\times m}$ with eigenvalues $\theta_{m+1}, \ldots, \theta_{2m}$; take $\mathbf{x} \in \mathbb{C}^m$ to be any vector.

3. Build the matrix and starting vector

$$\mathbf{A} = \begin{bmatrix} \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} \mathbf{x} \\ \mathbf{w} \end{bmatrix}.$$

We wish to compute the largest magnitude eigenvalues of $\mathbf{A}$, which by (5) are the eigenvalues $\theta_{m+1}, \ldots, \theta_{2m}$ of $\mathbf{T}$. Apply k steps of the Arnoldi process, resulting in the same Ritz values obtained in step 1.

4. The exact shift procedure will select Ritz values $\theta_{m+1}, \ldots, \theta_k$ as the roots of the filter polynomial. These shifts include the eigenvalues of $\mathbf{A}$ we seek, which are thus deflated from $\mathbf{v}$ and hence cannot be recovered by the restarted iteration.

Examples of this sort rely on Ritz values that fall well beyond all the eigenvalues of $\mathbf{D}$. If that matrix is normal, then all its Ritz values would be contained in the convex hull of its spectrum. Hence no normal $\mathbf{D}$ will give Ritz values larger in magnitude than (or further to the right of) its eigenvalues, and consequently no such matrix is suitable for use in the above construction. Nonnormality thus plays a central role in examples of this form. Both the example of section 2 and the one we shall next present have been designed so that the eigenvalues of interest are not unduly influenced by this nonnormality and hence would likely be meaningful in applications (e.g., $\mathbf{D}$ may be far from normal, but the desired eigenvalues of $\mathbf{A}$ are well separated from $\mathbf{D}$). As in [4, section 4.4], nonnormality associated with undesired eigenvalues complicates convergence to desired, ideally conditioned eigenvalues.

4. Larger example. Next we follow the procedure just described to produce a scenario that perhaps appears less contrived than the 4×4 example of section 2. To begin, consider the upper triangular matrix $\mathbf{D} = \mathbf{\Lambda} + \beta\mathbf{S}^\gamma$, where

$$\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad \mathbf{S} = \begin{bmatrix} 0 & s_1 & & \\ & 0 & \ddots & \\ & & \ddots & s_{n-1} \\ & & & 0 \end{bmatrix},$$

with

$$\lambda_j = -\alpha + \frac{2\alpha(j-1)}{n-1}, \qquad s_j = \frac{j}{n-1}$$

for parameters $\alpha, \beta, \gamma \geq 0$. The spectrum of this matrix is uniformly distributed over the interval $[-\alpha, \alpha]$. For $\alpha > 0$ the matrix is diagonalizable, with β and γ controlling the conditioning of the eigenvalues. Qualitatively, the growth of s_j with j causes eigenvalues on the right end of the spectrum to exhibit greater sensitivity than those on the left. Increasing β magnifies the ill-conditioning throughout the spectrum, while increasing γ improves the conditioning, most acutely for the leftmost eigenvalues.

In the example that follows, $\mathbf{D}$ has dimension $n = 100$ with $\alpha = 1/2$ and $\beta = \gamma = 4$. Figure 1(a) shows the spectrum, numerical range (2), and ε-pseudospectra (3) of $\mathbf{D}$. The matrix exhibits a moderate departure from normality, with the right half of the spectrum especially sensitive.

Figure 1(b) shows $k = 10$ Ritz values for $\mathbf{D}$ with the starting vector $\mathbf{w} = [1, \ldots, 1]^T$. If we seek $m = 5$ eigenvalues, then the exact shift strategy would choose

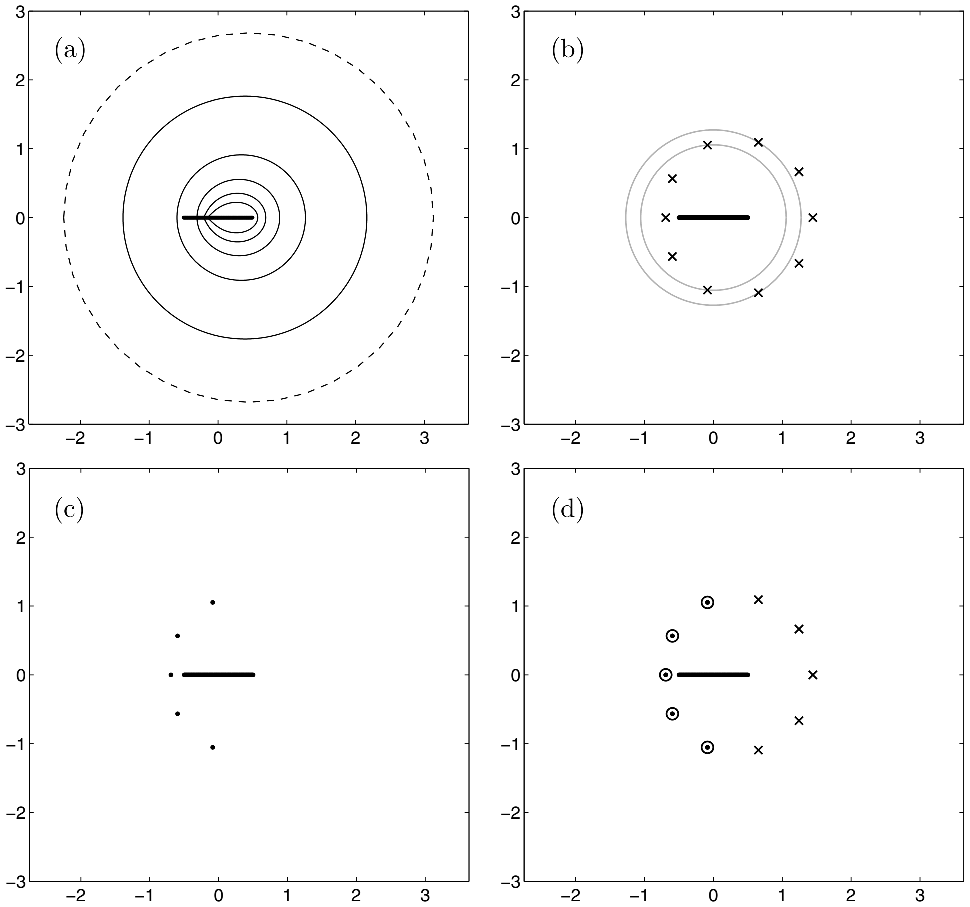


FIG. 1. *Illustration of the example discussed in section* 4. (a) *The eigenvalues (dots covering the segment* $[-1/2, 1/2]$*),* ε*-pseudospectra (boundaries shown as solid lines for* $\varepsilon = 10^{-1}, 10^{-3}, \ldots, 10^{-9}$*), and the numerical range (boundary shown as a dashed line) for* $\mathbf{D}$. (b) *The eigenvalues of* $\mathbf{D}$ *with the* $k = 10$ *Ritz values for* $\mathbf{D}$ *and* $\mathbf{w}$ *(crosses); the gray circles indicate the magnitude of the fifth and sixth largest Ritz values.* (c) *The eigenvalues of* $\mathbf{A}$. (d) *The eigenvalues of* $\mathbf{A}$ *(dots) with* $k = 10$ *Ritz values for* $\mathbf{A}$ *and* $\mathbf{v}$ *(circles and crosses); the smallest five in magnitude (circles) match the largest magnitude eigenvalues of* $\mathbf{A}$ *and will be used as exact shifts.*

the $k - m = 5$ smallest magnitude Ritz values as roots of the filter polynomial; these Ritz values are the leftmost ones shown in the figure. Comparing this with Figure 1(a), notice that these Ritz values are outside the ε-pseudospectra of $\mathbf{A}$ for all $\varepsilon \leq 10^{-3}$; equivalently, the “sep” of all these Ritz values from $\mathbf{D}$ is larger than 10^{-3}.

We now follow the recipe outlined above to obtain an $\mathbf{A}$ and $\mathbf{v}$ for which these shifts would be catastrophic. Let $\mathbf{T} \in \mathbb{C}^{5\times5}$ be a diagonal matrix whose eigenvalues equal those five smallest magnitude Ritz values, and set $\mathbf{A}$ to be the 105×105 matrix $\mathbf{A} = \mathrm{diag}(\mathbf{T}, \mathbf{D})$; the eigenvalues of $\mathbf{A}$ are shown in Figure 1(c). By design, the eigenvalues of $\mathbf{T}$ are the largest magnitude eigenvalues of $\mathbf{A}$, and they all fall beyond the $\varepsilon = 10^{-3}$ pseudospectrum of $\mathbf{D}$.

Now compute $k = 10$ Ritz values for $\mathbf{A}$ with $\mathbf{v} = [\mathbf{x}^T\ \mathbf{w}^T]^T$ for any choice of $\mathbf{x}$, shown in Figure 1(d). (The computation to produce this illustration used $\mathbf{x} = [1, \ldots, 1]^T$, but that choice has no influence on the figure.) As ensured by the theorem, these Ritz values are identical to those obtained from $\mathbf{D}$ and $\mathbf{w}$. To compute the largest magnitude eigenvalues of $\mathbf{A}$, the exact shift strategy picks as shifts the five smallest Ritz values—which coincide with the five largest magnitude eigenvalues of $\mathbf{A}$. Again, the exact shift procedure will deflate precisely those eigenvalues we wish to compute.

5. Discussion. Our aim here has been to address a theoretical question concerning the convergence of the restarted Arnoldi algorithm with exact shifts. In no way do we suggest that the behavior our examples exhibit is commonplace. The exact shift procedure remains the most robust general-purpose method for restarting the Arnoldi method, endorsed by years of successful computation. Our constructions rely on special choices for the starting vector (the components in $\mathbf{w}$) and a fixed number of steps; by changing $\mathbf{w}$, or increasing or decreasing k, one may well converge to the correct eigenvalues without incident. Even with problematic $\mathbf{w}$ and k, the rounding errors that occur in practical computations can allow the desired eigenvectors to emerge after numerous additional iterations. (Indeed, when applied to the example in section 4, `eigs` eventually finds the five desired eigenvalues.)

On those occasions when ARPACK fails to converge, the culprit is likely more mundane than the extreme failure exhibited here. For example, the requested eigenvalues may form part of a tight cluster, in which case any shift procedure would struggle to develop, in a tractable number of iterations, a polynomial filter that is large on the desired eigenvalues while being small on those nearby undesired eigenvalues. Such behavior is explained by the conventional restarted Arnoldi convergence analysis cited in the introduction.

The mere existence of examples for which the restarted Arnoldi algorithm with exact shifts deflates the desired eigenvalues underscores the need for a deeper understanding of the behavior of Ritz values of non-Hermitian matrices. Though these eigenvalue estimates must fall within the numerical range, little else is known, deterministically or stochastically, about their distribution. Progress on this important problem could provide a foundation for a robust convergence theory for the restarted Arnoldi algorithm and illuminate many other corners of iterative linear algebra.

Acknowledgments. I thank Rich Lehoucq for drawing my attention to Ruhe's paper [12], and Howard Elman for his kind hospitality during a visit to the University of Maryland, where this work was begun. The pseudospectra in the figure were computed using the EigTool package of T. G. Wright [18].

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] Z. Bai, J. Demmel, and A. McKenney, *On computing condition numbers for the nonsymmetric eigenproblem*, ACM Trans. Math. Software, 19 (1993), pp. 202–223.

[3] C. Beattie, M. Embree, and J. Rossi, *Convergence of restarted Krylov subspaces to invariant subspaces*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 1074–1109.

[4] C. A. Beattie, M. Embree, and D. C. Sorensen, *Convergence of polynomial restart Krylov methods for eigenvalue computations*, SIAM Review, 47 (2005), pp. 492–515.

[5] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[6] Z. Jia, *The convergence of generalized Lanczos methods for large unsymmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 843–862.

[7] L. Knizhnerman, *Error bounds for the Arnoldi method: A set of extreme eigenpairs*, Linear Algebra Appl., 296 (1999), pp. 191–211.

[8] R. B. Lehoucq, *Implicitly restarted Arnoldi methods and subspace iteration*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 551–562.

[9] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998.

[10] K. Meerbergen, A. Spence, and D. Roose, *Shift-invert and Cayley transforms for detection of rightmost eigenvalues of nonsymmetric matrices*, BIT, 34 (1994), pp. 409–423.

[11] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[12] A. Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, Linear Algebra Appl., 58 (1984), pp. 391–405.

[13] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[14] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[15] G. W. Stewart, *A Krylov–Schur algorithm for large eigenproblems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 601–614.

[16] G. W. Stewart, *Matrix Algorithms Volume* II: *Eigensystems*, SIAM, Philadelphia, 2001.

[17] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.

[18] T. G. Wright, *EigTool*, 2002. Software available at http://www.comlab.ox.ac.uk/pseudospectra/eigtool.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 1, pp. 11–17

RELATIVE PERTURBATION BOUNDS FOR EIGENVALUES OF SYMMETRIC POSITIVE DEFINITE DIAGONALLY DOMINANT MATRICES[*]

QIANG YE[†]

Abstract. For a symmetric positive semidefinite diagonally dominant matrix, if its off-diagonal entries and its diagonally dominant parts for all rows (which are defined for a row as the diagonal entry subtracted by the sum of absolute values of off-diagonal entries in that row) are known to a certain relative accuracy, we show that its eigenvalues are known to the same relative accuracy. Specifically, we prove that if such a matrix is perturbed in a way that each off-diagonal entry and each diagonally dominant part have relative errors bounded by some ϵ, then all its eigenvalues have relative errors bounded by ϵ. The result is extended to the generalized eigenvalue problem.

Key words. relative perturbation, eigenvalues, diagonal dominant matrix

AMS subject classifications. 65F35, 15A42

DOI. 10.1137/060676349

1. Introduction.

The study of relative perturbation theory and high relative accuracy algorithms has been a subject of great interest for many years; see [7, 12, 13] for an overview. For the matrix eigenvalue or singular value problems, by restricting perturbations to those that preserve certain structure and are small entrywise, the perturbation bounds could be strengthened, and thus even some small singular values or eigenvalues can be guaranteed to have small relative perturbations; see [5, 6, 7, 9, 10, 14, 16, 15, 19] for some of the references. We note that such results can only be established by considering matrices perturbed within certain classes, and in some cases, the matrices may need to be reparameterized; see [7, 9, 14] for examples.

In this paper, we develop a relative perturbation theory for eigenvalues of symmetric positive semidefinite diagonally dominant matrices (or symmetric diagonally dominant matrices with nonnegative diagonals). Diagonally dominant matrices arise in a large variety of applications and form one of the most well-studied classes of matrices; see [18] for some recent interest. While the property of diagonal dominance has traditionally been used more in solving linear systems, in recent years, this is also exploited for eigenvalue computations. In [3], Barlow and Demmel develop entrywise perturbation analysis and algorithms for the eigenvalues of symmetric scaled diagonally dominant matrices. Their perturbation results [3] show that the relative perturbations on eigenvalues, when each entry of the matrix has small relative perturbation, depend on a condition number, which is essentially related to the diagonal dominance. In [1, 2], Alfa, Xue, and Ye show that the smallest eigenvalue of a diagonally dominant M-matrix is determined and can be computed to high relative accuracy without any condition number if the row sums (i.e., the diagonally dominant parts) are known to high relative accuracy. Under the same assumptions, Demmel and Koev [8] show that all singular values of a diagonally dominant M-matrix are determined and can be computed to high relative accuracy. More refined results on

[*] Received by the editors November 30, 2006; accepted for publication (in revised form) by F. Dopico June 22, 2007; published electronically February 27, 2009. This research was supported in part by NSF grant DMS-0411502.

http://www.siam.org/journals/simax/31-1/67634.html

[†] Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 (qye@ms.uky.edu).

diagonally dominant M-matrices are given by Peña [17]. Other related perturbation results include those for M-matrices by Elsner [11], Xue [21], and Xue and Jiang [20], which all contain some condition numbers. Note that an M-matrix can be scaled to become a diagonally dominant M-matrix.

Here, we shall prove that if a symmetric positive semidefinite diagonally dominant matrix $A = [a_{ij}]$ is perturbed symmetrically with each off-diagonal entry a_{ij} and each diagonally dominant part ($v_i := a_{ii} - \sum_{j \neq i} |a_{ij}|$) having relative error bounded by ϵ, then the relative error of each eigenvalue is bounded exactly by ϵ. We shall also extend our results to the generalized eigenvalue problem. Compared with the results of [3], our perturbation bound is independent of any condition number. Compared with that of [2, 8], we do not require the matrix to be an M-matrix, but rather we require symmetry. Also our bound is sharp and is valid for all eigenvalues.

We remark that the key to obtain the strong bound is to consider the diagonal dominant parts, replacing the diagonal entries, as the parameters representing such matrices. Namely, the eigenvalues may not be determined to high relative accuracy by the entries of A, but they are so determined by its off-diagonal entries and the diagonal dominant parts. This parameterization is originally introduced in Alfa, Xue, and Ye [1, 2] for diagonally dominant M-matrices. We concentrate on the perturbation theory in this paper but consider algorithms that compute all eigenvalues to the order of machine precision in a separate work [22].

The rest of this paper is organized as follows. We first give in section 2 some definitions and preliminary results. We then present the perturbation results in section 3.

2. Preliminaries and notation.

Throughout this paper, we shall use the following notation. Given a matrix $A = [a_{ij}]$, we use $|A| = [|a_{ij}|]$ and $\operatorname{sign}(A) = [\operatorname{sign}(a_{ij})]$, where $\operatorname{sign}(x)$ denotes the sign of x with $\operatorname{sign}(0) = 1$. Given a vector $v = [v_i]$, $\operatorname{diag}\{v\}$ is the diagonal matrix with the entries of v on its diagonal. $A \geq 0$ denotes that A is symmetric positive semidefinite, and $A \geq B$ denotes that $A - B$ is symmetric positive semidefinite.

The basis of our relative perturbation theory is a reparameterization of the matrices by their off-diagonal entries and their diagonal dominant parts. This is originally introduced in [1, 2] for diagonal dominant M-matrices and can be done for a general matrix as follows.

Definition 2.1. [illegible] $n \times n$ [illegible] $M = [m_{ij}]$ [illegible] n [illegible] $v = [v_i]$ [illegible] $\mathcal{D}(M, v)$ [illegible] $A = [a_{ij}]$ [illegible] ff [illegible] M [illegible] i [illegible] $a_{ii} = v_i + \sum_{j \neq i} |m_{ij}|$ [illegible]

$$A = \mathcal{D}(M, v) \tag{2.1}$$

[illegible] A [illegible] v [illegible]

$$a_{ij} = m_{ij} \ [illegible] \ i \neq j; \ [illegible] \quad a_{ii} = v_i + \sum_{j \neq i} |m_{ij}|.$$

Note that the diagonal entries of M, if given, are not used in defining the matrix $\mathcal{D}(M, v)$. Now, given a matrix $A = [a_{ij}]$, we denote by A_D the matrix whose off-diagonal entries are the same as A and whose diagonal entries are zero. Then, letting $v_i = a_{ii} - \sum_{j \neq i} |a_{ij}|$ and $v = (v_1, v_2, \ldots, v_n)^T$, we have

$$A = \mathcal{D}(A_D, v)$$

as the representation of A by diagonally dominant parts. In this way, the parameters defining A are those of A_D (i.e., the off-diagonal entries of A) and v (i.e., the diagonally dominant parts). The diagonal entries are not used to define A in this representation.

DEFINITION 2.2. [illegible] $A = [a_{ij}]$ [illegible] $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$ [illegible] i

The diagonally dominant matrix defined here is sometimes referred to as being weakly diagonally dominant. A matrix represented in $A = \mathcal{D}(A_D, v)$ is diagonally dominant with nonnegative diagonals if and only if $v_i \geq 0$ for all i.

Throughout we shall consider symmetric matrices $\mathcal{D}(A_D, v)$ with $v_i \geq 0$ for all i, i.e., symmetric diagonally dominant matrices with nonnegative diagonals. Clearly, a matrix is symmetric diagonally dominant with nonnegative diagonals if and only if it is symmetric positive semidefinite and diagonally dominant.

3. Relative perturbation bounds. We present relative perturbation bounds for symmetric positive semidefinite diagonally dominant matrices, which are represented in the form $A = \mathcal{D}(A_D, v)$ with $v_i \geq 0$ for all i.

We first introduce some notation. For $A = [a_{ij}]$, let

$$T_i = \operatorname{diag}(a_{1,i+1}, a_{2,i+2}, \ldots, a_{n-i,n}), \tag{3.1}$$

$$D_i = \begin{array}{c} \begin{array}{cc} \scriptstyle n-i & \scriptstyle i \end{array} \\ \left(\begin{array}{cc} |T_i| & 0 \\ 0 & 0 \end{array} \right) \end{array} \begin{array}{c} \scriptstyle n-i \\ \scriptstyle i \end{array}, \quad N_i = \begin{array}{c} \begin{array}{cc} \scriptstyle n-i & \scriptstyle i \end{array} \\ \left(\begin{array}{cc} 0 & 0 \\ T_i & 0 \end{array} \right) \end{array} \begin{array}{c} \scriptstyle i \\ \scriptstyle n-i \end{array}, \tag{3.2}$$

and

$$M_i = \operatorname{sign}(N_i), \quad L_i = I + M_i. \tag{3.3}$$

In other words, T_i is the diagonal matrix whose diagonal is given by the ith diagonal of A above the main diagonal; D_i is the $n \times n$ diagonal matrix whose diagonals are absolute values of T_i, extended by zeros; and N_i is the matrix obtained from A by striking out all entries except the ith diagonal below the main diagonal. Clearly, L_i is lower triangular with ± 1 on the ith diagonal.

We first give a lemma that gives an LDL^T factorization of a symmetric banded matrix that has a single band and zero diagonal dominant part. A feature of this factorization is that the parameters defining the matrix are entirely contained in the D matrix.

LEMMA 3.1. [illegible] A_i [illegible] i [illegible] $a_{1,i+1}, a_{2,i+2}, \ldots, a_{n-i,n}$ [illegible] j [illegible] ff [illegible] j [illegible] $A_i = \mathcal{D}(N_i + N_i^T, 0)$ [illegible] (2.1) [illegible] $A_i = L_i D_i L_i^T$

[illegible] Since the diagonally dominant part is 0, we can write

$$A_i = N_i + N_i^T + \begin{pmatrix} 0 & 0 \\ 0 & |T_i| \end{pmatrix} + \begin{pmatrix} |T_i| & 0 \\ 0 & 0 \end{pmatrix}.$$

We have

$$\begin{aligned} L_i D_i L_i^T &= (I + M_i) D_i (I + M_i^T) \\ &= D_i + M_i D_i + D_i M_i^T + M_i D_i M_i^T \\ &= D_i + N_i + N_i^T + N_i M_i^T = A_i, \end{aligned}$$

where we note that

$$N_i M_i^T = \begin{pmatrix} 0 & 0 \\ 0 & |T_i| \end{pmatrix}. \qquad \square$$

There is a similar result for the case $i = 1$ presented in [4, Example 5.1, p. 196], which was the inspiration of this work. We next decompose A into a sum of banded matrices with a single band and hence, using the above lemma, decompose it into a sum of LDL^T factorizations.

Lemma 3.2. [illegible] $A = [a_{ij}]$ [illegible] $A = \mathcal{D}(A_D, v)$ [illegible] (2.1) [illegible] $v = [v_1, v_2, \ldots, v_n]^T$ [illegible]

$$A = V_0 + L_1 D_1 L_1^T + \cdots + L_{n-1} D_{n-1} L_{n-1}^T, \tag{3.4}$$

[illegible] $V_0 = \operatorname{diag}\{v_1, \ldots, v_n\}$ [illegible] L_i D_i [illegible] $1 \le i \le n-1$ [illegible] (3.3) [illegible] (3.2)

[illegible] Let $A_i = \mathcal{D}(N_i + N_i^T, 0)$ as defined in Lemma 3.1. Then the off-diagonal entries of $\sum_{i=1}^{n-1} A_i$ are the same as those of A. Since each A_i has zero diagonal dominant part, it is easy to see that $\sum_{i=1}^{n-1} A_i$ also has zero diagonal dominant part. Thus, we have

$$\sum_{i=1}^{n-1} A_i = \mathcal{D}(A, 0).$$

Now, with O denoting the zero matrix, we have

$$A = \mathcal{D}(O, v) + \mathcal{D}(A, 0) = V_0 + \sum_{i=1}^{n-1} A_i,$$

and hence (3.4) follows from Lemma 3.1. □

To clearly see the decomposition (3.4), we show an example of 3 × 3 matrix $A = \mathcal{D}(A_D, v)$ as

$$\begin{aligned}
A &= \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\
&= \begin{pmatrix} v_1 & & \\ & v_2 & \\ & & v_3 \end{pmatrix} + \begin{pmatrix} |a_{21}| + |a_{31}| & a_{21} & a_{31} \\ a_{21} & |a_{21}| + |a_{32}| & a_{32} \\ a_{31} & a_{32} & |a_{31}| + |a_{32}| \end{pmatrix} \\
&= \begin{pmatrix} v_1 & & \\ & v_2 & \\ & & v_3 \end{pmatrix} + \begin{pmatrix} 1 & & \\ s_{21} & 1 & \\ & s_{32} & 1 \end{pmatrix} \begin{pmatrix} |a_{21}| & & \\ & |a_{32}| & \\ & & 0 \end{pmatrix} \begin{pmatrix} 1 & s_{21} & \\ & 1 & s_{32} \\ & & 1 \end{pmatrix} \\
&\quad + \begin{pmatrix} 1 & & \\ 0 & 1 & \\ s_{31} & 0 & 1 \end{pmatrix} \begin{pmatrix} |a_{31}| & & \\ & 0 & \\ & & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & s_{31} \\ & 1 & 0 \\ & & 1 \end{pmatrix},
\end{aligned}$$

where $s_{ij} = \operatorname{sign}(a_{ij})$.

We are now ready to present our perturbation results.

Theorem 3.3. [illegible] $A = [a_{ij}]$ [illegible] $\widetilde{A} = [\widetilde{a}_{ij}]$ [illegible] $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ [illegible] $\widetilde{\lambda}_1 \le \widetilde{\lambda}_2 \le \cdots \le \widetilde{\lambda}_n$ [illegible] $0 \le \epsilon < 1$

$$|a_{ij} - \widetilde{a}_{ij}| \le \epsilon |a_{ij}| \quad \text{[illegible]} \quad i \ne j \tag{3.5}$$

[illegible]

$$|v_i - \widetilde{v}_i| \le \epsilon v_i \quad \text{[illegible]}\ i, \tag{3.6}$$

[illegible] $v_i = a_{ii} - \sum_{j\neq i}|a_{ij}|$ [illegible] $\widetilde{v}_i = \widetilde{a}_{ii} - \sum_{j\neq i}|\widetilde{a}_{ij}|$ [illegible] $A = \mathcal{D}(A_D, v)$ [illegible] $\widetilde{A} = \mathcal{D}(\widetilde{A}_D, \widetilde{v})$ [illegible] (2.1) [illegible] A [illegible] $\widetilde{A}$ [illegible] $v = [v_i]$ [illegible] $\widetilde{v} = [\widetilde{v}_i]$ [illegible] i

$$|\widetilde{\lambda}_i - \lambda_i| \le \epsilon\lambda_i. \tag{3.7}$$

[illegible]. Let D_i, L_i and $\widetilde{D}_i, \widetilde{L}_i$ be the matrices defined from A and $\widetilde{A}$, respectively, according to (3.2) and (3.3). By (3.5), a_{ij} and $\widetilde{a}_{ij}$ have the same sign. Since L_i and $\widetilde{L}_i$ are defined from the signs of a_{ij} and $\widetilde{a}_{ij}$, we have

$$L_i = \widetilde{L}_i.$$

Furthermore, it follows from (3.5) and (3.6) that

$$(1-\epsilon)|a_{ij}| \le |\widetilde{a}_{ij}| \le (1+\epsilon)|a_{ij}|, \quad (1-\epsilon)v_i \le \widetilde{v}_i \le (1+\epsilon)v_i,$$

where we note that $v_i \ge 0$ and $\widetilde{v}_i \ge 0$ by assumption. This leads to

$$(1-\epsilon)D_i \le \widetilde{D}_i \le (1+\epsilon)D_i, \quad (1-\epsilon)V_0 \le \widetilde{V}_0 \le (1+\epsilon)V_0,$$

where $V_0 = \mathrm{diag}(v)$ and $\widetilde{V}_0 = \mathrm{diag}(\widetilde{v})$. Now, applying Lemma 3.2, we have

$$A = V_0 + L_1 D_1 L_1^T + \cdots + L_{n-1} D_{n-1} L_{n-1}^T$$

and

$$\widetilde{A} = \widetilde{V}_0 + L_1 \widetilde{D}_1 L_1^T + \cdots + L_{n-1} \widetilde{D}_{n-1} L_{n-1}^T.$$

Thus

$$(1-\epsilon)A \le \widetilde{A} \le (1+\epsilon)A, \tag{3.8}$$

from which it follows that $(1-\epsilon)\lambda_i \le \widetilde{\lambda}_i \le (1+\epsilon)\lambda_i$. The theorem is proved. □

[illegible] 1. From the assumptions (3.5) and (3.6), we have that $|\widetilde{a}_{ii} - a_{ii}| \le \epsilon |a_{ii}|$; see [2]. The converse is not true; namely, $|\widetilde{a}_{ij} - a_{ij}| \le \epsilon|a_{ij}|$ for all i, j does not imply $|v_i - \widetilde{v}_i| \le \epsilon v_i$.

[illegible] 2. Our bound is sharp. This can be verified by considering a diagonal A.

[illegible] 3. We have assumed in the theorem that both A and $\widetilde{A}$ are symmetric positive semidefinite and diagonally dominant. However, we can assume only that A is symmetric positive semidefinite and diagonally dominant and that $\widetilde{A}$ is symmetric and satisfies (3.6), which imply that $\widetilde{A}$ is also positive semidefinite and diagonally dominant, because it follows from (3.6) that $\widetilde{v}_i \ge 0$.

The above theorem can be easily generalized to the definite symmetric pencil eigenvalue problem $Ax = \lambda Bx$.

THEOREM 3.4. [illegible] $A = [a_{ij}]$ [illegible] $\widetilde{A} = [\widetilde{a}_{ij}]$ [illegible] $B = [b_{ij}]$ [illegible] $\widetilde{B} = [\widetilde{b}_{ij}]$ [illegible] $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ [illegible] $\widetilde{\lambda}_1 \le \widetilde{\lambda}_2 \le \cdots \le \widetilde{\lambda}_n$ [illegible] $A - \lambda B$ [illegible] $\widetilde{A} - \lambda\widetilde{B}$ [illegible] $A = \mathcal{D}(A_D, v)$

$B = \mathcal{D}(B_D, w)$ $\widetilde{A} = \mathcal{D}(\widetilde{A}_D, \widetilde{v})$ [illegible] $\widetilde{B} = \mathcal{D}(\widetilde{B}_D, \widetilde{w})$ [illegible] (2.1) [illegible] $v = [v_i]$ $w = [w_i]$ $\widetilde{v} = [\widetilde{v}_i]$ [illegible] $\widetilde{w} = [\widetilde{w}_i]$ [illegible] $0 \le \epsilon, \epsilon' < 1$

$$|a_{ij} - \widetilde{a}_{ij}| \le \epsilon |a_{ij}|, \quad |v_i - \widetilde{v}_i| \le \epsilon v_i$$

[illegible]

$$|b_{ij} - \widetilde{b}_{ij}| \le \epsilon' |b_{ij}|, \quad |w_i - \widetilde{w}_i| \le \epsilon' w_i,$$

[illegible] $i \neq j$ [illegible] i

$$|\widetilde{\lambda}_i - \lambda_i| \le \frac{\epsilon + \epsilon'}{1 - \epsilon'} \lambda_i.$$

[illegible]. As in the proof of Theorem 3.3, we have the bound (3.8) for $\widetilde{A}$ and the following corresponding bound for $\widetilde{B}$:

$$(1 - \epsilon')B \le \widetilde{B} \le (1 + \epsilon')B. \tag{3.9}$$

Now, using the minimax theorem, we obtain

$$\frac{1 - \epsilon}{1 + \epsilon'} \lambda_i \le \widetilde{\lambda}_i \le \frac{1 + \epsilon}{1 - \epsilon'} \lambda_i,$$

which leads to the theorem. □

The theorems show that if the data $\mathcal{D}(A_D, v)$ representing A is known to a certain relative accuracy, then its eigenvalues are determined to the same relative accuracy. This is even true for the zero eigenvalue. For the smallest eigenvalue of a diagonally dominant M-matrix, our result improves the perturbation bound in [2], where only $|\lambda - \widetilde{\lambda}|/\lambda \le (2n-1)\epsilon + O(\epsilon^2)$ is obtained. It can be applied then to the electronic circuit application as in [2] to obtain significantly improved perturbation bounds on the circuit speed. The improvement is achieved of course with the condition that the matrix is symmetric.

Acknowledgments. I would like to thank Prof. Jim Demmel for many valuable discussions. This work was inspired by an illuminating example in his book [4]. I would also like to thank an anonymous referee and Prof. Froilan Dopico for their very careful reading of the paper and numerous comments that have improved the presentation of the paper.

REFERENCES

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

[2] A. S. Alfa, J. Xue, and Q. Ye, *Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix*, Math. Comp., 71 (2002), pp. 217–236.

[3] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

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

[5] J. Demmel and W. Gragg, *On computing accurate singular values and eigenvalues of matrices with acyclic graphs*, Linear Algebra Appl., 185 (1993), pp. 203–217.

[6] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[7] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[8] J. W. Demmel and P. Koev, *Accurate SVDs of weakly diagonally dominant M-matrices*, Numer. Math., 98 (2004), pp. 99–104.

[9] F. M. Dopico and P. Koev, *Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1126–1156.

[10] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[11] L. Elsner, *Bounds for determinants of perturbed M-matrices*, Linear Algebra Appl., 257 (1997), pp. 283–288.

[12] N. J. Higham, *A survey of componentwise perturbation theory in numerical linear algebra*, in Mathematics of Computation 1943–1993, Proceedings of Symposia in Applied Mathematics 48, W. Gautschi, ed., AMS, Providence, RI, 1994, pp. 49–77.

[13] I. C. F. Ipsen, *Relative perturbation bounds for matrix eigenvalues and singular values*, in Acta Numerica 1998, Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 151–201.

[14] P. Koev, *Accurate eigenvalues and SVDs of totally nonnegative matrices*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 1–23.

[15] R.-C. Li, *Relative perturbation theory:* I. *Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.

[16] R. C. Li, *Relative perturbation theory:* (III) *More bounds on eigenvalue variation*, Linear Algebra Appl., 266 (1997), pp. 337–345.

[17] J. M. Peña, *LDU decompositions with L and U well conditioned*, Electron. Trans. Numer. Anal., 18 (2004), pp. 198–208.

[18] D. A. Spielman and S. Teng, *Solving sparse, symmetric, diagonally-dominant linear systems in time $O(m^{1.31})$*, in Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Los Alamitos, CA, 2003, pp. 416–427.

[19] K. Veselic and I. Slapnicar, *Floating point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

[20] J. Xue and E. Jiang, *Entrywise relative perturbation theory for nonsingular M-matrices and applications*, BIT, 35 (1995), pp. 417–427.

[21] J. Xue, *Computing the smallest eigenvalue of an M-matrix*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 748–762.

[22] Q. Ye, *Computing singular values of diagonally dominant matrices to high relative accuracy*, Math. Comp., to appear.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 18–39

AN EFFICIENT METHOD FOR ESTIMATING THE OPTIMAL DAMPERS' VISCOSITY FOR LINEAR VIBRATING SYSTEMS USING LYAPUNOV EQUATION*

NINOSLAV TRUHAR† AND KREŠIMIR VESELIĆ‡

Abstract. This paper deals with an efficient algorithm for dampers' viscosity optimization in mechanical systems. Our algorithm optimizes the trace of the solution of the corresponding Lyapunov equation using an iterative method which calculates a low rank Cholesky factor for the solution of the corresponding Lyapunov equation. We have shown that the new algorithm calculates the trace in $\mathcal{O}(m)$ flops per iteration, where m is a dimension of matrices in the Lyapunov equation (our coefficient matrices are treated as dense).

Key words. damped vibration, Lyapunov equation, optimization of viscosities of dampers

AMS subject classifications. 70J25, 70J50, 11D04, 15A06, 15A24, 90C31

DOI. 10.1137/070683052

1. Introduction. This paper can be considered as a certain continuation of the paper [16]. In [16] we derived some new estimates for the eigenvalue decay rate of the Lyapunov equation $AX + XA^T = B$ with a low rank right-hand side B; we also proposed a new choice of the ADI parameters for calculating X. All this was based on newly established bounds on the trace of a solution to the Lyapunov equation with a general stable coefficient matrix. The trace itself was calculated from the solution of the Lyapunov equation which has been obtained using low rank Cholesky factor ADI (LRCF-ADI) proposed in [12], [8].

In this paper we use the results from [16] to develop an efficient algorithm for dampers' viscosity optimization in mechanical systems. Our penalty function is the trace of the Lyapunov solution X (advantages of this choice were discussed in [4], [17], [18]). Our main issue here is to calculate [illegible] and not the whole solution of the Lyapunov equation with obvious computational advantages.

We consider a damped linear vibrational system described by the differential equation

$$M\ddot{x} + D\dot{x} + Kx = 0, \tag{1.1}$$

$$x(0) = x_0, \quad \dot{x}(0) = \dot{x}_0, \tag{1.2}$$

where M, D, K (called mass, damping, stiffness matrices, respectively) are real, symmetric matrices of order n with M, K positive definite and $D = C_u + C$, where C_u is positive definite and represents the internal damping, which is usually taken to be a small multiple of the critical damping; that is,

$$C_u = \alpha C_{crit},$$

[*] Received by the editors February 20, 2007; accepted for publication (in revised form) by F. Dopico October 22, 2007; published electronically February 27, 2009.
http://www.siam.org/journals/simax/31-1/68305.html

[†] Department of Mathematics, J.J. Strossmayer University of Osijek, Trg Ljudevita Gaja 6, 31000 Osijek, Croatia (ntruhar@mathos.hr). Part of this work was done while this author was visiting the Department of Mathematics, University of Kentucky, Lexington, Kentucky, under the support of a postdoctoral research award from the National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia.

[‡] Lehrgebiet Mathematische Physik, Fernuniversität, 58084 Hagen, Germany (kresimir.veselic@fernuni-hagen.de).

$$C_{crit} = 2M^{1/2}\sqrt{M^{-1/2}KM^{-1/2}}M^{1/2}, \quad \alpha = 2\text{–}10\%$$

(see [11, pp. 26, 260]), and C is positive semidefinite.

A very important question arises in considerations of such systems: [illegible]

This optimization problem has been recently considered in [17], [14], [4], [15].

For such optimization one can use different optimization criteria (see [15]). One of the frequently used criteria is the so-called spectral abscissa criterion, which requires that a maximal real part of the eigenvalues λ_k be minimal, symbolically

$$sp := \max_k \mathrm{Re}\lambda_k \to \min, \tag{1.3}$$

where λ_k are the complex eigenvalues of the system

$$\left(\lambda^2 M + \lambda D + K\right) x = 0, \tag{1.4}$$

obtained from (1.1), simply using the substitution $x(t) = e^{\lambda t}x$.

For example, this criterion was used in [5] and [7]. In [5] a nice result on optimal damping was presented, but as the authors pointed out in section 4, "the only situation for which it is feasible to compute explicitly all possible solutions of the optimization problem (2.3) by hand is when n equals 2," which means that they present an exact form of the optimal damping matrix for systems that are 2×2. In [7] the problem of optimal damping (optimal dampers' positions, or, more precisely, optimal regions) has been solved for a string vibration.

Another criterion, used in [19], [17], [15], [4], is given by requirement of the minimization of the total energy of the system, that is,

$$\int_0^\infty E(t)\,dt \to \min. \tag{1.5}$$

The advantages of this criterion are (i) its obvious closeness to the total energy of the vibration and (ii) its smoothness as the function of the damping parameters, which allows standard methods of minimization via gradient or Hessian. Note that the latter property is not shared by the spectral penalty function (1.3). On the other hand, Veselić in [20], [21], [22] has shown that the solution of the Lyapunov equation provides rigorous bounds to the energy decay of a vibrating system.

Since the criterion (1.5) depends on the initial condition, the simplest way to correct this is to take the average of (1.5) over all initial states of the unit total energy and a given frequency range. It can be shown that this average is the trace of the solution of the corresponding Lyapunov equation.

A general algorithm for the optimization of damping does not exist. Available algorithms optimize only viscosities of dampers, not their positions. Two types of algorithms are currently in use. The first are the Newton-type algorithms for higher-dimensional (constrained or unconstrained) problems which use some Lyapunov solvers, and the second are the algorithms which explicitly calculate the trace of the solution of the corresponding Lyapunov equation.

An algorithm of the second type was presented in [19] for the case when $C_u = 0$ and the rank of the matrix C is one. Moreover, in [19] Veselić has given an efficient algorithm which calculates an optimal v, where $C = vcc^*$, and the optimal viscosity is given by a closed formula.

On the other hand, in [15] a Newton-type algorithm which calculates optimal viscosity v has been proposed. This algorithm covers the case with internal damping ($C_u \neq 0$) with $C = vC_0C_0^*$, where $r \equiv rank(C_0) > 1$; it calculates the trace of the solution of the corresponding Lyapunov equation as a function of viscosity v of dampers in $\mathcal{O}(r^3m^3)$ flops, where $m = 2n$ (dimension of the phase space). This means that if the number of degrees of freedom of dampers r is much less than n ($r = 2, 3, 4$), this algorithm can be more efficient than the standard methods which use Lyapunov solvers such as, e.g., Bartels–Stewart, which cost $\mathcal{O}(m^3)$ operations per iteration.

Unfortunately, all existing algorithms calculate the solution of the Lyapunov equation and do not take advantage of the fact that we need only the trace of the solution.

Thus, we propose a different approach for optimization of the trace of the solution of the corresponding Lyapunov equation. Our algorithm calculates only the trace of the solution of the Lyapunov equation using an iterative method for an LRCF of the solution of the corresponding Lyapunov equation. This fact allows a more efficient memory usage. Further, in the case when only a small part of undamped spectra (say, the first s smallest undamped eigenvalues) is dominant, our algorithm needs $\mathcal{O}(r^3)+\mathcal{O}(r\,m)+\mathcal{O}(s^3)$ flops per iteration. Since standard optimization processes, such as the golden section search (which has been implemented in the MATLAB function `fminbnd`), need 20–30 iterations, if $r \ll n$ and $s \ll n$, our algorithm minimizes the trace of the Lyapunov equation in $\mathcal{O}(r\,m)$ operations.

We also present a new error bound for the trace approximation, which shows that sometimes the structure of the right-hand side of the Lyapunov equation can greatly influence the accuracy of the solution.

This paper is organized as follows. Section 2 describes a mathematical model we will use and three different algorithms for optimization of the trace of the solution of the corresponding Lyapunov equation. Then, section 3 contains the algorithm which calculates the trace of the solution using LRCF-ADI proposed in [8] (we use the algorithm described in [12]). Since the proper choice of ADI parameters is crucial for efficiency of the LRCF-ADI method, we describe two different algorithms for selection of a suboptimal set of ADI parameters. One was proposed by Penzl in [12], and the other was proposed in [16] and is based on the result that the optimal set of ADI parameters in the case of "modal damping" is given by the set of $2s$ eigenvalues of the matrix $\mathbf{A}$ which correspond to $2s$ undamped eigenvalues (for more details, see [16]). In section 4 we present a new error bound for the trace obtained by the new algorithm.

Finally, in the last section we present two examples. The first example illustrates the efficiency and accuracy of the new algorithm with respect to the column rank of the right-hand side of the Lyapunov equation. The second example compares our new algorithm (applied by using two different suboptimal sets of ADI parameters) with algorithms from [18], [2], and [15].

We will use the following notation: matrices written in simple mathematical italic fonts (M, D, or K), for example, will have $\mathcal{O}(n^2)$ nonzero entries. Matrices written in mathematical bold fonts ($\mathbf{A}$, $\mathbf{B}$) will have $\mathcal{O}(m^2)$ nonzero entries, where $m = 2n$. The symbol $\|\cdot\|$ stands for the standard 2-norm, while $\|\cdot\|_F$ denotes a Frobenius norm. $\mathcal{R}(A)$ denotes a column space spanned by the columns of the matrix A.

2. Setting the scene. As described in [15], [14], [17], [4], minimization of the total energy (1.5) is equivalent to minimization of the trace of the solution of the

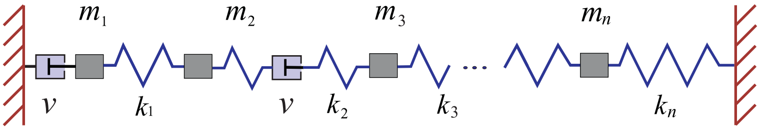


Fig. 2.1. *The n-mass oscillator with two dampers.*

Lyapunov equation. For the sake of completeness, we will shortly describe the basics of this approach.

We consider a damped linear vibrational system described by the differential equation

$$M\ddot{x} + D\dot{x} + Kx = 0, \tag{2.1}$$

where M, C, K (called mass, damping, stiffness matrix, respectively) are real, symmetric matrices of order n with M, K positive definite and $D = C_u + C$ positive semidefinite, where C_u describes internal damping. Often the matrix C has a small rank. An example is the so-called [illegible] or [illegible] (Figure 2.1), where

$$M = \mathrm{diag}(m_1, m_2, \dots, m_n),$$

$$K = \begin{bmatrix} k_0 + k_1 & -k_1 & & & \\ -k_1 & k_1 + k_2 & -k_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -k_{n-2} & k_{n-2} + k_{n-1} & -k_{n-1} \\ & & & -k_{n-1} & k_{n-1} + k_n \end{bmatrix},$$

$$D \equiv C_u + C = C_u + v e_1 e_1^T + v(e_3 - e_2)(e_3 - e_2)^T.$$

Here $m_i > 0$ are the masses, $k_i > 0$ are the spring constants or stiffnesses, e_i is the ith canonical basis vector, and v is the viscosity of the damper applied on the ith mass (in Figure 2.1, $k_0 = 0$). Note that all dampers have the same viscosity and that rank of the matrix C is two. In this paper we study the system with r equal dampers where we assume that $r \ll n$ (usually $r = 2, 3, 4$), which will allow us to use a one-dimensional optimization process (MATLAB function `fminbnd`).

To (2.1) there corresponds the eigenvalue problem

$$(\lambda^2 M + \lambda D + K)x = 0\,. \tag{2.2}$$

Obviously all eigenvalues of (2.2) lie in the left complex plane.

Using the eigenvalue decomposition

$$\Phi^T K \Phi = \Omega^2\,, \quad \Phi^T M \Phi = I, \tag{2.3}$$

where $\Omega = \mathrm{diag}(\omega_1, \dots, \omega_n)$, $\omega_1 < \cdots < \omega_n$, and setting

$$y_1 = \Omega\,\Phi^T x, \quad y_2 = \Phi^T \dot{x}, \tag{2.4}$$

(2.1) can be written as

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \tag{2.5}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & \Omega \\ -\Omega & -\Phi^T D \Phi \end{bmatrix} \tag{2.6}$$

(we are now in a $2n$-dimensional phase space), with the solution

$$\mathbf{y} = e^{\mathbf{A}t}\,\mathbf{y}_0\,, \quad \text{where } \mathbf{y}_0 \text{ is the initial data.} \tag{2.7}$$

Note that the numbers

$$\omega_1, \omega_2, \ldots, \omega_n \tag{2.8}$$

are the eigenvalues of the corresponding undamped system

$$(\lambda^2 M + K)x = 0,$$

and we call them (undamped) eigenfrequencies of the system.

The eigenvalue problem $\mathbf{A}\mathbf{y} = \lambda\,\mathbf{y}$ is equivalent to (2.2). The energy of the system is given by

$$E(t) = \frac{1}{2}\,\dot{x}(t)^T M \dot{x}(t) + \frac{1}{2}\,x(t)^T K x(t) = \frac{1}{2}\,y^T y.$$

Now (1.5) can be written as

$$\mathbf{y}_0^T \mathbf{X} \mathbf{y}_0 \to \min, \tag{2.9}$$

where

$$\mathbf{X} = \int_0^\infty e^{\mathbf{A}^T t}\, e^{\mathbf{A}t} dt \tag{2.10}$$

is the solution of the Lyapunov equation

$$\mathbf{A}^T \mathbf{X} + \mathbf{X}\mathbf{A} = -\mathbf{I}\,. \tag{2.11}$$

An inconvenience of the criterion (2.9) is its dependence on the initial data $\mathbf{y}_0$. Thus, similarly as in [17], instead of the quantity $\mathbf{y}_0^T \mathbf{X} \mathbf{y}_0$ we are going to take its mean value over all initial data $\mathbf{y}$ with the unit energy $\|\mathbf{y}\|^2$. Therefore, instead of (2.9) we require

$$\int_{\|\mathbf{y}_0\|=1} \mathbf{y}_0^T \mathbf{X} \mathbf{y}_0 \; d\sigma \to \min, \tag{2.12}$$

where $d\sigma$ is a chosen probability measure on the unit sphere $S^{2n} = \{\mathbf{y}_0 \in \mathbb{R}^{2n}; \|\mathbf{y}_0\| = 1\}$.

In [17], [10], and [15] it has been shown that (2.12) is equivalent to

$$Tr(\mathbf{Z}\mathbf{X}) \to \min, \tag{2.13}$$

where $\mathbf{Z}$ is a symmetric positive semidefinite matrix which may be normalized to have a unit trace. If we take for the measure σ the measure generated by the Lebesgue measure on $\mathbb{R}^{2n}$, we obtain $Z = \frac{1}{2n}I$.

Further, it is easy to show that

$$Tr(\mathbf{ZX}) = Tr(\mathbf{Y}),$$

where $\mathbf{Y}$ is a solution of the so-called dual Lyapunov equation

$$\mathbf{AY} + \mathbf{YA}^T = -\mathbf{Z}.$$

The structure of the matrix $\mathbf{Z}$ has been studied in detail in [10], and some of these results are presented in [15].

Throughout this paper we will assume that the matrix $\mathbf{Z}$ has the form

$$\mathbf{Z} = \begin{bmatrix} 0_{t_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0_{t_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0_{t_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_s & 0 \\ 0 & 0 & 0 & 0 & 0 & 0_{t_2} \end{bmatrix}, \tag{2.14}$$

where I_s is the s-dimensional identity matrix and 0_{t_i} is the t_i-dimensional ($i = 1, 2$) zero matrix, where t_1 and s are defined such that eigenfrequencies from (2.8) smaller than ω_{t_1} and greater than ω_{t_1+s} are not dangerous (observe that $t_2 = n - t_1 - s$).

Now we will briefly describe the existing algorithms for optimization (2.13).

In [19] a solution of problem (2.13) has been given in the case when $C_u = 0$ and $\text{rank}(C) = 1$. In particular,

$$Tr(\mathbf{ZX}(v)) = const + \frac{a}{v} + bv, \tag{2.15}$$

where $a, b > 0$ are constants which can be easily calculated (in $O(n)$ flops), which makes it possible to find the minimum explicitly by a simple formula. The case $\text{rank}(C) > 1$ seems to be essentially more difficult to handle.

In [15], problem (2.13) with $C_u \neq 0$ and $\text{rank}(C) > 1$ has been considered. In particular,

$$Tr(\mathbf{ZX}(v)) = -x_0 - v\, b_L^T(\mathbf{I} - v\,\mathbf{H}_s)^{-1} b_R, \tag{2.16}$$

where $\mathbf{H}_s$ denotes the upper Hessenberg matrix for whose construction one needs $\frac{112}{3}\, r^3 m^3 + \mathcal{O}(r^2 m^2)$ operations. Since from (2.16) one can find the first and the second derivative of the function $v \to Tr(\mathbf{ZX}(v))$ almost for free, the whole optimization process costs $\frac{112}{3}\, r^3 m^3 + \mathcal{O}(r^2 m^2)$.

On the other hand, a more general case with the damping matrix

$$D \equiv C_u + C = C_u + C_0 \text{diag}(v_1, \dots, v_r) C_0^T$$

has been considered in [2]. There, a Newton-type algorithm has been proposed, which uses the Bartels–Stewart Lyapunov solver.

As we will see in the last section, each of these algorithms has some advantages in certain situations. But all of them calculate the whole solution at every stage of the iteration and then use only the trace.

As we have mentioned in the introduction, our approach here consists of constructing an efficient algorithm which will derive the trace $Tr(\mathbf{ZX}(v))$ using the LRCF-ADI method, and then find the minimum of the function $v \rightarrowtail Tr(\mathbf{ZX}(v))$ using some standard minimization process such as the golden section search which has been implemented in the MATLAB function `fminbnd`. Since we calculate only the trace and not the whole solution, our algorithm is much faster than existing ones which calculate the whole solution first and then the trace. The next section contains a description of our new algorithm.

3. The main algorithm. As described in the previous section, our aim is to minimize the trace of the Lyapunov equation

$$\mathbf{AX} + \mathbf{XA}^T = -\mathbf{GG}^T, \tag{3.1}$$

where

$$\mathbf{A} \equiv \mathbf{A}_0 - v\,\mathbf{D} = \begin{bmatrix} 0 & \Omega \\ -\Omega & -\alpha\,\Omega^k \end{bmatrix} - v \begin{bmatrix} 0 & 0 \\ 0 & C_0 C_0^T \end{bmatrix}, \tag{3.2}$$

where $\text{rank}(\mathbf{G}) = 2s, s \ll n$, and

$$\mathbf{D} = \mathbf{D}_0\mathbf{D}_0^T, \quad \mathbf{D}_0 = \begin{bmatrix} 0 \\ C_0 \end{bmatrix}, \quad \text{and} \quad C_0 = \Phi^T \left[e_{i_1}, \dots e_{i_r}\right]. \tag{3.3}$$

The vector e_{i_j} is the i_jth canonical basis vector and r is the number of dampers. We assume that $\Omega = \text{diag}(\omega_1, \dots, \omega_n)$, where $\omega_1 < \cdots < \omega_n$.

Note that for $\mathbf{Z}$ defined as in (2.14), we have $\mathbf{Z} = \mathbf{GG}^T$, where

$$\mathbf{G} = \begin{bmatrix} 0 & I_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_s & 0 \end{bmatrix}^T, \tag{3.4}$$

$\mathbf{G} \in \mathbb{R}^{m\times 2s}$, $s \ll n$. This assumption and the fact that the solution of (3.1) is positive definite allow us to use the LRCF-ADI method proposed in [8] (see also [9]) and implemented in [12]. As we will see throughout this paper, the choice of this algorithm for computing the trace of the Lyapunov equation has many advantages. The most important fact is that by using this algorithm one can find the trace of the solution without calculating the whole solution, which can substantially speed up the calculation.

As we have mentioned above, since $s \ll n$, we are going to use the LRCF-ADI algorithm for solving the Lyapunov equation

$$\mathbf{AX} + \mathbf{XA}^T = -\mathbf{GG}^T.$$

The basic code taken from [12] reads as follows.

Algorithm 1 (LRCF-ADI).

```
INPUT: $\mathbf{A}$  $\mathbf{G}$  $\{p_1, p_2, \dots, p_{i_{max}}\}$
    OUTPUT: $\mathbf{V} = \mathbf{V}_{i_{max}} \in \mathbb{C}^{m\times 2\,s\,i_{max}}$ [illegible] $\mathbf{VV}^* \approx \mathbf{X}$
    1  $\mathbf{W}_1 = \sqrt{-2\text{Re}p_1}\,(\mathbf{A} + p_1\mathbf{I})^{-1}\mathbf{G}$
2  $\mathbf{V} = \mathbf{W}_1$
    FOR:  $i = 2, 3, \dots, i_{max}$
3  $\mathbf{W}_i = \sqrt{\text{Re}p_i/\text{Re}p_{i-1}}\,\left(\mathbf{W}_{i-1} - (p_i + \overline{p}_{i-1})(\mathbf{A} + p_i\mathbf{I})^{-1}\mathbf{W}_{i-1}\right)$
4  $\mathbf{V}_i = [\mathbf{V}_{i-1}\,\mathbf{W}_i]$
END
```

Here $\{p_1, p_2, \ldots, p_{i_{max}}\}$ denotes a set of ADI parameters. As pointed out in [12], the proper choice of ADI parameters is crucial for efficiency of the LRCF-ADI method. There exist several routines for selection of ADI parameters. We will describe two of them.

The first has been presented in [12] and is based on the following two ideas. First, we generate a discrete set, which "approximates" the spectrum, which is done by a pair of Arnoldi processes (we calculate the set of Ritz values). Then we choose a set of shift parameters which is a subset of the set of Ritz values by a heuristic that delivers a suboptimal set of ADI shifts. As we will see in the last section, sometimes this choice can yield a poor approximation of the trace, especially in cases when the viscosity is of small magnitude or in the case when s is not small enough.

The second routine has been proposed in [16] and contains the following four steps:

1. Find the indices of 1's on the right-hand side (i.e., find positions of 1's in the matrix GG^T).
2. Find the corresponding submatrix of A using these indices (i.e., form the submatrix A_s).
3. Take a "little bit bigger block" A_{block} (which depends on a particular problem) which includes the submatrix A_s.
4. Eigenvalues of the chosen matrix A_{block} are ADI parameters $(p_1, \ldots, p_l \in \sigma(A_{block}))$.

Figure 3.1 shows how we form the matrix $\mathbf{A}_{block}$.

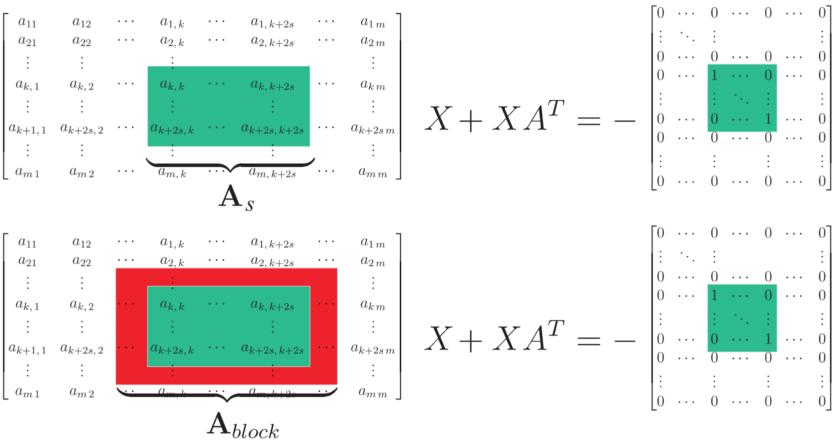


FIG. 3.1. *Choosing of A_{block}.*

Once we find a proper set of ADI parameters we can proceed with the implementation of Algorithm 1.

Before giving our algorithm for the trace of the solution of the Lyapunov equation (3.1), we will point out some facts and introduce some notation which will be used later.

First, in Algorithm 1 one has to compute the inverse of $(\mathbf{A} + p_i\mathbf{I})$. In our case (3.1)–(3.3)

$$\mathbf{A} \equiv \mathbf{A}_0 - v\,\mathbf{D}_0\,\mathbf{D}_0^T,$$

where

$$\mathbf{A}_0 = \begin{bmatrix} 0 & \Omega \\ -\Omega & -\alpha\,\Omega^k \end{bmatrix} \quad \text{and} \quad \mathbf{D}_0 = \begin{bmatrix} 0 \\ C_0 \end{bmatrix}.$$

Since we consider the problems with $\operatorname{rank}(\mathbf{D}_0) = r \ll n$, one can use the [illegible] for calculation of the inverse $(\mathbf{A} + p_i\mathbf{I})^{-1}$ [6, eq. (2.1.4), p. 51]. For this purpose we will need the notation

$$\mathbf{A}_0(p_i) = \mathbf{A}_0 + p_i\mathbf{I}\,. \tag{3.5}$$

Now, we can write

$$\begin{aligned} \mathbf{A}^{-1} &\equiv (\mathbf{A}_0(p_i) - v\mathbf{D}_0\,\mathbf{D}_0^T)^{-1} \\ &= \mathbf{A}_0(p_i)^{-1} + v\,\mathbf{A}_0(p_i)^{-1}\mathbf{D}_0\left(I_r - v\mathbf{D}_0^T\mathbf{A}_0(p_i)^{-1}\mathbf{D}_0\right)^{-1}\mathbf{D}_0^T\mathbf{A}_0(p_i)^{-1}\,. \end{aligned} \tag{3.6}$$

Note that the inverse $\mathbf{A}_0(p_i)^{-1}$ can be derived directly; that is,

$$\begin{aligned} \mathbf{A}_0(p_i)^{-1} &= \begin{bmatrix} p_i I & \Omega \\ -\Omega & p_i I - \alpha\,\Omega^k \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (\Omega^2 + p_i^2 I - p_i\alpha\,\Omega^k)^{-1}(p_i I - \alpha\,\Omega^k) & -(\Omega^2 + p_i^2 I - p_i\alpha\,\Omega^k)^{-1}\,\Omega \\ (\Omega^2 + p_i^2 I - p_i\alpha\,\Omega^k)^{-1}\,\Omega & p_i\,(\Omega^2 + p_i^2 I - p_i\alpha\,\Omega^k)^{-1} \end{bmatrix}. \end{aligned} \tag{3.7}$$

This means that all matrices in (3.6) can be computed directly, except

$$Inv(v, p_i) \equiv \left(I_r - v\mathbf{D}_0^T\mathbf{A}_0(p_i)^{-1}\mathbf{D}_0\right)^{-1}\,. \tag{3.8}$$

The matrix $I_r - v\mathbf{D}_0^T\mathbf{A}_0(p_i)^{-1}\mathbf{D}_0$ is of order r.

Using the above considerations, we can adapt Algorithm 1 for calculating the trace of the solution of the Lyapunov equation (3.1) in the following way.

Algorithm 2 (calculating the trace using LRCF-ADI).

```
INPUT: $\Omega$  $C_0$  $v$  $\mathbf{G}$  $\{p_1, p_2, \dots, p_l\}$
OUTPUT: $Tr$  $Tr$ [illegible] $trace(X)$
    0  $Tr = 0$
    1  $\mathbf{W}_1 = \sqrt{-2\mathrm{Re}p_1}\,\left(\mathbf{A}_0(p_1)^{-1}\mathbf{G} + v\,\mathbf{A}_0(p_1)^{-1}\mathbf{D}_0\,Inv(v,p_1)\,\mathbf{D}_0^T\mathbf{A}_0(p_1)^{-1}\mathbf{G}\right)$
2   $tr(1) = \sum_i^{2\,s} \|\mathbf{W}_1(:,i)\|^2$
    FOR  $j = 2, 3, \dots, l$
3   $\mathbf{W}_j = \sqrt{\mathrm{Re}p_j/\mathrm{Re}p_{j-1}}$
          $\cdot \left(\mathbf{W}_{j-1} - (p_j + \overline{p}_{j-1})\left(\mathbf{A}_0(p_j)^{-1}\right.\right.$
             $\left.\left. + v\,\mathbf{A}_0(p_j)^{-1}\mathbf{D}_0\,Inv(v,p_j)\,\mathbf{D}_0^T\mathbf{A}_0(p_j)^{-1}\right)\mathbf{W}_{j-1}\right)$
4   $tr(j) = \sum_i^{2\,s} \|\mathbf{W}_j(:,i)\|^2$
END
5   $Tr = \sum_i^{l} tr(i)$
```

Assuming that we have a proper set of ADI parameters, we can calculate the costs for Algorithm 2. Note that every step in Algorithm 2 contains $\mathbf{A}_0(p_j)^{-1}\mathbf{G}$. It is easy to see that this multiplication costs $2s \cdot \mathcal{O}(m)$ flops. Further, the inner loop contains a matrix $\mathbf{A}_0(p_j)^{-1}\mathbf{D}_0\, Inv(v,p_j)\, \mathbf{D}_0^T \mathbf{A}_0(p_j)^{-1}\mathbf{G}$, which can be calculated in $2\,s\,r\cdot(\,\mathcal{O}(m)+\mathcal{O}(r)\,)+2s(\,m\mathcal{O}(r)+2\mathcal{O}(m)\,)+\mathcal{O}(r^3\,)$ operations. Altogether this yields that Algorithm 2 calculates the trace of the solution of the Lyapunov equation in

$$l\cdot\big(\,s\,r\cdot(\,\mathcal{O}(m)+\mathcal{O}(r)\,)+s(\mathcal{O}(m\,r)+\mathcal{O}(m)\,)+\mathcal{O}(r^3)\big) \tag{3.9}$$

operations. In our applications we usually have $r\leq 12$. Now from (3.9) it follows that Algorithm 2 with such r needs less than $\mathcal{O}(s\,m)$ flops.

At this point it is important to emphasize that the fact that Algorithm 2 needs less operations than existing algorithms which calculate the whole solution of the Lyapunov equation ($\mathcal{O}(s\,m)$ contrary to $\mathcal{O}(m^3)$) is not its only advantage. The fact that Algorithm 2 calculates only the trace of the solution of the Lyapunov equation implies much more efficient memory usage. Indeed, in each iteration step of Algorithm 2 we have to save only one $m\times 2s$ matrix (in each step we overwrite the old one) instead of standard LRCF-ADI where we have to form the factor which is a matrix of dimension $m\times 2s\cdot l$, where l is the number of iteration steps.

Since the efficiency and accuracy of Algorithm 2 depend on a proper choice of ADI parameters, in the next section we will analyze accuracy of the solution obtained by Algorithm 2 using a new suboptimal set of ADI parameters.

4. Quality of the new choice of ADI parameters. In this section we present an error bound for the approximation of the trace of the solution of the Lyapunov equation obtained by Algorithm 1 (Algorithm 2) generated by ADI parameters $\{p_1,\ldots,p_l\}$ obtained by a new suboptimal choice proposed in the last section.

The error bound contains two parts: the first belongs to the approximation of the solution $\mathbf{X}$ of Lyapunov equation (3.1) with its lth approximation $\mathbf{X}_l$ obtained by Algorithm 1 (Algorithm 2) with the set of ADI parameters which corresponds to a certain subset of the spectrum of the matrix $\mathbf{A}$. This bound was presented in [16, Theorem 2.1].

The second part of the bound belongs to the approximation of a suboptimal set of ADI shifts ("exact eigenvalues" of the matrix $\mathbf{A}$) with some approximative values. This approximation has to be done since the location of eigenvalues which represent a suboptimal set of ADI shifts is still an open problem.

Thus, let the matrix $\widetilde{\mathbf{X}}_l$ be the approximation of the solution $\mathbf{X}_l$ by Algorithm 1 (Algorithm 2) with the set of ADI parameters $\{p_1,\ldots,p_l\}$ obtained by our new suboptimal choice.

Thus, we can write

$$|Tr(\mathbf{X})-Tr(\widetilde{\mathbf{X}}_l)|\leq|Tr(\mathbf{X})-Tr(\mathbf{X}_l)|+|Tr(\mathbf{X}_l)-Tr(\widetilde{\mathbf{X}}_l)|. \tag{4.1}$$

As pointed out above, the bound for $|Tr(\mathbf{X})-Tr(\mathbf{X}_l)|$ will be taken from [16, Theorem 2.1], assuming that $\mathbf{A}$ is diagonalizable with eigendecomposition:

$$\mathbf{A}=\mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}\,.$$

Let $\mathbf{X}_l$ be the lth approximation obtained by Algorithm 1 (Algorithm 2) with the set of ADI parameters which correspond to any subset of exact eigenvalues of the matrix $\mathbf{A}$ (i.e., $\lambda_{k_i}\in\sigma(\mathbf{A})$ for $i=1,\ldots,l$). Then the following bound holds [16, Theorem 2.1]:

$$|Tr(\mathbf{X})-Tr(\mathbf{X}_l)|\leq\|\mathbf{S}\|^2\sum_{j=l+1}^{m}(-2\mathrm{Re}(\lambda_{k_j}))\sum_{k=1}^{m}|\sigma(j,k)|^2\cdot\|\widehat{g}_k\|^2, \tag{4.2}$$

where

$$\sigma(1,k) = \frac{1}{\lambda_k + \overline{\lambda}_{k_1}} \quad \text{and} \quad \sigma(j,k) = \frac{1}{\lambda_k + \overline{\lambda}_{k_1}} \prod_{t=2}^{j} \frac{\lambda_k - \lambda_{k_{t-1}}}{\lambda_k + \overline{\lambda}_{k_t}} \quad \text{for} \quad j > 1, \tag{4.3}$$

and

$$\widehat{\mathbf{G}} = \mathbf{S}^{-1}\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1s} \\ g_{21} & g_{22} & \cdots & g_{2s} \\ \vdots & \vdots & \vdots & \vdots \\ g_{m1} & g_{m2} & \cdots & g_{ms} \end{bmatrix} = \begin{bmatrix} \widehat{g}_1 \\ \widehat{g}_2 \\ \vdots \\ \widehat{g}_m \end{bmatrix}; \tag{4.4}$$

that is, $\widehat{g}_i$ denotes the ith row of the matrix $\widehat{\mathbf{G}}$.

As shown in [16], the right-hand side of (4.2) strongly depends on the magnitude of $\|\widehat{g}_k\|_F^2$, $k = 1, \ldots, k_0$ (the structure of the matrix $\widehat{G}$ is important). For example, if

$$\|\widehat{g}_1\| \geq \cdots \geq \|\widehat{g}_l\| \gg \|\widehat{g}_{l+1}\|_F \approx \cdots \approx \|\widehat{g}_m\|_F \approx \sqrt{\varepsilon}, \tag{4.5}$$

then we can choose $\lambda_{k_1}, \ldots, \lambda_{k_l}$ such that $\sigma(j,1) = \cdots = \sigma(j,l) = 0$ for $j \geq 2$. This is fulfilled for $k_i = i$. If $\|S\|$, $\mathrm{Re}(\lambda_j)$, and the rest of $\sigma(j,k)$'s have modest magnitudes, then from (4.2) we have

$$|Tr(\mathbf{X}) - Tr(\mathbf{X}_l)| \leq \mathcal{O}(\varepsilon).$$

With this assumption we will continue with bounding the second part of the right-hand side of (4.1). Without loss off generality, we will assume that the matrix $\mathbf{G}$ from (3.1) has the form $\mathbf{G} = [I_s, 0]^T$, where I_s is an identity matrix of dimension s, that is,

$$\mathbf{G}\mathbf{G}^T = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.6}$$

It is important to note that in the case when $\mathbf{G}$ has the form defined as in (4.6), our choice of ADI parameters is given as the set of eigenvalues of the matrix $\mathbf{A}_{block}$, where $\mathbf{A}_{block} = (\mathbf{A}_0)_{11} - v\,\mathbf{d}_1\,\mathbf{d}_1^T$ and where, after perfect shuffle permutation, $\mathbf{A}$ has the following form:

$$\mathbf{A} \equiv \begin{bmatrix} (\mathbf{A}_0)_{11} - v\,\mathbf{d}_1\,\mathbf{d}_1^T & -v\,\mathbf{d}_1\,\mathbf{d}_2^T \\ -v\,\mathbf{d}_2\,\mathbf{d}_1^T & (\mathbf{A}_0)_{22} - v\,\mathbf{d}_2\,\mathbf{d}_2^T \end{bmatrix}, \quad \text{where} \quad \mathbf{D}_0 = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}.$$

Usually, the dimension l of the matrix $\mathbf{A}_{block}$ is taken as $3s \leq l \leq 5s$.

Note that we can write

$$\widetilde{\mathbf{A}} = \mathbf{A} - \Delta\mathbf{A} \equiv \begin{bmatrix} (\mathbf{A}_0)_{11} - v\,\mathbf{d}_1\,\mathbf{d}_1^T & 0 \\ 0 & (\mathbf{A}_0)_{22} - v\,\mathbf{d}_2\,\mathbf{d}_2^T \end{bmatrix},$$

where

$$\Delta\mathbf{A} = v \cdot \begin{bmatrix} 0 & \mathbf{d}_1\,\mathbf{d}_2^T \\ \mathbf{d}_2\,\mathbf{d}_1^T & 0 \end{bmatrix},$$

which means that our ADI shifts are exact eigenvalues of the matrix $\widetilde{\mathbf{A}}$.

Recall that $\mathbf{X}_l$ is the lth approximation of the solution of the Lyapunov equation (3.1) obtained by Algorithm 1 (Algorithm 2) generated by the set $\{\lambda_1, \lambda_2, \ldots, \lambda_l\}$, where $\lambda_i \in \sigma(\mathbf{A})$ for $i = 1, \ldots, l$, while $\widetilde{\mathbf{X}}_l$ is the lth approximation of the solution of the Lyapunov equation (3.1) obtained by Algorithm 1 (Algorithm 2) generated by the set of ADI parameters obtained by our new suboptimal choice of ADI parameters, that is, with $p_i \in \sigma(\mathbf{A}_{block}) \subset \sigma(\widetilde{\mathbf{A}})$.

Our choice of ADI parameters can be written as

$$p_i \equiv \overline{\overline{\lambda}}_i = \overline{\lambda}_i \pm \overline{\delta\lambda}_i\,, \qquad i = 1, \ldots, l. \tag{4.7}$$

Further, $\mathbf{X}_l$ and $\widetilde{\mathbf{X}}_l$ can be written as

$$\mathbf{X}_l = \sum_{j=1}^{l} \|\mathbf{W}_j\|_F^2\,, \qquad\qquad \widetilde{\mathbf{X}}_l = \sum_{j=1}^{l} \|\widetilde{\mathbf{W}}_j\|_F^2,$$

where $\mathbf{W}_j$ and $\widetilde{\mathbf{W}}_j$ are matrices obtained by Algorithm 1 (Algorithm 2).

Then, if we write

$$\widetilde{\mathbf{W}}_j = \mathbf{W}_j + \delta\mathbf{W}_j,$$

it is easy to show that the following first order bound holds:

$$Tr(\mathbf{X}_l) - Tr(\widetilde{\mathbf{X}}_l) \leq 2\sum_{j=1}^{l} \|\mathbf{W}_j\|_F\, \|\delta\mathbf{W}_j\|_F + \mathcal{O}(\|\delta\mathbf{W}_j\|_F{}^2)\,. \tag{4.8}$$

We will continue with bounding $\|\delta\mathbf{W}_j\|_F$.

Let $\mathbf{W}_j$ be the jth matrix obtained by Algorithm 1 (Algorithm 2) with ADI parameters $\{\lambda_1, \ldots, \lambda_l\}$, with input matrices $\mathbf{A}$ and $\mathbf{G}$, where $\mathbf{G}$ is defined as in (4.6). In [16] it has been shown that $\mathbf{W}_j$ can be written as

$$\mathbf{W}_j = \sqrt{-2\,\mathrm{Re}(\lambda_j)}\,\mathbf{S}\cdot \mathrm{diag}\left(\sigma(j,1), \sigma(j,2), \ldots, \sigma(j,m)\right)\,\mathbf{S}^{-1}\mathbf{G}, \tag{4.9}$$

where $\sigma(j,k)$ are given by

$$\sigma(1,k) = \frac{1}{\lambda_k + \overline{\lambda}_1} \quad \text{and} \quad \sigma(j,k) = \frac{1}{\lambda_k + \overline{\lambda}_j}\prod_{t=1}^{j-1}\frac{\lambda_k - \lambda_t}{\lambda_k + \overline{\lambda}_{t+1}} \quad \text{for} \quad j > 1.$$

Indeed, from Algorithm 1 (Algorithm 2) (for more details see the proof of Theorem 2.1 in [16]), it follows that

$$\begin{aligned}\mathbf{W}_j = \sqrt{-2\,\mathrm{Re}(\lambda_j)}\,\mathbf{S}&\cdot\left(\mathbf{I} - (\overline{\lambda}_j + \lambda_{j-1})(\mathbf{\Lambda} + \overline{\lambda}_j\mathbf{I})^{-1}\right)\\ &\cdot\left(\mathbf{I} - (\overline{\lambda}_{j-1} + \lambda_{j-2})(\mathbf{\Lambda} + \overline{\lambda}_{j-1}\mathbf{I})^{-1}\right)\cdots\left(\mathbf{I} - (\overline{\lambda}_2 + \lambda_1)(\mathbf{\Lambda} + \overline{\lambda}_2\mathbf{I})^{-1}\right)\\ &\cdot(\mathbf{\Lambda} + \overline{\lambda}_1\mathbf{I})^{-1}\mathbf{S}^{-1}\mathbf{G}\,,\end{aligned}$$

which together with the fact that in the above equality we have a $j-1$ diagonal matrix of the form

$$\left(\mathbf{I} - (\overline{\lambda}_k + \lambda_{k-1})(\mathbf{\Lambda} + \overline{\lambda}_k\mathbf{I})^{-1}\right) = \mathrm{diag}\left(\frac{\lambda_i - \lambda_{k-1}}{\lambda_i + \overline{\lambda}_k}\right)_i\,,$$

$i = 1, \ldots, m$, $k = 2, \ldots, j$, gives (4.9).

Here it is important to note that all eigenvalues of the matrix $\mathbf{A}$ from (3.2) are given in complex conjugate pairs. Thus, if we choose ADI parameters as the first l exact eigenvalues of $\mathbf{A}$, then the structure of $\sigma(j,k)$ implies

$$\sigma(j,k) = 0 \quad \text{for} \quad j = 1, \ldots, l, \quad j > k.$$

Similarly, let $\widetilde{\mathbf{W}}_j$ be the jth matrix obtained by Algorithm 1 (Algorithm 2) with ADI parameters $\{p_1, \ldots, p_l\}$, with the same input matrices $\mathbf{A}$ and $\mathbf{G}$, where p_i is defined by (4.7):

$$\widetilde{\mathbf{W}}_j = \sqrt{-2\,\mathrm{Re}(\lambda_j \pm \delta\lambda_j)}\,\mathbf{S} \cdot \mathrm{diag}\left(\widetilde{\sigma}(j,1), \widetilde{\sigma}(j,2), \ldots, \widetilde{\sigma}(j,m)\right)\, \mathbf{S}^{-1}\mathbf{G}, \tag{4.10}$$

where $\widetilde{\sigma}(j,k)$ are given by

$$\widetilde{\sigma}(1,k) = \frac{1}{\lambda_k + \overline{\lambda}_1 \pm \overline{\delta\lambda_1}} \quad \text{and}$$
$$\widetilde{\sigma}(j,k) = \frac{1}{\lambda_k + \overline{\lambda}_j \pm \overline{\delta\lambda_j}} \prod_{t=1}^{j-1} \frac{\lambda_k - \lambda_t \mp \delta\lambda_t}{\lambda_k + \overline{\lambda}_{t+1} \pm \overline{\delta\lambda}_{t+1}} \quad \text{for} \quad j > 1\,.$$

Now it is easy to see that from (4.9) and (4.10) it follows that

$$\delta\mathbf{W}_j = \mathbf{S} \cdot \mathrm{diag}\left(\delta\zeta(j,1) \ldots, \delta\zeta(j,k), \ldots, \delta\zeta(j,m)\right)\, \mathbf{S}^{-1}\mathbf{G},$$

where

$$\delta\zeta(j,k) = \sqrt{-2\,\mathrm{Re}(\lambda_j \pm \delta\lambda_j)} \cdot \widetilde{\sigma}(j,k) - \sqrt{-2\,\mathrm{Re}(\lambda_j)} \cdot \sigma(j,k). \tag{4.11}$$

Now $\delta\mathbf{W}_j$ can be written as

$$\delta\mathbf{W}_j = \mathbf{S} \begin{bmatrix} \delta\zeta(j,1)\,\widehat{g}_1 \\ \delta\zeta(j,2)\,\widehat{g}_2 \\ \vdots \\ \delta\zeta(j,m)\,\widehat{g}_m \end{bmatrix}. \tag{4.12}$$

Recall that we have assumed that $\widehat{\mathbf{G}}$ satisfy (4.5) and that all $\mathrm{Re}(\lambda_j)$, $\sigma(j,k)$'s for $k \geq l$ have modest magnitudes such that

$$|\sigma(j,k)| \cdot \|\widehat{g}_k\| = \mathcal{O}(\sqrt{\varepsilon}), \qquad k \geq l, \tag{4.13}$$

and $\sqrt{-2\,\mathrm{Re}(\lambda_j)}\,\mathcal{O}(\varepsilon) = \mathcal{O}(\varepsilon)$. Then, as we have already pointed out, there holds $|Tr(\mathbf{X}) - Tr(\mathbf{X}_l)| \leq \mathcal{O}(\varepsilon)$.

Thus, if

$$|\eta_j| \equiv \left|\frac{\delta\lambda_j}{\lambda_j}\right| < 1, \tag{4.14}$$

then we have the first order approximation

$$\sqrt{-2\,\mathrm{Re}(\lambda_j \pm \delta\lambda_j)} = \sqrt{-2\,\mathrm{Re}(\lambda_j)} \cdot \left(1 \pm \frac{|\eta_j|}{2}\right) + \mathcal{O}(|\eta_j|^2),$$

which implies

$$\delta\zeta(j,k) \approx \sqrt{-2\operatorname{Re}(\lambda_j)} \cdot (\widetilde{\sigma}(j,k) - \sigma(j,k)) \mp \frac{\eta_j}{2} \cdot \sqrt{-2\operatorname{Re}(\lambda_j)} \cdot \widetilde{\sigma}(j,k)\,.$$

Thus, the above consideration implies that we have to bound more carefully the first l components of the matrix $\delta\mathbf{W}_j$ from (4.4) than the rest of the components. That is, we are going to bound $|\delta\zeta(j,i)|$ for $j = 1,\dots,l$.

For $1 < j$ and $k \le l$ we have

$$\widetilde{\sigma}(j,k) = \frac{1}{\lambda_k + \overline{\lambda}_1 \pm \overline{\delta\lambda}_1} \prod_{\substack{t=1\\ t\neq k}}^{j-1} \frac{\lambda_k - \lambda_t \mp \delta\lambda_t}{\lambda_k + \overline{\lambda}_{t+1} \pm \overline{\delta\lambda}_{t+1}} \cdot \frac{\lambda_k \;\; \eta_k}{\lambda_k + \overline{\lambda}_{k+1} \pm \overline{\delta\lambda}_{k+1}},$$

which can be written as

$$\widetilde{\sigma}(j,k) = \delta\sigma(j,k) \cdot \eta_k, \qquad \text{where}$$

$$\delta\sigma(j,k) = \frac{1}{\lambda_k + \overline{\lambda}_1 \pm \overline{\delta\lambda}_1} \prod_{\substack{t=1\\ t\neq k}}^{j-1} \frac{\lambda_k - \lambda_t \mp \delta\lambda_t}{\lambda_k + \overline{\lambda}_{t+1} \pm \overline{\delta\lambda}_{t+1}} \cdot \frac{\lambda_k}{\lambda_k + \overline{\lambda}_{k+1} \pm \overline{\delta\lambda}_{k+1}},$$

Further, let $\delta\mathbf{W}_j(k,:)$ be the kth row of the matrix $\delta\mathbf{W}_j$ defined by (4.12). Assumption (4.5) implies that the entries in $\delta\mathbf{W}_j(k,:)$ will be small in magnitude for $k > l$. Thus for $k > l$ the following simple bound is quite satisfactory to us:

$$|\delta\zeta(j,k)| \lesssim \sqrt{-2\operatorname{Re}(\lambda_j)} \cdot |\sigma(j,k) + \widetilde{\sigma}(j,k)| + \frac{|\eta_j|}{2} \cdot \sqrt{-2\operatorname{Re}(\lambda_j)} \cdot |\widetilde{\sigma}(j,k)|\,.$$

Finally, we can bound the right-hand side in (4.8). Indeed,

$$|Tr(\mathbf{X}_l) - Tr(\widetilde{\mathbf{X}}_l)| \lesssim 2 \sum_{j=1}^{l} \|\mathbf{W}_j\|_F \, \|\delta\mathbf{W}_j\|_F, \tag{4.15}$$

where

$$\|\delta\mathbf{W}_j\|_F \le \|\mathbf{S}\| \sum_{k=1}^{m} |\delta\zeta(j,k)| \cdot \|\widehat{g}_k\|,$$

with

$$|\delta\zeta(j,k)| \lesssim \sqrt{-2\operatorname{Re}(\lambda_j)} \cdot \left(|\delta\sigma(j,k)| \cdot |\eta_k| + \frac{|\eta_j|}{2} \cdot |\widetilde{\sigma}(j,k)| \right), \quad k \le l, \tag{4.16}$$

and

$$|\delta\zeta(j,k)| \lesssim \sqrt{-2\operatorname{Re}(\lambda_j)} \cdot \left(|\sigma(j,k)| + |\widetilde{\sigma}(j,k)| + \frac{|\eta_j|}{2} \cdot |\widetilde{\sigma}(j,k)| \right), \quad k > l. \tag{4.17}$$

Now, from (4.2) and (4.15) it follows that

$$\begin{aligned} |Tr(\mathbf{X}) - Tr(\widetilde{\mathbf{X}}_l)| \le \|\mathbf{S}\|^2 \sum_{j=l+1}^{m} (-2\operatorname{Re}(\lambda_{k_j})) \sum_{k=1}^{m} |\sigma(j,k)|^2 \cdot \|\widehat{g}_k\|^2 \\ + 2\,\|\mathbf{S}\| \sum_{j=1}^{l} \|\mathbf{W}_j\|_F \left(\sum_{k=1}^{m} |\delta\zeta(j,k)| \cdot \|\widehat{g}_k\| \right), \end{aligned} \tag{4.18}$$

where all quantities used in the above bound are defined in the above consideration.

From (4.16) and (4.17) it follows that an important part in our bound is played by the perturbation of eigenvalues. Since we have assumed that all eigenvalues of the matrix $\mathbf{A}$ are simple, the following (see [3] or [13]) holds:

$$|\delta\lambda_k| \leq \frac{t_k^* \Delta A s_k}{t_k^* s_k} = \varepsilon_k, \tag{4.19}$$

where s_k and t_k are right and left eigenvectors belonging to λ_k normalized so that $\|s_k\| = \|t_k\| = 1$ and $|t_k^* s_k| = t_k^* s_k$. Now from (4.19) and (4.14) it follows that

$$|\eta_k| \leq \frac{t_k^* \Delta A s_k}{|\lambda_k| \; t_k^* s_k} \, .$$

As the last issue in this section, we are going to discuss how realistic is our assumption (4.5). It is obvious that assumption (4.5) will be fulfilled if $\mathcal{R}(\mathbf{G})$ is close to column space $\mathcal{R}(\mathbf{S})$.

We are going to derive a bound for viscosity v, from which will be possible to conclude when our assumption,

$$\|\widehat{g}_j\| \approx \sqrt{\varepsilon} \qquad \text{for} \quad j = l+1, \ldots, m,$$

will be feasible.

Recall that we have denoted

$$\widehat{\mathbf{G}} = \mathbf{S}^{-1}\mathbf{G} = \left[\widehat{g}_1, \ldots, \widehat{g}_l, \widehat{g}_{l+1}, \ldots \widehat{g}_m\right]^T .$$

Note that for (4.6) $\widehat{\mathbf{G}}$ contains only the first s columns of the matrix $\mathbf{T}^* = \mathbf{S}^{-1}$. Further, let $\widetilde{\mathbf{A}} = \widetilde{\mathbf{S}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{S}}^{-1}$ be the eigenvalue decomposition of the matrix $\widetilde{\mathbf{A}}$. Note that since $\widetilde{\mathbf{A}}$ is block diagonal, $\widetilde{\mathbf{S}}$ will be block diagonal, too. Thus for v of modest magnitude one can expect that $\mathbf{T}^*$ will have an almost block diagonal structure.

If we write

$$\mathbf{E}_S = \mathbf{S}^{-1}\widetilde{\mathbf{S}},$$

using $\widehat{\mathbf{G}} = \mathbf{T}^*\widetilde{\mathbf{S}}\widetilde{\mathbf{S}}^{-1}\mathbf{G} \equiv \mathbf{E}_S\widetilde{\mathbf{T}}^*_{(1:s,:)}$ we can bound $\widehat{\mathbf{G}}_2 \equiv \left[\widehat{g}_{l+1}, \ldots, \widehat{g}_m\right]^T$ in the following way:

$$\|\widehat{\mathbf{G}}_2\| \leq \|(\mathbf{E}_S)_{21}\| \, \|\widetilde{\mathbf{T}}^*_{(1:s,:)}\|,$$

where $(\mathbf{E}_S)_{21}$ denotes an off-diagonal block which contains rows from $l+1$ up to m and the first s columns of $\mathbf{E}_S$.

Now using the simple equality

$$\mathbf{\Lambda}\mathbf{E}_S - \mathbf{E}_S\widetilde{\mathbf{\Lambda}} = -\mathbf{S}^{-1}\Delta\mathbf{A}\widetilde{\mathbf{S}},$$

one can easily see that

$$|(\mathbf{E}_S)_{21}|_{ij} = \frac{\mathbf{T}^*_{(:,\,i)}\Delta\mathbf{A}\widetilde{\mathbf{S}}_{(:,\,j)}}{|\lambda_i - \widetilde{\lambda}_j|} = v \cdot \frac{\mathbf{T}^*_{(:,\,i)}\mathbf{d}_2\mathbf{d}_1^T\widetilde{\mathbf{S}}_{(:,\,j)}}{|\lambda_i - \widetilde{\lambda}_j|} \equiv v \cdot \frac{\Psi_{ij}}{|\lambda_i - \widetilde{\lambda}_j|}\,, \qquad \text{where}$$
$$\lambda_i \in \mathbf{\Lambda}(l+1:m, l+1:m)\,, \qquad \widetilde{\lambda}_j \in \widetilde{\mathbf{\Lambda}}(1:s, 1:s).$$

Altogether this implies

$$\|\widehat{\mathbf{G}}_2\|_F \le v \cdot \frac{\|\Psi\|_F \; \|\widetilde{\mathbf{T}}^*_{(1:s,:)}\|}{\operatorname{gap}(\widetilde{\mathbf{\Lambda}}(1:s,1:s),\mathbf{\Lambda}(l+1:m,l+1:m))}\,, \tag{4.20}$$

$$\operatorname{gap}(\widetilde{\mathbf{\Lambda}}(1:s,1:s),\mathbf{\Lambda}(l+1:m,l+1:m)) = \min_{i,j}|\lambda_i - \widetilde{\lambda}_j|, \qquad \text{where}$$

$$\lambda_i \in \mathbf{\Lambda}(l+1:m,l+1:m)\,, \qquad \widetilde{\lambda}_j \in \widetilde{\mathbf{\Lambda}}(1:s,1:s)\,.$$

Note that since $l > s$, the gap function can be large in magnitude; thus if $\|\widetilde{\mathbf{T}}^*_{(1:s,:)}\|$ has a modest magnitude, then for $l > 4\,s$ from (4.20) it follows that if

$$v \lesssim \sqrt{\varepsilon} \cdot \frac{\operatorname{gap}(\widetilde{\mathbf{\Lambda}}(1:s,1:s),\mathbf{\Lambda}(l+1:m,l+1:m))}{\|\Psi\|_F \; \|\widetilde{\mathbf{T}}^*_{(1:s,:)}\|},$$

our assumption (4.5) will be fulfilled, which means that $\|\widehat{g}_k\| \approx \sqrt{\varepsilon}$ for $k = l+1,\dots,m$.

5. Numerical illustration. The first example in this section illustrates the efficiency and accuracy of the new algorithm in respect to the column rank of the right-hand side of the Lyapunov equation (3.1).

[illegible] 1. We consider the Lyapunov equation

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^T = -\mathbf{G}\mathbf{G}^T,$$

where

$$\mathbf{A} = \mathbf{A}_0 - v\,\mathbf{D} = \begin{bmatrix} 0 & \Omega \\ -\Omega & 0 \end{bmatrix} - v \begin{bmatrix} 0 & 0 \\ 0 & C_0 C_0^T \end{bmatrix}\,,$$

where $C_0 = \operatorname{rand}(n,r)$ and $r = 4$. The matrix $\mathbf{A}$ is $m \times m$, with $m = 400$ (note that $n = 200$). The matrix $\mathbf{G}$ we construct as

$$\mathbf{G} = 0.001 \cdot \operatorname{rand}(m, 2\,s), \qquad \mathbf{G}(1:s,1:s) = \operatorname{rand}(s),$$
$$\mathbf{G}(n+1:n+s, s+1:2\,s) = \operatorname{rand}(s).$$

where $\operatorname{rank}(\mathbf{G}) = 2s$.

The ADI shifts will be chosen as the eigenvalues of the matrix $\mathbf{A}_p(1:l_p,1:l_p)$, where $\mathbf{A}_p$ is obtained from $\mathbf{A}$ using the perfect shuffle permutation.

We are going to compute a relative error `relerr` and the number of floating point operations `flops` obtained by the new algorithm, Algorithm 2, for the different s—the rank of the matrix $\mathbf{G}$.

The relative error is defined as

$$\text{relerr} = \frac{|Tr_X - Trnew_X|}{Tr_X},$$

where Tr_X is the trace of the solution of the Lyapunov equation (3.1) obtained by MATLAB function `Lyap` (which is based on the Bartels–Stewart method), while $Trnew_X$ is the trace of the solution obtained by the new algorithm, Algorithm 2. Further, the number of floating point operations `flops` is defined by

$$\texttt{flops} = l_p \cdot \big(s\,r\,(\,\mathcal{O}(m) + \mathcal{O}(r)\,) + s(\,\mathcal{O}(m\,r) + \mathcal{O}(m)\,) + \mathcal{O}(r^3)\big) + \mathcal{O}(l_p^3),$$

which has been obtained using (3.9) with additional $\mathcal{O}(l_p^3)$ flops, which corresponds to the number of calculations needed for the l_p suboptimal ADI shifts.

s	10	10	10	10
l_p	20	40	60	200
relerr	0.0016	3.2430e-004	2.1210e-004	3.3638e-005
`flops`	$\mathcal{O}(10^5)$	$\mathcal{O}(10^6)$	$\mathcal{O}(10^6)$	$\mathcal{O}(10^7)$

s	30	30	30	30
l_p	60	80	100	200
relerr	0.0015	6.2926e-004	4.0017e-004	1.1102e-004
`flops`	$\mathcal{O}(10^6)$	$\mathcal{O}(10^6)$	$\mathcal{O}(10^7)$	$\mathcal{O}(10^7)$

s	50	50	50	50
l_p	50	100	140	200
relerr	0.5525	0.0016	5.2693e-004	2.3062e-004
`flops`	$\mathcal{O}(10^6)$	$\mathcal{O}(10^7)$	$\mathcal{O}(10^7)$	$\mathcal{O}(10^7)$

As one can see from the above table, the accuracy and efficiency of the new algorithm, Algorithm 2, strongly depends on s—the number of damped modes. If we are interested in a result of a certain accuracy, then as s grows, the required number of ADI parameters grows, too, which slows performance of the new algorithm.

Further, we will compare different algorithms for dampers' viscosity optimization considering a simple mechanical system consisting of three rows of n masses connected with $n+1$ springs on the left-hand side on the fixed base and on the right-hand side on the mass m_0 connected to the fixed base with the spring with stiffness k_0 (see Figure 5.1).

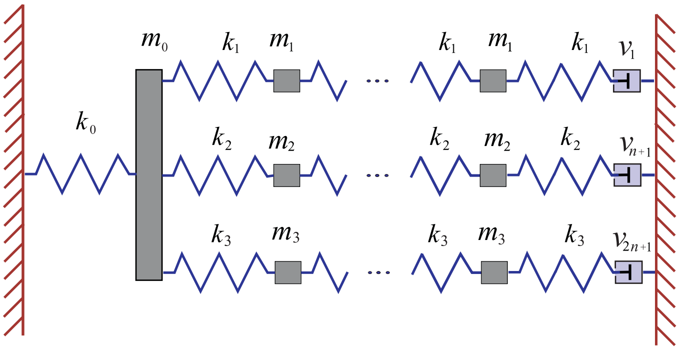


FIG. 5.1. *$(3n+1)$-mass oscillator with three dampers.*

2. Consider a damped linear vibrational mechanical system consisting of three rows of n masses connected with $n+1$ springs on the left-hand side on the fixed base and on the right-hand side on the mass m_0 connected to the fixed base with the spring with stiffness k_0 (see Figure 5.1). Then one can write

$$M\ddot{x} + D\dot{x} + Kx = 0,$$

where M, D, and K are defined as

$$M = \operatorname{diag}(m_1, \ldots, m_1, m_2, \ldots, m_2, m_3, \ldots, m_3, m_0)\,, \tag{5.1}$$

(5.2)

$$K = \begin{bmatrix} K_{11} & & & -\kappa_1 \\ & K_{22} & & -\kappa_2 \\ & & K_{33} & -\kappa_3 \\ -\kappa_1^T & -\kappa_2^T & -\kappa_3^T & k_1{+}k_2{+}k_3{+}k_0 \end{bmatrix}, \quad K_{ii} = k_i \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix},$$

and $\kappa_i = \begin{bmatrix} 0 & \ldots & 0 & k_i \end{bmatrix}^T$, $K_{ii} \in \mathbb{R}^{n\times n}$ and $\kappa_i \in \mathbb{R}^{n\times 1}$, for $i = 1, 2, 3$,

$$D \equiv C_u + C = C_u + v e_1 e_1^T + v e_n e_n^T + v e_{2n+1} e_{2n+1}^T.$$

Note that M, D, and K are matrices of order $3n + 1 \times 3n + 1$. We will set $n = 40$.

Let $m_2 = k_2 = 2$ and $m_3 = k_3 = 4$ be fixed, and let m_0, m_1, k_0, k_1 be chosen such that

$$m_0, k_0 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\} \quad \text{and} \quad m_1, k_1 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}\,.$$

This means that we have 1296 different configurations defined by different sets $\{m_0, m_1, m_2, m_3\}$ and $\{k_0, k_1, k_2, k_3\}$. For each of these configurations we have derived the optimal trace of the solution of the corresponding Lyapunov equation (3.1), where G is defined by (3.4) and $G(1 : s, 1 : s) = I_s$, $G(n + 1 : n + s, s + 1 : 2\,s) = I_s$ with $s = 20$ and the optimal viscosity. We have used the following four algorithms:

1. Minimization process based on Newton iteration process for higher-dimensional problems which use the Bartels–Stewart Lyapunov solver [1]. For the starting point we have used the one proposed in [2] ([illegible]).
2. Minimization process based on Newton iteration process for one-dimensional problems which use a new Lyapunov solver proposed in [15]. For the starting point we have used $v_0 = 0.01$ ([illegible]).
3. Minimization process with the standard MATLAB function `fminbnd` (using the Golden section search and parabolic interpolation) based on LRCF-ADI Lyapunov solver Algorithm 2 generated by the set of ADI parameters proposed by Penzl in [12] ([illegible]). Minimization has been performed on the interval $[0, 5000]$.
4. Minimization process with the standard MATLAB function `fminbnd` (using the Golden section search and parabolic interpolation) based on LRCF-ADI Lyapunov solver Algorithm 2 generated by the new set of ADI parameters proposed in [16] ([illegible]). Minimization has been performed on the interval $[0, 500]$.

Before we continue with the analysis of all 1296 configurations, we will illustrate the quality of error bound (4.18) on one particular configuration. The chosen configuration has some interesting properties (later the same configuration will be considered as the case of a nonconvex energy curve), and it is far away from the best possible case for our error analysis.

Let

$$\begin{aligned} &m_0 = 100, \quad m_1 = 0.01, \quad m_2 = 2, \quad m_4 = 4; \\ &k_0 = 100, \quad k_1 = 0.01, \quad k_2 = 2, \quad k_4 = 4\,. \end{aligned}$$

Let $v = 4$, $s = 20$ and let $l = 60$ be the number of ADI shifts generated by a new algorithm proposed in section 3. Simple calculation gives that the first part in the bound (4.18) is bounded with 0.7876, while the second part is bounded with 100.7001. Altogether this gives

$$\frac{|Tr(\mathbf{X}) - Tr(\widetilde{\mathbf{X}}_l)|}{Tr(\mathbf{X})} \leq 0.0205\,.$$

At the same time real relative error for the lth approximation of the trace is

$$\frac{|Tr(\mathbf{X}) - Tr(\widetilde{\mathbf{X}}_l)|}{Tr(\mathbf{X})} \leq 2.55 \cdot 10^{-4}\,.$$

It has to be pointed out that for the considered configuration a relative error in eigenvalues satisfies

$$\max_{1\leq k\leq m} |\eta_k| \leq \max_{1\leq k\leq m} \frac{t_k^* \Delta A s_k}{|\lambda_k|\; t_k^* s_k} \leq 0.76,$$

while for the norms of the rows of the matrix $\widehat{\mathbf{G}} = \mathbf{S}^{-1}\mathbf{G}$ it holds that

$$10^{-7} \leq |\widehat{g}_k| \leq 0.77\,, \qquad l+1 \leq k \leq m.$$

The above example shows that, although we do not have a very accurate approximation for all eigenvalues and eigenvectors (which was expected), we still have 4 exact digits in our approximation of the trace, while our bound predicts 2 exact digits.

We continue with the analysis of all 1296 configurations. Table 5.1 contains the ratios between optimal traces obtained by algorithms [illegible] and [illegible] (3 and 4) and algorithms [illegible] and [illegible] (2 and 1).

Table 5.1

	LRCF-ADI-Penzl/New.-new	*LRCF-ADI-new/New.-new*
> 1.02	15.7 %	5.2 %
< 0.98	2.7 %	2.7 %
	LRCF-ADI-Penzl/New.-Bart.-Stew.	*LRCF-ADI-new/New.-Bart.-Stew.*
> 1.05	42 %	37 %
< 1	0.75 %	6 %

As one can see from Table 5.1 in 5.2% of our experiments the optimal trace obtained by the algorithm [illegible] is more than 2% larger than the optimal trace obtained by [illegible], while in 15.7% of our experiments the optimal trace obtained by [illegible] is more than 2% larger than the optimal trace obtained by [illegible]. At the same time, in 2.7% of our experiments, both algorithms ([illegible] and [illegible]) obtain at least 2% a larger optimal trace than the optimal trace obtained by [illegible].

Since the discrepancies in both cases were not expected, we will carefully consider the cases in which they appear. It turns out that by the algorithm [illegible] in 5.2% of our experiments we have obtained a larger trace in comparison to the algorithm [illegible], whereas in 15.7% we have obtained a larger trace by the algorithm [illegible] in comparison to the algorithm [illegible]. This has been caused by using the wrong intervals: $[0, 500]$ for [illegible] and $[0, 5000]$ for [illegible], respectively. For example, in the abovementioned situations, the algorithm [illegible] has obtained the optimal trace for optimal viscosity $v = 500$, which is obviously wrong. On the other hand, in 2.7% of our experiments the optimal trace obtained by the algorithms [illegible] and [illegible] is smaller than the optimal trace obtained by algorithm [illegible]. The reason for this is a wrong starting point $v_0 = 0.01$ for Newton iterations.

For illustration consider the case with

$$\begin{aligned} &m_0 = 100, \quad m_1 = 0.01, \quad m_2 = 2, \quad m_4 = 4; \\ &k_0 = 100, \quad k_1 = 0.01, \quad k_2 = 2, \quad k_4 = 4. \end{aligned}$$

Figure 5.2 shows the trace as the function of viscosity v. It is obvious that starting point $v_0 = 0.01$ will lead to a wrong result. But if we take for the starting point any point $6 < v_0 < 10$, optimal viscosity is $v = 14.765$.

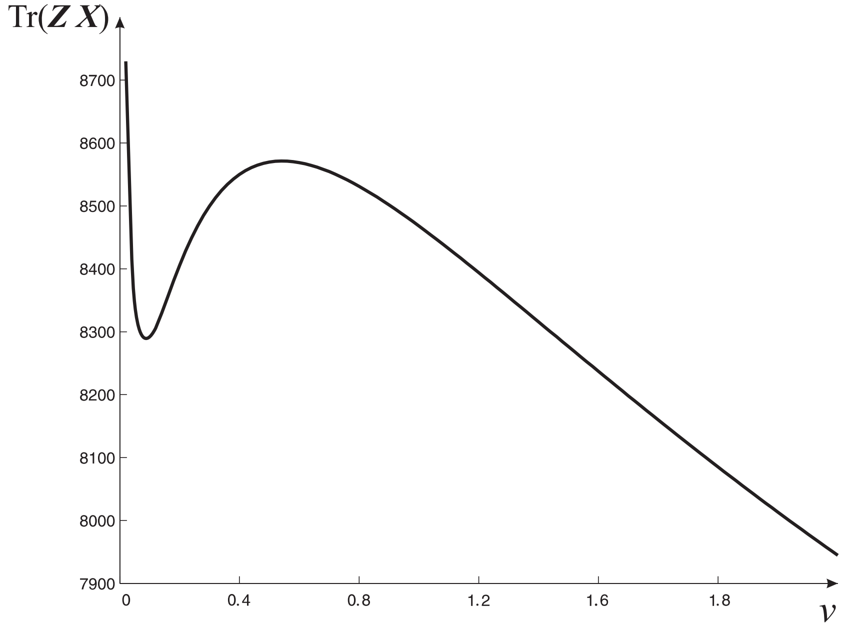


FIG. 5.2. *The graph of the trace function.*

Similar conclusions hold for the ratio between optimal traces obtained by algorithms [illegible] and [illegible] and the optimal trace obtained by multidimensional optimization using [illegible].

As expected (with a usage of correct intervals on which we perform minimization) in 30% of our experiments, the optimal trace obtained by [illegible] is more than 5% smaller than the optimal trace obtained by algorithm [illegible], while in 35% of our experiments, the optimal trace obtained by [illegible] is more than 5% smaller than optimal trace obtained by algorithm [illegible]. But in 6% of our experiments, the optimal trace obtained by algorithm [illegible] is smaller than the optimal trace obtained by [illegible], which was definitively unexpected. The reason for this lies in the fact that in these particular situations the starting point for [illegible], obtained by the algorithm proposed in [2], is wrong. For illustration we consider the case with

$$\begin{aligned} m_0 &= 100, \quad m_1 = 0.01, \quad m_2 = 2, \quad m_4 = 4;\\ k_0 &= 100, \quad k_1 = 10, \quad k_2 = 2, \quad k_4 = 4. \end{aligned}$$

The starting point for optimization process [illegible] obtained by routine `calcvisc` taken from [2] gives visc $= \left[0.0147, 2.7535, 5.5009\right]$, which corresponds to $Tr(\mathbf{ZX}_{opt}) = 4965.4$, at the same time the optimal trace for $v_{opt} = 16.41$ (obtained by algorithm [illegible]) is $Tr(\mathbf{ZX}) = 4062.6$. But if we change a starting point to visc$_2 = \left[16.4, 16.4, 16.4\right]$, then the algorithm [illegible] gives optimal trace $Tr(\mathbf{ZX}_{opt}) = 3997.1$ for viscosity visc$_{opt} = \left[20.6384, 11.5852, 23.183\right]$.

Considering the abovementioned, we can conclude that both algorithms based on the LRCF-ADI Lyapunov solver combined with some nonsmooth optimization give us very satisfactory results.

At the same time, the number of operations needed for one optimization with algorithm [illegible] is much smaller than the number needed for optimization with algorithm [illegible]. For illustration, to obtain optimal viscosity with algorithm [illegible] one usually needs ~ 20 iterations, which together with (3.9) gives $20 \cdot (280\mathcal{O}(m) + 10/3(3s)^3)$ operations, where the second number in the bracket $10/3(3s)^3$ stands for the number of operations needed for calculating the ADI parameters (eigenvalues of a $3s \times 3s$ nonsymmetric matrix (see [6])). On the other hand, as shown in [15], the algorithm [illegible] needs $14/3\,(2rm)^3 + \mathcal{O}(r^2m^2)$ operations, which is obviously much more.

Acknowledgment. We would like to thank the anonymous referees for their very careful reading of the manuscript and valuable comments.

REFERENCES

[1] R. H. Bartels and G. W. Stewart, *A solution of the matrix equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.

[2] K. Brabender, *Optimale Dämpfung von linearen Schwingungssystemen*, Ph.D. thesis, Fernuniversität Hagen, Hagen, Germany, 1998.

[3] R. Byers and D. Kressner, *On the condition of a complex eigenvalue under real perturbations*, BIT, 44 (2004), pp. 209–214.

[4] S. J. Cox, I. Nakić, A. Rittmann, and K. Veselić, *Minimization of energy of a damped system*, Systems Control Lett., 53 (2004), pp. 187-194.

[5] P. Freitas and P. Lancaster, *On the optimal value of the spectral abscissa for a system of linear oscillators*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 195–208.

[6] G. H. Golub and Ch.F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[7] P. Hebrard and A. Henrot, *Optimal shape and position of the actuators for the stabilization of a string*, Systems Control Lett., 48 (2003), pp. 199–209.

[8] J.-R. Li and J. White, *Low rank solution of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 260–280.

[9] J. Li, F. Wang, and J. White, *An efficient Lyapunov equation-based approach for generating reduced-order models of interconnect*, in Proceedings of the 36th Annual IEEE/ACM Design Automation Conference, New Orleans, LA, 1999.

[10] I. Nakić, *Optimal Damping of Vibrational Systems*, Ph.D. thesis, Fernuniversität Hagen, Hagen, Germany, 2002.

[11] M. Paz, *Structural Dynamics: Theory and Computation*, Van Nostrand Reinhold, New York, 1991.

[12] T. Penzl, *LYAPACK*, http://www.tu-chemnitz.de/sfb393/lyapack.

[13] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[14] N. Truhar and K. Veselić, *On some properties of the Lyapunov equation for damped systems*, Math. Commun., 9 (2004), pp. 189–197.

[15] N. Truhar, *An efficient algorithm for damper optimization for linear vibrating systems using Lyapunov equation*, J. Comput. Appl. Math., 172 (2004), pp. 169–182.

[16] N. Truhar and K. Veselić, *Bounds on the trace of solution to the Lyapunov equation with a general stable matrix*, Systems Control Lett., 56 (2007), pp. 493–503.

[17] K. Veselić, K. Brabender, and K. Delinić, *Passive control of linear systems*, in Applied Mathematics and Computation, M. Rogina et al., eds. Department of Mathematics, University of Zagreb, Zagreb, Croatia, 2001, pp. 39–68.

[18] K. Veselić, *On linear vibrational systems with one-dimensional damping*, Appl. Anal., 29 (1988), pp. 1–18.

[19] K. Veselić, *On linear vibrational systems with one-dimensional damping.* II, Integral Equations Operator Theory, 13 (1990), pp. 883–897.

[20] K. Veselić, *Exponential decay of semigroups in Hilbert space*, Semigroup Forum, 55 (1997), pp. 325–331.

[21] K. Veselić, *Estimating the operator exponential*, Linear Algebra Appl., 280 (1998), pp. 241–244.

[22] K. Veselić, *Bounds for exponentially stable semigroups*, Linear Algebra Appl., 358 (2003), pp. 309–333.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 1, pp. 40–53

REFINED PERTURBATION BOUNDS FOR EIGENVALUES OF HERMITIAN AND NON-HERMITIAN MATRICES*

I. C. F. IPSEN† AND B. NADLER‡

Abstract. We present eigenvalue bounds for perturbations of Hermitian matrices and express the change in eigenvalues in terms of a projection of the perturbation onto a particular eigenspace, rather than in terms of the full perturbation. The perturbations we consider are Hermitian of rank one, and Hermitian or non-Hermitian with norm smaller than the spectral gap of a specific eigenvalue. Applications include principal component analysis under a spiked covariance model, and pseudo-arclength continuation methods for the solution of nonlinear systems.

Key words. eigenvalues, Hermitian matrix, eigenvalue gap, perturbation bounds, non-Hermitian perturbations, principal components, numerical continuation

AMS subject classifications. 65F15, 65H10, 65H17, 65H20, 15A18, 15A42

DOI. 10.1137/070682745

1. Introduction.

We present perturbation bounds for eigenvalues of Hermitian matrices that were motivated by two applications: principal component analysis under a spiked covariance model [25], and pseudo-arclength continuation methods for the solution of systems of nonlinear equations [7].

Although these applications are very different, they share a common requirement: The change in the eigenvalues of interest should be determined not by the global norm of the full perturbation, which can be quite large, but rather by the norm of a projection of the perturbation on a particular eigenspace. In contrast, most existing eigenvalue bounds are expressed either in terms of the full perturbation or else in terms of a residual, and therefore do not provide sufficient information for our applications.

The paper is organized as follows. We start with the most specific class of perturbations, Hermitian rank one updates, and then generalize the perturbations first to Hermitian and then to non-Hermitian matrices. In section 2 we present bounds for Hermitian rank one updates, and explain why such bounds can be useful in pseudo-arclength continuation methods. In section 3 we consider Hermitian perturbations whose norm is smaller than the spectral gap of a specific eigenvalue, and we describe their use in principal component analysis. In section 4 we extend the bounds to non-Hermitian perturbations.

Notation. The identity matrix of order k is $I_k = \begin{pmatrix} e_1 & \dots & e_k \end{pmatrix}$. The norm $\|\cdot\|$ denotes the two norm. The eigenvalues of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ are numbered so that

$$\lambda_{\min}(A) \equiv \lambda_n(A) \le \cdots \le \lambda_{\max}(A) \equiv \lambda_1(A).$$

*Received by the editors February 16, 2007; accepted for publication (in revised form) by J. H. Brandts February 19, 2008; published electronically February 27, 2009. This research was supported in part by NSF grant DMS 0112069.

http://www.siam.org/journals/simax/31-1/68274.html

†Department of Mathematics, North Carolina State University, P.O. Box 8205, Raleigh, NC 27695-8205 (ipsen@ncsu.edu, http://www4.ncsu.edu/~ipsen/).

‡Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel (boaz.nadler@weizmann.ac.il). This author's research was supported by a grant from the Lord Sieff of Brimpton memorial fund.

The conjugate transpose of a matrix A is denoted by A^*; an overbar, as in $\overline{A}$, denotes elementwise complex conjugation.

We will use two measures for the separation between adjacent eigenvalues of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$: the distance of an eigenvalue $\lambda_i(A)$ to its right neighbor,

$$\mathrm{gap}_i \equiv \lambda_{i-1}(A) - \lambda_i(A), \qquad 2 \leq i \leq n,$$

and the minimum of the distance to left and right neighbors,

$$\mathrm{Gap}_i \equiv \min_{j\neq i} |\lambda_i(A) - \lambda_j(A)|.$$

The two measures are related,

$$\mathrm{Gap}_n = \mathrm{gap}_n, \qquad \mathrm{Gap}_1 = \mathrm{gap}_2, \qquad \mathrm{Gap}_i = \min\{\mathrm{gap}_i, \mathrm{gap}_{i+1}\}, \quad 1 < i < n.$$

2. Hermitian rank one updates.

We present improved perturbation bounds for eigenvalues of Hermitian matrices when the perturbation is Hermitian of rank one.

Before describing an application that requires such bounds, we mention that algorithms for computing eigenvalues and eigenvectors of Hermitian matrices modified by a rank one matrix are well established [3, 10, 14], [11, section 8.5.3, section 12.5.1]; the corresponding inverse eigenvalue problem has also been investigated [21].

2.1. Numerical continuation.

Numerical continuation is the process of solving systems of nonlinear equations $G(u, \lambda) = 0$ for various values of the real parameter λ. Here $G : \mathbb{R}^{N+1} \to \mathbb{R}^N$ is assumed to be sufficiently smooth [12, 20, 22, 29].

Parameter continuation is a method for tracing out a solution path by repeatedly incrementing λ until the desired value of λ is reached. In each iteration, the current solution u serves as an initial iterate for, say, Newton's method to compute a solution for the next value of λ. Although parameter continuation is simple and intuitive, it fails at points (u, λ) where the Jacobian G_u is singular.

One can try to circumvent singularities by reparameterizing the problem and introducing the arclength parameter s. Now both u and λ depend on s, and the original parameter λ is treated as an unknown. The resulting pseudo-arclength continuation method [12, 20, 22, 29] implements parameter continuation on $F(u(s), \lambda(s)) = 0$ with s as the parameter and solves

$$F(x, s) = \begin{pmatrix} G(x) \\ \mathcal{N}(x, s) \end{pmatrix} = 0, \qquad x = \begin{pmatrix} u(s) \\ \lambda(s) \end{pmatrix},$$

where $\mathcal{N}$ represents a normalization equation. Pseudo-arclength continuation requires that the Jacobian

$$F_x = \begin{pmatrix} G_u & G_\lambda \\ \mathcal{N}_u & \mathcal{N}_\lambda \end{pmatrix}$$

be nonsingular. The normalization equation is set up so that at a point where $G(u_0, \lambda_0) = 0$ the row $\begin{pmatrix} \mathcal{N}_u & \mathcal{N}_\lambda \end{pmatrix}$ has unit norm and is almost orthogonal to the rows of $\begin{pmatrix} G_u & G_\lambda \end{pmatrix}$. Hence F_x is nonsingular at (u_0, λ_0) if the rank of $\begin{pmatrix} G_u & G_\lambda \end{pmatrix}$ equals N. In other words, F_x is nonsingular if G_u is nonsingular, or if the nullspace of G_u has dimension 1 and G_λ is not in the range of G_u [20, 29]. The latter singularity is called a limit point, fold point, simple fold, or turning point [4, 5, 23, 27, 29].

Denote the partial derivatives at (u_0, λ_0) by $G_u = G_u(u_0, \lambda_0)$ and $y = G_\lambda(u_0, \lambda_0)$. Instead of the singular values of the Jacobian F_x we consider the eigenvalues of

$$F_x F_x^* = \begin{pmatrix} A + yy^* & 0 \\ 0 & 1 \end{pmatrix} + \mathcal{E},$$

where $A = G_u G_u^*$, and we have used the fact that the last row of F_x has unit norm and is almost orthogonal to the others, so that $\|\mathcal{E}\|$ is small. To estimate the condition number of F_x and the convergence rate of a Newton-GMRES method, it suffices to bound $\|F_x^{-1}\|$ by determining a nontrivial lower bound for the smallest eigenvalue $\lambda_{\min}(A + yy^*)$ [7]. For this positive semidefinite rank one update, Weyl's theorem implies [11, Theorem 8.1.8], [26, Corollary 10.3.1]

$$\lambda_{\min}(A) \le \lambda_{\min}(A + yy^*).$$

When A is nonsingular, this bound is adequate for our purposes. However, it is useless at a fold point, because there A is singular and $0 = \lambda_{\min}(A) = 0 < \lambda_{\min}(A + yy^*)$. We need a lower bound for $\lambda_{\min}(A)$ that takes into account that y is not in the range of A and has a nonzero contribution in the eigenspace of $\lambda_{\min}(A)$.

Our objective is to tighten our previous bound [7, Theorem 3.3] and the bounds in [13]. This is accomplished in Theorem 2.1 below. The results in section 2.2 may also be of benefit in the construction of nonsingular bordered matrices.

2.2. Smallest eigenvalue. For a given Hermitian matrix $A \in \mathbb{C}^{n\times n}$ and a column vector $y \in \mathbb{C}^n$, we improve the inclusion interval from Weyl's theorem for the smallest eigenvalue of Hermitian rank one updates $A \pm yy^*$,

$$\lambda_{\min}(A) \le \lambda_{\min}(A + yy^*) \le \lambda_{n-1}(A), \tag{2.1}$$
$$\lambda_{\min}(A) - \|y\|^2 \le \lambda_{\min}(A - yy^*) \le \lambda_{\min}(A), \tag{2.2}$$

by taking into account the contribution of y in the eigenspace of $\lambda_{\min}(A)$.

Let $A = V\Lambda V^*$ be an eigenvalue decomposition, where $V = \begin{pmatrix} v_1 & \dots & v_n \end{pmatrix}$ is unitary and

$$\Lambda = \begin{pmatrix} \lambda_1(A) & & \\ & \ddots & \\ & & \lambda_n(A) \end{pmatrix}, \qquad \lambda_{\max}(A) = \lambda_1(A) \ge \cdots \ge \lambda_n(A) = \lambda_{\min}(A).$$

Define the projections of the vector y onto the eigenvectors of A,

$$y_{i:j} \equiv \begin{pmatrix} v_i & \dots & v_j \end{pmatrix}^* y, \qquad 1 \le i \le j \le n.$$

Below we bound the smallest eigenvalues of the rank one updates in terms of eigenvalues of 2 × 2 matrices (which can be considered as rank one updates of projections onto two-dimensional subspaces). Explicit expressions for these eigenvalues are given in Corollary 2.2. A simpler upper bound in Corollary 2.3 emphasizes the influence of y_n and the separation of $\lambda_{\min}(A)$ from the next eigenvalue.

Theorem 2.1 (smallest eigenvalue). *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian, $y \in \mathbb{C}^n$, and*

$$L_\pm \equiv \begin{pmatrix} \lambda_{n-1}(A) & 0 \\ 0 & \lambda_n(A) \end{pmatrix} \pm \begin{pmatrix} \|y_{1:n-1}\| \\ y_n \end{pmatrix} \begin{pmatrix} \|y_{1:n-1}\| & \overline{y}_n \end{pmatrix},$$
$$U_\pm \equiv \begin{pmatrix} \lambda_{n-1}(A) & 0 \\ 0 & \lambda_n(A) \end{pmatrix} \pm \begin{pmatrix} y_{n-1} \\ y_n \end{pmatrix} \begin{pmatrix} \overline{y}_{n-1} & \overline{y}_n \end{pmatrix}.$$

Then $\lambda_{\min}(L_\pm) \le \lambda_{\min}(A \pm yy^*) \le \lambda_{\min}(U_\pm)$, *where*

$$\begin{aligned}\lambda_{\min}(A) \le \lambda_{\min}(L_+) \le \lambda_{\min}(U_+) \le \lambda_{n-1}(A),\\ \lambda_{\min}(A) - \|y\|^2 \le \lambda_{\min}(L_-) \le \lambda_{\min}(U_-) \le \lambda_{\min}(A).\end{aligned}$$

Proof. Abbreviate $\alpha_j \equiv \lambda_j(A)$, $1 \le j \le n$, and partition the eigenvalue decomposition of A so as to distinguish the smallest eigenvalue $\alpha_n = \lambda_{\min}(A)$.

$$\Lambda = \begin{pmatrix}\Lambda_1 & \\ & \alpha_n\end{pmatrix}, \qquad \Lambda_1 \equiv \begin{pmatrix}\alpha_1 & & \\ & \ddots & \\ & & \alpha_{n-1}\end{pmatrix},$$

and $V = \begin{pmatrix}V_1 & v_n\end{pmatrix}$ with $V_1 \equiv \begin{pmatrix}v_1 & \dots & v_{n-1}\end{pmatrix}$. We derive the bounds by "projecting" A onto a 2×2 matrix with eigenvalues α_n and α_{n-1}.

Lower bounds. We start with the positive semidefinite update. Let z be a unit-norm eigenvector associated with $\lambda_{\min}(A + yy^*)$, i.e., $(A + yy^*)z = \lambda_{\min}(A + yy^*)z$, $\|z\| = 1$. Express z in the V-basis,

$$\begin{pmatrix}z_{1:n-1}\\ z_n\end{pmatrix} = \begin{pmatrix}V_1^* z\\ v_n^* z\end{pmatrix} = V^* z.$$

Then

$$\begin{aligned}\lambda_{\min}(A + yy^*) &= z^*(A + yy^*)z = z_{1:n-1}^*\Lambda_1 z_{1:n-1} + \alpha_n|z_n|^2 + |y^*z|^2\\ &\ge \alpha_{n-1}\|z_{1:n-1}\|^2 + \alpha_n|z_n|^2 + |z_{1:n-1}^* y_{1:n-1} + \overline{z}_n y_n|^2\\ &= \begin{pmatrix}z_{1:n-1}^* & \overline{z}_n\end{pmatrix}\left[\begin{pmatrix}\alpha_{n-1}I_{n-1} & 0\\ 0 & \alpha_n\end{pmatrix} + \begin{pmatrix}y_{1:n-1}\\ y_n\end{pmatrix}\begin{pmatrix}y_{1:n-1}^* & \overline{y}_n\end{pmatrix}\right]\begin{pmatrix}z_{1:n-1}\\ z_n\end{pmatrix}.\end{aligned}$$

Let Q be a unitary matrix of order $n-1$ so that $Qy_{1:n-1} = \|y_{1:n-1}\|e_{n-1}$ and set $w \equiv \left(\begin{smallmatrix}Qz_{1:n-1}\\ z_n\end{smallmatrix}\right)$, where $\|w\| = 1$. Then

$$\begin{aligned}\lambda_{\min}(A + yy^*) &\ge w^*\left(\begin{pmatrix}\alpha_{n-1}I_{n-1} & 0\\ 0 & \alpha_n\end{pmatrix} + \begin{pmatrix}\|y_{1:n-1}\|e_{n-1}\\ y_n\end{pmatrix}\begin{pmatrix}\|y_{1:n-1}\|e_{n-1}^* & \overline{y}_n\end{pmatrix}\right)w\\ &\ge \lambda_{\min}\begin{pmatrix}\alpha_{n-1}I_{n-2} & 0\\ 0 & L_+\end{pmatrix} = \min\{\alpha_{n-1}, \lambda_{\min}(L_+)\}.\end{aligned}$$

Applying (2.1) to L_+ gives $\alpha_n \le \lambda_{\min}(L_+) \le \alpha_{n-1}$, and

$$\lambda_{\min}(A + yy^*) \ge \min\{\alpha_{n-1}, \lambda_{\min}(L_+)\} = \lambda_{\min}(L_+).$$

Now consider the negative semidefinite update, and let z be a unit-norm eigenvector associated with $\lambda_{\min}(A - yy^*)$, i.e., $(A - yy^*)z = \lambda_{\min}(A - yy^*)z$, $\|z\| = 1$. As above one shows $\lambda_{\min}(A - yy^*) \ge \min\{\alpha_{n-1}, \lambda_{\min}(L_-)\}$. Applying (2.2) to L_- gives $\alpha_n - \|y\|^2 \le \lambda_{\min}(L_-) \le \alpha_n$, and

$$\lambda_{\min}(A - yy^*) \ge \min\{\alpha_{n-1}, \lambda_{\min}(L_-)\} = \lambda_{\min}(L_-).$$

Upper bounds. Since $U_\pm$ are the respective trailing 2×2 principal submatrices of $V^*(A \pm yy^*)V$, Cauchy's interlace theorem [26, section 10.1] implies $\lambda_{\min}(A \pm yy^*) \le \lambda_{\min}(U_\pm)$. Applying (2.1) to U_+ and (2.2) to U_- gives $\lambda_{\min}(U_+) \le \alpha_{n-1}$ and $\lambda_{\min}(U_-) \le \alpha_n$. $\square$

Below we give explicit expressions for the bounds in Theorem 2.1 in terms of the absolute gap between the two smallest eigenvalues,

$$\mathrm{gap}_n \equiv \lambda_{n-1}(A) - \lambda_n(A) \geq 0.$$

Corollary 2.2 (smallest eigenvalue). *In Theorem* 2.1

$$\lambda_{\min}(L_\pm) = \lambda_{\min}(A) + \frac{1}{2}\left(\mathrm{gap}_n \pm \|y\|^2 - \sqrt{(\mathrm{gap}_n \pm \|y\|^2)^2 \mp 4\mathrm{gap}_n |y_n|^2}\right)$$

and

$$\lambda_{\min}(U_\pm) = \lambda_{\min}(A)+\frac{1}{2}\left(\mathrm{gap}_n \pm \|y_{n-1:n}\|^2 - \sqrt{(\mathrm{gap}_n \pm \|y_{n-1:n}\|^2)^2 \mp 4\mathrm{gap}_n |y_n|^2}\right).$$

Implications for numerical continuation. For the application to pseudo-arclength continuation in section 2.1, it is important to know how $|y_n|$ and gap_n influence $\lambda_{\min}(A + yy^*)$, provided $\lambda_{\min}(A) < \lambda_{n-1}(A)$, $y_n \neq 0$, and $y_{n-1} \neq 0$. This influence becomes clear in the next bound, which illustrates how much of the increase in the smallest eigenvalue can be due to the contribution of y in the eigenspace of $\lambda_{\min}(A)$.

Corollary 2.3. *Under the conditions of Theorem* 2.1*,*

$$\lambda_{\min}(A + yy^*) \leq \lambda_{\min}(A) + |y_n|\sqrt{\mathrm{gap}_n}.$$

Proof. Abbreviate $\beta = \mathrm{gap}_n + \|y_{n-1:n}\|^2$ and $\gamma = \mathrm{gap}_n|y_n|^2$, and in the expression for $\lambda_{\min}(U_+)$ from Corollary 2.2 write $\lambda_{\min}(U_+) = \lambda_{\min}(A) + \delta$, where

$$\delta = \frac{1}{2}\left(\beta - \sqrt{\beta^2 - 4\gamma}\right) = \frac{2\gamma}{\beta + \sqrt{\beta^2 - 4\gamma}} \leq 2\frac{\gamma}{\beta} \leq \sqrt{\gamma} = |y_n|\sqrt{\mathrm{gap}_n}.$$

The last inequality follows from the fact that the term under the square root is nonnegative, i.e., $\beta^2 \geq 4\gamma$. □

2.3. Largest eigenvalue. We improve the inclusion interval from Weyl's theorem for the largest eigenvalue of $A \pm yy^*$,

$$\begin{aligned} \lambda_{\max}(A) \leq \lambda_{\max}(A + yy^*) &\leq \lambda_{\max}(A) + \|y\|^2, \\ \lambda_2(A) \leq \lambda_{\max}(A - yy^*) &\leq \lambda_{\max}(A), \end{aligned}$$

by taking into account the contribution of y in the eigenspace of $\lambda_{\max}(A)$.

Theorem 2.4 (largest eigenvalue). *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian, $y \in \mathbb{C}^n$, and*

$$L_\pm \equiv \begin{pmatrix} \lambda_1(A) & 0 \\ 0 & \lambda_2(A) \end{pmatrix} \pm \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \begin{pmatrix} \bar{y}_1 & \bar{y}_2 \end{pmatrix},$$

$$U_\pm \equiv \begin{pmatrix} \lambda_1(A) & 0 \\ 0 & \lambda_2(A) \end{pmatrix} \pm \begin{pmatrix} y_1 \\ \|y_{2:n}\| \end{pmatrix} \begin{pmatrix} \bar{y}_1 & \|y_{2:n}\| \end{pmatrix}.$$

Then $\lambda_{\max}(L_\pm) \leq \lambda_{\max}(A \pm yy^*) \leq \lambda_{\max}(U_\pm)$, where

$$\begin{aligned} \lambda_{\max}(A) \leq \lambda_{\max}(L_+) \leq \lambda_{\max}(U_+) &\leq \lambda_{\max}(A) + \|y\|^2, \\ \lambda_2(A) \leq \lambda_{\max}(L_-) \leq \lambda_{\max}(U_-) &\leq \lambda_{\max}(A). \end{aligned}$$

Proof. Use the fact that $\lambda_{\max}(A) = -\lambda_{\min}(-A)$, and apply Theorem 2.1. □

As in section 2.2, we give explicit expressions for the bounds in Theorem 2.4 in terms of the absolute gap between the two largest eigenvalues,

$$\mathrm{gap}_2 \equiv \lambda_{\max}(A) - \lambda_2(A) \geq 0.$$

Corollary 2.5 (largest eigenvalue). *In Theorem* 2.4

$$\lambda_{\max}(L_{\pm}) = \lambda_{\max}(A) + \frac{1}{2}\left(-\mathrm{gap}_2 \pm \|y_{1:2}\|^2 + \sqrt{\left(\mathrm{gap}_2 \pm \|y_{1:2}\|^2\right)^2 \mp 4\mathrm{gap}_2|y_2|^2}\right)$$

and

$$\lambda_{\max}(U_{\pm}) = \lambda_{\max}(A) + \frac{1}{2}\left(-\mathrm{gap}_2 \pm \|y\|^2 + \sqrt{\left(\mathrm{gap}_2 \pm \|y\|^2\right)^2 \mp 4\mathrm{gap}_2\|y_{2:n}\|^2}\right).$$

2.4. Interior eigenvalues. We improve the inclusion intervals from Weyl's theorem for the interior eigenvalues of $A \pm yy^*$,

$$\lambda_i(A) \leq \lambda_i(A + yy^*) \leq \lambda_{i-1}(A), \qquad 2 \leq i \leq n-1, \tag{2.3}$$

$$\lambda_{i+1}(A) \leq \lambda_i(A - yy^*) \leq \lambda_i(A), \tag{2.4}$$

by using the bounds for the extreme eigenvalues in Theorems 2.1 and 2.4 on principal submatrices.

Theorem 2.6 (interior eigenvalues). *Let* $A \in \mathbb{C}^{n\times n}$ *be Hermitian,* $y \in \mathbb{C}^n$, *and*

$$L_{\pm}^{(i)} \equiv \begin{pmatrix} \lambda_{i-1}(A) & 0 \\ 0 & \lambda_i(A) \end{pmatrix} \pm \begin{pmatrix} \|y_{1:i-1}\| \\ y_i \end{pmatrix} \begin{pmatrix} \|y_{1:i-1}\| & \overline{y}_i \end{pmatrix},$$

$$U_{\pm}^{(i)} \equiv \begin{pmatrix} \lambda_i(A) & 0 \\ 0 & \lambda_{i+1}(A) \end{pmatrix} \pm \begin{pmatrix} y_i \\ \|y_{i+1:n}\| \end{pmatrix} \begin{pmatrix} \overline{y}_i & \|y_{i+1:n}\| \end{pmatrix}.$$

Then

$$\lambda_{\min}(L_{+}^{(i)}) \leq \lambda_{\min}(A \pm yy^*) \leq \min\{\lambda_{\max}(U_{+}^{(i)}), \lambda_{i-1}(A)\}, \qquad 2 \leq i \leq n-1,$$

where $\lambda_i(A) \leq \lambda_{\min}(L_{+}^{(i)}) \leq \lambda_{\max}(U_{+}^{(i)}) \leq \lambda_i(A) + \|y_{i:n}\|^2$. *Moreover*

$$\max\{\lambda_{i+1}(A), \lambda_{\min}(L_{-}^{(i)})\} \leq \lambda_i(A - yy^*) \leq \lambda_{\max}(U_{-}^{(i)}), \qquad 2 \leq i \leq n-1,$$

where $\lambda_i(A) - \|y_{1:i}\|^2 \leq \lambda_{\min}(L_{-}^{(i)}) \leq \lambda_{\max}(U_{-}^{(i)}) \leq \lambda_i(A)$.

Proof. As before, abbreviate $\alpha_j \equiv \lambda_j(A)$, $1 \leq j \leq n$.

Lower bounds. Partition the eigenvalue decomposition so that

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}, \qquad \Lambda_1 \equiv \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_i \end{pmatrix}, \qquad \Lambda_2 \equiv \begin{pmatrix} \alpha_{i+1} & & \\ & \ddots & \\ & & \alpha_n \end{pmatrix},$$

and $V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}$ with $V_1 \equiv \begin{pmatrix} v_1 & \dots & v_i \end{pmatrix}$ and $V_2 \equiv \begin{pmatrix} v_{i+1} & \dots & v_n \end{pmatrix}$. Since $V_1^*(A \pm yy^*)V_1$ is a principal submatrix of order i of $V^*(A \pm yy^*)V$, the Cauchy interlace theorem [26, section 10.1] implies

$$\lambda_i(A \pm yy^*) = \lambda_i(V^*(A \pm yy^*)V) \geq \lambda_i(V_1^*(A \pm yy^*)V_1) = \lambda_{\min}\left(\Lambda_1 \pm y_{1:i}y_{1:i}^*\right).$$

Apply the lower bounds in Theorem 2.1. The second term in the maximum follows from (2.4).

Upper bounds. Partition

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}, \qquad \Lambda_1 \equiv \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_{i-1} \end{pmatrix}, \qquad \Lambda_2 \equiv \begin{pmatrix} \alpha_i & & \\ & \ddots & \\ & & \alpha_n \end{pmatrix},$$

and $V = \begin{pmatrix} V_1 & V_2 \end{pmatrix}$ with $V_1 \equiv \begin{pmatrix} v_1 & \dots & v_{i-1} \end{pmatrix}$ and $V_2 \equiv \begin{pmatrix} v_i & \dots & v_n \end{pmatrix}$. Since $V_2^*(A+yy^*)V_2$ is a principal submatrix of order $n-(i-1)$ of $V^*(A+yy^*)V$, the Cauchy interlace theorem implies

$$\lambda_i(A+yy^*) = \lambda_i\left(V^*(A+yy^*)V\right) \le \lambda_1(V_2^*(A+yy^*)V_2) = \lambda_{\max}\left(\Lambda_2 + y_{i:n}y_{i:n}^*\right).$$

Applying the upper bound in Theorem 2.4 yields $\lambda_{\max}\left(\Lambda_2 + y_{i:n}y_{i:n}^*\right) \le \lambda_{\max}(U_+^{(i)})$. The second term in the bound follows from (2.3). □

We use the absolute gap between the ith eigenvalue and its right neighbor,

$$\mathrm{gap}_i \equiv \lambda_{i-1}(A) - \lambda_i(A) \ge 0, \qquad 2 \le i \le n,$$

to determine explicit expressions for the bounds in Theorem 2.6.

COROLLARY 2.7. *In Theorem* 2.6

$$\lambda_{\min}(L_\pm^{(i)}) = \lambda_i(A) + \frac{1}{2}\left(\mathrm{gap}_i \pm \|y_{1:i}\|^2 - \sqrt{\left(\mathrm{gap}_i \pm \|y_{1:i}\|^2\right)^2 \mp 4\mathrm{gap}_i|y_i|^2}\right)$$

and

$$\lambda_{\max}(U_+^{(i)}) = \lambda_i(A) + \frac{1}{2}\left(-\mathrm{gap}_{i+1} \pm \|y_{i:n}\|^2 + \sqrt{\left(\mathrm{gap}_{i+1} \pm \|y_{i:n}\|^2\right)^2 \mp 4\mathrm{gap}_{i+1}\|y_{i+1:n}\|^2}\right).$$

3. Hermitian perturbations.

We present improved perturbation bounds for well-separated eigenvalues of Hermitian matrices. As in the previous section, we start by presenting an application that motivates these bounds.

3.1. Principal component analysis under the spiked covariance model.

Principal component analysis is a common tool in the analysis of high-dimensional data [15, 17]. Given m samples $x_i \in \mathbb{R}^n$, stored in a (mean centered) $m \times n$ matrix X, principal component analysis proceeds in three steps: It computes the *empirical covariance matrix* $C = \frac{1}{m}X^*X$; it finds orthonormal directions with maximal variance of the data, represented by the largest eigenvalues and eigenvectors of the matrix C; and at last it determines a low-dimensional representation of the data from linear projections onto these directions associated with maximal variance.

A common model for the analysis of principal component analysis on high-dimensional data is a small rank linear mixture or "spiked covariance model" [8, 16, 25]. Under this model, each data sample x_i is an independent identically distributed random vector of the form

$$x = \sum_{j=1}^{k} u_j v_j + \sigma\xi,$$

where u_j are random variables, also referred to as *components* or *latent variables*, the vectors $v_j \in \mathbb{R}^n$ are the *responses*, $\xi \in \mathbb{R}^n$ is a multivariate Gaussian noise vector with identity covariance matrix, and the scalar σ is the level of noise.

If all k random variables u_j are uncorrelated with zero mean and unit variance, and all eigenvectors v_j are orthogonal, then the first k eigenvalues and eigenvector pairs of the population covariance matrix are $(\|v_j\|^2+\sigma^2, v_j)$. Given that we have only a finite dataset $\{x_i\}_{i=1}^m$, the question is how close are the eigenvalues and eigenvectors of the empirical noisy covariance matrix C to their limiting values?

We formulate this problem in terms of matrix perturbation theory by working in an orthonormal basis whose first k vectors are $v_j/\|v_j\|$ and by writing the empirical covariance matrix as

$$C = A + E, \qquad \text{where} \quad A = \begin{pmatrix} \|v_1\|^2+\sigma^2 & & & \\ & \ddots & & \\ & & \|v_k\|^2+\sigma^2 & \\ & & & \sigma^2 I_{n-k} \end{pmatrix}.$$

We want to determine under which conditions the first few (largest) eigenvalues and eigenvectors of C correspond to the first few latent variables and characteristic responses of A, and how close these noisy estimates are to the unperturbed eigenvalues and eigenvectors of A. Many papers in statistics have studied the asymptotic distribution of the eigenvalues and eigenvectors of C in the limit as $m \to \infty$; see [1, 2, 9, 15, 24] and the references therein. However, in our application we are interested in answers to these questions for a finite number of samples m.

In the context of matrix perturbation theory, we look for absolute normwise perturbation bounds for eigenvalues of Hermitian matrices A and $A+E$. In particular, assuming that the signals u_j have a significant signal-to-noise ratio, we want bounds for eigenvalues $\lambda_j(A)$ that are well separated from all others, in the sense that $\text{gap}_j > \|E\|$. Moreover, to obtain sharp bounds we cannot afford to deal with the global norm of E, but rather we need to restrict ourselves to the contribution of E in the relevant eigenspace of A. In the present paper, we derive such bounds that depend on the projection of E onto a space spanned by an eigenvector v_j. The analysis is completed in a second paper [25], where we derive probabilistic bounds of the type "$\|Ev_j\| \le f(m,n)$ with probability $1-\delta$."

3.2. Perturbation bounds. Two types of existing two-norm results could potentially be applicable for the application in section 3.1: two-norm bounds that hold for all eigenvalues, and residual bounds that hold for a few eigenvalues. The best known example of a two-norm bound for Hermitian matrices $A, A+E \in \mathbb{C}^{n\times n}$ is Weyl's theorem [11, Theorem 8.1.6], [26, Theorem 10.3.1],

$$|\lambda_j(A) - \lambda_j(A+E)| \le \|E\|, \qquad 1 \le j \le n. \tag{3.1}$$

The advantage of (3.1) is that it applies to all eigenvalues of A and $A+E$. The disadvantage is that the bound is the same for all eigenvalues and depends on the global norm of E, which can be quite large, specifically in high dimensions, $n \gg 1$ [16].

For a single perturbed eigenvalue, one can either tighten the Bauer–Fike theorem [6, section 4.6.1], [28, Corollary 3.3] or derive a residual bound from scratch [26, Theorem 4.5.1] as follows. If w_j is a unit norm eigenvector associated with an eigenvalue $\lambda_j(A+E)$, i.e., $(A+E)w_j = \lambda_j(A+E)w_j$, $\|w_j\| = 1$, then

$$\min_i |\lambda_i(A) - \lambda_j(A+E)| \le \|Ew_j\|. \tag{3.2}$$

The problem is that this bound depends on the a priori unknown projection of E onto the perturbed eigenvector w_j. However, by switching the roles of A and $A+E$, we obtain, for each eigenvalue $\lambda_i(A)$,

$$\min_j |\lambda_i(A) - \lambda_j(A+E)| \leq \|Ev_i\|, \qquad 1 \leq i \leq n. \tag{3.3}$$

While this bound depends on the projection of E onto an eigenspace of A, it doesn't pair up $\lambda_i(A)$ with the corresponding perturbed eigenvalue $\lambda_i(A+E)$.

Below we show that such a pairing is possible for eigenvalues $\lambda_i(A)$ that are well separated from the other eigenvalues of A, and that the distance between $\lambda_i(A)$ and $\lambda_i(A+E)$ is bounded only by the projection of E onto the eigenspace of $\lambda_i(A)$, rather than by the full perturbation E. Now we use the two-sided eigenvalue separation,

$$\mathrm{Gap}_i \equiv \min_{j \neq i} |\lambda_i(A) - \lambda_j(A)|, \qquad 1 \leq i \leq n.$$

In the following lemma we present a bound that is probably known, but we were not able to find it in the literature.

LEMMA 3.1. *If $A, A+E \in \mathbb{C}^{n\times n}$ are Hermitian, then for every eigenvalue $\lambda_i(A)$ with* $\mathrm{Gap}_i > 2\|E\|$

$$|\lambda_i(A+E) - \lambda_i(A)| \leq \|Ev_i\|.$$

Proof. According to (3.3) for every eigenvalue $\lambda_i(A)$ there exists an eigenvalue $\lambda_j(A+E)$ such that

$$|\lambda_i(A) - \lambda_j(A+E)| \leq \|Ev_i\|.$$

We now prove that under the gap condition, $j = i$. Weyl's theorem implies

$$|\lambda_j(A) - \lambda_j(A+E)| \leq \|E\|.$$

Moreover, all other eigenvalues of A are further from $\lambda_i(A)$, because for $j \neq i$,

$$|\lambda_i(A) - \lambda_j(A+E)| \geq |\lambda_i(A) - \lambda_j(A)| - |\lambda_j(A) - \lambda_j(A+E)| \geq \mathrm{Gap}_i - \|E\| > \|E\|.$$

This means that for each eigenvalue $\lambda_i(A)$ satisfying the gap condition there is exactly one eigenvalue of $A+E$ at distance less than $\|E\|$, and this eigenvalue must be $\lambda_i(A+E)$. Therefore, $j = i$. □

The condition $\mathrm{Gap}_i > 2\|E\|$ appears in many other contexts, because it is a sufficient condition that prevents the eigenvalue $\lambda_i(A+\epsilon E)$ from crossing other eigenvalues for $|\epsilon| < 1$ [19, Theorem II.3.9]. Now we improve the gap condition in Lemma 3.1 by a factor of 2, but at the expense of a multiplicative factor of $\sqrt{2}$ in the perturbation bound.

THEOREM 3.2. *If $A, A+E \in \mathbb{C}^{n\times n}$ are Hermitian, then for every eigenvalue $\lambda_i(A)$ with* $\mathrm{Gap}_i > \|E\|$

$$|\lambda_i(A) - \lambda_i(A+E)| \leq \sqrt{2}\|Ev_i\|.$$

Proof. Fix an index i, and let

$$V^*AV = \begin{pmatrix} \lambda_i(A) & \\ & \Lambda_i \end{pmatrix}$$

be an eigenvalue decomposition of A, where V is unitary with leading column $Ve_1 = v_i$. Partition

$$V^*EV = \begin{pmatrix} f_{ii} & f^* \\ f & E_i \end{pmatrix}$$

conformally with V^*AV. Then we can write $V^*(A+E)V = M + F$, where

$$M \equiv \begin{pmatrix} \lambda_i(A) & \\ & \Lambda_i + E_i \end{pmatrix}, \qquad F \equiv \begin{pmatrix} f_{ii} & f^* \\ f & \end{pmatrix}.$$

From $\sqrt{|f_{ii}|^2 + 4\|f\|^2} \leq |f_{ii}| + 2\|f\|$ follows

$$\|F\| = \frac{1}{2}\left(|f_{ii}| + \sqrt{|f_{ii}|^2 + 4\|f\|^2}\right) \leq |f_{ii}| + \|f\| \leq \sqrt{2}\|Ev_i\|.$$

Weyl's theorem (3.1) implies

$$|\lambda_j(M) - \lambda_j(M+F)| \leq \|F\| \leq \sqrt{2}\|Ev_i\|, \qquad 1 \leq j \leq n.$$

Let $\lambda_i(A)$ be the kth eigenvalue of M so that $\lambda_k(M) = \lambda_i(A)$ and $|\lambda_i(A) - \lambda_k(A+E)| \leq \sqrt{2}\|Ev_i\|$. We now prove that if $\text{Gap}_i > \|E\|$, then $\lambda_i(A)$ is indeed the ith eigenvalue of M.

Assume $\text{Gap}_i > \|E\|$. Since $\lambda_j(A)$ for $j \neq i$ are the eigenvalues of Λ_i, we can write the eigenvalues of $\Lambda_i + E_i$ as $\lambda_j(A) + \gamma_j$ for $j \neq i$. Weyl's theorem (3.1) implies $|\gamma_j| \leq \|E_i\| \leq \|E\|$. For $i+1 \leq j \leq n$ we have

$$\lambda_i(A) - (\lambda_j(A) + \gamma_j) \geq \text{Gap}_i - \|E\| > 0,$$

and for $1 \leq j \leq i-1$

$$\lambda_j(A) + \gamma_j - \lambda_i(A) \geq \text{Gap}_i - \|E\| > 0.$$

This means there are exactly $n-i$ eigenvalues of M that are smaller than $\lambda_i(A)$, and $i-1$ eigenvalues that are larger. Thus $\lambda_i(A) = \lambda_i(M)$. □

For every eigenvalue $\lambda_i(A)$, Theorem 3.2 bounds the distance to a perturbed eigenvalue in terms of $\|Ev_i\|$. Since E is Hermitian, $\|Ev_i\| = \|v_i v_i^* E\|$ is the orthogonal projection of E onto the eigenspace of $\lambda_i(A)$.

The bound in Theorem 3.2 is tighter than (3.1) for a particular eigenvalue $\lambda_i(A)$ if the projection of E on the eigenspace of $\lambda_i(A)$ is small compared to $\|E\|$, i.e., if $\sqrt{2}\|Ev_i\| \leq \|E\|$. This is typically the case for random perturbations in high dimensions, as in principal component analysis. In contrast to (3.1), which matches up all eigenvalues, Theorem 3.2 bounds the distance between corresponding eigenvalues of A and $A+E$ only for those eigenvalues of A that are sufficiently well separated from all other eigenvalues of A.

4. Non-Hermitian perturbations. In section 3 we showed that a small Hermitian perturbation E of a Hermitian matrix A changes a well-separated eigenvalue $\lambda_i(A)$ by at most $\|Ev_i\|$ rather than by the full norm $\|E\|$. We extend this approach to general non-Hermitian perturbations E and obtain bounds that are comparable to those for Hermitian perturbations.

Since a non-Hermitian perturbation of a Hermitian matrix may lead to a nondiagonalizable matrix, there is relatively little work on eigenvalue bounds for non-Hermitian perturbations. A notable exception is the work by Kahan [18, 30], who proved that all eigenvalues of $A+E$ are included in the union of the regions

$$\{z \in \mathbb{C} : |z-\lambda_k| \leq \|E\| \text{ and } |\Im(z)| \leq \|(E-E^*)/2\|\}.$$

If one of these regions is isolated from the others, then it contains exactly one eigenvalue, and if both matrices A and E are real, then this eigenvalue must also be real. Another type of eigenvalue bound for general matrices is a Gershgorin theorem [31], which in the simplest form states that for a diagonal matrix $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, all eigenvalues of $A+E$ are in the union of the disks

$$\left\{ z \in \mathbb{C} : |z - \lambda_k - E_{kk}| \leq \sum_{j \neq k} |E_{kj}| \right\}.$$

Below we derive a bound that is sharper whenever a perturbed eigenvalue is close to a well-separated eigenvalue, where the separation condition involves the two-sided gap

$$\mathrm{Gap}_i \equiv \min_{j \neq i} |\lambda_i(A) - \lambda_j(A)|.$$

The bound is almost, but not quite, the same as the one for Hermitian perturbations in Theorem 3.2.

THEOREM 4.1. *Let $A, E \in \mathbb{C}^{n\times n}$, where A is Hermitian, let μ be a (possibly complex) eigenvalue of $A+E$, and let $\lambda_i(A)$ be an eigenvalue of A closest to μ, i.e.,*

$$|\lambda_i(A) - \mu| = \min_{1 \leq l \leq n} |\lambda_l(A) - \mu|.$$

If $\mathrm{Gap}_i > 3\|E\|$, *then*

$$|\lambda_i(A) - \mu| \leq \sqrt{5}\|Ev_i\|.$$

Proof. Abbreviate $\lambda_i \equiv \lambda_i(A)$, and let w be a unit norm eigenvector of μ, i.e., $(A+E)w = \mu w$, $\|w\| = 1$. By assumption λ_i is, among all eigenvalues of A, an eigenvalue that is closest to μ. Thus the Bauer–Fike theorem (3.2) applied to the Hermitian matrix A and the perturbed eigenvalue μ of the matrix $A+E$ yields

$$|\lambda_i - \mu| \leq \|Ew\|. \tag{4.1}$$

We now perform a similarity transformation of A that makes it possible to express the perturbed eigenvector w in terms of the exact eigenvector v_i.

Let $W = \begin{pmatrix} w & W_2 \end{pmatrix}$ be a unitary matrix and perform the similarity transformation

$$W^*AW = \begin{pmatrix} \mu - w^*Ew & b^* \\ b & M \end{pmatrix},$$

where $\|b\| = \|W_2^*Ew\| \leq \|Ew\|$. Isolate the block diagonal part, $W^*AW = D + F$, where

$$D \equiv \begin{pmatrix} \mu - w^*Ew & \\ & M \end{pmatrix}, \qquad F \equiv \begin{pmatrix} & b^* \\ b & \end{pmatrix}.$$

The matrix D is Hermitian because it consists of principal submatrices of the Hermitian matrix W^*AW; in particular the scalar $\mu - w^*Ew$ is real. We show in two steps that $\mu - w^*Ew$ is the ith eigenvalue of D.

1. Under the gap condition $\mathrm{Gap}_i > 3\|E\|$, among all eigenvalues of A, λ_i is the only eigenvalue closest to $\mu - w^*Ew$, with all other eigenvalues at a distance of at least $\|E\|$.
   To show this, apply the Bauer–Fike theorem (3.2) to the leading diagonal element of D to conclude that there exists an eigenvalue $\lambda_k(A)$ so that

$$|\lambda_k(A) - (\mu - w^*Ew)| \leq \|F\| \leq \|Ew\|.$$

   We show that $k = i$ by showing that $\mu - w^*Ew$ is too far away from all eigenvalues of A but λ_i. The gap condition implies for $j \neq i$

$$\begin{aligned} |\lambda_j(A) - (\mu - w^*Ew)| &= |\lambda_j(A) - \lambda_i + \lambda_i - \mu + w^*Ew| \\ &\geq |\lambda_j(A) - \lambda_i| - \|Ew\| - |w^*Ew| \\ &\geq \mathrm{Gap}_i - 2\|Ew\| > \|E\|. \end{aligned}$$

   Therefore λ_i is the only eigenvalue of A that is close to $\mu - w^*Ew$. Hence $k = i$ and

$$|\lambda_i - (\mu - w^*Ew)| \leq \|Ew\|. \tag{4.2}$$

2. $\mu - w^*Ew$ is the ith eigenvalue of D.
   As in the proof of Theorem 3.2 we show that this follows from the gap condition. Weyl's theorem (3.1),

$$|\lambda_l(A) - \lambda_l(D)| \leq \|F\| \leq \|Ew\|, \qquad 1 \leq l \leq n, \tag{4.3}$$

   allows us to write the eigenvalues of D as $\lambda_j(A) + \gamma_j$, where $|\gamma_j| \leq \|Ew\|$. Assuming $\mathrm{Gap}_i > 3\|E\|$, we have for $i+1 \leq j \leq n$

$$\begin{aligned} \mu - w^*Ew - (\lambda_j(A) + \gamma_j) &= (\mu - w^*Ew - \lambda_i) + (\lambda_i - \lambda_j(A)) + \gamma_j \\ &\geq -|\mu - w^*Ew - \lambda_i| + (\lambda_i - \lambda_j(A)) - |\gamma_j| \\ &\geq \mathrm{Gap}_i - 2\|Ew\| > 0, \end{aligned}$$

   where the last inequality follows from (4.2). Similarly for $1 \leq j \leq i-1$

$$\lambda_j(A) + \gamma_j - (\mu - w^*Ew) \geq \mathrm{Gap}_i - 2\|Ew\| > 0.$$

   This means there are exactly $n - i$ eigenvalues of D that are smaller than $\lambda_i(A)$, and $i - 1$ eigenvalues that are larger. Thus $\mu - w^*Ew = \lambda_i(D)$.

Because $\mu - w^*Ew$ is the ith eigenvalue of D, no eigenvalue of M can be the ith eigenvalue of D. We use this fact to express w in terms of v_i. The partitioning of W^*AW provides a 2×2 system from which one can solve for v_i. Abbreviate

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} w^*v_i \\ W_2^*v_i \end{pmatrix} = W^*v_i.$$

From $Av_i = \lambda_i v_i$ follows

$$0 = (W^*AW - \lambda_i I)W^*v_i = \begin{pmatrix} \mu - w^*Ew - \lambda_i & b^* \\ b & M - \lambda_i I \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}. \tag{4.4}$$

We show that $M-\lambda_i I$ is nonsingular by showing that λ_i cannot be an eigenvalue of M. Above we established that no eigenvalue of M can be the ith eigenvalue of D. Hence $\lambda_j(M)=\lambda_{j_k}(D)$ for some $j_k \neq i$, and (4.3) and the gap condition imply

$$\begin{aligned}|\lambda_j(M)-\lambda_i| &= |\lambda_{j_k}(D)-\lambda_i| = |(\lambda_{j_k}(A)-\lambda_i)+(\lambda_{j_k}(D)-\lambda_{j_k}(A))| \\ &\geq \mathrm{Gap}_i - \|E\| > 2\|E\| > 0.\end{aligned}$$

Thus λ_i is not an eigenvalue of M, and $M-\lambda_i I$ is nonsingular.

As a consequence we can solve for z_2 in (4.4) and obtain $z_2=-z_1(M-\lambda_i I)^{-1}b$. Since $z_1=0$ would imply $z_2=0$, and then $v_i=0$, we must have $z_1\neq 0$. From the definition of z_1 and z_2 follows

$$W^* v_i = z_1 \begin{pmatrix} 1 \\ (M-\lambda_i I)^{-1}b \end{pmatrix}.$$

Multiplying on the left by the unitary matrix $W=\begin{pmatrix} w & W_2\end{pmatrix}$ yields

$$v_i = z_1\left(w + W_2(M-\lambda_i I)^{-1}b\right), \qquad z_1 = 1/\left\|\begin{pmatrix} 1 \\ (M-\lambda_i I)^{-1}b \end{pmatrix}\right\|.$$

Solving for w gives

$$w=\sqrt{1+\|(M-\lambda_i I)^{-1}b\|^2}\, v_i - W_2(M-\lambda_i I)^{-1}b,$$

and a subsequent multiplication by E yields

$$Ew=\sqrt{1+\|(M-\lambda_i I)^{-1}b\|^2}\, Ev_i - EW_2(M-\lambda_i I)^{-1}b.$$

Thus

$$\|Ew\| \leq \sqrt{1+\|(M-\lambda_i I)^{-1}b\|^2}\, \|Ev_i\| + \|E\|\|(M-\lambda_i I)^{-1}b\|.$$

To bound $\|(M-\lambda_i I)^{-1}\|$ from above, we use the fact from item 2 that $\mu - w^*Ew$ is the ith eigenvalue of D. As a consequence the eigenvalues of M correspond to $\lambda_j(D)$ for $j\neq i$. This means there is a $k\neq i$ so that

$$\begin{aligned}1/\|(M-\lambda_i I)^{-1}\| &\geq \min_j |\lambda_j(M)-\lambda_i| = |\lambda_k(D)-\lambda_i| \geq |\lambda_i - \lambda_k(A)| - |\lambda_k(D)-\lambda_k(A)| \\ &\geq \mathrm{Gap}_i - \|E\| > 2\|E\|,\end{aligned}$$

where the next-to-last inequality follows from (4.3). Hence

$$\|(M-\lambda_i I)^{-1}b\| \leq \frac{\|Ew\|}{2\,\|E\|} \leq \frac{1}{2}.$$

At last, we substitute this into the above bound for $\|Ew\|$ to obtain

$$\|Ew\| \leq \frac{\sqrt{5}}{2}\|Ev_i\| + \|E\|\frac{\|Ew\|}{2\,\|E\|} \leq \frac{\sqrt{5}}{2}\|Ev_i\| + \frac{\|Ew\|}{2},$$

so $\|Ew\|\| \leq \sqrt{5}\|Ev_i\|$. The result follows by substituting this bound for $\|Ew\|$ in (4.1). $\square$

Acknowledgments. We thank Tim Kelley and Alastair Spence for helpful discussions, and we are grateful to Jan Brandts whose valuable suggestions greatly improved the presentation of the paper.

REFERENCES

[1] T. W. Anderson, *Asymptotic theory for principal component analysis*, Ann. Math. Statist., 34 (1963), pp. 122–148.

[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1984.

[3] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenvalue problem*, Numer. Math., 31 (1978), pp. 31–48.

[4] T. F. Chan, *Newton-like pseudo-arclength methods for computing simple turning points*, SIAM J. Sci. and Stat. Comput., 5 (1984), pp. 135–148.

[5] T. F. C. Chan and H. B. Keller, *Arc-length continuation and multi-grid techniques for nonlinear elliptic eigenvalue problems*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 173–194.

[6] F. Chatelin, *Valeurs Propres de Matrices*, Masson, Paris, 1986.

[7] K. I. Dickson, C. T. Kelley, I. C. F. Ipsen, and I. G. Kevrekidis, *Condition estimates for pseudo-arclength continuation*, SIAM J. Numer. Anal., 45 (2007), pp. 263–276.

[8] R. B. Dozier and J. W. Silverstein, *On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices*, J. Multivariate Anal., 98 (2007), pp. 678–694.

[9] M. Eaton and D. E. Tyler, *On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix*, Ann. Statist., 19 (1991), pp. 260–271.

[10] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

[11] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[12] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamic Equilibria*, SIAM, Philadelphia, 2000.

[13] W. Govaerts and J. D. Pryce, *A singular value inequality for block matrices*, Linear Algebra Appl., 125 (1989), pp. 141–148.

[14] M. Gu and S. C. Eisenstat, *A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1266–1276.

[15] J. E. Jackson, *A User's Guide to Principal Components*, Wiley, New York, 1991.

[16] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Statist., 29 (2001), pp. 295–327.

[17] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.

[18] W. Kahan, *Spectra of nearly Hermitian matrices*, Proc. Amer. Math. Soc., 48 (1975), pp. 11–17.

[19] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1995.

[20] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Theory*, Tata Inst. Fund. Res. Lectures on Math. and Phys. 79, Springer, New York, 1987.

[21] M. Krupnik, *Changing the spectrum of an operator by a perturbation*, Linear Algebra Appl., 167 (1992), pp. 113–118.

[22] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer, New York, 1998.

[23] R. Menzel and H. Schwetlick, *Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen*, Numer. Math., 30 (1978), pp. 65–79.

[24] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.

[25] B. Nadler, *Finite sample approximation results for principal component analysis: A matrix perturbation approach*, Ann. Statist., to appear.

[26] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.

[27] W. C. Rheinboldt, *Solution fields of nonlinear equations and continuation methods*, SIAM J. Numer. Anal., 17 (1980), pp. 221–237.

[28] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, New York, 1992.

[29] A. Spence, *Numerical methods for bifurcation problems*, in The Graduate's Guide to Numerical Analysis, Springer Ser. Comput. Math. 26, Springer, Berlin, 1999, pp. 177–216.

[30] G. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.

[31] R. S. Varga, *Geršgorin and His Circles*, Springer, Heidelberg, 2004.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 54–67

SUBSPACE GAP RESIDUALS FOR RAYLEIGH–RITZ APPROXIMATIONS[*]

NELA BOSNER[†] AND ZLATKO DRMAČ[‡]

Abstract. Large-scale eigenvalue and singular value computations are usually based on extracting information from a compression of the matrix to suitably chosen low dimensional subspaces. This paper introduces new a posteriori relative error bounds based on a residual expressed using the largest principal angle (gap) between relevant subspaces. The eigenvector approximations are estimated using subspace gaps and relative separation of the eigenvalues.

Key words. eigenvalues, Rayleigh–Ritz approximation, singular values, subspace gap

AMS subject classifications. 15A18, 15A23, 15A42, 65F15

DOI. 10.1137/070689425

1. Introduction.

Let $\mathsf{A} = U\Sigma V^*$ be the SVD of $\mathsf{A} \in \mathbb{C}^{m\times n}$, $m \geq n$. To approximate $\ell \leq \min\{m, n\}$ desired singular triplets (σ_i, u_i, v_i) $(\mathsf{A}v_i = u_i\sigma_i)$, one can deploy the standard Rayleigh–Ritz procedure to extract approximations from suitably chosen ℓ-dimensional subspaces $\mathcal{X}$, $\mathcal{Y}$ defined as images of orthonormal matrices $X \in \mathbb{C}^{m\times\ell}$, $Y \in \mathbb{C}^{n\times\ell}$, respectively. For given X, Y, optimal approximation is obtained using the SVD of the Rayleigh quotient compression $\mathsf{C} = X^*\mathsf{A}Y$: the singular values $\gamma_1 \leq \cdots \leq \gamma_\ell$ of C approximate some of the singular values $\sigma_1 \leq \cdots \leq \sigma_n$ of A. The approximation error is bounded by the spectral norms of the residuals $R = \mathsf{A}Y - X\mathsf{C}$, $L = \mathsf{A}^*X - Y\mathsf{C}^*$. More precisely, for some singular values $\sigma_{i_1}, \ldots, \sigma_{i_\ell}$ of A we have

$$\max_{1\leq j\leq\ell} |\sigma_{i_j} - \gamma_j| \leq \max\{\|R\|_2, \|L\|_2\}. \tag{1.1}$$

This is a generalization of the Kahan [7] theorem for Hermitian Rayleigh–Ritz procedure with Hermitian H and $Y = X$. The Rayleigh quotient matrix C is optimal in the sense that it minimizes the residual in both the Frobenius and the spectral norm. For details see, e.g., [13], [17]. The SVD residual bound (1.1) is used, e.g., in Jacobi–Davidson type SVD computation [5].

Let $X_\perp$, $Y_\perp$ be orthonormal matrices such that $\begin{pmatrix} X & X_\perp \end{pmatrix} \in \mathbb{C}^{m\times m}$ and $\begin{pmatrix} Y & Y_\perp \end{pmatrix} \in \mathbb{C}^{n\times n}$ are unitary, and let $\mathsf{C}_\perp = X_\perp^*\mathsf{A}Y_\perp$ be the corresponding Rayleigh quotient. Define $\tilde{\mathsf{A}} = X\mathsf{C}Y^* + X_\perp\mathsf{C}_\perp Y_\perp^*$. It is convenient to write A and $\tilde{\mathsf{A}}$ in the block-matrix

[*] Received by the editors April 25, 2007; accepted for publication (in revised form) by I. C. F. Ipsen February 21, 2008; published electronically February 27, 2009. This work was supported by the Croatian MZOS through grant 0372783–2750.

http://www.siam.org/journals/simax/31-1/68942.html

[†] Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia (nela@math.hr).

[‡] Department of Mathematics, Virginia Polytechnic Institute and State University, 460 McBryde, Blacksburg, VA 24061-0123. On leave from Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia (drmac@math.hr). This author was also supported by the NSF grants DMS 050597 and DMS 0513542 and Air Force Office of Scientific Research grant FA9550-05-1-0449.

form

$$\tilde{\mathsf{A}} = \begin{pmatrix} X & X_\perp \end{pmatrix} \begin{pmatrix} \mathsf{C} & 0 \\ 0 & \mathsf{C}_\perp \end{pmatrix} \begin{pmatrix} Y^* \\ Y_\perp^* \end{pmatrix} = X\mathsf{C}Y^* + X_\perp \mathsf{C}_\perp Y_\perp^*, \tag{1.2}$$

$$\mathsf{A} = \begin{pmatrix} X & X_\perp \end{pmatrix} \begin{pmatrix} \mathsf{C} & X^*\mathsf{A}Y_\perp \\ X_\perp^*\mathsf{A}Y & \mathsf{C}_\perp \end{pmatrix} \begin{pmatrix} Y^* \\ Y_\perp^* \end{pmatrix} = X\mathsf{C}Y^* + X_\perp \mathsf{C}_\perp Y_\perp^* + \mathsf{E}, \tag{1.3}$$

where $\mathsf{E} = \begin{pmatrix} X & X_\perp \end{pmatrix} \left(\begin{smallmatrix} 0 & X^*\mathsf{A}Y_\perp \\ X_\perp^*\mathsf{A}Y & 0 \end{smallmatrix} \right) \left(\begin{smallmatrix} Y^* \\ Y_\perp^* \end{smallmatrix} \right)$ can be expressed as

$$\mathsf{E} = X_\perp X_\perp^* \mathsf{A} Y Y^* + X X^* \mathsf{A} Y_\perp Y_\perp^* = (\mathsf{A}Y - X\mathsf{C})Y^* + X(\mathsf{A}^*X - Y\mathsf{C}^*)^*.$$

The perturbation E depends on the residuals $R = \mathsf{A}Y - X\mathsf{C}$, $L = \mathsf{A}^*X - Y\mathsf{C}^*$. It is easily checked that in the trace scalar product $\langle \mathsf{E}, \tilde{\mathsf{A}} \rangle_F = 0$. Further, $\|\mathsf{E}\|_2 = \max\{\|R\|_2, \|L\|_2\}$, and $X^*R = 0$, $Y^*L = 0$. The latter two equalities represent Galerkin conditions with test spaces equal to the search spaces $\mathcal{X}$, $\mathcal{Y}$.

Proposition 1.1. [illegible] $X\mathsf{C}Y^* = XX^*\mathsf{A}YY^*$ [illegible] A [illegible] $\mathcal{S}(X,Y) = \{XSY^*,\ S \in \mathbb{C}^{\ell\times\ell}\}$ [illegible]

$$\min_{S\in\mathbb{C}^{\ell\times\ell}} \|\mathsf{A} - XSY^*\|_F = \|\mathsf{A} - X\mathsf{C}Y^*\|_F = \sqrt{\|\mathsf{A}\|_F^2 - \|\mathsf{C}\|_F^2}.$$

[illegible] $\tilde{\mathsf{A}}$ [illegible] A [illegible] $\langle\cdot,\cdot\rangle_F$ [illegible] $\mathcal{S}(X,Y) \oplus_F \mathcal{S}(X_\perp, Y_\perp)$

[illegible]. In the Frobenius (trace) scalar product $\langle\cdot,\cdot\rangle_F$, $\mathsf{A} - X\mathsf{C}Y^*$ is orthogonal to XSY^*, $\langle \mathsf{A} - X\mathsf{C}Y^*, XSY^* \rangle_F = 0$ for all S. Similarly, $\langle \mathsf{A} - \tilde{\mathsf{A}}, XSY^* + X_\perp \tilde{S} Y_\perp^* \rangle_F = 0$ for all $S \in \mathbb{C}^{\ell\times\ell}$, $\tilde{S} \in \mathbb{C}^{(m-\ell)\times(n-\ell)}$. □

The key idea then is to consider the given A as perturbed $\tilde{\mathsf{A}}$ and to take advantage of the fact that we can compute the SVD of C. Since the singular values $\gamma_1 \leq \cdots \leq \gamma_\ell$ of C coincide with some of the singular values $\tilde{\sigma}_1 \leq \cdots \leq \tilde{\sigma}_n$ of $\tilde{\mathsf{A}}$ ($\gamma_j = \tilde{\sigma}_{i_j}$ for some i_j), the comparison of the γ_j and the some of the σ_i is reduced to singular value perturbation analysis under the structured perturbation of $\tilde{\mathsf{A}}$ to $\mathsf{A} = \tilde{\mathsf{A}} + \mathsf{E}$. An application of the classical perturbation theory yields

$$\max_{1\leq i\leq n} \frac{|\tilde{\sigma}_i - \sigma_i|}{\|\mathsf{A}\|_2} \leq \frac{\|\mathsf{E}\|_2}{\|\mathsf{A}\|_2} \Longrightarrow \max_{1\leq j\leq \ell} \frac{|\gamma_j - \sigma_{i_j}|}{\|\mathsf{A}\|_2} \leq \frac{\|\mathsf{E}\|_2}{\|\mathsf{A}\|_2}. \tag{1.4}$$

Note that we can write the residuals as $R = (I - XX^*)\mathsf{A}Y$, $L = (I - YY^*)\mathsf{A}^*X$. It is easily seen that both residuals are zero if and only if $\mathsf{A}\mathcal{Y} \subseteq \mathcal{X}$ and $\mathsf{A}^*\mathcal{X} \subseteq \mathcal{Y}$. In that case $(\mathcal{X}, \mathcal{Y})$ is called a [illegible] [15, Definition 6.1]. (Here $\mathsf{A}\mathcal{Y} \equiv \{\mathsf{A}y : y \in \mathcal{Y}\}$.) In the Hermitian case, a zero residual corresponds to an H-invariant subspace $\mathcal{X}$, $\mathsf{H}\mathcal{X} \subseteq \mathcal{X}$.

Our approach in this paper is based on the following observations:

(i) The angles $\sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y})$ and $\sphericalangle(\mathcal{Y}, \mathsf{A}^*\mathcal{X})$ are natural measures of the departure of $(\mathcal{X}, \mathcal{Y})$ from being a pair of singular subspaces. (Here the angle between two subspaces of the same dimension is defined as the largest principal angle. For basics on angles between subspaces see [17, I.5].) To illustrate, consider the right residual $R = (I - XX^*)\mathsf{A}Y$ with full rank $\mathsf{A}Y$, compute the thin QR factorization $\mathsf{A}Y = QT$, and note that $R(T^{-1}Q^*) = R(\mathsf{A}Y)^\dagger = (I - XX^*)QQ^*$. In other words, the size of the residual R "relative to $\mathsf{A}Y$," expressed as the product of projectors, is related to the gap between the subspaces $\mathcal{X}$ and $\mathsf{A}\mathcal{Y}$. Such a [illegible], in addition to the classical residuals, may be useful in extracting valuable information from given subspaces, or even in constructing subspace corrections. We note here that in large-scale computation, information is always extracted from a subspace, and its quality can be tested using another subspace (e.g., Petrov–Galerkin framework). So, even though we compute with particular bases, the underlying theory is most naturally expressed using subspaces.

(ii) Classical residual bounds give an estimate of the type (1.4); that is, the approximation error is measured relative to the matrix norm. It would be useful to have tight bounds for the relative errors $|\sigma_{i_j} - \gamma_j|/\gamma_j$, $1 \le j \le \ell$. However, for sharp relative error bounds, modern perturbation theory requires that the relative size of the perturbation is expressed as the norm of the quotient of matrices (e.g., $\|\delta\mathsf{A}\mathsf{A}^\dagger\|_2$ in the multiplicative form of the perturbation $\mathsf{A} + \delta\mathsf{A} = (\mathsf{I} + \delta\mathsf{A}\mathsf{A}^\dagger)\mathsf{A}$), rather than the quotient of the norms ($\|\delta\mathsf{A}\|_2/\|\mathsf{A}\|_2$ in the additive form $\mathsf{A} + \delta\mathsf{A}$). The multiplicative form leads to scaled residuals with norms expressed as functions of the angles between certain subspaces.

(iii) The error bounds for the eigenvectors and singular vectors should preferably depend on relative gaps in the spectrum. For instance, if a relative gap is used, two different eigenvalues λ_i and λ_j of a Hermitian positive definite matrix H are considered well separated if $|\lambda_i - \lambda_j|/\sqrt{\lambda_i\lambda_j}$ is not too small. Take, e.g., $\lambda_i = 10^{-10}$, $\lambda_j = 2 \cdot 10^{-10}$. At the same time, if the absolute gap $|\lambda_i - \lambda_j|/\|\mathsf{H}\|_2$ is used, these two eigenvalues are considered pathologically close if both are much smaller that $\|\mathsf{H}\|_2$ (take here, e.g., $\|\mathsf{H}\|_2 = 1$). For the more favorable relative separation and better estimates for the Ritz vectors in such cases, the residual must be scaled as discussed above.

Such considerations are introduced in [1], [2], and here we improve them, provide new bounds for the Ritz vectors and harmonic Ritz values, and generalize them to the singular value decomposition.

2. New tan Θ residual bounds. The main shortcoming of estimate (1.4) is that it does not take advantage of the geometric structure of the perturbation. We exploit that structure and show how the angles $\sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y})$ and $\sphericalangle(\mathcal{Y}, \mathsf{A}^*\mathcal{X})$ naturally appear in the residual bounds.

PROPOSITION 2.1. [illegible] $\mathcal{X} \bigcap \mathrm{Im}(\mathsf{A})^\perp = \{0\}$ [illegible] $\sphericalangle(\mathsf{A}^*\mathcal{X}, \mathcal{Y}) < \pi/2$ [illegible] $\mathsf{C} = X^*\mathsf{A}Y$ [illegible]

[illegible] Because of $\mathrm{Im}(\mathsf{A})^\perp = \mathrm{Ker}(\mathsf{A}^*)$, A^*X is of full column rank. Using the thin QR factorization of A^*X, $\mathsf{A}^*X = QT$ (T $\ell \times \ell$ nonsingular), we can write $\mathsf{C} = T^*(Q^*Y)$. Now the second assumption implies nonsingularity of Q^*Y. □

THEOREM 2.2. [illegible] C [illegible] $\gamma_1 \le \cdots \le \gamma_\ell$ [illegible] C [illegible] ℓ [illegible] $\sigma_{i_1} \le \cdots \le \sigma_{i_\ell}$ [illegible] A [illegible]

$$\max_{1\le j\le \ell} \frac{|\sigma_{i_j} - \gamma_j|}{\sqrt{\sigma_{i_j}\gamma_j}} \le \frac{1}{2}\tan\sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y}) + \frac{1}{2}\tan\sphericalangle(\mathcal{Y}, \mathsf{A}^*\mathcal{X}). \tag{2.1}$$

[illegible] Let, in (1.3), $F = X^*\mathsf{A}Y_\perp$, $G = X_\perp^*\mathsf{A}Y$. The block elimination

$$\begin{pmatrix} I & 0 \\ -G\mathsf{C}^{-1} & I \end{pmatrix} \begin{pmatrix} \mathsf{C} & F \\ G & \mathsf{C}_\perp \end{pmatrix} \begin{pmatrix} I & -\mathsf{C}^{-1}F \\ 0 & I \end{pmatrix} = \begin{pmatrix} \mathsf{C} & 0 \\ 0 & \mathsf{C}_\perp - G\mathsf{C}^{-1}F \end{pmatrix} \tag{2.2}$$

represents a multiplicative perturbation with the corresponding additive form (cf. (1.2)) $\breve{\mathsf{A}} = X\mathsf{C}Y^* + X_\perp\mathsf{C}_\perp Y_\perp^* - X_\perp X_\perp^*\mathsf{A}Y(X^*\mathsf{A}Y)^{-1}X^*\mathsf{A}Y_\perp Y_\perp^*$. If $\breve{\sigma}_1 \le \cdots \le \breve{\sigma}_n$ are the singular values of $\breve{\mathsf{A}}$, then $\gamma_j = \breve{\sigma}_{i_j}$, $j = 1, \ldots, \ell$, and by [9, Corollary 5.5],

$$\max_{1\le i\le n} \frac{|\breve{\sigma}_i - \sigma_i|}{\sqrt{\breve{\sigma}_i \sigma_i}} \le \frac{1}{2}\|G\mathsf{C}^{-1}\|_2 + \frac{1}{2}\|\mathsf{C}^{-1}F\|_2.$$

It remains to give a geometric interpretation to the [illegible] $G\mathsf{C}^{-1}$ and $\mathsf{C}^{-1}F$. If $\mathsf{A}^*X = QT$ is the QR factorization, then $\mathsf{C}^{-1}F = (Q^*Y)^{-1}Q^*Y_\perp$. From the CS decomposition of the partitioned matrix $\begin{pmatrix} Y & Y_\perp \end{pmatrix}^* Q$, we conclude that $\|\mathsf{C}^{-1}F\|_2 = \tan\sphericalangle(\mathsf{A}^*\mathcal{X}, \mathcal{Y})$. Similarly, $\|G\mathsf{C}^{-1}\|_2 = \tan\sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y})$. □

Theorem 2.3. [illegible] C [illegible] $\gamma_1 \le \cdots \le \gamma_\ell$ [illegible] C [illegible] ℓ [illegible] $\sigma_{i_1} \le \cdots \le \sigma_{i_\ell}$ [illegible] A [illegible]

(2.3)

$$\max_{1\le j\le \ell} \frac{|\sigma_{i_j} - \gamma_j|}{\min(\sigma_{i_j}, \gamma_j)} \le \tan\sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y}) + \tan\sphericalangle(\mathcal{Y}, \mathsf{A}^*\mathcal{X}) + \tan\sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y}) \cdot \tan\sphericalangle(\mathcal{Y}, \mathsf{A}^*\mathcal{X}).$$

[illegible]. Relation (2.2) can be equivalently stated as

$$\begin{pmatrix} \mathsf{C} & F \\ G & \mathsf{C}_\perp \end{pmatrix} = \underbrace{\begin{pmatrix} I & 0 \\ G\mathsf{C}^{-1} & I \end{pmatrix}}_{S_1} \begin{pmatrix} \mathsf{C} & 0 \\ 0 & \mathsf{C}_\perp - G\mathsf{C}^{-1}F \end{pmatrix} \underbrace{\begin{pmatrix} I & \mathsf{C}^{-1}F \\ 0 & I \end{pmatrix}}_{S_2} \tag{2.4}$$

and an application of [3, Theorem 3.1] to (2.2) and (2.4) yields

$$\sigma_{\min}(S_1^{-1})\, \sigma_{\min}(S_2^{-1})\, \sigma_i \le \breve{\sigma}_i \le \sigma_i\, \sigma_{\max}(S_1^{-1})\, \sigma_{\max}(S_2^{-1})$$
$$\sigma_{\min}(S_1)\, \sigma_{\min}(S_2)\, \breve{\sigma}_i \le \sigma_i \le \breve{\sigma}_i\, \sigma_{\max}(S_1)\, \sigma_{\max}(S_2).$$

Now, $\sigma_{\max}(S_1) \le 1 + \|G\mathsf{C}^{-1}\|_2$, $\sigma_{\max}(S_2) \le 1 + \|\mathsf{C}^{-1}F\|_2$ and similar bounds with S_1^{-1}, S_2^{-1} give

$$\max_{1\le i\le n} \frac{|\breve{\sigma}_i - \sigma_i|}{\min(\breve{\sigma}_i, \sigma_i)} \le \|G\mathsf{C}^{-1}\|_2 + \|\mathsf{C}^{-1}F\|_2 + \|G\mathsf{C}^{-1}\|_2\|\mathsf{C}^{-1}F\|_2.$$

Finally, we note that with some indices $i_1, \ldots, i_\ell$, $\gamma_j = \breve{\sigma}_{i_j}$, $j = 1, \ldots, \ell$. □

If $\mathcal{U}$ and $\mathcal{V}$ are an exact pair of singular subspaces, then $\mathsf{A}\mathcal{V} \subseteq \mathcal{U}$. Thus, it seems natural to try to enforce this relation for the approximate pair $\mathcal{X}$, $\mathcal{Y}$, e.g., by taking $\mathcal{X} = \mathsf{A}\mathcal{Y}$. For given complex orthonormal Y, the corresponding X is defined by the QR factorization $\mathsf{A}Y = XP$. Without loss of generality we assume that $\mathsf{A}Y$ is of full column rank. (Else, a rank revealing QR factorization of $\mathsf{A}Y$ can be used to select columns $\hat{Y}$ of Y such that $\mathsf{A}\hat{Y}$ has full column rank.)

Corollary 2.4. [illegible] $\mathcal{X} = \mathsf{A}\mathcal{Y}$ [illegible] C [illegible] $\gamma_1 \le \cdots \le \gamma_\ell$ [illegible] C [illegible] ℓ [illegible] $\sigma_{i_1} \le \cdots \le \sigma_{i_\ell}$ [illegible] A [illegible]

(2.5)

$$(a)\quad \max_{1\le j\le \ell} \frac{|\sigma_{i_j} - \gamma_j|}{\sqrt{\sigma_{i_j}\gamma_j}} \le \frac{1}{2}\tan\sphericalangle(\mathsf{A}^*\mathcal{X}, \mathcal{Y}), \quad (b)\quad \max_{1\le j\le \ell} \frac{|\sigma_{i_j} - \gamma_j|}{\min\{\sigma_{i_j}, \gamma_j\}} \le \tan\sphericalangle(\mathsf{A}^*\mathcal{X}, \mathcal{Y}).$$

[illegible]. For given complex orthonormal Y, the corresponding X is defined by the QR factorization $\mathsf{A}Y = XP$. In that case $\mathsf{C} = X^*\mathsf{A}Y = P$, $G \equiv X_\perp^*\mathsf{A}Y = 0$, and $R = \mathsf{A}Y - X\mathsf{C} = 0$. Hence, (2.5.a), (2.5.b) hold as special cases of Theorem 2.2 and Theorem 2.3, respectively. □

[illegible] 2.5. In (2.4), if $G = \mathsf{0}$, then $S_1 = I$, $S_2 = \left(\begin{smallmatrix} I & \mathsf{C}^{-1}F \\ 0 & I \end{smallmatrix}\right)$. Let $\xi = \|\mathsf{C}^{-1}F\|_2$. Then $\sigma_{\max}(S_2) = \sqrt{1+\xi^2/2+\xi\sqrt{1+\xi^2/4}}$, and for small $\sphericalangle(\mathsf{A}^*\mathcal{X},\mathcal{Y})$, the relation (2.5)(b) reads

$$\max_{1\le i\le n}\frac{|\tilde{\sigma}_i-\sigma_i|}{\min\{\tilde{\sigma}_i,\sigma_i\}} \le \frac{1}{2}\tan\sphericalangle(\mathsf{A}^*\mathcal{X},\mathcal{Y}) + O(\tan^2\sphericalangle(\mathsf{A}^*\mathcal{X},\mathcal{Y})).$$

This shows that the above asymptotic formula is as sharp as (2.5)(a).

2.1. Using the complements $\mathcal{X}^\perp$, $\mathcal{Y}^\perp$. Next, we show that the complements of the search spaces contain information that can be accessed using the generalized inverse of A. In general, the matrix $\mathsf{C}_\perp = X_\perp^*\mathsf{A}Y_\perp$ is rectangular $(m-\ell)\times(n-\ell)$, and the following technical lemmas are used to reduce the general case to a square nonsingular one.

Lemma 2.6. [illegible] $\mathcal{X}\subseteq \mathrm{Im}(\mathsf{A})$ [illegible] $X_\perp$ [illegible] $\mathsf{C}_\perp = \left(\begin{smallmatrix}\hat{\mathsf{C}}_\perp \\ 0\end{smallmatrix}\right)$ [illegible] $G \equiv X_\perp^*\mathsf{A}Y = \left(\begin{smallmatrix}\hat{G} \\ 0\end{smallmatrix}\right)$ [illegible] $\alpha = \mathrm{rank}(\mathsf{A})$ $\hat{\mathsf{C}}_\perp \in \mathbb{C}^{(\alpha-\ell)\times(n-\ell)}$ [illegible] $\hat{G}\in\mathbb{C}^{(\alpha-\ell)\times\ell}$

[illegible]. Let $X_\perp$ be constructed so that in the block partition $X_\perp = \begin{pmatrix} X_{\perp,1} & X_{\perp,2}\end{pmatrix}$ the $\alpha-\ell$ columns of $X_{\perp,1}$ span the part of the orthogonal complement of $\mathcal{X}$ inside $\mathrm{Im}(\mathsf{A})$, and $X_{\perp,2}^*\mathsf{A} = \mathsf{0}$. □

Lemma 2.7. [illegible] $\mathcal{X}\subseteq \mathrm{Im}(\mathsf{A})$ [illegible] $\mathsf{A}^\dagger\mathcal{X} = (\mathsf{A}^*\mathcal{X}_{\perp,1}\oplus\mathrm{Ker}(\mathsf{A}))^\perp$ [illegible] A [illegible] $\mathsf{A}^{-1}\mathcal{X} = (\mathsf{A}^*\mathcal{X}^\perp)^\perp$

Lemma 2.8. [illegible] $\mathcal{X}\subseteq\mathrm{Im}(\mathsf{A})$ [illegible] $\sphericalangle(\mathsf{A}^\dagger\mathcal{X},\mathcal{Y}) < \pi/2$ [illegible] $\hat{\mathsf{C}}_\perp = X_{\perp,1}^*\mathsf{A}Y_\perp$ [illegible] 2.6 [illegible] $\mathrm{rank}(\hat{\mathsf{C}}_\perp) = \alpha-\ell$ [illegible] $\alpha = n$ $\hat{\mathsf{C}}_\perp$ [illegible]

[illegible]. Then $\hat{\mathsf{C}}_\perp = X_{\perp,1}^*\mathsf{A}Y_\perp = (\mathsf{A}^*X_{\perp,1})^*Y_\perp$, where the $(n\times(\alpha-\ell))$ matrix $\mathsf{A}^*X_{\perp,1}$ has full column rank. Let $A^*X_{\perp,1} = Q_1T_1$ be a thin QR factorization, where T_1 is $(\alpha-\ell)\times(\alpha-\ell)$ nonsingular. Then $\hat{\mathsf{C}}_\perp = T_1^*(Q_1^*Y_\perp)$, where $Q_1^*Y_\perp$ is full row rank by the assumption $\sphericalangle(\mathsf{A}^\dagger\mathcal{X},\mathcal{Y}) < \pi/2$ and by Lemma 2.7. □

Lemma 2.9. [illegible] $\mathcal{Y}\subseteq\mathrm{Im}(\mathsf{A}^*)$ [illegible] $Y_\perp$ [illegible] $F = \begin{pmatrix}\hat{F} & \mathsf{0}\end{pmatrix}$ $\hat{\mathsf{C}}_\perp = \begin{pmatrix}\check{\mathsf{C}}_\perp & \mathsf{0}\end{pmatrix}$ [illegible] $\hat{F}\in\mathbb{C}^{\ell\times(\alpha-\ell)}$ [illegible] $\check{\mathsf{C}}_\perp\in\mathbb{C}^{(\alpha-\ell)\times(\alpha-\ell)}$

[illegible]. Let $Y_\perp$ be such that in the partition $Y_\perp = \begin{pmatrix}Y_{\perp,1} & Y_{\perp,2}\end{pmatrix}$ the $\alpha-\ell$ columns of $Y_{\perp,1}$ span the part of $\mathcal{Y}^\perp$ inside $\mathrm{Im}(\mathsf{A}^*)$, and $\mathsf{A}Y_{\perp,2} = \mathsf{0}$ □

Theorem 2.10. [illegible] $\mathcal{X}\subseteq\mathrm{Im}(\mathsf{A})$ $\mathcal{Y}\subseteq\mathrm{Im}(\mathsf{A}^*)$ [illegible] $\sphericalangle(\mathsf{A}^*\mathcal{X},\mathcal{Y}) < \frac{\pi}{2}$ $\sphericalangle(\mathsf{A}^\dagger\mathcal{X},\mathcal{Y}) < \pi/2$ [illegible] $\gamma_1\le\cdots\le\gamma_\ell$ [illegible] C [illegible] ℓ [illegible] $\sigma_{i_1}\le\cdots\le\sigma_{i_\ell}$ [illegible] A [illegible]

$$\max_{1\le j\le\ell}\frac{|\sigma_{i_j}-\gamma_j|}{\gamma_j} \le \max\{\tan\sphericalangle(\mathsf{A}^*\mathcal{X},\mathcal{Y}),\tan\sphericalangle(\mathsf{A}^\dagger\mathcal{X},\mathcal{Y})\}. \tag{2.6}$$

[illegible]. Using Lemmas 2.6, 2.8, and 2.9, we represent A as

$$\mathsf{A} = \begin{pmatrix} X & X_{\perp,1} & X_{\perp,2}\end{pmatrix}\begin{pmatrix}\mathsf{C} & \hat{F} & \mathsf{0}\\ \hat{G} & \check{\mathsf{C}}_\perp & \mathsf{0}\\ \mathsf{0} & \mathsf{0} & \mathsf{0}\end{pmatrix}\begin{pmatrix}Y^*\\ Y_{\perp,1}^*\\ Y_{\perp,2}^*\end{pmatrix},$$

where $\hat{F} = X^*\mathsf{A}Y_{\perp,1}\in\mathbb{C}^{\ell\times(\alpha-\ell)}$, $\hat{G} = X_{\perp,1}^*\mathsf{A}Y\in\mathbb{C}^{(\alpha-\ell)\times\ell}$, and $\check{\mathsf{C}}_\perp = X_{\perp,1}^*\mathsf{A}Y_{\perp,1}\in\mathbb{C}^{(\alpha-\ell)\times(\alpha-\ell)}$ is nonsingular. (If $\mathsf{A}^*X_{\perp,1} = Q_1T_1$ is the QR factorization of $\mathsf{A}^*X_{\perp,1}$, then $\sigma_{\min}(Q_1^*Y_{\perp,1}) \ge \cos\sphericalangle(\mathsf{A}^*\mathcal{X}_{\perp,1}\oplus\mathrm{Ker}(\mathsf{A}),\mathcal{Y}^\perp) = \cos\sphericalangle(\mathsf{A}^\dagger\mathcal{X},\mathcal{Y}) > 0$.) Now, we can write

$$\begin{pmatrix} \mathsf{C} & \hat{F} & 0 \\ \hat{G} & \check{\mathsf{C}}_\perp & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathsf{C} & 0 & 0 \\ 0 & \check{\mathsf{C}}_\perp & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} I_\ell & \mathsf{C}^{-1}\hat{F} & 0 \\ \check{\mathsf{C}}_\perp^{-1}\hat{G} & I_{\alpha-\ell} & 0 \\ 0 & 0 & I_{n-\alpha} \end{pmatrix}, \tag{2.7}$$

where $\Omega = \left(\begin{smallmatrix} I_\ell & \mathsf{C}^{-1}\hat{F} \\ \check{\mathsf{C}}_\perp^{-1}\hat{G} & I_{\alpha-\ell} \end{smallmatrix}\right) \oplus I_{n-\alpha}$ is nonsingular with $1 - \epsilon \le \sigma_{\min}(\Omega) \le \sigma_{\max}(\Omega) \le 1 + \epsilon$, $\epsilon = \max\{\|\mathsf{C}^{-1}\hat{F}\|_2, \|\check{\mathsf{C}}_\perp^{-1}\hat{G}\|_2\}$. If the postmultiplication by Ω in (2.7) is taken as a multiplicative perturbation, then we can apply [3, Theorem 3.1] to conclude $\sigma_{\min}(\Omega) \le \sigma_{i_j}/\gamma_j \le \sigma_{\max}(\Omega)$, and thus

$$\max_{1\le j\le \ell} \frac{|\gamma_j - \sigma_{i_j}|}{\gamma_j} \le \epsilon. \tag{2.8}$$

Here by Theorem 2.2, $\|\mathsf{C}^{-1}\hat{F}\|_2 = \|\mathsf{C}^{-1}F\|_2 = \tan\sphericalangle(\mathsf{A}^*\mathcal{X}, \mathcal{Y})$. It remains to estimate $\check{\mathsf{C}}_\perp^{-1}\hat{G} = (Q_1^*Y_{\perp,1})^{-1}Q_1^*Y$. We first note that $\|(Q_1^*Y_{\perp,1})^{-1}Q_1^*Y\|_2 = \|Y^*Q_1(Y_\perp^*Q_1)^\dagger\|_2$. Using the CS decomposition $Y^*Q_1 = W_1\Phi Z$, $Y_\perp^*Q_1 = W_2\Psi Z$, we can conclude that $\|(Q_1^*Y_{\perp,1})^{-1}Q_1^*Y\|_2 \le \|\Phi\Psi^\dagger\|_2 = \tan\sphericalangle(\mathsf{A}^\dagger\mathcal{X}, \mathcal{Y})$. Finally, note that (2.8) provides useful information for $\epsilon < 1$, that is, if both angles in question are less than $\pi/4$. □

3. Using the Gram matrix of *A*. The SVD of A is conveniently analyzed using the Gram matrices $\mathsf{A}^*\mathsf{A}$ and $\mathsf{A}\mathsf{A}^*$. For simplicity, as in section 2.1, we may assume that A is square and nonsingular and we will consider the eigenvalue problem of $\mathsf{H} = \mathsf{A}^*\mathsf{A}$. By the QR factorization of A, H can be represented in the form of the Cholesky factorization $\mathsf{H} = LL^*$.

To introduce the key quantities for residual estimates in the Hermitian positive definite case, we start with a theorem which gives a simple and elegant proof of the existing linear residual bounds from [1], [2].

THEOREM 3.1. [illegible] H [illegible] $\mathsf{H} = LL^*$ [illegible] $0 < \lambda_1 \le \cdots \le \lambda_n$ [illegible] $\mathcal{X}$ [illegible] ℓ [illegible] $\mathbb{C}^n$ [illegible] $X \in \mathbb{C}^{n\times\ell}$ [illegible] $\mathsf{M} = X^*\mathsf{H}X$ [illegible] $R = \mathsf{H}X - X\mathsf{M}$ [illegible] $\delta\mathsf{H} = RX^* + XR^*$ [illegible] H [illegible] $\tilde{\mathsf{H}} = \mathsf{H} - \delta\mathsf{H}$ [illegible] $\tilde{\mathsf{H}}$ [illegible] $0 < \tilde{\lambda}_1 \le \cdots \le \tilde{\lambda}_n$ [illegible]

$$\max_{1\le i\le n} \frac{|\lambda_i - \tilde{\lambda}_i|}{\tilde{\lambda}_i} \le \sin\sphericalangle(L^*\mathcal{X}, L^{-1}\mathcal{X}), \tag{3.1}$$

[illegible] $\psi \equiv \sphericalangle(L^*\mathcal{X}, L^{-1}\mathcal{X}) = \sphericalangle(L^*\mathcal{X}, (L^*\mathcal{X}^\perp)^\perp) = \sphericalangle(L^*\mathcal{X}^\perp, L^{-1}\mathcal{X}^\perp)$ [illegible] L [illegible] L [illegible] H [illegible]. Let $X_\perp$ be an orthonormal matrix spanning $\mathcal{X}^\perp$ and let

$$\hat{\mathsf{H}} = \begin{pmatrix} X & X_\perp \end{pmatrix}^* \mathsf{H} \begin{pmatrix} X & X_\perp \end{pmatrix}, \; \mathsf{W} = X_\perp^*\mathsf{H}X_\perp, \; K = X_\perp^*\mathsf{H}X = (L^*X_\perp)^*(L^*X).$$

Then

$$\hat{\mathsf{H}} = \begin{pmatrix} \mathsf{M}^{1/2} & 0 \\ 0 & \mathsf{W}^{1/2} \end{pmatrix} \begin{pmatrix} I & \mathsf{M}^{-1/2}K^*\mathsf{W}^{-1/2} \\ \mathsf{W}^{-1/2}K\mathsf{M}^{-1/2} & I \end{pmatrix} \begin{pmatrix} \mathsf{M}^{1/2} & 0 \\ 0 & \mathsf{W}^{1/2} \end{pmatrix} \cong \mathsf{H},$$

where $\cong$ denotes (unitary) similarity. Let $\Psi \equiv \mathsf{W}^{-1/2}K\mathsf{M}^{-1/2}$ have SVD $\Psi = U_\psi\Sigma_\psi V_\psi^*$, $\Sigma_\psi = \mathrm{diag}(\cos\psi_i)_{i=1}^{\min(\ell,n-\ell)}$. The angles ψ_i are the acute principal angles between $L^*\mathcal{X}$ and $L^*\mathcal{X}^\perp$. It is easily checked that $\sin\psi = \|\Psi\|_2 < 1$ and that the eigenvalues of $\left(\begin{smallmatrix} I & \Psi^* \\ \Psi & I \end{smallmatrix}\right)$ are $1 \pm \psi_i$, $1 \le i \le \min(\ell, n-\ell)$ and 1 with multiplicity $n - 2\min(\ell, n-\ell)$. Then

$$\mathsf{H} \cong \hat{\mathsf{H}} \cong \begin{pmatrix} I & \Psi^* \\ \Psi & I \end{pmatrix}^{1/2} \begin{pmatrix} \mathsf{M} & 0 \\ 0 & \mathsf{W} \end{pmatrix} \begin{pmatrix} I & \Psi^* \\ \Psi & I \end{pmatrix}^{1/2} \quad \text{and} \quad \begin{pmatrix} \mathsf{M} & 0 \\ 0 & \mathsf{W} \end{pmatrix} \cong \tilde{\mathsf{H}}.$$

This argument is closed by an application of the Ostrowski theorem [6, Theorem 4.5.9]. □

[illegible] 3.2. An interesting feature of Theorem 3.1 is that it simultaneously estimates the quality of $\mathcal{X}$ and $\mathcal{X}^\perp$, with the same error bound which measures how much $\mathcal{X}$ or $\mathcal{X}^\perp$ moves under the action of H. This respects the fact that in the Hermitian case H-invariance of $\mathcal{X}$ is equivalent to H-invariance of $\mathcal{X}^\perp$. Further, in the H-scalar product ($<x, y>_{\mathsf{H}} = y^* \mathsf{H} x$), the angle $\sphericalangle(L^*\mathcal{X}, L^{-1}\mathcal{X})$ corresponds to the angle $\sphericalangle_{\mathsf{H}}(\mathcal{X}, \mathsf{H}^{-1}\mathcal{X})$. (For angles between subspaces in the H-scalar product see [8].) Similarly, in the H^{-1}-scalar product, $\sphericalangle(L^*\mathcal{X}, L^{-1}\mathcal{X}) = \sphericalangle_{\mathsf{H}^{-1}}(\mathcal{X}, \mathsf{H}\mathcal{X})$. This shows that ψ is a very natural gap measure for departure from H-invariance.

The gap residual $\sin\sphericalangle(L^*\mathcal{X}, L^{-1}\mathcal{X})$ also naturally measures the distance (in the subspace gap metric) between $\mathcal{X}$ and certain H-invariant subspace.

Theorem 3.3. [illegible] H [illegible] $\tilde{\mathsf{H}}$ [illegible] 3.1 [illegible]

$$\mathsf{H} = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix}, \quad \tilde{\mathsf{H}} = \begin{pmatrix} X & X_\perp \end{pmatrix} \begin{pmatrix} \mathsf{M} & 0 \\ 0 & \mathsf{W} \end{pmatrix} \begin{pmatrix} X^* \\ X_\perp^* \end{pmatrix},$$

[illegible] $(U_1 \quad U_2)$ [illegible] $(X \quad X_\perp)$ [illegible] $(\lambda_i^{(1)})_{i=1}^{\ell}$ [illegible] Λ_1 [illegible] $(\omega_j)_{j=1}^{n-\ell}$ [illegible] W [illegible] $\gamma = \min_{i,j} |\ln \frac{\lambda_i^{(1)}}{\omega_j}|$ [illegible]

$$\sin\sphericalangle(\mathrm{Im}(U_1), \mathcal{X}) \le \frac{\pi}{\gamma} \frac{\sin\psi}{2\sqrt{1-\sin\psi}}, \quad \psi = \sphericalangle(L^*\mathcal{X}, L^{-1}\mathcal{X}).$$

[illegible]. The proof is based on a perturbation result by Li [11, Theorem 1], who elegantly used the Sylvester equation with a structured right-hand side. In our case, the perturbation itself is nicely structured and we exploit that structure. First note that $\tilde{\mathsf{H}} = L(I - L^{-1}\delta H L^{-*})L^* = L\Xi\Xi^* L^*$, where Ξ is the square root of $I - L^{-1}\delta H L^{-*}$. Let $\tilde{L} = L\Xi$. Then, following [11], $\tilde{\mathsf{H}} - \mathsf{H} = \tilde{L}(\Xi^* - \Xi^{-1})L^*$ and premultiplying by $X_\perp^*$ and postmultiplying by U_1 we obtain $\mathsf{W}(X_\perp^* U_1) - (X_\perp^* U_1)\Lambda_1 = X_\perp^* \tilde{L}(\Xi^* - \Xi^{-1})L^* U_1$. Clearly, $X_\perp^* \tilde{L} = \mathsf{W}^{1/2}\tilde{Q}^*$ with some orthonormal $\tilde{Q}$. In the same way, $L^* U_1 = Q\Lambda_1^{1/2}$ with some orthonormal Q.

Consider the structure of the scaled perturbation $\Omega \equiv L^{-1}\delta\mathsf{H}L^{-*}$. (Recall that $\mathcal{X}$ is $\tilde{\mathsf{H}}$-invariant and $X^*\tilde{\mathsf{H}}X = \mathsf{M}$.) Set $Y = L^*X$, $Z = L^{-1}X$, $\mathcal{Y} = \mathrm{Im}(Y)$, $\mathcal{Z} = \mathrm{Im}(Z)$. It holds that $\mathcal{Y} \bigcap \mathcal{Z}^\perp = \mathcal{Z} \bigcap \mathcal{Y}^\perp = \{0\}$ and

$$\Omega = (I - ZY^*)YZ^* + ZY^*(I - YZ^*) = (I - P_{\mathcal{Z},\mathcal{Y}})P_{\mathcal{Y},\mathcal{Z}} + P_{\mathcal{Z},\mathcal{Y}}(I - P_{\mathcal{Y},\mathcal{Z}}),$$

where $P_{\mathcal{Y},\mathcal{Z}} = (P_{\mathcal{Z}}P_{\mathcal{Y}})^\dagger$, $P_{\mathcal{Z},\mathcal{Y}} = (P_{\mathcal{Y}}P_{\mathcal{Z}})^\dagger$ are the oblique projections (e.g., $P_{\mathcal{Y},\mathcal{Z}}$ projects on $\mathcal{Y}$ along $\mathcal{Z}^\perp$). In fact, in a suitable orthogonal basis $\mathcal{B}$ of $\mathbb{C}^n$, the projectors $P_{\mathcal{Y}}$ and $P_{\mathcal{Z}}$ can be represented as (see Wedin [19])

$$P_{\mathcal{Y}} = \left(\begin{array}{c|c|c} I_k & & \\ \hline & \bigoplus_{i=1}^{d} \Psi_i^0 & \\ \hline & & J_Y \end{array}\right), \quad P_{\mathcal{Z}} = \left(\begin{array}{c|c|c} I_k & & \\ \hline & \bigoplus_{i=1}^{d} \Psi_i & \\ \hline & & J_Z \end{array}\right), \quad \boxed{k = \dim(\mathcal{Y} \bigcap \mathcal{Z})},$$

where $\Psi_i^0 = \binom{1}{0}\begin{pmatrix}1 & 0\end{pmatrix}$, $\Psi_i = \binom{\cos\psi_i}{\sin\psi_i}\begin{pmatrix}\cos\psi_i & \sin\psi_i\end{pmatrix}$, $\psi_i \in (0, \frac{\pi}{2})$. For general subspaces, J_Y and J_Z are diagonal matrices with diagonal entries from $\{0,1\}$ and with $J_Y J_Z = 0$. In our case it is necessary that $J_Y = J_Z = 0$. Namely, an instance of, e.g., $(J_Y)_{ii} = 1$, $(J_Z)_{ii} = 0$ corresponds to a direction in $\mathcal{Y}$ orthogonal to entire $\mathcal{Z}$, which we already have excluded as impossible. The perturbation Ω is in the basis $\mathcal{B}$ represented as

$$\Omega = \left(\begin{array}{c|c|c} 0 & & \\ \hline & \bigoplus_{i=1}^{d}\Omega_i & \\ \hline & & 0 \end{array}\right), \quad \Omega_i = -\begin{pmatrix} 0 & \tan\psi_i \\ \tan\psi_i & 2\tan^2\psi_i \end{pmatrix} \cong \begin{pmatrix} \dfrac{-\sin\psi_i}{1-\sin\psi_i} & 0 \\ 0 & \dfrac{\sin\psi_i}{1+\sin\psi_i} \end{pmatrix},$$

where $\cong$ denotes (unitary) similarity. Note that this special structure of Ω implies that $I-\Omega$ is always positive definite (with eigenvalues 1, $\frac{1}{1-\sin\psi_i}$, $\frac{1}{1+\sin\psi_i}$, $1 \le i \le d$); thus it can be written as $I - \Omega = \Xi\Xi^*$, where Ξ is the definite square root or, e.g., Cholesky factor of $I - \Omega$. The singular values of $\Xi^{-1} - \Xi^*$ are zero and $\sin\psi_i/\sqrt{1+\sin\psi_i}$, $\sin\psi_i/\sqrt{1-\sin\psi_i}$, $1 \le i \le d$.

The proof is completed by an application of [11, Theorem 1], which states that $\|X_\perp^* U_1\|_2 \le (\pi/2)\|\tilde{Q}^*(\Xi^* - \Xi^{-1})Q\|_2/\gamma$. □

[illegible] 3.4. Since $\mathsf{H} = LL^*$ is similar to L^*L and $\tilde{\mathsf{H}} = L\Xi\Xi^*L^*$ is similar to $\Xi^*L^*L\Xi$, we can apply [10, Theorem 3.1] to conclude

$$\max_{1\le i\le n} \frac{|\lambda_i - \tilde{\lambda}_i|}{\sqrt{\lambda_i\tilde{\lambda}_i}} \le \|\Xi^{-1} - \Xi^*\|_2 \le \frac{\sin\psi}{\sqrt{1-\sin\psi}}. \tag{3.2}$$

The matrix $\Psi = \mathsf{W}^{-1/2}K\mathsf{M}^{-1/2}$ from the proof of Theorem 3.1 determines the accuracy of the Ritz values, as shown in the following corollaries.

Corollary 3.5. [illegible] $\mu_1 \le \cdots \le \mu_\ell$ [illegible] M [illegible] 3.1 [illegible] ℓ [illegible] $\lambda_{i_1} \le \cdots \le \lambda_{i_\ell}$ [illegible] H [illegible] $|\lambda_{i_j} - \mu_j| \le \|\Psi\|_2\mu_j$ [illegible] $1 \le j \le \ell$

It is interesting to note that in the framework of Theorem 3.1, the locking (see, e.g., [14]) of converged eigenvectors allows the same accuracy for the remaining wanted eigenvalues. This is because the same bound holds for $\mathcal{X}$ and $\mathcal{X}^\perp$.

Corollary 3.6. [illegible] $\|\Psi\|_2$ [illegible] $\varepsilon > 0$ [illegible] X [illegible] $\mathsf{H}_1 = (I - XX^*)\mathsf{H}(I - XX^*)$ [illegible] H_1 [illegible] $\|\Psi\|_2$

Corollary 3.7. [illegible] $\sqrt{\mu_1} \ge \cdots \ge \sqrt{\mu_\ell}$ [illegible] ℓ [illegible] $\sigma_{i_1} \ge \cdots \ge \sigma_{i_\ell}$ [illegible] A [illegible]

$$\max_{1\le j\le\ell} \frac{|\sigma_{i_j} - \sqrt{\mu_j}|}{\sqrt{\mu_j}} \le \frac{\|\Psi\|_2}{2 - \|\Psi\|_2}. \tag{3.3}$$

4. Quadratic residual bounds. Our new quadratic residual bounds for eigenvalues and singular values, which we present in this section, are complementary to those given in [16], [18], [12]. They provide relative error bounds in terms of relative gaps in the spectrum and certain angles which measure the quality of the search subspaces. We start with the Hermitian positive definite case.

4.1. Hermitian positive definite case. We consider positive definite operators H and $\tilde{\mathsf{H}}$ in block-matrix form,

$$\mathsf{H} = \begin{pmatrix} \mathsf{M} & K^* \\ K & \mathsf{W} \end{pmatrix}, \quad \tilde{\mathsf{H}} = \begin{pmatrix} \mathsf{M} & 0 \\ 0 & \mathsf{W} \end{pmatrix}, \quad \mathsf{M} \in \mathbb{C}^{\ell \times \ell}, \;\; \mathsf{W} \in \mathbb{C}^{w \times w}, \;\; \ell + w = n, \tag{4.1}$$

where the spectral decompositions of M and W are assumed to be known. Let $\mu_1 \leq \cdots \leq \mu_\ell$ and $\omega_1 \leq \cdots \leq \omega_w$ be the eigenvalues of M and W, respectively. The eigenvalues of H and $\tilde{\mathsf{H}}$ are, respectively, $\lambda_1 \leq \cdots \leq \lambda_n$ and $\tilde{\lambda}_1 \leq \cdots \leq \tilde{\lambda}_n$, where the $\tilde{\lambda}_i$ are obtained by merging the μ_j and the ω_k. Since M and W are principal submatrices of H, the [illegible] [6, Theorem 4.3.15] implies

$$\lambda_j \leq \mu_j, \; 1 \leq j \leq \ell; \;\; \omega_k \leq \lambda_{\ell+k}, \; 1 \leq k \leq w. \tag{4.2}$$

In some situations, the spectra of M and W are separated by an interval. In that case, there is an interval (α, β) such that $\lambda_{\max}(\mathsf{M}) \leq \alpha$ and $\beta \leq \lambda_{\min}(\mathsf{W})$. We shall denote this situation by $\mathsf{M} \overset{(\alpha,\beta)}{\leftrightsquigarrow} \mathsf{W}$. In that case, $\tilde{\lambda}_i = \mu_i$, $1 \leq i \leq \ell$, and $\tilde{\lambda}_{\ell+k} = \omega_k$, $1 \leq k \leq w$. This is known to be, for many good reasons, a favorable distribution. For now, we note that as a consequence of (4.2), it imposes similar separation inside the spectrum of H, independent of K. To measure the degree of separation of a scalar ζ from the spectrum $\mathfrak{S}(A)$ of a matrix A, we will use the function

$$\varrho(\zeta, A) = \min_{\lambda \in \mathfrak{S}(A)} \left| \frac{\zeta - \lambda}{\lambda} \right|.$$

The technique we use is, again, the construction of a particularly structured perturbation which nicely fits as input into the state-of-the-art perturbation theory. The motivation for this development stems from [2], [12].

Theorem 4.1. [illegible] (4.1) [illegible] M [illegible] W [illegible] (α, β) [illegible] $\Psi = \mathsf{W}^{-1/2} K \mathsf{M}^{-1/2}$ [illegible] 3.1 [illegible]

- [illegible] $i \in \{1, \ldots, \ell\}$ λ_i [illegible] W [illegible]
$$0 \leq \frac{\mu_i - \lambda_i}{\mu_i} \leq \min\left\{ \|\Psi\|_2, \frac{\|\Psi\|_2^2}{\varrho(\lambda_i, \mathsf{W})} \right\}; \tag{4.3}$$
- [illegible] $k \in \{1, \ldots, w\}$ $\lambda_{\ell+k}$ [illegible] M [illegible]
$$0 \leq \frac{\lambda_{\ell+k} - \omega_k}{\omega_k} \leq \min\left\{ \|\Psi\|_2, \frac{\|\Psi\|_2^2}{\varrho(\lambda_{\ell+k}, \mathsf{M})} \right\}. \tag{4.4}$$

[illegible] M [illegible] W [illegible] i

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{\tilde{\lambda}_i} \leq \min\left\{ \|\Psi\|_2, \frac{\|\Psi\|_2^2}{\max\{\varrho(\lambda_i, \mathsf{M}), \varrho(\lambda_i, \mathsf{W})\}} \right\}. \tag{4.5}$$

[illegible] The proof is based on the Schur factorizations

$$\mathsf{H} - \zeta I = \begin{pmatrix} I & 0 \\ K(\mathsf{M} - \zeta I)^{-1} & I \end{pmatrix} \begin{pmatrix} \mathsf{M} - \zeta I & 0 \\ 0 & \hat{\mathsf{M}}(\zeta) \end{pmatrix} \begin{pmatrix} I & (\mathsf{M} - \zeta I)^{-1} K^* \\ 0 & I \end{pmatrix}, \tag{4.6}$$

$$\mathsf{H} - \zeta I = \begin{pmatrix} I & K^*(\mathsf{W} - \zeta I)^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \hat{\mathsf{W}}(\zeta) & 0 \\ 0 & \mathsf{W} - \zeta I \end{pmatrix} \begin{pmatrix} I & 0 \\ (\mathsf{W} - \zeta I)^{-1} K & I \end{pmatrix} \tag{4.7}$$

with the Schur complements $\hat{\mathsf{M}}(\zeta) = \mathsf{W} - \zeta I - K(\mathsf{M} - \zeta I)^{-1}K^*$ (for ζ not in the spectrum of M), $\hat{\mathsf{W}}(\zeta) = \mathsf{M} - \zeta I - K^*(\mathsf{W} - \zeta I)^{-1}K$ (for ζ outside the spectrum of W).

Let λ_i be an eigenvalue of H such that it is not in the spectrum of W. Then the congruence (4.7) is well defined at $\zeta = \lambda_i$, and

$$\mathsf{H} - \lambda_i I \text{ is congruent to } \begin{pmatrix} \mathsf{M} - \lambda_i I & 0 \\ 0 & \mathsf{W} - \lambda_i I \end{pmatrix} - \begin{pmatrix} K^*(\mathsf{W} - \lambda_i I)^{-1}K & 0 \\ 0 & 0 \end{pmatrix}.$$

Since the ith eigenvalue of $\mathsf{H} - \lambda_i I$ is zero, Sylvester's inertia theorem implies that both matrices in the last relation have zero as the ith eigenvalue. This implies that λ_i is the ith eigenvalue $\tilde{\tilde{\lambda}}_i$ of the matrix

$$\begin{pmatrix} \mathsf{M} & 0 \\ 0 & \mathsf{W} \end{pmatrix} - \begin{pmatrix} K^*(\mathsf{W} - \lambda_i I)^{-1}K & 0 \\ 0 & 0 \end{pmatrix} \equiv \tilde{\mathsf{H}} - \delta\tilde{\mathsf{H}}_1.$$

Thus, comparing $\tilde{\lambda}_i$ and λ_i amounts to comparing $\tilde{\lambda}_i$ and its perturbation $\tilde{\tilde{\lambda}}_i$. Here is the key point where our approach differs from [12]—we consider the scaled residual with a geometric interpretation and relative eigenvalue perturbations. Recall $\Psi = \mathsf{W}^{-1/2}K\mathsf{M}^{-1/2}$ from the proof of Theorem 3.1. We know that $\|\Psi\|_2 < 1$ and that $|\tilde{\lambda}_i - \lambda_i| \le \|\Psi\|_2\tilde{\lambda}_i$. Then we can write

$$\tilde{\mathsf{H}} - \delta\tilde{\mathsf{H}}_1 = \begin{pmatrix} \mathsf{M}^{1/2} & 0 \\ 0 & \mathsf{W}^{1/2} \end{pmatrix} \left\{ I - \begin{pmatrix} \Psi^*(I - \lambda_i\mathsf{W}^{-1})^{-1}\Psi & 0 \\ 0 & 0 \end{pmatrix} \right\} \begin{pmatrix} \mathsf{M}^{1/2} & 0 \\ 0 & \mathsf{W}^{1/2} \end{pmatrix}, \tag{4.8}$$

which in the case $\|\Psi^*(I - \lambda_i\mathsf{W}^{-1})^{-1}\Psi\|_2 < 1$ implies (as in the proof of Theorem 3.1)

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{\tilde{\lambda}_i} \le \|\Psi^*(I - \lambda_i\mathsf{W}^{-1})^{-1}\Psi\|_2 \le \frac{\|\Psi\|_2^2}{\varrho(\lambda_i, \mathsf{W})}, \quad \varrho(\lambda_i, \mathsf{W}) = \min_k \frac{|\lambda_i - \omega_k|}{\omega_k}. \tag{4.9}$$

If $1 \le i \le \ell$, and if the spectra of M and W are separated, then $\tilde{\lambda}_i = \mu_i$. Further, by (4.2), $\lambda_i \le \mu_i < \omega_1$, which means that λ_i is strictly left from the spectrum of W and that the perturbation $\delta\tilde{\mathsf{H}}_1$ is positive semidefinite. Thus, $\tilde{\lambda}_i \ge \lambda_i$.

If λ_i does not belong to the spectrum of M, then, by a similar argument as above, it must be the ith eigenvalue of the perturbed matrix

$$\tilde{\mathsf{H}} - \delta\tilde{\mathsf{H}}_2 = \begin{pmatrix} \mathsf{M}^{1/2} & 0 \\ 0 & \mathsf{W}^{1/2} \end{pmatrix} \left\{ I - \begin{pmatrix} 0 & 0 \\ 0 & \Psi(I - \lambda_i\mathsf{M}^{-1})^{-1}\Psi^* \end{pmatrix} \right\} \begin{pmatrix} \mathsf{M}^{1/2} & 0 \\ 0 & \mathsf{W}^{1/2} \end{pmatrix}.$$

In that case $\|\Psi(I - \lambda_i\mathsf{M}^{-1})^{-1}\Psi^*\|_2 < 1$ implies

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{\tilde{\lambda}_i} \le \|\Psi(I - \lambda_i\mathsf{M}^{-1})^{-1}\Psi^*\|_2 \le \frac{\|\Psi\|_2^2}{\varrho(\lambda_i, \mathsf{M})}, \quad \varrho(\lambda_i, \mathsf{M}) = \min_j \frac{|\lambda_i - \mu_j|}{\mu_j}. \tag{4.10}$$

If $i = \ell + k$, and $\mathsf{M} \overset{(\alpha,\beta)}{\leftrightsquigarrow} \mathsf{W}$, then $\lambda_i \ge \tilde{\lambda}_i = \omega_k > \mu_m$. □

Combined with LR iterations and spectral monotonicity, Theorem 4.1 gives residual bounds for the harmonic Ritz values.

Theorem 4.2. [illegible] $\chi_1 \le \cdots \le \chi_\ell$ [illegible] H^{-1} [illegible] $\mathcal{X}$ [illegible] $(X^*\mathsf{H}^{-1}X)^{-1}$ [illegible] χ_j [illegible] λ_j [illegible] 4.1 [illegible]

$$0 \le \frac{\chi_j - \lambda_j}{\chi_j} \le \frac{\mu_j - \lambda_j}{\mu_j}, \quad 1 \le j \le \ell. \tag{4.11}$$

[illegible] $\varpi_1 \leq \cdots \leq \varpi_{n-\ell}$ [illegible] $X_\perp^* \mathsf{H}^2 X_\perp - \lambda X_\perp^* \mathsf{H} X_\perp$ [illegible] H^{-1} [illegible] $\mathsf{H}\mathcal{X}^\perp$ [illegible]

$$0 \leq \frac{\lambda_{\ell+j} - \varpi_j}{\varpi_j} \leq \frac{\lambda_{\ell+j} - \omega_j}{\omega_j}, \quad 1 \leq j \leq n - \ell. \tag{4.12}$$

[illegible] Recall, $\mathsf{H} = LL^*$, $\mathsf{M} = X^*\mathsf{H}X$, $K = X_\perp^*\mathsf{H}X$, $\mathsf{W} = X_\perp^*\mathsf{H}X_\perp$. In the factorization

$$\begin{pmatrix} \mathsf{M} & K^* \\ K & \mathsf{W} \end{pmatrix} = \begin{pmatrix} \sqrt{\mathsf{M} - K^*\mathsf{W}^{-1}K} & K^*\mathsf{W}^{-1/2} \\ 0 & \mathsf{W}^{1/2} \end{pmatrix} \begin{pmatrix} \sqrt{\mathsf{M} - K^*\mathsf{W}^{-1}K} & 0 \\ \mathsf{W}^{-1/2}K & \mathsf{W}^{1/2} \end{pmatrix}$$

reverse the order of the factors to obtain the similar matrix

$$\mathsf{H}_1 = \begin{pmatrix} \mathsf{M} - K^*\mathsf{W}^{-1}K & \sqrt{\mathsf{M} - K^*\mathsf{W}^{-1}K}K^*\mathsf{W}^{-1/2} \\ \mathsf{W}^{-1/2}K\sqrt{\mathsf{M} - K^*\mathsf{W}^{-1}K} & \mathsf{W} + \mathsf{W}^{-1/2}KK^*\mathsf{W}^{-1/2} \end{pmatrix} \equiv \begin{pmatrix} \mathsf{M}_\downarrow & K_1^* \\ K_1 & \mathsf{W}_\uparrow \end{pmatrix}.$$

Let $L^*X_\perp = QR$ be the thin QR factorization. Then $\mathsf{M}_\downarrow \equiv \mathsf{M} - K^*\mathsf{W}^{-1}K$ satisfies $\mathsf{M}_\downarrow = X^*L(I - QQ^*)L^*X$, where $I - QQ^*$ is the orthogonal projection onto $(L^*\mathcal{X}^\perp)^\perp = L^{-1}\mathcal{X}$. Using the thin QR factorization $L^{-1}X = Q_1R_1$, we can write $\mathsf{M}_\downarrow = X^*LQ_1Q_1^*L^*X^* = R_1^{-1}R_1^{-*} = (X^*\mathsf{H}^{-1}X)^{-1}$. Thus, in the Loewner partial order $\mathsf{M}_\downarrow \preccurlyeq \mathsf{M}$ ($\mathsf{M} - \mathsf{M}_\downarrow$ is positive semidefinite) and by monotonicity, $\chi_j \leq \mu_j$ for all j. Since H_1 and H are similar, an application of the inclusion principle yields $\lambda_j \leq \chi_j$ for all $1 \leq j \leq \ell$. Thus, we have $0 < \lambda_j \leq \chi_j \leq \mu_j$, which implies (4.11). Further, it is easily shown that $\mathsf{W} \preccurlyeq \mathsf{W}_\uparrow = \mathsf{W}^{-1/2}X_\perp^*\mathsf{H}^2X_\perp\mathsf{W}^{-1/2}$, which means that the eigenvalues of the pencil $X_\perp^*\mathsf{H}^2X_\perp - \lambda X_\perp^*\mathsf{H}X_\perp$ dominate the eigenvalues of W, while forced below the corresponding λ_j by the inclusion principle. Hence, $\omega_j \leq \varpi_j \leq \lambda_{\ell+j}$, yielding (4.12). Finally, note that the Ritz values of H^{-1} with respect to $\mathsf{H}\mathcal{X}^\perp$ are the eigenvalues of the pencil $X_\perp^*\mathsf{H}X_\perp - \lambda X_\perp^*\mathsf{H}^2X_\perp$.

We should also note that the error bound as well is better (in the sense of Theorems 3.1 and 4.1) for the diagonal blocks of H_1. Recall $\Psi = \mathsf{W}^{-1/2}K\mathsf{M}^{-1/2}$ and let $\Psi_1 = \mathsf{W}_\uparrow^{-1/2}K_1M_\downarrow^{-1/2}$. Then $\Psi_1 = \mathsf{W}_\uparrow^{-1/2}\Psi\mathsf{M}^{1/2}$ and $\|\Psi_1\|_2 \leq \|\Psi\|_2(\frac{\lambda_{\max}(\mathsf{M})}{\lambda_{\min}(\mathsf{W}_\uparrow)})^{1/2} < \|\Psi\|_2$. Moreover, the relative separation increases because $\lambda_{\max}(\mathsf{M}_\downarrow) \leq \lambda_{\max}(\mathsf{M})$ and $\lambda_{\min}(\mathsf{W}) \leq \lambda_{\min}(\mathsf{W}_\uparrow)$. □

[illegible] 4.3. Note that one can also reverse the order of the factors in the factorization

$$\mathsf{H} = \begin{pmatrix} \sqrt{\mathsf{M}} & 0 \\ K\mathsf{M}^{-1/2} & \sqrt{\mathsf{W} - K\mathsf{M}^{-1}K^*} \end{pmatrix} \begin{pmatrix} \sqrt{\mathsf{M}} & \mathsf{M}^{-1/2}K^* \\ 0 & \sqrt{\mathsf{W} - K\mathsf{M}^{-1}K^*} \end{pmatrix}$$

and obtain diagonal blocks $\mathsf{M}_\uparrow = \mathsf{M} + \mathsf{M}^{-1/2}K^*K\mathsf{M}^{-1/2}$ and $\mathsf{W}_\downarrow = \mathsf{W} - K\mathsf{M}^{-1}K^*$.

It is interesting that separation of the spectra of M and W implies a certain relation between $\mathrm{Im}(X)$ and a particular spectral subspace of H.

Proposition 4.4. [illegible] H [illegible] X [illegible]

$$\hat{\mathsf{H}} = \begin{pmatrix} X & X_\perp \end{pmatrix}^* \mathsf{H} \begin{pmatrix} X & X_\perp \end{pmatrix} = \begin{pmatrix} \mathsf{M} & K^* \\ K & \mathsf{W} \end{pmatrix}, \text{[illegible]} \quad \lambda_{\max}(\mathsf{M}) < \lambda_{\min}(\mathsf{W}).$$

[illegible] $\mathcal{U}_1 = \mathrm{Im}(U_1)$ [illegible] ℓ [illegible] H [illegible] $\mathrm{Im}(X)$ [illegible] $\mathcal{U}_1$ [illegible] $\pi/4$

[illegible] Pick $\xi \in (\lambda_{\max}(\mathsf{M}), \lambda_{\min}(\mathsf{W}))$ and note that $\xi I - \mathsf{M}$ and $\mathsf{W} - \xi I$ are both positive definite. Exactly ℓ eigenvalues of H are smaller than ξ. In the eigenvector matrix $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$ of H let the ℓ columns of U_1 belong to those eigenvalues below ξ. Since the eigenvectors are shift invariant, we can study the eigenvectors of H by looking at $\hat{\mathsf{H}} - \xi I$, which is in fact a [illegible] matrix. The eigenvectors of quasi-definite matrices have a special dominance property. If a unitary eigenvector matrix Q of $\hat{\mathsf{H}}$ is partitioned as

$$Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}, \quad Q_i = \begin{pmatrix} Q_{1i} \\ Q_{2i} \end{pmatrix}, \quad i = 1, 2,$$

then Q_1 and Q_2 are determined up to postmultiplication by certain block-diagonal unitary matrices. Further, $Q_{11}^* Q_{11} - Q_{21}^* Q_{21}$ is positive definite, which, because Q is unitary, implies that the minimal singular value of Q_{11} is bounded from below by $1/\sqrt{2}$, $\sigma_{\min}(Q_{11}) > 1/\sqrt{2}$. (See [4] for more on eigenvector structure of quasi-definite matrices.) If the eigenvector matrix U of H is partitioned in the same way as Q, then $X^* U_1 = Q_{11} C$ with some unitary C. Thus, the cosines of the canonical angles between the ranges of U_1 and X are bigger than $1/\sqrt{2}$. □

4.2. Quadratic SVD residual bound. One obvious usage of Theorem 4.1 is to apply it separately to $\mathsf{A}^*\mathsf{A}$, $\mathcal{Y}$ and $\mathsf{A}\mathsf{A}^*$, $\mathcal{X}$ and thus obtain quadratic residual bound for the singular values of A. Another approach is to use A, $\mathcal{X}$, and $\mathcal{Y}$ fused into $\mathsf{C} = X^* \mathsf{A} Y$. As in Theorem 2.10, we can reduce the problem to the square nonsingular case, i.e., in what follows we replace A with the square block matrix (cf. (1.3))

$$\mathsf{A}' = \begin{pmatrix} \mathsf{C} & F \\ G & \mathsf{C}_\perp \end{pmatrix}, \quad \mathsf{C} = X^* \mathsf{A} Y, \quad F = X^* \mathsf{A} Y_\perp, \quad G = X_\perp^* \mathsf{A} Y, \quad \mathsf{C}_\perp = X_\perp^* \mathsf{A} Y_\perp, \tag{4.13}$$

where A' and both diagonal blocks $\mathsf{C} \in \mathbb{C}^{\ell \times \ell}$ and $\mathsf{C}_\perp \in \mathbb{C}^{(n-\ell)\times(n-\ell)}$ are nonsingular.

By an application of the results of section 4.1 to the cross products $(\mathsf{A}')^*\mathsf{A}'$ and $\mathsf{A}'(\mathsf{A}')^*$, one easily obtains estimates for the singular values of $\begin{pmatrix} \mathsf{C} & F \end{pmatrix}$, $\begin{pmatrix} \mathsf{C}^* & G^* \end{pmatrix}$, $\begin{pmatrix} G & \mathsf{C}_\perp \end{pmatrix}$, $\begin{pmatrix} F^* & \mathsf{C}_\perp^* \end{pmatrix}$. The following theorem shows that we can also work with C and $\mathsf{C}_\perp$.

Theorem 4.5. [illegible] A' [illegible] (4.13) [illegible]

$$(a) \quad \sigma_{\max}(\mathsf{C}) \le \alpha < \beta \le \sigma_{\min}(\mathsf{C}_\perp), \quad (b) \quad \|G\mathsf{C}^{-1}\|_2^2 < 2\frac{\beta - \alpha}{\alpha}. \tag{4.14}$$

- [illegible] $i \in \{1, \ldots, \ell\}$ [illegible] $\rho_i \equiv 1 - \sigma_i^2/\sigma_{\min}^2(\mathsf{C}_\perp) > 0$ [illegible] σ_i [illegible] $\mathsf{C}_\perp$ [illegible]

$$\xi \equiv \frac{(\tan \sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y}) + \tan \sphericalangle(\mathcal{X}, \mathsf{A}^{-*}\mathcal{Y}))^2}{\rho_i} < 1,$$

[illegible] $\gamma_i = \sigma_i(\mathsf{C})$ [illegible]

$$\frac{|\gamma_i - \sigma_i|}{\gamma_i} \le \frac{\xi}{2 - \xi}. \tag{4.15}$$

- [illegible] $i = \ell + k$ $k = 1, \ldots, n - \ell$ [illegible] $\varrho_i \equiv \sigma_i^2/\sigma_{\max}^2(\mathsf{C}) - 1 - \tan^2 \sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y}) > 0$ [illegible]

$$\zeta \equiv \tan^2 \sphericalangle(\mathcal{X}, \mathsf{A}^{-*}\mathcal{Y}) + \frac{(\tan \sphericalangle(\mathcal{X}, \mathsf{A}\mathcal{Y}) + \tan \sphericalangle(\mathcal{X}, \mathsf{A}^{-*}\mathcal{Y}))^2}{\varrho_i} < 1,$$

[illegible]

$$\frac{|\sigma_k(\mathsf{C}_\perp)-\sigma_i|}{\sigma_k(\mathsf{C}_\perp)} \leq \frac{\zeta}{2-\zeta}. \tag{4.16}$$

[illegible] Let $i \in \{1,\ldots,\ell\}$. Then, by the Poincaré separation theorem and (4.14),

$$\sigma_i \equiv \sigma_i(\mathsf{A}') \leq \sigma_i\left(\begin{pmatrix}\mathsf{C}\\ G\end{pmatrix}\right) \leq \sigma_i(\mathsf{C})\left(1+\frac{\|G\mathsf{C}^{-1}\|_2^2}{2}\right) < \beta \leq \sigma_{\min}(\mathsf{C}_\perp) \leq \sigma_{\min}\left(\begin{pmatrix}F\\ \mathsf{C}_\perp\end{pmatrix}\right).$$

Thus, $\rho_i > 0$, σ_i is not a singular value of $\left(\begin{smallmatrix}F\\ \mathsf{C}_\perp\end{smallmatrix}\right)$, and, using the congruence (4.7) as in Theorem 4.1, we conclude that σ_i^2 is the ith eigenvalue of

$$\breve{H} = \begin{pmatrix}\mathsf{C}^*\mathsf{C} & 0\\ 0 & \mathsf{C}_\perp^*\mathsf{C}_\perp + F^*F\end{pmatrix} + \begin{pmatrix}E & 0\\ 0 & 0\end{pmatrix} \equiv \tilde{H} + \begin{pmatrix}E & 0\\ 0 & 0\end{pmatrix}, \tag{4.17}$$

where $E = G^*G - (\mathsf{C}^*F + G^*\mathsf{C}_\perp)(\mathsf{C}_\perp^*\mathsf{C}_\perp + F^*F - \sigma_i^2 I)^{-1}(F^*\mathsf{C} + \mathsf{C}_\perp^* G)$. Now write $\breve{H}$ as

$$\breve{H} = \begin{pmatrix}\mathsf{C}^* & 0\\ 0 & \sqrt{\mathsf{C}_\perp^*\mathsf{C}_\perp + F^*F}\end{pmatrix}\begin{pmatrix}I+\Xi & 0\\ 0 & I\end{pmatrix}\begin{pmatrix}\mathsf{C} & 0\\ 0 & \sqrt{\mathsf{C}_\perp^*\mathsf{C}_\perp + F^*F}\end{pmatrix},\quad \Xi = \mathsf{C}^{-*}E\mathsf{C}^{-1},$$

to conclude that the ith eigenvalue $\tilde{\lambda}_i$ of $\tilde{H}$ satisfies $|\sigma_i^2 - \tilde{\lambda}_i| \leq \tilde{\lambda}_i\|\Xi\|_2$, provided that $\|\Xi\|_2 < 1$. The separation condition (4.14.a) implies that $\tilde{\lambda}_i = \sigma_i^2(\mathsf{C})$, and thus

$$\frac{|\sigma_i - \sigma_i(\mathsf{C})|}{\sigma_i(\mathsf{C})} \leq \frac{\|\Xi\|_2}{2-\|\Xi\|_2}.$$

To compute $\|\Xi\|_2$, set $\Phi = F\mathsf{C}_\perp^{-1}$, $\Gamma = G\mathsf{C}^{-1}$ and write Ξ as

$$\Xi = \Gamma^*\Gamma - (\Phi + \Gamma^*)\left(I + \Phi^*\Phi - \sigma_i^2\mathsf{C}_\perp^{-*}\mathsf{C}_\perp^{-1}\right)^{-1}(\Phi^* + \Gamma).$$

Now, $i \leq \ell$ and (4.14) imply that Ξ is the difference of two semidefinite matrices, and

$$\|\Xi\|_2 \leq \max\left\{\|\Gamma\|_2^2, \frac{\|\Phi+\Gamma^*\|_2^2}{\lambda_{\min}(I+\Phi^*\Phi - \sigma_i^2\mathsf{C}_\perp^{-*}\mathsf{C}_\perp^{-1})}\right\},$$

where

$$\lambda_{\min}(I+\Phi^*\Phi - \sigma_i^2\mathsf{C}_\perp^{-*}\mathsf{C}_\perp^{-1}) \geq \lambda_{\min}(I+\Phi^*\Phi) - \frac{\sigma_i^2}{\sigma_{\min}^2(\mathsf{C}_\perp)} \geq 1 - \frac{\sigma_i^2}{\sigma_{\min}^2(\mathsf{C}_\perp)} > 0.$$

Now, let $i = \ell + k$ for some $k \in \{1,\ldots,n-\ell\}$. As before, it can be shown that σ_i^2 is the ith eigenvalue of

$$\breve{\mathsf{H}} = \begin{pmatrix}\mathsf{C}^*\mathsf{C} + G^*G & 0\\ 0 & \mathsf{C}_\perp^*\mathsf{C}_\perp\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & E\end{pmatrix} \equiv \tilde{\mathsf{H}} + \begin{pmatrix}0 & 0\\ 0 & E\end{pmatrix},$$

where $E = F^*F - (F^*\mathsf{C} + \mathsf{C}_\perp^* G)(\mathsf{C}^*\mathsf{C} + G^*G - \sigma_i^2 I)^{-1}(\mathsf{C}^*F + G^*\mathsf{C}_\perp)$. Thus, the ith eigenvalue $\tilde{\lambda}_i = \sigma_k^2(\mathsf{C}_\perp)$ of $\tilde{\mathsf{H}}$ satisfies $|\sigma_i^2 - \tilde{\lambda}_i| \leq \tilde{\lambda}_i\|\Xi\|_2$, provided that the norm of $\Xi \equiv \Phi^*\Phi - (\Phi^* + \Gamma)(I + \Gamma^*\Gamma - \sigma_i^2\mathsf{C}^{-*}\mathsf{C}^{-1})^{-1}(\Phi + \Gamma^*)$ is less than one. Note that

$$\|\Xi\|_2 \leq \|\Phi\|_2^2 + \frac{\|\Phi+\Gamma^*\|_2^2}{\sigma_{\min}(I + \Gamma^*\Gamma - \sigma_i^2\mathsf{C}^{-*}\mathsf{C}^{-1})} \leq \|\Phi\|_2^2 + \frac{\|\Phi+\Gamma^*\|_2^2}{|1 - \sigma_i^2/\sigma_{\max}^2(\mathsf{C}) + \|\Gamma\|_2^2|}. \qquad \square$$

[illegible] 4.6. Note that the results of Theorem 4.5 are simplified if A' is block-triangular (e.g., $\mathcal{X} = \mathsf{A}\mathcal{Y}$ or $\mathcal{Y} = \mathsf{A}^*\mathcal{X}$, implying that one of the two tangents is zero). Another set of estimates can be obtained by using $\mathsf{A}'(\mathsf{A}')^*$.

Acknowledgments. The authors are indebted to the anonymous referees for valuable comments and corrections and in particular to Ilse Ipsen for constructive criticism.

REFERENCES

[1] Z. Drmač, *On relative residual bounds for the eigenvalues of a Hermitian matrix*, Linear Algebra Appl., 244 (1996), pp. 155–163.

[2] Z. Drmač and V. Hari, *Relative residual bounds for the eigenvalues of a Hermitian semidefinite matrix*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 21–29.

[3] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[4] A. George, K. Ikramov, and A. B. Kucherov, *Some properties of symmetric quasi-definite matrices*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1318–1323.

[5] M. E. Hochstenbach, *A Jacobi–Davidson type SVD method*, SIAM J. Sci. Comput., 23 (2001), pp. 606–628.

[6] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, London, 1990.

[7] W. Kahan, *Inclusion Theorems for Clusters of Eigenvalues of Hermitian Matrices*, Tech. report CS42, Department of Computer Science, University of Toronto, Toronto, Canada, 1967.

[8] A. V. Knyazev and M. E. Argentati, *Principal angles in an A-based scalar product: Algorithms and perturbation estimates*, SIAM J. Sci. Comput., 23 (2002), pp. 2009–2040.

[9] C.-K. Li and R. Mathias, *The Lidskii–Mirsky–Wielandt theorem—additive and multiplicative versions*, Numer. Math., 81 (1999), pp. 377–413.

[10] R.-C. Li, *Relative perturbation theory:* I. *Eigenvalue and singular value variations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 956–982.

[11] R.-C. Li, *A bound on the solution to a structured Sylvester equation with an application to relative perturbation theory*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 440–445.

[12] R. Mathias, *Quadratic residual bounds for the Hermitian eigenvalue problem*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 541–550.

[13] B. N. Parlett, *The Symmetric Eigenvalue Problem*, in Classics in Appl. Math. 20, SIAM, Philadelphia, 1998.

[14] A. Stathopoulos, *Locking Issues for Finding a Large Number of Eigenvectors of Hermitian Matrices*, Tech. report, Department of Computer Science, College of William and Mary, Williamsburg, VA, 2006.

[15] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[16] G. W. Stewart, *Two simple residual bounds for the eigenvalues of a Hermitian matrix*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 205–208.

[17] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[18] J.-G. Sun, *Eigenvalues of Rayleigh quotient matrices*, Numer. Math., 59 (1991), pp. 603–614.

[19] P. A. Wedin, *On angles between subspaces of a finite dimensional inner product space*, in Matrix Pencils, Springer-Verlag, New York, 1983, pp. 263–285.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 1, pp. 68–74

THE RESULT OF TWO STEPS OF THE *LR* ALGORITHM IS DIAGONALLY SIMILAR TO THE RESULT OF ONE STEP OF THE *HR* ALGORITHM[*]

JASON SLEMONS[†]

Abstract. Real nonsymmetric tridiagonal matrices arise in various applications. When one is asked to find the eigenvalues of such a matrix, the QR algorithm is used, but this destroys tridiagonal form by converting the matrix to Hessenberg form, resulting in increased storage requirements and numerical operations. The HR algorithm, based on the HR factorization of the matrix into a (Δ, Δ_1)-orthogonal part H, where $H^T \Delta H = \Delta_1$, and an upper triangular part R, solves this problem. In a result proved by Hongguo Xu, two steps of the LR algorithm are equivalent to one step of the QR algorithm for symmetric matrices. The first object of this paper is to use the HR algorithm to extend Hongguo Xu's result to the nonsymmetric case. Since an HR factorization does not always exist, so we also consider an extension to it called XHR factorization. We then prove a similar result about it.

Key words. LR, QR, triangular factorization, HR

AMS subject classifications. 15A18, 15A23, 15A57, 65F15

DOI. 10.1137/070686457

1. Introduction.

An attractive homework exercise in a course on matrix computations is to show that for an SPD (symmetric positive definite) matrix A, the result of one step of the QR algorithm is equal to the result of two steps of the Cholesky LR algorithm. In [7] Xu proves that if a symmetric matrix A admits a triangular factorization, then the result of one step of the QR algorithm applied to A is equivalent to the result of two steps of the LR algorithm applied to A. The goal of this paper is to extend Xu's result to the nonsymmetric case. This goal seems doomed because the QR algorithm does not preserve bandwidth while the LR algorithm does. The way out of this difficulty is to find another algorithm that is less restrictive than QR. This useful step was taken in 1981 by Bunse-Gerstner in [3] with the introduction of the HR algorithm.

2. *HR*.

The HR factorization of a matrix requires the use of matrices that when squared yield the identity. In this paper we consider two such classes of matrices: signature matrices, the set of which is denoted by Δ, and signed symmetric permutation matrices.

DEFINITION 2.1. Ω SSP π

$$\Omega_{i,j} = \begin{cases} \pm 1 & j = \pi(i), \\ 0 & . \end{cases} \tag{2.1}$$

[*] Received by the editors March 28, 2007; accepted for publication (in revised form) by Z. Drmač March 3, 2008; published electronically February 27, 2009. This research was supported by NSF grant DMS 0354131.

http://www.siam.org/journals/simax/31-1/68645.html

[†] Department of Applied Mathematics, University of Washington, Seattle, WA 98195 (slemons@amath.washington.edu).

For example,

$$\Omega = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \tag{2.2}$$

is an SSP matrix. Clearly signature matrices are also SSP matrices.

In this paper Δ and Ω will denote a signature matrix and an SSP matrix, respectively.

DEFINITION 2.2. [illegible] $H \in \mathbb{R}^{n\times n}$ [illegible] Ω_0, Ω [illegible] H [illegible] (Ω, Ω_0)-orthogonal [illegible] $H^T\Omega H = \Omega_0$ [illegible] $O(\Omega, \Omega_0)$ [illegible] (Ω, Ω_0) [illegible] Δ [illegible] Δ_0 [illegible] (Δ, Δ_0)-orthogonal [illegible] $O(\Delta, \Delta_0)$ [illegible] H [illegible] $H^T\Delta H = \Delta_0$

Observe that if H is an element of $O(\Omega, \Omega_0)$, then $H^{-1} = \Omega_0 H^T \Omega$. This will be useful in Theorem 2.5.

DEFINITION 2.3. [illegible] $A \in \mathbb{R}^{n\times n}$ [illegible] A [illegible] Ω-symmetric [illegible] $A^T\Omega = \Omega A$ [illegible] Δ-symmetric [illegible] $A^T\Delta = \Delta A$

DEFINITION 2.4. [illegible] $A \in \mathbb{R}^{n\times n}$ [illegible] Δ [illegible] HR factorization [illegible] A [illegible] $A = HR$ [illegible] $H \in O(\Delta, \Delta_0)$ [illegible] $R \in \mathbb{R}^{n\times n}$ [illegible]

The HR factorization exists if and only if no principle minor of $A^T\Delta A$ vanishes and the product of the first i diagonal entries of Δ_0 coincides with the sign of the ith principle minor of $A^T\Delta A$ for each $i \in \{1 \ldots n\}$. The HR factorization is also unique. Analyses of its use in the HR algorithm can be found in [1], [2], and [3].

2.1. *HR* and *LR*. Given the factorization of a matrix $A = LU$, where L is a unit lower triangular matrix and U is an upper triangular matrix, we define the result of one step of the LR algorithm applied to A to be the matrix given by UL. This matrix is similar to the original by $UL = L^{-1}AL$. The result of one step of the HR algorithm, on a matrix A given its HR factorization, $A = HR$, is the matrix RH, which is similar to A by $RH = H^{-1}AH$. Throughout this section we will denote the matrix UL as $\dot{A}$. The result of two steps of the LR algorithm applied to A means finding $\dot{A}$, factoring it as $\dot{A} = \hat{L}\hat{U}$, and then forming $\ddot{A} = \hat{U}\hat{L}$. $\ddot{A}$ is similar to $\dot{A}$, which is similar to A, and in turn $\ddot{A}$ is similar to A. As we continue the algorithm, at each step the result has the same eigenvalues as the original matrix.

The real use of the LR or HR algorithms is in their application to tridiagonal matrices. If T_G is an unreduced tridiagonal matrix, we can assume without loss of generality that it is balanced.[1] Next we consider $T_G = \Delta T$, its factorization into a signature matrix, Δ, and a symmetric tridiagonal matrix T. This can be done by simply factoring out ±1 from each row. T_G is then Δ-symmetric, and assuming it has the required factorizations, we can apply the following theorem to it.

THEOREM 2.5. [illegible] T_G [illegible] Δ [illegible] T_G [illegible] T_G [illegible] T_G [illegible] Δ [illegible] HR [illegible] T_G [illegible] $\ddot{T}_G$

[1] This means T_G's ith absolute row sum equals its ith absolute column sum. Every tridiagonal matrix is similar by a diagonal matrix to a balanced tridiagonal matrix, meaning $|T_{G_{i,i+1}}| = |T_{G_{i+1,i}}|$.

[illegible]. If T_G is Δ-symmetric, then $T_G^T\Delta = \Delta T_G$. $T = \Delta T_G$ is a symmetric matrix. Since $T_G = \Delta T$, the existence of a triangular factorization of T_G implies the existence of a triangular factorization for T. Let $T = L_1 D_1 L_1^T$, where L_1 is a unit lower triangular matrix and D_1 is a diagonal matrix. We factorize T_G as

$$\begin{aligned} T_G &= \Delta T \\ &= (\Delta L_1 \Delta)(\Delta D_1 L_1^T), \end{aligned}$$

which is the LR factorization of T_G. The result of one step of the [illegible] algorithm applied to T_G is

$$\dot{T}_G = (\Delta D_1 L_1^T)(\Delta L_1 \Delta). \tag{2.3}$$

Since $\dot{T}_G$ permits triangular factorization then so does $L_1^T \Delta L_1$. We decompose $L_1^T\Delta L_1 = L_2 D_2 L_2^T$. Given $L_1^T \Delta L_1 = L_2 D_2 L_2^T$ we can now explicitly separate the triangular factors of $\dot{T}_G$ to get

$$\begin{aligned} \dot{T}_G &= \Delta D_1 L_1^T \Delta L_1 \Delta \\ &= \Delta D_1 L_2 D_2 L_2^T \Delta \\ &= (\Delta D_1 L_2 D_1^{-1} \Delta)(\Delta D_1 D_2 L_2^T \Delta). \end{aligned} \tag{2.4}$$

The matrix $\Delta D_1 L_2 D_1^{-1}\Delta$ is unit lower triangular and $\Delta D_1 D_2 L_2^T \Delta$ is an upper triangular matrix so (2.4) is the LR factorization of $\dot{T}_G$. Applying another step of the [illegible] algorithm to $\dot{T}_G$ we get

$$\ddot{T}_G = (\Delta D_1 D_2 L_2^T \Delta)(\Delta D_1 L_2 D_1^{-1}\Delta). \tag{2.5}$$

The matrix $\ddot{T}_G$ denotes the result of two steps of the [illegible] algorithm applied to T_G. It remains to show that (2.5) is diagonally similar to one step of the HR algorithm applied to T_G. For this purpose, we rewrite $D_2 = \bar{D}_2 \Delta_2 \bar{D}_2$, where $\bar{D}_2 > 0$ is a diagonal matrix, and Δ_2 is a signature matrix. This allows us to define the matrix Q as $L_1 L_2^{-T} \bar{D}_2^{-1}$. Notice that

$$\begin{aligned} \bar{D}_2^{-1} L_2^{-1} L_1^T \Delta L_1 L_2^{-T} \bar{D}_2^{-1} &= \Delta_2, \\ Q^T \Delta Q &= \Delta_2, \end{aligned} \tag{2.6}$$

implies Q is (Δ, Δ_2)-orthogonal. Another factorization of Q, coming from (2.6), is

$$\begin{aligned} Q &= \Delta Q^{-T} \Delta_2 \\ &= \Delta L_1^{-T} L_2 \bar{D}_2 \Delta_2. \end{aligned} \tag{2.7}$$

Next, we gradually manipulate (2.5) to introduce Q and Q^{-1}. First replace the diagonal matrix $\Delta D_1 \bar{D}_2 \Delta_2$ with $\mathcal{D}$. Then from (2.5) and (2.7) we have

$$\begin{aligned} \ddot{T}_G &= \mathcal{D} \bar{D}_2 L_2^T D_1 L_2 \bar{D}_2 \Delta_2 \mathcal{D}^{-1} \\ &= \mathcal{D}(\bar{D}_2 L_2^T L_1^{-1})(L_1 D_1 L_2 \bar{D}_2 \Delta_2)\mathcal{D}^{-1} \\ &= \mathcal{D} Q^{-1} L_1 D_1 (L_1^T L_1^{-T}) L_2 \bar{D}_2 \Delta_2 \mathcal{D}^{-1} \\ &= \mathcal{D} Q^{-1} (L_1 D_1 L_1^T)(L_1^{-T} L_2 \bar{D}_2 \Delta_2)\mathcal{D}^{-1} \\ &= \mathcal{D} Q^{-1} T \Delta Q \mathcal{D}^{-1} \\ &= \mathcal{D}(\Delta Q)^{-1} T_G (\Delta Q)\mathcal{D}^{-1}. \end{aligned} \tag{2.8}$$

Since $\Delta^3 = \Delta$, it follows that $(\Delta Q)^T \Delta (\Delta Q) = \Delta_2$, so ΔQ is also (Δ, Δ_2)-orthogonal. Finally we manipulate $T_G = \Delta T$ into its HR factorization,

$$\begin{aligned} T_G &= \Delta L_1 D_1 L_1^T \\ &= \Delta L_1 (L_2^{-T} \bar{D}_2^{-1})(\bar{D}_2 L_2^T) D_1 L_1^T \\ &= \Delta Q (\bar{D}_2 L_2^T D_1 L_1^T). \end{aligned} \tag{2.9}$$

Write D_1 as $|D_1| sign(D_1)$ to find

$$T_G = \Delta Q sign(D_1)[sign(D_1) \bar{D}_2 L_2^T sign(D_1)]|D_1|L_1^T. \tag{2.10}$$

Notice that $H = \Delta Q sign(D_1)$ is (Δ, Δ_2)-orthogonal and that $R = sign(D_1)\bar{D}_2 L_2^T sign(D_1)]|D_1|L_1^T$ is an upper triangular matrix with positive diagonal elements. Therefore (2.10) exhibits the HR factorization of T_G. Using (2.8) we discover

$$\ddot{T}_G = sign(D_1)\mathcal{D} H^{-1} T_G H (sign(D_1)\mathcal{D})^{-1}. \tag{2.11}$$

The result of one step of the HR algorithm applied to T_G is $H^{-1} T_G H = RH$. Note that (2.11) exhibits the diagonal similarity mentioned in the statement of the theorem between $\ddot{T}_G$ and this matrix. □

In an effort to extend this result to multiple steps of the HR algorithm we define $sign(D_1)\mathcal{D} = \mathcal{D}_1$, $H = H_1$, and $X = \mathcal{D}_1^{-1} \ddot{T}_G \mathcal{D}_1 = H_1^{-1} T_G H_1$, using the same notation as in the proof of Theorem 2.5. Since $H_1 \in O(\Delta, \Delta_2)$ and $H_1^{-1} T_G H_1 = \Delta_2 H_1^T T H_1$, X is Δ_2-symmetric. Therefore we can apply the HR algorithm to X and get $H_2^{-1} X H_2$, where $H_2 \in O(\Delta_2, \Delta_3)$. By Theorem 2.5,

$$\ddot{X} = \mathcal{D}_2 H_2^{-1} X H_2 \mathcal{D}_2^{-1},$$

where $\ddot{X}$ is the result of two steps of the LR algorithm applied to X, and $\mathcal{D}_2$ is a diagonal matrix. It is a straightforward exercise to show that

$$\ddot{X} = \mathcal{D}_1^{-1} \ddot{\ddot{T}}_G \mathcal{D}_1,$$

where $\ddot{\ddot{T}}_G$ is the result of four steps of the LR algorithm applied to T_G. This result then leads to

$$\begin{aligned} \ddot{\ddot{T}}_G &= \mathcal{D}_1 \mathcal{D}_2 H_2^{-1} X H_2 \mathcal{D}_2^{-1} \mathcal{D}_1^{-1} \\ &= \mathcal{D}_1 \mathcal{D}_2 H_2^{-1} H_1^{-1} T_G H_1 H_2 \mathcal{D}_2^{-1} \mathcal{D}_1^{-1}. \end{aligned}$$

Since $H_2 \in O(\Delta_2, \Delta_3)$ and $H_1 \in O(\Delta, \Delta_2)$ then $H_1^{-1} = \Delta_2 H_1^T \Delta$ and $H_2^{-1} = \Delta_3 H_2^T \Delta_2$. Therefore

$$H_2^{-1} H_1^{-1} T_G H_1 H_2 = \Delta_3 (H_1 H_2)^T T (H_1 H_2).$$

The result of two steps of the HR algorithm, $H_2^{-1} H_1^{-1} T_G H_1 H_2$, is Δ_3-symmetric and so we could apply the theorem yet again. One can prove that the result, on a nonsingular Δ-symmetric matrix, of $2k$ steps of the LR algorithm is diagonally similar to the result of k steps of the HR algorithm on the same matrix, provided that at each stage the factorizations exist.

2.2. *XHR* and *SSPLR*. Unfortunately the HR factorization does not always exist for a matrix A. The condition that H be an element of $O(\Delta, \Delta_0)$ is too restrictive. Fortunately if the space is expanded to $H \in O(\Omega, \Omega_0)$ such a factorization always exists; see [6] and [4].

THEOREM 2.6. [illegible] A [illegible] Ω [illegible] $A = HR$ [illegible] $H \in O(\Omega, \Omega_0)$ [illegible] Ω_0 [illegible] $R \in \mathbb{R}^{n\times n}$ [illegible] XHR factorization [illegible] A [illegible] Ω [illegible] Ω_0 [illegible] A [illegible] Ω

[illegible]. See Liu [4]. □

Is there an algorithm based on triangular factorization other than the LR algorithm that is related to a step of the XHR algorithm and not the HR algorithm? Since every nonsingular matrix has an XHR factorization, the algorithm and the triangular factorization it comes from will be more flexible than the LR factorization. For this reason we relax the conditions on our triangular factors. The following theorem is very much like the modified Bruhat decomposition in [5], but differs in that it is specific to SSP matrices.

THEOREM 2.7. [illegible] A [illegible] A [illegible] $A = L\Omega L^T$ [illegible] L [illegible] $\Omega = D\bar{\Omega}D$ [illegible] D [illegible] $\bar{\Omega}$ [illegible] $\bar{\Omega}$ [illegible]

[illegible]. This is slightly modified from the factorization in [4], but the proof is the same. See Liu [4]. □

We define the $SSPLR$ factorization of a general matrix A as a factorization of A into $L\Omega U$, but this may not always exits. In this context L is a unit lower triangular matrix, U is an upper triangular matrix, and Ω is an SSP matrix. The result of a single step of the $SSPLR$ algorithm applied to A is the matrix $\dot{A} = \Omega UL$. Notice that $\dot{A} = LAL^{-1}$, so the eigenvalues of $\dot{A}$ and A are the same. The $SSPLR$ factorization is a triangular factorization that is more flexible than the LR factorization since it always exists for nonsingular symmetric matrices. The XHR factorization is a similar generalization of HR. We define a single step of the XHR algorithm applied to a matrix A given its XHR factorization $A = HR$ to be the matrix RH.

Now the question of whether or not a single step of the XHR algorithm is equivalent to a combination of steps of the $SSPLR$ and LR algorithms can be answered.

THEOREM 2.8. [illegible] T_G [illegible] LR [illegible] Δ [illegible] T_G [illegible] Δ [illegible] XHR [illegible] T_G [illegible] A [illegible]

[illegible]. As before, we can write either an unreduced tridiagonal matrix or a Δ-symmetric matrix T_G as ΔT with T symmetric. Since T_G has an LR factorization we may factor T into unit triangular matrices L_1 and L_1^T, and a diagonal matrix D_1 as $L_1 D_1 L_1^T$. Consider one step of the LR algorithm applied to $T_G = \Delta T$. We get

$$\begin{aligned} T_G &= \Delta T \\ &= \Delta L_1 \Delta \Delta D_1 L_1^T, \\ \dot{T}_G &= \Delta D_1 L_1^T \Delta L_1 \Delta. \end{aligned} \tag{2.12}$$

The $SSPLR$ factorization of the symmetric matrix $L_1^T \Delta L_1$ is $L_2 \Omega_2 L_2^T$. The matrix L_2 is unit lower triangular and $\Omega_2 = \bar{D}_2 \bar{\Omega}_2 \bar{D}_2$, where $\bar{D}_2$ is a positive diagonal matrix and $\bar{\Omega}_2$ is an SSP matrix. This factorization always exists by Theorem 2.7. We take another step to get

$$\begin{aligned} \dot{T}_G &= \Delta D_1 L_2 \Omega_2 L_2^T \Delta \\ &= (\Delta D_1 L_2 D_1^{-1} \Delta)(\Delta D_1 \Omega_2 L_2^T \Delta), && (2.13) \\ \ddot{T}_G &= (\Delta D_1 \Omega_2 L_2^T \Delta)(\Delta D_1 L_2 D_1^{-1} \Delta). && (2.14) \end{aligned}$$

Notice that $\Delta D_1 L_2 D_1^{-1} \Delta$ is a unit lower triangular matrix and $\Delta D_1 \Omega_2 L_2^T \Delta$ can be written as an SSP matrix times an upper triangular matrix. Thus, (2.13) exhibits the $SSPLR$ factorization of the matrix $\dot{T}_G$. $\ddot{T}_G$ is the result of one step of the $SSPLR$ algorithm applied to $\dot{T}_G$. It remains to show that $\ddot{T}_G$ is similar to the result of one step of the XHR algorithm, as opposed to one step of the HR algorithm. We manipulate the triangular factorization of T_G by inserting the identity to get its XHR factorization,

$$T_G = \Delta L_1 (L_2^{-T} \bar{D}_2^{-1} sign(D_1))(sign(D_1) \bar{D}_2 L_2^T) D_1 L_1^T, \tag{2.15}$$

where $sign(D_1)|D_1| = D_1$. Observe that $R = sign(D_1)\bar{D}_2 L_2^T D_1 L_1^T$ is an upper triangular matrix with positive diagonal elements. The rest of the factorization above is H since

$$\begin{aligned} (\Delta L_1 L_2^{-T} \bar{D}_2^{-1} sign(D_1))^T \; \Delta \; (\Delta L_1 L_2^{-T} \bar{D}_2^{-1} sign(D_1)) \\ = sign(D_1) Q^T \Delta\Delta\Delta Q sign(D_1) \\ = sign(D_1) Q^T \Delta Q sign(D_1) \\ = sign(D_1) \bar{\Omega}_2 sign(D_1) = \Omega_3. \end{aligned}$$

We use the definition of Q from the proof of Theorem 2.5. If $H = \Delta L_1 L_2^{-T} \bar{D}_2^{-1} sign(D_1)$, then the above calculation implies H is (Δ, Ω_3)-orthogonal. This fact along with (2.15) implies that H and R are the XHR factors of T_G. Since Ω_3 is an SSP matrix and not a signature matrix this cannot be an HR factorization. It remains to show that $H^{-1} T_G H$ is similar, by an SSP matrix times a diagonal matrix, to $\ddot{T}_G$. Evaluating (2.14) we get

$$\begin{aligned} \ddot{T}_G &= (\Delta D_1 \Omega_2 L_2^T \Delta)(\Delta D_1 L_2 D_1^{-1} \Delta) \\ &= \Delta D_1 \bar{D}_2 \bar{\Omega}_2 \bar{D}_2 L_2^T D_1 L_2 D_1^{-1} \Delta \end{aligned}$$

since $\Omega_2 = \bar{D}_2 \bar{\Omega}_2 \bar{D}_2$. Inserting $sign(D_1)^2$ and $L_1^{-1} \Delta^2 L_1$ we get

$$\begin{aligned} \ddot{T}_G &= \Delta D_1 \bar{D}_2 \bar{\Omega}_2 sign(D_1)(sign(D_1) \bar{D}_2 L_2^T L_1^{-1} \Delta) \Delta L_1 D_1 L_2 D_1^{-1} \Delta \\ &= \Delta D_1 \bar{D}_2 \bar{\Omega}_2 sign(D_1) H^{-1} \Delta L_1 D_1 L_2 D_1^{-1} \Delta \end{aligned}$$

because $H^{-1} = sign(D_1) \bar{D}_2 L_2^T L_1^{-1} \Delta$. Define $\mathcal{D} = sign(D_1) \bar{\Omega}_2 \bar{D}_2^{-1} D_1^{-1} \Delta$ and $\mathcal{D}^{-1} = \Delta D_1 \bar{D}_2 \bar{\Omega}_2 sign(D_1)$ to get

$$\begin{aligned} \ddot{T}_G &= \mathcal{D}^{-1} H^{-1} \Delta L_1 D_1 L_2 \bar{D}_2 \bar{\Omega}_2 sign(D_1) \mathcal{D} \\ &= \mathcal{D}^{-1} H^{-1} (\Delta L_1 D_1 L_1^T)(L_1^{-T} L_2 \bar{D}_2 \bar{\Omega}_2 sign(D_1)) \mathcal{D}. \end{aligned}$$

As a result of $H^T \Delta H = \Omega_3$ we also know that $H = \Delta H^{-T} \Omega_3$ This implies

$$\begin{aligned} \ddot{T}_G &= \mathcal{D}^{-1} H^{-1} (\Delta L_1 D_1 L_1^T) H \mathcal{D}, && (2.16) \\ &= \mathcal{D}^{-1} H^{-1} T_G H \mathcal{D}. \end{aligned}$$

The similarity matrix is $\mathcal{D}$, which can be written as an SSP matrix times a diagonal matrix. $\square$

Unfortunately $H^{-1}T_G H = \Omega_3 H^T T H$ is an Ω_3-symmetric matrix but Ω_3 is not a signature matrix, and so we cannot apply this theorem again in the way we did for Theorem 2.5.

Acknowledgments. The author would like to thank Beresford Parlett and Loyce Adams for their helpful comments on the manuscript. The author would also like to thank the referees for their helpful suggestions.

REFERENCES

[1] P. Benner, H. Faßbender, and D. S. Watkins, *Two connections between the SR and HR eigenvalue algorithms*, Linear Algebra Appl., 272 (1998), pp. 17–32.

[2] M. A. Brebner and J. Grad, *Eigenvalues of $Ax = B\lambda x$ for real symmetric matrices A and B by reduction to pseudosymmetric form and the HR process*, Linear Algebra Appl., 43 (1982), pp. 99–118.

[3] A. Bunse-Gerstner, *An analysis of the HR algorithm for computing the eigenvalues of a matrix*, Linear Algebra Appl., 35 (1981), pp. 155–173.

[4] Z.-S. Liu, *On the Extended HR Algorithm*, Technical Report PAM-564, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, 1992.

[5] E. Tyrtyshnikov, *Matrix Bruhat decompositions with a remark on the $QR(GR)$ algorithm*, Linear Algebra Appl., 250 (1997), pp. 61–68.

[6] P. Wiberg, *A Study of the HR and Extended HR Methods for the Standard Eigenvalue Problem*, Technical Report TR/PA/97/33, CERFACS, Toulouse, France, 1997.

[7] H. Xu, *The relation between the QR and LR algorithms*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 551–555.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 75–91

PERTURBATION SPLITTING FOR MORE ACCURATE EIGENVALUES[*]

RUI RALHA[†]

Abstract. Let T be a symmetric tridiagonal matrix with entries and eigenvalues of different magnitudes. For some T, small entrywise relative perturbations induce small errors in the eigenvalues, independently of the size of the entries of the matrix; this is certainly true when the perturbed matrix can be written as $\widetilde{T} = X^TTX$ with small $||X^TX - I||$. Even if it is not possible to express in this way the perturbations in every entry of T, much can be gained by doing so for as many as possible entries of larger magnitude. We propose a technique which consists of splitting multiplicative and additive perturbations to produce new error bounds which, for some matrices, are much sharper than the usual ones. Such bounds may be useful in the development of improved software for the tridiagonal eigenvalue problem, and we describe their role in the context of a mixed precision bisection-like procedure. Using the very same idea of splitting perturbations (multiplicative and additive), we show that when T defines well its eigenvalues, the numerical values of the pivots in the usual decomposition $T - \lambda I = LDL^T$ may be used to compute approximations with high relative precision.

Key words. symmetric tridiagonal matrices, eigenvalues, perturbation theory

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

DOI. 10.1137/070687049

1. Introduction.

Let A and E be n-by-n symmetric matrices. Let $\lambda_1 \leq \cdots \leq \lambda_n$ and $\widetilde{\lambda}_1 \leq \cdots \leq \widetilde{\lambda}_n$ be the eigenvalues of A and $\widetilde{A} = A + E$, respectively. Then $|\lambda_k - \widetilde{\lambda}_k| \leq \|E\|_2$. This is a classical result in the perturbation theory (see [44, pp. 101–102]), which is usually referred to as Weyl's theorem (see, for instance, [9, p. 198]).

Weyl's theorem can be used to get error bounds for the eigenvalues computed by any backward stable algorithm since such an algorithm computes eigenvalues $\widetilde{\lambda}_k$ that are the exact eigenvalues of $\widetilde{A} = A + E$, where $\|E\|_2 = O(\epsilon)\|A\|_2$. (Here and throughout the paper we will use ϵ to denote the rounding error unit.) This is a very satisfactory error bound for large eigenvalues, especially those of magnitude close to $\|A\|_2$, but eigenvalues much smaller than $\|A\|_2$ will have fewer correct digits (eventually none in extreme cases).

The decade starting in 1990 was fertile in new results on bounds for relative errors of eigenvalues and several authors have contributed to this [1], [4], [5], [16], [17], [22], [32], [33], [34], [35], [42]. In [22], Ipsen presents a good survey of the work done until 1998. Not surprisingly, many of the published results are for the Hermitian positive definite case. For an Hermitian indefinite matrix A and, more generally, for normal matrices, the Hermitian positive-semidefinite factor H in the polar decomposition $A = HU$, with U unitary, may be used to derive bounds for the eigenvalues of A (see [22, Theorems 2.4 and 2.10] and the references therein).

The first relative perturbation bound for eigenvalues is due to Ostrowski. Let $\widehat{A} = XAX^*$, with X nonsingular, be a multiplicative perturbation of an Hermitian matrix

[*] Received by the editors April 2, 2007; accepted for publication (in revised form) by Z. Drmač March 3, 2008; published electronically February 27, 2009. This research was supported by Portuguese Foundation for Science and Technology research program POCI 2010.

http://www.siam.org/journals/simax/31-1/68704.html

[†] Departamento de Matemática, Universidade do Minho, 4710-057 Braga, Portugal (r_ralha@math.uminho.pt).

A; for the eigenvalues λ_k and $\widehat{\lambda}_k$, of A and $\widehat{A}$, respectively, we have [21, Theorem 4.5.9]

$$\lambda_k \cdot \lambda_{\min}\left(XX^*\right) \leq \widehat{\lambda}_k \leq \lambda_k \cdot \lambda_{\max}\left(XX^*\right).$$

This result is at the heart of high relative accuracy theory for the eigenvalues of Hermitian matrices (and singular values). An immediate consequence, for real symmetric matrices, is the following (Theorem 2.1 in [16]): let A have eigenvalues λ_k and $\widehat{A} = X^T A X$ have eigenvalues $\widehat{\lambda}_k$. Then $|\widehat{\lambda}_k - \lambda_k| \leq |\lambda_k| \; ||X^T X - I||_2$. Following Demmel [9, p. 208], we will refer to this result as the relative Weyl's theorem.

Some types of matrices are known to define well their eigenvalues and/or singular values. In 1990, Demmel and Kahan [4] showed that small relative perturbations in the entries of any bidiagonal matrix cause small relative errors in the singular values, independent of their magnitudes. They also proposed the zero-shifted QR algorithm to compute such singular values with high relative accuracy. Another remarkable development in this area of fast and highly accurate computation of the singular values of bidiagonal matrices was the dqds algorithm [18], [38]. Furthermore, any matrix with an acyclic graph (bidiagonals and many others) defines well its singular values, and these may be computed to high accuracy using bisection [6].

In [11], Demmel et al. showed that it is possible to compute efficiently a highly accurate SVD of a dense rectangular matrix A from a rank-revealing decomposition (RRD) $A = XDY^T$, i.e., a decomposition where D is diagonal and X and Y are well conditioned (but otherwise arbitrary); furthermore, also in [11], a variety of matrix classes were described for which a special form of Gaussian elimination with complete pivoting does provide the necessary accuracy of the computed factors $\widetilde{X}$, $\widetilde{D}$, and $\widetilde{Y}$. For some structured matrices (these include, among others, Cauchy matrices, Vandermonde matrices, M-matrices, and totally nonnegative matrices), forward stable algorithms have been proposed for the computation of highly accurate RRD. See [10], [11], [12], [15], and [26], [27], [28], [29].

Congruence transformations play an important role in the perturbation theory of the eigenvalues of an Hermitian positive-definite matrix A (see [22, Corollary 2.2] and [34, Theorem 2.4]). For scaled diagonally dominant ( ) matrices, diagonal congruence transformations may be used to pull the grading out of the matrix [1], [5], [35], [9]. If A is indefinite, the error bounds are the same as the error bounds for the eigenvalues of the best scaled version of the positive-definite polar factor of A (see [22, Corollary 2.6] and [42, Theorem 2.13]).

Symmetric tridiagonal matrices do not always define well their eigenvalues, not even in the positive-definite case. In this paper, we focus our attention on symmetric tridiagonal matrices with entries of different magnitudes. Our matrices, however, are not necessarily .

Suppose that we are given a symmetric matrix A which has entries of different orders of magnitude and assume small relative perturbations of size $O(\varepsilon)$ in its entries (or, at least, small relative perturbations in the entries of larger size). With $\widetilde{A} = A+E$, it is clear that $\|E\|_2$ is proportional to the size of the largest entries in A, and the classical error bound, provided by Weyl's theorem, may not be very satisfactory for small eigenvalues, if they arise. For this reason, we attack $\widetilde{A}$ with a congruence $X^T \widetilde{A} X$ to get $\widehat{A} = A+F$ with $\|F\|_2 < \|E\|_2$ and $||X^T X - I||_2$ of size $O(\varepsilon)$; the relative Weyl's theorem gives

$$|\widetilde{\lambda}_k - \widehat{\lambda}_k| \leq ||X^T X - I||_2 \cdot |\widetilde{\lambda}_k|, \tag{1.1}$$

and we get

$$|\widetilde{\lambda}_k - \lambda_k| \le |\widetilde{\lambda}_k - \widehat{\lambda}_k| + |\widehat{\lambda}_k - \lambda_k| \le ||X^T X - I||_2 \cdot |\widetilde{\lambda}_k| + \|F\|_2\,, \tag{1.2}$$

which, in some cases, is a much sharper bound than

$$|\widetilde{\lambda}_k - \lambda_k| \le \|E\|_2\,. \tag{1.3}$$

In the following sections, we exploit this idea in the context of symmetric tridiagonal matrices, although it can also be applied to dense symmetric matrices. In section 2, we analyze the perturbation of the eigenvalues of affine transformations of Golub–Kahan matrices. Section 3 contains the main perturbation result, Theorem 3.1, which states that a symmetric tridiagonal matrix T, with diagonals a_j, defines well the eigenvalues whose magnitude is not much smaller than $\max |a_j|$. In section 4 we present a detailed numerical example to show that for matrices with entries of different magnitudes, depending upon the location of the entries of larger size, the eigenvalues may or may not be all well defined. In section 5 we describe a fast procedure that will produce an estimate for the value of $\|F\|_2$ in the bound (1.2). In sections 6 and 7 we present applications of our perturbation results; in section 6 we show that the numerical values of the pivots in the decomposition $T - \lambda I = LDL^T$, computed in the usual way, may be used to determine the eigenvalues with high relative accuracy, if the matrix T defines them well, and in section 7 we show that our results are useful in the context of a mixed precision bisection algorithm.

2. Constant main diagonal. It is well known that, for n even, the eigenvalues of the Golub–Kahan matrix

$$T(0) = \begin{bmatrix} 0 & b_1 & & & \\ b_1 & 0 & b_2 & & \\ & b_2 & 0 & \ddots & \\ & & \ddots & \ddots & b_{n-1} \\ & & & b_{n-1} & 0 \end{bmatrix} \tag{2.1}$$

are

$$-\sigma_1 \le \cdots \le -\sigma_{\frac{n}{2}} \le \sigma_{\frac{n}{2}} \le \cdots \le \sigma_1, \tag{2.2}$$

where σ_k $(k = 1, \dots, \frac{n}{2})$ are the singular values of

$$B = \begin{bmatrix} b_1 & b_2 & & \\ & b_3 & \ddots & \\ & & \ddots & b_{n-2} \\ & & & b_{n-1} \end{bmatrix} \tag{2.3}$$

(see, for instance, Lemma 5.5 in [9]). This relation may be used in both directions; that is, one may compute singular values of B as the corresponding positive eigenvalues of $T(0)$ or one may compute eigenvalues of $T(0)$ from the corresponding singular values of B. This last option may also be used for the computation of the eigenvalues of a skew-symmetric tridiagonal matrix with high relative accuracy (see [41]). We will therefore be interested in matrices with the structure given in (2.1), with n even or odd. We have the following result.

Proposition 2.1. [illegible] $T(0)$ [illegible] (2.1) [illegible] D_{2k-1} $k = 1, \dots, \frac{n}{2}$ [illegible] n [illegible] $k = 1, \dots, \frac{n+1}{2}$ [illegible] n [illegible] D_{2k} $k = 1, \dots, \frac{n}{2}$ [illegible] n [illegible] $k = 1, \dots, \frac{n-1}{2}$ [illegible] n [illegible] $T(0)$ [illegible]

$$D_{2k-1} = 0,\ D_{2k} = (-1)^k \cdot \prod\nolimits_{j=1}^{k} b_{2j-1}^2. \tag{2.4}$$

[illegible] n [illegible] $T(0)$ [illegible] $b_{2j-1} = 0$ [illegible] j, $1 \le j \le k$; [illegible] n [illegible] $D_n = 0$. [illegible] $T(0)$ [illegible]

[illegible]. The proof follows easily from $D_1 = 0$, $D_2 = -b_1^2$, and the relation $D_j = -b_{j-1}^2 \times D_{j-2}$ for $j \ge 3$. □

When n is odd, we may keep relating $T(0)$ to a bidiagonal matrix. For this, we construct a matrix of even order by adding a row and a column of zeros to $T(0)$. The resulting matrix has a double eigenvalue equal to zero. The corresponding bidiagonal in (2.3) is now replaced by the singular matrix with diagonal entries $b_1, \dots, b_{n-2}$, $b_n = 0$ and superdiagonal entries $b_2, \dots, b_{n-1}$.

Small relative perturbations of the off-diagonal pairs of $T(0)$ may be expressed in terms of a congruence transformation $X^T T(0) X$ with X diagonal very close to identity (see [1], [16], and [22, Example 5.1]). Therefore, $T(0)$ defines well its eigenvalues (even when n is odd, because the zero eigenvalue is unchanged by perturbations in the off-diagonal entries). From [9, Theorem 5.13] we may conclude the following.

Corollary 2.2. [illegible] $T(0)$ [illegible] (2.1) [illegible] $\widetilde{T}(0)$ [illegible] $T(0)$ [illegible] b_k [illegible] $\widetilde{b}_k = b_k(1 + \delta_k)$ [illegible] $|\delta_k| \le \varepsilon \ll 1$ [illegible] $\lambda_1 \le \dots \le \lambda_n$ [illegible] $T(0)$ [illegible] $\widetilde{\lambda}_1 \le \dots \le \widetilde{\lambda}_n$ [illegible] $\widetilde{T}(0)$ [illegible]

$$|\widetilde{\lambda}_k - \lambda_k| \le \xi\left(n, \varepsilon\right) |\lambda_k|, \tag{2.5}$$

[illegible]

$$\xi\left(n, \varepsilon\right) = (2n - 1)\varepsilon + O\left(\varepsilon^2\right). \tag{2.6}$$

Now, we consider affine transformations of $T(0)$. If $T(c)$ is a symmetric tridiagonal matrix whose main diagonal entries are equal to a constant c, then $T(0) = T(c) - cI$ has zeros in the main diagonal and Corollary 2.2 does apply. We have the following.

Proposition 2.3. [illegible] $\lambda_k(0)$ [illegible] $\lambda_k(c)$ [illegible] $T(0)$ [illegible] $T(c) = T(0) + cI$ [illegible] $\widetilde{\lambda}_k(0)$ [illegible] $\widetilde{T}(0)$ [illegible] 2.2 [illegible] $\widetilde{\lambda}_k(c)$ [illegible] $\widetilde{T}(c) = \widetilde{T}(0) + cI$ [illegible] $\lambda_k(c) \ne 0$ [illegible]

$$|\widetilde{\lambda}_k(c) - \lambda_k(c)| \le \xi(n, \varepsilon) \left| 1 - \frac{c}{\lambda_k(c)} \right| |\lambda_k(c)|, \tag{2.7}$$

[illegible] $\xi(n, \varepsilon)$ [illegible] (2.6)

[illegible]. Since $\lambda_k(c) = \lambda_k(0) + c$ and $\widetilde{\lambda}_k(c) = \widetilde{\lambda}_k(0) + c$, we have $\widetilde{\lambda}_k(c) - \lambda_k(c) = \widetilde{\lambda}_k(0) - \lambda_k(0)$; using (2.5), we get

$$|\widetilde{\lambda}_k(c) - \lambda_k(c)| \le \xi(n, \varepsilon) |\lambda_k(0)|,$$

which, for $\lambda_k(c) \ne 0$, can be written as

$$|\widetilde{\lambda}_k(c) - \lambda_k(c)| \le \xi(n, \varepsilon) \left| \frac{\lambda_k(0)}{\lambda_k(c)} \right| |\lambda_k(c)|. \tag{2.8}$$

Replacing $\lambda_k(0)$ with $\lambda_k(c) - c$ gives (2.7). $\square$

Small relative perturbations in the off-diagonal entries of $T(c)$ cause relative errors in the eigenvalues which depend upon the ratio

$$\frac{\lambda_k(0)}{\lambda_k(c)} = 1 - \frac{c}{\lambda_k(c)}. \tag{2.9}$$

Therefore, we see that the relative errors will be small except for those eigenvalues $\lambda_k(c)$ such that $|\lambda_k(0)| \gg |\lambda_k(c)|$, i.e.,

$$|\lambda_k(c)| \ll |c|\,. \tag{2.10}$$

Furthermore, (2.7) shows that the relative error of $\widetilde{\lambda}_k(c)$ approaches zero when $\lambda_k(c)$ gets close to c.

[illegible] 1. Consider the matrix

$$T(1) = \begin{bmatrix} 1 & 10^6 & & & & \\ 10^6 & 1 & 1 & & & \\ & 1 & 1 & 1 & & \\ & & 1 & 1 & 1 & \\ & & & 1 & 1 & 10^6 \\ & & & & 10^6 & 1 \end{bmatrix}. \tag{2.11}$$

The function eig of MATLAB (version 7.4) produces the following approximations for the eigenvalues (note that with a previous version of MATLAB we got much worse values for $\widetilde{\lambda}_3(1)$ and $\widetilde{\lambda}_4(1)$):

$$\begin{aligned} \widetilde{\lambda}_1(1) &= -9.999990000005000e{+}005, & \widetilde{\lambda}_2(1) &= -9.999990000005000e{+}005, \\ \widetilde{\lambda}_3(1) &= 1.139421890172798e{-}012, & \widetilde{\lambda}_4(1) &= 1.999999999999141e{+}000, \\ \widetilde{\lambda}_5(1) &= 1.000001000000500e{+}006, & \widetilde{\lambda}_6(1) &= 1.000001000000500e{+}006. \end{aligned}$$

The classical error analysis gives us, with $\epsilon = 2^{-52}$, for all $k = 1, \dots, 6$,

$$|\widetilde{\lambda}_k(1) - \lambda_k(1)| \le O(\epsilon)\, \|T(1)\|_2 = O(10^{-10}).$$

Thus, for $k \neq 3$ and $k \neq 4$, $\widetilde{\lambda}_k(1)$ is an accurate approximation of the corresponding true eigenvalue $\lambda_k(1)$, and $\widetilde{\lambda}_4(1)$ has at least 9 or 10 correct decimal digits. Interestingly, we may improve upon the computed values $\widetilde{\lambda}_3(1)$ and $\widetilde{\lambda}_4(1)$. Since we know the exact value $\det(T(1)) = 2 \times 10^{12} - 1$, we use the relation

$$\lambda_3(1) = \det(T(1)) / \prod_{k=1, k\neq 3}^{6} \lambda_k(1)$$

to compute an approximation

$$\overline{\lambda}_3(1) = fl\left(\det(T(1)) / \prod_{k=1, k\neq 3}^{6} \widetilde{\lambda}_k(1) \right) \tag{2.12}$$

which has at least nine correct decimal significant digits. We have

$$\begin{aligned}\overline{\lambda}_3(1) &= \det(T(1))/\left(\prod_{k=1,k\neq 3}^{6} \lambda_k(1)\,(1+\phi_k)\right)(1+\kappa\epsilon)\\ &= \lambda_3(1)\cdot\left((1+\kappa\epsilon)\prod_{k=1,k\neq 3}^{6}(1+\phi_k)^{-1}\right),\end{aligned}$$

where ϕ_k for $k \neq 3$ is the relative error in $\widetilde{\lambda}_k(1)$ and the term $\kappa\epsilon$, with $\kappa \leq 5.05$, accounts for the rounding errors in the four multiplications and one division. Since the relative errors ϕ_k in the four eigenvalues of larger size are all bounded by $O\left(2^{-52}\right)$, the size of the relative error in $\overline{\lambda}_3(1)$ is determined essentially by the size of ϕ_4, which we know to be bounded by $O\left(10^{-10}\right)$. The computation of (2.12) in MATLAB produces $\overline{\lambda}_3(1) = 9.999999999999297e{-}013$. Since the interval $[\lambda_3(1),\ \lambda_4(1)]$ of the true eigenvalues is known to be centered in $c = 1$, we compute $\overline{\lambda}_4(1) = 2 - \overline{\lambda}_3(1) = 1.999999999999000e{+}000$ with 16 correct digits. Again, we may use (2.12), replacing $\widetilde{\lambda}_4(1)$ with $\overline{\lambda}_4(1)$ to compute $\overline{\overline{\lambda}}_3(1) = 1.000000000000000e{-}012$ with a relative error bounded by $O(\epsilon)$. Now, according to Proposition 2.3, if MATLAB could deliver the exact eigenvalues of a matrix differing from $T(1)$ by relative perturbations of size $O(\epsilon)$ in the off-diagonal entries,[1] $\widetilde{\lambda}_4(1)$ would be closer to $\overline{\lambda}_4(1)$, and for $\widetilde{\lambda}_3(1)$ we would have

$$\left|\widetilde{\lambda}_3(1) - \lambda_3(1)\right| \leq \left|1 - \frac{1}{\lambda_3(1)}\right| O(\epsilon)\,|\lambda_3(1)| \approx 10^{-4}\,|\lambda_3(1)|\,,$$

and such approximation, although not as good as $\overline{\overline{\lambda}}_3(1)$ or even $\overline{\lambda}_3(1)$, is significantly better than the computed $\widetilde{\lambda}_3(1)$. It is also worth mentioning that in MATLAB, svd(T) and [L,U]=lu(T); eig(U*L), where T is the matrix in our example, both produce approximations $\widetilde{\lambda}_4(1)$ and $\widetilde{\lambda}_3(1)$ which do satisfy the error bound (2.7).[2]

We conclude this section by emphasizing that the matrix in our example is not and the theory of Barlow and Demmel does not apply here.

3. A perturbation theory result. In the previous section, we showed that small relative changes in the off-diagonal entries of a symmetric tridiagonal matrix with constant main diagonal c do not cause too much perturbation in those eigenvalues of magnitude not much smaller than the constant $|c|$. To this end, we have used a simple affine transformation of the given matrix to produce a Golub–Kahan matrix whose relative perturbations in the off-diagonal pairs may be entirely expressed in terms of a congruence transformation $X^TT(0)X$, with X very close to identity. A similar result may be obtained without the affine transform by directly expressing the perturbations in the off-diagonal entries in terms of a congruence transformation. This is a more general procedure since it applies to any symmetric tridiagonal matrix. We have the following theorem.

[1] This is what the bisection method can actually deliver; see section 6.

[2] Zlatko Drmač has brought to our attention the accuracy of these approximations.

THEOREM 3.1.

$$T = \begin{bmatrix} a_1 & b_1 & & \\ b_1 & & & \\ & & & b_{n-1} \\ & & b_{n-1} & a_n \end{bmatrix} \tag{3.1}$$

$$\widetilde{T} = \begin{bmatrix} a_1\,(1+\eta_1) & b_1\,(1+\delta_1) & & \\ b_1\,(1+\delta_1) & & & \\ & & & b_{n-1}\,(1+\delta_{n-1}) \\ & & b_{n-1}\,(1+\delta_{n-1}) & a_n\,(1+\eta_n) \end{bmatrix}, \tag{3.2}$$

δ_k η_k $|\delta_k| \le \varepsilon$ $|\eta_k| \le \varepsilon$ λ_k $\widetilde{\lambda}_k$ T $\widetilde{T}$ $k = 1, \ldots, n$

$$|\lambda_k - \widetilde{\lambda}_k| < 2.02 n\varepsilon \left(\max_j |a_j| + |\widetilde{\lambda}_k| \right) \tag{3.3}$$

. We use a diagonal congruence to account for all the off-diag perturbations and then just see what it does to the diagonal entries: lo and behold, it makes just a few more changes from what was there initially. Concretely, if we write

$$\widehat{T} = X^T \widetilde{T} X \tag{3.4}$$

with X diagonal, $X(1,1) = 1$, $X(2,2) = (1+\delta_1)^{-1}$, and

$$X(j,j) = (1+\delta_{j-1})^{-1}\, X(j-1,j-1)^{-1}, \qquad j = 3, \ldots, n, \tag{3.5}$$

we get $\widehat{T}(i,j) = T(i,j)$ for $i \neq j$, $\widehat{T}(1,1) = a_1\,(1+\eta_1)$, and

$$\widehat{T}(j,j) = a_j\,(1+\eta_j) \cdot X(j,j)^2, \qquad j = 2, \ldots, n. \tag{3.6}$$

We write

$$X(j,j)^2 = 1 + \phi_j, \tag{3.7}$$

and since $\phi_1 = 0$ and $|\delta_j| \le \varepsilon$, from (3.5), assuming that $2\,(n-1)\,\varepsilon \le 0.01$, we get

$$|\phi_j| \le 2.02\,(j-1)\,\varepsilon, \qquad j = 2, \ldots, n, \tag{3.8}$$

and $||X^T X - I||_2 \le \max_j |\phi_j| < 2.02 n\varepsilon$. From (3.6)–(3.8) and taking into account that $|\eta_j| \le \varepsilon$, we may write, for each $j = 1, \ldots, n$, assuming that $(2n-1)\,\varepsilon \le 0.01$, $\widehat{T}(j,j) = a_j\,(1+\theta_j)$ with $|\theta_j| \le 1.01\,(2j-1)\,\varepsilon$. Therefore, we have $\widehat{T} = T + F$ with F a diagonal matrix such that

$$\|F\|_2 = \max_j |a_j|\,|\theta_j| < 2.02 n\varepsilon \cdot \max_j |a_j|\,. \tag{3.9}$$

Applying the relative Weyl's theorem to matrices $\widehat{T}$ and $\widetilde{T}$ in (3.4), we get $|\widehat{\lambda}_k - \widetilde{\lambda}_k| \leq |\widetilde{\lambda}_k| \cdot ||X^T X - I||_2$, and we may finally write $|\lambda_k - \widetilde{\lambda}_k| \leq |\lambda_k - \widehat{\lambda}_k| + |\widehat{\lambda}_k - \widetilde{\lambda}_k| \leq \|F\|_2 + |\widetilde{\lambda}_k| \cdot ||X^T X - I||_2$, which, after some simplifications, gives (3.3). □

If $\widetilde{\lambda}_k \neq 0$, the bound (3.3) may be written as

$$\frac{\left|\lambda_k - \widetilde{\lambda}_k\right|}{\left|\widetilde{\lambda}_k\right|} < 2.02 n\varepsilon \left(1 + \frac{\max\limits_j |a_j|}{\left|\widetilde{\lambda}_k\right|}\right). \tag{3.10}$$

Part of the novelty of Theorem 3.1 for relative perturbation theory is that, as expressed in (3.10), a general symmetric tridiagonal matrix T defines well those eigenvalues whose magnitude is not much smaller than $\max |a_j|$.

For the case of a matrix with zeros in the main diagonal, we get from (3.3)

$$|\lambda_k - \widetilde{\lambda}_k| \leq 2.02 n\varepsilon |\widetilde{\lambda}_k| \tag{3.11}$$

and we note that this is essentially the bound given in (2.5), with $|\lambda_k|$ replaced with $|\widetilde{\lambda}_k|$.

It must be observed that there are many distinct congruences X which are able to produce $\widehat{T}$ with unperturbed off-diagonal entries. We have used X with $X(1,1) = 1$, but it is possible to use a different X, setting $X(k,k) = 1$, for any $k = 1, \ldots, n$; then, we choose the values of $X(k-1, k-1), \ldots, X(1,1)$ to remove perturbations from entries $\widetilde{b}_{k-1}, \ldots, \widetilde{b}_1$, by this order, and $X(k+1, k+1), \ldots, X(n,n)$ to remove perturbations from entries $\widetilde{b}_k, \ldots, \widetilde{b}_{n-1}$. In particular, by choosing $k = n/2$ we may reduce the bounds (3.3) and (3.10) by a factor of 2.

Finally, we remark that there is a diagonal X which, besides the off-diagonal perturbations, also expresses, in multiplicative terms, the perturbation in any diagonal entry $\widetilde{a}_k$: $X(k,k)$ is chosen to remove the perturbation in $\widetilde{a}_k$, and the remaining entries of X are determined as we have just described. So, in the bounds (3.3) and (3.10) we may replace $\max |a_j|$ with the second largest absolute value of the diagonal entries of T.

4. More general perturbations: An example. There are matrices for which the bound (3.10) is sharp. This is the case with matrices of constant main diagonal c since, as we have seen in section 2, the relative error in $\widetilde{\lambda}_k(c)$ depends upon the ratio $c/\lambda_k(c)$. In discussing the relative errors of small eigenvalues computed with the bisection method, Wilkinson also observed (see [44, p. 307]) that the method (which we know to be able to compute accurately the eigenvalues if the matrix defines them well) could not compute accurately the small eigenvalues of such a matrix.

However, we know that there are other matrices for which the bound (3.10) is too pessimistic. This is the case of the   matrices. We now show that there are other matrices, not , which define well their eigenvalues, even in cases where their size is much smaller than that of some of the diagonal entries.

In the previous section, we expressed the perturbations in the off-diagonal entries in terms of a diagonal congruence,

$$\widehat{T} = X^T \widetilde{T} X. \tag{4.1}$$

Although X does not account for the perturbations in the diagonal entries, the key point of our analysis is based upon the fact that

$$\widehat{T} = T + F \tag{4.2}$$

with $\|F\|_2$ independent of the size of the off-diagonal entries.

In a more general situation, T may have entries of different order of magnitude, and we are interested in expressing the perturbations in the entries of larger size, independently of their location, in terms of the transformation expressed in (4.1). We point out that in the general case, F in (4.2) does not need to be a diagonal matrix. Again, we start with a numerical example to motivate the general procedure that will be proposed in the next section.

2. Consider the matrices

$$T_1 = \begin{bmatrix} 1 & 10^5 & 0 \\ 10^5 & 10^5 & 10^5 \\ 0 & 10^5 & 1 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 10^5 & 10^5 & 0 \\ 10^5 & 10^5 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The approximations for the eigenvalues of T_1 and T_2, computed with MATLAB, are

$$\begin{aligned} \lambda_1(T_1) &= -9.999933333407408e{+}004, \\ \lambda_2(T_1) &= 1.000000000014616e{+}000, \\ \lambda_3(T_1) &= 2.000003333340741e{+}005 \end{aligned}$$

and

$$\begin{aligned} \lambda_1(T_2) &= -3.660259320914954e{-}001, \\ \lambda_2(T_2) &= 1.366023432085007e{+}000, \\ \lambda_3(T_2) &= 2.000000000025000e{+}005. \end{aligned}$$

In both cases, we know that the absolute errors in these approximations have a bound of size $O\left(10^{-11}\right)$ because the norm of the matrices is $O\left(10^5\right)$ and ϵ is $O\left(10^{-16}\right)$. To gain insight into the influence of perturbations, we used again the function eig of MATLAB to compute the eigenvalues of the matrices

$$\widetilde{T}_1 = \begin{bmatrix} 1(1+\eta_1) & 10^5(1+\delta_1) & 0 \\ 10^5(1+\delta_1) & 10^5(1+\eta_2) & 10^5(1+\delta_2) \\ 0 & 10^5(1+\delta_2) & 1(1+\eta_3) \end{bmatrix}$$

and

$$\widetilde{T}_2 = \begin{bmatrix} 10^5(1+\eta_1') & 10^5(1+\delta_1') & 0 \\ 10^5(1+\delta_1') & 10^5(1+\eta_2') & 1(1+\delta_2') \\ 0 & 1(1+\delta_2') & 1(1+\eta_3') \end{bmatrix}$$

with $\eta_k, \eta_k', \delta_k$, and δ_k' randomly generated, all bounded by $\varepsilon = 10^{-7}$ in absolute value. We got the errors

$$\begin{aligned} \lambda_1(T_1) - \lambda_1(\widetilde{T}_1) &\approx 3.1e{-}008, \\ \lambda_2(T_1) - \lambda_2(\widetilde{T}_1) &\approx -9.9e{-}009, \\ \lambda_3(T_1) - \lambda_3(\widetilde{T}_1) &\approx 8.9e{-}010 \end{aligned}$$

and

$$\begin{aligned} \lambda_1(T_2) - \lambda_1(\widetilde{T}_2) &\approx -7.1e{-}003, \\ \lambda_2(T_2) - \lambda_2(\widetilde{T}_2) &\approx 5.1e{-}004, \\ \lambda_3(T_2) - \lambda_3(\widetilde{T}_2) &\approx -1.4e{-}008, \end{aligned}$$

which are clearly due to the perturbations, not to the numerical errors in the function eig. We see that the eigenvalues of $\widetilde{T}_1$ exhibit absolute errors much smaller than $\|T_1\|_2\,\varepsilon \approx 2 \times 10^{-2}$, which do correspond to relative errors smaller than $\varepsilon = 10^{-7}$, but the error in $\lambda_1(\widetilde{T}_2)$ is close to $\|T_2\|_2\,\varepsilon \approx 2 \times 10^{-2}$. Why does T_1 define well its eigenvalues? First, we note that T_1 is not ; therefore [1, Theorem 4] does not apply. Furthermore, we computed the polar factor H of $T1 = T_1$, in MATLAB, from $[V, D] = \mathrm{eig}(T1); H = V * \mathrm{abs}(D) * V'$, and observed that the results of [22, section 2.8] are also unable to explain the good results obtained for T_1. Now, take

$$X = \begin{bmatrix} (1+\delta_1)^{-1}(1+\eta_2)^{1/2} & & \\ & (1+\eta_2)^{-1/2} & \\ & & (1+\delta_2)^{-1}(1+\eta_2)^{1/2} \end{bmatrix}$$

and verify that for $\widehat{T}_1 := X^T \widetilde{T}_1 X$ we get

$$\widehat{T}_1 = \begin{bmatrix} (1+\eta_1)(1+\delta_1)^{-2}(1+\eta_2) & 10^5 & 0 \\ 10^5 & 10^5 & 10^5 \\ 0 & 10^5 & (1+\eta_3)(1+\delta_2)^{-2}(1+\eta_2) \end{bmatrix}.$$

As in the example given in section 2, we have managed to produce a matrix $\widehat{T}_1$ with no perturbations in the entries of larger size and, as a consequence, we have $\widehat{T}_1 = T_1 + F$ with $\|F\|_2$ much smaller than $||T_1 - \widetilde{T}_1||_2$; furthermore, since X is close to the identity matrix, the relative Weyl's theorem guarantees that the eigenvalues of $\widehat{T}_1$ and $\widetilde{T}_1$ are close. The situation is quite different with T_2 because it is not possible to express the perturbations in all the larger entries $T_2(1,1)$, $T_2(2,2)$, $T_2(1,2)$, and $T_2(2,1)$ in terms of a multiplicative perturbation $X^T \widetilde{T}_2 X$, with some X close to the identity matrix.

5. A fast procedure to compute the error bound. In general, given a symmetric tridiagonal T with entries of different magnitudes and small relative perturbations, as expressed in $\widetilde{T}$ given in (3.2), we want to find a diagonal matrix X, with entries very close to the unity, such that the relations (4.1) and (4.2) hold, with $\|F\|_2$ as small as possible.

The example in the previous section shows that the rate of success of the procedure depends upon the locations of the entries of larger magnitude relatively to each other. Since our goal is to minimize, as much as possible, the size of the perturbed entries in $\widehat{T}$, we start by producing a sequence of $2n-1$ numbers, sorting the entries of T by decreasing order of their absolute values and "clean" as many entries as possible in this sequence. To simplify the presentation, we say that we clean the entry (i,j) when, in the course of the transformation (4.1), we get $\widehat{T}(i,j) = T(i,j)$, getting rid of the perturbation in $\widetilde{T}(i,j)$. In practice, we do not carry out such an operation, we just need to assume that it has been done. (This is in fact a combinatorial task and does not require any arithmetic at all.) By "operation of index k," $k = 1, \ldots, n$, we will mean the transformation that multiplies the kth row and the kth column of $\widetilde{T}$, i.e., the diagonal congruence associated with $X(k,k)$ in (4.1). We illustrate the cleaning procedure with the following example.

3. Suppose that our matrix T, of order $n = 5$, is such that

$$|a_1| \geq |b_3| \geq |b_2| \geq |a_4| \geq |b_4| \geq |a_2| \geq |b_1| \geq |a_5| \geq |a_3|\,. \tag{5.1}$$

First, we remove the perturbation from $a_1\,(1+\eta_1)$, the entry of largest size, by setting $X(1,1) = (1+\eta_1)^{-1/2}$; because we want a_1 to remain unperturbed, we close the index 1; i.e., it is removed from the set of indices allowed for subsequent operations. Next, to clean $b_3\,(1+\delta_3)$, we have two options: an operation of index 3 or an operation of index 4. Note that after cleaning $b_3\,(1+\delta_3)$ the indices 3 and 4 will be closed; therefore, before cleaning $b_3\,(1+\delta_3)$, we clean $a_4\,(1+\eta_4)$, since $|a_4| \geq |a_3|$, and close index 4. Then, we clean $b_3\,(1+\delta_3)\,(1+\eta_4)^{-1/2}$ and close index 3. The next entry in (5.1) is b_2, and the set of indices still open is $\{2,5\}$. So, we clean $b_2\,(1+\delta_2)\,(1+\delta_3)^{-1}\,(1+\eta_4)^{1/2}$. At this point, it is still possible to clean $b_4\,(1+\delta_4)\,(1+\eta_4)^{-1/2}$, which is next in (5.1), and this is the last entry to be cleaned. To summarize, with X diagonal such that

$$\begin{aligned}
X(1,1) &= (1+\eta_1)^{-1/2}\,,\\
X(2,2) &= (1+\delta_2)^{-1}\,(1+\delta_3)\,(1+\eta_4)^{-1/2}\,,\\
X(3,3) &= (1+\delta_3)^{-1}\,(1+\eta_4)^{1/2}\,,\\
X(4,4) &= (1+\eta_4)^{-1/2}\,,\\
X(5,5) &= (1+\delta_4)^{-1}\,(1+\eta_4)^{1/2}\,,
\end{aligned}$$

we get the following entries for $\widehat{T} = X^T\widetilde{T}X$:

$$\begin{aligned}
&\widehat{a}_1 = a_1, \quad \widehat{a}_4 = a_4, \quad \widehat{b}_2 = b_2, \quad \widehat{b}_3 = b_3, \quad \widehat{b}_4 = b_4,\\
&\widehat{a}_2 = a_2\,(1+\eta_2)\,(1+\delta_2)^{-2}\,(1+\delta_3)^2\,(1+\eta_4)^{-1}\,,\\
&\widehat{a}_3 = a_3\,(1+\eta_3)\,(1+\delta_3)^{-2}\,(1+\eta_4)\,,\\
&\widehat{a}_5 = a_5\,(1+\eta_5)\,(1+\delta_4)^{-1}\,(1+\eta_4)^{1/2}\,,\\
&\widehat{b}_1 = b_1\,(1+\delta_1)\,(1+\eta_1)^{-1/2}\,(1+\delta_2)^{-1}\,(1+\delta_3)\,(1+\eta_4)^{-1/2}\,.
\end{aligned}$$

Therefore, we may write $\widehat{T} = T + F$ with

$$F = \begin{bmatrix} 0 & b_1\delta_1' & & & \\ b_1\delta_1' & a_2\eta_2' & 0 & & \\ & 0 & a_3\eta_3' & 0 & \\ & & 0 & 0 & 0 \\ & & & 0 & a_5\eta_5' \end{bmatrix},$$

where δ_1', η_2', η_3', and η_5' are all of magnitude $O(\varepsilon)$ and the null entries do correspond to those positions that have been cleaned. In our example, if $|a_1| \gg |a_2|$ (remember that a_2 is the entry of largest size that has not been possible to clean), then $||T-\widetilde{T}||_2$ is much larger than $\|F\|_2$ and the bound (1.2) will be much sharper than the bound (1.3) for the eigenvalues of size significantly smaller than $\|T\|_2$. The gain, in terms of the sharpness of the bound that we get for the absolute errors in the eigenvalues, depends roughly on how large the ratio $|a_1|\,/\,|a_2|$ is.

We should remark that the described procedure is not optimal for symmetric tridiagonal matrices whose entries satisfy the condition $\max|a_j| < \min|b_j|$. In fact, by closing indices $1,\ldots,n$, in this ordering, we may clean all off-diagonal elements, as we did in Theorem 3.1; however, the procedure, as presented before, will clean first the off-diagonal entries of larger size and will not allow, in general, all off-diagonal entries to be cleaned. There are other cases for which our cleaning algorithm is not optimal and where it may be possible to use combinatorial analysis to improve the technique.

It should be noted that it is not possible to clean every entry of a submatrix

$$\begin{bmatrix} a_j & b_j \\ b_j & a_{j+1} \end{bmatrix}$$

for any $j = 1, \ldots, n-1$. Therefore, the error bound in (1.2) will never be smaller than $M \cdot O(\varepsilon)$, where

$$M = \max_{1 \le j \le n-1} \min \left\{ |a_j| \, , \, |b_j| \, , \, |a_{j+1}| \right\} .$$

In particular, for the matrix T_2 in the example given in section 4, we have $M = 10^5$.

We finish this section by noting that our procedure can be readily adapted for general symmetric matrices A to clean up to n entries. As for the tridiagonal case, after ordering the nonzero entries of A, in decreasing absolute values, we clean as many entries as possible in this sequence. To clean a pair of off-diagonal entries, say, $A(i,j)$ and $A(j,i)$, there may be a choice for the index to use (if both i and j are open). Because after cleaning $A(i,j)$ and $A(j,i)$, both indices will be closed, we may, similarly to the procedure that we have used in the tridiagonal case, clean first $A(i,i)$ or $A(j,j)$, the one of larger absolute value. A better solution may consist in looking at the size of the remaining entries in the ith and jth columns (or rows) and trying to clean the one of bigger size. Let this be the pair $A(i,p)$ and $A(p,i)$ for some $p \neq i$ and $p \neq j$. If the index p is already close, then it is certainly a good decision to clean entries $A(i,p)$ and $A(p,i)$ before cleaning entries $A(i,j)$ and $A(j,i)$. However, if index p is still open, the cleaning of the entries $A(i,p)$ and $A(p,i)$ closes p, and this may prevent the eventual cleaning of a bigger entry in the pth column (row). For this reason, it appears to be sensible to clean the pair $A(i,q)$ and $A(q,i)$ such that

$$|A(i,q)| = |A(q,i)| = \max_{r \in C} \left\{ |A(r,i)| \, , \, |A(r,j)| \right\} ,$$

where , denotes the set of indices which are already closed at this point. As it happens with the tridiagonal case, we cannot claim that this always produces the best possible X. Nevertheless, this procedure is very fast and may improve significantly the error bounds for the eigenvalues.

6. Accurate computation of the pivots. Using the very same idea of combining additive perturbations with multiplicative perturbations, we now show that the numerical values of the pivots of a symmetric tridiagonal matrix, computed through the formulae (6.1), may be used to determine eigenvalues with high relative accuracy. This may be of interest in the practical development of a parallel implementation of an algorithm which combines bisection with a faster zerofinder. Even in the context of sequential processing, there may still be room for new codes to take advantage of special features of matrices like those exploited in this paper. For instance, the state-of-the-art dqds algorithm, described in [40] and now implemented in the DSTEMR routine of the latest release of LAPACK, cannot guarantee high relative accuracy for the eigenvalues of symmetric tridiagonal indefinite matrices that define well their eigenvalues. In such cases, the only LAPACK routine that warrants full precision is DSTEBZ which implements the bisection method.

For a matrix T as given in Theorem 3.1, bisection (and related methods) is based upon the decomposition $T - \lambda I = LDL^T$, where L is unit lower bidiagonal and $D = \mathrm{diag}(q_1, \ldots, q_n)$ is diagonal. The numbers q_k are computed through

$$\begin{aligned} & q_1(\lambda) = a_1 - \lambda, \\ (6.1) \qquad & q_k(\lambda) = a_k - \lambda - b_{k-1}^2 / q_{k-1}(\lambda) \, , \qquad k = 2, \ldots, n. \end{aligned}$$

For each λ, the inertia of $T - \lambda I$, which is given by the signs of the $q_k(\lambda)$, can be used to locate eigenvalues. It is well known (see [25, p. 35] and [9, p. 230]) that the bisection method is able to compute the eigenvalues of a symmetric matrix which is very close to the exact one. In fact, the values $q_k(\lambda)$ computed with (6.1) in floating point arithmetic have the same signs as the values $\widetilde{q}_k(\lambda)$ that would be obtained if exact arithmetic was carried out with the matrix $\widetilde{T}$ such that[3]

$$\begin{aligned} &\widetilde{a}_k = a_k, \\ &\widetilde{b}_k = b_k\left(1+\delta_k\right), \text{ where } \left|\delta_k\right| \leq 2.5\epsilon + O\left(\epsilon^2\right). \end{aligned} \tag{6.2}$$

However, if one is to use not only the signs but also the numerical values of $q_k(\lambda)$, in the context of a method with a faster convergence rate, the previous result does not apply because it does not guarantee that the computed values of $q_k(\lambda)$ do correspond to a matrix $\widetilde{T}$ with entries satisfying the relations (6.2). In the context of the computation of singular values of bidiagonal matrices with relative accuracy, Demmel and Kahan [4, p. 24] briefly mentioned the possible use of zerofinders, different from simple bisection, to refine intervals; however, no details were given on the accuracy of the computed values $q_k(\lambda)$.

It is not true, in general, that the computed pivots are the exact ones for a matrix with small relative changes in its entries. However, an analysis similar to that used by Wilkinson for the leading principal minors (see [44, p. 303]) allows us to show that the computed values $q_k(\lambda)$ are the exact ones corresponding to off-diagonal entries with small relative perturbations and diagonal entries with additive perturbations of size $(a_k - \lambda)\, O(\epsilon)$. Writing the perturbed diagonal entries in the form

$$\widetilde{a}_k = a_k\left(1 + O\left(\epsilon\right)\right) - \lambda O\left(\epsilon\right),$$

we see that the computed $q_k(\lambda)$ do correspond to a matrix with small relative perturbations in its entries plus a diagonal additive perturbation of size $|\lambda|\, O(\epsilon)$. More precisely, we have the next theorem.

THEOREM 6.1. [illegible] T [illegible] (3.1) [illegible] λ [illegible] $q_1(\lambda), \ldots, q_n(\lambda)$ [illegible] (6.1) [illegible] $\widetilde{a}_k = a_k(1+\eta_k) - \lambda\eta_k$ [illegible] ff [illegible] $\widetilde{b}_{k-1} = b_{k-1}(1+\delta_{k-1})$ [illegible]

$$\left\{ \begin{array}{l} |\eta_k| \leq 2.02\epsilon, \\ |\delta_{k-1}| \leq 3.03\epsilon. \end{array} \right. \tag{6.3}$$

[illegible]. The proof is by induction. The result is obviously true for $k = 1$. Let us assume that the computed

$$\widetilde{q}_1\left(\lambda\right), \ldots, \widetilde{q}_{r-1}\left(\lambda\right)$$

are exact for a matrix having modified elements up to $\widetilde{a}_{r-1}$ and $\widetilde{b}_{r-2}$ and then show that the computed $\widetilde{q}_r\left(\lambda\right)$ is the exact value for a matrix having those modified elements and also the elements $\widetilde{a}_r$ and $\widetilde{b}_{r-1}$. If we assume that $\widetilde{q}_{r-1}\left(\lambda\right) \neq 0$ and represent by ε_1, ε_2, ε_3, and ε_4 the individual errors in the four operations involved in (6.1), we get

[3] In [6], it is shown that a similar result holds for symmetric matrices with acyclic graphs.

for the computed value of $q_r(\lambda)$

$$\begin{aligned} \widetilde{q}_r(\lambda) &= \left[(a_r - \lambda)(1+\varepsilon_1) - \frac{b_{r-1}^2 (1+\varepsilon_2)}{\widetilde{q}_{r-1}(\lambda)} (1+\varepsilon_3) \right] (1+\varepsilon_4) \\ &= a_r(1+\eta_r) - \lambda(1+\eta_r) - \frac{b_{r-1}^2 (1+\delta_{r-1})}{\widetilde{q}_{r-1}(\lambda)}, \end{aligned} \tag{6.4}$$

where $\eta_r = (1+\varepsilon_1)(1+\varepsilon_4) - 1$ and $\delta_{r-1} = (1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_4) - 1$, so that we get

$$\left\{ \begin{array}{l} |\eta_r| \le 2.02\epsilon, \\ |\delta_{r-1}| \le 3.03\epsilon. \end{array} \right. \tag{6.5}$$

Now, we may get $\widetilde{q}_{r-1}(\lambda) = 0$. If the arithmetic can handle the division by zero, as IEEE arithmetic does, then it gives $\widetilde{q}_r(\lambda) = -\infty$, independently of the value of $b_{r-1} \neq 0$, and we can write, in this case,

$$\eta_r = \delta_{r-1} = 0, \tag{6.6}$$

which, of course, satisfy the bounds (6.3). Furthermore, with $\widetilde{q}_r(\lambda) = -\infty$ in (6.1), we get that

$$\widetilde{q}_{r+1}(\lambda) = (a_{r+1} - \lambda)(1+\eta_{r+1})$$

does not depend upon the value of b_r and we can write

$$|\eta_{r+1}| \le \epsilon, \ \delta_r = 0. \tag{6.7}$$

In case the arithmetic in use does not handle the division by zero, we may replace $\widetilde{q}_{r-1}(\lambda) = 0$ with $\widetilde{q}_{r-1}(\lambda) = a_{r-1}\epsilon$ since this corresponds to perturbing a_{r-1} to $a_{r-1}(1+\epsilon)$, in (6.1), for $k = r-1$. □

So, for a given λ, the computed $\widetilde{q}_k(\lambda)$ do correspond to a matrix

$$\widetilde{T} = \widehat{T} + D, \tag{6.8}$$

where $\widehat{T}$ differs from T by small relative perturbations in its (diagonal and off-diagonal) entries and D is a diagonal matrix with entries of size bounded by $2.02\epsilon|\lambda|$. Therefore, if T defines well its eigenvalues so that for $\lambda_k \neq 0$ and some small constant γ, we may write, denoting by $\widehat{\lambda}_k$ the eigenvalues of $\widehat{T}$,

$$\left| \lambda_k - \widehat{\lambda}_k \right| \le \gamma\epsilon |\lambda_k|,$$

we get, denoting by $\widetilde{\lambda}_k$ the eigenvalues of $\widetilde{T}$ and taking into account that $\|D\| \le 2.02\epsilon|\lambda|$,

$$\left| \lambda_k - \widetilde{\lambda}_k \right| \le \gamma\epsilon |\lambda_k| + 2.02\epsilon|\lambda|$$

or

$$\left| \lambda_k - \widetilde{\lambda}_k \right| \le \left(\gamma + 2.02 \frac{|\lambda|}{|\lambda_k|} \right) \epsilon |\lambda_k|, \tag{6.9}$$

which shows that the relative error in $\widetilde{\lambda}_k$ is small whenever the ratio $\frac{|\lambda|}{|\lambda_k|}$ is not large. In practice, the bisection method, based upon the inertia of $T - \lambda I$, is used until we have a good approximation for the target eigenvalue; therefore, if one starts using the numerical values $\widetilde{q}_k(\lambda)$ only when a few significant digits are correct, we have that $\frac{|\lambda|}{|\lambda_k|} \approx 1$ and, in this case, the bound in (6.9) guarantees small relative errors. Note that from the point of view of convergence speed, it is premature to switch from bisection to a method with a better asymptotic rate of convergence before we have an approximation with a few correct digits anyway. So we claim that the numerical values of the pivots may be used to compute the eigenvalues with high relative accuracy whenever T defines them well.

7. Toward a mixed precision bisection algorithm. Another practical application that we envisage for our results is a mixed precision bisection algorithm. Processors are arriving on the market that are much faster for single precision floating point operations than for double precision arithmetic. Examples include the Intel Pentium IV and M processors, AMD's Opteron architectures, and the IBM Cell Broad Engine processor. When working in single precision, floating point operations can be performed up to two times faster on the Pentium and up to 10 times faster on the Cell than for double precision [31]. This technological change is likely to have a significant impact in the design of many numerical algorithms. Some work has already been carried out in the context of iterative refinement for linear systems (see [2], [30], [31]).

In an implementation of the bisection method, tailored for such processors, single precision arithmetic may be used to deliver intervals that are refined using double precision arithmetic. Because each interval produced in single precision is not guaranteed to contain the desired eigenvalue (unless some form of interval arithmetic is implemented), it cannot be accepted blindly and may need to be corrected in double precision.

Now, a critical issue is to decide when to switch from single to double precision. If we switch too soon, we will be using expensive double precision arithmetic that could have been carried out in the single format; on the other hand, if we go too far in single precision, an incorrect interval will be produced and we pay a penalty for correcting the interval. It is for this reason that a good stopping criterion for the single precision phase is much more important than a stopping criterion in the usual situation where double precision is used from the very beginning.

For a matrix T with diagonal elements a_j of size much smaller than $\|T\|$, we may, taking the relation (3.3) into account, switch from single to double precision immediately after locating an eigenvalue in the interval $[y, z]$ such that

$$z - y \leq O(\epsilon_s) \max |a_j| , \tag{7.1}$$

where ϵ_s denotes the single precision roundoff error unit. More generally, for a matrix with entries of different magnitudes, we may use the procedure described in section 5 to compute the largest size M of the entries that cannot be cleaned and replace $\max|a_j|$ with M in (7.1). For matrices, this does not provide a good stopping criteria; therefore, a different test would be required in conjunction with the one proposed here.

8. Conclusions and further work. We have combined well-known results of the perturbation theory to derive new error bounds for the eigenvalues of symmetric tridiagonal matrices. Our bounds are sharper than the usual bounds in the case of certain matrices with entries and eigenvalues of varying size. As an application of this idea, we have shown that a symmetric tridiagonal matrix T, with diagonal entries a_j, defines well the eigenvalues whose magnitude is not much smaller than $\max|a_j|$. This can be understood as a generalization of the well-known fact that a Golub–Kahan matrix defines well all its eigenvalues. As a practical application of our perturbation technique, we have shown that the numerical values and not only the signs of the pivots, computed in the usual way, may be used to find, with high relative accuracy, those eigenvalues which are well defined. Also, we have briefly considered a mixed precision bisection algorithm and have shown that our perturbation technique may help in the critical issue of determining when to switch from single to double precision. We are currently working in this line of research.

Acknowledgments. The author is indebted to Beresford Parlett, Zlatko Drmač, and the anonymous referees for several criticisms and suggestions that improved the paper significantly. The author also expresses his gratitude to Froilán Dopico for fruitful discussions and acknowledges that he proved Theorem 3.1 independently.

REFERENCES

[1] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[2] A. Buttari, J. Dongarra, J. Kurzak, P. Luszcez, and S. Tomov, *Using Mixed Precision for Sparse Matrix Computations to Enhance the Performance while Achieving* 64-*Bit Accuracy*, LAPACK Working Note 180, 2006.

[3] P. Deift, J. Demmel, L.-C. Li, and C. Tomei, *The bidiagonal singular value decomposition and Hamiltonian mechanics*, SIAM J. Numer. Anal., 28 (1991), pp. 1463–1516.

[4] J. Demmel and W. Kahan, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[5] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[6] J. Demmel and W. Gragg, *On computing accurate singular values and eigenvalues of acyclic matrices*, Linear Algebra Appl., 185 (1993), pp. 203–218.

[7] J. Demmel, *The Inherent Inaccuracy of Implicit Tridiagonal QR*, LAPACK Working Note 45, 1992.

[8] J. Demmel, I. Dhillon, and H. Ren, *On the correctness of some bisection-like parallel eigenvalue algorithms in floating point arithmetic*, Electronic Trans. Numer. Anal., 3 (1995), pp. 116–149.

[9] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[10] J. Demmel, *Accurate SVDs of structured matrices*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 562–508.

[11] J. Demmel, M. Gu, S. Eisenstat, I. Slapničar, K. Veselić, and Z. Drmač, *Computing the singular value decomposition with high relative accuracy*, Linear Algebra Appl., 299 (1999), pp. 21–80.

[12] J. Demmel and P. Koev, *Accurate SVDs of weakly diagonally dominant M-matrices*, Numer. Math., 98 (2004), pp. 99–104.

[13] I. Dhillon and B. Parlett, *Multiple representations to compute orthogonal eigenvectors of symmetric matrices*, Linear Algebra Appl., 387 (2004), pp. 1–28.

[14] I. Dhillon and B. Parlett, *Orthogonal eigenvectors and relative gaps*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 858–899.

[15] F. Dopico and P. Koev, *Accurate symmetric rank revealing and eigendecompositions of symmetric structured matrices*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1126–1156.

[16] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–88.

[17] S. Eisenstat and I. Ipsen, *Absolute Perturbation Bounds for Matrix Eigenvalues Imply Relative Bounds*, Technical report CRSC-TR97-16, Center for Research in Scientific Computation, Department of Mathematics, North Carolina State University, Raleigh, NC, 1997.

[18] V. Fernando and B. Parlett, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.

[19] B. Grosser and B. Lang, *An $O(n^2)$ algorithm for the bidiagonal SVD*, Linear Algebra Appl., 358 (2003), pp. 45–70.

[20] B. Grosser and B. Lang, *On symmetric eigenproblems induced by the bidiagonal SVD*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 599–620.

[21] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[22] I. Ipsen, *Relative perturbation bounds for matrix eigenvalues and singular values*, in Acta Numerica, 1998, Acta Numer. 7, Cambridge University Press, Cambridge, UK, 1998, pp. 151–201.

[23] I. Ipsen, *A note on unifying absolute and relative perturbation bounds*, Linear Algebra Appl., 358 (2003), pp. 239–53.

[24] I. C. F. Ipsen and B. Nadler, *Refined perturbation bounds for eigenvalues of Hermitian and non-Hermitian matrices*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 40–53.

[25] W. Kahan, *Accurate Eigenvalues of a Symmetric Tridiagonal Matrix*, Technical Report CS41, Computer Science Department, Stanford University, Palo Alto, CA, 1966.

[26] P. Koev, *Accurate and Efficient Computations with Structured Matrices*, Ph.D. thesis, University of California, Berkeley, CA, 2002.

[27] P. Koev, *Accurate eigenvalues and SVDs of totally nonnegative matrices*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 1–23.

[28] P. Koev and F. Dopico, *Accurate eigenvalues of certain sign regular matrices*, Linear Algebra Appl., 424 (2007), pp. 435–447.

[29] P. Koev, *Accurate computations with totally nonnegative matrices*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 731–751.

[30] J. Kurkzak and J. Dongarra, *Implementation of a Mixed-Precision High Performance LINPACK Benchmark on the Cell Processor*, LAPACK Working Note 177, 2006.

[31] J. Langou, J. Langou, P. Luszczek, J. Kurzak, A. Buttari, and J. Dongarra, *Exploiting the Performance of* 32 *Bit Floating Point Arithmetic in Obtaining* 64 *Bit Accuracy*, LAPACK Working Note 175, 2006.

[32] R. Li, *Relative Perturbation Theory:* (I) *Eigenvalue Variations*, LAPACK Working Note 84, 1994.

[33] C. Li and R. Mathias, *On the Lidskii-Mirsky-Wielandt Theorem*, Technical report, Department of Mathematics, College of William and Mary, Williamsburg, VA, 1997.

[34] R. Mathias, *Spectral Perturbation Bounds for Positive Definite Matrices*, Technical report, Department of Mathematics, College of William and Mary, Williamsburg, VA, 1994.

[35] R. Mathias, *Accurate eigensystem computations by Jacobi methods*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 977–1003.

[36] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, New York, 1980.

[37] B. Parlett and B. Nour-Omid, *The use of a refined error bound when updating eigenvalues of tridiagonals*, Linear Algebra Appl., 68 (1985), pp. 179–219.

[38] B. Parlett, *The new qd algorithms*, in Acta Numerica, 1995, Acta Numer. 4, Cambridge University Press, Cambridge, UK, 1995, pp. 459–491.

[39] B. Parlett and I. Dhillon, *Relatively robust representations of symmetric tridiagonals*, Linear Algebra Appl., 309 (2000), pp. 121–151.

[40] B. Parlett and O. Marques, *An implementation of the dqds algorithm (positive case)*, Linear Algebra Appl., 309 (2000), pp. 217–259.

[41] S. Singer and S. Singer, *Skew-symmetric differential qd algorithm*, Appl. Numer. Anal. Comput. Math., 2 (2005), pp. 134–151.

[42] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195, pp. 81–116.

[43] P. Willems, B. Lang, and C. Vomel, *Computing the Bidiagonal SVD Using Multiple Relatively Robust Representations*, LAPACK Working Note 166, 2005.

[44] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 1, pp. 92–113

INEXACT INVERSE SUBSPACE ITERATION WITH PRECONDITIONING APPLIED TO NON-HERMITIAN EIGENVALUE PROBLEMS*

MICKAËL ROBBÉ†, MILOUD SADKANE†, AND ALASTAIR SPENCE‡

Abstract. Convergence results are provided for inexact inverse subspace iteration applied to the problem of finding the invariant subspace associated with a small number of eigenvalues of a large sparse matrix. These results are illustrated by the use of block-GMRES as the iterative solver. The costs of the inexact solves are measured by the number of inner iterations needed by the iterative solver at each outer step of the algorithm. It is shown that for a decreasing tolerance the number of inner iterations should not increase as the outer iteration proceeds, but it may increase for preconditioned iterative solves. However, it is also shown that an appropriate small rank change to the preconditioner can produce significant savings in costs and, in particular, can produce a situation where there is no increase in the costs of the iterative solves even though the solve tolerances are reducing. Numerical examples are provided to illustrate the theory.

Key words. eigenvalue approximation, inverse subspace iteration, iterative methods, preconditioning

AMS subject classifications. 65F10, 65F15

DOI. 10.1137/060673795

1. Introduction.

Inverse subspace iteration is a block version of the inverse iteration. It computes an approximation of the invariant subspace of a large matrix $A \in \mathbb{C}^{n\times n}$ corresponding to the eigenvalues in an isolated cluster around a given shift σ. The corresponding algorithm is very simple and can formally be written as

$$X_i = (A - \sigma I)^{-1} X_{i-1}, \ \ i = 1, 2, \dots, \tag{1.1}$$

where $X_0 \in \mathbb{C}^{n\times p}$ is full rank with $p \ll n$. As the iterations unfold, the invariant subspace and hence the eigenvectors corresponding to eigenvalues near σ eventually dominate X_i. The method is known to be reliable [17, 26, 20, 27] and, although its convergence is linear, only a few iterations are needed to converge, provided that the target eigenvalues lie in a cluster well separated from the rest of the spectrum and p is chosen as large as the number of eigenvalues in the cluster. The drawback of this method is that each iteration necessitates the exact solution of a block linear system, that is, a linear system with multiple right-hand sides of the form

$$(A - \sigma I)Y = X, \quad Y, X \in \mathbb{C}^{n\times p}, \tag{1.2}$$

which is a challenge when n is large. The first aim of this paper is to analyze the convergence of (1.1) when the underlying block linear systems (1.2) are solved inexactly by an iterative method. The method obtained this way belongs to the wide class of "inner-outer" iterative methods. The outer iteration is the inverse subspace iteration and the inner iteration is the iterative solution of the block linear system (1.2). The

*Received by the editors October 31, 2006; accepted for publication (in revised form) by Z. Drmač March 10, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/simax/31-1/67379.html

†Département de Mathématiques, Université de Bretagne Occidentale, 6, Av. Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France (robbe@univ-brest.fr, sadkane@univ-brest.fr).

‡Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom (a.spence@maths.bath.ac.uk).

results in this paper extend the results in [14, 12, 3, 4] on inexact inverse iteration to inexact inverse subspace iteration.

The second aim of this paper is to discuss the performance of unpreconditioned and preconditioned block-GMRES as the inexact solver. If P denotes a preconditioner for $(A - \sigma I)$, the (right) preconditioned form of (1.2) is

$$(A - \sigma I)P^{-1}\widetilde{Y} = X, \quad Y = P^{-1}\widetilde{Y}, \tag{1.3}$$

with the aim that (1.3) is solved more efficiently than (1.2). For inexact inverse iteration, [3, 5] consider the costs of the inner solves for Krylov solvers and analyze cases where the number of inner iterations may remain approximately constant or may increase as the outer iteration proceeds. In this paper we extend these results to the block case. Moreover, we show how a rank-p modification of P gives a "tuned" preconditioner which eliminates the increase in the number of inner iterations as the outer iteration proceeds.

Recently, inexact inverse iteration has been discussed by [14, 12, 4], and for the symmetric case by [25, 3]. The idea of tuning the preconditioner for eigenvalue problems was introduced in [8] as a means of obtaining quadratic outer convergence in a variation of inexact inverse iteration with a certain variable shift, but no analysis of the inexact solver was presented. In [7] tuning the preconditioner for inexact inverse iteration applied to a Hermitian eigenvalue problem was analyzed, and certain specialized results, comparing a tuned with a standard preconditioner, were obtained by exploiting the Hermitian structure. The fact that, for a fixed shift and a variable tolerance, tuning removes the dependence on the number of inner iterations was also presented, though the analysis relied on the construction of an "ideal preconditioner," first introduced in an earlier version of this paper. Recent work in [30] contains a detailed convergence analysis of preconditioned MINRES as the solver in inexact Rayleigh quotient iteration which shows that tuning the preconditioner leads to a major reduction in inner iteration costs over the untuned case. There is considerable interest in inexact solves for subspace-based methods, especially in relation to the Jacobi–Davidson method (JD) [24, 2, 16] and the Riccati-based methods as developed in [18, 6]; the latter may be viewed as the block analogue of JD and are useful for computing invariant subspaces. JD is an important inner-outer iteration that improves the current approximate eigenvector through inexact solves and preconditioning. It exhibits good inner solve performance because of the gradient form of the right-hand side. The performance of preconditioned inner solves for inexact inverse iteration is often poorer than for JD, but we shall show that the tuning idea discussed here substantially improves the performance of the preconditioned iterative solver. In fact, [9] contains an equivalence result between preconditioned simplified JD (where the correction equation in JD is used without expanding the search space) and a special tuned inexact Raleigh quotient (IRQ) iteration when a Galerkin–Krylov solver is used. This extends a result of [23] and suggests that inverse iteration–based methods might become more competitive if tuned preconditioners are used. The link between IRQ and simplified JD has also been noted in [15], and another useful method which uses inexact solves within inner-outer iterations is the trace minimization method [22]. A method which uses preconditioned iterative solves on subspaces is the truncated-CG–based trust-region method [1], which is particularly successful for finding extremal eigenvalues of symmetric matrices and is also related to simplified JD [1].

In section 2 we present the inexact inverse subspace iteration algorithm and some preliminary results. In particular, we discuss some tools for measuring the closeness between subspaces. Section 3 presents a convergence theory for the inexact (and exact) inverse subspace iteration. We shall show that if these linear systems are solved to an appropriately chosen decreasing tolerance, then the method attains a linear rate of convergence just as in the case of exact solves (Theorem 3.1). In section 3.1 we consider the use of block-GMRES as the (unpreconditioned) solver. We show that for a decreasing tolerance the number of inner iterations should not increase as the outer iteration converges. The case of preconditioned solves is discussed in section 4. Our main result, presented in section 4.2, is that if a standard preconditioner is modified by a small rank change, then there is again no increase in the number of inner iterations as the outer iteration proceeds. We call the process of modifying the preconditioner in this way "tuning." In section 5, numerical tests are given to illustrate the theory. In particular, it is shown that significant savings are obtained when a tuned preconditioner is used.

This paper contains several extensions of the work in [14, 12, 3, 5, 7, 8]. First, there is the extension to the block case. Second, our convergence theory for inexact subspace iteration and for block-GMRES works under rather weak assumptions; for example, the matrix A can be defective. Third, the tuning theory, discussed in [7] for Hermitian problems, is extended to the non-Hermitian case, and strong supporting numerical evidence for its effectiveness is presented. Finally, the introduction of the "ideal" preconditioner in section 4 allows a complete and rigorous proof that the tuned preconditioner removes the dependence on the number of inner iterations. This construction, developed here first, was utilized in [7] and is a rather powerful theoretical tool. It is likely to have application in other related areas.

2. Inexact inverse subspace iteration.

In this section we describe the inexact inverse subspace iteration algorithm, and in section 2.1 revise some background material, especially relating to the angle between two subspaces.

In many applications interest centers on the invariant subspace corresponding to the eigenvalues nearest zero, and from now on we shall choose the shift in (1.2) to be zero. Much of what we say extends to the case of a nonzero fixed shift.

Inexact inverse subspace iteration is described in the following inner-outer algorithm.

Algorithm 1 (inexact subspace iteration).

```
$\delta \geq 0$   $X_0 \in \mathbb{C}^{n\times p}$   $X_0^* X_0 = I$
$i = 0, 1, \ldots$
  1. $L_i = X_i^* A X_i$
  2. $R_i = AX_i - X_i L_i$
  3. $AY_i = X_i$   $Y_i$

        $X_i - AY_i = \Delta_i$  with  $\|\Delta_i\| \leq \tau_i = \delta \|R_i\|,$

  4. $Y_i$   $X_{i+1}$
$i$
```

In Algorithm 1 and throughout this paper, the symbol $\| \; \|$ denotes the Euclidean norm or its induced matrix norm.

In section 3 we first analyze the convergence of Algorithm 1 with no particular solver in mind, and in section 3.1 we discuss the case when the block linear systems in step 3 of Algorithm 1 are solved by block-GMRES. Note that if the block systems are solved exactly, then the (exact) inverse subspace iteration (1.1) is recovered.

The next section gathers some technical details which will be used throughout this paper.

2.1. Notation and preliminaries. We assume that the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of A are such that

$$0 < |\lambda_1| \leq \cdots \leq |\lambda_p| \; < \; |\lambda_{p+1}| \leq \cdots \leq |\lambda_n|. \tag{2.1}$$

By Schur's theorem we may decompose the matrix A to upper triangular form by a unitary matrix $\begin{pmatrix} V_1 & V_1^\perp \end{pmatrix}$:

$$A = \begin{pmatrix} V_1 & V_1^\perp \end{pmatrix} \begin{pmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} V_1 & V_1^\perp \end{pmatrix}^*, \tag{2.2}$$

where $V_1 \in \mathbf{C}^{n\times p}$, $V_1^\perp \in \mathbf{C}^{n\times(n-p)}$, $T_{11} \in \mathbf{C}^{p\times p}$, and $T_{22} \in \mathbf{C}^{(n-p)\times(n-p)}$. The spectra of T_{11} and T_{22} are, respectively, $\lambda_1, \ldots, \lambda_p$ and $\lambda_{p+1}, \ldots, \lambda_n$. Let $Q \in \mathbf{C}^{p\times(n-p)}$ be the unique solution of the Sylvester equation

$$QT_{22} - T_{11}Q = T_{12}. \tag{2.3}$$

Then A can be block-diagonalized as follows (see, e.g., [11]):

$$A = \begin{pmatrix} V_1 & V_1^\perp \end{pmatrix} \begin{pmatrix} I & Q \\ 0 & I \end{pmatrix} \begin{pmatrix} T_{11} & 0 \\ 0 & T_{22} \end{pmatrix} \begin{pmatrix} I & -Q \\ 0 & I \end{pmatrix} \begin{pmatrix} V_1 & V_1^\perp \end{pmatrix}^*. \tag{2.4}$$

Let

$$V_2 = \left(V_1 Q + V_1^\perp\right)\left(I + Q^* Q\right)^{-\frac{1}{2}},$$

$$L = T_{11}, \quad M = \left(I + Q^*Q\right)^{-\frac{1}{2}} T_{22} \left(I + Q^*Q\right)^{\frac{1}{2}}.$$

Then the block-diagonalization in (2.4) can be written

$$A = \begin{pmatrix} V_1 & V_2 \end{pmatrix} \begin{pmatrix} L & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} V_1 & V_2 \end{pmatrix}^{-1}. \tag{2.5}$$

Note that M and T_{22} have the same spectra and that V_1 and V_2 have orthonormal columns. The subspaces $\mathcal{V}_1 = \mathcal{R}(V_1)$ and $\mathcal{V}_2 = \mathcal{R}(V_2)$ spanned by the columns of V_1 and V_2 are complementary invariant subspaces of A associated, respectively, with the eigenvalues $\lambda_1, \ldots, \lambda_p$ of L and $\lambda_{p+1}, \ldots, \lambda_n$ of M. Our main task in this paper is to compute the invariant subspace $\widetilde{\mathcal{V}}_1 \subset \mathcal{V}_1$ associated with the $q \leq p$ smallest (in modulus) eigenvalues of A by Algorithm 1.

The smallest (largest) singular value of a matrix B is denoted by $\sigma_{min}(B) = \min_{\|x\|=1} \|Bx\|$ ($\sigma_{max}(B) = \|B\| = \max_{\|x\|=1} \|Bx\|$). The separation $\operatorname{sep}(E, F)$ between two matrices $E \in \mathbb{C}^{p\times p}$ and $F \in \mathbb{C}^{q\times q}$ is defined as (see [28])

$$\operatorname{sep}(E, F) = \min_{\|X\|=1} \|EX - XF\|.$$

It is known that $\operatorname{sep}(E, F) > 0$ if and only if E and F have disjoint spectra. Our analysis will lead us to use either $\operatorname{sep}(T_{22}, L)$ or $\operatorname{sep}(M, L)$. These quantities are equivalent since (see [28])

$$\operatorname{sep}(T_{22}, L)/\kappa \leq \operatorname{sep}(M, L) \leq \kappa \operatorname{sep}(T_{22}, L),$$

where $\kappa = \sqrt{\frac{1+\sigma^2_{\max}(Q)}{1+\sigma^2_{\min}(Q)}}$, with Q defined by (2.3).

Let

$$W_1 = V_1 - V_1^{\perp} Q^* \quad \text{and} \quad W_2 = V_1^{\perp} \left(I + Q^*Q\right)^{\frac{1}{2}}. \tag{2.6}$$

Then $\mathcal{R}(W_1)$ and $\mathcal{R}(W_2)$ are complementary invariant subspaces of A^* corresponding to the eigenvalues $\bar{\lambda}_1, \ldots, \bar{\lambda}_p$ of L^* and $\bar{\lambda}_{p+1}, \ldots, \bar{\lambda}_n$ of M^*, and we have

$$\begin{pmatrix} W_1 & W_2 \end{pmatrix}^* = \begin{pmatrix} V_1 & V_2 \end{pmatrix}^{-1}. \tag{2.7}$$

The spectral projection on $\mathcal{V}_1$ is defined by

$$\mathcal{P} = V_1 W_1^*. \tag{2.8}$$

Note that

$$\|\mathcal{P}\| = \|W_1\| = \sqrt{1 + \|Q\|^2}. \tag{2.9}$$

To understand the performance of Algorithm 1 we need to measure the deviation of X_i from V_1. This can be done by monitoring the angle between the subspaces $\mathcal{V}_1$ and $\mathcal{X}_i = \mathcal{R}(X_i)$. One tool is the sine of the largest canonical angle between $\mathcal{V}_1$ and $\mathcal{X}_i$ defined by (see [11, p. 584])

$$\sin \angle(\mathcal{X}_i, \mathcal{V}_1) = \left\|(V_1^{\perp})^* X_i\right\|. \tag{2.10}$$

We assume that the subspaces $\mathcal{X}_i$ and $\mathcal{V}_1$ have the same dimension. Then (see [11, p. 76])

$$\begin{aligned} \sin \angle(\mathcal{X}_i, \mathcal{V}_1) &= \sin \angle(\mathcal{V}_1, \mathcal{X}_i) = \|X_i X_i^* - V_1 V_1^*\| && (2.11) \\ &= \min_{Z \in \mathbb{C}^{p \times p}} \|X_i - V_1 Z\| = \min_{Z \in \mathbb{C}^{p \times p}} \|V_1 - X_i Z\|. && (2.12) \end{aligned}$$

We also assume that the matrix X_i can be decomposed as

$$X_i = V_1 C_i + V_2 S_i \quad \text{with } \|S_i\| < 1. \tag{2.13}$$

Using (2.7), we see that the matrices C_i and S_i are given by

$$C_i = W_1^* X_i \in \mathbb{C}^{p \times p}, \quad S_i = W_2^* X_i \in \mathbb{C}^{(n-p) \times p}. \tag{2.14}$$

From (2.13), formula (2.10) becomes

$$\sin \angle(\mathcal{X}_i, \mathcal{V}_1) = \|(V_1^{\perp})^* V_2 S_i\|, \tag{2.15}$$

which shows that $\|S_i\|$ can also be used to measure the deviation between $\mathcal{V}_1$ and $\mathcal{X}_i$. In fact, we will cast our results in terms of the quantities

$$\sin \angle(\mathcal{X}_i, \mathcal{V}_1), \; t_i = \left\|S_i C_i^{-1}\right\| \quad \text{or} \quad s_i = t_i \left\|C_i\right\|.$$

Note that in the case when A is Hermitian, then t_i and $\|S_i\|$ represent, respectively, the tangent and the sine of the largest angle between $\mathcal{X}_i$ and $\mathcal{V}_i$. The following proposition shows that all these quantities behave like $\|S_i\|$. So $\mathcal{X}_i \to \mathcal{V}_1$ if and only if $t_i \to 0$, $s_i \to 0$, or $\|S_i\| \to 0$.

Proposition 2.1. [illegible] X_i [illegible] (2.13) [illegible]

(1) C_i [illegible] t_i [illegible] C_i [illegible]

$$0 < 1 - \|S_i\| \leq \sigma_k(C_i) \leq 1 + \|S_i\|, \; k = 1, \ldots, p, \tag{2.16}$$

[illegible] C_i [illegible]

$$C_i = U_i + \Upsilon_i, \tag{2.17}$$

[illegible] U_i [illegible] $\|\Upsilon_i\| \leq \|S_i\| < 1$

(2) $\sin \angle(\mathcal{X}_i, \mathcal{V}_1) \leq \|S_i\| \leq s_i \leq \|S_i\| \frac{1+\|S_i\|}{1-\|S_i\|}$

(3) $\sin \angle(\mathcal{X}_i, \mathcal{V}_1) \leq t_i \leq \frac{\|S_i\|}{1-\|S_i\|}$

(4) $\|S_i\| \leq \|\mathcal{P}\| \sin \angle(\mathcal{X}_i, \mathcal{V}_1)$

[illegible]

(1) Assume C_i is singular and let u be a nonzero vector such that $C_i u = 0$. Then

$$\|u\| = \|X_i u\| = \|V_2 S_i u\| \leq \|S_i\| \|u\| < \|u\|,$$

a contradiction. Hence C_i is nonsingular. The kth singular values of X_i and $V_1 C_i$ satisfy (see [11, p. 428])

$$|\sigma_k(X_i) - \sigma_k(V_1 C_i)| \leq \|X_i - V_1 C_i\| \leq \|S_i\|,$$

and hence

$$|1 - \sigma_k(C_i)| \leq \|S_i\|.$$

Let $C_i = W_i^{(l)} \Sigma_i W_i^{(r)}$ be the singular value decomposition of C_i. Then C_i can be written as in (2.17) with $U_i = W_i^{(l)} W_i^{(r)}$ and $\Upsilon_i = W_i^{(l)} \left(\Sigma_i - I\right) W_i^{(r)}$.

(2) The first bound follows from (2.15) and the other ones from the definition of s_i and (2.16).

(3)

$$\begin{aligned} \sin \angle(\mathcal{X}_i, \mathcal{V}_1) &= \|(X_i^{\perp})^* V_1\| = \|(X_i^{\perp})^* (V_1 - X_i C_i^{-1})\| \\ &\leq \|V_1 - X_i C_i^{-1}\| = \|V_2 S_i C_i^{-1}\| = t_i, \\ t_i &= \|S_i C_i^{-1}\| \leq \frac{\|S_i\|}{\sigma_{\min}(C_i)} \leq \frac{\|S_i\|}{1 - \|S_i\|}. \end{aligned}$$

(4)

$$\begin{aligned} \|S_i\| = \|W_2^* X_i\| &= \left\| (I + Q^* Q)^{\frac{1}{2}} \; (V_1^{\perp})^* X_i \right\| \\ &\leq \|\mathcal{P}\| \sin \angle(\mathcal{X}_i, \mathcal{V}_1). \qquad \Box \end{aligned}$$

The following proposition gives bounds on the residual norm.

PROPOSITION 2.2. [illegible]

$$\operatorname{sep}(T_{22}, L_i) \; \sin \angle(\mathcal{X}_i, \mathcal{V}_1) \leq \|R_i\| \leq ||\mathcal{S}|| s_i,$$

[illegible] $\mathcal{S}$ [illegible] $X \to \mathcal{S}(X) = MX - XL$ [illegible]

$$||\mathcal{S}|| = \sup_{\|X\|=1} \|\mathcal{S}(X)\|.$$

[illegible]

$$\begin{aligned}\|R_i\| &\ge \|(V_1^{\perp})^* R_i\| \\ &= \|(V_1^{\perp})^* A X_i - (V_1^{\perp})^* X_i L_i\|.\end{aligned}$$

From (2.2) we have $(V_1^{\perp})^* A = T_{22}(V_1^{\perp})^*$. Then

$$\|R_i\| \ge \|T_{22}(V_1^{\perp})^* X_i - (V_1^{\perp})^* X_i L_i\| \ge \operatorname{sep}(T_{22}, L_i)\ \|(V_1^{\perp})^* X_i\|.$$

Also,

$$\begin{aligned}\|R_i\| &= \min_{Z\in\mathbb{C}^{p\times p}} \|AX_i - X_i Z\| \quad (\text{see } [28, \text{Thm. } 1.15]) \\ &\le \|AX_i - X_i C_i^{-1} L C_i\| \\ &= \|MS_i - S_i C_i^{-1} L C_i\| \\ &= \|\left(MS_i C_i^{-1} - S_i C_i^{-1} L\right) C_i\| \le \|\mathcal{S}\|\ \|S_i C_i^{-1}\|\|C_i\|. \qquad \square\end{aligned}$$

3. Convergence analysis of Algorithm 1. In this section we analyze the convergence of Algorithm 1 when the inner iterations are solved inexactly. First, we make no assumption on the inexact solver except that step 3 in Algorithm 1 is satisfied. Then, in section 3.1, we assume that a block-GMRES method is the inexact solver.

THEOREM 3.1 (convergence of Algorithm 1). [illegible] X_0 [illegible] V_1 [illegible] $\tau_i = \delta\|R_i\|$ [illegible] $\delta < \left(\|(C_i^{-1}\|\|\mathcal{P}\|\|R_i\|\right)^{-1}$ [illegible]

$$t_{i+1} \le \|M^{-1}\|\|L\|\frac{t_i + \alpha_i\tau_i}{1-\alpha_i\tau_i}, \tag{3.1}$$

[illegible] $\alpha_i = \|C_i^{-1}\|\|\mathcal{P}\| \le \frac{\|\mathcal{P}\|}{1-\|S_i\|}$ [illegible] M [illegible] L [illegible] (2.5) [illegible] $\|M^{-1}\|\|L\| < 1$ [illegible] 1 [illegible]

[illegible] Note first that

$$t_{i+1} = \left\|S_{i+1}C_{i+1}^{-1}\right\| = \left\|(W_2^* X_{i+1})(W_1^* X_{i+1})^{-1}\right\| = \left\|(W_2^* X_{i+1}K)(W_1^* X_{i+1}K)^{-1}\right\|,$$

where $K \in \mathbb{C}^{p\times p}$ is an arbitrary nonsingular matrix. In particular $t_{i+1} = \|(W_2^* Y_i)(W_1^* Y_i)^{-1}\|$ and therefore

$$\begin{aligned}t_{i+1} &= \left\|W_2^* A^{-1}(X_i - \Delta_i)\left(W_1^* A^{-1}(X_i-\Delta_i)\right)^{-1}\right\| \\ &= \left\|M^{-1}W_2^*(X_i-\Delta_i)\left(L^{-1}W_1^*(X_i-\Delta_i)\right)^{-1}\right\| \\ &\le \|M^{-1}\|\|L\|\left\|W_2^*(X_i-\Delta_i)\left(W_1^*(X_i-\Delta_i)\right)^{-1}\right\| \\ &\le \|M^{-1}\|\|L\|\left\|W_2^*(X_i-\Delta_i)C_i^{-1}\left(I - W_1^*\Delta_i C_i^{-1}\right)^{-1}\right\| \\ &\le \|M^{-1}\|\|L\|\frac{t_i + \|W_2^*\Delta_i\|\|C_i^{-1}\|}{1-\|W_1^*\Delta_i\|\|C_i^{-1}\|}.\end{aligned}$$

Note that $\|W_2^*\Delta_i\| \le \|W_2\|\tau_i$. From (2.6) and (2.9) we have $\|W_2\| = \|\mathcal{P}\|$, and from Proposition 2.2, $\|C_i^{-1}\| \le 1/(1-\|S_i\|)$. The expression $\|W_1^*\Delta_i\|$ is bounded in a similar way, and (3.1) is obtained. Since $\tau_i \le \delta\|R_i\| \le \delta\|\mathcal{S}\|\|C_i\|t_i$, linear convergence is established for small δ and $\|M^{-1}\|\|L\| < 1$. $\square$

When A is Hermitian, $\|M^{-1}\|\|L\| = |\lambda_p|/|\lambda_{p+1}| < 1$, so the condition $\|M^{-1}\|\|L\| < 1$ is automatically satisfied; moreover, in this case $Q = 0$ in (2.4) and $\|\mathcal{P}\| = 1$ (see (2.9)), $\|C_i^{-1}\| = 1/\cos\angle(\mathcal{X}_i, \mathcal{V}_1)$ and $\|W_1^*\Delta_i\| \le \|\Delta_i\|$, $\|W_2^*\Delta_i\| \le \|\Delta_i\|$. Thus (3.1) becomes

$$\tan\angle(\mathcal{X}_{i+1}, \mathcal{V}_1) \le \frac{|\lambda_p|}{|\lambda_{p+1}|}\,\frac{\sin\angle(\mathcal{X}_i, \mathcal{V}_1) + \tau_i}{\cos\angle(\mathcal{X}_i, \mathcal{V}_1) - \tau_i}. \tag{3.2}$$

If we now take $\delta = 0$ in Theorem 3.1, then we recover two convergence results for the exact solves case (see [20, Thm. 5.2] and [27, p. 383]). This point is clarified in the next corollaries.

Corollary 3.2. [illegible] 3 [illegible] 1 [illegible] $\tau_i = 0$ [illegible]

$$t_i \le \|M^{-1}\|\|L\|t_{i-1}$$

[illegible]

$$t_i \le \|M^{-i}\|\|L^i\|t_0 \le \left(\frac{|\lambda_p|}{|\lambda_{p+1}|} + \eta_i\right)^i t_0,$$

[illegible] $\lim_{i\to\infty}\eta_i = 0$

[illegible] The first inequality follows directly from Theorem 3.1. For the second one, we have

$$\begin{aligned} t_i &= \left\|\left(W_2^* A^{-1} X_i\right)\left(W_1^* A^{-1} X_i\right)^{-1}\right\| \\ &= \left\|\left(W_2^* A^{-i} X_0\right)\left(W_1^* A^{-i} X_0\right)^{-1}\right\| \\ &= \|\left(M^{-i} S_0\right)\left(L^{-i} C_0\right)^{-1}\| \le \|M^{-i}\|\|L^i\|t_0. \end{aligned}$$

Now, using the fact that for any square matrix E, $\|E^i\| \le (\rho(E) + \eta_E^{(i)})^i$, where $\rho(E)$ is the spectral radius of E and $\lim_{i\to\infty}\eta_E^{(i)} = 0$, we obtain with obvious notation

$$\|M^{-i}\|\|L^i\| \le \left(\rho(M^{-1}) + \eta_{M^{-1}}^{(i)}\right)^i \left(\rho(L) + \eta_L^{(i)}\right)^i = \left(\frac{|\lambda_p|}{|\lambda_{p+1}|} + \eta_i\right)^i$$

with

$$\eta_i = \eta_{M^{-1}}^{(i)}|\lambda_p| + \eta_L^{(i)}/|\lambda_{p+1}| + \eta_{M^{-1}}^{(i)}\eta_L^{(i)} \to 0 \quad \text{as} \quad i \to \infty. \qquad \square$$

In practice, the block size p is chosen to enable the computation of invariant subspaces corresponding to close/multiple/complex pairs of eigenvalues. Therefore, to speed up the convergence, it is desirable to choose p larger than the dimension of the sought invariant subspace. Thus an estimate of the angle between $\mathcal{X}_i$ and a subspace $\widetilde{\mathcal{V}}_1 \subset \mathcal{V}_1$ is needed. Corollary 3.2 does not give such an estimate because t_i relates $\mathcal{X}_i$ to $\mathcal{V}_1$ not to a subspace $\widetilde{\mathcal{V}}_1 \subset \mathcal{V}_1$. The following corollary treats this point.

Corollary 3.3. [illegible] 3 [illegible] 1 [illegible] $V_1 = \begin{pmatrix}\widetilde{V}_1 & \widetilde{\widetilde{V}}_1\end{pmatrix}$ [illegible] $\widetilde{V}_1 \in \mathbb{C}^{n\times q}$, $q \le p$ [illegible] $\widetilde{\mathcal{V}}_1 := \mathcal{R}(\widetilde{V}_1)$ [illegible] A [illegible] $\lambda_1, \ldots, \lambda_q$ [illegible]

$$\sin\angle(\widetilde{\mathcal{V}}_1, \mathcal{X}_i) \le \left(\frac{|\lambda_q|}{|\lambda_{p+1}|} + \eta_i\right)^i \left\|(I - \mathcal{P})\widetilde{X}_0\right\|$$

[illegible] $\lim_{i\to\infty}\eta_i = 0$ [illegible] $\widetilde{X}_0 = X_0 C_0^{-1}\left(\begin{smallmatrix} I_q \\ 0 \end{smallmatrix}\right)$.

[illegible] From the proof of Corollary 3.2, we have

$$\left\|S_iC_i^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\| = \left\|\left(M^{-i}S_0\right)\left(L^{-i}C_0\right)^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\|$$
$$= \left\|M^{-i}\left(S_0C_0^{-1}\right)L^i\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\|.$$

Since L is upper triangular, $L^i\left(\begin{smallmatrix} I_q \\ 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} I_q \\ 0 \end{smallmatrix}\right)\left(L_{1:q,1:q}\right)^i$. Then

$$\left\|S_iC_i^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\| \le \|M^{-i}\|\|\left(L_{1:q,1:q}\right)^i\|\left\|S_0C_0^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\|$$
$$\le \left(\frac{|\lambda_q|}{|\lambda_{p+1}|} + \eta_i\right)^i\left\|S_0C_0^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\| \quad \text{as in Corollary 3.2.}$$

The proof is completed by noting that

$$\sin\angle(\widetilde{\mathcal{V}}_1, \mathcal{X}_i) = \left\|(I - X_iX_i^*)\widetilde{V}_1\right\|$$
$$\le \left\|\widetilde{V}_1 - X_iC_i^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\| = \left\|S_iC_i^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\|$$

and that

$$\left\|S_0C_0^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix}\right\| = \left\|X_0C_0^{-1}\begin{pmatrix} I_q \\ 0 \end{pmatrix} - \widetilde{V}_1\right\| = \left\|(I - \mathcal{P})\widetilde{X}_0\right\|. \qquad \square$$

Note that this corollary generalizes [21, Thm. 5.2] in the sense that the estimate on $\sin\angle(\widetilde{\mathcal{V}}_1, \mathcal{X}_i)$ deals with invariant subspaces rather than eigenvectors.

3.1. Use of block-GMRES as inner iteration. In this section we restrict our attention to the use of block-GMRES as inner solver in Algorithm 1. Block-GMRES belongs to the family of block Krylov subspace methods (see [21]), and it is attractive for large (sparse) linear systems with multiple right-hand sides, as in the case of interest.

Assume that block-GMRES is used to solve a linear system with multiple right-hand sides of the form

$$AZ = B, \quad B \in \mathbb{C}^{n\times p}, \tag{3.3}$$

and that B can be decomposed as

$$B = V_1B_1 + V_2B_2, \tag{3.4}$$

with V_1 and V_2 given in (2.5), $B_1 \in \mathbb{C}^{p\times p}$ nonsingular, and $B_2 \in \mathbb{C}^{(n-p)\times p}$. Then we have the following theorem.

Theorem 3.4. [illegible] $B - AZ_k$ [illegible] Z_k [illegible] (3.3) [illegible] k [illegible] $Z_0 = 0$ [illegible]

$$\|B - AZ_k\| \le \min_{p\in\bar{\mathbf{P}}_{k-1}} \|p(M)\|\|\mathcal{S}\|\|L^{-1}\|\|B_2B_1^{-1}\|\|B_1\|, \tag{3.5}$$

[illegible] $\bar{\mathbf{P}}_l$ [illegible] p_l [illegible] l [illegible] $p_l(0) = 1$ [illegible] $\mathcal{S}$ [illegible] 2.2

[illegible] Since $Z_k \in \mathcal{R}\left(B \; AB \; \dots \; A^{k-1}B\right)$ and block-GMRES minimizes the residual, we have

$$\|B - AZ_k\| = \min_{G_1,\dots,G_k \in \mathbb{C}^{p\times p}} \left\| B + \sum_{i=1}^{k} A^i B G_i \right\| .$$

Let $f_1, \dots, f_{k-1} \in \mathbb{C}$. Set $F = B_1^{-1} L^{-1} B_1$ and choose

$$G_1 = f_1 I - F, \quad G_i = f_i I - f_{i-1} F, \; i = 2, \dots, k-1, \quad \text{and} \quad G_k = -f_{k-1} F.$$

Then

$$\|B - AZ_k\| \le \min_{f_1,\dots,f_{k-1} \in \mathbb{C}} \left\| B - ABF + \sum_{i=1}^{k-1} f_i A^i (B - ABF) \right\| .$$

Now observe that the decomposition (3.4) yields

$$\begin{aligned} B - ABF &= V_2 \left(B_2 - M B_2 F\right) \\ &= V_2 \left(B_2 B_1^{-1} L - M B_2 B_1^{-1}\right) L^{-1} B_1 \\ &= -V_2 \mathcal{S}(B_2 B_1^{-1}) \; L^{-1} B_1 \end{aligned}$$

and therefore

$$\|B - AZ_k\| \le \min_{f_1,\dots,f_{k-1} \in \mathbb{C}} \left\| \left(I + \sum_{i=1}^{k-1} f_i M^i \right) \mathcal{S}(B_2 B_1^{-1}) \; L^{-1} B_1 \right\| .$$

The proof is completed by noting that

$$\min_{f_1,\dots,f_p \in \mathbb{C}} \left\| I + \sum_{i=1}^{k-1} f_i M^i \right\| = \min_{p \in \bar{\mathbf{P}}_{k-1}} \|p(M)\|$$

and that

$$\left\| \mathcal{S}(B_2 B_1^{-1}) \; L^{-1} B_1 \right\| \le \|\mathcal{S}\| \|B_2 B_1^{-1}\| \|L^{-1}\| \|B_1\|. \qquad \square$$

Note that the minimum in (3.5) is taken with respect to the matrix M and not A as in the usual theory. Also note that according to Proposition 2.1 the quantity $\|B_2 B_1^{-1}\| \|B_1\|$ behaves like the sine of the largest canonical angle between $\mathcal{V}_1$ and $\mathcal{R}(B)$.

To estimate $\min_{p \in \bar{\mathbf{P}}_{k-1}} \|p(M)\|$ we use the following lemma, whose proof can be read off from that of, e.g., [13, Lemma 1].

Proposition 3.5. [illegible] E [illegible] ϕ [illegible] E [illegible] $\phi(\infty) = \infty$ [illegible] $\phi'(\infty) > 0$ [illegible] M [illegible] E [illegible] $0 \notin E$ [illegible]

$$\min_{p \in \bar{\mathbf{P}}_{k-1}} \|p(M)\| \le N \left(\frac{1}{|\phi(0)|} \right)^{k-1}, \tag{3.6}$$

[illegible] $N = \frac{3\, l(\partial E)}{2\pi d(\partial E)}$ [illegible] $l(\partial E)$ [illegible] ∂E [illegible] E [illegible] $d(\partial E)$ [illegible] M [illegible] ∂E

An immediate corollary is as follows.

COROLLARY 3.6. [illegible] M [illegible] $M + \delta M$ [illegible] $\|\delta M\| < d(\partial E)$ [illegible]

$$\min_{p\in\bar{\mathbf{P}}_{k-1}} \|p(M+\delta M)\| \le N_\delta \left(\frac{1}{|\phi(0)|}\right)^{k-1}, \tag{3.7}$$

[illegible] $N_\delta = \frac{3\,l(\partial E)}{2\pi(d(\partial E)-\|\delta M\|)}$

A favorable situation for the bound in Proposition 3.5 is when the numerical range of M is well separated from 0. Then $|\phi(0)| \gg 1$ and $\min_{p\in\bar{\mathbf{P}}_{k-1}} \|p(M)\|$ goes to 0 quickly as k increases.

Proposition 3.5 remains valid if the numerical range of M is replaced by the ϵ-spectrum of M defined, for $\epsilon > 0$, by

$$\Lambda_\epsilon(M) = \{\lambda \in \mathbb{C} : \ \sigma_{\min}(\lambda I - M) \le \epsilon\}.$$

Then the constant N in (3.6) should be replaced by the larger one $N = \frac{3\,l(\partial E)}{2\pi\,\epsilon}$ (see [13]). The advantage here is that the set $\Lambda_\epsilon(M)$ is generally smaller than the numerical range of M (see [29]). Thus the set E can be chosen far from 0, which leads to the favorable condition $|\phi(0)| \gg 1$. For the perturbed case a similar change is needed in (3.7).

A combination of Theorem 3.4 and Proposition 3.5 gives the following result.

THEOREM 3.7. [illegible] Z_k [illegible] (3.3) [illegible] k [illegible] $Z_0 = 0$ [illegible] 3.5 [illegible]

$$k \ge 1 + \frac{1}{\log|\phi(0)|}\left(\log\left(N\|\mathcal{S}\|\|L^{-1}\|\right) + \log\frac{\|B_2B_1^{-1}\|\|B_1\|}{\tau}\right), \tag{3.8}$$

[illegible] $\|B - AZ_k\| \le \tau$

Note that the bound in (3.8) is only a sufficient condition which guarantees that the norm of the residual is less than τ. It is clear that the required accuracy may be reached for k smaller than the bound (3.8) suggests.

In step 3 of Algorithm 1, the system to be solved by block-GMRES is $AY_i = X_i$. The right-hand side X_i decomposes as in (2.13), which is of the same form as (3.4). In this context, Theorem 3.7 tells us that the residual obtained with k_i iterations of block-GMRES starting with 0 is less than $\tau_i = \delta\|R_i\|$ if

$$\begin{aligned} k_i &\ge 1 + \frac{1}{\log|\phi(0)|}\left(\log\left(N\|\mathcal{S}\|\|L^{-1}\|\right) + \log\frac{\|S_iC_i^{-1}\|\|C_i\|}{\tau_i}\right) \\ &= 1 + \frac{1}{\log|\phi(0)|}\left(\log\left(N\|\mathcal{S}\|\|L^{-1}\|\right) + \log\frac{s_i}{\delta\|R_i\|}\right). \end{aligned} \tag{3.9}$$

The next proposition shows that as $\mathcal{X}_i$ starts to approximate $\mathcal{V}_1$, the ratio $s_i/\|R_i\|$ is bounded independent of i, and thus the number of inner iterations needed by block-GMRES is bounded independent of i.

PROPOSITION 3.8. [illegible] X_i [illegible]

$$X_i = V_1C_i + V_2S_i \quad [illegible] \quad \|S_i\| < \epsilon. \tag{3.10}$$

[illegible] $0 \le \epsilon < \min\left(1, -\frac{1}{2} + \frac{1}{2}\sqrt{1 + \mathrm{sep}(T_{22}, L)/\|A\|}\right)$ [illegible]

$$\frac{s_i}{\|R_i\|} \le \frac{1+\epsilon}{1-\epsilon}\,\frac{\|\mathcal{P}\|}{\mathrm{sep}(T_{22}, L) - 4\|A\|\epsilon(\epsilon+1)}. \tag{3.11}$$

[illegible] 3 [illegible] 1 [illegible] i [illegible]. Note first that the condition on ϵ ensures that $\mathrm{sep}(T_{22}, L) - 4\|A\|\epsilon(\epsilon+1) > 0$. Using Propositions 2.1 and 2.2, we have

$$s_i \leq \|S_i\| \, \frac{1 + \|S_i\|}{1 - \|S_i\|} \leq \|\mathcal{P}\| \sin \angle(\mathcal{X}_i, \mathcal{V}_1) \, \frac{1+\epsilon}{1-\epsilon},$$

$$\|R_i\| \geq \mathrm{sep}\,(T_{22}, L_i) \sin \angle(\mathcal{X}_i, \mathcal{V}_1).$$

Then

$$\frac{s_i}{\|R_i\|} \leq \frac{1+\epsilon}{1-\epsilon} \, \frac{\|\mathcal{P}\|}{\mathrm{sep}(T_{22}, L_i)}.$$

As in Proposition 2.1, the decomposition (3.10) allows us to write $C_i = U_i + \Upsilon_i$ with U_i unitary and $\|\Upsilon_i\| < \epsilon$. Then

$$\begin{aligned} \mathrm{sep}(T_{22}, L_i) &= \mathrm{sep}\,(T_{22}, U_i^* L_i U_i) \\ &\geq \mathrm{sep}(T_{22}, L) - \|L - U_i^* L_i U_i\| \quad \text{(see [28, p. 234])}. \end{aligned}$$

Using the decomposition of X_i in $L_i = X_i^* A X_i$ and the expression of C_i given above, we obtain the bound

$$\|L - U_i^* L_i U_i\| \leq 4\|A\|\epsilon(1+\epsilon),$$

from which (3.11) is obtained. □

This proposition is illustrated in Figure 3.1 on a convection-diffusion problem (see Example 1 in section 5 for details) where after an initial increase in k_i, the number of inner iterations needed at each outer iteration settles down to an approximately constant value.

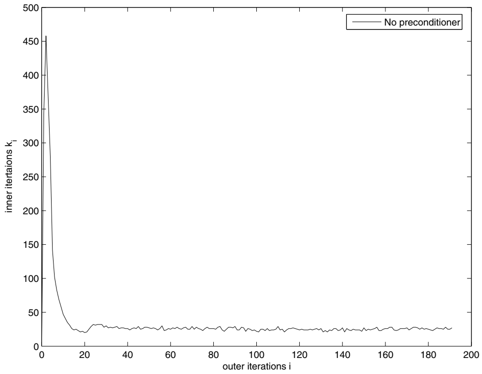


FIG. 3.1. *Outer iterations against inner iterations (Example* 1*).*

Our aim in the next section is to see if the nice property that k_i is bounded independent of i holds when system (3.3) is preconditioned.

4. Preconditioning the inexact inverse subspace iteration. A good preconditioner helps to accelerate the computations in step 3 of Algorithm 1 and hence the convergence of this algorithm. A standard way to accomplish this task is to find an approximation P of A such that the systems with the matrix P are cheap to solve. Then the matrix Y_i in step 3 of Algorithm 1 is obtained by applying block-GMRES to the (right) preconditioned block system

$$AP^{-1}Z_i = X_i, \quad Y_i = P^{-1}Z_i. \tag{4.1}$$

Let us denote by Z_{k_i} the approximation of Z_i obtained at iteration k_i of block-GMRES and satisfying

$$\left\| X_i - AP^{-1}Z_{k_i} \right\| \leq \tau_i = \delta \left\| R_i \right\|, \tag{4.2}$$

and so, with $Y_{k_i} = P^{-1}Z_{k_i}$, step 3 in Algorithm 1 is satisfied. The natural question is whether k_i can be bounded independent of i as $\mathcal{X}_i$ approaches $\mathcal{V}_1$, as for the unpreconditioned case in Proposition 3.8.

To answer this question, we attempt to repeat the analysis in the previous section. So, analogously to (2.5), assume that AP^{-1} is block-diagonalized as

$$AP^{-1} = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} K_1 & 0 \\ 0 & K_2 \end{pmatrix} \begin{pmatrix} U_1 & U_2 \end{pmatrix}^{-1}, \quad \text{with } U_i^* U_i = I, \ i = 1, 2, \tag{4.3}$$

where K_1 and K_2 have disjoint spectra. Assume further that X_i is decomposed as

$$X_i = U_1 \widetilde{C}_i + U_2 \widetilde{S}_i \quad \text{with } \|\widetilde{S}_i\| < 1, \tag{4.4}$$

and that hypotheses analogous to Proposition 3.5 and Theorem 3.7 hold. Let

$$\tilde{s}_i = \|\widetilde{S}_i \widetilde{C}_i^{-1}\| \|\widetilde{C}_i\|.$$

The question now is, can the ratio $\tilde{s}_i / \|R_i\|$ be bounded independent of i as $\mathcal{X}_i$ approaches $\mathcal{V}_1$? The a priori answer is no, as the following analysis shows.

From Proposition 2.1, (2.13), and (4.4) we have

$$\tilde{s}_i + s_i \geq \|\widetilde{S}_i\| + \|S_i\| \geq \|U_2 \widetilde{S}_i - V_2 S_i\| = \|U_1 \widetilde{C}_i - V_1 C_i\|$$

and

$$\|U_1 \widetilde{C}_i - V_1 C_i\| \geq \|U_1 \widetilde{C}_i C_i^{-1} - V_1\| \, \sigma_{\min}(C_i).$$

Denoting $\mathcal{U}_1 = \mathcal{R}(U_1)$ and using (2.12) and (2.16), we obtain

$$\|U_1 \widetilde{C}_i - V_1 C_i\| \geq (1 - s_i) \sin \angle(\mathcal{U}_1, \mathcal{V}_1).$$

Therefore

$$\tilde{s}_i \geq \sin \angle(\mathcal{U}_1, \mathcal{V}_1) - s_i \left(1 + \sin \angle(\mathcal{U}_1, \mathcal{V}_1)\right).$$

As $\mathcal{X}_i \to \mathcal{V}_1$, $s_i \to 0$, but there is no reason why $\tilde{s}_i \to 0$. In fact, $\tilde{s}_i / \|R_i\|$ may increase as $\sin \angle(\mathcal{U}_1, \mathcal{V}_1) / \|R_i\|$, leading to a corresponding increase in k_i given by (3.9). Such an increase is shown in Figures 5.1 and 5.5, where an ILU preconditioner is applied to two different examples. The above analysis shows that we do not have a result like (3.11) for preconditioned solves. It also shows that as $\mathcal{X}_i \to \mathcal{V}_1$, a necessary (but not sufficient) condition for a bound similar to (3.11) to hold for preconditioned solves is $\sin \angle(\mathcal{U}_1, \mathcal{V}_1) \approx 0$; that is, $\mathcal{V}_1$ is almost an invariant subspace of AP^{-1}. In other words, the right-hand side of (4.1) is an approximate invariant subspace of the iteration matrix AP^{-1}.

4.1. An “ideal” preconditioner. In this subsection we discuss the theoretical case of $\mathcal{U}_1 = \mathcal{V}_1$. We shall see that a preconditioner which satisfies $\mathcal{U}_1 = \mathcal{V}_1$ is

$$\mathbf{P} = AV_1V_1^* + P(I - V_1V_1^*), \tag{4.5}$$

which we call an “ideal” preconditioner. First, it is easy to see that $\mathbf{P}V_1 = AV_1 = V_1L$. Thus, $\mathcal{V}_1$ is an invariant subspace of both A and $\mathbf{P}$. Moreover, the following proposition shows that if P is a good approximation of A, then the spectrum of $A\mathbf{P}^{-1}$ should be clustered near 1.

PROPOSITION 4.1. [illegible] $\mathbf{P}$ [illegible] (4.5) [illegible] A [illegible] (2.2) [illegible] $A\mathbf{P}^{-1}$ [illegible]

$$\begin{pmatrix} I & V_1^*A\mathbf{P}^{-1}V_1^{\perp} \\ 0 & T_{22}((V_1^{\perp})^*PV_1^{\perp})^{-1} \end{pmatrix}.$$

[illegible]. We have

$$\left(V_1 \;\; V_1^{\perp}\right)^* A\mathbf{P}^{-1}\left(V_1 \;\; V_1^{\perp}\right) = \begin{pmatrix} V_1^*A\mathbf{P}^{-1}V_1 & V_1^*A\mathbf{P}^{-1}V_1^{\perp} \\ (V_1^{\perp})^*A\mathbf{P}^{-1}V_1 & (V_1^{\perp})^*A\mathbf{P}^{-1}V_1^{\perp} \end{pmatrix}.$$

Now observe that $A\mathbf{P}^{-1}V_1 = V_1$ and $(V_1^{\perp})^*A = T_{22}(V_1^{\perp})^*$. Then

$$\left(V_1 \;\; V_1^{\perp}\right)^* A\mathbf{P}^{-1}\left(V_1 \;\; V_1^{\perp}\right) = \begin{pmatrix} I & V_1^*A\mathbf{P}^{-1}V_1^{\perp} \\ 0 & T_{22}(V_1^{\perp})^*\mathbf{P}^{-1}V_1^{\perp} \end{pmatrix}.$$

Finally, since $\mathbf{P}V_1^{\perp} = PV_1^{\perp}$, we have

$$\begin{aligned}\left((V_1^{\perp})^*\mathbf{P}^{-1}V_1^{\perp}\right)\left((V_1^{\perp})^*PV_1^{\perp}\right) &= \left((V_1^{\perp})^*\mathbf{P}^{-1}V_1^{\perp}\right)\left((V_1^{\perp})^*\mathbf{P}V_1^{\perp}\right) \\ &= (V_1^{\perp})^*\mathbf{P}^{-1}(I - V_1V_1^*)\mathbf{P}V_1^{\perp} = I.\end{aligned}$$

Hence, $\left((V_1^{\perp})^*\mathbf{P}^{-1}V_1^{\perp}\right) = \left((V_1^{\perp})^*PV_1^{\perp}\right)^{-1}$. □

If P is a good approximation of A, then $(V_1^{\perp})^*PV_1^{\perp}$ will be a good approximation of T_{22}, and hence the eigenvalues of $A\mathbf{P}^{-1}$ should be clustered around 1.

Now, assume that $\mathcal{V}_1$ is a simple invariant subspace of $A\mathbf{P}^{-1}$. This ensures the existence of a block-diagonalization of the form

$$A\mathbf{P}^{-1} = \begin{pmatrix} V_1 & U \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & K \end{pmatrix}\begin{pmatrix} V_1 & U \end{pmatrix}^{-1} \text{ with } U^*U = I, \;\; \operatorname{sep}(I, K) > 0. \tag{4.6}$$

Assume also that X_i can decomposed, for all $i \geq 0$, in the form

$$X_i = V_1\widetilde{C}_i + U\widetilde{S}_i \;\text{ with }\; \widetilde{\epsilon}_i := \|\widetilde{S}_i\| < 1. \tag{4.7}$$

Multiplying (2.13) and (4.7) on the left by W_2^* gives $S_i = W_2^*U\widetilde{S}_i$. It is easy to see that W_2^*U is nonsingular and therefore that $\widetilde{S}_i = (W_2^*U)^{-1}S_i$. Assume $\mathcal{X}_i \to \mathcal{V}_1$, so that $\widetilde{\epsilon}_i \to 0$ and there exists $\widetilde{\epsilon} < 1$ such that $\widetilde{\epsilon}_i \leq \widetilde{\epsilon}$ for all $i \geq 0$. Then, from Proposition 2.1, we have

$$\begin{aligned}\tilde{s}_i &\leq \frac{1 + \|\widetilde{S}_i\|}{1 - \|\widetilde{S}_i\|}\,\|\widetilde{S}_i\| \\ &\leq \frac{1 + \widetilde{\epsilon}}{1 - \widetilde{\epsilon}}\,\|(W_2^*U)^{-1}\|\; s_i.\end{aligned}$$

Now a proof similar to that of Proposition 3.8 shows that $s_i/\|R_i\|$, and therefore $\tilde{s}_i/\|R_i\|$, is bounded independent of i. This analysis shows that if the ideal preconditioner were available, then we would be able to show that the iterations used by block-GMRES should be independent of i as in Proposition 3.8.

4.2. The "tuned" preconditioner. Of course, the ideal preconditioner cannot be used in practice since V_1 is unknown, so we replace $\mathbf{P}$ by the "tuned" preconditioner

$$\mathbf{P}_i = AX_iX_i^* + P(I - X_iX_i^*), \tag{4.8}$$

where the V_1 in (4.5) is replaced by X_i computed by Algorithm 1. This preconditioner satisfies the tuning condition

$$\mathbf{P}_iX_i = AX_i.$$

This is a generalization of the condition proposed in [8] and [7] in the context of inexact inverse iteration, but the motivation given here is different. Note that the tuned preconditioner changes at each iteration i of Algorithm 1 and its quality improves with that of X_i. Since $\mathbf{P}_iX_i = AX_i = X_iL_i + R_i$. We see that as $\mathcal{X}_i \to \mathcal{V}_1$, $R_i \to 0$, and so, in the limit, $\mathbf{P}_i$ has $\mathcal{V}_1$ as a invariant subspace. Also, the tuning condition can be written as

$$A\mathbf{P}_i^{-1}(AX_i) = AX_i, \tag{4.9}$$

which means that $A\mathcal{X}_i$ is an invariant subspace of $A\mathbf{P}_i^{-1}$ corresponding to the eigenvalue 1, which is a property shared with the ideal preconditioner given by (4.5). So the tuned preconditioner also has the nice property of clustering around 1 at least a part of the spectrum of $A\mathbf{P}_i^{-1}$. Asymptotically, that is, when $\mathcal{X}_i \to \mathcal{V}_1$, the tuned preconditioner $\mathbf{P}_i$ will behave like the ideal preconditioner.

We now prove a result for the tuned preconditioner corresponding to that given by Proposition 3.8 for unpreconditioned solves, namely, that the number of inner iterations needed to achieve (4.2) will be independent of i. This is to be expected given the closeness of $A\mathbf{P}_i^{-1}$ to $A\mathbf{P}^{-1}$, and is exactly what is observed in the numerical examples discussed in section 5.

Assume that X_i decomposes as

$$X_i = V_1C_i + V_2S_i \quad \text{with} \quad \|S_i\| \to 0 \quad \text{as} \quad i \to \infty \tag{4.10}$$

and define ϵ_i by

$$\epsilon_i = \max_{j \geq i} \|S_j\|. \tag{4.11}$$

Note that the sequence ϵ_i is decreasing.

In order to prove a result similar to that of Proposition 3.8 for the tuned preconditioned system, i.e., system (4.1) with $\mathbf{P}_i$ as preconditioner, we need the following three lemmas.

LEMMA 4.2. [illegible]

$$X_iX_i^* = V_1V_1^* + E_i, \quad [illegible] \quad \|E_i\| \leq \|S_i\| \leq \epsilon_i, \tag{4.12}$$

[illegible] ϵ_i [illegible] α [illegible] β [illegible] i [illegible]

$$\alpha \sin\angle(A\mathcal{X}_i, \mathcal{X}_i) \leq \|R_i\|, \tag{4.13}$$

$$\|A\mathbf{P}^{-1} - A\mathbf{P}_i^{-1}\| \leq \beta\, \epsilon_i. \tag{4.14}$$

[illegible] We use the same notation as in the proof of Proposition 3.8. The property (4.12) is a consequence of Proposition 2.1 and the fact that $\|E_i\| = \sin\angle(\mathcal{X}_i, \mathcal{V}_1)$ (see (2.11)).

The columns of $AX_i\,(L_i^*L_i+R_i^*R_i)^{-\frac{1}{2}}$ form an orthonormal basis of $A\mathcal{X}_i$. Therefore

$$\begin{aligned}\sin\angle(A\mathcal{X}_i,\mathcal{X}_i) &= \|(I-X_iX_i^*)AX_i\,(L_i^*L_i+R_i^*R_i)^{-\frac{1}{2}}\,\|\\ &= \|R_i\,(L_i^*L_i+R_i^*R_i)^{-\frac{1}{2}}\,\|\\ &\le \|R_i\|/\sigma_{\min}(L_i),\end{aligned}$$

and as in Proposition 3.8,

$$\sigma_{\min}(L_i)\ge\sigma_{\min}(L)-4\|A\|\epsilon_i(1+\epsilon_i).$$

Then since ϵ_i is decreasing, there exists $\alpha>0$ independent of i, such that for ϵ_i small enough

$$\sigma_{\min}(L)-4\|A\|\epsilon_i(1+\epsilon_i)\ge\alpha$$

and then

$$\sin\angle(A\mathcal{X}_i,\mathcal{X}_i)\le\|R_i\|/\alpha.$$

From (4.5), (4.8), and (4.12) we have $\mathbf{P}_i=\mathbf{P}+(A-P)E_i$. Then

$$\begin{aligned}A\mathbf{P}^{-1}-A\mathbf{P}_i^{-1} &= A\mathbf{P}^{-1}\left(\mathbf{P}_i-\mathbf{P}\right)\mathbf{P}_i^{-1}\\ &= A\mathbf{P}^{-1}\left(A-P\right)E_i\left(\mathbf{P}+(A-P)E_i\right)^{-1},\\ \|A\mathbf{P}^{-1}-A\mathbf{P}_i^{-1}\| &\le \|A\mathbf{P}^{-1}\|\|A-P\|\|E_i\|\|\mathbf{P}^{-1}\|\left\|\left(I+\mathbf{P}^{-1}(A-P)E_i\right)^{-1}\right\|\\ &\le \frac{\|A\mathbf{P}^{-1}\|\|A-P\|\|\mathbf{P}^{-1}\|}{1-\|\mathbf{P}^{-1}(A-P)\|\,\|E_i\|}\|E_i\|,\end{aligned}$$

and the same argument used for α shows the existence of β independent of i such that

$$\frac{\|A\mathbf{P}^{-1}\|\|A-P\|\|\mathbf{P}^{-1}\|}{1-\|\mathbf{P}^{-1}(A-P)\|\,\|E_i\|}\le\frac{\|A\mathbf{P}^{-1}\|\|A-P\|\|\mathbf{P}^{-1}\|}{1-\|\mathbf{P}^{-1}(A-P)\|\,\epsilon_i}\le\beta.\qquad\square$$

The following lemma shows that under some natural hypotheses, $A\mathbf{P}_i^{-1}$ will have a block-diagonalization close to that of $A\mathbf{P}^{-1}$ given in (4.6).

Lemma 4.3. [illegible] $\mathcal{V}_1$ [illegible] $A\mathbf{P}^{-1}$ [illegible] ϵ_i [illegible] $A\mathbf{P}_i^{-1}$ [illegible]

$$A\mathbf{P}_i^{-1}=\begin{pmatrix}\widetilde{U}_i & U_i\end{pmatrix}\begin{pmatrix}I & 0\\ 0 & K_i\end{pmatrix}\begin{pmatrix}\widetilde{U}_i & U_i\end{pmatrix}^{-1}\tag{4.15}$$

[illegible] $\widetilde{U}_i^*\widetilde{U}_i=I$ [illegible] $U_i^*U_i=I$ [illegible]

$$\begin{aligned}\|K-K_i\| &\le c_1\epsilon_i,\\ \sin\angle(\mathcal{U},\mathcal{U}_i) &\le c_2\epsilon_i \qquad \mathcal{U}=\mathcal{R}(U) \quad \mathcal{U}_i=\mathcal{R}(U_i \quad ,\\ \sin\angle(\mathcal{V}_1,\widetilde{\mathcal{U}}_i) &\le c_3\epsilon_i \qquad \widetilde{\mathcal{U}}_i=\mathcal{R}(\widetilde{U}_i)=A\mathcal{X}_i),\end{aligned}$$

[illegible] c_1 c_2 [illegible] c_3 [illegible] i [illegible]

[illegible]. Since $\mathcal{V}_1$ is a simple invariant subspace, we know that the block-diagonalization (4.6) exists and [28, Thm. 2.8] and Lemma 4.2 can be used to compare the invariant subspaces of $A\mathbf{P}^{-1}$ and $A\mathbf{P}_i^{-1}$. Thus for ϵ_i sufficiently small, there exist matrices U_i and K_i such that

$$A\mathbf{P}_i^{-1}U_i = U_iK_i \quad \text{with} \quad U_i^*U_i = I$$

and positive constants c_1 and c_2 independent of i such that

$$\|K - K_i\| \leq c_1\epsilon_i,$$
$$\sin\angle(\mathcal{U},\mathcal{U}_i) \leq c_2\epsilon_i.$$

From (4.9) it is clear that

$$\widetilde{U}_i = AX_i\left((AX_i)^*(AX_i)\right)^{-1/2}$$

satisfies $A\mathbf{P}_i^{-1}\widetilde{U}_i = \widetilde{U}_i$ and $\widetilde{U}_i^*\widetilde{U}_i = I$. Moreover, there exists c_3 independent of i such that, for ϵ_i sufficiently small,

$$\sin\angle(\mathcal{V}_1,\widetilde{\mathcal{U}}_i) \equiv \sin\angle(\mathcal{V}_1, A\mathcal{X}_i) \leq c_3\epsilon_i,$$

because $\sin\angle(\mathcal{V}_1, A\mathcal{X}_i) \leq \sin\angle(\mathcal{V}_1,\mathcal{X}_i) + \sin\angle(\mathcal{X}_i, A\mathcal{X}_i) \leq \epsilon_i + \|R_i\|/\alpha$, using(4.13), and

$$\|R_i\| = \|(I - X_iX_i^*)AX_i\| \leq \|M\|\|S_i\| + \|E_i\|\|A\| \leq 2\|A\|\epsilon_i.$$

Since 1 is not an eigenvalue of K, then for ϵ_i sufficiently small 1 cannot be an eigenvalue of K_i. This shows the existence of the decomposition (4.15). □

The next lemma shows the continuous dependence of a spectral projection on the matrix.

Lemma 4.4. [illegible] B [illegible] C [illegible] $P_\gamma(B)$ [illegible] $P_\gamma(C)$ [illegible] B [illegible] C [illegible] γ [illegible] $\|B - C\| \leq \xi$ [illegible] $m_\gamma(B) = \max_{\lambda\in\gamma}\|(\lambda I - B)^{-1}\|$ [illegible] $\xi m_\gamma(B) < 1$ [illegible]

$$\|P_\gamma(B) - P_\gamma(C)\| \leq \frac{1}{2\pi}l_\gamma\frac{\xi m_\gamma^2(B)}{1 - \xi m_\gamma(B)},$$

[illegible] l_γ [illegible] γ

[illegible]. See, e.g., [10, sect. 8.2]. □

We are now in a position to state and prove the key result in this paper.

Theorem 4.5. [illegible] X_i [illegible] (4.10)–(4.11) [illegible] ϵ_i [illegible] $\mathcal{V}_1$ [illegible] $A\mathbf{P}^{-1}$ [illegible] K [illegible] (4.6) [illegible] 3.5 [illegible] ϵ_i [illegible] k_i [illegible] Z_{k_i} [illegible]

$$\left\|X_i - A\mathbf{P}_i^{-1}Z_{k_i}\right\| \leq \tau_i = \delta\,\|R_i\| \tag{4.16}$$

[illegible] i

[illegible]. Let ϕ and E be given by Proposition 3.5 applied to K (instead of M).

For small enough ϵ_i, Lemma 4.3 shows that the decomposition (4.15) holds and Corollary 3.6 can be used with K_i to obtain a constant $\widehat{N}$ independent of i such that

$$\min_{p\in\bar{\mathbf{P}}_{k-1}}\|p(K_i)\| \leq \widehat{N}\left(\frac{1}{|\phi(0)|}\right)^{k-1}.$$

Decompose X_i in $\mathcal{R}(\widetilde{U}_i \;\; U_i)$ as

$$X_i = \widetilde{U}_i \widehat{C}_i + U_i \widehat{S}_i$$

and, for ϵ_i small enough, define $\widehat{s}_i$ by $\widehat{s}_i = \|\widehat{S}_i \widehat{C}_i^{-1}\| \|\widehat{C}_i\|$.

It is a simple task to show that the residual obtained with k_i iterations of block-GMRES starting with 0 is less than $\tau_i = \delta \|R_i\|$ if

$$k_i \geq 1 + \frac{1}{\log |\phi(0)|} \left(\log \left(\widehat{N} \|\widehat{\mathcal{S}}_i\| \right) + \log \frac{\widehat{s}_i}{\delta \|R_i\|} \right), \tag{4.17}$$

where $\|\widehat{\mathcal{S}}_i\| = \|I - K_i\| \leq \|I - K\| + c_1 \epsilon_i$ can be bounded independent of i since ϵ_i is decreasing.

Now in order to show that k_i can be bounded independent of i for small enough ϵ_i, it remains only to show that the ratio $\widehat{s}_i / \|R_i\|$ possesses this property.

From Proposition 2.1 we have

$$\|\widehat{S}_i\| \leq \|\widehat{\mathcal{P}}_i\| \sin \angle(\widetilde{\mathcal{U}}_i, \mathcal{X}_i) \equiv \|\widehat{\mathcal{P}}_i\| \sin \angle(A\mathcal{X}_i, \mathcal{X}_i),$$

where $\widehat{\mathcal{P}}_i$ is the spectral projection of $A\mathbf{P}_i^{-1}$ onto $\widetilde{\mathcal{U}}_i$.

We have

$$\sin \angle(A\mathcal{X}_i, \mathcal{X}_i) \leq \sin \angle(A\mathcal{X}_i, \mathcal{V}_1) + \sin \angle(\mathcal{V}_1, \mathcal{X}_i) \leq (c_3 + 1)\epsilon_i.$$

The term $\|\widehat{\mathcal{P}}_i\|$ is bounded as

$$\|\widehat{\mathcal{P}}_i\| \leq \|\widehat{\mathcal{P}} - \widehat{\mathcal{P}}_i\| + \|\widehat{\mathcal{P}}\|,$$

where $\widehat{\mathcal{P}}$ is the spectral projection of $A\mathbf{P}^{-1}$ onto $\mathcal{V}_1$. For small enough ϵ_i, (4.14) shows that Lemma 4.4 can be applied. Taking, in this lemma, γ as the circle of center 1 and radius ϵ_i, we obtain

$$\|\widehat{\mathcal{P}} - \widehat{\mathcal{P}}_i\| \leq \frac{\beta m_\gamma^2(A\mathbf{P}^{-1})}{1 - \beta m_\gamma(A\mathbf{P}^{-1})\epsilon_i} \, \epsilon_i.$$

Since ϵ_i is decreasing, we have for ϵ_i small enough

$$\frac{\beta m_\gamma^2(A\mathbf{P}^{-1})}{1 - \beta m_\gamma(A\mathbf{P}^{-1})\epsilon_i} \leq c_4$$

with c_4 independent of i and hence

$$\|\widehat{S}_i\| \leq (c_4 \epsilon_i + \|\widehat{\mathcal{P}}\|)(c_3 + 1)\epsilon_i \leq c_5 \epsilon_i \text{ with } c_5 = \left(\|\widehat{\mathcal{P}}\| + c_4 \right)(c_3 + 1).$$

Finally from Proposition 2.1 and Lemma 4.2, we have for ϵ_i small enough

$$\begin{aligned} \frac{\widehat{s}_i}{\|R_i\|} &\leq \frac{1 + \|\widehat{S}_i\|}{1 - \|\widehat{S}_i\|} \, \frac{\|\widehat{S}_i\|}{\alpha \, \sin \angle(A\mathcal{X}_i, \mathcal{X}_i)} \\ &\leq \frac{1 + \|\widehat{S}_i\|}{1 - \|\widehat{S}_i\|} \, \frac{\|\widehat{\mathcal{P}}_i\|}{\alpha} \\ &\leq \frac{1 + c_5 \epsilon_i}{1 - c_5 \epsilon_i} \, \frac{c_4 \epsilon_i + \|\widehat{\mathcal{P}}\|}{\alpha}. \end{aligned}$$

Since ϵ_i is decreasing, the last inequality shows that the ratio $\frac{\hat{s}_i}{\|R_i\|}$ is bounded independent of i for small enough ϵ_i. □

The numerical results illustrate this theorem, namely, that the number of iterations is asymptotically independent of i; see Figures 5.1 and 5.5.

Note that instead of using right preconditioning, as in (4.1), left preconditioning could have been used. A left preconditioner would also yield the necessary condition that $\mathcal{V}_1$ should be an approximate invariant subspace of $P^{-1}A$. However, the right-hand side is now $P^{-1}X_i$ and so a "tuned" preconditioner (see (4.8)) is still necessary to alter the eigendirections of the preconditioned right-hand side. The theory for left preconditioning is essentially the same as for right preconditioning and is omitted. Numerical results for tuned left preconditoning show the same improvements as for tuned right preconditioning; see [8]. Finally, we note that the numerical results in [7] are for symmetrically preconditioned systems.

5. Numerical tests. In this section we present some numerical tests to illustrate the performance of Algorithm 1 when step 3 is replaced by the preconditioned block system

$$AP^{-1}Z_i = X_i, \quad Y_i = P^{-1}Z_i, \tag{5.1}$$

solved by block-GMRES with the tolerance $\tau_i = \min(\delta, \delta\|R_i\|)$, $\delta = 10^{-3}$.

Any version of block-GMRES can be used to illustrate the theory. We have chosen to use a new variant of block-GMRES which detects the near-dependence in the corresponding block-Arnoldi basis and then adapts the block sizes accordingly. As a consequence, this variant selects appropriate directions for convergence. See [19] for details.

We compare two preconditioners:

- [illegible]. P is obtained from the incomplete LU factorization of A with a drop tolerance fixed at 10^{-1}.
- [illegible]. $\mathbf{P}_i = P + F_iX_i^*$, where $F_i = AX_i - PX_i$ and P is as above. In this case the computation of $\mathbf{P}_i^{-1}$ in $Y_i = \mathbf{P}_i^{-1}Z_i$ uses the Woodbury formula (see [11]). Note that the application of $\mathbf{P}_i^{-1}$ within block-GMRES requires little extra work compared with the application of P^{-1}, with the additional work mainly needed at the outer step.

For each example, we give information on the spectrum of A, the block size p, the dimension q of the computed invariant subspace $\widetilde{\mathcal{V}}_1$ associated to the eigenvalues near 0. We show the inner iterations for the two preconditioners and the norm of the residuals, denoted by Γ_i, associated to the computed invariant subspaces.

[illegible] 1. A is obtained with a five-point stencil and centered difference discretization of the convection diffusion operator (see [12]):

$$\begin{cases} \mathcal{A}u = \Delta u + 10\frac{\partial u}{\partial x} + 10\frac{\partial u}{\partial y} & \text{on } \Omega = [0,1] \times [0,1], \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

The matrix A is of order $n = 2025$ and has $nz = 9945$ nonzero elements, $\|A\| = 16152$, $\|A - P\| = 1400$. We use $p = 6$ and look for the invariant subspace $\widetilde{\mathcal{V}}_1$ of dimension $q = 4$. The computations stop when $\|\Gamma_i\| < 10^{-8}$.

Figure 5.1 shows the inner iterations k_i for the two preconditioners, and Figures 5.2 and 5.3 show the behavior of $\|\Gamma_i\|$ during the outer iterations and compared with the total number of inner iterations. The spectrum of A and the computed eigenvalues are shown in Figure 5.4. Figure 5.1 illustrates well the theory: it shows that as the

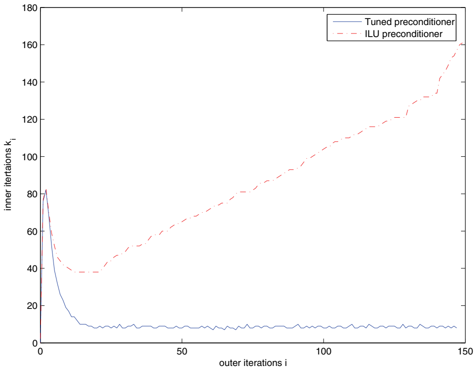


FIG. 5.1. *Outer iterations against inner iterations (Example* 1*).*

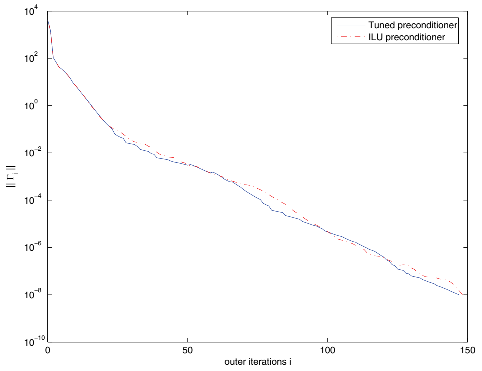


FIG. 5.2. *Residual norms against outer iterations (Example* 1*).*

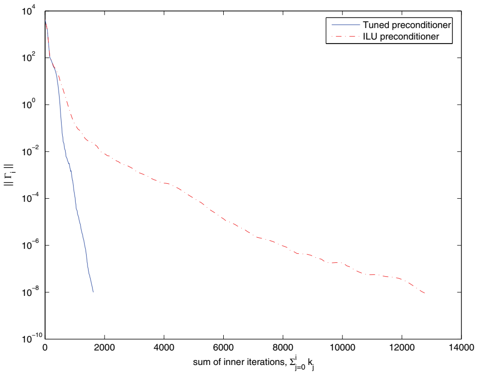


FIG. 5.3. *Residual norms against the total number of inner iterations (Example* 1*).*

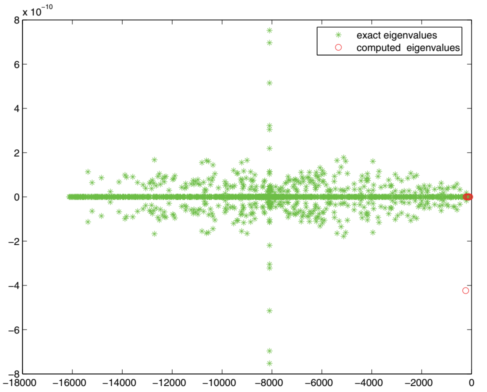


FIG. 5.4. *Eigenvalues A (Example* 1*).*

outer convergence proceeds, the number of inner iterations becomes independent of i when the tuned preconditioner is used but increases when the standard ILU preconditioner is used. Figure 5.2 illustrates that there is little difference in the performance of the two preconditioners with regard to the residual norms in step 3 of Algorithm 1. Figure 5.3 shows the dramatic improvement in overall cost achieved by the tuned preconditioner, with the required tolerance being achieved at 12.65% of the cost needed for the untuned preconditioner.

Example 2. A is the matrix $QC2534$ from the NEP set.[1] This matrix is complex, symmetric, and non-Hermitian. It is of order $n = 2534$ and has $nz = 463360$ nonzero elements, $\|A\| = 3.32$, $\|A - P\| = 0.41$. We use $p = 16$ and look for the invariant subspace $\widetilde{\mathcal{V}}_1$ of dimension $q = 10$. The computations stop when $\|\Gamma_i\| < 10^{-8}$. Figure 5.5 compares the number of inner iterations for the ILU and tuned preconditioners. Figures 5.6 and 5.7 show the norm of the residual of the invariant subspace associated to the q eigenvalues near 0. Figure 5.8 shows the spectrum of A and the computed eigenvalues. Similar comments apply as in Example 1. The tuned preconditioner requires a roughly constant number of inner iterations per outer iteration (see Figure 5.5).

[1] see http://math.nist.gov/MatrixMarket/collections/NEP.html

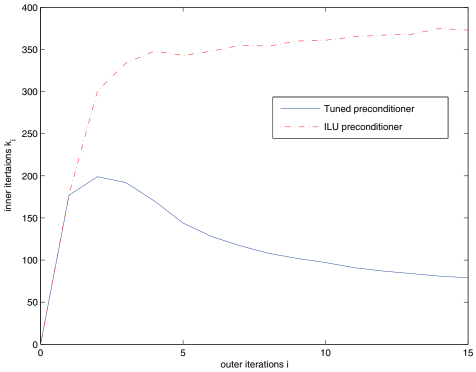


FIG. 5.5. *Outer iterations against inner iterations (Example 2).*

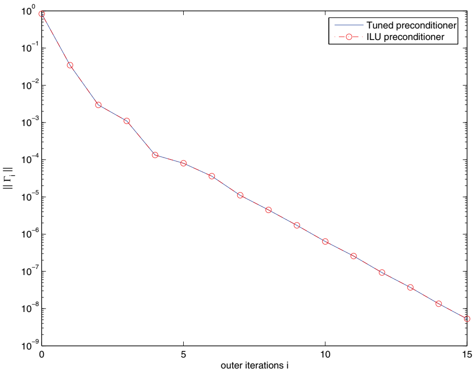


FIG. 5.6. *Residual norms against outer iterations (Example 2).*

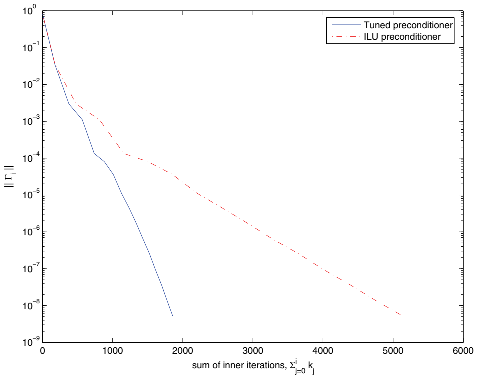


FIG. 5.7. *Residual norms against the total number of inner iterations (Example 2).*

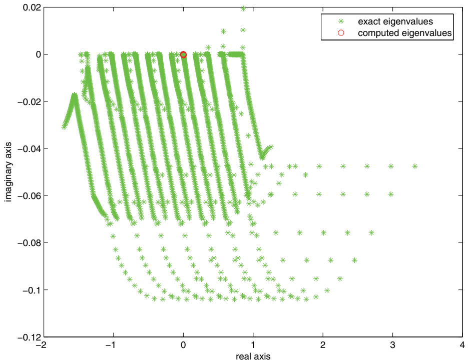


FIG. 5.8. *Eigenvalues of A (Example 2).*

Finally, the overall costs for the tuned preconditioned system to achieve $\|\Gamma_i\| < 10^{-8}$ are about 36.18% of the costs for the untuned preconditioner (see Figure 5.7).

REFERENCES

[1] P.A. Absil, C.G. Baker, and K.A. Gallivan, *A truncated-CG style method for symmetric generalized eigenvalue problems*, J. Comput. Appl. Math., 189 (2006), pp. 274–285.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems*, SIAM, Philadelphia, 2000.

[3] J. Berns-Müller, I.G. Graham, and A. Spence, *Inexact inverse iteration for symmetric matrices*, Linear Algebra Appl., 416 (2006), pp. 389–413.

[4] J. Berns-Müller and A. Spence, *Inexact inverse iteration with variable shift for nonsymmetric eigenvalue problems*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1069–1082.

[5] J. Berns-Müller and A. Spence, *Inexact inverse iteration and GMRES*, submitted (2006).

[6] J. Brandts, *The Riccati algorithm for eigenvalues and invariant subspaces of matrices with inexpensive action*, Linear Algebra Appl., 358 (2003), pp. 335–365.

[7] M.A. Freitag and A. Spence, *A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems*, IMA J. Numer. Anal., to appear.

[8] M.A. Freitag and A. Spence, *Convergence of inexact inverse iteration with application to preconditioned inexact solves*, BIT, 47 (2007), pp. 27–44.

[9] M.A. Freitag and A. Spence, *Rayleigh quotient iteration and simplified Jacobi-Davidson method with preconditioned iterative solves*, Linear Algebra Appl., to appear.

[10] S.K. Godunov, *Modern Aspects of Linear Algebra*, Transl. Math. Monogr. 175, AMS, Providence, RI, 1998.

[11] G.H. Golub and Ch. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[12] G.H. Golub and Q. Ye, *Inexact inverse iteration for generalised eigenvalue problems*, BIT, 40 (2000), pp. 671–684.

[13] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[14] Y.-L. Lai, K.-Y. Lin, and L. Wen-Wei, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 1 (1997), pp. 1–13.

[15] Y. Notay, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[16] Y. Notay, *Inner Iterations in Eigenvalue Solvers*, Report GANMN 05-01, Université Libre de Bruxelles, Brussels, Belgium, 2005.

[17] B.N. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, 1998.

[18] M. Robbé and M. Sadkane, *Riccati-based preconditioner for computing invariant subspaces of large matrices*, Numer. Math., 92 (2002), pp. 129–159.

[19] M. Robbé and M. Sadkane, *Exact and inexact breakdowns in the block GMRES method*, Linear Algebra Appl., 49 (2006), pp. 265–285.

[20] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.

[21] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.

[22] A. Sameh and Z. Tong, *The trace minimization method for the symmetric generalized eigenvalue problem*, J. Comput. Appl. Math., 123 (2000), pp. 155–175.

[23] V. Simoncini and Eldén, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.

[24] 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.

[25] 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.

[26] G.W. Stewart, *Simultaneous iteration for computing invariant subspaces of non-Hermitian matrices*, Numer. Math., 25 (1976), pp. 123–136.

[27] G.W. Stewart, *Matrix Algorithms. Volume* II: *Eigensystems*, SIAM, Philadelphia, 2001.

[28] G.W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.

[29] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.

[30] F. Xue and H.C. Elman, *Convergence Analysis of Iterative Solvers in Inexact Rayleigh Quotient Iteration*, preprint, 2008.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 114–132

THE SPECTRUM OF A GLUED MATRIX[*]

BERESFORD N. PARLETT[†] AND CHRISTOF VÖMEL[‡]

Abstract. A glued matrix can be obtained from a direct sum of p copies of an unreduced symmetric tridiagonal matrix T by modifying the junctions by a glue γ, in one of two ways, so that the new tridiagonal matrix has no zero off-diagonal entries. Despite being unreduced, a glued matrix can have some eigenvalues agreeing to hundreds of decimal places. This makes glued matrices practically useful as test matrices for tridiagonal eigensolvers such as inverse iteration and the MRRR algorithm. However, the eigenvalue distribution of a glued matrix is a fascinating subject of theoretical interest in its own right. By means of secular equations, this paper studies how width and placement of the eigenvalue clusters of a glued matrix depend on T, on p, and on γ. Interlacing properties and the question of eigenvalue repetition between T and a glued matrix are also investigated.

Key words. glued matrix, rank-1 gluing, rank-2 gluing, eigenvalue clusters, secular equation

AMS subject classifications. 15A18, 15A23

DOI. 10.1137/070687062

1. Introduction.

This paper analyzes the eigenvalue distribution of real symmetric tridiagonal matrices of dimension pm that are constructed by *gluing* p copies of a (smaller) tridiagonal $T \in \mathcal{R}^{m\times m}$ together. Gluing is formally defined in Definition 2.1. Here we give a symbolic illustration. For $p = 2$, the direct sum is held together by $\gamma \neq 0$ at the junction in one of the two following ways:

$$G_1 = \left[\begin{array}{cccc|cccc} \ddots & \ddots & \ddots & & & & & \\ & * & * & * & & & & \\ & & * & *+\gamma & \gamma & & & \\ \hline & & & \gamma & *+\gamma & * & & \\ & & & & * & * & * & \\ & & & & & \ddots & \ddots & \ddots \end{array}\right], \tag{1.1}$$

$$G_2 = \left[\begin{array}{cccc|cccc} \ddots & \ddots & \ddots & & & & & \\ & * & * & * & & & & \\ & & * & * & \gamma & & & \\ \hline & & & \gamma & * & * & & \\ & & & & * & * & * & \\ & & & & & \ddots & \ddots & \ddots \end{array}\right].$$

For $p > 2$, that is, more than two copies, the glue is inserted at each junction. We want to consider γ as a perturbation, so we assume it to be reasonably small compared to the other matrix entries.

Glued matrices are of great practical relevance as test matrices because their spectra can have very tight clusters even when the glue γ is not particularly small. We first encountered this phenomenon while investigating the behavior of the MRRR algorithm on a matrix obtained from gluing five copies of the Wilkinson matrix W_{201}^{+};

[*] Received by the editors April 2, 2007; accepted for publication (in revised form) by F. Dopico April 30, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/simax/31-1/68706.html

[†] Mathematics Department and Computer Science Division, University of California, Berkeley, CA 94720 (parlett@math.berkeley.edu).

[‡] Institute of Computational Science, ETH Zürich, CAB, Universitätsstraße 6, 8092 Zürich, Switzerland (cvoemel@inf.ethz.ch).

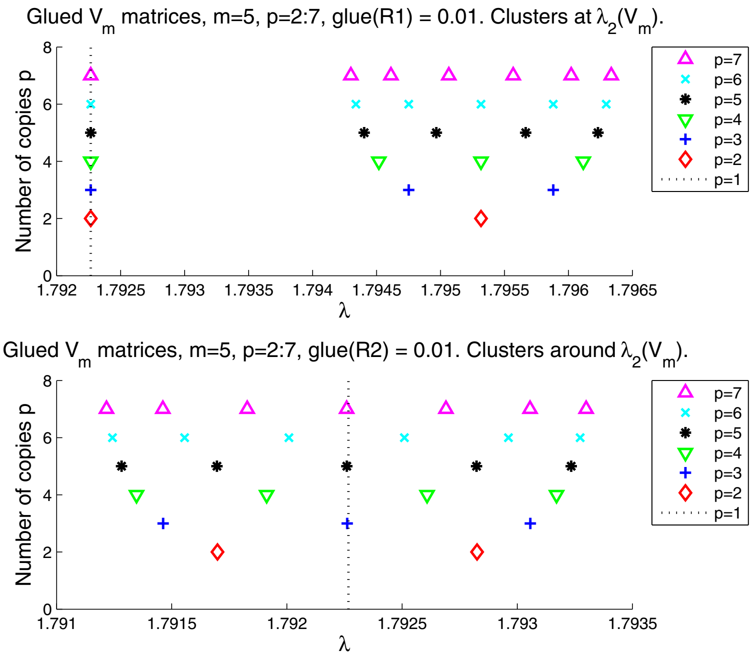


FIG. 1.1. *An example of the structure of the eigenvalue clusters obtained for G_1 (top) and G_2 (bottom) when gluing $p = 2$ to $p = 7$ copies of a matrix V_m to itself, where $\gamma = 10^{-2}$. (The matrix V_m is defined in section 2.) Shown are the resulting clusters of the respective glued matrices in the neighborhood of an isolated eigenvalue $\bar{\lambda}$ of the original matrix V_m (which is indicated by the dashed vertical line). For G_1, all except one eigenvalue form a cluster to the right of $\bar{\lambda}$, with the single remaining eigenvalue being seemingly very close to the location of $\bar{\lambda}$. In the case of G_2, the eigenvalues are distributed around $\bar{\lambda}$.*

see [10]. The experiments in that paper showed that gluing can constitute a powerful alternative to constructing unreduced tridiagonal matrices with tight eigenvalue clusters via LAPACK's test matrix generator [1, 9], Lanczos without reorthogonalization [8, 16], or the solution of a tridiagonal inverse eigenvalue problem [11, 14, 19].

The original motivation for the present work was to provide theoretical understanding of the strong eigenvalue clustering that can be found in glued matrices. However, this paper goes beyond this goal: we not only analyze the localization of eigenvalues but also investigate interlacing properties and study the question of eigenvalue repetition between T and a glued matrix. To intrigue and motivate the reader, Figure 1.1 is an illustrative example of the differences in the eigenvalue distributions of G_1 and G_2 for increasing p and fixed γ.

Section 3 gives a first, qualitative picture of the spectrum of glued matrices, emphasizing interlacing properties. It is also investigated when the original T and a glued matrix have eigenvalues in common. We prove that all eigenvalues of T are eigenvalues of G_1 and that T and G_2 only share eigenvalues under special conditions stated there.

The quantitative analysis of G_1 and G_2 starts with section 4. We show the existence of secular functions Γ_1 and Γ_2 whose zeros give the eigenvalues of G_1 and G_2, respectively. Interestingly, Γ_1 and Γ_2 are the respective determinants of tridiagonal matrices of rational functions, the former with a Toeplitz [6, 12] and the latter with a rank-1 perturbed quasi-, or pseudo-, Toeplitz [2, 4, 15] structure. This is the key to finding formulae for the eigenvalues of G_1 and G_2. The contribution of T to a cluster

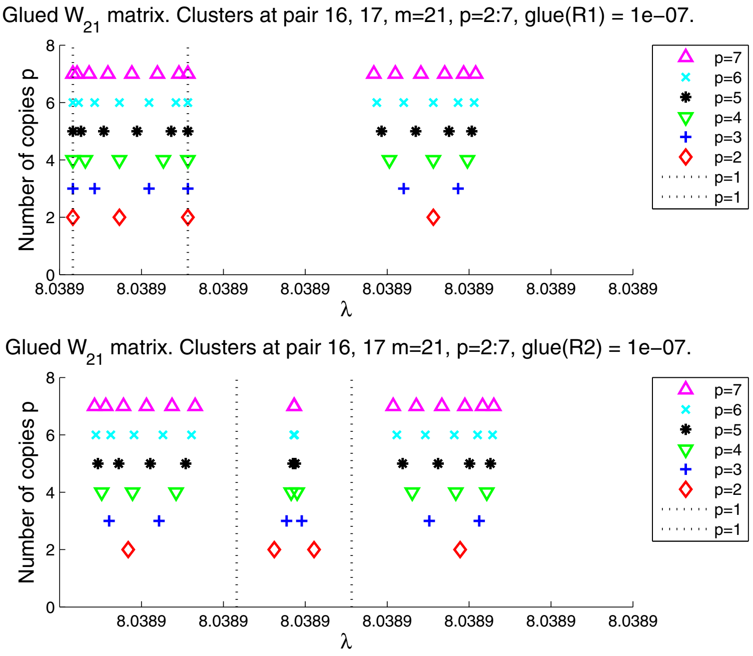


Fig. 1.2. *An example of the cluster structures obtained for G_1 (top) and G_2 (bottom) when gluing $p = 2$ to $p = 7$ copies of a matrix T to itself, where $\gamma = 10^{-7}$. Shown are the resulting clusters of the respective glued matrices in the neighborhood of a close pair of eigenvalues of the original matrix T, indicated by dashed vertical lines. (The Wilkinson matrix W^+_{2n+1} is defined in section 2.)*

of the glued matrix can be described by its spectrum $\{\lambda_j\}$ and certain associated weights $\{\omega_j\}$. Together with γ and the number of copies p, these determine the location and width of the cluster, or clusters, of G_1 and G_2.

Determining the cluster structure of G_1 from the secular function Γ_1 is quite easy. However, we found it more difficult to analyze Γ_2 and determine the clusters for G_2. A considerable part of this paper is thus devoted to elucidating the structure of the clusters of close eigenvalues for G_2. It pleased and surprised us that in the analysis, it is possible to perform a local change of variable in each cluster. This yields an equation with an interesting universal part S_p which captures the role of p in the distribution of zeros within the cluster; see section 5. The clusters of zeros of G_2 near an isolated eigenvalue of T occur where the graph of S_p intersects with a certain hyperbola. These intersections interlace with the poles of S_p which have a Chebyshev distribution, and each zero sticks out further from its pole the closer it is to the center; see section 6.

This paper mainly focuses on the structure of an eigenvalue cluster of a glued matrix close to an isolated eigenvalue of T. It is also interesting to study the spectrum of a glued matrix when T has two close eigenvalues. To give the reader another glimpse, Figure 1.2 plots the cluster structures for G_1 and G_2 around a close pair of eigenvalues of the original tridiagonal. Section 3 makes some qualitative remarks about this setting; see also the conclusions in section 7 for additional comments and the technical report [17] for some analysis.

2. Notation, basic facts, and the definition of glued matrices.

This section introduces the notation and establishes basic facts about the eigenpairs of tridiagonal matrices needed for the analysis of glued matrices. We try to follow Householder notational conventions: capital Roman letters for matrices, lowercase Roman letters for column vectors, and lowercase Greek letters for scalars and functions. We denote by v^t the transpose of column vector v.

Throughout this paper, we consider a real, symmetric tridiagonal matrix

$$T := \text{tridiag}\begin{pmatrix} & \beta_1 & & \beta_2 & & \cdots & & \beta_{m-2} & & \beta_{m-1} & \\ \alpha_1 & & \alpha_2 & & \cdots & & \cdots & & \alpha_{m-1} & & \alpha_m \\ & \beta_1 & & \beta_2 & & \cdots & & \beta_{m-2} & & \beta_{m-1} & \end{pmatrix} \tag{2.1}$$

that is unreduced (entries $\beta_i \neq 0$, $i = 1, 2, \ldots, m-1$). Its spectral factorization is

$$T = Z\Lambda Z^t, \quad \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m), \tag{2.2}$$

where the eigenvector matrix $Z = [z_1, z_2, \ldots, z_m]$ is orthogonal, $Z^t = Z^{-1}$.

Fact 1 (see [16, Lemma 7.7.1]). *The eigenvalues of T are simple, and we number them such that*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_m.$$

Note, however, that some eigenvalues may be indistinguishable on a computer; see [10].

Fact 2 (see [16, Theorem 7.9.3]). *Each eigenvector z_i, $i = 1, 2, \ldots m$, has nonvanishing top and bottom entries $z_i(1)$ and $z_i(m)$.*

Fact 3 (see [16, Corollary 7.9.1]). *The product satisfies*

$$z_j(1)z_j(m) = \prod_{i=1}^{m-1} \beta_i / \chi'(\lambda_j) =: \frac{\beta_\pi}{\chi'(\lambda_j)} \tag{2.3}$$

with β_i from (2.1) *and where $\chi(\zeta) := \det[\zeta I - T]$ is the characteristic polynomial of T and χ' its derivative. Note that*

$$\text{sign}\,(\chi'(\lambda_j)) = \text{sign}\left(\prod_{i \neq j}(\lambda_j - \lambda_i)\right) = (-1)^{m-j}.$$

Fact 4. *Define the "weight"*

$$\omega_j := |z_j(1)z_j(m)|,\ j = 1, \ldots, m. \tag{2.4}$$

Then

$$\sum_1^m (-1)^{m-j}\omega_j = \pm e_1^t Z Z^t e_m = 0,$$

$$0 < \sum_1^m \omega_j \leq \frac{1}{2}\sum_1^m (z_j(1)^2 + z_j(m)^2) = 1.$$

When T is persymmetric ($T(i,j) = T(m+1-j, m+1-i)$), then $\sum_1^m \omega_j = 1$.

Next, we define the glued matrices G_1 and G_2, formalizing the illustration in (1.1). For any square matrix M, let $M^{(p)} := \text{diag}(M, M, \ldots, M)$ denote the direct sum of p copies. Define also two auxiliary matrices $E_1, E_m \in \mathcal{R}^{(mp)\times(p-1)}$:

$$E_1 := \begin{bmatrix} 0 & & & & \\ e_1 & 0 & & & \\ & e_1 & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 0 \\ & & & & e_1 \end{bmatrix}, \quad E_m := \begin{bmatrix} e_m & & & & \\ 0 & e_m & & & \\ & 0 & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & e_m \\ & & & & 0 \end{bmatrix}, \tag{2.5}$$

where $e_1^t = (1, 0, \ldots, 0)$, $e_m^t = (0, 0, \ldots, 0, 1)$ are vectors with m entries.

Definition 2.1. *For unreduced T, $p \geq 2$, and $\gamma \neq 0$ define*

$$G_1(T, p, \gamma) := T^{(p)} + \gamma(E_1 + E_m)(E_1 + E_m)^t, \tag{2.6}$$

$$G_2(T, p, \gamma) := T^{(p)} + \gamma(E_m E_1^t + E_1 E_m^t). \tag{2.7}$$

Furthermore, $G_1(T, 1, \gamma) := T =: G_2(T, 1, \gamma)$.

Thus, at each junction in $T^{(p)}$ between two Ts, we add a rank-1 (rank-2) update to obtain G_1 (G_2). Correspondingly, we speak of rank-1 (rank-2) gluing of T.

Section 4 will show that the glued matrix G_1 is easier to analyze than G_2; much of the theory resembles the analysis of the divide-and-conquer algorithm [3, 5, 13].

Remark 1. Because we wish to consider G_1 and G_2 as low rank modifications of $T^{(p)}$ and not the other way around, we assume throughout this paper that $|\gamma|$ is reasonably bounded, for example, by the geometric mean of $|\beta_i|$, $i = 1, \ldots, m-1$.

Fact 5 (see [12, Example 7.4], [15]). *The spectral decomposition of a symmetric tridiagonal Toeplitz matrix of dimension m is*

$$Toep\,(a, b, a) = SDS^t, \quad D = \operatorname{diag}\left(b + 2a\, \cos\left(\frac{j\pi}{m+1}\right)\right), \tag{2.8}$$

and S is orthonormal with

$$s_{ij} = \sqrt{\frac{2}{m+1}} \sin\left(\frac{ij\,\pi}{m+1}\right), \; i, j = 1, \ldots, m. \tag{2.9}$$

Last, we recall the definition of a matrix introduced by Wilkinson in [18]:

$$W_{2n+1}^{+} = \operatorname{tridiag}\left(\begin{smallmatrix} & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & \\ n & & n-1 & & \cdots & & 1 & & 0 & & 1 & & n-1 & & n \\ & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & & 1 & \end{smallmatrix}\right). \tag{2.10}$$

W_{2n+1}^{+} is useful because its eigenvalues appear in pairs of varying closeness; the largest pair differ by approximately $2/((n-2)!)^2$. It also is persymmetric and thus has $|z(1, j)| = |z(2n+1, j)|$, $j = 1, 2, \ldots, 2n+1$, and $\sum \omega_j = 1$. The leading n-dimensional submatrix of W_{2n+1}^{+} is called V_n; its eigenvalues are all well separated from each other.

3. A qualitative analysis.

Using results on eigenvalue monotonicity from [16], this section gives a first crude picture of the effects of gluing and refines it.

3.1. Introductory comments.

The first observation concerns the role of $\operatorname{sign}(\gamma)$.

Remark 2. $G_2(T, p, \gamma)$ and $G_2(T, p, -\gamma)$ have the same eigenvalues since they are orthogonally similar to each other with respect to $\operatorname{diag}(I_m, -I_m, I_m, -I_m, \ldots)$. Here $I_m = [e_1, e_2, \ldots, e_m]$ denotes the $m \times m$ identity matrix.

Furthermore, since $\operatorname{trace}(G_2(T, p, \gamma)) = p \cdot \operatorname{trace}(T)$, rank-2 gluing does not change the average (arithmetic mean) of the eigenvalues. On the other hand, rank-1 gluing does change the average by $2\gamma(p-1)/(mp)$, and the sign of γ does matter.

The next result gives a first crude idea of the eigenvalues' locations.

Remark 3. By (2.6), $G_1(T,p,\gamma)$ is a rank $p-1$ modification of $T^{(p)}$ whose nonzero eigenvalues equal 2γ. By (2.7), $G_2(T,p,\gamma)$ is a rank $2(p-1)$ modification of $T^{(p)}$ whose nonzero eigenvalues are $\pm\gamma$. Weyl's theorem [16, Theorem 10.3.1]) yields, for $\gamma > 0$,

$$\lambda_j(G_1) \in [\lambda_j(T^{(p)}), \lambda_j(T^{(p)}) + 2\gamma], \;\; \lambda_j(G_2) \in [\lambda_j(T^{(p)}) - \gamma, \lambda_j(T^{(p)}) + \gamma]. \tag{3.1}$$

We note that G_1 alters the first and last T differently from the others. One can construct G_2 by rank-1 gluing from the direct sum

$$\operatorname{diag}\left(T - \gamma e_m e_m^t, T - \gamma(e_1 e_1^t + e_m e_m^t), \ldots, T - \gamma(e_1 e_1^t + e_m e_m^t), T - \gamma e_1 e_1^t\right).$$

If one knows the eigenvalues of the perturbed Ts, then no rank-2 gluing theory is needed. However, we do not consider this a realistic point of view and assume that only the spectrum of T is known.

3.2. Interlacing properties. This section gives some qualitative understanding of Figures 1.1 and 1.2. Key is the following inductive construction of a glued matrix.

Remark 4. $G_1(T,p+1,\gamma)$ ($G_2(T,p+1,\gamma)$) can be written as a rank-1 (rank-2) modification of the direct sum S_1 (S_2) of T and $G_1(T,p,\gamma)$ ($G_2(T,p,\gamma)$).

A special case of the rank theorem [16, Theorem 10.3.1, Corollary 10.3.1] is needed.

THEOREM 3.1. *Let S_1 and S_2 be as in Remark 4, $\gamma > 0$, and let $n := (p+1)\cdot m$. Then*

$$\lambda_i\left(G_1(T,p+1,\gamma)\right) \in \begin{cases} [\lambda_i(S_1), \lambda_{i+1}(S_1)], & i \neq n, \\ [\lambda_n(S_1), \lambda_n(S_1) + 2\gamma], & i = n. \end{cases} \tag{3.2}$$

$$\lambda_i\left(G_2(T,p+1,\gamma)\right) \in [\lambda_{i-1}(S_2), \lambda_{i+1}(S_2)], \; i \neq 1, n. \tag{3.3}$$

First consider rank-1 gluing for $p = 2$. By definition (2.6),

$$G_1(T,2,\gamma) = T^{(2)} + \gamma \begin{pmatrix} e_m \\ e_1 \end{pmatrix} \left(e_m^t \; e_1^t\right).$$

Since all eigenvalues of $T^{(2)}$ have multiplicity 2, all eigenvalues of T are also eigenvalues of $G_1(T,2,\gamma)$ by (3.2). Next, using the same argument together with Remark 4, we also find that each eigenvalue of T is an eigenvalue of $G_1(T,3,\gamma)$. Continuing by induction, we obtain the following theorem.

THEOREM 3.2. *Any eigenvalue λ of a real unreduced $m \times m$ symmetric tridiagonal T is also an eigenvalue of the rank-1 glued matrix $G_1(T,p,\gamma)$ for any γ and for $p \geq 2$.*

Application of the interlacing property (3.2) also explains the "chandelier" shape of the clusters in rank-1 gluing with increasing p that one can see in Figure 1.1. Furthermore, it shows that when T has two close eigenvalues, there will be a cluster of G_1 that is "squished" between them, as seen in Figure 1.2.

Rank-2 gluing satisfies weaker interlacing properties. Consider $G_2(T,2,\gamma)$, a rank-2 modification of $T^{(2)}$. By (3.3), $G_2(T,2,\gamma)$ has an eigenvalue in each interval $\left[\lambda_{i-1}(T^{(2)}), \lambda_{i+1}(T^{(2)})\right]$. Since all eigenvalues of $T^{(2)}$ have multiplicity 2, it follows that $G_2(T,2,\gamma)$ has two eigenvalues in each interval $[\lambda_j(T), \lambda_{j+1}(T)]$. Now, using the inductive construction from Remark 4, we find that the direct sum of $G_2(T,p,\gamma)$ and T has four eigenvalues in the interval $[\lambda_j(T), \lambda_{j+1}(T)]$; hence at least two eigenvalues of $G_2(T,p+1,\gamma)$ must lie in it. This proves the following lemma.

LEMMA 3.3. *For any $p \geq 2$, there lie at least two eigenvalues of $G_2(T,p,\gamma)$ in each closed interval $[\lambda_j(T), \lambda_{j+1}(T)]$, $j = 1, \ldots, m-1$.*

3.3. When G_2 shares an eigenvalue with T. By Theorem 3.2, each eigenvalue of T is an eigenvalue of G_1. For G_2, the situation is more complicated and is illuminated in Theorem 3.4. The first part gives a necessary and sufficient condition for $\lambda(T)$ to be an eigenvalue of G_2, independent of γ. By giving an explicit construction of the associated eigenvector, the second part shows that when T and G_2 have a common eigenvalue for a certain p, then by gluing additional copies one can find more matrices with this property. We use M'' to denote the submatrix of a given matrix M obtained by deleting the first and last rows and columns.

THEOREM 3.4. *Let λ, z denote an eigenpair of a real unreduced $m \times m$ symmetric tridiagonal T.*

- *λ is an eigenvalue of $G_2(T, p, \gamma)$ for every $\gamma \neq 0$ if and only if (λ, w'') is an eigenpair of $[G_2(T, p-2, \gamma)]''$.*
- *If the previous condition holds for a certain p, then λ also is an eigenvalue of $G_2(T, q, \gamma)$ for $q = p + (p-1),\ p + 2(p-1), \ldots,$ with eigenvector*

$$(z^t\phi_1, \gamma w^t\phi_2, z^t\phi_3, \gamma w^t\phi_4, \ldots, \gamma w^t\phi_{2k}, z^t\phi_{2k+1})^t, \tag{3.4}$$

where $w := (0, (w'')^t, 0)^t$ and $\phi_1, \ldots, \phi_{2k+1}$ are nonzero scalars.

Proof. Consider $(G_2 - \lambda I_{mp})x = 0$ and, without loss of generality, choose $x(1) := z(1) \neq 0$. The first equation determines $x(2)$, the second $x(3)$, and so on until equation $m-1$ determines $x(m)$. This m-vector, call it $\tilde{z}$, satisfies $(T - \lambda I)\tilde{z} = 0$, and since λ is simple, $\tilde{z} = z$. Now examine equation m of $(G_2 - \lambda I_{mp})x = 0$;

$$\beta_{m-1}x(m-1) + (\alpha_m - \lambda)x(m) + \gamma x(m+1) = 0.$$

Since $\beta_{m-1}x(m-1) + (\alpha_m - \lambda)x(m) = 0$ and $\gamma \neq 0$, we find $x(m+1) = 0$. Analogously, starting from the bottom up with $x(m \cdot p) := z(m) \neq 0$ shows that the entry in the row above the last copy of T, $x((m-1)\cdot p)$, is zero. Note that when starting from the top, from equation $(m+1)$

$$w''(1) := x(m+2) = -\gamma x(m)/\beta_1 = -\gamma z(m)/\beta_1 \neq 0.$$

Equations $m+2$ to $(m-1)\cdot p - 1$ now have to hold, showing that (λ, w'') is an eigenpair of $[G_2(T, p-2, \gamma)]''$.

Conversely, if (λ, w'') is an eigenpair of $[G_2(T, p-2, \gamma)]''$, it suffices to verify that

$$(z^t\phi_1, 0, \gamma(w'')^t\phi_2, 0, z^t\phi_3)^t$$

with $\phi_1 \neq 0$ is a nonzero vector from the kernel of $(G_2 - \lambda I_{mp})$.

For the last part of the theorem, observe that the construction of an eigenvector for λ with alternating scalar multiples of z and w can be continued. Let the scaling factors ϕ_i obey $\phi_1 \neq 0$, and

$$\phi_{2k} = -\phi_{2k-1}(z(m)/\beta_1\, w(2)), \quad \phi_{2k+1} = -\phi_{2k}(\beta_{m-1}\, w((m-1)\cdot p - 1)/z(1)); \tag{3.5}$$

then (3.4) is an eigenvector of $G_2(T, q, \gamma)$, $q = p + (k-1)(p-1)$. □

Remark 5. $G_2(T, 2, \gamma)$ cannot have eigenvalues in common with T. The simplest example of the special conditions from Theorem 3.4 occurs for the middle eigenvalue $\lambda_2 = 0$ of

$$\widetilde{T} := \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Then 0 is also an eigenvalue of $G_2(\widetilde{T}, 3, \gamma)$ for all γ. In general, if λ is an eigenvalue of both T and T'', then for odd p, it is also an eigenvalue of $G_2(T, p, \gamma)$ for all γ.

Remark 6. It is interesting to examine what happens when we do not consider G_2 simultaneously for all $\gamma \neq 0$. Take a tridiagonal Toeplitz matrix $T = \text{Toep}\,(a, b, a) \in \mathcal{R}^{m\times m}$, glued to itself p times, with the special glue $\gamma = a$. $G_2(T, p, a)$ is (again) a symmetric Toeplitz matrix of order mp. By (2.8), the number of eigenvalues that T and G_2 have in common equals $|\,\{i/(m+1)|1 \leq i < m+1\} \cap \{i/(mp+1)|1 \leq i < mp+1\}\,|$. Thus, when g denotes the greatest common divisor of $m+1$ and $mp+1$, the number of eigenvalues that T and G_2 have in common equals $g - 1$. In particular,

- when both $m+1$ and $mp+1$ are even, the two matrices have at least one common eigenvalue, and
- when $mp+1$ is a multiple of $m+1$, all eigenvalues of T are also eigenvalues of G_2.

4. Governing rational functions. We derive rational functions whose zeros yield those eigenvalues of $G_1(T, p, \gamma)$ and $G_2(T, p, \gamma)$ that are not already eigenvalues of T. Let $f := Z^t e_1$ and $l := Z^t e_m$ denote the first and the last row of the eigenvector matrix Z of T. Further, from (2.5) obtain two $mp \times (p-1)$ matrices

$$F = Z^{(p)t}E_1 = \begin{bmatrix} 0 & & \\ f & \cdot & \\ & \cdot & 0 \\ & & f \end{bmatrix}, \quad L := Z^{(p)t}E_m = \begin{bmatrix} l & & \\ 0 & \cdot & \\ & \cdot & l \\ & & 0 \end{bmatrix}.$$

Again, F stands for first and L for last. Thus, Definition 2.1 yields

$$G_1 = Z^{(p)}[\Lambda^{(p)} + \gamma(F+L)(F+L)^t]Z^{(p)t}, \tag{4.1}$$

$$G_2 = Z^{(p)}[\Lambda^{(p)} + \gamma(LF^t + FL^t)]Z^{(p)t}. \tag{4.2}$$

To replace $mp \times mp$ matrices by matrices of order $(p-1)$ or $2(p-1)$, we invoke the identity $\det[I + XY^t] = \det[I + Y^t X]$ to find, for any ζ that is not an eigenvalue of T,

$$\det\,[G_1 - \zeta I] = \det\,[\Lambda - \zeta I]^p \det\left[I_{(p-1)} + \gamma(F+L)^t(\Lambda^{(p)} - \zeta I)^{-1}(F+L)\right]$$

and

$$\begin{aligned}\det\,[G_2 - \zeta I] &= \det\left[\Lambda^{(p)} - \zeta I\right] \det\left[I_{mp} + \gamma(\Lambda^{(p)} - \zeta I)^{-1}(LF)\begin{pmatrix} F^t \\ L^t \end{pmatrix}\right] \\ &= \det\,[\Lambda - \zeta I]^p \det\left[I_{2(p-1)} + \gamma\begin{pmatrix} F^t \\ L^t \end{pmatrix}(\Lambda^{(p)} - \zeta I)^{-1}(LF)\right].\end{aligned}$$

Let the nilpotent $(p-1)\times(p-1)$ matrix

$$N = \text{bidiag}\begin{pmatrix} & 1 & & 1 & & \cdot & & 1 & \\ 0 & & 0 & & \cdot & & \cdot & & 0 \end{pmatrix}.$$

Further, for all possible combinations $x, y \in \{f, l\}$, define the rational functions

$$\rho_{xy}(\zeta) := x^t(\Lambda - \zeta I)^{-1}y = \sum_{j=1}^{m} \frac{x_j y_j}{\lambda_j - \zeta}. \tag{4.3}$$

Note that $\rho_{fl} = \rho_{lf}$; hence we simply write ρ. The special structure of F and L yields

$$\det[G_1 - \zeta I] = \det[\Lambda - \zeta I]^p \det\left[(1 + \gamma\rho_{ff} + \gamma\rho_{ll})I + (\gamma\rho)(N + N^t)\right] \tag{4.4}$$

and

$$\begin{aligned}\det[G_2 - \zeta I] &= \det[\Lambda - \zeta I]^p \det\begin{bmatrix} I + \gamma\rho N & \gamma\rho_{ff} I \\ \gamma\rho_{ll} I & I + \gamma\rho N^t \end{bmatrix} \\ &= \det[\Lambda - \zeta I]^p \det\left[(I + \gamma\rho N^t)(I + \gamma\rho N) - \gamma^2\rho_{ff}\rho_{ll} I\right].\end{aligned} \tag{4.5}$$

The last matrix on the right of (4.4) is a tridiagonal Toeplitz matrix. Define

$$\Gamma_1(T, p, \gamma) := \det[\mathrm{Toep}\,(\gamma\rho, 1 + \gamma(\rho_{ff} + \rho_{ll}), \gamma\rho)]; \tag{4.6}$$

then using the spectral decomposition (2.8), we have proved the next theorem.

THEOREM 4.1. *The eigenvalues of $G_1(T, p, \gamma)$ that are not eigenvalues of T are zeros of the rational function*

$$\Gamma_1(T, p, \gamma) = \prod_{k=1}^{p-1} \{1 + \gamma\big(\rho_{ff}(\zeta) + \rho_{ll}(\zeta) + 2\rho(\zeta)\cos(k\pi/p)\big)\}. \tag{4.7}$$

COROLLARY 4.2. *When*

$$\gamma \ll \min\{\lambda_{j+1} - \lambda_j, \lambda_j - \lambda_{j-1}\} := \mathrm{gap}(\lambda_j), \tag{4.8}$$

then the eigenvalues of G_1 close to λ_j, other than λ_j itself, are to first order in γ,

$$\lambda_j + \gamma\{f_j^2 + l_j^2 + 2(-1)^{m-j}\,|f_j l_j| \cos(k\pi/p)\},\ k = 1, 2, \ldots, p-1. \tag{4.9}$$

In words, the cluster near λ_j is a Chebyshev distribution of radius $2|\gamma|\omega_j$ and center $\lambda_j + \gamma(f_j^2 + l_j^2)$, $f_j = z_j^{(1)}$, $l_j = z_j^{(m)}$.

Proof. All rational functions from (4.3) share the same poles. Use the leading terms of expansion (A.8) in Appendix A.2 for $\rho_{ff}(\zeta)$, $\rho_{ll}(\zeta)$, and $\rho(\zeta)$ in (4.7). Note that because of (4.8), each factor from (4.7) contributes one of the zeros to find (4.9). □

The complication with G_2 comes from the fact that in (4.5), $N^t N = I_{p-1} - e_1 e_1^t$. Instead of a Toeplitz matrix of rational functions as in (4.6), a rank-1 perturbed quasi-, or pseudo-, Toeplitz [2, 4, 15] matrix governs rank-2 gluing. We have the next theorem.

THEOREM 4.3. *The eigenvalues of $G_2(T, p, \gamma)$ that are not eigenvalues of T are zeros of the rational function*

$$\Gamma_2(T, p, \gamma) := \det\left[\mathrm{Toep}_{p-1}\left\{\gamma\rho, 1 - \gamma^2(\rho_{ff}\rho_{ll} - \rho^2), \gamma\rho\right\} - (\gamma\rho)^2 e_1 e_1^t\right]. \tag{4.10}$$

Unfortunately, there does not seem to be a simple representation for Γ_2 in (4.10) that would be as instructive as the one for Γ_1 in Theorem 4.1. The characteristic polynomial of $N + N^t - \kappa e_1 e_1^t$, for fixed κ, may be described as a sum of Chebyshev polynomials of the second kind; see also [6, 7, 15]. However, the roots are not known explicitly, which would be necessary for our application where κ is not fixed but is a rational function.

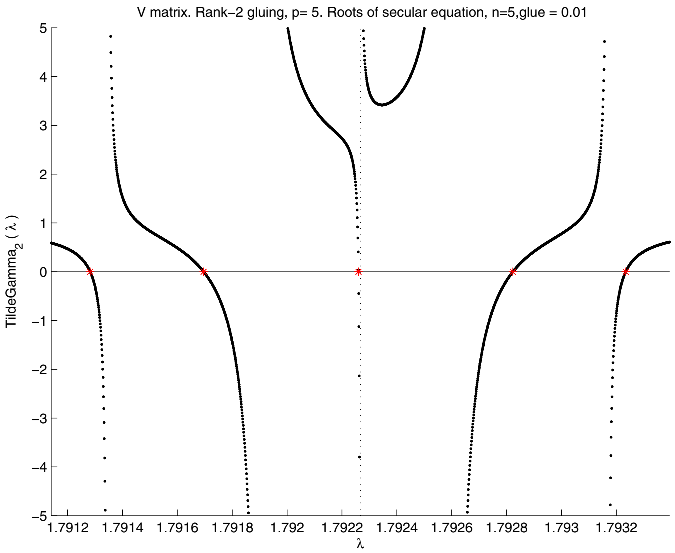


FIG. 5.1. *The secular function* (5.3) *around an isolated eigenvalue of the original matrix.*

5. Approximative roots for rank-2 gluing. Even if they are not available explicitly, one can at least try to compute approximations of the roots of Γ_2 from (4.10). Because of the rank-1 perturbation, the following approach is guided by rank-1 gluing, respectively, updating, theory. Observe that the poles of Γ_2, which are the eigenvalues of an (unperturbed) Toeplitz matrix, have a Chebyshev distribution. With

$$\epsilon(\zeta) := \rho_{ff}(\zeta)\rho_{ll}(\zeta) - \rho^2(\zeta), \tag{5.1}$$

we find by (2.8) that $\mathrm{Toep}(\gamma\rho, 1-\gamma^2\epsilon, \gamma\rho) = SMS^t$, with S according to (2.9), and with $M := \mathrm{diag}(\mu_1, \ldots, \mu_{p-1})$,

$$\mu_j = \mu_j(\zeta) := 1 - \gamma^2\epsilon(\zeta) + 2\gamma\rho(\zeta)\cos j\psi, j = 1, \ldots, p-1, \psi := \pi/p. \tag{5.2}$$

From (4.10), we obtain with $s := S^t e_1$

$$\begin{aligned}\Gamma_2(T,p,\gamma) &= \det[\mathrm{Toep} - (\gamma\rho)^2 e_1 e_1^t] = \det[S(M - (\gamma\rho)^2 ss^t)S^t]\\ &= \det[I_{p-1} - (\gamma\rho)^2 M^{-1}ss^t]\det[M] = [1 - (\gamma\rho)^2 s^t M^{-1} s]\prod_{i=1}^{p-1}\mu_i(\zeta).\end{aligned}$$

The previous derivation assumes that $M = M(\zeta)$ is nonsingular at the eigenvalues of G_2. This is the generic and difficult case. We thus obtain the secular function

$$\widetilde{\Gamma}_2(T,p,\gamma) := 1 - (\gamma\rho)^2 s^t M^{-1} s = 1 - (\gamma\rho(\zeta))^2 \frac{2}{p}\sum_{k=1}^{p-1}\frac{\sin^2 k\psi}{\mu_k(\zeta)} \tag{5.3}$$

whose zeros are the eigenvalues of G_2 we seek. An illustration is given in Figure 5.1.

Note that ρ plays no role for $p = 2$. Section 4 directly shows $\widetilde{\Gamma}_2 = 1 - \gamma^2\rho_{ff}(\zeta)\rho_{ll}(\zeta)$. In this case, one can prove, similarly to Corollary 4.2, the following theorem.

THEOREM 5.1. *When*

$$\gamma \ll \min\{\lambda_{j+1} - \lambda_j, \lambda_j - \lambda_{j-1}\} := \operatorname{gap}(\lambda_j), \tag{5.4}$$

then the eigenvalues of $G_2(T, p = 2, \gamma)$ close to λ_j are to first order in γ,

$$\lambda_j \pm |\gamma f_j l_j|. \tag{5.5}$$

In words, the cluster near λ_j is symmetrically distributed around λ_j.

For $p > 2$, one can determine the approximate location of the $p-1$ poles of $\widetilde{\Gamma}_2$ stemming from the zeros of $\mu_1, \dots, \mu_{p-1}$. Under assumption (5.4), an expansion of (5.2) in the neighborhood of λ_j yields to first order

$$\mu_k(\zeta) \approx 1 + \frac{2\gamma f_j l_j \cos k\psi}{\lambda_j - \zeta}.$$

Plugging this into (5.3) yields

$$\widetilde{\Gamma}_2 \approx 1 - (\lambda_j - \zeta)\,(\gamma\rho(\zeta))^2 \frac{2}{p} \sum_{k=1}^{p-1} \frac{\sin^2 k\psi}{(\lambda_j + 2\gamma f_j l_j \cos k\psi) - \zeta}. \tag{5.6}$$

This is a very crude approximation, and a more formal treatment of $\widetilde{\Gamma}_2$ is the topic of section 6. Nevertheless, note that everything in the numerator of the second term is positive, with the possible exception of the factor $(\lambda_j - \zeta)$. Thus, one can expect mirroring behavior of the eigenvalues of G_2 close to poles that are approximately opposite to each other with respect to λ_j. This is a microscopic version of Remark 2 that observed that rank-2 gluing does not change the arithmetic mean of the eigenvalues.

5.1. A change of variables. A further analysis of $\widetilde{\Gamma}_2$ for $p > 2$ is quite complicated. We rephrase (5.3) using the following change of variable $\zeta \to \alpha$:

$$\alpha = \alpha(\zeta) := \frac{1 - \gamma^2\epsilon(\zeta)}{-2\gamma\rho(\zeta)}. \tag{5.7}$$

It is valid locally in each cluster.

Since $\cos(p-k)\psi = -\cos k\psi$, $\sin(p-k)\psi = \sin k\psi$ for $k < p/2$, we find

$$\begin{aligned}\frac{\sin^2 k\psi}{\mu_k(\zeta)} + \frac{\sin^2(p-k)\psi}{\mu_{p-k}(\zeta)} &= \sin^2 k\psi \left(\frac{1}{1 - \gamma^2\epsilon + 2\gamma\rho\cos k\psi} + \frac{1}{1 - \gamma^2\epsilon - 2\gamma\rho\cos k\psi} \right) \\ &= \frac{\sin^2 k\psi\, 2(1-\gamma^2\epsilon)}{(1-\gamma^2\epsilon)^2 - (2\gamma\rho\cos k\psi)^2}.\end{aligned}$$

Hence, using (5.7) in (5.3), with the usual $\psi = \pi/p$, yields the secular equation

$$\begin{aligned} 0 = \widetilde{\Gamma}_2(T, p, \gamma) &= 1 - \frac{2}{p}(\gamma\rho)^2 \sum_{k=1}^{p-1} \frac{\sin^2 k\psi}{\mu_k(\zeta)} \\ &= 1 - (1-\gamma^2\epsilon) \left\{ \frac{1}{p} \sum_{k=1}^{\lfloor p/2 \rfloor} \frac{\sin^2 k\psi}{\alpha^2 - \cos^2 k\psi} \right\} =: 1 - (1-\gamma^2\epsilon) S_p(\alpha). \end{aligned} \tag{5.8}$$

$S_p(\alpha)$ is a function of α^2 and thus is symmetric around $\alpha = 0$ with $S_p(\pm 1) = \lfloor p/2 \rfloor / p$. Its center $\alpha = 0$ is given for ζ such that $1 = \gamma^2\epsilon(\zeta)$. By (5.3), there are $(p-1)$

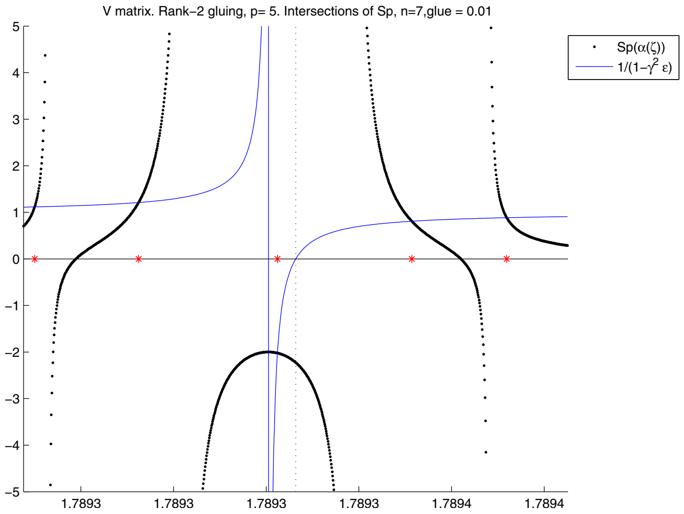


FIG. 5.2. *Intersections of $S_p(\alpha)$ and $(1-\gamma^2\epsilon)^{-1}$. The marks on the horizontal axis show the eigenvalues of G_2.*

simple poles of $\widetilde{\Gamma}_2$, each from one μ_k, around the eigenvalue of T. When changing to the α-variable, we become interested in the special ζ-intervals that include the poles of $S_p(\alpha)$. By (5.8), these are the intervals in which α varies between ± 1. For small enough γ, the zeros of $\widetilde{\Gamma}_2$ occur close to solutions of $S_p(\alpha) = 1$, that is, close to the poles. Otherwise, the zeros are the intersections of $S_p(\alpha)$ and $(1-\gamma^2\epsilon)^{-1}$; see the illustration in Figure 5.2.

6. Approximative zeros of $\widetilde{\boldsymbol{\Gamma}}_2$. In this section, we express approximations to the zeros of $\widetilde{\Gamma}_2$ in the α-coordinate. When p is even, all zeros are of the form $\lambda_j + O(\gamma)$. In contrast, when p is odd, there is one interior eigenvalue within $O(\gamma)^2$ of λ_j. Note that the zeros are roughly, but in general not exactly, symmetric about λ_j; see also Figure 5.2.

In more detail, a zero of $\widetilde{\Gamma}_2$ sticks out further from its pole the closer the pole is to the center. The outer zeros, analyzed in section 6.1, are located at distance $\pm[\cos^2 k\psi + O(\sin^2 k\psi)]^{1/2}$, $k = 1, 2, \ldots, \lfloor p/2 \rfloor - 1$. The description of the innermost zeros is given in section 6.2 and requires an examination of ρ, ϵ, and the parity of p.

6.1. The outer zeros of $\widetilde{\boldsymbol{\Gamma}}_2$. Consider (5.8). Except very close to the eigenvalue λ_j, say, $k = p/2 - 1$ when p is even or $(p-1)/2$ when p is odd, $(1-\gamma^2\epsilon)^{-1} = 1 + O(\gamma)$ in the intervals $\alpha \in\,]-1, 1[$. For simplicity, we abbreviate $(1-\gamma^2\epsilon)^{-1}$ by ν in what follows. It varies only slightly in each subinterval $]\cos k\psi, \cos(k-1)\psi[$.

With $s_k := \sin k\psi, c_k := \cos k\psi$, S_p from (5.8) becomes

$$S_p(\alpha) = \frac{1}{p} \sum_{1}^{\lfloor p/2 \rfloor} \frac{s_k^2}{\alpha^2 - c_k^2}\,.$$

To find an intelligible approximation to the solution of $\widetilde{\Gamma}_2 = 0$ between the poles $]c_k, c_{k-1}[$, we model S_p by keeping the two neighboring poles and replacing the rest by a value independent of α:

$$\widetilde{S_p}(\alpha) := \frac{1}{p}\left(\frac{s_k^2}{\alpha^2-c_k^2}+\frac{s_{k-1}^2}{\alpha^2-c_{k-1}^2}\right)+S_p''(\beta_k), \tag{6.1}$$

$$S_p''(\beta) := \frac{1}{p}\sum_{i\neq k,k-1}^{\lfloor p/2\rfloor}{}'' \frac{s_i^2}{\beta^2-c_i^2}. \tag{6.2}$$

Here β_k is a constant at our disposal; its default value is $\beta_k = c_k$, except when $k = p/2$. We say more about $S_p''(\beta_k)$ in Appendix B.

For $\alpha \in \,]c_k, c_{k-1}[$, write $\alpha = \alpha_k = [(1-\sigma_k)c_k^2 + \sigma_k c_{k-1}^2]^{1/2}$, $0 < \sigma_k < 1$. Next, we solve the quadratic (6.1) for σ_k. Substitute into $\widetilde{S_p}(\alpha) = \nu$ to find

$$p(\nu - S_p'') = \frac{s_k^2}{\sigma_k(c_{k-1}^2-c_k^2)}+\frac{s_{k-1}^2}{(\sigma_k-1)(c_{k-1}^2-c_k^2)},\ S_p'' = S_p''(\beta_k),$$

or

$$p(s_k^2 - s_{k-1}^2)(\nu - S_p'') = \frac{s_k^2}{\sigma_k}-\frac{s_{k-1}^2}{1-\sigma_k}.$$

Thus σ_k is the smaller zero $(0<\sigma<1)$ of the quadratic

$$A_k\sigma_k^2 - B_k\sigma_k + s_k^2 = 0 \tag{6.3}$$

with $A_k = p(s_k^2 - s_{k-1}^2)(\nu - S_p'')$, $B_k = A_k + (s_k^2 + s_{k-1}^2)$. The discriminant is

$$\Delta := B_k^2 - 4A_k s_k^2 = (A_k')^2 + (2s_k s_{k-1})^2 \tag{6.4}$$

with $A_k' := A_k - (s_k^2 - s_{k-1}^2) = p(s_k^2 - s_{k-1}^2)(\nu - S_p'' - 1/p)$. The difficulty with (6.4) is that the first term dominates for small k and the second one dominates for k close to $p/2$.

Case $k = 1$, the outermost pair. The quadratic (6.3) can be solved exactly, giving $\sigma_1 = s_1^2/A_1 = 1/p(\nu - S_p'')$ and $\alpha_1^2 = c_1^2 + \sigma_1(s_1^2 - s_0^2) = c_1^2 + s_1^2/p(\nu - S_p'')$. With $\nu = 1 + O(\gamma)$ and $S_p'' \approx 1/2$ (see Appendix B), one has

$$\alpha_j^{(p-1)}, \alpha_j^{(1)} \approx \pm\sqrt{\cos^2\psi + 2\sin^2\psi/p}. \tag{6.5}$$

Case $1 < k$ and $2|s_k s_{k-1}| < |A_k'|$ in (6.4). Use

$$\Delta = (A_k')^2 + (2s_k s_{k-1})^2 = (A_k')^2\left[1 + \frac{1}{2}\left(\frac{2s_k s_{k-1}}{A_k'}\right)^2 + O\left(\left(\frac{2s_k s_{k-1}}{A_k'}\right)^4\right)\right]$$

to find

$$\sigma_k = (A_k' + 2s_k^2 - \sqrt{\Delta})/2A_k = \frac{s_k^2}{A_k}\left[1 - \frac{s_{k-1}^2}{A_k'} + O\left(\frac{s_k^2 s_{k-1}^4}{(A_k')^3}\right)\right]. \tag{6.6}$$

The desired zeros of $\widetilde{\Gamma}_2$ satisfy

$$\alpha_k^2 == c_k^2 + \sigma_k(s_k^2 - s_{k-1}^2) = c_k^2 + \frac{s_k^2}{p(\nu - S_p''(c_k))}\left[1 - \frac{s_{k-1}^2}{A_k'} + O\left(\frac{s_k^2 s_{k-1}^4}{(A_k')^3}\right)\right]. \tag{6.7}$$

Case $k < p/2$ *and* $|A'_k| < 2|s_k s_{k-1}|$ *in* (6.4). We use $\Delta = (2s_k s_{k-1})^2[1 + (A'_k/2s_k s_{k-1})^2]$ so that

$$\begin{aligned}\sigma_k &= \frac{A_k + s_k^2 + s_{k-1}^2 - (2s_k s_{k-1})[1 + \frac{1}{2}(A'_k/2s_k s_{k-1})^2 + \cdots]}{2A_k} \\ &= \frac{1}{2} + \frac{(s_k - s_{k-1})^2}{2A_k} - \frac{1}{4}\left(\frac{A'_k}{A_k}\right)\left(\frac{A'_k}{2s_k s_{k-1}}\right) + O\left(\left(\frac{A'_k}{2s_k s_{k-1}}\right)^3\right). \qquad (6.8)\end{aligned}$$

We bound the second term on the right in (6.8). With $A'_k < s_k s_{k-1}$,

$$s_k^2 - s_{k-1}^2 < \frac{s_k s_{k-1}}{p(\nu - S''_p - 1/p)} < \frac{(s_k + s_{k-1})^2}{4p(\nu - S''_p - 1/p)},$$

since $\frac{1}{2}s_k s_{k-1} \le \frac{1}{4}(s_k^2 + s_{k-1}^2)$ and $\frac{1}{2}s_k s_{k-1} = \frac{1}{4}(2s_k s_{k-1})$. Thus

$$\frac{s_k - s_{k-1}}{s_k + s_{k-1}} < \frac{1}{4p(\nu - S''_p - 1/p)}. \qquad (6.9)$$

Hence one finds

$$\frac{(s_k - s_{k-1})^2}{A_k} < \left(\frac{1}{2p(\nu - S''_p - 1/p)}\right)^2.$$

Since the third term in (6.8) is bounded by $\frac{1}{4} \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{8}$, omitting the O term gives

$$3/8 < \sigma_k < \frac{1}{2} + \frac{1}{2(2p(\nu - S''_p - 1/p))^2},$$

$$\begin{aligned}(\alpha_j^{(p-1-k)})^2 = (\alpha_j^{(k)})^2 &= (1-\sigma_k)c_k^2 + \sigma_k c_{k-1}^2 = c_k^2 + \sigma_k(s_k^2 - s_{k-1}^2) \\ &\approx c_k^2 + \frac{1}{2}(s_k^2 - s_{k-1}^2). \qquad (6.10)\end{aligned}$$

We collect in one place the results of this section.

Theorem 6.1. *The cluster of eigenvalues of* $G_2(T, p, \gamma)$ *around an isolated eigenvalue* λ_j *of* T*, in the* α *variable, is given by* (6.5), (6.7), (6.10) *depending on the regime. As* k *increases, the zero moves farther away from its pole toward pole* $k - 1$.

6.2. The innermost zeros of $\widetilde{\Gamma}_2$.

Case 1: p *odd.* The previous section presented the $p - 1 = 2[(p-1)/2]$ zeros of $\widetilde{\Gamma}_2$ associated with the $p - 1$ poles. There remains one more zero, $\zeta_j^{(p)}$.

$S_p(0)$ has the interesting negative finite value, from (5.8),

$$S_p(0) = \frac{-1}{p}\sum_{k=1}^{(p-1)/2} \tan^2 k(\pi/p),$$

and, as is easily verified, $S'_p(0) = 0$. ϵ now enters the scene. In general $\zeta = \lambda_j$ is not a zero of $\widetilde{\Gamma}_2$ because, from (A.6), $\alpha(\lambda_j) = O(\gamma)$ so that S_p is finite while $\epsilon(\lambda_j) = \infty$. Yet close to λ_j, when $\alpha = 0$, then $1 - \gamma^2\epsilon = 0$, and $\widetilde{\Gamma}_2$ sinks down to 1.

Theorem 6.2. *When* p *is odd, the zero* $\zeta_j^{(p)}$ *of* $\widetilde{\Gamma}_2(T, p, \gamma)$ *that is closest to an isolated eigenvalue* λ_j *of* T *has the form*

$$\zeta_j^{(p)} = \lambda_j - \frac{\gamma^2\omega_j S_p(0)E'_j}{1 - S_p(0) + \gamma^2 S_p(0)K'_j} + O(\gamma^4\omega_j^2).$$

See (A.11) *for* E'_j *and* K'_j.

Proof. From (A.10), near λ_j,

$$\epsilon(\zeta) = \frac{\omega_j E_j'}{\lambda_j - \zeta} + K_j' + O\left(\frac{\gamma^2 \omega_j}{\mathrm{gap}(\lambda_j)}\right)$$

and, since $\alpha(\lambda_j) = O(\gamma)$,

$$S_p(\alpha) = S_p(0) + O(\alpha^2) = S_p(0) + O(\gamma^2).$$

Thus

$$\widetilde{\Gamma}_2(\zeta) = 1 - S_p(0)(1 - \gamma^2 K_j') + \frac{\gamma^2 \omega_j S_p(0) E_j'}{\lambda_j - \zeta} + O\left(\gamma^2 + \frac{\gamma^2 \omega_j}{\mathrm{gap}(\lambda_j)}\right).$$

Solve $\widetilde{\Gamma}_2 = 0$ for ζ, and the result follows. □

Case 2: *p even.* There are $p - 2 = 2(p/2 - 1)$ zeros of $\widetilde{\Gamma}_2$ associated with the poles $\pm \cos k\pi/p$, $k = 1, 2, \ldots, p/2 - 1$. The bounds just above (6.10) apply, but the dramatic difference from $k < p/2$ is that $S_p''(\alpha) < 0$ for $0 < \alpha < s_1 = \sin \pi/p$, and $|S_p''| = O(p)$. Since $\nu = 1 + O(p\gamma)$ we find that, for large p, $\nu - S_p'' \approx |S_p''|$ and in these cases $\alpha_{p/2}$ is very close to the smallest zero of S_p (i.e., $\nu = 0$):

$$\begin{aligned} \sigma_{p/2} &< 1/2[1 + (2p\,(\nu + |S_p''| - 1/p))^{-2}], \\ \alpha_{p/2} &= \pm(O + \sqrt{\sigma_{p/2}} \sin \pi/p) \approx \pm \frac{1}{\sqrt{2}} \sin \frac{\pi}{p}. \end{aligned}$$

Theorem 6.3. *When p is even, the two zeros of $\widetilde{\Gamma}_2(T, p, \gamma)$ closest to an isolated eigenvalue λ_j of T are, to first order in $\gamma\omega_j$,*

$$\lambda_j \pm \gamma \omega_j \frac{1}{\sqrt{2}} \sin \frac{\pi}{p}. \tag{6.11}$$

7. Summary and conclusions. Our initial interest in glued matrices stemmed from their importance as test matrices for tridiagonal eigensolvers such as inverse iteration and the MRRR algorithm. However, beyond their practical significance, these matrices offer a number of interesting theoretical points of study.

In this paper, we first investigated qualitative issues such as interlacing and eigenvalue repetition. Then we derived secular equations for rank-1 and rank-2 gluing and exhibited connections to tridiagonal Toeplitz matrices. Under the assumption of a small enough glue, we established expressions for the location of the eigenvalues clustered around an isolated eigenvalue of T. In the difficult analysis of rank-2 gluing, we used a change of variables to find approximations.

Even though it is interesting, this paper was only briefly concerned with the spectrum of a glued matrix when T has two close eigenvalues. We refer the interested reader to the technical report [17] for some (quite technical) analysis of this case.

Appendix A. Analysis of important rational functions.

A.1. Analysis of ρ and ϵ. Use of the weights ω_j from (2.4) in (4.3) yields

$$\rho(\zeta) = \rho_{fl}(\zeta) = f^t(\Lambda - \zeta I)^{-1} l = \sum_{i=1}^{m} \frac{(-1)^{m-i}\omega_i}{\lambda_i - \zeta}. \tag{A.1}$$

Use (2.3) to write

$$\rho(\zeta) = -\sum_{i=1}^{m} \frac{\beta_\pi}{\chi'(\lambda_i)(\zeta - \lambda_i)} = -\frac{\beta_\pi}{\chi(\zeta)}. \tag{A.2}$$

The last part of (A.2) follows by recognizing that the sum is a partial fraction expansion of $\chi(\zeta)$. As a consequence, $\rho(\zeta)$ never vanishes.

From (5.1), find

$$\epsilon(\zeta) = \sum_\mu \frac{z_\mu(1)^2}{\lambda_\mu - \zeta} \sum_\nu \frac{z_\nu(m)^2}{\lambda_\nu - \zeta} - \sum_\mu \frac{(-1)^{m-\mu}\omega_\mu}{\lambda_\mu - \zeta} \sum_\nu \frac{(-1)^{m-\nu}\omega_\nu}{\lambda_\nu - \zeta}. \tag{A.3}$$

Observe that terms in $(\lambda_\mu - \zeta)^{-2}$ in (A.3) cancel, leaving terms in $(\lambda_\mu - \zeta)^{-1}(\lambda_\nu - \zeta)^{-1}$, $\nu < \mu$, where the numerator is

$$\begin{aligned}
&z_\mu(1)^2 z_\nu(m)^2 + z_\nu(1)^2 z_\mu(m)^2 - 2z_\mu(1)z_\mu(m)z_\nu(1)z_\nu(m)\\
&= [z_\mu(1)z_\nu(m) - z_\mu(m)z_\nu(1)]^2\\
&= \omega_\mu\omega_\nu \left[(-1)^{m-\nu}\sqrt{\left|\frac{z_\mu(1)z_\nu(m)}{z_\mu(m)z_\nu(1)}\right|} - (-1)^{m-\mu}\sqrt{\left|\frac{z_\mu(m)z_\nu(1)}{z_\mu(1)z_\nu(m)}\right|}\right]^2\\
&= \omega_\mu\omega_\nu \left(\left|\frac{z_\mu(1)z_\nu(m)}{z_\mu(m)z_\nu(1)}\right| + \left|\frac{z_\mu(m)z_\nu(1)}{z_\mu(1)z_\nu(m)}\right| - 2(-1)^{\mu+\nu}\right)\\
&= \omega_\mu\omega_\nu g_{\mu\nu}, \text{ defining } g_{\mu\nu} = g_{\nu\mu}.
\end{aligned} \tag{A.4}$$

Note that $g_{\mu\nu} \geq 0$ with equality when $\frac{z_\nu(m)}{z_\nu(1)} = \frac{z_\mu(m)}{z_\mu(1)}$ and $\mu + \nu := 0 \pmod 2$. In general $g_{\mu\nu} \geq 4$ when $\mu + \nu := 1 \pmod 2$. In the persymmetric case

$$g_{\mu\nu} = \begin{cases} 0, & \mu + \nu := 0 \pmod 2, \\ 4, & \mu + \nu := 1 \pmod 2. \end{cases}$$

Thus

$$\epsilon(\zeta) = \sum_{\nu<\mu}^{\mathrm{m}}\sum^{\mathrm{m}} \frac{\omega_\mu\omega_\nu g_{\mu\nu}}{(\lambda_\mu - \zeta)(\lambda_\nu - \zeta)}. \tag{A.5}$$

Since ϵ and ρ have the same poles, α from (5.7) is well defined at λ_j.

For isolated $\lambda_j, \gamma \ll \text{gap}(\lambda_j)$,

$$\begin{aligned}
\alpha(\lambda_j) = \lim_{\zeta\to\lambda_j} \frac{1 - \gamma^2\epsilon}{-2\gamma\rho} &= \gamma\omega_j \sum_{i\neq j} \frac{\omega_i g_{ij}}{\lambda_i - \lambda_j} \div 2\omega_j(-1)^{m-j}\\
&= \frac{1}{2}\gamma(-1)^{m-j} \sum_{i\neq j} \frac{\omega_i g_{ij}}{\lambda_i - \lambda_j}.
\end{aligned} \tag{A.6}$$

Thus $|\alpha(\lambda_j)| = O(\gamma) < 1$ and λ_j is inside a special interval.

A.2. $\boldsymbol{\rho}$ and $\boldsymbol{\epsilon}$ near an isolated zero $\boldsymbol{\lambda_j}$. Throughout this analysis $\sum'$ indicates omission of one index value in the sum, usually j. For ζ near λ_j write

$$(\lambda_i - \zeta)^{-1} = (\lambda_i - \lambda_j)^{-1}\left[1 + \frac{\zeta - \lambda_j}{\lambda_i - \lambda_j} + \left(\frac{\zeta - \lambda_j}{\lambda_i - \lambda_j}\right)^2 + O\left(\left|\frac{\zeta - \lambda_j}{\text{gap}(\lambda_j)}\right|^3\right)\right]. \tag{A.7}$$

Insertion into (4.3) yields

$$\rho_{xy}(\zeta) = \frac{x_k y_k}{\lambda_j - \zeta} + \sum_{i\neq j} \frac{x_i y_i}{\lambda_i - \lambda_j} + O\left(\left|\frac{\lambda_j - \zeta}{\mathrm{gap}^2(\lambda_j)}\right|\right). \tag{A.8}$$

Hence, from (A.1),

$$\rho(\zeta) = \frac{(-1)^{m-j}\omega_j}{\lambda_j - \zeta} + R_j' + O\left(\left|\frac{\zeta - \lambda_j}{\mathrm{gap}(\lambda_j)^2}\right|\right), \quad R_j' := \sum_{i\neq j} \frac{(-1)^{m-i}\omega_i}{\lambda_i - \lambda_j}, \tag{A.9}$$

and, from (A.5),

$$\begin{aligned}
\epsilon(\zeta) &= \frac{\omega_j}{\lambda_j - \zeta} \sum{}' \frac{\omega_i g_{ij}}{\lambda_i - \lambda_j}\left[1 + \frac{\zeta - \lambda_j}{\lambda_i - \lambda_j} + O\left(\left|\frac{\zeta - \lambda_j}{\lambda_i - \lambda_j}\right|^2\right)\right] \\
&\quad + \sum_{\nu<\mu}{}' \sum{}' \frac{\omega_\mu \omega_\nu g_{\mu\nu}}{(\lambda_\mu - \lambda_j)(\lambda_\nu - \lambda_j)}\left[1 + \frac{\zeta - \lambda_j}{\lambda_\mu - \lambda_j} + O\left(\left|\frac{\lambda_j - \zeta}{\mathrm{gap}(\lambda_j)}\right|^2\right)\right] \\
&\qquad\qquad \times \left[1 + \frac{\zeta - \lambda_j}{\lambda_\mu - \lambda_j} + O\left(\left|\frac{\lambda_j - \zeta}{\mathrm{gap}(\lambda_j)}\right|^2\right)\right] \\
&= \frac{\omega_j}{\lambda_j - \zeta} E_j' + K_j' + O\left(\left|\frac{\lambda_j - \zeta}{\mathrm{gap}(\lambda_j)}\right|\right) \qquad \text{(A.10)}
\end{aligned}$$

with

$$E_j' := \sum{}' \frac{\omega_i g_{ij}}{\lambda_i - \lambda_j}, \tag{A.11a}$$

$$K_j' := \sum_{\nu<\mu}{}' \sum{}' \frac{\omega_\mu \omega_\nu g_{\mu\nu}}{(\lambda_\mu - \lambda_j)(\lambda_\nu - \lambda_j)} - \omega_j \sum{}' \frac{\omega_i g_{ij}}{(\lambda_i - \lambda_j)^2}. \tag{A.11b}$$

Note that in this case, ρ and ϵ have comparable residues at the pole λ_j and comparable constant terms. We can find ζ_0 such that $\alpha = 0$: using $1 - \gamma^2\epsilon(\zeta_0) = 0$ and the first two terms of (A.10),

$$\lambda_j - \zeta_0 = \gamma^2 \omega_j E_j' / (1 - \gamma^2 K_j'). \tag{A.12}$$

Thus $\zeta_0 = \lambda_j + O(\gamma^2\omega_j)$.

Appendix B. Analysis of S_p'', p even.

This section gives supporting material for the analysis in section 6.1. Recall the abbreviations $c_k = \cos k\psi, s_k := \sin k\psi$; then from (6.2),

$$S_p''(\alpha) := \frac{1}{p} \sum_{\substack{j\neq k-1,k\\ j=1}}^{p/2} \frac{s_j^2}{\alpha^2 - c_j^2} \text{ when } c_k < \alpha < c_{k-1}.$$

Consider the case $k = 1$, the outermost pair:

$$L_1 := \frac{1}{p}(1 + 1 + \cdots + 1) < S_p'' < \frac{1}{p}\left(\frac{4}{3} + \frac{9}{8} + \frac{16}{15} + \cdots + \frac{1}{\cos^2 \pi/p}\right) =: U_1,$$

$$\begin{aligned} U_1 &< \frac{1}{p}\left[\frac{p}{2} - 1 + \sum_2^\infty (n^2 - 1)^{-1}\right] \\ &= \frac{1}{p}\left[\frac{p}{2} - 1 + \sum_2^\infty (n^{-2} + n^{-4} + n^{-6} + \cdots)\right] \\ &= \frac{1}{p}\left[\frac{p}{2} - 1 + \left(\frac{\pi^2}{6} - 1\right) + \left(\frac{\pi^4}{90} - 1\right) + \cdots\right] \\ &\approx \frac{1}{p}\left[\frac{p}{2} - 1 + \frac{2}{3} + \frac{1}{9} + \cdots\right] \approx \frac{1}{2}, \end{aligned}$$

$$\frac{1}{2} - \frac{1}{2p} < S_p'' < \frac{1}{2}.$$

At the other extreme, $k = p/2$, on $(\cos\frac{\pi}{p}, 1)$ each term is negative:

$$\begin{aligned} L_{p/2} := \frac{1}{p}\left\{\frac{c_2^2}{s_2^2} + \frac{c_3^2}{s_3^2} + \cdots + \frac{s_1^2}{c_1^2}\right\} &< -S_p'' \\ &< \frac{1}{p}\left(\frac{c_2^2}{s_2^2 - s_1^2} + \frac{c_3^2}{s_3^2 - s_1^2} + \cdots + \frac{s_1^2}{c_1^2 - s_1^2}\right) =: U_{p/2}, \end{aligned}$$

$$\begin{aligned} L_{p/2} &\approx \frac{1}{p}\left\{\left(\frac{p^2}{4\pi^2} - 1\right) + \left(\frac{p^2}{9\pi^2} - 1\right) + \cdots \frac{\pi^2}{p^2}\right\} \\ &= \frac{1}{p}\left\{\frac{p^2}{\pi^2}\left[\frac{1}{4} + \frac{1}{9} + \cdots + \left(\frac{4}{p}\right)^2\right] - \frac{p}{4} + O(1)\right\} \\ &\approx \frac{1}{p}\left\{\frac{p^2}{\pi^2}\left(\frac{\pi^2}{6} - 1\right) - \frac{p}{4} + O(1)\right\} \approx O(p/15). \end{aligned}$$

Computer studies (for $p \leq 100$) show that as k increases, S_p'' rises gently from just below $\frac{1}{2}$ to just above $\frac{1}{2}$ at $k = p/4$ and then declines toward 0. Only at $k = p/2$ is $S_p'' < 0$.

Acknowledgments. Many thanks to the anonymous referees and the editor for their help with this manuscript.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK User's Guide*, 3rd ed., SIAM, Philadelphia, 1999.

[2] 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.

[3] J. Bunch, P. Nielsen, and D. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[4] B. Cahlon, D. M. Kulkarni, and P. Shi, *Stepwise stability for the heat equation with a nonlocal constraint*, SIAM J. Numer. Anal., 32 (1995), pp. 571–593.

[5] J. J. M. Cuppen, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[6] C. M. da Fonseca, *On the location of the eigenvalues of Jacobi matrices*, Appl. Math. Lett., 19 (2006), pp. 1168–1174.

[7] C. M. da Fonseca, *On the eigenvalues of some tridiagonal matrices*, J. Comput. Appl. Math., 200 (2007), pp. 283–286.

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

[9] J. W. Demmel, O. A. Marques, B. N. Parlett, and C. Vömel, *A Testing Infrastructure for Symmetric Tridiagonal Eigensolvers*, Technical report LBNL-61831, Lawrence Berkeley National Laboratory, Berkeley, CA, 2006.

[10] I. S. Dhillon, B. N. Parlett, and C. Vömel, *Glued matrices and the MRRR algorithm*, SIAM J. Sci. Comput., 27 (2005), pp. 496–510.

[11] W. B. Gragg and W. J. Harrod, *The numerically stable reconstruction of Jacobi spectral data*, Numer. Math., 44 (1984), pp. 317–336.

[12] R. Gregory and D. Karney, *A Collection of Matrices for Testing Computational Algorithms*, John Wiley, New York, 1969.

[13] 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.

[14] M. Hegland and J. T. Marti, *Algorithms for the reconstruction of special Jacobi matrices from their eigenvalues*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 219–228.

[15] D. Kulkarni, D. Schmidt, and S.-K. Tsui, *Eigenvalues of tridiagonal pseudo-Toeplitz matrices*, Linear Algebra Appl., 297 (1999), pp. 63–80.

[16] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.

[17] B. N. Parlett and C. Vömel, *The Spectrum of a Glued Matrix*, Technical report 579, Computer Science Department, ETH Zurich, Zurich, Switzerland, 2007.

[18] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

[19] S.-F. Xu, *On the Jacobi matrix inverse eigenvalue problem with mixed given data*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 632–639.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 1, pp. 133–153

ACCURACY OF THE JACOBI METHOD ON SCALED DIAGONALLY DOMINANT SYMMETRIC MATRICES[*]

J. MATEJAŠ[†]

Abstract. This paper proves that the two-sided Jacobi method computes the eigenvalues of the indefinite symmetric matrix to high relative accuracy, provided that the initial matrix is scaled diagonally dominant. It proves sharp eigenvalue perturbation bounds coming from a single Jacobi step and from the whole sweep defined by the serial pivot strategies.

Key words. symmetric matrices, eigenvalues, Jacobi method, scaled diagonally dominant matrices, relative accuracy

AMS subject classifications. 65F15, 65G05

DOI. 10.1137/070685993

1. Introduction and notation.

It is known (see [2]) that the two-sided Jacobi method computes the eigenvalues and eigenvectors of the positive definite symmetric matrices to high relative accuracy, as high as the initial matrix and the computer arithmetic allow. Numerical tests lead to the presumption that the same is true if the positive definite matrices are replaced with indefinite symmetric, but sufficiently almost diagonal matrices. This paper proves that the two-sided Jacobi method computes the eigenvalues of an α-scaled diagonally dominant (α-s.d.d.) symmetric matrix to high relative accuracy, provided that α is sufficiently small. Let us recall that the Hermitian matrix H is α-s.d.d., $0 \leq \alpha < 1$, with respect to a norm $||\cdot||$ if $||D^{-1}\text{off}(H)D^{-1}|| \leq \alpha$, where $D = |\text{diag}(H)|^{1/2}$. Here $\text{diag}(H)$ is the diagonal matrix with the same diagonal as H and $\text{off}(H) = H - \text{diag}(H)$ (see [1]).

This fact is not a surprise for several reasons. First, almost diagonal symmetric matrices share many properties with positive definite matrices. Say, changing the signs of all negative diagonal elements of a strictly diagonally dominant symmetric matrix makes the matrix positive definite. Then, for small enough α, α-s.d.d. symmetric matrices are well behaved. This result was proved in [1] and was later improved in [7]. This means that certain classes of small symmetric perturbations can cause only small relative errors in all eigenvalues and eigenvectors. Finally, when the Jacobi method is applied to a well-behaved symmetric positive definite matrix in finite arithmetic, it produces rounding errors which fit well into the new theory of relative perturbations. Therefore, one can expect that the Jacobi method behaves similarly for the indefinite almost diagonal symmetric matrices. Indeed, from [7, Theorem 3.1] it follows that for α-s.d.d. symmetric matrices, with small enough α, the proper measure for the relative error in the eigenvalues is $\eta = ||D^{-1}\delta H D^{-1}||_2$, where $D = |\text{diag}(H)|^{1/2}$. Together with Theorems 6 and 7 from this paper, it implies the relative accuracy of the Jacobi method on such matrices (Corollary 11).

It is known that the Jacobi method is not accurate for general indefinite symmetric matrices. If accurate eigenvalues and eigenvectors are wanted for such matrices, one can either follow the idea of Veselić to factorize the indefinite symmetric H as

[*] Received by the editors March 22, 2007; accepted for publication (in revised form) by J. L. Barlow May 12, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/simax/31-1/68599.html

[†] Faculty of Economics, University of Zagreb, Kennedyjev trg 6, 10000 Zagreb, Croatia (josip.matejas@kr.htnet.hr).

FJF^T and then apply the J-symmetric Jacobi-type method to the pair (F^TF, J) (see [10], [12], [11]), or follow the ideas from [3] to devise an SVD (singular value decomposition)-based algorithm. Differences, shortcomings, and advantages of each of these approaches are discussed in [3].

Here we show that the simple two-sided Jacobi method is accurate if the starting indefinite symmetric matrix is sufficiently almost diagonal. In particular, we show that for the α-s.d.d. H, the errors in the eigenvalues, produced during one sweep are bounded by $(2n-3)\,(258.67\,\alpha+2.47)\,u$, where u is the roundoff unit. The bound is somewhat better than the one obtained in [2] for positive definite matrices, especially if one takes into account that α tends to zero as the process advances. In addition, as is explained in Remark 12, the constant part of the bound, $2.47(2n-3)u$, can be reduced to just u if the known trick of Rutishauser [8] is used. For Jacobi to achieve this accuracy we use the stopping criterion "if $|h_{ij}| \leq 2u \cdot |h_{ii}h_{jj}|^{1/2}$, then set $h_{ij}=0$" (here u is a small threshold value, usually machine precision). This is a slightly modified version of the stopping criterion for positive definite matrices from [2], [8].

Although the presented error analysis corresponds to real arithmetic, similar results can be expected for complex Hermitian scaled diagonally dominant matrices.

Throughout this paper, we use the following notation. For any square matrix X, $\mathrm{diag}(X)$ stands for the diagonal matrix having the same diagonal as X, and $\mathrm{off}(X) = X - \mathrm{diag}(X)$ denotes the off-diagonal part of X. By $\|\cdot\|_2$ and $\|\cdot\|_F$ we denote the spectral and the Frobenius (Euclidean) matrix norm, respectively. The Euclidean vector norm is also denoted by $\|\cdot\|_2$.

This paper is organized as follows. In section 2 we present the simple code of the Jacobi algorithm which we analyze in the paper. In section 3 we derive some auxiliary accuracy results. The main accuracy results are proved in section 4.

2. The Jacobi algorithm. Let H be a symmetric matrix of order n. A single Jacobi step annihilates the pivot element at position (i,j), $i<j$, by the similarity transformation $H' = U^THU$, where U is a rotation in the plane (i,j). On the level of 2×2 pivot submatrices, we have $\widetilde{H}'_{ij} = \widetilde{U}_{ij}^T\widetilde{H}_{ij}\widetilde{U}_{ij}$, where $\widetilde{H}'_{ij} = \mathrm{diag}(h'_{ii}, h'_{jj})$,

$$(2.1)\quad \widetilde{H}_{ij} = \begin{bmatrix} h_{ii} & h_{ij} \\ h_{ji} & h_{jj} \end{bmatrix} \equiv \begin{bmatrix} a & c \\ c & b \end{bmatrix}, \quad \widetilde{U}_{ij} \equiv \begin{bmatrix} u_{ii} & u_{ij} \\ u_{ji} & u_{jj} \end{bmatrix} = \begin{bmatrix} cs & sn \\ -sn & cs \end{bmatrix}.$$

Thus, the sine and cosine of the rotation angle are denoted by sn and cs and in general $H' = (h'_{lm})$. The Jacobi algorithm has the following form (see [2]).

Algorithm 1.

```
repeat
    select the pivot pair $(i,j)$ with $i<j$        % according to the pivot strategy
    $a = h_{ii};\ b = h_{jj};\ c = h_{ij} = h_{ji}$
      % compute the Jacobi rotation
        $\xi = (b-a)/(2c)$
        $t = \mathrm{sgn}(\xi)/(|\xi| + \sqrt{1+\xi^2})$
        $\omega = \sqrt{1+t^2}$
        $cs = 1/\omega\ ;\quad sn = t/\omega$
      % update the 2 × 2 submatrix
        $d = c * t$
        $h'_{ii} = a - d\ ;\quad h'_{jj} = b + d$
        $h'_{ij} = 0;\quad h'_{ji} = 0$
    % update the rest of rows and columns $i$ and $j$
      for $k = 1$ to $n$ except $i$ and $j$
        $tmp = h_{ik}$
        $h'_{ik} = cs * tmp - sn * h_{jk}$
        $h'_{jk} = sn * tmp + cs * h_{jk}$
        $h'_{ki} = h'_{ik}\,;\quad h'_{kj} = h'_{jk}$
      endfor
until convergence
```

3. Some auxiliary accuracy results. We assume the following notation:

- u denotes the machine precision according to the IEEE standard,

$$u \in \left\{2^{-23},\ 2^{-24},\ 2^{-52},\ 2^{-53},\ 2^{-64}\right\}. \tag{3.1}$$

- ε is a quantity whose absolute value is bounded above by u. It may be sub- or superscripted. It mostly appears after one arithmetic operation.
- $\boldsymbol{\epsilon}$ is a quantity whose absolute value can be larger than u. It may be sub- or superscripted.
- η is a quantity whose absolute value is $O(u^2)$. It may be sub- or superscripted. It usually denotes the nonlinear part of the error.

We use a standard model of computer arithmetic where the floating point result $\mathrm{fl}(a \circ b)$ of the basic arithmetic operation $\circ$ is given by

$$\begin{aligned} \mathrm{fl}(a \pm b) &= (a \pm b)(1 + \varepsilon_1), & \mathrm{fl}(a \cdot b) &= (a \cdot b)(1 + \varepsilon_2), \\ \mathrm{fl}(a/b) &= (a/b)(1 + \varepsilon_3), & \mathrm{fl}(\sqrt{a}) &= \sqrt{a}(1 + \varepsilon_4), \end{aligned}$$

where, according to our notation, $|\varepsilon_i| \le u$, $i = 1, 2, 3, 4$. Here ε_i depends on the operand(s) and the operation.

In the following results we give exact error estimates without rounding or neglecting nonlinear terms. For example we shall write $(1 + \varepsilon_1)(1 + \varepsilon_2) = 1 + \boldsymbol{\epsilon}_1 + \eta_1$, where $\boldsymbol{\epsilon}_1 = \varepsilon_1 + \varepsilon_2$ stands for the linear part and $\eta_1 = \varepsilon_1 \varepsilon_2$ denotes the nonlinear part of the computed error. To make further analysis easier and clearer we also separate the error of the initial data in a similar way, say $y = (1 + \boldsymbol{\epsilon}_x + \eta_x)x$. Here $\boldsymbol{\epsilon}_x$ denotes the main (say, linear) part of the error and $\eta_x = O(u^2)$. But the results also hold if we assume that the whole error is $\boldsymbol{\epsilon}_x$ ($\eta_x = 0$) or η_x ($\boldsymbol{\epsilon}_x = 0$).

LEMMA 2. [illegible] $y = (1 + \boldsymbol{\epsilon}_x + \eta_x)x$ [illegible] $z = \sqrt{\frac{x^2}{1+x^2}}$ [illegible]

(i) $\mathrm{fl}(y^2) = (1 + \boldsymbol{\epsilon}_1 + \eta_1)x^2,$

[illegible]
$$\begin{aligned} \boldsymbol{\epsilon}_1 &= \varepsilon_1 + 2\boldsymbol{\epsilon}_x, \\ \eta_1 &= 2\eta_x + 2\varepsilon_1(\boldsymbol{\epsilon}_x + \eta_x) + (1 + \varepsilon_1)(\boldsymbol{\epsilon}_x + \eta_x)^2, \end{aligned}$$

(ii) $\mathrm{fl}(1 + y^2) = (1 + \boldsymbol{\epsilon}_2 + \eta_2)(1 + x^2),$

[illegible]
$$\begin{aligned} \boldsymbol{\epsilon}_2 &= \varepsilon_2 + z^2\boldsymbol{\epsilon}_1 = \varepsilon_2 + z^2(\varepsilon_1 + 2\boldsymbol{\epsilon}_x), \\ \eta_2 &= z^2[\eta_1 + \varepsilon_2(\boldsymbol{\epsilon}_1 + \eta_1)], \end{aligned}$$

(iii) $\mathrm{fl}\left(\sqrt{1 + y^2}\right) = (1 + \boldsymbol{\epsilon}_3 + \eta_3)\sqrt{1 + x^2},$

[illegible]
$$\begin{aligned} \boldsymbol{\epsilon}_3 &= \varepsilon_3 + \frac{\boldsymbol{\epsilon}_2}{2} = \varepsilon_3 + \frac{\varepsilon_2}{2} + z^2\left(\frac{\varepsilon_1}{2} + \boldsymbol{\epsilon}_x\right), \\ \eta_3 &= \frac{\eta_2}{2} + \frac{\boldsymbol{\epsilon}_2 + \eta_2}{2}\varepsilon_3 - \frac{(1 + \varepsilon_3)(\boldsymbol{\epsilon}_2 + \eta_2)^2}{4(1 + \sqrt{1 + \boldsymbol{\epsilon}_2 + \eta_2}) + 2(\boldsymbol{\epsilon}_2 + \eta_2)}, \end{aligned}$$

(iv) $\mathrm{fl}\left(|y|+\sqrt{1+y^2}\right)=(1+\boldsymbol{\epsilon}_4+\eta_4)\left(|x|+\sqrt{1+x^2}\right)$,

[illegible]
$$\begin{aligned}
\boldsymbol{\epsilon}_4 &= \varepsilon_4+\frac{z}{1+z}\boldsymbol{\epsilon}_x+\frac{1}{1+z}\boldsymbol{\epsilon}_3\\
&= \varepsilon_4+\frac{1}{1+z}\left(\varepsilon_3+\frac{\varepsilon_2}{2}\right)+\frac{z^2}{1+z}\cdot\frac{\varepsilon_1}{2}+z\boldsymbol{\epsilon}_x,\\
\eta_4 &= \frac{1}{1+z}\eta_3+\frac{z}{1+z}\eta_x+\frac{z(\boldsymbol{\epsilon}_x+\eta_x)+\boldsymbol{\epsilon}_3+\eta_3}{1+z}\varepsilon_4,
\end{aligned}$$

(v) $\mathrm{fl}\left(\dfrac{y}{\sqrt{1+y^2}}\right) \;=\; (1+\boldsymbol{\epsilon}_5+\eta_5)\dfrac{x}{\sqrt{1+x^2}}$,

[illegible]
$$\begin{aligned}
\boldsymbol{\epsilon}_5 &= \varepsilon_5-\boldsymbol{\epsilon}_3+\boldsymbol{\epsilon}_x=\varepsilon_5-\varepsilon_3-\frac{\varepsilon_2}{2}-\frac{z^2}{2}\varepsilon_1+\frac{1}{1+x^2}\boldsymbol{\epsilon}_x,\\
\eta_5 &= \eta_x+\varepsilon_5(\boldsymbol{\epsilon}_x+\eta_x-\boldsymbol{\epsilon}_3)-(1+\varepsilon_5)\boldsymbol{\epsilon}_3(\boldsymbol{\epsilon}_x+\eta_x)\\
&\quad+(1+\varepsilon_5)(1+\boldsymbol{\epsilon}_x+\eta_x)\left(\frac{(\boldsymbol{\epsilon}_3+\eta_3)^2}{1+\boldsymbol{\epsilon}_3+\eta_3}-\eta_3\right).
\end{aligned}$$

[illegible]. (i) This assertion follows from the equality

$$\begin{aligned}
\mathrm{fl}(y^2)=(1+\varepsilon_1)y^2 \;&=\; (1+\varepsilon_1)(1+\boldsymbol{\epsilon}_x+\eta_x)^2x^2\\
&=\left[1+\varepsilon_1+2\boldsymbol{\epsilon}_x+2\eta_x+2\varepsilon_1(\boldsymbol{\epsilon}_x+\eta_x)+(1+\varepsilon_1)(\boldsymbol{\epsilon}_x+\eta_x)^2\right]x^2.
\end{aligned}$$

(ii) Here we use the equality

$$\begin{aligned}
\mathrm{fl}(1+y^2)=(1+\varepsilon_2)[1+\mathrm{fl}(y^2)] \;&=\; (1+\varepsilon_2)[1+(1+\boldsymbol{\epsilon}_1+\eta_1)x^2]\\
&=(1+\varepsilon_2)\left[1+\frac{x^2}{1+x^2}(\boldsymbol{\epsilon}_1+\eta_1)\right](1+x^2)\\
&=\left[1+\varepsilon_2+\frac{x^2}{1+x^2}(\boldsymbol{\epsilon}_1+\eta_1)+\frac{x^2}{1+x^2}\varepsilon_2(\boldsymbol{\epsilon}_1+\eta_1)\right](1+x^2).
\end{aligned}$$

(iii) Using the identity $\sqrt{1+\tau}=1+\frac{\tau}{2}-\frac{\tau^2}{4(1+\sqrt{1+\tau})+2\tau}, \tau\geq -1,$ the assertion follows from

$$\begin{aligned}
\mathrm{fl}\left(\sqrt{1+y^2}\right) \;&=\; (1+\varepsilon_3)\sqrt{\mathrm{fl}(1+y^2)} \;=\; (1+\varepsilon_3)\sqrt{(1+\boldsymbol{\epsilon}_2+\eta_2)(1+x^2)}\\
&=(1+\varepsilon_3)\left[1+\frac{\boldsymbol{\epsilon}_2+\eta_2}{2}-\frac{(\boldsymbol{\epsilon}_2+\eta_2)^2}{4(1+\sqrt{1+\boldsymbol{\epsilon}_2+\eta_2})+2(\boldsymbol{\epsilon}_2+\eta_2)}\right]\sqrt{1+x^2}\\
&=\left[1+\varepsilon_3+\frac{\boldsymbol{\epsilon}_2}{2}+\frac{\eta_2}{2}+\frac{\boldsymbol{\epsilon}_2+\eta_2}{2}\varepsilon_3-\frac{(1+\varepsilon_3)(\boldsymbol{\epsilon}_2+\eta_2)^2}{4(1+\sqrt{1+\boldsymbol{\epsilon}_2+\eta_2})+2(\boldsymbol{\epsilon}_2+\eta_2)}\right]\sqrt{1+x^2}.
\end{aligned}$$

(iv) Here we use

$$\begin{aligned}
\mathrm{fl}\left(|y|+\sqrt{1+y^2}\right) \;&=\; (1+\varepsilon_4)\left[|y|+\mathrm{fl}\left(\sqrt{1+y^2}\right)\right]\\
&=(1+\varepsilon_4)\left[(1+\boldsymbol{\epsilon}_x+\eta_x)|x|+(1+\boldsymbol{\epsilon}_3+\eta_3)\sqrt{1+x^2}\right]
\end{aligned}$$

$$= (1+\varepsilon_4)\left[1+\frac{(\boldsymbol{\epsilon}_x+\eta_x)|x|+(\boldsymbol{\epsilon}_3+\eta_3)\sqrt{1+x^2}}{|x|+\sqrt{1+x^2}}\right](|x|+\sqrt{1+x^2})$$

$$= (1+\varepsilon_4)\left[1+\frac{z(\boldsymbol{\epsilon}_x+\eta_x)+\boldsymbol{\epsilon}_3+\eta_3}{1+z}\right](|x|+\sqrt{1+x^2})$$

$$= \left[1+\varepsilon_4+\frac{z\boldsymbol{\epsilon}_x+\boldsymbol{\epsilon}_3}{1+z}+\frac{z\eta_x+\eta_3}{1+z}+\frac{z(\boldsymbol{\epsilon}_x+\eta_x)+\boldsymbol{\epsilon}_3+\eta_3}{1+z}\varepsilon_4\right](|x|+\sqrt{1+x^2}).$$

(v) Using the identity

$$\frac{1}{1+\tau}=1-\tau+\frac{\tau^2}{1+\tau},\quad \tau\neq-1, \tag{3.2}$$

the desired assertion follows from

$$\mathrm{fl}\left(\frac{y}{\sqrt{1+y^2}}\right) = (1+\varepsilon_5)\frac{y}{\mathrm{fl}\left(\sqrt{1+y^2}\right)} = (1+\varepsilon_5)\frac{1+\boldsymbol{\epsilon}_x+\eta_x}{1+\boldsymbol{\epsilon}_3+\eta_3}\cdot\frac{x}{\sqrt{1+x^2}}$$

$$= (1+\varepsilon_5)(1+\boldsymbol{\epsilon}_x+\eta_x)\left[1-\boldsymbol{\epsilon}_3-\eta_3+\frac{(\boldsymbol{\epsilon}_3+\eta_3)^2}{1+\boldsymbol{\epsilon}_3+\eta_3}\right]\frac{x}{\sqrt{1+x^2}}$$

$$= \Bigg[(1+\varepsilon_5)(1+\boldsymbol{\epsilon}_x+\eta_x)(1-\boldsymbol{\epsilon}_3)$$

$$+ (1+\varepsilon_5)(1+\boldsymbol{\epsilon}_x+\eta_x)\left(\frac{(\boldsymbol{\epsilon}_3+\eta_3)^2}{1+\boldsymbol{\epsilon}_3+\eta_3}-\eta_3\right)\Bigg]\frac{x}{\sqrt{1+x^2}}$$

$$= \Bigg[1+\varepsilon_5-\boldsymbol{\epsilon}_3+\boldsymbol{\epsilon}_x+\eta_x+\varepsilon_5(\boldsymbol{\epsilon}_x+\eta_x-\boldsymbol{\epsilon}_3)-(1+\varepsilon_5)\boldsymbol{\epsilon}_3(\boldsymbol{\epsilon}_x+\eta_x)$$

$$+ (1+\varepsilon_5)(1+\boldsymbol{\epsilon}_x+\eta_x)\left(\frac{(\boldsymbol{\epsilon}_3+\eta_3)^2}{1+\boldsymbol{\epsilon}_3+\eta_3}-\eta_3\right)\Bigg]\frac{x}{\sqrt{1+x^2}},$$

which completes the proof. □

The error estimates for the reciprocal expressions of those in Lemma 2, i.e., for $1/y^2$, $1/(1+y^2)$, $1/\sqrt{1+y^2}$, and $1/(|y|+\sqrt{1+y^2})$, can be obtained from the following more general result.

Lemma 3. [illegible] $y=(1+\boldsymbol{\epsilon}_x+\eta_x)x$ [illegible] $\mathrm{fl}\,(F(y))=(1+\boldsymbol{\epsilon}_F+\eta_F)F(x)$ [illegible] F [illegible] $F(x)\neq0$ [illegible] $G(x)=1/F(x)$ [illegible]

$$\mathrm{fl}\,(G(y))=(1+\boldsymbol{\epsilon}_G+\eta_G)G(x),$$

[illegible] $\boldsymbol{\epsilon}_G=\varepsilon_6-\boldsymbol{\epsilon}_F, \eta_G=(1+\varepsilon_6)\big(\frac{(\boldsymbol{\epsilon}_F+\eta_F)^2}{1+\boldsymbol{\epsilon}_F+\eta_F}-\eta_F\big)-\varepsilon_6\boldsymbol{\epsilon}_F.$

[illegible]. The proof uses a similar argument as in the proof of Lemma 2(v). Using the identity (3.2), we have

$$\mathrm{fl}\left(\frac{1}{F(y)}\right) = (1+\varepsilon_6)\frac{1}{\mathrm{fl}(F(y))} = (1+\varepsilon_6)\frac{1}{1+\boldsymbol{\epsilon}_F+\eta_F}\cdot\frac{1}{F(x)}$$

$$= (1+\varepsilon_6)\left[1-\boldsymbol{\epsilon}_F-\eta_F+\frac{(\boldsymbol{\epsilon}_F+\eta_F)^2}{1+\boldsymbol{\epsilon}_F+\eta_F}\right]\frac{1}{F(x)}$$

$$= \left[1+\varepsilon_6-\boldsymbol{\epsilon}_F-\varepsilon_6\boldsymbol{\epsilon}_F+(1+\varepsilon_6)\left(\frac{(\boldsymbol{\epsilon}_F+\eta_F)^2}{1+\boldsymbol{\epsilon}_F+\eta_F}-\eta_F\right)\right]\frac{1}{F(x)},$$

which yields the desired estimate. □

4. Accuracy of the Jacobi method.

4.1. The estimates for one step. We know that the sequence of matrices $H^{(k)}$, $k \geq 0$, and $H^{(0)} = H$ are generated by the Jacobi method converges to a diagonal matrix. Thus we have $||\mathrm{off}(H^{(k)})||_F \to 0$. It means that the off-diagonal elements $h_{pq}^{(k)}, p \neq q$, become smaller in modulus. Since $\xi_k = (h_{jj}^{(k)} - h_{ii}^{(k)})/(2h_{ij}^{(k)}) = \cot 2\varphi_k$, the same is true for the rotation angle φ_k, provided that $h_{ii}^{(k)}$ and $h_{jj}^{(k)}$ (a and b in Algorithm 1) do not converge to the same limit. Thus, $|\xi_k|$ ($|\xi|$ in Algorithm 1) will increase. So, it appears desirable to bound $|\tan 2\varphi_k| = 1/|\xi_k|$ above by a fixed value τ and to express the error estimates as a function of τ. As we know that τ can be decreased with each cycle of the method, we shall be able to trace the movements of the error bounds depending on τ. We also introduce in ξ_k the initial error which is caused by previous computations.

In the following lemma we use the same notation for the errors as in Lemma 2. Since $\boldsymbol{\epsilon}_i, \eta_i, i = 1, 2, 3, 4, 5$ from Lemma 2 are obtained starting with $y = (1+\boldsymbol{\epsilon}_x+\eta_x)x$, we can consider them to be the functions, i.e., $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_i(x, \boldsymbol{\epsilon}_x, \eta_x)$, $\eta_i = \eta_i(x, \boldsymbol{\epsilon}_x, \eta_x)$, $i = 1, 2, 3, 4, 5$. In the same way we treat $\boldsymbol{\epsilon}_G$ and η_G from Lemma 3, $\boldsymbol{\epsilon}_G = \boldsymbol{\epsilon}_G(\boldsymbol{\epsilon}_F, \eta_F)$, $\eta_G = \eta_G(\boldsymbol{\epsilon}_F, \eta_F)$.

Lemma 4. [illegible] τ [illegible] $1/|\xi| \leq \tau$ [illegible] 1 [illegible] $\mathrm{fl}(\xi) = (1 + \boldsymbol{\epsilon}_\xi + \eta_\xi)\xi$ [illegible] $\phi = \min\left\{\frac{1}{2}, \frac{\tau^2}{4+\tau^2}\right\}$ [illegible]

$$\mathrm{fl}(t) = (1 + \boldsymbol{\epsilon}_t + \eta_t)t, \quad \mathrm{fl}(cs) = (1 + \boldsymbol{\epsilon}_{cs} + \eta_{cs})cs, \quad \mathrm{fl}(sn) = (1 + \boldsymbol{\epsilon}_{sn} + \eta_{sn})sn,$$

[illegible]

$$\text{(i)}\quad |\boldsymbol{\epsilon}_t| \leq (3 + \phi)\, u + |\boldsymbol{\epsilon}_\xi| \leq \frac{7}{2} u + |\boldsymbol{\epsilon}_\xi|, \qquad \eta_t = \eta_G\,(\, \boldsymbol{\epsilon}_4(\xi, \boldsymbol{\epsilon}_\xi, \eta_\xi),\, \eta_4(\xi, \boldsymbol{\epsilon}_\xi, \eta_\xi)\,),$$

$$\text{(ii)}\quad |\boldsymbol{\epsilon}_{cs}| \leq \left[\frac{5}{2} + \frac{\tau^2}{4 + 2\tau^2}\left(\frac{7}{2} + \phi\right)\right] u + \frac{\tau^2}{4 + 2\tau^2}|\boldsymbol{\epsilon}_\xi| \leq \frac{9}{2} u + \frac{1}{2}|\boldsymbol{\epsilon}_\xi|,$$

$$\eta_{cs} = \eta_G\,(\, \boldsymbol{\epsilon}_3(t, \boldsymbol{\epsilon}_t, \eta_t),\, \eta_3(t, \boldsymbol{\epsilon}_t, \eta_t)\,),$$

$$\text{(iii)}\quad |\boldsymbol{\epsilon}_{sn}| \leq \left(\frac{11}{2} + \phi\right) u + |\boldsymbol{\epsilon}_\xi| \leq 6u + |\boldsymbol{\epsilon}_\xi|, \qquad \eta_{sn} = \eta_5(t, \boldsymbol{\epsilon}_t,\, \eta_t).$$

[illegible] $\boldsymbol{\epsilon}_3$, $\boldsymbol{\epsilon}_4$, η_3, η_4, η_5 [illegible] η_G [illegible] 2 [illegible] 3 [illegible]

[illegible] (i) Let $\nu = \nu(\xi) = |\xi| + \sqrt{1 + \xi^2}$. Using Lemma 2(iv) with $\boldsymbol{\epsilon}_x = \boldsymbol{\epsilon}_\xi$ and $\eta_x = \eta_\xi$, we have $\mathrm{fl}(\nu) = (1 + \boldsymbol{\epsilon}_\nu + \eta_\nu)\nu$, where

$$\boldsymbol{\epsilon}_\nu = \boldsymbol{\epsilon}_4(\xi, \boldsymbol{\epsilon}_\xi, \eta_\xi) = \varepsilon_4 + \frac{1}{1+z}\left(\varepsilon_3 + \frac{\varepsilon_2}{2}\right) + \frac{z^2}{1+z} \cdot \frac{\varepsilon_1}{2} + z\boldsymbol{\epsilon}_\xi, \quad z^2 = \frac{\xi^2}{1+\xi^2},$$

$$\eta_\nu = \eta_4(\xi, \boldsymbol{\epsilon}_\xi, \eta_\xi).$$

Now we have

$$|\boldsymbol{\epsilon}_\nu| \leq u + \frac{1}{1+z} \cdot \frac{3}{2}u + \frac{z^2}{1+z} \cdot \frac{u}{2} + |z\boldsymbol{\epsilon}_\xi| = \left(1 + z + \frac{4}{1+z}\right)\frac{u}{2} + |z\boldsymbol{\epsilon}_\xi|. \tag{4.1}$$

Since $0 \leq z \leq 1$, the function $f(z) = 1 + z + \frac{4}{1+z}$ attains the maximum at $z = 0$ and $f(0) = 5$. Thus we obtain

$$|\boldsymbol{\epsilon}_\nu| \leq \frac{5}{2}u + |\boldsymbol{\epsilon}_\xi| = \left(2 + \frac{1}{2}\right)u + |\boldsymbol{\epsilon}_\xi|. \tag{4.2}$$

Since $z^2 = \frac{1}{1+1/\xi^2} \geq \frac{1}{1+\tau^2}$, we have $z \geq \frac{1}{\sqrt{1+\tau^2}} \geq \frac{1}{1+\frac{\tau^2}{2}} = \frac{2}{2+\tau^2}$ and, consequently, $\frac{1}{1+z} \leq \frac{1}{1+\frac{2}{2+\tau^2}} = \frac{2+\tau^2}{4+\tau^2}$. Thus, from the inequality (4.1), we have

$$|\boldsymbol{\epsilon}_\nu| \;\leq\; \left(1+1+4\cdot\frac{2+\tau^2}{4+\tau^2}\right)\frac{u}{2} + |\boldsymbol{\epsilon}_\xi| \;=\; \left(2+\frac{\tau^2}{4+\tau^2}\right)u + |\boldsymbol{\epsilon}_\xi|. \tag{4.3}$$

Using now (4.2) and (4.3), since $t = t(\xi) = 1/\nu(\xi)$, the desired estimates for $|\boldsymbol{\epsilon}_t|$ and η_t are obtained by applying Lemma 3 with $F(x)$ replaced by $\nu(\xi)$.

(ii) Let $\operatorname{fl}(\omega) = (1+\boldsymbol{\epsilon}_\omega+\eta_\omega)\omega$, where $\omega = \omega(t) = \sqrt{1+t^2}$. Using Lemma 2(iii) with $\boldsymbol{\epsilon}_x = \boldsymbol{\epsilon}_t$ and $\eta_x = \eta_t$, we obtain

$$\boldsymbol{\epsilon}_\omega = \boldsymbol{\epsilon}_3(t,\boldsymbol{\epsilon}_t,\eta_t) \;=\; \varepsilon_3 + \frac{\varepsilon_2}{2} + z^2\left(\frac{\varepsilon_1}{2}+\boldsymbol{\epsilon}_t\right), \quad z^2 = \frac{t^2}{1+t^2}, \tag{4.4}$$

$$\eta_\omega = \eta_3(t,\boldsymbol{\epsilon}_t,\eta_t).$$

We first estimate z^2 to be

$$z^2 = \frac{1}{1+\frac{1}{t^2}} \;=\; \frac{1}{1+\left(|\xi|+\sqrt{1+\xi^2}\right)^2} \;\leq\; \frac{1}{1+\left(\frac{1}{\tau}+\sqrt{1+\frac{1}{\tau^2}}\right)^2}$$

$$= \frac{1}{2+\frac{2}{\tau^2}+\frac{2}{\tau^2}\sqrt{1+\tau^2}} \;=\; \frac{\tau^2}{2\left(1+\tau^2+\sqrt{1+\tau^2}\right)} \;\leq\; \frac{\tau^2}{2\left(1+\tau^2+1\right)}$$

$$= \frac{\tau^2}{4+2\tau^2}, \tag{4.5}$$

which together with (4.4) and the assertion (i) yields

$$\begin{aligned}
|\boldsymbol{\epsilon}_\omega| &\leq u + \frac{u}{2} + \frac{\tau^2}{4+2\tau^2}\left[\frac{u}{2} + (3+\phi)\,u + |\boldsymbol{\epsilon}_\xi|\right] \\
&= \left[\frac{3}{2} + \frac{\tau^2}{4+2\tau^2}\left(\frac{7}{2}+\phi\right)\right]u + \frac{\tau^2}{4+2\tau^2}|\boldsymbol{\epsilon}_\xi| \\
&\leq \left[\frac{3}{2} + \frac{2\tau^2}{2+\tau^2}\right]u + \frac{\tau^2}{4+2\tau^2}|\boldsymbol{\epsilon}_\xi| \;\leq\; \frac{7}{2}\,u + \frac{1}{2}\,|\boldsymbol{\epsilon}_\xi|.
\end{aligned}$$

Since $cs = cs(t) = 1/\omega(t)$, to obtain the estimates for $|\boldsymbol{\epsilon}_{cs}|$ and η_{cs}, we apply Lemma 3 for $\omega(t)$.

(iii) We have $sn = sn(t) = t/\omega(t) = t/\sqrt{1+t^2}$. Using Lemma 2(v) we obtain

$$\boldsymbol{\epsilon}_{sn} = \boldsymbol{\epsilon}_5(t,\boldsymbol{\epsilon}_t,\eta_t) \;=\; \varepsilon_5 - \varepsilon_3 - \frac{\varepsilon_2}{2} - \frac{z^2}{2}\varepsilon_1 + \frac{1}{1+t^2}\boldsymbol{\epsilon}_t, \quad z^2 = \frac{t^2}{1+t^2},$$

$$\eta_{sn} = \eta_5(t,\boldsymbol{\epsilon}_t,\eta_t).$$

Using the assertion (i) we have

$$\begin{aligned}
|\boldsymbol{\epsilon}_{sn}| &\leq u + u + \frac{u}{2} + \frac{u}{2}\cdot\frac{t^2}{1+t^2} + |\boldsymbol{\epsilon}_t|\cdot\frac{1}{1+t^2} \;\leq\; \frac{5}{2}\,u + \max\left\{\frac{u}{2},\, |\boldsymbol{\epsilon}_t|\right\} \\
&\leq \frac{5}{2}\,u + (3+\phi)u + |\boldsymbol{\epsilon}_\xi| \;=\; \left(\frac{11}{2}+\phi\right)u + |\boldsymbol{\epsilon}_\xi| \;\leq\; 6u + |\boldsymbol{\epsilon}_\xi|,
\end{aligned}$$

which completes the proof. □

We can see from Lemma 4 that the accumulation of the errors in t, cs, and sn decreases when the rotation angle becomes small enough. Since $\boldsymbol{\epsilon}_\xi$ is the main (linear) part of the error that is caused by the previous computations, the current computation increases the error of t for at most $3u + O(\tau^2 u)$, the error of cs for $2.5u + O(\tau^2 u)$, and the error of sn for $5.5u + O(\tau^2 u)$. Note also that the initial error $\boldsymbol{\epsilon}_\xi$ appears in $\boldsymbol{\epsilon}_{cs}$ with a factor less than $\tau^2/2$.

For further analysis we shall now compute the general error bounds from Lemmas 4 and 2 by using the relation (3.1).

Lemma 5. [illegible] (3.1) [illegible] 1 [illegible]

(i) $\mathrm{fl}(\xi) = (1 + \boldsymbol{\epsilon}_\xi + \eta_\xi)\xi$, [illegible]
$|\boldsymbol{\epsilon}_\xi| \le 2u, \quad |\eta_\xi| \le u^2 < 0.000\,000\,12\,u,$

(ii) $\mathrm{fl}(t) = (1 + \boldsymbol{\epsilon}_t + \eta_t)t$, [illegible]
$|\boldsymbol{\epsilon}_t| \le 5.5u, \quad |\eta_t| < 43.750\,046\,7\,u^2 < 0.000\,005\,3\,u,$

(iii) $\mathrm{fl}(cs) = (1 + \boldsymbol{\epsilon}_{cs} + \eta_{cs})cs$, [illegible]
$|\boldsymbol{\epsilon}_{cs}| \le 5.5\,u, \quad |\eta_{cs}| < 69.562\,743\,u^2 < 0.000\,008\,3\,u,$

(iv) $\mathrm{fl}(sn) = (1 + \boldsymbol{\epsilon}_{sn} + \eta_{sn})sn$, [illegible]
$|\boldsymbol{\epsilon}_{sn}| \le 8u, \quad |\eta_{sn}| < 143.562\,755\,3\,u^2 < 0.000\,017\,2\,u.$

[illegible]. In this proof we shall use the assumption $u \le 2^{-23} < 1.192\,092\,897 \cdot 10^{-7}$ which is seen in (3.1).

(i) We have $\mathrm{fl}(\xi) = (1 + \varepsilon_1)(1 + \varepsilon_2)\frac{b-a}{2c} = [1 + (\varepsilon_1 + \varepsilon_2) + \varepsilon_1\varepsilon_2] \cdot \xi$, and thus $\boldsymbol{\epsilon}_\xi = \varepsilon_1 + \varepsilon_2$, $\eta_\xi = \varepsilon_1\varepsilon_2$, which yields the assertion.

(ii) The bound for $|\boldsymbol{\epsilon}_t|$ is easily obtained using (i) and Lemma 4(i). To bound $|\eta_t|$, we first use (i) and Lemma 2(i)–(iv) with $y = (1 + \boldsymbol{\epsilon}_\xi + \eta_\xi)\xi$. We obtain $|\boldsymbol{\epsilon}_1| \le 5u$, $|\eta_1| < 10.0000012u^2$; $|\boldsymbol{\epsilon}_2| \le 6u$, $|\eta_2| < 15.0000024u^2$; $|\boldsymbol{\epsilon}_3| \le 4u$, $|\eta_3| < 15.0000089u^2$; $|\boldsymbol{\epsilon}_4| \le 4.5u$, $|\eta_4| < 19.0000107u^2$. Then we apply Lemma 3 with $\boldsymbol{\epsilon}_F = \boldsymbol{\epsilon}_4$, $\eta_F = \eta_4$ and we obtain the bound for $|\eta_t|$ ($\eta_t = \eta_G$).

(iii) Similarly, (i) and Lemma 4(ii) yield the bound for $|\boldsymbol{\epsilon}_{cs}|$. To bound $|\eta_{cs}|$, we first use (ii) and Lemma 2(i)–(iii) with $y = (1 + \boldsymbol{\epsilon}_t + \eta_t)t$. Note that here $z^2 = t^2/(1 + t^2)$, and according to the inequality (4.5) we have $z^2 < \tau^2/(4 + 2\tau^2) < 1/2$. We then obtain $|\boldsymbol{\epsilon}_1| \le 12u$, $|\eta_1| < 128.750165u^2$; $|\boldsymbol{\epsilon}_2| \le 7u$, $|\eta_2| < 70.375091u^2$; $|\boldsymbol{\epsilon}_3| \le 4.5u$, $|\eta_3| < 44.8125673u^2$. Finally, $|\eta_{cs}|$ is bounded by Lemma 3 with $\boldsymbol{\epsilon}_F = \boldsymbol{\epsilon}_3$, $\eta_F = \eta_3$, and $\eta_{cs} = \eta_G$.

(iv) The assertion (i) and Lemma 4(iii) yield the bound for $|\boldsymbol{\epsilon}_{sn}|$. To obtain the bound for $|\eta_{sn}|$, we use the bounds for $|\boldsymbol{\epsilon}_1|, \ldots, |\eta_3|$ from the proof of (iii) and then apply Lemma 2(v). □

We prove now the main accuracy result. It shows that [7, Theorem 3.1] can be applied.

Theorem 6. [illegible] H [illegible] n [illegible] H' [illegible] $\mathrm{fl}(H')$ [illegible] H [illegible] 1 [illegible] (i,j) [illegible] δH [illegible] $\mathrm{fl}(H')$ [illegible] $H + \delta H$ [illegible] $D = |\mathrm{diag}(H)|^{1/2}$ $A = D^{-1}HD^{-1}$

$\delta A = D^{-1}\delta H D^{-1}$ [illegible]

$$\alpha_{ij} = \sqrt{2\sum_{\substack{k=1\\k\neq i,j}}^{n} a_{ik}^2 + 2\sum_{\substack{k=1\\k\neq i,j}}^{n} a_{jk}^2 + 2a_{ij}^2} = ||E_{ij}\mathrm{off}(A) + \mathrm{off}(A)E_{ij} - E_{ij}\mathrm{off}(A)E_{ij}||_F,$$

[illegible] $E_{ij} = e_i e_i^T + e_j e_j^T$ [illegible] $\alpha_{ij} < 1,$ [illegible]

$$||\delta A||_F < (22.54\alpha_{ij} + 2.51)\,u,$$

[illegible] u [illegible] (3.1)

[illegible]. The idea of the proof is based on the proof of [2, Theorem 3.1] but includes significant improvements. Let

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} \equiv \begin{bmatrix} \mathrm{sgn}(h_{ii})d_i^2 & a_{ij}d_id_j \\ a_{ij}d_id_j & \mathrm{sgn}(h_{jj})d_j^2 \end{bmatrix},$$

where $d_i = \sqrt{|h_{ii}|}$, $d_j = \sqrt{|h_{jj}|}$. Note that $a_{ij} = h_{ij}/\sqrt{|h_{ii}h_{jj}|}$ is an element of the scaled matrix A. According to our assumption we have

$$a_{ij}^2 \;\le\; \frac{\alpha_{ij}^2}{2} \;<\; \frac{1}{2}. \tag{4.6}$$

We consider first the part of Algorithm 1 that is indicated by "update the rest of rows and columns i and j." We have

$$\begin{aligned} \mathrm{fl}(h'_{ik}) = (1+\varepsilon_1)[(1+\varepsilon_2)(1+\boldsymbol{\epsilon}_{cs}+\eta_{cs})\cdot cs\cdot h_{ik} \\ -(1+\varepsilon_3)(1+\boldsymbol{\epsilon}_{sn}+\eta_{sn})\cdot sn\cdot h_{jk}] \;=\; h'_{ik}+\delta h'_{ik}, \quad k\neq i,j, \end{aligned}$$

where

$$\begin{aligned} \delta h'_{ik} &= [(\boldsymbol{\epsilon}_{cs}+\eta_{cs})(1+\varepsilon_1)(1+\varepsilon_2)+\varepsilon_1+\varepsilon_2+\varepsilon_1\varepsilon_2]\cdot cs\cdot h_{ik} \\ &\quad -[(\boldsymbol{\epsilon}_{sn}+\eta_{sn})(1+\varepsilon_1)(1+\varepsilon_3)+\varepsilon_1+\varepsilon_3+\varepsilon_1\varepsilon_3]\cdot sn\cdot h_{jk}. \end{aligned} \tag{4.7}$$

Thus we have

$$|\delta h'_{ik}| \le \mu_1|cs|\cdot|h_{ik}| + \mu_2|sn|\cdot|h_{jk}|, \quad k\neq i,j, \tag{4.8}$$

where

$$\mu_1 = |\boldsymbol{\epsilon}_{cs}+\eta_{cs}|(1+u)^2+2u+u^2, \qquad \mu_2 = |\boldsymbol{\epsilon}_{sn}+\eta_{sn}|(1+u)^2+2u+u^2. \tag{4.9}$$

Similarly, we obtain $\mathrm{fl}(h'_{jk}) = h'_{jk}+\delta h'_{jk},\ k\neq i,j$, where

$$|\delta h'_{jk}| \le \mu_1|cs|\cdot|h_{jk}| + \mu_2|sn|\cdot|h_{ik}|, \quad k\neq i,j. \tag{4.10}$$

To obtain the backward error, we write

$$\begin{aligned} \begin{bmatrix} \mathrm{fl}(h'_{ik}) \\ \mathrm{fl}(h'_{jk}) \end{bmatrix} &= \widetilde{U}^T \begin{bmatrix} h_{ik} \\ h_{jk} \end{bmatrix} + \begin{bmatrix} \delta h'_{ik} \\ \delta h'_{jk} \end{bmatrix} \;=\; \widetilde{U}^T\left(\begin{bmatrix} h_{ik} \\ h_{jk} \end{bmatrix} + \widetilde{U}\begin{bmatrix} \delta h'_{ik} \\ \delta h'_{jk} \end{bmatrix}\right) \\ &= \widetilde{U}^T\left(\begin{bmatrix} h_{ik} \\ h_{jk} \end{bmatrix} + \begin{bmatrix} \delta h_{ik} \\ \delta h_{jk} \end{bmatrix}\right), \end{aligned}$$

where $\delta h_{ik} = cs\cdot \delta h'_{ik} + sn\cdot \delta h'_{jk}$, $\delta h_{jk} = cs\cdot \delta h'_{jk} - sn\cdot \delta h'_{ik}$, and $\widetilde{U}$ is defined by (2.1). This together with the relations (4.8) and (4.10) yields

$$|\delta a_{ik}| = \frac{|\delta h_{ik}|}{d_i d_k} \le \mu_1\, cs^2 |a_{ik}| + \mu_2\, |cs|\,|sn|\,|a_{jk}|\,\frac{d_j}{d_i} + \mu_1\, |sn|\,|cs|\,|a_{jk}|\,\frac{d_j}{d_i} + \mu_2\, sn^2 |a_{ik}|,$$

$$|\delta a_{jk}| = \frac{|\delta h_{jk}|}{d_j d_k} \le \mu_1\, cs^2 |a_{jk}| + \mu_2\, |cs|\,|sn|\,|a_{ik}|\,\frac{d_i}{d_j} + \mu_1\, |sn|\,|cs|\,|a_{ik}|\,\frac{d_i}{d_j} + \mu_2\, sn^2 |a_{jk}|.$$

Let $q = d_j/d_i \le 1$, else we set $q = d_i/d_j$. Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned} (\delta a_{ik})^2 &\le \left[(\mu_1\, cs^2 + \mu_2\, sn^2)|a_{ik}| + (\mu_1+\mu_2)|cs|\,|sn|\,|a_{jk}|\,q\right]^2 \\ &\le \left[(\mu_1\, cs^2 + \mu_2\, sn^2)^2 + (\mu_1+\mu_2)^2 cs^2 sn^2 q^2\right]\cdot (a_{ik}^2 + a_{jk}^2) \\ &\le \left[\mu^2 + (\mu_1+\mu_2)^2 \cdot \frac{1}{4}\sin^2 2\varphi \cdot q^2\right]\cdot (a_{ik}^2 + a_{jk}^2), \end{aligned} \tag{4.11}$$

where φ is the rotation angle ($\tan 2\varphi = 2c/(b-a)$) and $\mu = \max\{\mu_1,\, \mu_2\}$. Similarly, we obtain

$$(\delta a_{jk})^2 \le \left[\mu^2 + (\mu_1+\mu_2)^2 \cdot \frac{1}{4}\sin^2 2\varphi \cdot \frac{1}{q^2}\right]\cdot (a_{ik}^2 + a_{jk}^2). \tag{4.12}$$

We now have

$$|\xi| = \left|\frac{b-a}{2c}\right| = \left|\frac{\operatorname{sgn}(h_{jj})d_j^2 - \operatorname{sgn}(h_{ii})d_i^2}{2a_{ij}d_i d_j}\right| \ge \left|\frac{d_j^2 - d_i^2}{2a_{ij}d_i d_j}\right| = \frac{1-q^2}{2|a_{ij}|q} \tag{4.13}$$

and

$$\frac{1}{4}\sin^2 2\varphi \le \frac{1}{4}\tan^2 2\varphi = \frac{1}{4\xi^2} \le \frac{a_{ij}^2 q^2}{(1-q^2)^2}.$$

Since $q \le 1$, we have $\frac{1}{4}\sin^2 2\varphi \cdot q^2 \le \frac{1}{4}\cdot 1 \cdot 1 = 0.25$ and

$$\frac{1}{4}\sin^2 2\varphi \cdot \frac{1}{q^2} \le \begin{cases} \dfrac{a_{ij}^2}{(1-q^2)^2} < \dfrac{1}{2(1-q^2)^2} < 0.933013 & \text{for } q \le \dfrac{\sqrt{6}-\sqrt{2}}{2}, \\[2ex] \dfrac{1}{4q^2} < 0.933013 & \text{for } \dfrac{\sqrt{6}-\sqrt{2}}{2} \le q \le 1. \end{cases}$$

Here we have used the assumption (4.6). Thus, we have obtained

$$\frac{1}{4}\sin^2 2\varphi \cdot q^2 \le 0.25, \qquad \frac{1}{4}\sin^2 2\varphi \cdot \frac{1}{q^2} \le 0.933\,013. \tag{4.14}$$

If we apply these bounds in the inequalities (4.11) and (4.12), then we obtain

$$(\delta a_{ik})^2 + (\delta a_{jk})^2 < \left[2\mu^2 + 1.183013(\mu_1+\mu_2)^2\right]\cdot (a_{ik}^2 + a_{jk}^2), \quad k \ne i, j. \tag{4.15}$$

Note that the same inequality holds for $(\delta a_{ki})^2 + (\delta a_{kj})^2$. We proceed now by considering the part of Algorithm 1 that is indicated by "update the 2×2 submatrix." We have

$$\begin{aligned} \operatorname{fl}(h'_{ii}) &= (1+\varepsilon_4)[a - (1+\varepsilon_5)(1+\boldsymbol{\epsilon}_t+\eta_t)\cdot c\cdot t] \\ &= (1+\varepsilon_4)[h_{ii} - (1+\varepsilon_5)(1+\boldsymbol{\epsilon}_t+\eta_t)\cdot h_{ij}t] \;=\; h'_{ii} + \delta h'_{ii}, \end{aligned}$$

where $\delta h'_{ii} = \varepsilon_4 h_{ii} - [(\boldsymbol{\epsilon}_t + \eta_t)(1+\varepsilon_4)(1+\varepsilon_5) + \varepsilon_4 + \varepsilon_5 + \varepsilon_4\varepsilon_5] \cdot h_{ij} t\,.$
In the same way, we obtain $\mathrm{fl}(h'_{jj}) = h'_{jj} + \delta h'_{jj}$, where

$$\delta h'_{jj} = \varepsilon_6 h_{jj} + [(\boldsymbol{\epsilon}_t + \eta_t)(1+\varepsilon_6)(1+\varepsilon_7) + \varepsilon_6 + \varepsilon_7 + \varepsilon_6\varepsilon_7] \cdot h_{ij} t.$$

Thus we have

$$|\delta h'_{ii}| \le |h_{ii}| u + \mu_3 |h_{ij}| \cdot t, \qquad |\delta h'_{jj}| \le |h_{jj}| u + \mu_3 |h_{ij}| \cdot t, \tag{4.16}$$

where

$$\mu_3 = |\boldsymbol{\epsilon}_t + \eta_t|(1+u)^2 + 2u + u^2. \tag{4.17}$$

To obtain the backward error, we have

$$\widetilde{U}^T \left(\begin{bmatrix} h_{ii} & h_{ij} \\ h_{ij} & h_{jj} \end{bmatrix} + \begin{bmatrix} \delta h_{ii} & \delta h_{ij} \\ \delta h_{ij} & \delta h_{jj} \end{bmatrix} \right) \widetilde{U} = \begin{bmatrix} h'_{ii} & 0 \\ 0 & h'_{jj} \end{bmatrix} + \begin{bmatrix} \delta h'_{ii} & 0 \\ 0 & \delta h'_{jj} \end{bmatrix}$$

and thus

$$\begin{aligned} \begin{bmatrix} \delta h_{ii} & \delta h_{ij} \\ \delta h_{ij} & \delta h_{jj} \end{bmatrix} &= \widetilde{U} \begin{bmatrix} \delta h'_{ii} & 0 \\ 0 & \delta h'_{jj} \end{bmatrix} \widetilde{U}^T \\ &= \begin{bmatrix} cs^2 \cdot \delta h'_{ii} + sn^2 \cdot \delta h'_{jj} & sn \cdot cs \cdot (\delta h'_{jj} - \delta h'_{ii}) \\ sn \cdot cs \cdot (\delta h'_{jj} - \delta h'_{ii}) & cs^2 \cdot \delta h'_{jj} + sn^2 \cdot \delta h'_{ii} \end{bmatrix}. \end{aligned}$$

Using now the inequality (4.16), we obtain

$$|\delta a_{ii}| = \frac{|\delta h_{ii}|}{d_i^2} \le cs^2 \left(u + \mu_3 |t| \cdot |a_{ij}| \, \frac{d_j}{d_i} \right) + sn^2 \left(\frac{d_j^2}{d_i^2} u + \mu_3 |t| \cdot |a_{ij}| \, \frac{d_j}{d_i} \right)$$

$$= (cs^2 + sn^2 \cdot q^2)\, u + \mu_3 |a_{ij}| \cdot |t| q, \tag{4.18}$$

$$|\delta a_{jj}| = \frac{|\delta h_{jj}|}{d_j^2} \le cs^2 \left(u + \mu_3 |t| \cdot |a_{ij}| \, \frac{d_i}{d_j} \right) + sn^2 \left(\frac{d_i^2}{d_j^2} u + \mu_3 |t| \cdot |a_{ij}| \, \frac{d_i}{d_j} \right)$$

$$= \left(cs^2 + sn^2 \cdot \frac{1}{q^2} \right) u + \mu_3 |a_{ij}| \cdot |t| \frac{1}{q}, \tag{4.19}$$

$$|\delta a_{ij}| = \frac{|\delta h_{ij}|}{d_i d_j} \le |cs \cdot sn| \left(\frac{d_i}{d_j} u + \mu_3 |t| \cdot |a_{ij}| + \frac{d_j}{d_i} u + \mu_3 |t| \cdot |a_{ij}| \right)$$

$$= |cs| \cdot |sn| \left(q + \frac{1}{q} \right) u + 2\mu_3 \cdot sn^2 |a_{ij}|. \tag{4.20}$$

Note that the change $q \to 1/q$ interchange the bounds for $|\delta a_{ii}|$ and $|\delta a_{jj}|$. We shall now estimate the terms in the relations (4.18)–(4.20). First, from (4.13), we deduce

$$|sn| \le |t| \le \frac{1}{2} |\tan 2\varphi| = \frac{1}{2|\xi|} \le \frac{|a_{ij}| q}{1 - q^2}, \tag{4.21}$$

and thus

$$|sn| \, \frac{1}{q} \le |t| \, \frac{1}{q} \le \begin{cases} \dfrac{|a_{ij}|}{1-q^2} \le \dfrac{\sqrt{2}}{2(1-q^2)} \le \sqrt{2} \ \text{ for } \ q \le \dfrac{\sqrt{2}}{2}, \\[2ex] \dfrac{1}{q} \le \sqrt{2} \ \text{ for } \ \dfrac{\sqrt{2}}{2} \le q \le 1. \end{cases} \tag{4.22}$$

Now we have

$$cs^2 + sn^2 q^2 \le cs^2 + sn^2 \;=\; 1,$$

$$cs^2 + sn^2 \cdot \frac{1}{q^2} = 1 - sn^2 + sn^2 \cdot \frac{1}{q^2} \;=\; 1 + sn^2 \cdot \frac{1-q^2}{q^2}$$

$$\le 1 + |sn| \frac{1}{q} \cdot |a_{ij}| \;\le\; 1 + \sqrt{2} \cdot |a_{ij}|\,,$$

$$|cs| \cdot |sn| \left(q + \frac{1}{q} \right) = \frac{1}{2} |\sin 2\varphi|\, q + \frac{1}{2} |\sin 2\varphi| \cdot \frac{1}{q} \;\le\; 0.5 + \sqrt{0.933\,013} \;<\; 1.465\,926,$$

where we have used the inequalities (4.21) and (4.14). Note also that $|t|q \le 1$ and $sn^2 \le 0.5$ because $\varphi \in [-\pi/4,\, \pi/4]$. If we apply these bounds in the relations (4.18)–(4.20), then we obtain

$$\begin{gathered} |\delta a_{ii}| \le u + \mu_3 |a_{ij}|, \qquad |\delta a_{jj}| \le u + \sqrt{2}(\mu_3 + u)|a_{ij}|, \\ |\delta a_{ij}| \le 1.465\,926\, u + \mu_3 |a_{ij}|. \end{gathered} \tag{4.23}$$

The obtained expressions (4.15) and (4.23) are the estimates for the elements of the matrix δA. Note that the only nonzero elements of δA are those in ith, jth row, and column. If we want to compute the upper bound for the norm of δA, then we need to estimate μ_1, μ_2, μ, and μ_3. The computation will be performed to, at least, six significant digits. Using the relation (4.9) together with Lemma 5(iii), (iv) and then the relation (4.17) together with Lemma 5(ii), we obtain

$$\mu_1 < 7.500\,009\,74\, u, \quad \mu_2 < 10.000\,019\,24\, u, \quad \mu_3 < 7.500\,006\,74\, u. \tag{4.24}$$

The bound for μ_2 is also the bound for $\mu = \max\{\mu_1,\, \mu_2\}$. We apply these bounds in (4.23):

$$\begin{aligned} &(\delta a_{ii})^2 + (\delta a_{jj})^2 + 2(\delta a_{ij})^2 \\ &\quad \le 6.297\,878\,076\, u^2 + (10.692\,131\,13\, \mu_3 + 2\sqrt{2}\, u)\, |a_{ij}|\, u + [3\mu_3^2 + 2(\mu_3 + u)^2]\, a_{ij}^2 \\ &\quad \le \left[6.297\,878\,076 + 83.019\,482\,67\, |a_{ij}| + 313.250\,532\,6\, a_{ij}^2\right] u^2, \end{aligned}$$

and then in (4.15):

$$(\delta a_{ik})^2 + (\delta a_{jk})^2 \le 562.299\,700\,9\, (a_{ik}^2 + a_{jk}^2)\, u^2, \quad k \ne i, j. \tag{4.25}$$

Note again that the same inequality holds for $(\delta a_{ki})^2 + (\delta a_{kj})^2$.
Since $\sum_{\substack{k=1 \\ k \ne i,j}}^{n} (2a_{ik}^2 + 2a_{jk}^2) = \alpha_{ij}^2 - 2a_{ij}^2$, $|a_{ij}| \le \alpha_{ij}/\sqrt{2}$ and $\alpha_{ij}^2 \le \alpha_{ij}$, we have

$$\begin{aligned} ||\delta A||_F^2 &= 2 \sum_{\substack{k=1 \\ k \ne i,j}}^{n} \left[(\delta a_{ik})^2 + (\delta a_{jk})^2 \right] + (\delta a_{ii})^2 + (\delta a_{jj})^2 + 2(\delta a_{ij})^2 \\ &\le \Bigg[562.299\,700\,9 \sum_{\substack{k=1 \\ k \ne i,j}}^{n} (2a_{ik}^2 + 2a_{jk}^2) \\ &\qquad + 6.297\,878\,076 + 83.019\,482\,67\, |a_{ij}| + 313.250\,532\,6\, a_{ij}^2 \Bigg] u^2 \end{aligned}$$

$$
\begin{aligned}
&\le [\, 562.299\,700\,9\, \alpha_{ij}^2 + 83.019\,482\,67|a_{ij}| + 6.297\,878\,076 \,]\, u^2 \\
&\le [507.890\,439\,6\, \alpha_{ij}^2 + 54.409\,261\,29\, \alpha_{ij}^2 + 58.703\,639\,2\, \alpha_{ij} + 6.297\,878\,076 \,]\, u^2 \\
&\le [507.890\,439\,6\, \alpha_{ij}^2 + 113.112\,900\,5\, \alpha_{ij} + 6.297\,878\,076 \,]\, u^2 \\
&= [22.536\,424\,73\, \alpha_{ij} + 2.509\,557\,347]^2 \cdot u^2,
\end{aligned}
$$

which completes the proof. □

We can see that Theorem 6 gives a much sharper bound for $||\delta A||_F$ than [2, Theorem 3.1] that of positive definite matrices. The assumption of Theorem 6 naturally holds for the important class of α-s.d.d. matrices because $\alpha_{ij} \le ||\text{off}(A)||_F = \alpha < 1$. But the question is, If the initial matrix H is α-s.d.d., does the sequence of matrices, which is generated by the Jacobi method, retain this property? The affirmative answer to this question for positive definite matrices is given in [4, Lemma 3] and [5, Lemma 3]. It is also true for indefinite matrices (see Lemma 8 in section 4.3).

4.2. The estimates for one sweep. The most common pivot strategies which are used in the Jacobi method are the row- and the column-cyclic[1] strategy. We shall consider now an equivalent parallel strategy which was first proposed in [9]. Each cycle of this strategy consists of $2n-3$ batches of the Jacobi transformations (which are indicated by arrows in Figure 4.1) where all of the transformations in a batch can be performed simultaneously. Since the Jacobi transformation is determined by a pair of pivot indices, each batch will be identified by a set of pivot pairs. Thus we have

$$\text{cycle} = \{\mathcal{B}_1,\ \mathcal{B}_2, \ldots, \mathcal{B}_{2n-4},\ \mathcal{B}_{2n-3}\}, \tag{4.26}$$

where

$$
\begin{aligned}
\mathcal{B}_l &= \{(1,l+1),(2,l),(3,l-1),\ldots,(k,k+1)\} \text{ for } 1 \le l = 2k-1 \le n-1, \\
\mathcal{B}_l &= \{(1,l+1),(2,l),(3,l-1),\ldots,(k,k+2)\} \text{ for } 2 \le l = 2k \le n-1, \\
\mathcal{B}_l &= \{(l',n),(l'+1,n-1),(l'+2,n-2),\ldots,(k,k+1)\} \\
&\qquad \text{for } n \le l = 2k-1 \le 2n-3,\ l' = l-n+2, \\
\mathcal{B}_l &= \{(l',n),(l'+1,n-1),(l'+2,n-2),\ldots,(k,k+2)\} \\
&\qquad \text{for } n \le l = 2k \le 2n-3,\ l' = l-n+2.
\end{aligned} \tag{4.27}
$$

For example we have

$$
\begin{aligned}
&\mathcal{B}_1 = \{(1,2)\}, \quad \mathcal{B}_2 = \{(1,3)\}, \quad \mathcal{B}_3 = \{(1,4),(2,3)\}, \\
&\mathcal{B}_4 = \{(1,5),(2,4)\}, \quad \mathcal{B}_5 = \{(1,6),(2,5),(3,4)\}, \\
&\mathcal{B}_6 = \{(1,7),(2,6),(3,5)\}, \quad \mathcal{B}_7 = \{(1,8),(2,7),(3,6),(4,5)\}, \\
&\cdots \qquad \cdots \qquad \cdots \\
&\mathcal{B}_{2n-7} = \{(n-5,n),(n-4,n-1),(n-3,n-2)\}, \\
&\mathcal{B}_{2n-6} = \{(n-4,n),(n-3,n-1)\}, \quad \mathcal{B}_{2n-5} = \{(n-3,n),(n-2,n-1)\}, \\
&\mathcal{B}_{2n-4} = \{(n-2,n)\}, \quad \mathcal{B}_{2n-3} = \{(n-1,n)\}.
\end{aligned}
$$

Let us consider now the accuracy result for an arbitrarily chosen batch $\mathcal{B} \in$ cycle. We consider the element at arbitrary position $(i,r), i < r$ before and after the batch $\mathcal{B}$ is performed. If $(i,r) \in \mathcal{B}$, then it is the pivot element and it changes only once (it

[1] The details about cyclic pivot strategies are given in [4].

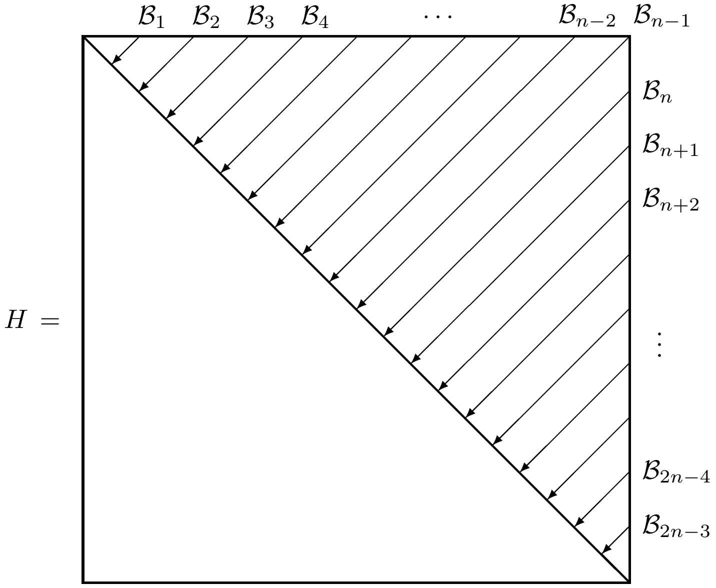


FIG. 4.1. *Parallel pivot strategy (Sameh).*

is annihilated). If $(i,r) \notin \mathcal{B}$, then it may not change or it may change once or twice. Indeed, if there is no pivot pair in $\mathcal{B}$ which belongs to the ith row or rth column, then this element does not change. If there is only one pivot pair in $\mathcal{B}$ corresponding to the ith row or rth column, then the element changes once. This case was considered in Theorem 6. If there are two pivot pairs in $\mathcal{B}$, one corresponding to the ith row, and one to the rth column, then it changes twice. Thus, it remains to consider this last case.

Suppose that $\mathcal{B}$ contains two pivot pairs $(i,j), i < j$ and $(l,r), l < r$ (see Figure 4.2). Let H be the matrix iterate at the stage just before applying the batch $\mathcal{B}$. Since all of the pairs from $\mathcal{B}$ are disjoint (as sets), it is irrelevant in which order the Jacobi transformations, defined by the pairs from $\mathcal{B}$, are applied. We can assume that the rotation with pivot pair (i,j) is applied as first and with the pair (l,r) as second. So, let

$$H \xrightarrow{(i,j)} H' \xrightarrow{(l,r)} H'', \qquad H' = U^T H U, \qquad H'' = (U\,')^T H' U\,'.$$

Using the transformation formulas from Algorithm 1, we have

$$\begin{aligned} h'_{ik} &= c h_{ik} - s h_{jk} = h'_{ki}, \quad & h'_{jk} &= c h_{jk} + s h_{ik} = h'_{kj}, \quad & k &\neq i,j; \\ h''_{kl} &= c' h'_{kl} - s' h'_{kr} = h''_{lk}, \quad & h''_{kr} &= c' h'_{kr} + s' h'_{kl} = h''_{rk}, \quad & k &\neq l,r, \end{aligned} \tag{4.28}$$

where $c = cs$, $s = sn$ and $c' = cs'$, $s' = sn'$ are the sine and the cosine of the rotation angle in the first and second transformation, respectively. Thus, after these two transformations, the elements h''_{il}, h''_{jl}, h''_{ir}, and h''_{jr} are some linear combination of h_{il}, h_{jl}, h_{ir}, and h_{jr} (see again Figure 4.2). As performed earlier, we set $\mathrm{fl}(h'_{pq}) = h'_{pq} + \delta h'_{pq}$ and $\mathrm{fl}(h''_{pq}) = h''_{pq} + \delta h''_{pq}$. Using the notation from Lemma 5, we have

$$\begin{aligned} \mathrm{fl}(h''_{il}) = (1+\varepsilon_1)\,&[(1+\varepsilon_2)(1+\boldsymbol{\epsilon}_{c'}+\eta_{c'}) \cdot c' \cdot (h'_{il} + \delta h'_{il}) \\ &- (1+\varepsilon_3)(1+\boldsymbol{\epsilon}_{s'}+\eta_{s'}) \cdot s' \cdot (h'_{ir} + \delta h'_{ir})] \;=\; c' h'_{il} - s' h'_{ir} + \delta h''_{il}, \end{aligned}$$

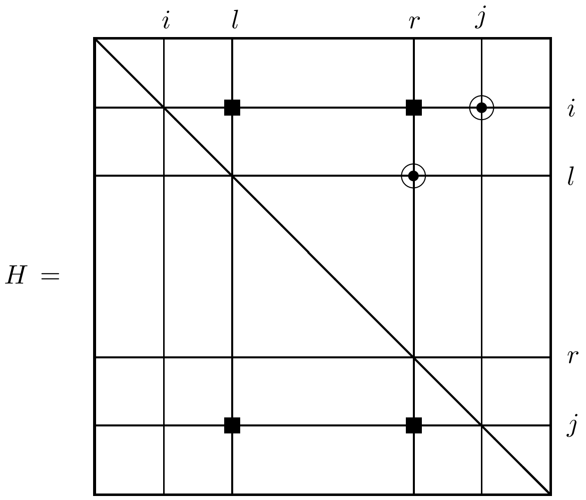


FIG. 4.2. *Two subsequent transformations.*

where

$$\begin{aligned}\delta h''_{il} = {} & [(\boldsymbol{\epsilon}_{c'} + \eta_{c'})(1+\varepsilon_1)(1+\varepsilon_2) + \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2] \cdot c' \cdot h'_{il} \\ & - [(\boldsymbol{\epsilon}_{s'} + \eta_{s'})(1+\varepsilon_1)(1+\varepsilon_3) + \varepsilon_1 + \varepsilon_3 + \varepsilon_1\varepsilon_3] \cdot s' \cdot h'_{ir} \\ & + (1+\varepsilon_1)(1+\varepsilon_2)(1+\boldsymbol{\epsilon}_{c'} + \eta_{c'}) \cdot c' \cdot \delta h'_{il} \\ & - (1+\varepsilon_1)(1+\varepsilon_3)(1+\boldsymbol{\epsilon}_{s'} + \eta_{s'}) \cdot s' \cdot \delta h'_{ir}.\end{aligned}$$

Using the relations (4.7)–(4.10), we obtain

$$|\delta h''_{il}| \leq \mu_1 |c'| \cdot |h'_{il}| + \mu_2 |s'| \cdot |h'_{ir}| + (1+\mu_1)|c'| \cdot |\delta h'_{il}| + (1+\mu_2)|s'| \cdot |\delta h'_{ir}|. \tag{4.29}$$

We obtain similar expressions for $\delta h''_{ir}$, $\delta h''_{jl}$, and $\delta h''_{jr}$. For $\delta h''_{ir}$ we just replace $il \to ir$ and $ir \to il$ in (4.29). For $\delta h''_{jl}$ we change $il \to jl$ and $ir \to jr$. For $\delta h''_{jr}$ we change $il \to jr$ and $ir \to jl$.

To obtain the backward error, we write

$$\begin{aligned}\begin{bmatrix} \mathrm{fl}(h''_{il}) & \mathrm{fl}(h''_{ir}) \\ \mathrm{fl}(h''_{jl}) & \mathrm{fl}(h''_{jr}) \end{bmatrix} &= \widetilde{U}^T \begin{bmatrix} h_{il} & h_{ir} \\ h_{jl} & h_{jr} \end{bmatrix} \widetilde{U}\,' + \begin{bmatrix} \delta h''_{il} & \delta h''_{ir} \\ \delta h''_{jl} & \delta h''_{jr} \end{bmatrix} \\ &= \widetilde{U}^T \left(\begin{bmatrix} h_{il} & h_{ir} \\ h_{jl} & h_{jr} \end{bmatrix} + \begin{bmatrix} \Delta h_{il} & \Delta h_{ir} \\ \Delta h_{jl} & \Delta h_{jr} \end{bmatrix} \right) \widetilde{U}\,',\end{aligned}$$

which yields

$$\begin{aligned}\begin{bmatrix} \Delta h_{il} & \Delta h_{ir} \\ \Delta h_{jl} & \Delta h_{jr} \end{bmatrix} &= \widetilde{U} \begin{bmatrix} \delta h''_{il} & \delta h''_{ir} \\ \delta h''_{jl} & \delta h''_{jr} \end{bmatrix} \left(\widetilde{U}\,'\right)^T \\ &= \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \delta h''_{il} & \delta h''_{ir} \\ \delta h''_{jl} & \delta h''_{jr} \end{bmatrix} \begin{bmatrix} c' & -s' \\ s' & c' \end{bmatrix}. \end{aligned} \tag{4.30}$$

Similarly, as in Theorem 6 we need to estimate the backward errors for scaled elements which requires a lot of elementary computations. We shall illustrate it for the element $a_{il} = h_{il}/(d_i d_l)$, where $d_i = \sqrt{|h_{ii}|}$, $d_l = \sqrt{|h_{ll}|}$.

Using (4.30), we have $\Delta h_{il} = cc'\delta h''_{il} + sc'\delta h''_{jl} + cs'\delta h''_{ir} + ss'\delta h''_{jr}$, which together with the relations (4.29), (4.28), (4.8), and (4.10), yields

$$|\Delta a_{il}| \le \frac{1}{d_i d_l}\left(|cc'| \cdot \Theta_{iljr} + |sc'| \cdot \Theta_{jlir} + |cs'| \cdot \Theta_{irjl} + |ss'| \cdot \Theta_{jril}\right), \tag{4.31}$$

where

$$\begin{aligned}\Theta_{iljr} \;=\; & \mu_1|c'| \cdot (|ch_{il}| + |sh_{jl}|) + \mu_2|s'| \cdot (|ch_{ir}| + |sh_{jr}|) \\ & + (1+\mu_1)|c'| \cdot (\mu_1|ch_{il}| + \mu_2|sh_{jl}|) + (1+\mu_2)|s'| \cdot (\mu_1|ch_{ir}| + \mu_2|sh_{jr}|).\end{aligned} \tag{4.32}$$

The expression for Θ_{jlir} is obtained by replacing indices, indicated by the transition $(i,l,j,r) \to (j,l,i,r)$ and similarly for Θ_{irjl} and Θ_{jril}. In the inequality (4.31) the following terms appear:

$$\frac{sh_{jl}}{d_i d_l} = s\frac{d_j}{d_i}a_{jl}, \quad \frac{s'h_{ir}}{d_i d_l} = s'\,\frac{d_r}{d_l}a_{ir}, \quad \text{and} \quad \frac{ss'h_{jr}}{d_i d_l} = s\frac{d_j}{d_i} \cdot s'\,\frac{d_r}{d_l} \cdot a_{jr}.$$

Thus we need to estimate $|sd_j/d_i|$ and $|s'd_r/d_l|$. Similarly, in the expressions for $|\Delta a_{ir}|$, $|\Delta a_{jl}|$, and $|\Delta a_{jr}|$ we additionally need to estimate $|sd_i/d_j|$ and $|s'd_l/d_r|$. Let

$$q = \min\left\{\frac{d_j}{d_i},\, \frac{d_i}{d_j}\right\}, \qquad q' = \min\left\{\frac{d_l}{d_r},\, \frac{d_r}{d_l}\right\}.$$

Since the rotation angles belong to $[-\pi/4,\, \pi/4]$, we have $|s| \le \sqrt{2}/2$ and $|s'| \le \sqrt{2}/2$. Using the inequalities (4.22), we obtain

$$|s|\,\frac{1}{q} \le \begin{cases} \dfrac{\sqrt{2}}{2(1-q^2)} \le \dfrac{\sqrt{2}(\sqrt{5}+1)}{4} = \tau \;\text{ for }\; 0 \le q \le \dfrac{\sqrt{5}-1}{2}, \\[2ex] \dfrac{\sqrt{2}}{2} \cdot \dfrac{1}{q} \le \tau \;\text{ for }\; \dfrac{\sqrt{5}-1}{2} \le q \le 1. \end{cases}$$

This estimate holds also for $|s'| \cdot (1/q')$. Thus we have obtained

$$\left|s\,\frac{d_j}{d_i}\right| \le |s| \cdot \frac{1}{q} \le \tau, \qquad \left|s'\,\frac{d_r}{d_l}\right| \le |s'| \cdot \frac{1}{q'} \le \tau < 1.144\,122\,807, \tag{4.33}$$

and the same for $|sd_i/d_j|$ and $|s'd_l/d_r|$. If we insert the bounds (4.33) in (4.31), then by using (4.32) we obtain

$$\begin{aligned}|\Delta a_{il}| \le\; & \mu_1|a_{il}| + \mu_1\tau|a_{jl}| + \mu_2\tau|a_{ir}| + \mu_2\tau^2|a_{jr}| + (1+\mu_1)\mu_1|a_{il}| \\ & + (1+\mu_1)\mu_2\tau|a_{jl}| + (1+\mu_2)\mu_1\tau|a_{ir}| + (1+\mu_2)\mu_2\tau^2|a_{jr}| \\ & + \mu_1\tau|a_{jl}| + \frac{\mu_1}{2}|a_{il}| + \mu_2\tau^2|a_{jr}| + \frac{\tau}{2}\mu_2|a_{ir}| + (1+\mu_1)\mu_1\tau|a_{jl}| \\ & + \frac{1}{2}(1+\mu_1)\mu_2|a_{il}| + (1+\mu_2)\mu_1\tau^2|a_{jr}| + \frac{\tau}{2}(1+\mu_2)\mu_2|a_{ir}| \\ & + \mu_1\tau|a_{ir}| + \mu_1\tau^2|a_{jr}| + \frac{\mu_2}{2}|a_{il}| + \frac{\tau}{2}\mu_2|a_{jl}| + (1+\mu_1)\mu_1\tau|a_{ir}| \\ & + (1+\mu_1)\mu_2\tau^2|a_{jr}| + \frac{1}{2}(1+\mu_2)\mu_1|a_{il}| + \frac{\tau}{2}(1+\mu_2)\mu_2|a_{jl}| \\ & + \mu_1\tau^2|a_{jr}| + \frac{\tau}{2}\mu_1|a_{ir}| + \frac{\tau}{2}\mu_2|a_{jl}| + \frac{1}{4}\mu_2|a_{il}| + (1+\mu_1)\mu_1\tau^2|a_{jr}| \\ & + \frac{\tau}{2}(1+\mu_1)\mu_2|a_{ir}| + \frac{\tau}{2}(1+\mu_2)\mu_1|a_{jl}| + \frac{1}{4}(1+\mu_2)\mu_2|a_{il}|.\end{aligned}$$

Here each of the terms Θ_{iljr}, Θ_{jlir}, Θ_{irjl}, and Θ_{jril} from the relation (4.31) are estimated by two rows in the above relation in the same order. Thus, we have obtained

$$|\Delta a_{il}| \leq \nu_1|a_{il}| + \nu_2|a_{jl}| + \nu_2|a_{ir}| + \nu_3|a_{jr}|, \tag{4.34}$$

and similarly,

$$\begin{aligned} |\Delta a_{jl}| &\leq \nu_1|a_{jl}| + \nu_2|a_{il}| + \nu_2|a_{jr}| + \nu_3|a_{ir}|, \\ |\Delta a_{ir}| &\leq \nu_1|a_{ir}| + \nu_2|a_{jr}| + \nu_2|a_{il}| + \nu_3|a_{jl}|, \\ |\Delta a_{jr}| &\leq \nu_1|a_{jr}| + \nu_2|a_{ir}| + \nu_2|a_{jl}| + \nu_3|a_{il}|, \end{aligned} \tag{4.35}$$

where

$$\begin{aligned} \nu_1 &= 3\mu_1 + 1.5\mu_2 + \mu_1^2 + \mu_1\mu_2 + 0.25\mu_2^2, \\ \nu_2 &= (3.5\mu_1 + 2.5\mu_2 + \mu_1^2 + 1.5\mu_1\mu_2 + 0.5\mu_2^2\,)\,\tau, \\ \nu_3 &= (4\mu_1 + 4\mu_2 + \mu_1^2 + 2\mu_1\mu_2 + \mu_2^2\,)\,\tau^2. \end{aligned} \tag{4.36}$$

If we use the inequalities (4.24) and $u \leq 2^{-23}$ in the equations (4.36), we obtain

$$\nu_1 < 37.500\,076\,72\,u, \quad \nu_2 < 58.636\,417\,67\,u, \quad \nu_3 < 91.631\,389\,15\,u. \tag{4.37}$$

Finally, using (4.37) in the relations (4.34) and (4.35), we obtain

$$\begin{aligned} &|\Delta a_{il}|^2 + |\Delta a_{jl}|^2 + |\Delta a_{ir}|^2 + |\Delta a_{jr}|^2 \\ &\leq 4(\nu_1^2 + 2\nu_2^2 + \nu_3^2)\left(|a_{il}|^2 + |a_{jl}|^2 + |a_{ir}|^2 + |a_{jr}|^2\right) \\ &\leq 66\,716.1048\,u^2\left(|a_{il}|^2 + |a_{jl}|^2 + |a_{ir}|^2 + |a_{jr}|^2\right). \end{aligned} \tag{4.38}$$

We can now prove the accuracy result for an arbitrarily chosen batch of Jacobi rotations.

THEOREM 7. [illegible] H [illegible] n [illegible] $\mathcal{B} \in$ *cycle* [illegible] (4.26) [illegible] (4.27) [illegible] $\widehat{H}$ [illegible] $\mathrm{fl}(\widehat{H})$ [illegible] H [illegible] $\mathcal{B}$ [illegible] 1 [illegible] ΔH [illegible] $\mathrm{fl}(\widehat{H})$ [illegible] $H + \Delta H$ [illegible] $\mathcal{B}$ [illegible] $D = |\mathrm{diag}(H)|^{1/2}$ $A = D^{-1}HD^{-1}$ $\Delta A = D^{-1}\Delta H D^{-1}$ [illegible]

$$\alpha = ||\mathrm{off}(A)||_F = \sqrt{\sum_{\substack{k,l=1 \\ k\neq l}}^{n} a_{kl}^2}.$$

[illegible] $\alpha < 1$ [illegible]

$$||\Delta A||_2 \;<\; (258.67\,\alpha + 2.47)\,u,$$

[illegible] u [illegible] (3.1)

[illegible] We consider first the error in an arbitrarily chosen off-diagonal element a_{ir}, $i < r$. If $(i, r) \in \mathcal{B}$, then a_{ir} is the pivot element and the inequalities (4.23) and (4.24) yield

$$|\Delta a_{ir}| = |\delta a_{ir}| \leq 1.465\,926\,u + 7.500\,006\,74|a_{ir}|u. \tag{4.39}$$

If there exists j such that $(i,j) \in \mathcal{B}$ and $(k,r) \notin \mathcal{B}$ for all $1 \le k \le n$, then a_{ir} is changed together with a_{jr} once and the inequality (4.25) holds. This gives

$$(\Delta a_{ir})^2 + (\Delta a_{jr})^2 \le 562.299\,700\,9\,(a_{ir}^2 + a_{jr}^2)\,u^2. \tag{4.40}$$

A similar situation appears if there exists l such that $(l,r) \in \mathcal{B}$ and $(i,k) \notin \mathcal{B}$ for all $1 \le k \le n$. If there exist j and l such that $(i,j),(l,r) \in \mathcal{B}$, then a_{ir} together with a_{il}, a_{jl}, and a_{jr} are changed twice, so the relation (4.38) holds. Note that the bound from (4.38) is larger than the one from (4.40), and we can use it in both cases.

Now, we consider the diagonal element a_{ii}. If $(i,k),(k,i) \notin \mathcal{B}$ for all $1 \le k \le n$, then a_{ii} does not change. If there exists j such that $(i,j) \in \mathcal{B}$, then (4.23) and (4.24) yield

$$\begin{aligned} |\Delta a_{ii}| &= |\delta a_{ii}| \le u + 7.500\,006\,74\,|a_{ij}|\,u, \\ |\Delta a_{jj}| &= |\delta a_{jj}| \le u + 12.020\,824\,82\,|a_{ij}|\,u. \end{aligned} \tag{4.41}$$

We obtain a similar conclusion, with i and j interchanged, if $(j,i) \in \mathcal{B}$.

Now, we construct ΔA of three parts, $\Delta A = \Delta_D + \Delta_P + \Delta_O$, where

- Δ_D is a diagonal matrix which contains only the errors of diagonal elements which are estimated by the inequalities (4.41).
- Δ_P contains the errors of pivot elements from $\mathcal{B}$ and their transposes according to (4.39). Thus, Δ_P is symmetric and zero at positions other than those defined in $\mathcal{B}$.
- Δ_O contains the errors of elements a_{ir}, where $i \ne r$ and neither (i,r) nor (r,i) belongs to $\mathcal{B}$. They are estimated by the inequalities (4.40) and (4.38).

Now, if we set $x = \max_{(i,j)\in\mathcal{B}} |a_{ij}|$, then we have

$$\begin{aligned} ||\Delta A||_2 &\le ||\Delta_D||_2 + ||\Delta_P||_2 + ||\Delta_O||_F \\ &\le u + 12.020\,824\,82\,xu + 1.465\,926\,u + 7.500\,006\,74\,xu \\ &\quad + u \cdot \sqrt{66716.1048} \cdot \sqrt{\alpha^2 - 2 \sum_{(i,j)\in\mathcal{B}} a_{ij}^2} \\ &\le 2.465\,926\,u + 19.520\,831\,6\,xu + 258.294\,608\,6\,u \cdot \sqrt{\alpha^2 - 2x^2} \\ &\le 2.465\,926\,u + 258.294\,608\,6\,u \cdot \left(0.075\,575\,84\,x + \sqrt{\alpha^2 - 2x^2}\,\right). \end{aligned}$$

Let us consider for the moment the function $f(x) = \mu x + \sqrt{\alpha^2 - 2x^2}$, $0 \le x \le \alpha/\sqrt{2}$, where $\mu \in \langle 0,1\rangle$. It attains the maximum at the point $x_* = \mu\alpha/\sqrt{4+2\mu^2}$ and $f(x_*) = \alpha\sqrt{1+\mu^2/2}$. Thus we have

$$\begin{aligned} ||\Delta A||_2 &\le 2.465\,926\,u + 258.294\,608\,6\,u \cdot \alpha \cdot \sqrt{1 + \frac{1}{2} \cdot 0.075\,575\,84^2} \\ &\le 2.465\,926\,u + 258.663\,172\,u\,\alpha, \end{aligned}$$

which completes the proof. □

Let us make some observations. In the proof of Theorem 7 we have used the inequality (4.38). This means that we have assumed that all off-diagonal elements (except the rotated ones) undergo two changes. This is almost true for the batches with about $n/2$ elements (pivot pairs). For example, if n is even, then such batches are $\mathcal{B}_{n-1}$ with $n/2$ elements, $\mathcal{B}_{n-2}$ and $\mathcal{B}_n$ with $n/2-1$ elements, $\mathcal{B}_{n-3}$ and $\mathcal{B}_{n+1}$ with

$n/2-2$ elements etc. In these cases the bound of Theorem 7 is realistic. If the number of elements of a batch is much smaller than $n/2$, then the bound is too large. The reason lies in the fact that for such batches most of the matrix elements change only once and for them the sharper bound (4.40) can be applied. Moreover, some parts of the matrix do not change. For example, the batches $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_{2n-4}$, and $\mathcal{B}_{2n-3}$ have only one pivot pair and Theorem 6 holds. For $\mathcal{B}_3$, $\mathcal{B}_4$, $\mathcal{B}_{2n-6}$, and $\mathcal{B}_{2n-5}$ only four elements in the upper triangle change twice. For $\mathcal{B}_5$, $\mathcal{B}_6$, $\mathcal{B}_{2n-8}$, and $\mathcal{B}_{2n-7}$ there are 12 such elements etc. It means that for certain batches the constant 258.67 from Theorem 7 can be improved.

4.3. Accuracy of the computed eigenvalues. We consider now one sweep of the Jacobi method under the column-cyclic strategy or under the equivalent parallel pivot strategy which is given by the relations (4.26)–(4.27) and Figure 4.1. If we apply the method to the initial matrix H of order n, then such a sweep contains $N = n(n-1)/2$ Jacobi steps, which are given by Algorithm 1. Thus, the Jacobi method generates the sequence of matrices $H^{(0)} = H, H^{(1)}, H^{(2)}, \ldots, H^{(N)}$. Let $\alpha_k = \|\mathrm{off}(A^{(k)})\|_F$ be the scaled off-norm, where $A^{(k)} = |\mathrm{diag}(H^{(k)})|^{-1/2}H^{(k)}|\mathrm{diag}(H^{(k)})|^{-1/2}$, $k \geq 0$. It may happen that $\alpha_k > \alpha_0$ for some k. But, if α_0 is small enough, then this growth is not significant and α-s.d.d. property is retained within the whole cycle. This follows from the following lemma

LEMMA 8 (see [6, Lemma 3]). [illegible] $\alpha_0 \leq 1/(10n)$, [illegible] $\alpha_k^2 \leq c_k\alpha_0^2$, $0 \leq k \leq N$ [illegible] $c_k = (1 + \frac{0.00126}{n^2})^k < 1.0007$.

The bound $1/(10n)$ has been chosen in [6] as the assumption for the quadratic convergence considerations and Lemma 8 is just an auxiliary result. The bound could have been chosen much larger, but that would result in a larger bound for c_k (see also [4, Lemma 3]). Although α_k may increase during the Jacobi process, we have the quadratic reduction at the end of the cycle provided that α_0 is small enough.

THEOREM 9 (see [6, Theorem 6]). [illegible] H [illegible] $n \geq 3$ [illegible]

$$[illegible]\quad \alpha_0 \leq \frac{1}{10}\min\left\{\frac{1}{n}, \gamma\right\}, \quad [illegible] \quad \alpha_N \ \leq \ 2.8\,\frac{\alpha_0^2}{\gamma}\,,$$

[illegible]

$$\gamma = \min_{\substack{\lambda,\mu \in spectrum(H)\\ \lambda \neq \mu}} \frac{|\lambda - \mu|}{|\lambda| + |\mu|}. \tag{4.42}$$

We estimate the errors in the computed eigenvalues, which are caused by floating point arithmetic, by using the following perturbation result.

THEOREM 10 (see [7, Theorem 3.1]). [illegible] H [illegible] δH [illegible] n [illegible] $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ [illegible] $\lambda_1' \geq \lambda_2' \geq \cdots \geq \lambda_n'$ [illegible] H [illegible] $H + \delta H$ [illegible] $D = |\,\mathrm{diag}(H)|^{1/2}$ [illegible] $A = D^{-1}HD^{-1}$ $\delta A = D^{-1}\delta HD^{-1}$ [illegible] α, η [illegible] $\|\mathrm{off}(A)\| \leq \alpha$ [illegible] $\|\delta A\| \leq \eta$

$$[illegible]\quad \eta + 2\alpha < 1\,, \quad [illegible] \quad \left|\frac{\lambda_i'}{\lambda_i} - 1\right| \ \leq \ \frac{\eta}{1 - \left(1 + \frac{\eta}{1-2\alpha}\right)\alpha} \ \leq \ \frac{\eta}{1 - 2\alpha}\,, \quad 1 \leq i \leq n\,.$$

Using now Lemma 8 and Theorem 10 together with Theorems 6 and 7, we have the following conclusions.

Corollary 11. [illegible] 10 δH [illegible] H [illegible]

(i) $\eta = (22.54\alpha + 2.51)\,u, \quad \alpha = \alpha_0$ [illegible],

(ii) $\eta = (258.67\,\alpha + 2.47)\,u, \quad \alpha = \alpha_0$ [illegible],

(iii) $\eta = (2n-3)\,(258.67\,\alpha + 2.47)\,u, \quad \alpha = \max_{0\le k\le N} \alpha_k$ [illegible]

[illegible] $\alpha_0 \le 1/(10n)$, [illegible] $\alpha \le 1.00035\,\alpha_0$.

For example, if $\alpha_0 \le 10^{-2}$ ($\alpha_0 \le 10^{-3}$), then the relative error in the computed eigenvalues after we have applied the Jacobi method is bounded above by $2.8u$ ($2.6u$) for one step, $5.1u$ ($2.8u$) for one batch, and $10.2nu$ ($5.5nu$) for one sweep.

[illegible] 12. As has been suggested by Rutishauser [8], the diagonal elements can be updated just once in each cycle. To do that, a vector z with n components is used. At the beginning of each cycle it is set to zero, $z = 0$. Within a cycle all contributions to the diagonal elements are accumulated in z. If the current value of the diagonal element $h_{ii}^{((r-1)N+k)}$, $0 \le k \le N-1$, $N = n(n-1)/2$ is needed to compute the rotation matrix, then one can use the value of $h_{ii}^{((r-1)N)} + z_i$. At the end of cycle r the diagonal elements are updated $h_{ii}^{(rN)} = h_{ii}^{((r-1)N)} + z_i$, $1 \le i \le n$, and z is set to zero. If we write h'_{ii} for $h_{ii}^{(rN)}$ and h_{ii} for $h_{ii}^{((r-1)N)}$, then at the end of cycle r, we have $h'_{ii} = h_{ii} + z_i$. If $\mathrm{fl}(z_i) = (1+\boldsymbol{\epsilon}_{z_i})z_i$, then

$$\mathrm{fl}(h'_{ii}) = (1+\varepsilon)[h_{ii} + (1+\boldsymbol{\epsilon}_{z_i})z_i] = (1+\boldsymbol{\epsilon}_i)h'_{ii}, \qquad \boldsymbol{\epsilon}_i = \varepsilon + (1+\varepsilon)\cdot\frac{z_i}{h_{ii}+z_i}\boldsymbol{\epsilon}_{z_i}.$$

If α is small enough, then $|z_i| \ll |h_{ii}|$ and thus $\boldsymbol{\epsilon}_{z_i}$ (which can be as large as nu) is suppressed by the factor $z_i/(h_{ii}+z_i)$. Therefore, after one cycle, the constant $(2n-3)\cdot 2.47u$ implied by Corollary 11(iii) becomes insignificant.

Usually, the quadratic convergence begins when $\alpha \le \gamma$, where γ is defined by (4.42) (see [4], [5], and Theorem 9). Due to the quadratic reduction of α (see [5] and Theorem 9), the process is terminated after several cycles. Now Corollary 11 and Remark 12 claim that the inaccuracy in the computed eigenvalues, coming from these last few cycles, will hardly be several ulps (units in the last place).

4.4. The stopping criterion. The known stopping strategy for positive definite matrices from [2] (which is actually implied by Rutishauser's stopping criterion from [8]) reads: if $|a_{ij}| \le u$, then set $h_{ij} = 0$. After all off-diagonal elements become zero, the process is terminated. Using Theorem 6, we can slightly modify the stopping criterion for indefinite matrices.

If we apply the rotation with pivot pair (i,j), which is given by Algorithm 1, then the backward error is estimated by Theorem 6. If we do not apply the rotation, but we set the pivot element to zero, then we produce the perturbation matrix δH for which the scaled matrix δA has only two nonzero elements, $(\delta A)_{ij} = (\delta A)_{ji} = -a_{ij}$ and we have $||\delta A||_F = \sqrt{2}|a_{ij}|$. If the error thus produced is less than the error which is caused by the rotation, then the rotation need not be performed. Thus, we suggest the stopping strategy

$$\text{(4.43)} \qquad \text{if} \quad \sqrt{2}|a_{ij}| \le (22.54\alpha_{ij} + 2.51)u, \quad \text{then set} \quad h_{ij} = 0,$$

where α_{ij} is defined in the statement of Theorem 6. However, the computation of α_{ij} requires about $4n$ operations. So, we shall simplify (4.43). Since $\alpha_{ij} \ge \sqrt{2}|a_{ij}|$, the inequality in (4.43) holds provided that $\sqrt{2}|a_{ij}| \leq (22.54\sqrt{2}|a_{ij}| + 2.51)u$, or

$$|a_{ij}| \leq \frac{2.51u}{\sqrt{2} - 22.54\sqrt{2}u} \leq 1.774\,843\,u\,.$$

Here we have used $u \leq 2^{-23}$. We can replace the bound $1.774\,843\,u$ by $2\,u$ since α_{ij} will generally be much larger than $\sqrt{2}|a_{ij}|$. Thus, instead of (4.43), we suggest the following simple stopping strategy for the indefinite Jacobi method:

$$\text{if} \quad |a_{ij}| \leq 2\,u, \quad \text{then set} \quad h_{ij} = 0, \tag{4.44}$$

and terminate the process after all off-diagonal elements become zero.

Acknowledgments. The author is indebted to V. Hari for suggesting the theme, as well as for reading and improving this paper. He is also thankful to the anonymous referees whose comments and requirements have made this paper better.

REFERENCES

[1] J. Barlow and J. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[2] J. Demmel and K. Veselić, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.

[3] F. M. Dopico, J. M. Molera, and J. Moro, *An orthogonal high relative accuracy algorithm for the symmetric eigenproblem*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 301–351.

[4] J. Matejaš, *Quadratic convergence of scaled matrices in Jacobi method*, Numer. Math., 87 (2000), pp. 171–199.

[5] J. Matejaš, *Convergence of scaled iterates by the Jacobi method*, Linear Algebra Appl., 349 (2002), pp. 17–53.

[6] J. Matejaš, *Quadratic convergence bounds of scaled iterates by the serial Jacobi method for indefinite Hermitian matrices*, Electron. J. Linear Algebra, 17 (2008), pp. 62–87.

[7] J. Matejaš and V. Hari, *Relative Eigenvalue and Singular Value Perturbations of Scaled Diagonally Dominant Matrices*, BIT, to appear.

[8] H. Rutishauser, *Handbook series linear algebra: The Jacobi method for real symmetric matrices*, Numer. Math., 9 (1996), pp. 1–10.

[9] A. H. Sameh, *On Jacobi and Jacobi-like algorithms for parallel computer*, Math. Comp., 25 (1971), pp. 579–590.

[10] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph.D. thesis, University of Hagen, Hagen, Germany, 1992.

[11] K. Veselić, *A Jacobi eigenreduction algorithm for definite matrix pairs*, Numer. Math., 64 (1993), pp. 241–269.

[12] K. Veselić and I. Slapničar, *Floating-point perturbations of Hermitian matrices*, Linear Algebra Appl., 195 (1993), pp. 81–116.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 154–174

A QUASI-SEPARABLE APPROACH TO SOLVE THE SYMMETRIC DEFINITE TRIDIAGONAL GENERALIZED EIGENVALUE PROBLEM[*]

RAF VANDEBRIL[†], GENE GOLUB[‡], AND MARC VAN BAREL[†]

Abstract. We present a new fast algorithm for solving the generalized eigenvalue problem $T\mathbf{x} = \lambda S\mathbf{x}$, in which both T and S are real symmetric tridiagonal matrices and S is positive definite. A method for solving this problem is to compute a Cholesky factorization $S = LL^T$ and solve the equivalent symmetric standard eigenvalue problem $L^{-1}TL^{-T}(L^T\mathbf{x}) = \lambda(L^T\mathbf{x})$. We prove that the matrix $L^{-1}TL^{-T}$ is quasi-separable; that is, all submatrices taken out of its strictly lower triangular part have rank at most 1. We show how to efficiently compute the $\mathcal{O}(n)$ parameters defining $L^{-1}TL^{-T}$ and review eigensolvers for quasi-separable matrices. Our approach shows that by fully exploiting the structure, the eigenvalues of $T\mathbf{x} = \lambda S\mathbf{x}$ can be computed in $\mathcal{O}(n^2)$ operations, as opposed to the $\mathcal{O}(n^3)$ operations for standard methods such as the so-called Cholesky-QR method. It will be shown that the computation of the representation of this quasi-separable matrix is only linear in time, and numerical experiments will illustrate the effectiveness of the presented approach.

Key words. generalized eigenvalue problem, quasi-separable, tridiagonal matrices

AMS subject classifications. 65F15, 15A18

DOI. 10.1137/070679491

1. Introduction.

In this paper we consider generalized eigenvalue problems of the form

$$T\mathbf{x} = \lambda S\mathbf{x},$$

where both T and S are symmetric tridiagonal matrices and S is positive definite. This problem arises in several applications such as the numerical solution of the radial Schrödinger and Sturm–Liouville equations and in vibrational analysis [4, 47]. More references to applications can be found in [38].

A method for solving this problem is the reduction to a standard eigenvalue problem in the following sense:

$$L^{-1}TL^{-T}(L^T\mathbf{x}) = \lambda(L^T\mathbf{x}),$$

where $S = LL^T$ is the Cholesky decomposition of the matrix S (see [33, 44]). This approach is considered less attractive since the generated matrix $L^{-1}TL^{-T}$ is dense.

[*] Received by the editors January 9, 2007; accepted for publication (in revised form) by Z. Drmač September 9, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/simax/31-1/67949.html

[†] Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Leuven (Heverlee), Belgium (Raf.Vandebril@cs.kuleuven.be, Marc.VanBarel@cs.kuleuven.be). The research of these authors was partially supported by the Research Council K.U.Leuven, project OT/05/40 (Large rank structured matrix computations), Center of Excellence: Optimization in Engineering, by the Fund for Scientific Research–Flanders (Belgium), Iterative methods in numerical Linear Algebra, G.0455.0 (RHPH: Riemann–Hilbert problems, random matrices and Padé–Hermite approximation), G.0423.05 (RAM: Rational modelling: Optimal conditioning and stable algorithms), and by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, project IUAP V-22 (Dynamical Systems and Control: Computation, Identification & Modelling). The first author also received a grant of "Postdoctoraal Onderzoeker" from the Funds of Scientific Research Flanders (FWO-Vlaanderen).

[‡] This author is deceased. Former address: Department of Computer Science, Stanford University, Stanford, CA 94305.

This leads hence to an $\mathcal{O}(n^3)$ method for computing the eigenvalues. Moreover, the accuracy of this method also depends on the condition number of S, as the inverse of its Cholesky factors is required. In [12] a detailed error analysis of the Cholesky-QR method is presented, providing error bounds potentially much smaller than $\kappa_2(S)u$, in which u is the unit round-off. Moreover, [12] suggests using for the symmetric definite generalized eigenvalue problem the Cholesky-QR method in which complete pivoting is used to compute the Cholesky factorization. In this paper we opt to use the Cholesky decomposition, but unfortunately we cannot use pivoting since this would destroy the structure and increase the computational complexity of the method.

Also other different techniques exist, working directly on the $T\mathbf{x} = \lambda S\mathbf{x}$ problem. These methods take advantage of both the tridiagonal and symmetric structure and lead to $\mathcal{O}(n^2)$ methods. In [38] one computes the eigenvalues by applying Laguerre's iteration on the associated characteristic polynomial of the pencil. In [6, 29] a divide-and-conquer method is presented. Also methods for the banded symmetric generalized matrix eigenvalue problem exist [36]. More related references can be found, e.g., in [38].

In this paper we prove that the dense matrix $L^{-1}TL^{-T}$ is quasi-separable; that is, all submatrices taken out of the strictly lower triangular part of this matrix are of rank at most one (see [26, 24]).

We show how to efficiently compute the representation of the quasi-separable matrix $L^{-1}TL^{-T}$. As the quasi-separable matrix is highly structured, only $\mathcal{O}(n)$ parameters are needed to define it. We will represent the quasi-separable matrix using the Givens-vector representation [55]. A fast $\mathcal{O}(n)$ method for transforming the generalized eigenvalue problem into an eigenvalue problem involving a quasi-separable matrix will be given.

Based on the $\mathcal{O}(n)$ representation of the quasi-separable matrix, alternative fast methods for computing the whole spectrum are reviewed. We refer to section 3 for an overview.

The manuscript is organized as follows. In the first section some definitions and a theoretical proof of the structure of the matrix $L^{-1}TL^{-T}$ are given. To conclude this section a description is given of the representation used for the quasi-separable matrix and a method for effectively computing the quasi-separable matrix representation of $L^{-1}TL^{-T}$. Section 3 discusses some methods for computing the eigenvalues of quasi-separable matrices. The final section of this manuscript presents the implementation and some numerical experiments related to the accuracy of the new technique for computing the eigenvalues of the generalized eigenvalue problem.

2. Transforming the generalized eigenvalue problem.

We begin with a couple of definitions.

DEFINITION 2.1. [illegible] $A \in \mathbb{R}^{n\times n}$ [illegible] [1] [illegible]

$$\operatorname{rank} A(i:n, 1:i-1) \leq 1 \quad [illegible] \quad i = 2, \dots, n,$$
$$\operatorname{rank} A(1:i-1, i:n) \leq 1 \quad [illegible] \quad i = 2, \dots, n.$$

We will also need to refer to lower/upper triangular semiseparable matrices, whose definition is given below.

[1] We use the colon notation.

Definition 2.2. [illegible] $A \in \mathbb{R}^{n\times n}$ [illegible] 1 [illegible]

$$\begin{aligned} \operatorname{rank} A(i:n,1:i) &\leq 1 \quad \text{[illegible]} \quad i = 1,\ldots,n,\\ \operatorname{rank} A(1:i,i:n) &\leq 1 \quad \text{[illegible]} \quad i = 1,\ldots,n. \end{aligned}$$

A matrix is called lower (resp., upper) semiseparable or quasi-separable if only the lower (resp., upper) triangular part satisfies the rank constraints.

The only difference between quasi-separable and semiseparable is the fact that a quasi-separable matrix does not have the diagonal included in the low rank structure, whereas a semiseparable matrix does have this diagonal included in the structure. Even though these definitions seem little different from each other, there are quite significant differences. For example, the inverse of a semiseparable matrix is of tridiagonal form, whereas the inverse of a quasi-separable matrix is again a quasi-separable matrix. The quasi-separable class of matrices contains also the semiseparable and the tridiagonal matrices.

In this paper some of the discussed references deal with semiseparable plus diagonal matrices instead of quasi-separable matrices. The techniques presented can, however, be adapted easily to be applicable on quasi-separable matrices.

2.1. Theoretical proof of the structure. Let us reconsider now the matrix product $L^{-1}TL^{-T}$, with L satisfying $S = LL^T$. Due to the fact that the matrix S is tridiagonal, the matrix L will be of lower bidiagonal form, and the matrix $L^{-1}TL^{-T}$ will be of quasi-separable form. We will formulate this as a theorem.

Theorem 2.3. [illegible] T [illegible] L [illegible]

$$A = L^{-1}TL^{-T}$$

[illegible]

[illegible]. We assume the considered matrices to be of size n. It is well known that the inverse of a lower bidiagonal matrix is a lower triangular semiseparable matrix (see, e.g., [31]).

The QR-factorization of such a lower semiseparable matrix can be computed by performing a bottom-up sequence of Givens transformations, $n-1$ in total. More precisely the first Givens transformation is performed on the bottom two rows of the matrix L^{-1} to remove the complete last row up to the diagonal. Note that it is possible to remove this complete row with one transformation as the last and second-to-last rows are dependent on each other due to the semiseparable structure. The second Givens transformation acts on rows $n-2$ and $n-1$ and removes the whole row $n-1$ up to the diagonal. This procedure can easily be repeated and gives us the following factorization:

$$G_1^T G_2^T \ldots G_{n-2}^T G_{n-1}^T L^{-1} = R,$$

with R an upper triangular matrix. This means that $L^{-1} = G_{n-1}G_{n-2}\ldots G_2G_1R$, which is the QR-factorization of the considered matrix.

Let us now look closer at the structure of the matrix

$$L^{-1}T = G_{n-1}G_{n-2}\ldots G_2G_1RT.$$

The matrix RT is upper Hessenberg.

The remainder of the proof proceeds by finite induction. Let $\tilde{A}^{(k)} = G_k \dots G_1 RT$. The Givens transformations G_i equal the identity matrix except for the diagonal block $G_i(i : i+1, i : i+1)$, which is of the following form:

$$G_i(i : i+1, i : i+1) = \begin{bmatrix} c_i & -s_i \\ s_i & c_i \end{bmatrix}, \quad \text{where } c_i^2 + s_i^2 = 1.$$

For every k (with $1 \leq k \leq n-1$) we will prove that the constraints

$$\operatorname{rank} \tilde{A}^{(k)}(i : n, 1 : i-1) \leq 1 \quad \text{for all} \quad i = 2, \dots, n \tag{2.1}$$

are satisfied. Therefore, every intermediate matrix $\tilde{A}^{(k)}$ as well as $\tilde{A}^{(n-1)} = L^{-1}T$ will be of lower quasi-separable form.

- Consider $k = 1$. The matrix $\tilde{A}^{(1)} = G_1 RT$ is a Hessenberg matrix, in which we denote the first subdiagonal element as v_1:

$$\tilde{A}^{(1)} = \begin{bmatrix} \times & \times & \times & \times & \cdots \\ \times & \times & \times & \times & \cdots \\ & \times & \times & \times & \\ & & \times & \times & \\ & & & \ddots & \ddots \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \times & \cdots \\ v_1 & \times & \times & \times & \cdots \\ & \times & \times & \times & \\ & & \times & \times & \\ & & & \ddots & \ddots \end{bmatrix}.$$

  The constraints (2.1) are clearly satisfied. For simplicity we include the case $k = 2$.

- Assume $k = 2$. The matrix $\tilde{A}^{(2)} = G_2 G_1 RT$ will be of the following form (in the right-hand matrix we denote the element in position $(3, 2)$ by v_2):

$$\tilde{A}^{(2)} = \begin{bmatrix} \times & \times & \times & \times & \cdots \\ c_2 v_1 & \times & \times & \times & \cdots \\ s_2 v_1 & \times & \times & \times & \\ & & \times & \times & \\ & & & \ddots & \ddots \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \times & \cdots \\ c_2 v_1 & \times & \times & \times & \cdots \\ s_2 v_1 & v_2 & \times & \times & \\ & & \times & \times & \\ & & & \ddots & \ddots \end{bmatrix}.$$

- By induction we assume that the statement holds for $k-1$. Let us show rows $k-2, k-1, k$, and $k+1$ of the matrix $\tilde{A}^{(k-1)}$:

$$\begin{bmatrix} & & & \ddots & & & & \\ c_{k-2}s_{k-3}\cdots s_3 s_2 v_1 & c_{k-2}s_{k-3}\cdots s_3 v_2 & \cdots & c_{k-2}v_{k-3} & \times & & & \\ c_{k-1}s_{k-2}\cdots s_3 s_2 v_1 & c_{k-1}s_{k-2}\cdots s_3 v_2 & \cdots & c_{k-1}s_{k-2}v_{k-3} & c_{k-1}v_{k-2} & \times & & \\ s_{k-1}s_{k-2}\cdots s_3 s_2 v_1 & s_{k-1}s_{k-2}\cdots s_3 v_2 & \cdots & s_{k-1}s_{k-2}v_{k-3} & s_{k-1}v_{k-2} & v_{k-1} & \times & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \times & \ddots \end{bmatrix}$$

  Performing the Givens transformation G_k on $\tilde{A}^{(k-1)}$ gives us the matrix $\tilde{A}^{(k)}$ which is of the following form (only rows $k-1, k, k+1$, and $k+1$ are depicted):

$$\begin{bmatrix} & & & \ddots & & & & \\ c_{k-1}s_{k-2}\cdots s_3 s_2 v_1 & c_{k-1}s_{k-2}\cdots s_3 v_2 & \cdots & c_{k-1}v_{k-2} & \times & & & \\ c_k s_{k-1}s_{k-2}\cdots s_3 s_2 v_1 & c_k c_{k-1}s_{k-2}\cdots s_3 v_2 & \cdots & c_k s_{k-1}v_{k-2} & c_k v_{k-1} & \times & & \\ s_k s_{k-1}s_{k-2}\cdots s_3 s_2 v_1 & s_k s_{k-1}s_{k-2}\cdots s_3 v_2 & \cdots & s_k s_{k-1}v_{k-2} & s_k v_{k-1} & v_k & \times & \\ 0 & 0 & \dots & 0 & 0 & 0 & \times & \ddots \end{bmatrix}$$

  One can check that the resulting matrix is lower quasi-separable. This concludes the induction procedure.

Based on the induction procedure we can conclude that the matrix $L^{-1}T$ is lower quasi-separable.

Due to the fact that the matrix L^{-T} is upper triangular, it is obvious that a multiplication of $L^{-1}T$ on the right with the matrix L^{-T} does not change the low rank structure below the diagonal. Hence we have proved that our matrix $A = L^{-1}TL^{-T}$ has the lower triangular part of quasi-separable form. Due to symmetry also the upper triangular part satisfies these constraints, and hence the complete matrix is quasi-separable. □

Let us now see how we can effectively represent a quasi-separable matrix and how to compute this representation.

2.2. The Givens-vector representation. We proved in the previous theorem that the resulting matrix is of quasi-separable form. To be able to work with the matrix an effective representation of the low rank part is necessary. A straightforward choice might be to represent the low rank part as coming from a rank-one matrix. This means representing the lower triangular part as coming from $\mathbf{u}\mathbf{v}^T$, with $\mathbf{u}$ and $\mathbf{v}$ two vectors. This is, however, a bad choice. First, this representation does not cover all kinds of quasi-separable matrices (consider, e.g., a tridiagonal matrix), and second, it suffers heavily from numerical instabilities, when computing, e.g., the spectrum via a QR-method for quasi-separable matrices. More information on the problems with this representation can be found in [55]. There exist various kinds of other suitable representations, such as the quasi-separable [26, 19], diagonal-subdiagonal [46, 32], and Givens-vector representation [55, 16].

In this manuscript we will focus on the Givens-vector representation. There are several reasons to prefer this representation above the other ones. First, it uses $3n-3$ parameters for representing a symmetric quasi-separable matrix, whereas the quasi-separable representation uses $4n-2$ parameters. These extra $n+1$ parameters used in the quasi-separable representation are therefore not uniquely determined. A good choice of these $n+1$ (almost free) parameters is important for numerical stability reasons. Second, it can be considered as an extension of the straightforward $\mathbf{u}\mathbf{v}^T$ representation of a rank-one matrix. A third reason for using the Givens-vector representation is that it is based on unitary transformations, which are stable, and a single vector. Possible numerical instabilities occur only in the presence of small or large elements in the vector and are therefore easily recognized. In the numerical experiments we will see that the representation provides accurate results, even for some very ill-conditioned problems.

It is important to note that the Givens-vector representation is strongly related to the QR-factorization of the considered low rank part. We will come back to this at the end of this section. A final reason for choosing the Givens-vector representation is the ease of constructing the representation in this case. This will be shown further in the text.

Let us briefly recapitulate some of the results for this representation. We will first show how to reconstruct a low rank part based on Givens transformations and a vector. Secondly we will show how to retrieve the representation given a matrix of, e.g., quasi-separable form.

To represent the strictly lower triangular part of the quasi-separable matrix, we will use a representation consisting of $n-2$ Givens transformations $G = [G_1, \ldots, G_{n-2}]$ and a vector $\mathbf{v} = [v_1, \ldots, v_{n-1}]$ of length $n-1$. The diagonal of the quasi-separable matrix is stored separately, leading to a global storage of $2n-1$ elements and $n-2$ Givens transformations. To represent the matrices T and S, $4n-2$ elements are needed; to represent the quasi-separable matrix $L^{-1}TL^{-T}$, essentially only $3n-3$ parameters are used. For actual implementations, however, we will not store the Givens transformations by a single parameter, but both the sine and cosine will be stored. So, in the practical implementation $4n-5$ parameters are used.

The following figures denote how the strictly lower triangular part of the matrix can be reconstructed. We show here only the strictly lower triangular part of the quasi-separable matrix. Initially one starts on the first two rows of the strictly lower triangular part. The element v_1 is placed in the upper left position, then a Givens transformation is applied, and finally, to complete the first step, element v_2 is added in position $(2,1)$. Only the first two columns and rows are shown here:

$$\begin{bmatrix} v_1 & 0 \\ 0 & 0 \end{bmatrix} \to G_1 \begin{bmatrix} v_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & v_2 \end{bmatrix} \to \begin{bmatrix} c_1v_1 & 0 \\ s_1v_1 & v_2 \end{bmatrix}.$$

The second step consists of applying the Givens transformation G_2 to the second and the third row; furthermore, v_3 is added in position $(3,3)$. Here only the first three columns are shown and the second and third row. This leads to

$$\begin{bmatrix} s_1v_1 & v_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \to G_2 \begin{bmatrix} s_1v_1 & v_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & v_3 \end{bmatrix} \to \begin{bmatrix} c_2s_1v_1 & c_2v_2 & 0 \\ s_2s_1v_1 & s_2v_2 & v_3 \end{bmatrix}.$$

This process can be repeated by applying the Givens transformation G_3 to the third and the fourth row of the matrix, and afterwards adding the diagonal element v_4. After applying all the Givens transformations and adding all the diagonal elements, the strictly lower triangular part of a quasi-separable matrix has been constructed. Because of the symmetry also the strictly upper triangular part is known. Finally one obtains a lower triangular matrix which equals the strictly lower triangular part of the matrix $L^{-1}TL^{-T}$:

$$\begin{bmatrix} c_1v_1 & & & \\ c_2s_1v_1 & c_2v_2 & & \\ c_3s_2s_1v_1 & c_3s_2v_2 & c_3v_3 & \\ \vdots & \vdots & & \ddots \end{bmatrix}. \tag{2.2}$$

We remark that the construction of the quasi-separable part of the matrix resembles the application of the Givens transformations to the Hessenberg matrix in the proof of Theorem 2.3.

Once the Givens-vector representation of, e.g., quasi-separable or semiseparable matrices is known, many techniques exist for solving systems of equations and computing eigenvalues [53]. How to retrieve the representation is, however, not always trivial. A general technique for retrieving a low rank representation, of undetermined rank, is based on cross-approximation. This is a relatively cheap technique and suitable for a wide variety of structured rank matrices [49, 48]. In many cases, however, the Givens-vector, the $\mathbf{u}\mathbf{v}^T$, or the quasi-separable representation is already known, e.g., as the result of inverting a band matrix [43] or applying an orthogonal similarity reduction to a semiseparable form [50]. We will not go into the details (see [57] for a complete exposition on several representations and how to compute them), but for semiseparable matrices as in Definition 2.2 there is a close relation between the QR-factorization and the Givens-vector representation. More precisely if one computes the QR-factorization of the semiseparable matrix by a sequence of bottom-up Givens transformations (exactly $n-1$ are needed), then these Givens transformations correspond to the Givens transformations of the Givens-vector representation. The vector can then be computed by some minor calculations. The fact that the Givens transformations of the QR-factorization coincide with the Givens transformations of the Givens-vector representation will come in handy when computing the representation of the quasi-separable matrix. For a quasi-separable matrix A also such a relation holds, but in this case one needs to compute the QR-factorization of $A(2:n, 1:n-1)$.

2.3. Computing the Givens-vector representation. Let us now show how to compute the representation of the quasi-separable matrix. Several things need to be computed. To obtain the Givens-vector representation of the matrix $L^{-1}TL^{-T}$, we need to compute intermediate representations of the matrices L^{-1} and TL^{-T}. These representations are used for computing the Givens-vector representation of $L^{-1}TL^{-T}$.

We will depict now all consecutive steps for computing the representation. Pseudocode for the implementation can be found in subsection 4.1.

- **Compute the Givens-vector representation for L^{-1}.** The matrix L is lower bidiagonal; hence its inverse is a lower semiseparable matrix. Therefore, this matrix can be represented by a Givens-vector representation.
    - [illegible] ($8n-8$ operations.[2]) There exists a sequence of Givens transformations $G_{n-1} \dots G_1$ applied on the right of the matrix L such that $LG_{n-1} \dots G_1 = B$ is an upper bidiagonal matrix. This corresponds to computing the RQ-factorization of the matrix[3] $L = B(G_1^T \dots G_{n-1}^T)$. The Givens transformation G_{n-1} works on the last two columns, the transformation G_{n-2} on columns $n-2$ and $n-1$ and so forth. Inverting L leads to

      $$L^{-1} = G_{n-1} \dots G_2 G_1 B^{-1}.$$

      Since the right-hand side is in fact a QR-factorization of the matrix L^{-1}, the Givens transformations $G_1, \dots, G_{n-1}$ coincide with the Givens transformations in the Givens-vector representation of L^{-1}. In the remaining calculations often the tangent of the corresponding Givens transformations is desired. Let us denote s_i as the sine, c_i as the cosine, and $\tau_i = s_i/c_i$ as the tangent corresponding to Givens transformation G_i.
    - [illegible] As the diagonal elements of L^{-1} are the inverses of the diagonal elements of L, one has as representation the Givens transformations from above, with the corresponding vector:

      $$\mathbf{v}_L = \left[\frac{1}{l_{11}c_1}, \frac{1}{l_{22}c_2}, \dots, \frac{1}{l_{n-1,n-1}c_{n-1}}, \frac{1}{l_{nn}}\right], \tag{2.3}$$

      with $L = (l_{ij})$. We remark that these computations are well defined, as one can easily verify that all cosines in the different Givens transformations are different from zero. A cosine equal to zero translates to the fact that a diagonal element of L needed to be zero, which is not possible due to the positive definiteness of S.

      However, in the next bullets we will notice that the division by the cosines c_i is not necessary. It creates extra operations and can cause numerical instabilities in case of small cosines. In fact, in most of the

[2] An operation consists of performing one of the following operations: $+, -, \times, /$.

[3] One can also consider $L\,(G_{n-1} \dots G_1) = B$ as the LQ-factorization of the matrix B.

computations only the diagonal of L^{-1} is involved. Hence, we do not compute the vector from the Givens-vector representation. We keep this in mind and compute these elements only when absolutely necessary.

- **Compute the representation of the matrix TL^{-T}.** The matrix TL^{-T} is a Hessenberg matrix having the upper triangular part of quasi-separable form. Hence we have to compute the Givens-vector representation for the strictly upper triangular part, the diagonal and subdiagonal elements of the Hessenberg-matrix $H = TL^{-T}$. Define $T = (t_{ij})$ and $H = (h_{ij})$. Some intermediate values for reducing the computations are stored in $\alpha, \beta, \gamma, \delta$. These extra vectors require an additional $4n$ memory but can reduce the computational cost from $37n - 49$ operations to $30n - 49$, an improvement of almost 20%.
  - [illegible] ($5n - 5$ operations.) Straightforward computations lead to

$$\begin{aligned} h_{11} &= \beta_1, \quad \text{with} \quad \beta_1 = \frac{t_{11}}{l_{11}}, \\ h_{ii} &= \alpha_i \; c_i + \beta_i, \quad \text{with} \quad \alpha_i = \frac{t_{i,i-1} \; \tau_{i-1}}{l_{i-1,i-1}}, \; \beta_i = \frac{t_{ii}}{l_{ii}}, \\ & \qquad \text{(for all} \quad i = 2, \dots, n-1) \\ h_{nn} &= \alpha_n + \beta_n, \quad \text{with} \quad \alpha_n = \frac{t_{n,n-1} \; \tau_{n-1}}{l_{n-1,n-1}}, \; \beta_n = \frac{t_{nn}}{l_{nn}}. \end{aligned}$$

    The above equations show that using the vector $\mathbf{v}_L$ in (2.3) from the Givens-vector representation of L^{-1} would only complicate matters.
  - [illegible] (Not computed.) The subdiagonal elements can be computed as follows:

$$h_{i+1,i} = \frac{t_{i+1,i}}{l_{ii}} \qquad \text{(for all} \quad i = 1, \dots, n-1).$$

    However, they are not essential for the construction of the quasi-separable matrix, due to symmetry of the final matrix $L^{-1}TL^{-T}$.
  - [illegible] ($7n - 12$ operations.) The Givens transformations are exactly the same as the one used above, only one fewer is required: $G_2, G_3, \dots, G_{n-1}$. The vector of the representation of this upper triangular part can be obtained by computing the superdiagonal elements of the matrix TL^{-T} and dividing them by the corresponding cosines of the Givens transformations. As before, we will not, however, divide these elements by the cosines, but continue working with the superdiagonal elements and perform the division only when absolutely necessary.
    The following formulas compute the superdiagonal elements:

$$\begin{aligned} h_{1,2} &= \beta_1 \; \tau_1 \; c_2 \; + \frac{t_{12}}{l_{22}}, \\ h_{i,i+1} &= \alpha_i \; \gamma_i + \beta_i \; \tau_i \; c_{i+1} + \frac{t_{i,i+1}}{l_{i+1,i+1}}, \quad \text{with} \quad \gamma_i = s_i c_{i+1} \\ & \qquad \text{(for} \quad i = 2, \dots, n-2), \\ h_{n-1,n} &= \alpha_{n-1} \; s_{n-1} + \beta_{n-1} \; \tau_{n-1} + \frac{t_{n-1,n}}{l_{n,n}}. \end{aligned}$$

- **Compute the representation of the matrix $A = L^{-1}TL^{-T}$.** It was proven before that the matrix is a symmetric quasi-separable matrix. Hence we need the Givens-vector representation of the strictly lower triangular part as well as the diagonal elements.
  - [illegible] $(7n-11$ operations.) To compute the diagonal elements, one can use the following loop (with t a temporary variable):

    $$\begin{aligned} a_{11} &= \frac{h_{11}}{l_{11}}, \\ t &= \delta_1, \quad \text{with} \quad \delta_1 = \frac{\tau_1\, h_{12}}{l_{11}}. \end{aligned}$$

    Iterate now for $i = 2, \ldots, n-2$ the following equations:

    $$\begin{aligned} a_{ii} &= \frac{h_{ii}}{l_{i,i}} + t\, c_i, \\ t &= (t\, \gamma_i + \delta_i\ )\, \tau_i, \quad \text{with} \quad \delta_i = \frac{h_{i,i+1}}{l_{i,i}}. \end{aligned}$$

    Finally we have the last two diagonal elements:

    $$\begin{aligned} a_{n-1,n-1} &= \frac{h_{n-1,n-1}}{l_{n-1,n-1}} + t\, c_{n-1}, \\ t &= (t\, s_{n-1} + \delta_{n-1})\, \tau_{n-1}, \quad \text{with} \quad \delta_{n-1} = \frac{h_{n-1,n}}{l_{n-1,n-1}}, \\ a_{nn} &= \frac{h_{nn}}{l_{nn}} + t. \end{aligned}$$
  - [illegible] $(3n-3$ operations.) Based on the relations discussed before we obtain

    $$\begin{aligned} L^{-1}TL^{-T} &= \left(BG_1^TG_2^T \ldots G_{n-1}^T\right)^{-1} TL^{-T} \\ &= G_{n-1} \ldots G_2 \left(G_1 B^{-1} T L^{-T}\right). \end{aligned}$$

    Combining the factors $G_1B^{-1}TL^{-T} = H$, we get an upper Hessenberg matrix. Hence it is clear due to construction that the Givens transformations G_2 up to G_{n-1} are the Givens transformations needed for the Givens-vector representation of the quasi-separable part in the matrix. The multiplication between the lower semiseparable matrix L^{-1} and the strictly upper triangular part of (TL^{-T}), which is of quasi-separable form, can be done in $\mathcal{O}(n)$ operations. Writing down the lower semiseparable matrix and the strictly upper triangular part of the matrix TL^{-T} as in (2.2), one can easily deduce a simple loop which computes the subdiagonal elements of the new quasi-separable matrix. Based on these subdiagonal elements and on the fact that the cosines of the Givens transformations are different from zero, one can easily obtain the representation of the strictly lower triangular part.

    The sub- or superdiagonal elements can be computed as follows:

    $$\begin{aligned} a_{12} &= \delta_1, \\ a_{i+1,i} &= a_{i,i-1}\, \tau_{i-1}\, \gamma_i + \delta_i \qquad (\text{for} \quad i = 2, \ldots, n-1), \\ a_{n,n-1} &= a_{n-1,n-2}\, \tau_{n-2}\, s_{n-1} + \delta_{n-1}. \end{aligned}$$

Summarizing, the complexity of computing the Cholesky decomposition of the positive definite tridiagonal matrix takes $5n - 3$ operations. Computing the representation of the quasi-separable matrix takes $30n - 49$ operations. Hence the total cost for computing the representation of the quasi-separable matrix is $35n - 52$ operations.

Traditionally, one assumed that the approach of computing the eigenvalues via $L^{-1}TL^{-T}$ was too expensive because the reduction to tridiagonal form of a dense matrix already took $\mathcal{O}(n^3)$ operations. This reduction was essential before being able to compute the spectrum in $\mathcal{O}(n^2)$ operations, via, for example, a divide-and-conquer technique or a QR-algorithm. Using the method presented above, however, we see that it takes $\mathcal{O}(n)$ operations to obtain the representation of the quasi-separable matrix. For this quasi-separable matrix there exist techniques $\mathcal{O}(n^2)$ for computing the whole spectrum. In the next section we will briefly discuss some of these methods.

3. Computing the eigenvalues of a quasi-separable matrix. We do not go into the details of how to compute the eigenvalues of quasi-separable matrices. Pointers to manuscripts in which all the essential information can be found will be given. For solving the overall problem, in fact, only the eigenvalues of the quasi-separable matrix are necessary. The eigenvectors can be computed afterwards, based on the equation $T\mathbf{x} = \lambda S\mathbf{x}$, once the eigenvalues are known. This can be done in $\mathcal{O}(n^2)$ operations.

Alternatively, one can also compute the eigenvectors of the quasi-separable matrix and then transform them back to the original problem at a cost of $\mathcal{O}(n^2)$, exploiting the bidiagonal structure of L. The cost of this second approach is dependent on how efficiently the eigenvectors of the quasi-separable matrix are computed. This can lead to $\mathcal{O}(n^3)$ as well as $\mathcal{O}(n^2)$ methods; in the next subsection some complexities for different approaches are given.

3.1. Reduction to tridiagonal form. Due to the specific rank structure of the matrix A, we can reduce this matrix to tridiagonal form in $\mathcal{O}(n^2)$ operations instead of the traditional reduction, which needs $\mathcal{O}(n^3)$ operations. There exist several variants to reduce a quasi-separable matrix to tridiagonal form [30, 40]. In fact, the traditional algorithms are adapted to fully exploit the rank structure in the involved matrices. Also a parallel method to reduce a quasi-separable matrix to tridiagonal form was developed in [39], starting the reduction to tridiagonal form simultaneously at two sides of the matrix. Recently also more general reduction schemes, to reduce arbitrary structured rank matrices to tridiagonal (Hessenberg in the nonsymmetric case), were proposed [15, 25]. All presented algorithms are of order $\mathcal{O}(rn^2)$, where r is a factor related to the rank of the structured rank parts. Remember that in our case we consider a quasi-separable matrix of quasi separability rank $r = 1$.

The benefit of reducing the quasi-separable matrix to tridiagonal form is that one can use all available solvers for tridiagonal matrices. The reduction to tridiagonal form costs, however, $\mathcal{O}(n^2)$ operations, which is of the same order as computing the full spectrum. This reduction cost is not needed, however, if one applies methods which are directly applicable to the quasi-separable matrix.

On the other hand, plenty of robust and efficient methods for tridiagonal matrices exist (LAPACK/`Matlab` implementations are available); e.g., the QR-method for tridiagonal matrices is discussed in many textbooks [33, 44, 17, 58] ($\mathcal{O}(n^2)$ for computing the eigenvalues, $\mathcal{O}(n^3)$ when computing the eigenvectors by accumulating the unitary transformations performed as is done in the current LAPACK implementation), divide-and-conquer methods [5, 6, 35, 11, 8] ($\mathcal{O}(n^2)$ if only eigenvalues are desired, in the worst case $\mathcal{O}(n^3)$ for the eigenvectors; in practice, however, the exponent is less than 3), bisection and inverse iteration [42, 13] ($\mathcal{O}(n^2)$ for the whole spectrum; in case of well separated eigenvalues it takes $\mathcal{O}(n^2)$ operations for all eigenvectors; otherwise, in case of clusters it might result in $\mathcal{O}(n^3)$ operations), the MRRR algorithm, which is an adapted version of inverse iteration, not using Gram–Schmidt orthogonalization [23, 20, 21, 22, 45] (this method requires $\mathcal{O}(n^2)$ for computing both eigenvalues and eigenvectors), etc. A recent overview of these methods, comparing them in LAPACK, was given in [18]. When applying the QR-algorithm to the resulting tridiagonal matrix, one applies in a certain sense a tuned Cholesky-QR method, exploiting the quasi-separable structure for reducing the matrix to tridiagonal form. This tuned Cholesky-QR method also involves only $\mathcal{O}(n^2)$ operations for computing the whole spectrum as well as the eigenvectors, when using, e.g., MRRR.

3.2. Applying the QR-algorithm directly on the quasi-separable matrix. The last few years people have intensively studied QR-algorithms for structured rank matrices. Let us present some of these results. There is an implicit QR-algorithm for semiseparable matrices [54] and and implicit one for semiseparable plus diagonal matrices [51]. Explicit QR-algorithms for higher order quasi-separable matrices can be found in [28, 14]. The algorithm for semiseparable plus diagonal matrices is after minor modifications also suitable for quasi-separable matrices. Recently more general types of QR-algorithms exist for low rank perturbations of unitary matrices and so forth [1, 2, 3]. Computing the eigenvalues using the QR-method for quasi-separable matrices results in an order $\mathcal{O}(n^2)$ method, whereas computing the eigenvectors by accumulating the transformations performed imposes a cubical complexity $\mathcal{O}(n^3)$.

One might be in favor of the QR-method because it is widely spread and these methods are directly applicable on the quasi-separable matrices. Unfortunately QR-methods for quasi-separable matrices have a higher computational complexity (50%) w.r.t. the QR-method than do tridiagonal matrices. Nevertheless they provide accurate results. Recently, however, also a faster algorithm for computing the spectrum of structured rank matrices was proposed [56]. The method is based on a QR-iteration driven by a rational function. The resulting method is much faster than the traditional QR-methods for rank structured matrices.

3.3. Other methods. As mentioned in the introduction also different techniques for computing the eigenvalues of quasi-separable matrices exist. For example, the bisection method and a method based on Sturm sequences can be found in [27]. With the bisection method one can compute a single eigenvalue in an interval, whereas the Sturm sequences method is an adaptation of the bisection method to compute the kth largest eigenvalue. Computing a single eigenvalue involves $\mathcal{O}(n)$ operations, whereas computing the full spectrum requires $\mathcal{O}(n^2)$ flops.

Another technique is based on halving the problem size at every step of the algorithm. These so called divide-and-conquer methods are based on solving the secular equation [41, 10].

Both methods mentioned above need $\mathcal{O}(n^2)$ operations for computing the whole spectrum. Up till now the divide-and-conquer method for quasi-separable and semiseparable plus diagonal matrices is the fastest and most accurate available method for computing the spectrum of quasi-separable matrices. Hence we choose to use this method for computing the eigenvalues of the corresponding quasi-separable matrix in the upcoming numerical experiments. Computing all eigenvectors increases the computational complexity and can lead to $\mathcal{O}(n^3)$. Using specialized techniques based on the FMM method [9, 34] as discussed in [10] can reduce the complexity of computing the eigenvectors to $\mathcal{O}(n^2 \log(n))$.

4. Numerical experiments. In this section we will present results concerning the timings and accuracy of the presented approach. We chose to use the divide-and-conquer method for computing the eigenvalues and eigenvectors of the considered quasi-separable matrices. To our knowledge this algorithm is among the fastest available methods and provides accurate results when computing the eigenvalues of quasi-separable matrices. The software was implemented in Matlab and executed on a Linux platform. The divide-and-conquer approach used is the one from [41], where we needed to adapt the software, as the implementation as presented in [41] (the straightforward solver) was based on the generator representation whereas the resulting quasi-separable matrix in our approach is represented using the Givens-vector representation.

In the following experiments the relative forward error and relative backward error are computed. First the pseudocode for generating the quasi-separable matrix is given.

Recall that the presented method and Laguerre's iteration take $\mathcal{O}(n^2)$ operations for computing the eigenvalues. The standard Cholesky-QR and QZ-approach take $\mathcal{O}(n^3)$ operations for computing the eigenvalues. We use the Matlab built-in functions for performing the QZ and Cholesky-QR approach, which are of the order $\mathcal{O}(n^3)$. We remark, however, that one can reduce the complexity count of the Cholesky-QR method from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$ when using an adapted reduction of the quasi-separable matrix to tridiagonal form. However, in this case also one has to exploit the quasi-separable structure. Even though one can reduce the Cholesky-QR complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$, the constant will be larger than the one in the quasi-separable approach combined with the divide-and-conquer method. Once the eigenvalues are known various methods exist ($\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$) for computing the eigenvectors; hence we will not go into the details related to eigenvectors.

We can, therefore, already conclude that the quasi-separable approach is as fast as Laguerre's iteration and faster than the Cholesky-QR and the QZ by a factor n (from $\mathcal{O}(n^2)$ to $\mathcal{O}(n^3)$). Keep in mind that quasi-separable techniques can be used for tuning the Cholesky-QR method in order to obtain an $\mathcal{O}(n^2)$ approach.

We will therefore not focus on the complexity of the methods, but on a comparison of the accuracy of the four methods.

4.1. The pseudocode. We will present the pseudocode for computing the representation of the quasi-separable matrix $L^{-1}TL^{-T}$. The case where $n = 1$ and $n = 2$ is trivial and not covered.

Explanation of some variables used in the pseudocode:

- dX stands for the diagonal of the matrix X;
- sd$X$ stands for super(sub)diagonal of the matrix X;
- uh stands for the upper Hessenberg matrix TL^{-T};
- Combinations lead to, for example, sbduh, which means the subdiagonal of the upper Hessenberg matrix TL^{-T}.
- G will contain the Givens transformations of the Givens-vector representation. This is a $2 \times (n - 1)$ matrix. Each column corresponds to a Givens transformation. The first element in each column contains the sine, the second element the cosine.
- v will contain the vector from the Givens-vector representation.

```
1. % Compute the Cholesky decomposition S=LL^T
     Store the diagonal in dL, the subdiagonal in sdL
2. % Compute the RQ-factorization of the matrix L
     Initialize: G(1,i-1)=1;
```

```
      for i=n,n-1,n-2,...,2
        [cosine,sine,tau(i-1)]=givens(Giv(1,i-1)*dL(i),sdL(i-1));
        Store the cosine and the sine in column i-1 of G;
      end for;
   3. % Construct the represention of the upper Hessenberg matrix T L^{-T}
      % Compute the diagonal elements of T L^{-T}
      Initialize: beta(1)=dT(1)/dL(1); duh(1)=beta(1);
      for i=2,3,...,n-1
        alpha(i)=sdT(i-1)/dL(i-1)*tau(i-1)
        beta(i)=dT(i)/dL(i);
        duh(i)=alpha(i)*G(1,i)+beta(i);
      end for
      duh(n)=sdT(n-1)/dL(n-1)*tau(n-1)+dT(n)/dL(n);
   4. % Compute the superdiagonal elements of T L^{-T}
      Initialize: sduh(1)=beta(1)*tau(1)*G(1,2)+sdT(1)/dL(2);
      for i=2,3,...,n-2
        gamma(i)=G(2,i)*G(1,i+1);
        sduh(i)=alpha(i)*gamma(i)+beta(i)*tau(i)*G(1,i+1)+sdT(i)/dL(i+1);
      end for
      sduh(n-1)=alpha(n-1)*G(2,n-1)+beta(n-1)*tau(n-1)+sdT(n-1)/dL(n);
   5. % Compute the representation for L^{-1} T L^{-T} and the diagonal
      Initialize: d(1)=duh(1)/dL(1);
      Initialize: delta(1)=sduh(1)/dL(1); tmp=tau(1)*delta(1);
      for i=2,3,...,n-2
        d(i)=duh(i)/dL(i)+tmp*G(1,i);
        tmp=(tmp*gamma(i)+delta(i))*tau(i);
      end for
      delta(n-1)=sduh(n-1)/dL(n-1);
      d(n-1)=duh(n-1)/dL(n-1)+tmp*G(1,n-1);
      tmp=(tmp*G(2,n-1)+delta(n-1))*tau(n-1);
      d(n)=duh(n)/dL(n)+tmp;
   6. % Compute the subdiagonal
      Initialize: sd(1)=delta(1);
      for i=2,3,...,n-2
        sd(i)=sd(i-1)*tau(i-1)*gamma(i)+delta(i);
      end for
      sd(n-1)=sd(n-2)*tau(n-2)*G(2,n-1)+delta(i);
   7. % Assign the Givens-vector representation
      G=G(:,2:n-1); v(1:n-2)=sd(1:n-2)./G(1,:);
```

NOTE 4.1. [illegible] sd [illegible] v [illegible] G [illegible] v [illegible] G [illegible] sd [illegible] v [illegible] $2n-2$ [illegible]

4.2. Relative forward error. In the first experiment we compared the computed eigenvalues with known eigenvalues of the problem. We solved the definite symmetric generalized eigenvalue problem with two tridiagonal Toeplitz matrices. The matrix T is constructed with a random element defining the diagonal and a random element defining the subdiagonal (generated using `rand` from MATLAB). The second matrix S has a random subdiagonal element, whereas the diagonal element is chosen $s_{ii} = 2\max(s_{i,i-1}, s_{i,i+1}) + 1$. In this way we know that the matrix S is positive definite and moreover is well conditioned.

As both Toeplitz matrices commute, we can explicitly compute the spectrum of the generalized eigenvalue problem as it equals the ratios of the eigenvalues of T and

S. To compute the eigenvalues of both Toeplitz matrices we used the explicit formulas derived in [7]. For a tridiagonal Toeplitz matrix having subdiagonal t_{-1}, diagonal t_0, and superdiagonal t_1 the eigenvalues are given by

$$t_0 + \sqrt{t_1 t_{-1}} \cos \frac{\pi i}{n+1} \qquad (\text{for} \quad i = 1, \ldots, n).$$

Based on these "correct" eigenvalues of the generalized eigenvalue problem, we performed experiments for sizes ranging from 100 to 1500, and for each experiment we performed fifteen random tests (with the constraints on S as mentioned above). Moreover, the eigenvalues of the Toeplitz matrix were computed via variable precision arithmetic in MATLAB to ensure they are correct up to machine precision (16 digits). The relative forward errors

$$\max_i \frac{|\lambda_i - \tilde{\lambda}_i|}{|\lambda_i|},$$

where λ_i denote the "correct" eigenvalues and $\tilde{\lambda}_i$ the computed eigenvalues, are plotted in Figure 4.1.

Figure 4.1 shows that the forward error can be pretty large, up to 10^{-11}. When comparing, however, against the condition number of the symmetric quasi-separable matrix, we see that the loss of digits is due to the condition number, which sometimes reaches approximately 10^5.

Comparing the quasi-separable approach with both other approaches, we see that overall it has the best forward error. In the upcoming experiment we see that the bad results are due to the conditioning of the problem, since all backward errors will be relatively good. The experiment was run for three approaches: the quasi-separable approach, the QZ-method, and Laguerre's iteration from [38].

4.3. Relative backward error. In the following set of experiments we compute a relative error involving the eigenvalues and eigenvectors. We first solve the eigenproblem involving the quasi-separable matrix A:

$$A\mathbf{y} = \lambda \mathbf{y},$$

where $A = L^{-1} T L^{-T}$ and $\mathbf{y} = L^T \mathbf{x}$.

We hence compute the eigenvalues λ_i corresponding to the eigenvectors $\mathbf{y}_i$. To obtain the eigenvectors $\mathbf{x}_i$ of the generalized eigenvalue problem, we need to compute the following:

$$\mathbf{x}_i = L^{-T} \mathbf{y}_i.$$

As we know from the theoretical results, the matrix L^{-T} is an upper triangular semiseparable matrix. Moreover, the representation in terms of Givens transformations and a vector is known, since it was computed in the algorithm explained in subsection 2.3. The multiplication between the matrix $L^{-T}\mathbf{y}$ can easily be performed in $\mathcal{O}(n)$ operations (see, e.g., [52]).

For the following set of experiments we took matrix sizes ranging from 100 to 2000, and 15 experiments for each size were considered. The tridiagonal matrix T has random diagonal and subdiagonal elements. The matrix S has random subdiagonal elements, and the diagonal elements were taken $s_{ii} = 2\max(s_{i,i-1}, s_{i,i+1}) + 1$, in order to make the matrix positive definite and well conditioned.

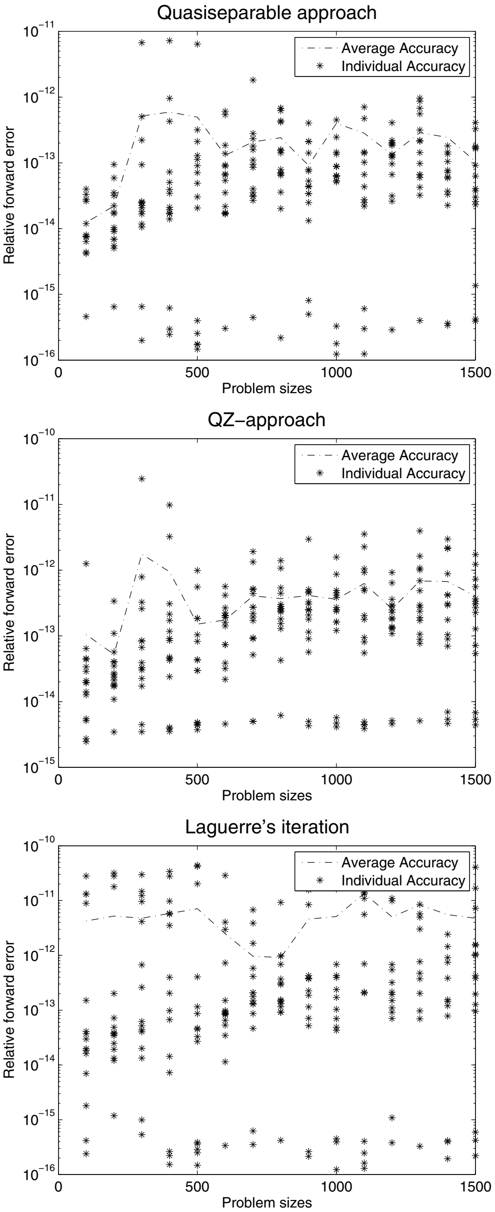


Fig. 4.1. *Relative forward errors on the eigenvalues.*

We considered the following relative backward error [37, 12]:

$$\max_i \left(\frac{\|T\mathbf{x}_i - \lambda_i S\mathbf{x}_i\|_2}{(\|T\|_2 + |\lambda_i|\ \|S\|_2)\ \|\mathbf{x}_i\|_2} \right),$$

where the eigenvectors $\mathbf{x}_i = L^{-T}\mathbf{y}_i$, eigenvalues λ_i, and eigenvectors $\mathbf{y}_i$ were computed using the presented method. The norm of the matrices T and S was estimated using the Matlab built-in function `normest`.

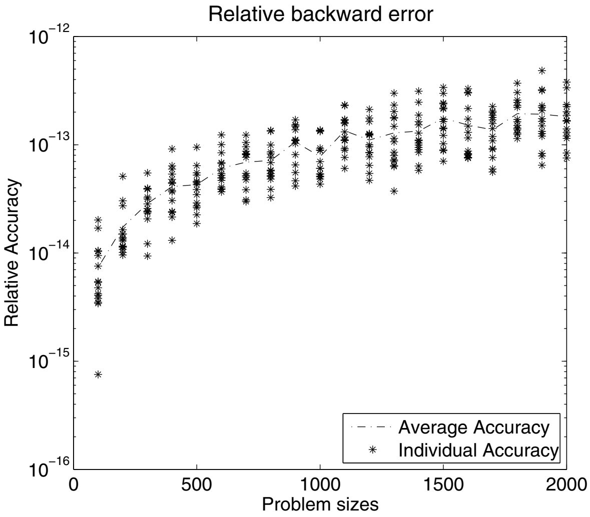


FIG. 4.2. *Relative backward error.*

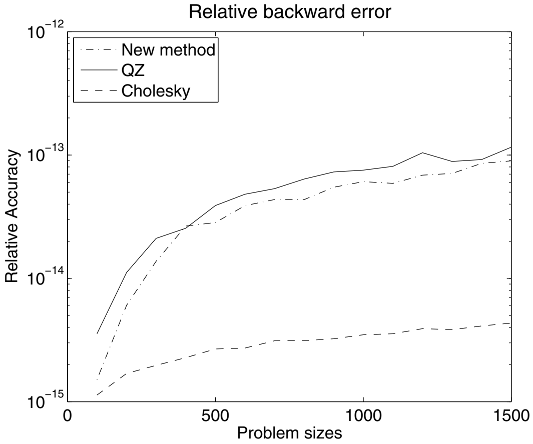


FIG. 4.3. *Comparison of three approaches.*

Figure 4.2 clearly illustrates that the quasi-separable method returns a relatively good backward error for all experiments.

4.4. Comparison of different approaches. In this experiment we will perform more specific tests for comparing the accuracy of the proposed method with some methods available in MATLAB. The matrices were generated similarly as in the previous experiment. For every matrix the eigenvalues and eigenvectors were computed via the quasi-separable method, the Cholesky-QR approach, and the QZ-method. A comparison of the relative backward error is depicted in Figure 4.3. It can be seen that the quasi-separable method has an error comparable to the QZ-method but performs worse than the Cholesky-QR method. Keep in mind that for this experiment all three methods are $\mathcal{O}(n^3)$ methods, since both the eigenvalues and eigenvectors are required. When only the eigenvalues are desired, however, the complexity of the

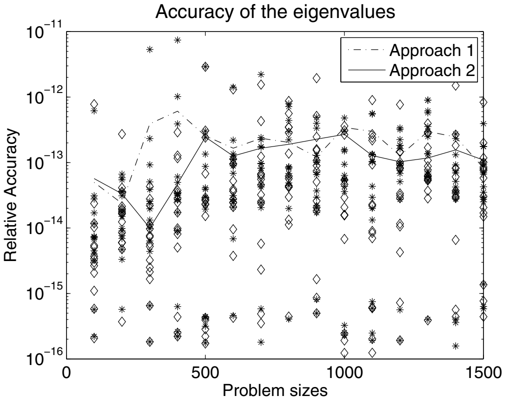


FIG. 4.4. *Comparison with another $\mathcal{O}(n^2)$ method.*

divide-and-conquer reduces to $\mathcal{O}(n^2)$, whereas both other methods remain cubical in the number of operations.

The experiments show that for this type of well-conditioned problems the Cholesky-QR method is the most accurate one. The quasi-separable method is equally accurate as the QZ-method.

4.5. Comparison with another $\mathcal{O}(n^2)$ method. In this section we will compare the method from [38] with the quasi-separable method presented in this paper. The implementation from [38] was written in Fortran, whereas the code from this article was written in MATLAB. In order to compare the accuracy of both methods we choose to embed the Fortran code in MATLAB via a MEX-file. Since the routine from [38] only computes the eigenvalues, it is impossible to use the backward error measure norm from above. The experiments from subsection 4.2 were redone, and the relative forward error for both approaches is plotted in Figure 4.4.

We see that both methods are equally accurate for the chosen experiments. Approach 1 corresponds to the quasi-separable method combined with divide-and-conquer, and Approach 2 is based on an evaluation of the determinant via Laguerre's iteration as presented in [38].

4.6. The matrix S is ill conditioned. Let us reconsider in this section Experiment 2 from [38]. We compute the eigenvalues of the generalized definite eigenproblem in which T is a tridiagonal Toeplitz matrix having 4 on the diagonal and 1 on the subdiagonal. The matrix S is almost Toeplitz, having the diagonal elements equal to $2 \cdot 10^{-10}$, the subdiagonal elements equal to 10^{-10}, and $s_{11} = s_{nn} = 1$. The condition number of the matrix S is approximately 10^{12}. First we run tests for several sizes of the matrices, ranging from 100 to 2000 via steps of size 50. The relative backward error was computed and plotted for the divide-and-conquer based on quasi-separable matrices, the QZ-method, and the Cholesky-QR method. Results are presented in Figure 4.5.

The QZ-method performs the best in this case. The Cholesky-QR method seems to have a lot of difficulties in finding approximations of the eigenvalues and seems to provide less accurate results. The quasi-separable method performs significantly better than the Cholesky-QR method.

We reconsidered this type of experiment, but now we compared the quasi-separable

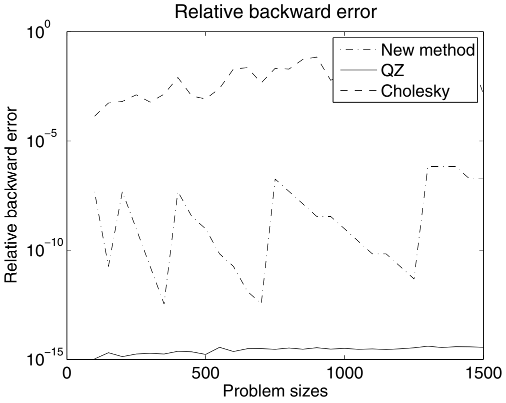


Fig. 4.5. *Comparison for ill-conditioned S (subdiagonal size* 10^{-10}*).*

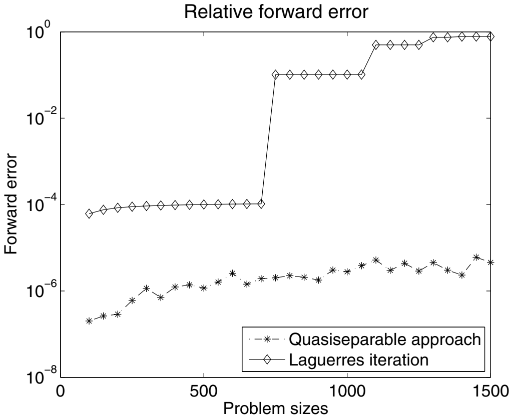


Fig. 4.6. *Comparison for ill-conditioned S.*

approach with the one based on Laguerre's iteration from [38]. Based on the previous experiment, we consider the eigenvalues computed by the QZ-method to be the correct ones λ_i. In Figure 4.6 you see the relative forward error.

When comparing the forward error, we see that the quasi-separable method performs the best. Moreover, when the dimension of the problem increases, Laguerre's iteration method seems to get into problems and no longer provides accurate results.

5. Conclusions. In this manuscript we showed that one can solve the definite generalized tridiagonal symmetric eigenvalue problem by transforming it into a standard symmetric eigenvalue problem and exploiting the quasi-separable structure of the coefficient matrix. The eigenvalues and eigenvectors can be computed using various methods. In this manuscript we used the divide-and-conquer method for computing the eigenvalues in $\mathcal{O}(n^2)$ operations.

It was shown that the quasi-separable method involves $\mathcal{O}(n^2)$ operations, the traditional Cholesky-QR method as well as the QZ-method $\mathcal{O}(n^3)$ operations, and the tuned Cholesky-QR method exploiting the quasi-separable structure uses $\mathcal{O}(n^2)$ operations. Among all these methods the quasi-separable approach is the fastest one and, moreover, more accurate than the Cholesky-QR approach in case of an ill-conditioned matrix S.

A final comparison between the quasi-separable approach and another $\mathcal{O}(n^2)$ method was given, showing that both methods are of a comparable accuracy in the well-conditioned case, but the quasi-separable method performs better in the ill-conditioned case.

Acknowledgments. The authors wish to thank Françoise Tisseur for fruitful discussions at the ILAS meeting in The Netherlands, 2006, and for the useful comments which improved the overall presentation of the manuscript. We also thank Prof. Zhonggang Zeng for providing us the Fortran code of [38].

REFERENCES

[1] D. A. Bini, F. Daddi, and L. Gemignani, *On the shifted QR iteration applied to companion matrices*, Electron. Trans. Numer. Anal., 18 (2004), pp. 137–152.

[2] D. A. Bini, Y. Eidelman, L. Gemignani, and I. Gohberg, *Fast QR eigenvalue algorithms for Hessenberg matrices which are rank-one perturbations of unitary matrices*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 566–585.

[3] D. A. Bini, L. Gemignani, and V. Y. Pan, *Fast and stable QR eigenvalue algorithms for generalized companion matrices and secular equations*, Numer. Math., 100 (2005), pp. 373–408.

[4] R. F. Boisvert, *Families of high order accurate discretizations of some elliptic problems*, SIAM J. Sci. Stat. Comput., 2 (1981), pp. 268–284.

[5] C. F. Borges and W. B. Gragg, *Divide and conquer for generalized real symmetric definite tridiagonal eigenproblems*, in Proceedings of the '92 Shanghai International Numerical Algebra and Its Applications Conference, E.-X. Jiang, ed., China Science and Technology Press, Beijing, China, 1992, pp. 70–76.

[6] C. F. Borges and W. B. Gragg, *A parallel divide and conquer algorithm for the generalized real symmetric definite tridiagonal eigenproblem*, in Numerical Linear Algebra and Scientific Computing, L. Reichel, A. Ruttan, and R. S. Varga, eds., De Gruyter, Berlin, Germany, 1993, pp. 11–29.

[7] A. Böttcher and S. M. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM, Philadelphia, PA, 2005.

[8] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenvalue problem*, Numer. Math., 31 (1978), pp. 31–48.

[9] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 669–686.

[10] S. Chandrasekaran and M. Gu, *A divide-and-conquer algorithm for the eigendecomposition of symmetric block-diagonal plus semiseparable matrices*, Numer. Math., 96 (2004), pp. 723–731.

[11] J. J. M. Cuppen, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[12] P. I. Davies, N. J. Higham, and F. Tisseur, *Analysis of the Cholesky method with iterative refinement for solving the symmetric definite generalized eigenproblem*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 472–493.

[13] J.-M. Delosme and I. C. F. Ipsen, *From Bareiss' algorithm to the stable computation of partial correlations*, J. Comput. Appl. Math., 27 (1989), pp. 53–91.

[14] S. Delvaux and M. Van Barel, *The Explicit QR-algorithm for Rank Structured Matrices*, Technical Report TW459, Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium, 2006.

[15] S. Delvaux and M. Van Barel, *A Hessenberg reduction algorithm for rank structured matrices*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 895–926.

[16] S. Delvaux and M. Van Barel, *A Givens-weight representation for rank structured matrices*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1147–1170.

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

[18] J. W. Demmel, O. A. Marques, B. N. Parlett, and C. Vömel, *Performance and accuracy of LAPACK's symmetric tridiagonal eigensolvers*, SIAM J. Sci. Comput., 30 (2008), pp. 1508–1526.

[19] P. Dewilde and A.-J. van der Veen, *Time-Varying Systems and Computations*, Kluwer Academic Publishers, Boston, MA, 1998.

[20] I. S. Dhillon, *A New $O(n^2)$ Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Ph.D. thesis, Department of Computer Science, University of California, Berkeley, 1989.

[21] I. S. Dhillon and B. N. Parlett, *Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices*, Linear Algebra Appl., 387 (2004), pp. 1–28.

[22] I. S. Dhillon and B. N. Parlett, *Orthogonal eigenvectors and relative gaps*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 858–899.

[23] I. S. Dhillon, B. N. Parlett, and C. Vömel, *The design and implementation of the MRRR algorithm*, ACM Trans. Math. Software, 32 (2006), pp. 533–560.

[24] Y. Eidelman, *Fast recursive algorithm for a class of structured matrices*, Appl. Math. Lett., 13 (2000), pp. 57–62.

[25] Y. Eidelman, L. Gemignani, and I. C. Gohberg, *On the fast reduction of a quasi-separable matrix to Hessenberg and tridiagonal forms*, Linear Algebra Appl., 420 (2007), pp. 86–101.

[26] Y. Eidelman and I. C. Gohberg, *On a new class of structured matrices*, Integral Equations Operator Theory, 34 (1999), pp. 293–324.

[27] Y. Eidelman, I. C. Gohberg, and V. Olshevsky, *Eigenstructure of order-one-quasi-separable matrices. Three-term and two-term recurrence relations*, Linear Algebra Appl., 405 (2005), pp. 1–40.

[28] Y. Eidelman, I. C. Gohberg, and V. Olshevsky, *The QR iteration method for Hermitian quasi-separable matrices of an arbitrary order*, Linear Algebra Appl., 404 (2005), pp. 305–324.

[29] L. Elsner, A. Fasse, and E. Langmann, *A divide-and-conquer method for the tridiagonal generalized eigenvalue problem*, J. Comput. Appl. Math., 86 (1997), pp. 141–148.

[30] D. Fasino, N. Mastronardi, and M. Van Barel, *Fast and stable algorithms for reducing diagonal plus semiseparable matrices to tridiagonal and bidiagonal form*, in Fast Algorithms for Structured Matrices: Theory and Applications, Contemp. Math. 323, American Mathematical Society, Providence, RI, 2003, pp. 105–118.

[31] M. Fiedler and T. L. Markham, *Completing a matrix when certain entries of its inverse are specified*, Linear Algebra Appl., 74 (1986), pp. 225–237.

[32] M. Fiedler and Z. Vavřín, *Generalized Hessenberg matrices*, Linear Algebra Appl., 380 (2004), pp. 95–105.

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

[34] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.

[35] 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.

[36] L. Kaufman, *An algorithm for the banded symmetric generalized matrix eigenvalue problem*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 372–389.

[37] S.-L. Lei and X.-Q. Jin, *BCCB preconditioners for systems of BVM-based numerical integrators*, Numer. Linear Algebra Appl., 11 (2004), pp. 25–40.

[38] K. Li, T.-Y. Li, and Z. Zeng, *An algorithm for the generalized symmetric tridiagonal eigenvalue problem*, Numer. Algorithms, 8 (1994), pp. 269–291.

[39] N. Mastronardi, S. Chandrasekaran, and S. Van Huffel, *Fast and stable two-way algorithm for diagonal plus semi-separable systems of linear equations*, Numer. Linear Algebra Appl., 8 (2001), pp. 7–12.

[40] N. Mastronardi, S. Chandrasekaran, and S. Van Huffel, *Fast and stable reduction of diagonal plus semi-separable matrices to tridiagonal and bidiagonal form*, BIT, 41 (2003), pp. 149–157.

[41] N. Mastronardi, M. Van Barel, and E. Van Camp, *Divide-and-conquer algorithms for computing the eigendecomposition of symmetric diagonal-plus-semiseparable matrices*, Numer. Algorithms, 39 (2005), pp. 379–398.

[42] N. Mastronardi, M. Van Barel, E. Van Camp, and R. Vandebril, *On computing the eigenvectors of a class of structured matrices*, J. Comput. Appl. Math., 189 (2006), pp. 580–591.

[43] G. Meurant, *A review of the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.

[44] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Appl. Math. 20, SIAM, Philadelphia, PA, 1998.

[45] B. N. Parlett and I. S. Dhillon, *Relatively robust representations of symmetric tridiagonals*, Linear Algebra Appl., 309 (2000), pp. 121–151.

[46] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*, Akademie-Verlag, Berlin, Germany, 1991.

[47] Y. M. Ram and G. M. L. Gladwell, *Constructing a finite element model of a vibratory rod from eigendata*, J. Sound Vibration, 169 (1994), pp. 229–237.

[48] E. E. Tyrtyshnikov, *Incomplete cross approximation in the mosaic-skeleton method*, Computing, 64 (2000), pp. 367–380.

[49] E. E. Tyrtyshnikov, *Mosaic ranks for weakly semiseparable matrices*, in Large-Scale Scientific Computations of Engineering and Environmental Problems II, M. Griebel, S. Margenov, and P. Y. Yalamov, eds., Notes on Numer. Fluid Mech. 73, Vieweg, Braunschweig, Germany, 2000, pp. 36–41.

[50] M. Van Barel, R. Vandebril, and N. Mastronardi, *An orthogonal similarity reduction of a matrix into semiseparable form*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 176–197.

[51] E. Van Camp, M. Van Barel, R. Vandebril, and N. Mastronardi, *An Implicit QR-Algorithm for Symmetric Diagonal-Plus-Semiseparable Matrices*, Technical Report TW419, Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium, 2005.

[52] R. Vandebril, *Semiseparable Matrices and the Symmetric Eigenvalue Problem*, Ph.D. thesis, Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium, 2004.

[53] R. Vandebril, M. Van Barel, G. H. Golub, and N. Mastronardi, *A bibliography on semiseparable matrices*, Calcolo, 42 (2005), pp. 249–270.

[54] R. Vandebril, M. Van Barel, and N. Mastronardi, *An implicit QR-algorithm for symmetric semiseparable matrices*, Numer. Linear Algebra Appl., 12 (2005), pp. 625–658.

[55] R. Vandebril, M. Van Barel, and N. Mastronardi, *A note on the representation and definition of semiseparable matrices*, Numer. Linear Algebra Appl., 12 (2005), pp. 839–858.

[56] R. Vandebril, M. Van Barel, and N. Mastronardi, *A New Iteration for Computing the Eigenvalues of Semiseparable (Plus Diagonal) Matrices*, Technical Report TW507, Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium, 2007.

[57] R. Vandebril, M. Van Barel, and N. Mastronardi, *Matrix Computations and Semiseparable Matrices, Volume* I: *Linear Systems*, Johns Hopkins University Press, Baltimore, MD, 2008.

[58] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 1, pp. 175–201

STRUCTURED HÖLDER CONDITION NUMBERS FOR MULTIPLE EIGENVALUES*

DANIEL KRESSNER[†], MARÍA JOSÉ PELÁEZ[‡], AND JULIO MORO[‡]

Abstract. The sensitivity of a multiple eigenvalue of a matrix under perturbations can be measured by its Hölder condition number. Various extensions of this concept are considered. A meaningful notion of structured Hölder condition numbers is introduced, and it is shown that many existing results on structured condition numbers for simple eigenvalues carry over to multiple eigenvalues. The structures investigated in more detail include real, Toeplitz, Hankel, symmetric, skew-symmetric, Hamiltonian, and skew-Hamiltonian matrices. Furthermore, unstructured and structured Hölder condition numbers for multiple eigenvalues of matrix pencils are introduced. Particular attention is given to symmetric/skew-symmetric, Hermitian, and palindromic pencils. It is also shown how matrix polynomial eigenvalue problems can be covered within this framework.

Key words. spectral condition number, sensitivity, structured matrices, Toeplitz matrices, Hankel matrices, generalized eigenvalue problem, matrix pencils, palindromic pencils

AMS subject classifications. 65F15, 15A18

DOI. 10.1137/060672893

1. Introduction.

Eigenvalue condition numbers asymptotically measure the sensitivity of an eigenvalue with respect to perturbations. If λ is a simple eigenvalue of a matrix $A \in \mathbb{C}^{n\times n}$, then it is well known that λ is differentiable with respect to perturbations in A and that the eigenvalue $\hat{\lambda}(\epsilon)$ of the perturbed matrix $A + \epsilon E$ admits the expansion

$$\hat{\lambda} = \lambda + (y^H E x)\epsilon + O(\epsilon^2), \quad \epsilon \to 0, \tag{1}$$

where x and y are, respectively, a right and a left eigenvector of A (normalized so that $|y^H x| = 1$) corresponding to λ. Then the absolute condition number for λ, defined as

$$\kappa(A, \lambda) = \lim_{\epsilon\to 0} \sup_{\substack{\|E\|\le 1\\ E\in\mathbb{C}^{n\times n}}} \frac{|\hat{\lambda} - \lambda|}{\epsilon}, \tag{2}$$

is given by $\kappa(A, \lambda) = \|x\|_2\, \|y\|_2$ for any unitarily invariant norm $\|\cdot\|$. One way to show this is to consider $E = yx^H/(\|x\|_2\, \|y\|_2)$, which—by inserting in (1)—can be seen to attain the supremum in (2).

Much of the recent research on eigenvalue condition numbers has been devoted to the case when the perturbation E is known to be in a set $\mathbb{S} \subset \mathbb{C}^{n\times n}$ of structured matrices. In this case, it is more appropriate to restrict the supremum in (2) to $E \in \mathbb{S}$, giving rise to the structured eigenvalue condition number $\kappa(A, \lambda; \mathbb{S})$. In [9, 40], computable expressions of $\kappa(A, \lambda; \mathbb{S})$ for general linear structures $\mathbb{S}$ have been developed.

[*] Received by the editors October 20, 2006; accepted for publication (in revised form) by J. L. Barlow November 24, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/simax/31-1/67289.html

[†] Seminar für angewandte Mathematik, ETH Zürich, 8092 Zürich, Switzerland (kressner@math.ethz.ch). The research of this author was supported by the DFG Emmy Noether fellowship KR 2950/1-1 and the Swedish Foundation for Strategic Research under grant A3 02:128.

[‡] Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad, 30, 28911-Leganés, Spain (mpelaezmath@gmail.com, jmoro@math.uc3m.es). The research of the second and third authors was partially supported by the Spanish Ministerio de Ciencia y Tecnología under grants BFM2003-00223 and MTM2006-05361.

This has been extended to smooth nonlinear structures in [16]. A simplified expression for zero-structured matrices can be found in [28]. Trivially, $\kappa(A,\lambda;\mathbb{S}) \le \kappa(A,\lambda)$. It naturally gives rise to the question of whether $\kappa(A,\lambda;\mathbb{S})$ can be much smaller than $\kappa(A,\lambda)$, or, in other words, whether λ can be much less sensitive to structured perturbations than to unstructured ones. For surprisingly many structures $\mathbb{S}$ the answer to this question is negative in the sense that $\kappa(A,\lambda;\mathbb{S})$ is always at most within a small factor of $\kappa(A,\lambda)$. This has been shown for $\mathbb{S} = \mathbb{R}^{n\times n}$ in [3], as well as for real skew-symmetric, skew-Hermitian, Hankel, Toeplitz, Hamiltonian, persymmetric, circulant, orthogonal, unitary, and related structures in [16, 32]. Practically relevant examples for which $\kappa(A,\lambda;\mathbb{S}) \ll \kappa(A,\lambda)$ is possible include complex skew-symmetric [32], zero-structured [28], and symplectic matrices [16].

If λ is a multiple eigenvalue of algebraic multiplicity m, there is generally not an expansion of the form (1). Instead, λ bifurcates into m perturbed eigenvalues $\hat\lambda_k(\epsilon)$, each admitting a fractional expansion

$$\hat\lambda_k = \lambda + \alpha_k^{\gamma_k}\epsilon^{\gamma_k} + o(\epsilon), \quad \epsilon \to 0, \quad k = 1,\dots,m, \tag{3}$$

with $\alpha_k > 0$ and $0 < \gamma_k \le 1$ [20, 42, 27]. Under generically satisfied conditions on E, Lidskii's theory [20] states that each Jordan block $J_{n_j}(\lambda)$ of size $n_j \times n_j$ gives rise to n_j perturbed eigenvalues satisfying the expansion (3) with $\gamma_k = 1/n_j$. Motivated by these results, the *Hölder condition number* for λ is defined in [27] as the pair

$$\kappa(A,\lambda) = (n_1, \alpha), \tag{4}$$

where $1/n_1$ is the smallest possible power γ_k of ϵ in (3) for any perturbation E. The scalar $\alpha^{1/n_1} > 0$ is the largest possible magnitude of the coefficient of ϵ^{1/n_1} for all E with $\|E\| \le 1$. While n_1 happens to be the size of the largest Jordan block belonging to λ, we have

$$\alpha^{1/n_1} = \lim_{\epsilon\to 0} \sup_{\substack{\|E\|\le 1\\ E\in\mathbb{C}^{n\times n}}} \max_{k=1,\dots,m} \frac{|\hat\lambda_k - \lambda|}{\epsilon^{1/n_1}} \tag{5}$$

(see also [5, p. 156] for a similar definition of condition number for multiple eigenvalues, and [4] for its relationship with $\kappa(A,\lambda)$). An explicit formula for α can be found in [27]; see also section 2. Let us emphasize that for certain nongeneric perturbations E, the value of γ_k can be larger than $1/n_1$ for all $\hat\lambda_k$. This is demonstrated by the following example [43, 27]. The characteristic polynomial of

$$A + \epsilon E = \left[\begin{array}{ccc|cc} 0 & 1 & 0 & & \\ & 0 & 1 & & \\ & & 0 & \epsilon & \\ \hline & & & 0 & 1 \\ \epsilon & & & & 0 \end{array}\right] \tag{6}$$

is $\epsilon^2 - \lambda^5$. Thus, $\gamma_k = 2/5$ for all $\hat\lambda_k$ in (3) while $1/n_1 = 1/3$.

The purpose of this paper is to investigate various extensions of the condition number (4). In particular, we are interested in the case when E is restricted to a set $\mathbb{S}$ of structured matrices, leading to the notion of a *structured Hölder condition number* $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$. We begin by noting that there exist structures $\mathbb{S}$ for which $n_{\mathbb{S}}$ can be smaller than n_1. Consider, for instance, the following example, taken from

[29]: $\mathbb{S}$ is the set of complex skew-symmetric matrices,

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ -1 & 0 & -i & 0 \\ 0 & i & 0 & i \\ -1 & 0 & -i & 0 \end{bmatrix} \in \mathbb{S}, \tag{7}$$

and $\lambda = 0$ with geometric multiplicity two and largest Jordan block of size 3, i.e., $n_1 = 3$. However, any complex skew-symmetric perturbation $E \in \mathbb{S}$ gives rise to $O(\epsilon^{1/2})$ perturbed eigenvalues so, according to Definition 3.1 below, $n_{\mathbb{S}} = 2 < n_1$.

However, for most structures under consideration we have $n_{\mathbb{S}} = n_1$. In this case, it makes sense to compare $\alpha_{\mathbb{S}}$ with its unstructured counterpart α. As will be shown in section 3, many of the results from [3, 16, 32] on structured condition numbers for simple eigenvalues carry over to multiple eigenvalues. A notable exception to $n_{\mathbb{S}} = n_1$ are complex skew-symmetric matrices, whose zero eigenvalues may exhibit $n_{\mathbb{S}} < n_1$; this is exemplified by (7). The relation $n_{\mathbb{S}} = n_1$ holds for nonzero eigenvalues of a complex skew-symmetric matrix, but $\alpha_{\mathbb{S}}$ can be significantly smaller than α, a fact that has already been observed in [32, 16]. In this paper, we not only provide additional insight by deriving explicit expressions for $\alpha_{\mathbb{S}}$, but we also cover the more general class of matrices that are skew-symmetric with respect to an orthosymmetric bilinear form.

Hölder condition numbers for the generalized eigenvalues of a regular matrix pencil $A - \lambda B$ can be defined similarly employing the perturbation expansions of Langer and Najman [17, 18, 19]; see also [6]. Structured Hölder condition numbers for eigenvalues of pencils can be defined analogously, and they have lately received some attention: results for simple eigenvalues of linearly structured pencils can be found in [9] and for multiple eigenvalues of definite Hermitian matrix pencils in [35, 36]. The problem of estimating the (multiple) eigenvalue sensitivities for parameter-dependent matrix pencils is closely related; see [1, 37, 45] and the references therein. To our knowledge, the results provided in this paper on structured Hölder condition numbers for real, symmetric/skew-symmetric, Hermitian, as well as palindromic matrix pencils are new, even for simple eigenvalues. Furthermore, this framework also allows us to cover matrix polynomial eigenvalue problems by imposing block companion structure.

The rest of this paper is organized as follows. In section 2 we recall definitions and provide some basic results on unstructured and structured Hölder condition numbers for multiple eigenvalues of matrices. Section 3 is devoted to structured Hölder condition numbers for real, Toeplitz, and Hankel matrices, as well as for matrix classes that form Jordan or Lie algebras associated with an orthosymmetric bilinear or sesquilinear form. Section 4 is concerned with Hölder condition numbers for multiple eigenvalues of generalized eigenvalue problems, first for (structured) matrix pencils and then for matrix polynomials via companion form. Finally, some conclusions and open issues not addressed in this paper can be found in section 5.

2. Preliminaries. In the following, we summarize the part of the discussion of Lidskii's results in [27] that leads to the condition number (4). Let λ be an eigenvalue of $A \in \mathbb{C}^{n\times n}$, and let n_1 be the size of the largest Jordan block corresponding to λ. The Jordan canonical form of A can be written as

$$\left[\begin{array}{c|c} J & 0 \\ \hline 0 & \widetilde{J} \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right] A \left[\begin{array}{c|c} P & \widetilde{P} \end{array}\right], \tag{8}$$

where

$$\left[\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right] \left[\begin{array}{c|c} P & \widetilde{P} \end{array}\right] = I \tag{9}$$

and J consists of all $n_1 \times n_1$ Jordan blocks corresponding to λ. Specifically, we have

$$J = \mathrm{diag}(\Gamma_1^1, \ldots, \Gamma_1^{r_1}), \qquad \Gamma_1^1 = \cdots = \Gamma_1^{r_1} = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \in \mathbb{C}^{n_1 \times n_1}. \tag{10}$$

The block $\widetilde{J}$ contains all Jordan blocks corresponding to λ with dimension smaller than n_1, as well as all Jordan blocks corresponding to eigenvalues different from λ.

The columns of P form r_1 linearly independent Jordan chains of length n_1, each of which starts with an eigenvector of A. Collecting these starting vectors in an $n \times r_1$ matrix X yields

$$X = \left[\ Pe_1,\ Pe_{n_1+1},\ \ldots,\ Pe_{(r_1-1)n_1+1}\ \right], \tag{11}$$

where e_i denotes the ith unit vector of length n. Similarly we collect in

$$Y = \left[\ Q^H e_{n_1},\ Q^H e_{2n_1},\ \ldots,\ Q^H e_{r_1 n_1}\ \right] \tag{12}$$

the left eigenvectors chosen from the r_1 independent Jordan chains of length n_1 appearing as rows of Q. Note that each column of Y represents a left eigenvector of A belonging to λ. Notice also that the relation (9) implies $Y^H X = I$ if $n_1 = 1$, and $Y^H X = 0$ otherwise. With these preparations we can state a highly abridged version of Lidskii's result.

THEOREM 2.1 (see [20, 27]). [illegible] $E \in \mathbb{C}^{n\times n}$ [illegible] $Y^H EX$ [illegible] X [illegible] Y [illegible] $n_1 r_1$ [illegible] $A + \epsilon E$ [illegible]

$$\hat{\lambda}_k = \lambda + (\xi_k)^{1/n_1} \epsilon^{1/n_1} + o(\epsilon^{1/n_1}), \quad k = 1, \ldots, r_1, \tag{13}$$

[illegible] $\xi_1, \ldots, \xi_{r_1}$ [illegible] $Y^H EX$

Since X and Y have linearly independent columns, the invertibility of $Y^H EX$ is generically satisfied for a general perturbation E in $\mathbb{C}^{n\times n}$. Within a set $\mathbb{S}$ of structured perturbations E, however, it may happen that $Y^H EX$ is not generically invertible. Fortunately, the result of Theorem 2.1 remains valid even if $Y^H EX$ is singular. This follows from a very general theory by Moro, Burke, and Overton [27] on the connection between Newton diagrams and eigenvalue perturbation expansions.

[illegible] 2.2 (see [27, p. 809]). Let $E \in \mathbb{C}^{n\times n}$ such that $Y^H EX$ is singular. Then each of the $\beta < r_1$ nonzero eigenvalues $\xi_1, \ldots, \xi_\beta$ of $Y^H EX$ gives rise to n_1 perturbation expansions of the form (13). The remaining $r_1 - \beta$ zero eigenvalues correspond to expansions where the exponent of the leading nonzero perturbation term is strictly larger than $1/n_1$.

Theorem 2.1 implies that, for sufficiently small ϵ, the worst-case change in λ is caused by an eigenvalue of $Y^H EX$ that is as large as possible in magnitude. This motivates the following definition.

Definition 2.3 (see [27]). [illegible] λ [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] A [illegible] (8) [illegible] Hölder condition number [illegible] λ [illegible] $\kappa(A,\lambda) = (n_1,\alpha)$, [illegible] n_1 [illegible] λ [illegible] (8) [illegible]

$$\alpha = \sup_{\substack{\|E\|\le 1\\ E\in\mathbb{C}^{n\times n}}} \rho(Y^H E X), \tag{14}$$

[illegible] $\rho(\cdot)$ [illegible]

We have $\rho(Y^HEX) = \rho(EXY^H) \le \|EXY^H\|_2 \le \|XY^H\|_2$ for the matrix 2-norm $\|\cdot\|_2$. To show equality, we have to construct a perturbation E so that $\rho(Y^HEX) = \|XY^H\|_2$ is attained. The following basic lemma helps identify such perturbations.

Lemma 2.4. [illegible]

$$XY^H = U\Sigma V^H$$

[illegible] $U \in \mathbb{C}^{n\times r_1}$ [illegible] $V \in \mathbb{C}^{n\times r_1}$ [illegible] $\Sigma = \operatorname{diag}(\sigma_1,\dots,\sigma_{r_1})$, [illegible] $\sigma_1 \ge \cdots \sigma_{r_1} \ge 0$ [illegible] $E = VDU^H$, [illegible] $D = \operatorname{diag}(1,\delta_2,\dots,\delta_{r_1})$ [illegible] $\delta_j \le 1$ [illegible] $\rho(Y^HEX) = \|XY^H\|_2$

[illegible] The result follows from

$$\rho(Y^HEX) = \rho(EXY^H) = \rho(VD\Sigma V^H) = \rho(D\Sigma) = \|D\Sigma\|_2 = \|\Sigma\|_2 = \|XY^H\|_2. \qquad \square$$

Note that the definition of α in (14) depends on the norm $\|\cdot\|$ used in the constraint $\|E\| \le 1$. For unitarily invariant norms, we have the following result.

Theorem 2.5 (see [27]). [illegible] (n_1,α) [illegible] λ [illegible] $\alpha = \|XY^H\|_2$ [illegible] $\|\cdot\|$ [illegible] (14).

[illegible] Setting $D = \operatorname{diag}(1,0,\dots,0)$ and $E = VDU^H$ in Lemma 2.4 gives $\|E\| = 1$ and thus proves $\alpha = \|XY^H\|_2$. $\square$

It is important to note that for a specific norm, choices of D other than the one used in the above proof are possible; e.g., for $\|\cdot\| \equiv \|\cdot\|_2$, any E in the sense of Lemma 2.4 gives $\|E\|_2 = 1$. In particular, setting $D = \Sigma/\sigma_1$ yields

$$E = \frac{YX^H}{\|XY^H\|_2}, \tag{15}$$

which resembles the classical perturbation matrix for simple eigenvalues [43]. This type of perturbation will often be used when proving that the structured and the unstructured condition numbers coincide for the 2-norm. Another class of perturbations which turns out to be very useful is given by the following lemma.

Lemma 2.6. [illegible] $u_1, v_1 \in \mathbb{C}^{n\times n}$, [illegible] $\|u_1\|_2 = \|v_1\|_2 = 1$ [illegible] XY^H [illegible] $E \in \mathbb{C}^{n\times n}$ [illegible] $Eu_1 = \beta v_1$, [illegible] $|\beta| = 1$ [illegible] $\rho(EXY^H) \ge \|XY^H\|_2$

[illegible] Let $XY^H = U\Sigma V^H$ be a singular value decomposition with $U = [u_1,\dots,u_n]$, $V = [v_1,\dots,v_n]$. Then

$$\begin{aligned}\rho(EXY^H) &= \rho(EU\Sigma V^H) = \rho(V^HEU\Sigma) = \rho(V^H[\beta v_1, Ev_2,\dots,Ev_n]\Sigma)\\ &= \rho\left(\begin{bmatrix} \beta\|XY^H\|_2 & \star \\ 0 & \star \end{bmatrix}\right) \ge \|XY^H\|_2. \qquad \square\end{aligned}$$

3. Structured Hölder condition numbers. Throughout the whole section, λ denotes an eigenvalue of A with Hölder condition number $\kappa(A,\lambda) = (n_1, \alpha)$. The matrices X and Y are defined as in (11) and (12), respectively.

Restricting the range of admissible perturbations E from $\mathbb{C}^{n\times n}$ to a subset $\mathbb{S} \subset \mathbb{C}^{n\times n}$ leads to a corresponding structured condition number $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$.

DEFINITION 3.1. [illegible] λ [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\mathbb{S}$ [illegible] $\mathbb{C}^{n\times n}$ [illegible] structured Hölder condition number [illegible] λ [illegible] $\kappa(A,\lambda;\mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$, [illegible] $1/n_{\mathbb{S}}$ [illegible] γ_k [illegible] ϵ [illegible] (3) [illegible] $E \in \mathbb{S}$ [illegible] $\alpha_{\mathbb{S}} > 0$ [illegible] α_k [illegible] (3) [illegible] $E \in \mathbb{S}$ [illegible] $\|E\| \leq 1$

As shown by example (7), it may happen that $n_{\mathbb{S}} < n_1$, but in this paper we focus on the cases when $n_{\mathbb{S}} = n_1$. If so, then by Theorem 2.1 and Remark 2.2 we can write

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|E\|\leq 1\\ E\in\mathbb{S}}} \rho(Y^H E X). \tag{16}$$

Note that the right-hand side in this expression becomes zero if and only if $n_{\mathbb{S}} < n_1$.

It turns out that the presence of the spectral radius in (16) considerably complicates the task of finding explicit formulas or reasonable bounds for $\alpha_{\mathbb{S}}$. However, we will see that it is often possible to identify structures with $\alpha_{\mathbb{S}} \approx \alpha$ by constructing a perturbation $E \in \mathbb{S}$ for which $\rho(Y^H E X)$ is close to α.

3.1. Real matrices. As a first example, we point out that restricting the perturbation to be real can, at best, mildly improve the sensitivity of λ. This has been shown for a simple eigenvalue λ in [3]. The following lemma is a generalization to multiple eigenvalues.

LEMMA 3.2. [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] $\kappa(A,\lambda;\mathbb{R}^{n\times n}) = (n_1, \alpha_{\mathbb{R}})$ [illegible]

(i) $\alpha/2 \leq \alpha_{\mathbb{R}} \leq \alpha$ [illegible] $\|\cdot\|$.

(ii) [illegible] $\alpha_{\mathbb{R}} = \alpha$ [illegible] 2 [illegible] $\|\cdot\| \equiv \|\cdot\|_2$ [illegible] A [illegible]

[illegible] Decomposing $XY^H = M_R + \imath M_I$ with $M_R, M_I \in \mathbb{R}^{n\times n}$ gives $\|M_R\|_2 \geq \|XY^H\|_2/2$ or $\|M_I\|_2 \geq \|XY^H\|_2/2$. Without loss of generality, we may assume $\|M_R\|_2 \geq \|XY^H\|_2/2$. Let us consider the perturbation $E = v_1 u_1^T \in \mathbb{R}^{n\times n}$, where u_1 and v_1 are normalized left and right singular vectors belonging to the largest singular value of M_R. Then $\|E\| = 1$ and

$$\begin{aligned}\rho(EXY^H) = \rho(v_1 u_1^T(M_R + \imath M_I)) &= |u_1^T(M_R + \imath M_I)v_1| \\ &\geq |u_1^T M_R v_1| = \|M_R\|_2 \geq \|XY^H\|_2/2,\end{aligned}$$

which proves $\alpha_{\mathbb{R}} \geq \alpha/2$. To show the second part, note that we can choose $Y = X$ if A is normal and thus $E = I \in \mathbb{R}^{n\times n}$ gives $\rho(EXX^H) = \alpha = \alpha_{\mathbb{R}}$. □

[illegible] 3.3. In the case $r_1 = 1$ (one single Jordan block of largest size n_1), we can use the same arguments as in [3] to improve the lower bound of Lemma 3.2(i) to $\alpha/\sqrt{2} \leq \alpha_{\mathbb{R}}$. It is not clear to us whether this slightly stronger result holds for $r_1 > 1$.

Suppose that $\mathbb{S}$ is a structure such that for any $E \in \mathbb{S}$ the real and imaginary parts of E are both in $\mathbb{S} \cap \mathbb{R}^{n\times n}$. For a simple eigenvalue λ, Rump [32] has extended the bounds of Lemma 3.2 to structured condition numbers in the sense that restricting the perturbations from $\mathbb{S}$ to $\mathbb{S} \cap \mathbb{R}^{n\times n}$ improves the condition number by at most a factor of $1/\sqrt{2}$. By a trivial extension of [32, Lemma 3.1], this result holds for the case $r_1 = 1$, but it is difficult to show that a similarly general result holds for an eigenvalue having multiple Jordan blocks of largest size. The following lemma is only a first step in this direction.

Lemma 3.4. [illegible] $\mathbb{SR}$ [illegible] $\mathbb{R}^{n\times n}$ [illegible] $\mathbb{S} = \mathbb{SR} + \imath\mathbb{SR}$ [illegible] $\mathbb{SR}$ [illegible] rank one [illegible] $E \in \mathbb{S}$ [illegible] $\|E\| = 1$ [illegible] $\alpha_{\mathbb{S}} = \rho(Y^H E X)$ [illegible]

$$\alpha_{\mathbb{S}}/4 \le \alpha_{\mathbb{SR}} \le \alpha_{\mathbb{S}}$$

[illegible] 2 [illegible] $\|\cdot\| \in \{\|\cdot\|_F, \|\cdot\|_2\}$

[illegible]. We can write $E = vu^H$ for $u, v \in \mathbb{C}^{n\times n}$ with $\|u\|_2 = \|v\|_2 = 1$. Decomposing $u = u_R + \imath u_I$ and $v = v_R + \imath v_I$ with $u_R, u_I, v_R, v_I \in \mathbb{R}^n$ gives

$$\alpha_{\mathbb{S}} = |u^H XY^H v| = |(u_R^T XY^H v_R + u_I^T XY^H v_I) - \imath(u_I^T XY^H v_R - u_R^T XY^H v_I)|.$$

At least one of the two bracketed terms in this sum is not smaller than $\alpha_{\mathbb{S}}/2$ in magnitude. Suppose $|u_R^T XY^H v_R + u_I^T XY^H v_I| \ge \alpha_{\mathbb{S}}/2$ and set $U = [u_R, u_I]$, $V = [v_R, v_I]$. Then $|\mathrm{trace}(U^T XY^H V)| \ge \alpha_{\mathbb{S}}/2$, implying $\rho(U^T XY^H V) \ge \alpha_{\mathbb{S}}/4$. Thus, the real perturbation $E_R = VU^T$ (which is the real part of E) yields $\alpha_{\mathbb{SR}} \ge \rho(E_R XY^H) \ge \alpha_{\mathbb{S}}/4$ while $\|E_R\|_2 \le 1$ and $\|E_R\|_F \le 1$, which completes the proof. The case $|u_I^T XY^H v_R - u_R^T XY^H v_I| \ge \alpha_{\mathbb{S}}/2$ is treated analogously. □

3.2. General linear structures. Let us briefly investigate the rather general case that $\mathbb{S}$ is a linear matrix space in $\mathbb{F}^{n\times n}$ with $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. Using an approach developed by Higham and Higham [9], we consider a fixed basis $\{M_1, \dots, M_l\}$ of $\mathbb{S}$ that is orthonormal with respect to the matrix inner product. Then for each perturbation $E \in \mathbb{S}$ there is a unique vector $p = [p_1, \dots, p_l]^T \in \mathbb{F}^l$ so that $E = p_1 M_1 + \cdots + p_l M_l$ and $\|E\|_F = \|p\|_2$. If $n_{\mathbb{S}} = n_1$, the structured condition number $\kappa(A, \lambda; \mathbb{S}) = (n_1, \alpha_{\mathbb{S}})$ satisfies

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|p\|_2 \le 1 \\ p \in \mathbb{F}^l}} \rho\left(p_1 Y^H M_1 X + \cdots + p_l Y^H M_l X\right) \tag{17}$$

for the Frobenius norm $\|\cdot\| \equiv \|\cdot\|_F$. Maximizing a nonsymmetric spectral function is known to be a nontrivial optimization problem; see, e.g., [2]. We therefore see little hope in finding an explicit expression for $\alpha_{\mathbb{S}}$ in general. There are two special cases for which $\alpha_{\mathbb{S}}$ can be (nearly) determined.

1. The case $r_1 = 1$ (X and Y are vectors) can be treated the same way as the case of simple eigenvalues [9, 40]. Defining the [illegible]

   $$\mathcal{M} = [\mathrm{vec}(M_1), \dots, \mathrm{vec}(M_l)], \tag{18}$$

   where vec stacks the columns of a matrix into a long vector, we can write $\mathrm{vec}(E) = \mathcal{M}p$, and therefore

   $$\alpha_{\mathbb{S}} = \sup_{\substack{\|p\|_2 \le 1 \\ p \in \mathbb{F}^l}} |p_1 Y^H M_1 X + \cdots + p_l Y^H M_l X| = \|(X^T \otimes Y^H)\mathcal{M}\|_2$$

   when $\mathbb{F} = \mathbb{C}$, or when $\mathbb{F} = \mathbb{R}$ and $X, Y \in \mathbb{R}^n$. For $\mathbb{F} = \mathbb{R}$ and $X, Y \notin \mathbb{R}^n$, we can show as in [16, section 2] that $\|(X^T \otimes Y^H)\mathcal{M}\|_2/\sqrt{2} \le \alpha_{\mathbb{S}} \le \|(X^T \otimes Y^H)\mathcal{M}\|_2$.

2. If $\mathbb{F} = \mathbb{C}$ and all matrices $N_j = Y^H M_j X$ are Hermitian, then

$$\begin{aligned} \alpha_{\mathbb{S}} &= \sup_{\substack{\|p\|_2 \le 1 \\ p \in \mathbb{C}^l}} \|p_1 N_1 + \cdots + p_l N_l\|_2 \\ &= \sup_{\substack{\|x\|_2 = 1 \\ x \in \mathbb{C}^n}} \|[x^H N_1 x, \ldots, x^H N_l x]\|_2. \end{aligned}$$

It follows that

$$\max_i \|N_i\|_2 \le \alpha_{\mathbb{S}} \le \sqrt{l} \max_i \|N_i\|_2.$$

3.3. Toeplitz and Hankel matrices. In [32], it is proven that the structured pseudospectrum of a matrix $A \in \mathbb{S}$ coincides with the unstructured pseudospectrum for the following [illegible] structures $\mathbb{S}$: symmetric, persymmetric, Toeplitz, symmetric Toeplitz, Hankel, persymmetric Hankel, and circulant. This implies, in particular, that $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda)$ for all these structures. Hence, some of the results that follow could be stated without proof. However, the proofs provided here explicitly construct structured perturbations that attain $\kappa(A, \lambda)$, which might lead to additional insight.

A [illegible] takes the form

$$T = \begin{bmatrix} t_0 & t_{-1} & \ldots & t_{-n+1} \\ t_1 & t_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \ldots & t_1 & t_0 \end{bmatrix} \in \mathbb{C}^{n \times n}$$

and $H \in \mathbb{C}^{n \times n}$ is a [illegible] if $F_n H$ is Toeplitz, where F_n is the $n \times n$ [illegible] with ones on the antidiagonal and zeros everywhere else.

THEOREM 3.5. [illegible] $\mathbb{T}$ [illegible] $\mathbb{SY}$ [illegible] 2 [illegible]

(i) $\kappa(A, \lambda; \mathbb{T}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ [illegible] $A \in \mathbb{T}$.

(ii) $\kappa(A, \lambda; \mathbb{T} \cap \mathbb{SY}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ [illegible] $A \in \mathbb{T} \cap \mathbb{SY}$.

(iii) $\kappa(A, \lambda; \mathbb{T} \cap \mathbb{R}^{n \times n}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ [illegible] $A \in \mathbb{T} \cap \mathbb{R}^{n \times n}$ [illegible] $\lambda \in \mathbb{R}$.

(iv) $\kappa(A, \lambda; \mathbb{T} \cap \mathbb{SY} \cap \mathbb{R}^{n \times n}) = \kappa(A, \lambda) = (1, 1)$ [illegible] $A \in \mathbb{T} \cap \mathbb{SY} \cap \mathbb{R}^{n \times n}$

[illegible] A Toeplitz matrix is complex [illegible], meaning that $F_n T$ is complex symmetric. We can therefore apply Corollary A.2(i) to conclude $Y = F_n \overline{X}$. A Takagi factorization [13, section 4.4] of the complex symmetric matrix XX^T is a special singular value decomposition $XX^T = U \Sigma U^T$, where $U \in \mathbb{C}^{n \times r_1}$ has orthonormal columns and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{r_1})$ with $\sigma_1 \ge \cdots \ge \sigma_{r_1} > 0$. By [31, Lemma 10.1], there is a Hankel matrix H with $\|H\|_2 = 1$ and $Hu_1 = \bar{u}_1$, where u_1 denotes the first column of U. Setting $E = F_n H \in \mathbb{T}$ gives $\|E\|_2 = 1$ with $Eu_1 = F_n \bar{u}_1$, which completes the proof of (i) by Lemma 2.6.

A symmetric Toeplitz matrix is persymmetric and symmetric; it can thus be block diagonalized by a simple orthogonal transformation:

$$G^T A G = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix},$$

where $A_{11} \in \mathbb{R}^{\lfloor n/2 \rfloor \times \lfloor n/2 \rfloor}$, $A_{22} \in \mathbb{R}^{\lceil n/2 \rceil \times \lceil n/2 \rceil}$ are complex symmetric and

$$G = \frac{1}{\sqrt{2}} \begin{bmatrix} I & F_{n/2} \\ -F_{n/2} & I \end{bmatrix} \text{(even } n\text{)}, \quad G = \frac{1}{\sqrt{2}} \begin{bmatrix} I & 0 & F_{(n-1)/2} \\ 0 & \sqrt{2} & 0 \\ -F_{(n-1)/2} & 0 & I \end{bmatrix} \text{(odd } n\text{)}.$$

This folklore result, which can be found, for example, in [44], shows that $X = [X_1, X_2]$ with $X_1 = -F_n X_1$ and $X_2 = F_n X_2$. The eigenvectors contained in X_1 and X_2 stem from Jordan blocks in A_{11} and A_{22}, respectively. Moreover, $Y = F_n[\overline{X}_1, \overline{X}_2]$ and

$$\begin{aligned} \alpha_{\mathbb{T}\cap\mathbb{SY}} &= \sup_{\substack{\|E\|_2=1 \\ E\in\mathbb{T}\cap\mathbb{SY}}} \max\big(\rho(EX_1X_1^TF_n), \rho(EX_2X_2^TF_n)\big) \\ &= \sup_{\substack{\|E\|_2=1 \\ E\in\mathbb{T}\cap\mathbb{SY}}} \max\big(\rho(EX_1X_1^T), \rho(EX_2X_2^T)\big)\,. \end{aligned}$$

From $X_2^H X_1 = X_2^H F_n F_n X_1 = -X_2^H X_1$ it follows that $X_2^H X_1 = 0$ and hence

$$\|XX^T\|_2 = \|[X_1, X_2][X_1, X_2]^T\|_2 = \max(\|X_1X_1^T\|_2, \|X_2X_2^T\|_2).$$

Let us assume $\|X_1X_1^T\|_2 \geq \|X_2X_2^T\|_2$ (the other case is treated analogously), and let $X_1X_1^T = U\Sigma U^T$ be a Takagi factorization. Then $U = -F_nU$ and by [32, Lemma 2.4] there is a symmetric Toeplitz matrix E such that $\|E\|_2 = 1$ and $Eu_1 = \overline{u}_1$. The proof of (ii) is completed by applying Lemma 2.6.

Parts (iii) and (iv) are shown by noting that $\lambda \in \mathbb{R}$ implies $X \in \mathbb{R}^{n\times r_1}$ and hence the perturbations constructed above can be chosen to be real [32]. □

Theorem 3.5 can be easily extended to Hankel matrices.

Corollary 3.6. [illegible] $\mathbb{HA}$ [illegible] $\mathbb{PS}$ [illegible] 2 [illegible]

(i) $\kappa(A, \lambda; \mathbb{HA}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ [illegible] $A \in \mathbb{HA}$.

(ii) $\kappa(A, \lambda; \mathbb{HA} \cap \mathbb{PS}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$ [illegible] $A \in \mathbb{HA} \cap \mathbb{PS}$.

(iii) $\kappa(A, \lambda; \mathbb{HA} \cap \mathbb{R}^{n\times n}) = \kappa(A, \lambda) = (1, 1)$ [illegible] $A \in \mathbb{HA} \cap \mathbb{R}^{n\times n}$.

(iv) $\kappa(A, \lambda; \mathbb{HA} \cap \mathbb{PS} \cap \mathbb{R}^{n\times n}) = \kappa(A, \lambda) = (1, 1)$ [illegible] $A \in \mathbb{HA} \cap \mathbb{PS} \cap \mathbb{R}^{n\times n}$

[illegible]. A Hankel matrix is complex symmetric, which implies $Y = \overline{X}$ by Corollary A.2(i). The rest of the proof is along the lines of the proof of Theorem 3.5 and is therefore omitted. □

3.4. Symmetric, skew-symmetric, and Hermitian matrices.

The construction used in the proof of Theorem 3.5 exploits only the persymmetry of Toeplitz matrices, and it can thus also be used to show $\kappa(A, \lambda; \mathbb{PS}) = \kappa(A, \lambda)$ for $A \in \mathbb{PS}$. More generally, we have the following result, which besides persymmetric ($M = F_n$) also includes symmetric ($M = I$) and pseudosymmetric ($M = \mathrm{diag}(I, -I)$) matrices.

Theorem 3.7. [illegible] $M \in \mathbb{R}^{n\times n}$ [illegible] $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^TM = MA\}$. [illegible] $A \in \mathbb{S}$ [illegible]

(i) $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda) = (n_1, \|XX^T\|_2)$.

(ii) $\kappa(A, \lambda; \mathbb{S} \cap \mathbb{R}^{n\times n}) = (n_1, \alpha_{\mathbb{S}\cap\mathbb{R}^{n\times n}})$ [illegible] $\|XX^T\|_2/2 \leq \alpha_{\mathbb{S}\cap\mathbb{R}^{n\times n}} \leq \|XX^T\|_2$

[illegible]. Corollary A.2(i) gives $Y = M\overline{X}$. Let $XX^T = U\Sigma U^T$ be a Takagi factorization; then we set $u_1 = Ue_1$ and $E = M\overline{u}_1u_1^H$ to obtain $\|E\| = 1$ and

$$\alpha_{\mathbb{S}} \geq \rho(EXX^TM) = \rho(M\overline{u}_1u_1^HXX^TM) = \rho(u_1^HXX^T\overline{u}_1) = \|XX^T\|_2 = \alpha.$$

This completes the proof of the first part. The proof of the second part is virtually identical with the proof of Lemma 3.2(i). □

Using the terminology of [16], Theorem 3.7 is concerned with Jordan algebras associated with the symmetric bilinear form $\langle x, y\rangle = x^TMy$. For the corresponding Lie algebras, which are given by $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : A^TM = -MA\}$, it is known that the structured and unstructured condition numbers for simple eigenvalues may widely differ [16, 32]. We have already shown in the introduction a skew-symmetric matrix (7) (corresponding to $M = I$) such that $n_{\mathbb{S}} < n_1$. This shows that multiple eigenvalues can have a [illegible] better behavior under structured perturbations. The following theorem identifies one such situation and proves that, in this setting, it can happen only under very specific conditions (namely, for zero eigenvalues with one single Jordan block of largest odd size). Additionally, Theorem 3.8 provides some insight on the expected difference between $\alpha_{\mathbb{S}}$ and α whenever $n_{\mathbb{S}} = n_1$.

Theorem 3.8. [illegible] $M \in \mathbb{R}^{n\times n}$ [illegible] $\mathbb{S} = \{A \in \mathbb{C}^{n\times n} : \ A^TM = -MA\}$ [illegible] $\kappa(A, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ [illegible] $A \in \mathbb{S}$ [illegible] 2 [illegible]

(i) [illegible] $\lambda = 0$ n_1 [illegible] $r_1 = 1$ [illegible] $n_{\mathbb{S}} < n_1$.

(ii) [illegible] $\lambda = 0$ n_1 [illegible] $r_1 > 1$ [illegible] $n_{\mathbb{S}} = n_1$ [illegible] $\alpha_{\mathbb{S}} = \sqrt{\sigma_1\sigma_2}$ [illegible] σ_1, σ_2 [illegible] XX^T [illegible] $\alpha = \sigma_1$.

(iii) [illegible] $\lambda = 0$ [illegible] n_1 [illegible] r_1 [illegible] $n_{\mathbb{S}} = n_1$ [illegible]

$$\alpha_{\mathbb{S}} = \alpha = \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2 ;$$

(iv) [illegible] $\lambda \neq 0$ [illegible] $r_1 = 1$ [illegible] $n_{\mathbb{S}} = n_1$ [illegible] $\alpha_{\mathbb{S}} = \sqrt{\|X\|_2^2\|Y\|_2^2 - |Y^TMX|^2}$.

(v) [illegible] $\lambda \neq 0$ [illegible] $r_1 > 1$ [illegible] $n_{\mathbb{S}} = n_1$

[illegible]. For $\lambda = 0$ with n_1 odd, Corollary A.4(i)(a) implies $Y = M\overline{X}$. Now, if $n_{\mathbb{S}}$ was equal to n_1, then there would exist some $E \in \mathbb{S}$ with $\rho(Y^HEX) = \rho(X^TMEX) > 0$. This is impossible for $r_1 = 1$, since X is a vector and ME is skew-symmetric, so $\rho(X^TMEX) = |X^TMEX| = 0$. This shows (i).

To show $n_{\mathbb{S}} = n_1$ for $\lambda = 0$ when n_1 is odd and $r_1 > 1$, it is sufficient to construct a perturbation $E \in \mathbb{S}$ such that $\rho(X^TMEX) > 0$. For this purpose, consider a Takagi factorization $XX^T = U\Sigma U^T$, where $U = [u_1, \ldots, u_{r_1}]$ has orthonormal columns and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{r_1})$ with $\sigma_1 \geq \cdots \geq \sigma_{r_1} > 0$. Setting $E = M[\overline{u}_1, \overline{u}_2] \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right] [u_1, u_2]^H$ gives $E \in \mathbb{S}$, $\|E\|_2 = 1$, and

$$\alpha_{\mathbb{S}} \geq \rho(X^TMEX) = \rho\left(\begin{bmatrix} 0 & \sigma_2 \\ -\sigma_1 & 0 \end{bmatrix}\right) = \sqrt{\sigma_1\sigma_2} > 0.$$

On the other hand, letting $\mathbb{SK}$ denote the set of complex skew-symmetric matrices, we have

$$\begin{aligned} \alpha_{\mathbb{S}} = \sup_{\substack{\|\widetilde{E}\|_2\leq 1 \\ \widetilde{E}\in\mathbb{SK}}} \rho(\widetilde{E}XX^T) &= \sup_{\substack{\|G\|_2\leq 1 \\ G\in\mathbb{SK}}} \rho(G\Sigma) = \sup_{\substack{\|G\|_2\leq 1 \\ G\in\mathbb{SK}}} \rho(\Sigma^{1/2}G\Sigma^{1/2}) \\ &\leq \sup_{\substack{\|G\|_2\leq 1 \\ G\in\mathbb{SK}}} \|\Sigma^{1/2}G\Sigma^{1/2}\|_2 = \sup_{\substack{\|G\|_2\leq 1 \\ G\in\mathbb{SK}}} \|\tilde{\Sigma} \circ G\|_2, \end{aligned}$$

where $\tilde{\Sigma} = [\sqrt{\sigma_i\sigma_j}]_{i,j=1}^{r_1}$ and $\circ$ denotes the Hadamard product. A result by Mathias [24, Corollary 2.6] implies $\|\tilde{\Sigma} \circ G\|_2 \leq \sqrt{\sigma_1\sigma_2}\|G\|_2$, which concludes the proof of (ii).

For $\lambda = 0$ with n_1 odd, Corollary A.4(i)(b) implies $Y = M\overline{X}\left[\begin{smallmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{smallmatrix}\right]$. To attain $\rho(Y^HEX) = \alpha = |X\left[\begin{smallmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{smallmatrix}\right]X^T|_2$ we may just use a perturbation of the form (15), which in this case turns out to be $E = \frac{1}{\alpha}M\overline{X}\left[\begin{smallmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{smallmatrix}\right]X^H \in \mathbb{S}$. This proves (iii).

If $\lambda \neq 0$, then $-\lambda$ is also an eigenvalue with the same Jordan structure as λ. If we let $\widetilde{X}, \widetilde{Y}$ denote the matrices of right/left eigenvectors stemming from the $n_1 \times n_1$ Jordan blocks belonging to $-\lambda$, then Corollary A.4(i)(c) yields $\widetilde{X} = -M\overline{Y}$ and $\widetilde{Y} = M\overline{X}$. This implies not only $\kappa(A, -\lambda) = \kappa(A, \lambda)$ as well as $\kappa(A, -\lambda; \mathbb{S}) = \kappa(A, \lambda; \mathbb{S})$ but also that $[X, M\overline{Y}]$ has full column rank. If $r_1 = 1$, then X, Y are vectors and we have

$$|Y^H E X| = \rho\left([M\overline{Y}, X]^T M E [M\overline{Y}, X]\right) = \rho\left(M E [M\overline{Y}, X][M\overline{Y}, X]^T\right)$$

for any $E \in \mathbb{S}$. Hence, using the arguments from the proof of (ii), we have $\alpha_{\mathbb{S}} = \sqrt{\sigma_1 \sigma_2}$, where σ_1 and σ_2 are the two largest singular values of the symmetric matrix $[M\overline{Y}, X][M\overline{Y}, X]^T$. This shows (iv) since

$$\begin{aligned}
\sigma_1 \sigma_2 &= \sqrt{\det\left([M\overline{Y}, X][M\overline{Y}, X]^T [MY, \overline{X}][MY, \overline{X}]^T\right)} \\
&= \sqrt{\det\left([MY, \overline{X}]^T [M\overline{Y}, X][M\overline{Y}, X]^T [MY, \overline{X}]\right)} \\
&= \left|\det\left([MY, \overline{X}]^T [M\overline{Y}, X]\right)\right| \\
&= \|X\|_2^2 \|Y\|_2^2 - |Y^T M X|^2.
\end{aligned}$$

Unfortunately, the technique of this proof does not extend to the case $r_1 > 1$. Still, we can show $\alpha_{\mathbb{S}} > 0$, but it is not clear how to obtain a good lower or upper bound on $\alpha_{\mathbb{S}}$. The full column rank of $[X, M\overline{Y}]$ implies the existence of an invertible matrix L such that

$$L^{-1}[X, M\overline{Y}] = \begin{bmatrix} I_{r_1} & \star \\ 0 & I_{r_1} \\ 0 & 0 \end{bmatrix}.$$

Setting

$$E = M L^{-T} \begin{bmatrix} 0 & I_{r_1} & 0 \\ -I_{r_1} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} L^{-1} \in \mathbb{S}$$

yields $\rho(Y^H E X) = \rho(I_r) = 1$ and thus $\alpha_{\mathbb{S}} > 0$, completing the proof of (v). □

3.9. Note that Theorem 3.8(iv) also improves the results in [16, Theorem 4.3] and [32, Theorem 3.2], which state only bounds but no explicit formula for the structured condition number of a simple nonzero eigenvalue. Recently, Karow [15] described the limit sets of the structured pseudospectra for complex skew-symmetric matrices, from which Theorem 3.8(iv) could also be derived.

Fortunately, the matter of structured condition numbers is much less complicated for Jordan and Lie algebras associated with a sesquilinear form $\langle x, y \rangle = x^H M y$.

LEMMA 3.10. $M \in \mathbb{R}^{n \times n}$ $\mathbb{S} = \{A \in \mathbb{C}^{n \times n} : A^H M = \gamma M A\}$ $\gamma \in \{1, -1\}$ $A \in \mathbb{C}^{n \times n}$ $\kappa(A, \lambda; \mathbb{S}) = \kappa(A, \lambda)$ 2

Let $XY^H = U \Sigma V^H$ be a singular value decomposition and set $u_1 = U e_1$, $v_1 = V e_1$. Then $\|u_1\|_2 = \|v_1\|_2 = 1$ and by [22, Theorem 8.6] we can find a Hermitian matrix H such that $\|H\|_2 = 1$ and $H u_1 = \mu M v_1$ for some $\mu \in \mathbb{C}$ with $|\mu| = 1$. Set $E = \sqrt{\gamma} M H$ if $M = M^T$, and $E = \sqrt{-\gamma} M H$ if $M = -M^T$. Then $E \in \mathbb{S}$ satisfies $\|E\|_2 = 1$ and $E u_1 = \beta v_1$ for some $|\beta| = 1$, which implies the result by Lemma 2.6. □

3.5. *J*-symmetric and *J*-skew-symmetric matrices. For $M = J_{2n} = \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$, the structure $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} :\ A^H M = \gamma M A\}$ considered in Lemma 3.10 coincides with the set of skew-Hamiltonian matrices if $\gamma = 1$, and with the set of Hamiltonian matrices if $\gamma = -1$. The following two theorems provide results for the closely related structures $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} :\ A^T J_{2n} = \gamma J_{2n} A\}$, including bounds on structured condition numbers for skew-Hamiltonian and Hamiltonian matrices.

Theorem 3.11. $M \in \mathbb{R}^{2n\times 2n}$ $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} :\ A^T M = MA\}$. $A \in \mathbb{S}$ 2

(i) $\kappa(A,\lambda;\mathbb{S}) = \kappa(A,\lambda) = (n_1, \|XJ_{r_1}X^T\|_2)$.

(ii) $\kappa(A,\lambda;\mathbb{S}\cap\mathbb{R}^{2n\times 2n}) = (n_1, \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}})$ $\|XJ_{r_1}X^T\|_2/4 \le \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} \le \|XJ_{r_1}X^T\|_2$

Corollary A.6 reveals the relation $Y = -M\overline{X}J_{r_1}$. Using a perturbation as in (15), namely $E = (M^T\overline{X}J_{r_1}X^H)/\|XJ_{r_1}X^T\|_2$, yields $E \in \mathbb{S}$ with $\|E\|_2 = 1$ and

$$\alpha_{\mathbb{S}} \ge \rho(J_{r_1}^T X^T M E X)/\|XJ_{r_1}X^T\|_2 = \|XJ_{r_1}X^T\|_2 = \alpha.$$

To prove the second part, let $u = u_R + \imath u_I$ and $v = v_R + \imath v_I$, with $u_R, u_I, v_R, v_I \in \mathbb{R}^n$ and $\|u\|_2 = \|v\|_2 = 1$, be left/right singular vectors corresponding to the largest singular value of $K = XJ_{r_1}X^T$. Then

$$\alpha_{\mathbb{S}} = \|K\|_2 = u^H K v = (u_R^T K v_R + u_I^T K v_I) + \imath(u_R^T K v_I - u_I^T K v_R).$$

At least one of the four terms in this sum is not smaller in magnitude than $\alpha_{\mathbb{S}}/4$. Choose this term and let the columns of $W = [w_1, w_2] \in \mathbb{R}^{2n\times 2}$ contain the two vectors corresponding to it. For example, if $|u_R^T K v_R| \ge \alpha_{\mathbb{S}}/4$, then $W = [u_R, v_R]$. By the skew-symmetry of K, we may assume that u and v satisfy $v^T u = 0$, which implies $\|W\|_2 \le 1$. Setting $E = M^T W J_2 W^T \in \mathbb{S}\cap\mathbb{R}^{2n\times 2n}$ yields $\|E\|_2 \le 1$ and

$$\begin{aligned}\alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} &\ge \rho(J_{r_1}^T X^T M E X) = \rho(J_{r_1}^T X^T W J_2 W^T X)\\ &= \rho(K W J_2 W^T) = \rho(J_2 W^T K W)\\ &= \rho(\operatorname{diag}(w_2^T K w_1, -w_1^T K w_2)) \ge \alpha_{\mathbb{S}}/4,\end{aligned}$$

where we used the fact that $w_1^T K w_1 = w_2^T K w_2 = 0$ due to the skew-symmetry of K. □

Theorem 3.12. $M \in \mathbb{R}^{2n\times 2n}$ $\mathbb{S} = \{A \in \mathbb{C}^{2n\times 2n} :\ A^T M = -MA\}$. $A \in \mathbb{C}^{n\times n}$

(i) $\kappa(A,\lambda;\mathbb{S}) = (n_1, \alpha_{\mathbb{S}})$ $\alpha/\sqrt{2} \le \alpha_{\mathbb{S}} \le \alpha$ $\alpha_{\mathbb{S}} = \alpha$ 2

(ii) $\kappa(A,\lambda;\mathbb{S}\cap\mathbb{R}^{2n\times 2n}) = (n_1, \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}})$ $\alpha/8 \le \alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} \le \alpha_{\mathbb{S}}$ 2

Let u_1, v_1 with $\|u_1\|_2 = \|v_1\|_2 = 1$ be the left/right singular vectors belonging to the largest singular value of $XY^H M$ and define $\widetilde{E} = [v_1, \overline{u}_1]\left[\begin{smallmatrix} 0 & 1 \\ 1 & -v_1^T u_1 \end{smallmatrix}\right][v_1, \overline{u}_1]^T$. Then $\widetilde{E}$ is symmetric and one can show that $\|\widetilde{E}\|_F = \sqrt{2 - |u_1^T v_1|^2} \le \sqrt{2}$; see also [23, Theorem 5.6]. Setting $E = M\widetilde{E}/\sqrt{2}$, we obtain $E \in \mathbb{S}$ and

$$\alpha_{\mathbb{S}} \ge \rho(Y^H E X) = \rho(\widetilde{E} X Y^H M)/\sqrt{2} \ge \|XY^H\|_2/\sqrt{2},$$

where we applied Lemma 2.6, using the fact that $\widetilde{E}$ maps u_1 to v_1. In the matrix 2-norm, Theorem 5.7 in [23] implies the existence of a symmetric matrix $\widetilde{E}$ which maps u_1 to v_1 and satisfies $\|\widetilde{E}\|_2 = 1$. Thus, setting $E = M\widetilde{E}$ shows the second part of (i).

To show (ii), let us decompose $XY^HM = S + W$, where $S = (XY^H + \overline{Y}X^T)/2$ and $W = (XY^H - \overline{Y}X^T)/2$. Then $\alpha = \|XY^H\|_2 \leq \|S\|_2 + \|W\|_2$. We distinguish two cases, depending on whether the skew-symmetric part W dominates the symmetric part S.

1. $\|S\|_2 \geq \|W\|_2/3$: Decompose $S = S_R + \imath S_I$ with real symmetric matrices S_R, S_I. Then $\|S_R\|_2 \geq \|S\|_2/2$ or $\|S_I\|_2 \geq \|S\|_2/2$. In the first case, let u_1 be a normalized eigenvector belonging to an eigenvalue of S_R that has magnitude $\|S_R\|_2$. Then $E = Mu_1u_1^T \in \mathbb{S} \cap \mathbb{R}^{2n\times 2n}$ with $\|E\|_2 = 1$ and

$$\begin{aligned}\alpha_{\mathbb{S}\cap\mathbb{R}^{n\times n}} &\geq \rho(EXY^H) = |u_1^TXY^HMu_1| = |u_1^TSu_1| \geq |u_1^TS_Ru_1| \\ &\geq \frac{\|S\|_2}{2} = \frac{\|S\|_2 + 3\|S\|_2}{8} \geq \frac{\|S\|_2 + \|W\|_2}{8} \geq \frac{\alpha}{8}.\end{aligned}$$

   The case $\|S_I\|_2 \geq \|S\|_2/2$ can be shown analogously.

2. $\|S\|_2 \leq \|W\|_2/3$: Decompose $W = W_R + \imath W_I$ with real skew-symmetric matrices W_R, W_I. Suppose that $\|W_R\|_2 \geq \|W\|_2/2$ (once again, the case $\|W_I\|_2 \geq \|W\|_2/2$ is treated in an analogous manner). Let u_1, v_1 with $\|u_1\|_2 = \|v_1\|_2 = 1$ be left/right singular vectors belonging to the largest singular value of W_R. Since W_R is skew-symmetric, we have $v_1^Tu_1 = 0$. Setting

$$E = M[u_1, v_1]\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}[u_1, v_1]^T \in \mathbb{S} \cap \mathbb{R}^{2n\times 2n}$$

   yields $\|E\|_2 = 1$ and

$$\alpha_{\mathbb{S}\cap\mathbb{R}^{2n\times 2n}} \geq \rho(EXY^H) = \rho\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}[u_1, v_1]^T(S + W)[u_1, v_1]\right) = \rho(\Phi),$$

   where

$$\Phi = \begin{bmatrix} -\beta & 0 \\ 0 & \beta \end{bmatrix} + \begin{bmatrix} u_1^TSv_1 & v_1^TSv_1 \\ u_1^TSu_1 & u_1^TSv_1 \end{bmatrix}$$

   with $\beta = \|W_R\|_2 + \imath u_1^TW_Iv_1$. We have $\det(\Phi) = -(\beta + \gamma)(\beta - \gamma)$ with

$$\gamma = \sqrt{(u_1^TSu_1)(v_1^TSv_1) - (u_1^TSv_1)^2}$$

   satisfying $|\gamma| \leq \|S\|_2$. This shows

$$\begin{aligned}\rho(\Phi) &\geq |\beta| - |\gamma| \geq \|W_R\|_2 - \|S\|_2 \geq \frac{\|W\|_2}{2} - \|S\|_2 \\ &= \frac{\|W\|_2}{2} - \frac{9}{8}\|S\|_2 + \frac{1}{8}\|S\|_2 \geq \frac{\|S\|_2 + \|W\|_2}{8} \geq \frac{\alpha}{8},\end{aligned}$$

   which concludes the proof. □

Theorem 3.12(ii) reveals that forcing the perturbations in a real Hamiltonian matrix to respect the structure will generally have only a mild positive effect on the accuracy of multiple eigenvalues. However, it should be emphasized that condition numbers provide little insight on the direction in which perturbed eigenvalues are likely to move, an issue which is crucial in deciding whether a purely imaginary eigenvalue of a Hamiltonian matrix stays on the imaginary axis or not under (structured) perturbations, something which is often important in applications. For results in this direction, see [1, 26] and the references therein.

4. Generalized eigenvalue problems.

4.1. Matrix pencils.

4.1. Matrix pencils. Langer and Najman [17, 18, 19] extended Lidskii's perturbation theory, obtaining eigenvalue perturbation expansions for analytic matrix functions $L(\lambda)$. They used the local Smith form of $L(\lambda)$ in much the same way as the Jordan canonical form was used by Lidskii for matrices. In a recent paper, de Terán, Dopico, and Moro [6] have investigated the special case $L(\lambda) = A - \lambda B$, relating the results by Langer and Najman to the Kronecker–Weierstraß form, which is a more natural canonical form when $L(\lambda)$ is a matrix pencil. Let us briefly recall these results, restricting our attention to regular matrix pencils (A and B are square, $\det(L(\lambda)) \not\equiv 0$).

In the following, we denote the regular matrix pencil $A - \lambda B$ by (A, B). For a finite eigenvalue λ of (A, B), the Kronecker–Weierstraß form implies

$$\left[\begin{array}{c|c} J & 0 \\ \hline 0 & \widetilde{J}_A \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right] A \left[\begin{array}{c|c} P & \widetilde{P} \end{array}\right], \quad \left[\begin{array}{c|c} I & 0 \\ \hline 0 & \widetilde{J}_B \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right] B \left[\begin{array}{c|c} P & \widetilde{P} \end{array}\right], \tag{19}$$

where $[P, \widetilde{P}]$, $\left[\begin{smallmatrix} Q \\ \widetilde{Q} \end{smallmatrix}\right]$ are invertible and J contains all r_1 Jordan blocks of largest size n_1; see also (10). Similarly for an infinite eigenvalue of (A, B), we have

$$\left[\begin{array}{c|c} I & 0 \\ \hline 0 & \widetilde{J}_A \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right] A \left[\begin{array}{c|c} P & \widetilde{P} \end{array}\right], \quad \left[\begin{array}{c|c} N & 0 \\ \hline 0 & \widetilde{J}_B \end{array}\right] = \left[\begin{array}{c} Q \\ \hline \widetilde{Q} \end{array}\right] B \left[\begin{array}{c|c} P & \widetilde{P} \end{array}\right], \tag{20}$$

where N contains all r_1 nilpotent blocks of largest nilpotency index n_1. As for the standard eigenvalue problem, we collect the (generalized) right and left eigenvectors contained in P and Q:

$$\begin{array}{rcl} X & = & \left[\; Pe_1,\; Pe_{n_1+1},\; \ldots,\; Pe_{(r_1-1)n_1+1} \;\right], \\ Y & = & \left[\; Q^H e_{n_1},\; Q^H e_{2n_1},\; \ldots,\; Q^H e_{r_1 n_1} \;\right]. \end{array} \tag{21}$$

As in the standard eigenvalue problem, this relationship between X, Y and P, Q imposes some normalization on X, Y. For $n_1 = 1$, we have $Y^H BX = I$ if λ is finite and $Y^H AX = I$ if λ is infinite. For $n_1 > 1$, we have $Y^H AX = Y^H BX = 0$.

The following theorem summarizes results from [6].

THEOREM 4.1. [illegible] λ [illegible] (A, B) [illegible] $(E, F) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$ [illegible] $Y^H(E - \lambda F)X$ [illegible] X [illegible] Y [illegible] (21) [illegible] $n_1 r_1$ [illegible] $\hat{\lambda}_k$ [illegible] $(A + \epsilon E, B + \epsilon F)$ [illegible]

$$\hat{\lambda}_k = \lambda + (\xi_k)^{1/n_1}\epsilon^{1/n_1} + o(\epsilon^{1/n_1}), \quad k = 1, \ldots, r_1, \tag{22}$$

[illegible] $\xi_1, \ldots, \xi_{r_1}$ [illegible] $Y^H(E - \lambda F)X$ [illegible] (A, B) [illegible] $F \in \mathbb{C}^{n\times n}$ [illegible] $Y^H FX$ [illegible] $n_1 r_1$

[illegible] $\hat{\lambda}_k$ [illegible] $(A + \epsilon E, B + \epsilon F)$ [illegible]

$$\frac{1}{\hat{\lambda}_k} = (\xi_k)^{1/n_1}\epsilon^{1/n_1} + o(\epsilon^{1/n_1}), \quad k = 1, \dots, r_1, \tag{23}$$

[illegible] ξ_k [illegible] $Y^H F X$

4.1.1. Hölder condition numbers for multiple eigenvalues of matrix pencils. Throughout the rest of this section, λ denotes a finite or infinite eigenvalue of a regular matrix pencil (A, B) with the matrices X and Y defined as in (21).

Based on Theorem 4.1, we can define a condition number for a multiple eigenvalue of a matrix pencil as follows.

Definition 4.2. [illegible] Hölder condition number for a finite eigenvalue λ [illegible] $\kappa(A, B, \lambda) = (n_1, \alpha)$, [illegible] n_1 [illegible] λ [illegible]

$$\alpha = \sup_{\substack{\|E\| \le w_A, \|F\| \le w_B \\ E, F \in \mathbb{C}^{n \times n}}} \rho(Y^H(E - \lambda F)X).$$

[illegible] Hölder condition number for $\lambda = \infty$ [illegible] $\kappa(A, B, \infty) = (n_1, \alpha)$, [illegible] n_1 [illegible] (A, B) [illegible]

$$\alpha = \sup_{\substack{\|F\| \le w_B \\ F \in \mathbb{C}^{n \times n}}} \rho(Y^H F X).$$

[illegible] 4.3. Note that Definition 4.2 depends not only on the employed matrix norm $\|\cdot\|$ but also on the choice of nonnegative weights w_A and w_B. It is implicitly assumed that w_A or w_B is strictly larger than zero; otherwise $\kappa(A, B, \lambda) = (0, 0)$. More specifically, we require $w_A > 0$ for $\lambda = 0$, $w_B > 0$ for $\lambda = \infty$, and $\max\{w_A, w_B\} > 0$ for any other eigenvalue.

The role of the weights w_A and w_B is to balance the influence of perturbations on A and B. For example, if each of the perturbations E and F is known to be small compared to the norm of A and B, respectively, then it is reasonable to set $w_A = \|A\|/\sqrt{\|A\|^2 + \|B\|^2}$ and $w_B = \|B\|/\sqrt{\|A\|^2 + \|B\|^2}$.

The following lemma represents a direct extension of [27, Theorem 4.2].

Lemma 4.4. [illegible]

$$\kappa(A, B, \lambda) = (n_1, (w_A + w_B|\lambda|)\|XY^H\|_2)$$

[illegible] λ [illegible] $\kappa(A, B, \lambda) = (n_1, w_B\|XY^H\|)$ [illegible] $\lambda = \infty$

[illegible] On the one hand,

$$\begin{aligned} \rho(Y^H(E - \lambda F)X) = \rho((E - \lambda F)XY^H) &\le \|(E - \lambda F)XY^H\|_2 \\ &\le \|(E - \lambda F)XY^H\| \le (w_A + w_B|\lambda|)\|XY^H\|_2 \end{aligned}$$

holds for any E, F satisfying $\|E\| \le w_A$, $\|F\| \le w_B$. Hence, $\alpha \le (w_A + w_B|\lambda|)\|XY^H\|_2$. On the other hand, let u_1, v_1 with $\|u_1\|_2 = \|v_1\|_2 = 1$ be the left/right singular vectors belonging to the largest singular value of XY^H. Setting $E = w_A v_1 u_1^H$ and $F = -\frac{\bar{\lambda}}{|\lambda|} w_B v_1 u_1^H$ ($F = 0$ if $\lambda = 0$) yields $\|E\| \le w_A$, $\|F\| \le w_B$, and

$$\alpha \ge \rho((w_A + w_B|\lambda|)v_1 u_1^H XY^H) = (w_A + w_B|\lambda|)\rho(u_1^H XY^H v_1) = (w_A + w_B|\lambda|)\|XY^H\|_2.$$

The proof for $\kappa(A, B, \infty)$ is analogous. □

Definition 4.2 is based on the distance $|\hat{\lambda} - \lambda|$ between an eigenvalue λ and a perturbed eigenvalue $\hat{\lambda}$. This distance lacks mathematical elegance for generalized eigenvalue problems, since infinite eigenvalues must be treated separately and $|\hat{\lambda} - \lambda|$ is not invariant under an interchange of A and B, i.e., $|\hat{\lambda} - \lambda| \neq |1/\hat{\lambda} - 1/\lambda|$. A more elegant distance concept is offered by the chordal metric

$$\chi(\hat{\lambda}, \lambda) = \frac{|\hat{\lambda} - \lambda|}{\sqrt{|\hat{\lambda}|^2 + 1}\sqrt{|\lambda|^2 + 1}},$$

which naturally includes infinite eigenvalues

$$\chi(\hat{\lambda}, \infty) = \lim_{|\mu| \to \infty} \chi(\hat{\lambda}, \mu) = \frac{1}{\sqrt{|\hat{\lambda}|^2 + 1}};$$

see [34] for more details. Inserting the perturbation expansions (22) and (23) yields

$$\chi(\hat{\lambda}_k, \lambda) = \frac{|\xi_k \epsilon|^{1/n_1}}{|\lambda|^2 + 1} + o(\epsilon^{1/n_1})$$

and

$$\chi(\hat{\lambda}_k, \infty) = |\xi_k \epsilon|^{1/n_1} + o(\epsilon^{1/n_1}),$$

respectively. This shows that when working in the chordal metric, the α-part in the Hölder condition number for a finite eigenvalue needs to be divided by $|\lambda|^2 + 1$ while the Hölder condition number for an infinite eigenvalue remains the same. It is simple to see that this modified condition number has the pleasant property of being continuous at $|\lambda| = \infty$.

Whether $|\hat{\lambda} - \lambda|$ or $\chi(\hat{\lambda}, \lambda)$ is more appropriate depends on the application. If the ultimate goal of a computation is a finite eigenvalue λ, it can be suspected that $|\hat{\lambda} - \lambda|$ is practically more relevant. All the following results employ $|\hat{\lambda} - \lambda|$, but the discussion above reveals that it is rather easy to translate them into the chordal metric setting.

4.1.2. Structured Hölder condition numbers for eigenvalues of matrix pencils.

The structured Hölder condition number $\kappa(A, B, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ for some subset $\mathbb{S} \subset \mathbb{C}^{n \times n} \times \mathbb{C}^{n \times n}$ can be defined in the same way as for the standard eigenvalue problem. In particular, if $n_{\mathbb{S}} = n_1$, then

$$\alpha_{\mathbb{S}} = \sup_{\substack{\|E\| \leq w_A, \|F\| \leq w_B \\ (E,F) \in \mathbb{S}}} \rho(Y^H (E - \lambda F) X).$$

Some proofs from section 3 can be rather directly extended to yield results on generalized eigenvalue problems if the structure is separable, i.e., $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2$ with $\mathbb{S}_1, \mathbb{S}_2 \subset \mathbb{C}^{n \times n}$. The following theorem collects such results.

THEOREM 4.5. $\kappa(A, B, \lambda) = (n_1, \alpha)$ $\kappa(A, B, \lambda; \mathbb{S}_1 \times \mathbb{S}_2) = (n_{\mathbb{S}_1 \times \mathbb{S}_2}, \alpha_{\mathbb{S}_1 \times \mathbb{S}_2})$

(i) Real matrix pencils $\mathbb{S}_1 = \mathbb{S}_2 = \mathbb{R}^{n \times n}$ $n_{\mathbb{S}_1 \times \mathbb{S}_2} = n_1$ $\alpha/4 \leq \alpha_{\mathbb{S}_1 \times \mathbb{S}_2} \leq \alpha$ $A, B \in \mathbb{C}^{n \times n}$

(ii) Symmetric matrix pencils $\mathbb{S}_1 = \mathbb{S}_2 = \{A \in \mathbb{C}^{n \times n} : A^T = A\}$ $n_{\mathbb{S}_1 \times \mathbb{S}_2} = n_1$ $\alpha_{\mathbb{S}_1 \times \mathbb{S}_2} = \alpha$ $A, B \in \mathbb{S}_1$

(iii) Real symmetric matrix pencils $\mathbb{S}_1 = \mathbb{S}_2 = \{A \in \mathbb{R}^{n\times n} : A^T = A\}$ $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ $\alpha/4 \le \alpha_{\mathbb{S}_1\times\mathbb{S}_2} \le \alpha$ $A, B \in \mathbb{C}^{n\times n}$ $A^T = A$, $B^T = B$

(iv) Symmetric/skew-symmetric matrix pencils $\mathbb{S}_1 = \{A \in \mathbb{C}^{n\times n} : A^T = A\}$ $\mathbb{S}_2 = \{B \in \mathbb{C}^{n\times n} : B^T = -B\}$ $(A,B) \in \mathbb{S}_1 \times \mathbb{S}_2$ 2

(a) $\lambda = \infty$ n_1 $r_1 = 1$ $n_{\mathbb{S}} < n_1$.

(b) $\lambda = \infty$ n_1 $r_1 > 1$ $n_{\mathbb{S}} = n_1$ $\alpha_{\mathbb{S}} = w_B\sqrt{\sigma_1\sigma_2}$ σ_1, σ_2 XX^T $\alpha = w_B\sigma_1$.

(c) $\lambda = \infty$ n_1 r_1 $n_{\mathbb{S}} = n_1$

$$\alpha_{\mathbb{S}} = \alpha = w_B \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2;$$

(d) $\lambda = 0$ n_1 $n_{\mathbb{S}} = n_1$ $\alpha_{\mathbb{S}} = \alpha = w_A\|XX^T\|_2$.

(e) $\lambda = 0$ n_1 r_1 $n_{\mathbb{S}} = n_1$

$$\alpha_{\mathbb{S}} = \alpha = w_A \left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2;$$

(f) $\lambda \neq \infty, \lambda \neq 0$ $r_1 = 1$ $n_{\mathbb{S}} = n_1$ $\alpha_{\mathbb{S}} = w_A\alpha_1 + w_B|\lambda|\alpha_2$ $\alpha_1 = \|X\|_2\|Y\|_2$ $\alpha_2 = \sqrt{\|X\|_2^2\|Y\|_2^2 - |Y^TX|^2}$.

(g) $\lambda \neq \infty, \lambda \neq 0$ $r_1 > 1$ $n_{\mathbb{S}} = n_1$ $\alpha_{\mathbb{S}} \ge w_A\|XY^H\|_2$

(v) Skew-symmetric matrix pencils $\mathbb{S}_1 = \mathbb{S}_2 = \{A \in \mathbb{C}^{n\times n} : A^T = -A\}$ $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ r_1 $\alpha_{\mathbb{S}_1\times\mathbb{S}_2} = \alpha$ $A, B \in \mathbb{S}_1$ 2

(vi) Hermitian matrix pencils $\mathbb{S}_j = \{A \in \mathbb{C}^{n\times n} : A^H = \gamma_j A\}$ $j \in \{1,2\}$ $\gamma_1, \gamma_2 \in \{1, -1\}$ $n_{\mathbb{S}_1\times\mathbb{S}_2} = n_1$ $\alpha/\sqrt{2} \le \alpha_{\mathbb{S}_1\times\mathbb{S}_2} \le \alpha$ $A, B \in \mathbb{C}^{n\times n}$ 2 $\gamma_1 = \gamma_2$ $\lambda \in \mathbb{R}$ $\alpha_{\mathbb{S}_1\times\mathbb{S}_2} = \alpha$

If not stated otherwise, it is tacitly assumed that λ is finite (the proofs can be easily modified to cover $\lambda = \infty$).

(i) As in the proof of Lemma 3.2, we can find a real matrix $\widetilde{E}$ with $\|\widetilde{E}\| \le 1$ such that $\rho(Y^H\widetilde{E}X) \ge \|XY^H\|_2/2$. We set $E = w_A\widetilde{E}, F = 0$ if $w_A \ge w_B|\lambda|$, and $E = 0, F = w_B\widetilde{E}$ otherwise. Then

$$\rho(Y^H(E - \lambda F)X) \ge \frac{w_A + w_B|\lambda|}{2}\rho(Y^H\widetilde{E}X) \ge \frac{w_A + w_B|\lambda|}{4}\|XY^H\|_2,$$

which proves (i).

(ii) and (iii) Corollary A.2(ii) implies $Y = \overline{X}$, and hence assertions (ii) and (iii) can be shown along the lines of the proof of Theorem 3.7.

(iv) For $\lambda = \infty$, the structured canonical form of a symmetric/skew-symmetric pencil imposes the same structure on X and Y as for the zero eigenvalue of a skew-symmetric matrix; see Corollary A.4. This implies that α/w_B coincides with the structured condition number for the zero eigenvalue of B, and hence (a)–(c) follow from Theorem 3.8(i)–(iii).

For $\lambda = 0$ and n_1 even, Corollary A.4(ii)(d) yields $Y = \overline{X}$, so taking $E = w_A\overline{X}X^H/\|XX^T\|_2$, $F = 0$ shows (d). If $\lambda = 0$ and n_1 is odd, then r_1 is even and $Y = \overline{X}\left[\begin{smallmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{smallmatrix}\right]$; see Corollary A.4(ii)(c). Let u_1, v_1 be, respectively, left and right singular vectors corresponding to the largest singular value σ_1 of the skew-symmetric matrix $XY^H = X[\begin{smallmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{smallmatrix}]X^T$. Then, the pencil $E - \lambda F$ with $E = w_A[\overline{u_1}, v_1] \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] [\overline{u_1}, v_1]^T$ and $F = 0$ is such that $E \in \mathbb{S}_1$, $F \in \mathbb{S}_2$, $\|E\|_2 = w_A$, and

$$\alpha_{\mathbb{S}} \geq \rho\left(EX \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T\right) = \rho(V^H E U \Sigma) = \rho\left(\left[\begin{smallmatrix} w_A\sigma_1 & 0 \\ 0 & * \end{smallmatrix}\right]\right) \geq w_A\sigma_1 = \alpha,$$

where $U\Sigma V^H$ stands for a singular value decomposition of $X[\begin{smallmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{smallmatrix}]X^T$, $*$ denotes a nonzero $(r_1 - 1) \times (r_1 - 1)$ matrix, and we have used that $v_1^T u_1 = 0$, since $X[\begin{smallmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{smallmatrix}]X^T$ is skew-symmetric. This proves (e).

For finite nonzero λ and $r_1 = 1$, we have $\rho(Y^H(E - \lambda F)X) = |Y^H(E - \lambda F)X|$, so

$$\alpha_{\mathbb{S}} = \sup_{\substack{||E|| \leq w_A, ||F|| \leq w_B \\ (E,F) \in \mathbb{S}}} \rho(Y^H(E - \lambda F)X) \leq w_A \sup_{\substack{||E|| \leq 1 \\ E \in \mathbb{S}_1}} \rho(Y^H E X) + w_B|\lambda| \sup_{\substack{||F|| \leq 1 \\ F \in \mathbb{S}_2}} \rho(Y^H F X).$$

The supremum over $E \in \mathbb{S}_1$ is clearly bounded by α_1, while the supremum over $F \in \mathbb{S}_2$ is equal to α_2 by Theorem 3.8(iv). Thus $\alpha_{\mathbb{S}} \leq w_A\alpha_1 + w_B|\lambda|\alpha_2$. By [23, Theorem 5.7] there exists a matrix $\tilde{E} \in \mathbb{S}_1$ with $\|\tilde{E}\|_2 = \|Y\|_2/\|X\|_2$ such that $\tilde{E}X = Y$. Hence, the symmetric matrix $E_1 = \frac{\|X\|_2}{\|Y\|_2}\tilde{E}$ has unit 2-norm and attains the upper bound α_1. Let $F_2 \in \mathbb{S}_2$ be a matrix with unit 2-norm attaining the maximal value α_2. Then we may choose $\gamma_1, \gamma_2 \in \mathbb{C}$, $|\gamma_1| = |\gamma_2| = 1$ in such a way that the pair $(E, F) = (\gamma_1 w_A E_1, \gamma_2 w_B F_2) \in \mathbb{S}_1 \times \mathbb{S}_2$ satisfies $\rho(Y^H(E - \lambda F)X) = w_A\alpha_1 + w_B|\lambda|\alpha_2$. This proves (f).

For finite nonzero λ and $r_1 > 1$, recall that, according to the proof of Theorem 3.12, there is a symmetric matrix E_1 with $\|E_1\|_2 = 1$ and $\rho(Y^H E_1 X) \geq \|XY^H\|_2$. Thus, taking $E = w_A E_1$ and $F = 0$ leads to (g).

(v) For skew-symmetric/skew-symmetric pencils, Theorem A.5 shows that every eigenvalue has r_1 even. Furthermore, Corollary A.6(ii) reveals the relationship $Y = \overline{X}[\begin{smallmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{smallmatrix}]$. Hence, if we set $\widetilde{E} = \overline{X}[\begin{smallmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{smallmatrix}]X^H$, the perturbation matrices $E = \frac{w_A}{\|\widetilde{E}\|_2}\widetilde{E}$, $F = -\frac{w_B}{\|\widetilde{E}\|_2}\frac{\overline{\lambda}}{|\lambda|}\widetilde{E}$ are such that $\|E\|_2 = w_A$, $\|F\|_2 = w_B$, and

$$\rho(Y^H(E - \lambda F)X) = (w_A + |\lambda|w_B)\left\|X^T \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X\right\|_2 = \alpha.$$

(vi) As in the proof of Lemma 3.10, we can construct a Hermitian matrix $\widetilde{E}$ such that $\|\widetilde{E}\|_2 = 1$ and $\rho(\widetilde{E}XY^H) = \|XY^H\|_2$. Let us choose $\delta \in \{1, -1\}$ such that δ matches the sign of λ_R if $\gamma_1 = \gamma_2$ and the sign of $-\lambda_I$ otherwise. Then $E = w_A\sqrt{\gamma_1}\widetilde{E} \in \mathbb{S}_1$ and $F = -\delta w_B\sqrt{\gamma_2}\widetilde{E} \in \mathbb{S}_2$ yield

$$\alpha_{\mathbb{S}_1 \times \mathbb{S}_2} \geq \rho((E - \lambda F)XY^H) = |w_A\sqrt{\gamma_1} + \delta w_B\sqrt{\gamma_2}\lambda| \, \|XY^H\|_2 \geq \frac{w_A + w_B|\lambda|}{\sqrt{2}}\|XY^H\|_2.$$

Note that the last inequality follows from the fact that

$$\begin{aligned} 2|w_A\sqrt{\gamma_1} + \delta w_B\sqrt{\gamma_2}\lambda|^2 - (w_A + w_B|\lambda|)^2 &\geq w_A^2 - 2w_A w_B|\lambda| + w_B^2|\lambda|^2 \\ &= (w_A - w_B|\lambda|)^2 \geq 0. \end{aligned}$$

If $\gamma_1 = \gamma_2 = 1$ and $\lambda \in \mathbb{R}$, then $|w_A + \delta w_B\lambda| = w_A + w_B|\lambda|$ and the factor $1/\sqrt{2}$ can be removed. □

[illegible] 4.6. For definite Hermitian matrix pencils, the result of Theorem 4.5(vi) can be found in [35, 37] for semisimple λ and in [9] for simple λ.

[illegible], which are addressed by the following theorem, provide a practically relevant example for a structure that is not separable; see [12, 21] for more details and applications. Without loss of generality, we may assume $w_A = w_B = 1$ in this case, since $B = A^T$.

THEOREM 4.7. [illegible] $\mathbb{S} = \{(A, A^T) : A \in \mathbb{C}^{n\times n}\}$ [illegible] $w_A = w_B = 1$ [illegible] $\kappa(A, A^T, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ [illegible] $A \in \mathbb{C}^{n\times n}$ [illegible] 2 [illegible]

(i) [illegible] $\lambda = 1$ n_1 [illegible] $r_1 = 1$ [illegible] $n_{\mathbb{S}} < n_1$.

(ii) [illegible] $\lambda = 1$ n_1 [illegible] $r_1 > 1$ [illegible] $n_{\mathbb{S}} = n_1$ [illegible] $\alpha_{\mathbb{S}} = 2\sqrt{\sigma_1\sigma_2}$ [illegible] σ_1, σ_2 [illegible] XX^T [illegible] $\alpha = 2\sigma_1$.

(iii) [illegible] $\lambda = 1$ [illegible] n_1 [illegible] r_1 [illegible] $n_{\mathbb{S}} = n_1$ [illegible]

$$\alpha_{\mathbb{S}} = \alpha = 2\left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2 ;$$

(iv) [illegible] $\lambda = -1$ [illegible] n_1 [illegible] r_1 [illegible] $n_{\mathbb{S}} = n_1$ [illegible]

$$\alpha_{\mathbb{S}} = \alpha = 2\left\| X \begin{bmatrix} 0 & -I_{r_1/2} \\ I_{r_1/2} & 0 \end{bmatrix} X^T \right\|_2 ;$$

(v) [illegible] $\lambda = -1$ [illegible] n_1 [illegible] $n_{\mathbb{S}} = n_1$ [illegible] $\alpha_{\mathbb{S}} = \alpha = 2\|XX^T\|_2$.

(vi) [illegible] $\lambda \neq \pm 1$ [illegible] $r_1 = 1$ [illegible] $n_{\mathbb{S}} = n_1$ [illegible]

$$\frac{1}{2}(|1-\lambda|\,\alpha_1 + |1+\lambda|\,\alpha_2) \leq \alpha_{\mathbb{S}} \leq |1-\lambda|\,\alpha_1 + |1+\lambda|\,\alpha_2,$$

[illegible] $\alpha_1 = \|X\|_2\|Y\|_2$ [illegible] $\alpha_2 = \sqrt{\|X\|_2^2\|Y\|_2^2 - |Y^TX|^2}$.

(vii) [illegible] $\lambda \neq \pm 1$ [illegible] $r_1 > 1$ [illegible] $n_{\mathbb{S}} = n_1$ [illegible] $\frac{|1-\lambda|}{1+|\lambda|}\alpha \leq \alpha_{\mathbb{S}} \leq \alpha$.

(viii) [illegible] $\lambda = \infty$ [illegible] $n_{\mathbb{S}} = n_1$ [illegible] $\alpha_{\mathbb{S}} = \alpha$

[illegible]. If λ is finite and $n_{\mathbb{S}} = n_1$, then

$$\begin{aligned} \alpha_{\mathbb{S}} &= \sup_{\substack{\|E\|_2 \leq 1 \\ E \in \mathbb{C}^{n\times n}}} \rho(Y^H(E - \lambda E^T)X) \\ &= \frac{1}{2} \sup_{\substack{\|E\|_2 \leq 1 \\ E \in \mathbb{C}^{n\times n}}} \rho\left((1-\lambda)Y^H(E+E^T)X + (1+\lambda)Y^H(E-E^T)X\right). \qquad (24) \end{aligned}$$

This relation indicates that the analysis of palindromic matrix pencils is closely tied to the analysis of symmetric/skew-symmetric pencils. In fact, it has been shown [14, 30, 33] that the structured canonical form of a palindromic matrix pencil (A, A^T) can be extracted from the structured canonical form [38] of the symmetric/skew-symmetric pencil $(A + A^T, A - A^T)$. In particular, the relation between X and Y for an eigenvalue λ of (A, A^T) coincides with the relation between X and Y for the eigenvalue $-(1+\lambda)/(1-\lambda)$ of $(A + A^T, A - A^T)$.

Consequently, if $\lambda = 1$ and n_1 is odd, then Corollary A.4(ii)(a) implies $Y = \overline{X}$. If $\lambda = 1$ and n_1 is even, then r_1 is even and $Y = \overline{X}\left[\begin{smallmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{smallmatrix}\right]$. It follows from (24) that

$$\begin{aligned} \alpha_{\mathbb{S}} = \sup_{\substack{\|E\|_2 \leq 1 \\ E \in \mathbb{C}^{n\times n}}} \rho(Y^H(E - E^T)X) &\leq \sup_{\substack{\|E - E^T\|_2 \leq 2 \\ E \in \mathbb{C}^{n\times n}}} \rho(Y^H(E - E^T)X) \\ &= 2 \sup_{\substack{\|G\|_2 \leq 1 \\ G \text{ skew-symmetric}}} \rho(Y^HGX). \end{aligned}$$

On the other hand,

$$\begin{aligned} 2 \sup_{\substack{\|G\|_2 \le 1 \\ G \text{ skew-symmetric}}} \rho(Y^H G X) &= \sup_{\substack{\|G\|_2 \le 1 \\ G \text{ skew-symmetric}}} \rho(Y^H (G - G^T) X) \\ &\le \sup_{\substack{\|E\|_2 \le 1 \\ E \in \mathbb{C}^{n \times n}}} \rho(Y^H (E - E^T) X) \le \alpha_{\mathbb{S}}. \end{aligned}$$

This shows that the structured Hölder condition number for $\lambda = 1$ of (A, A^T) essentially coincides with the structured Hölder condition number for the eigenvalue $\lambda = 0$ of the skew-symmetric matrix $A - A^T$. In particular, Theorem 3.8(i)–(iii) yield assertions (i)–(iii) of this theorem.

If $\lambda = -1$ and n_1 is odd, then Corollary A.4(ii)(c) implies that r_1 is even and $Y = \overline{X}\left[\begin{smallmatrix} 0 & I_{r_1/2} \\ -I_{r_1/2} & 0 \end{smallmatrix}\right]$. If $\lambda = -1$ and n_1 is even, then $Y = \overline{X}$. As above, it follows from (24) that the situation in assertions (iv) and (v) is completely parallel to the one in items (e) and (d), respectively, of Theorem 4.5(iv). Thus, an analogous choice of symmetric E proves (iv) and (v).

For finite λ with $r_1 = 1$ (X and Y are vectors), relation (24) implies

$$\alpha_{\mathbb{S}} \le |1 - \lambda| \sup_{\substack{\|E_1\|_2 \le 1 \\ E_1 \text{ is symmetric}}} |Y^H E_1 X| + |1 + \lambda| \sup_{\substack{\|E_2\|_2 \le 1 \\ E_2 \text{ is skew-symmetric}}} |Y^H E_2 X|.$$

As shown in the proof of Theorem 4.5(iv)(f), the supremum over symmetric E_1 is equal to α_1 and, according to Theorem 3.8(iv), the one over skew-symmetric E_2 is equal to α_2. This shows $\alpha_{\mathbb{S}} \le |1 - \lambda|\, \alpha_1 + |1 + \lambda|\, \alpha_2$. Now, let E_1 be a symmetric matrix with $\|E_1\|_2 \le 1$ and $|Y^H E_1 X| = \alpha_1$, and let E_2 be a skew-symmetric matrix with $\|E_2\|_2 \le 1$ and $|Y^H E_2 X| = \alpha_2$. Then, the matrix $E = \gamma_1 E_1 + \gamma_2 E_2$ with suitable scalars γ_1, γ_2 satisfying $|\gamma_1| = |\gamma_2| = 1$ gives $|Y^H (E - \lambda E) X| = |1 - \lambda|\, \alpha_1 + |1 + \lambda|\, \alpha_2$ with $\|E\|_2 \le 2$, which yields

$$\alpha_{\mathbb{S}} \ge \frac{1}{2} (|1 - \lambda|\, \alpha_1 + |1 + \lambda|\, \alpha_2)$$

and concludes the proof of (vi).

For assertion (vii) we recall that there is a symmetric matrix E such that $\|E\|_2 = 1$ and $\rho(Y^H E X) \ge \|X Y^H\|_2$; see the proof of Theorem 3.12. Thus

$$\alpha_{\mathbb{S}} \ge \rho(Y^H (E - \lambda E^T) X) = |1 - \lambda|\, \rho(Y^H E X) \ge |1 - \lambda|\, \|X Y^H\|_2.$$

This proves the assertion since $\alpha = (1 + |\lambda|)\, \|X Y^H\|_2$.

Finally, (viii) is verified by observing that imposing palindromic structure does not change the definition of α for an infinite eigenvalue. □

Summarizing the statements of Theorem 4.7, one may conclude that the structured and unstructured (Hölder) condition numbers of a palindromic matrix pencil may differ significantly only for eigenvalues close to 1.

4.2. Matrix polynomials. Some seemingly more general variants of the generalized eigenvalue problem, such as polynomial and product eigenvalue problems, can be addressed with the concepts introduced above. We illustrate this point for a matrix polynomial

$$P(\lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \cdots + \lambda A_1 + A_0, \quad A_i \in \mathbb{C}^{n \times n}.$$

Nonzero vectors $x, y \in \mathbb{C}^{n\times n}$ are called, respectively, right and left eigenvectors belonging to an eigenvalue λ if $P(\lambda)x = 0$ and $y^H P(\lambda)x = 0$, respectively. In the following, we assume that P is regular, i.e., $\det P(\cdot) \not\equiv 0$. The $mn \times mn$ matrix pencil

(25)
$$A - \lambda B = \begin{bmatrix} -A_{m-1} & -A_{m-2} & \cdots & -A_1 & -A_0 \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix} - \lambda \begin{bmatrix} A_m & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & I \end{bmatrix}$$

is called the [illegible] of P and represents one of its most common linearizations. It is well known [7] that the eigenvalues of (A, B) coincide with those of P.

Because of this one-to-one relation between the eigenvalues, condition numbers for the eigenvalues of P can be derived from [illegible] eigenvalue condition numbers for (A, B) if the structure admits perturbations only in the blocks $A_0, \ldots, A_m$ of $A - \lambda B$. A consequence of this restriction on the perturbations is that the resulting eigenvalue condition numbers for the polynomial do not depend on the particular type of linearization chosen. The described approach has the advantage that we can make use of Theorem 4.5 and do not require more general concepts for matrix polynomials.

Following this approach, let us consider the perturbed matrix polynomial $P + \epsilon \triangle P$ with

$$\triangle P(\lambda) = \lambda^m E_m + \lambda^{m-1} E_{m-1} + \cdots + \lambda E_1 + E_0.$$

Equivalently, we can consider the correspondingly perturbed linearization $(A + \epsilon E, B + \epsilon F)$, where

$$E = -V \begin{bmatrix} E_{m-1} & E_{m-2} & \cdots & E_1 & E_0 \end{bmatrix}, \quad F = V E_m V^T, \tag{26}$$

and $V = [I_n, 0, \ldots, 0]^T$. To measure the perturbations, we allow n nonnegative weights $w_1, \ldots, w_n$, each corresponding to a coefficient of the matrix polynomial. As in Remark 4.3, to avoid degenerate situations we require $w_0 > 0$ for $\lambda = 0$, $w_m > 0$ for $\lambda = \infty$, and $\max\{w_0, \ldots, w_m\} > 0$ for any other eigenvalue.

DEFINITION 4.8. [illegible] λ [illegible] P [illegible] (A, B) [illegible] (25) [illegible] X [illegible] Y [illegible] (21) [illegible] (A, B) [illegible] (E, F) [illegible] (26) [illegible] Hölder condition number [illegible] λ [illegible] $\kappa(P, \lambda) = (n_1, \alpha)$, [illegible] n_1 [illegible] (A, B) [illegible] λ [illegible]

$$\alpha = \sup_{\substack{\|E_i\| \le w_i \\ E_i \in \mathbb{C}^{n\times n}}} \rho(Y^H (E - \lambda F) X).$$

[illegible] Hölder condition number [illegible] ∞ [illegible] $\kappa(P, \infty) = (n_1, \alpha)$, [illegible] n_1 [illegible] (A, B) [illegible]

$$\alpha = \sup_{\substack{\|E_i\| \le w_i \\ E_i \in \mathbb{C}^{n\times n}}} \rho(Y^H F X).$$

The results in [11, Lemma 7.2] and [10, Lemma 3.7] show that x_1 and y_1 are right and left eigenvectors belonging to a finite eigenvalue λ of P if and only if

$$x = \begin{bmatrix} \lambda^{m-1}x_1 \\ \vdots \\ \lambda x_1 \\ x_1 \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ (\lambda A_m + A_{m-1})^H y_1 \\ \vdots \\ (\lambda^{m-1}A_m + \lambda^{m-2}A_{m-1} + \cdots + A_1)^H y_1 \end{bmatrix} \tag{27}$$

are right and left eigenvectors of (A, B), respectively. For $\lambda = \infty$, the eigenvectors of (A, B) are given by $x = [x_1^H, 0, \dots, 0]^H$ and $y = [y_1^H, 0, \dots, 0]^H$. This shows that the matrices X and Y defined in (21), containing right and left eigenvectors belonging to a finite (multiple) eigenvalue λ of (A, B), take the form

$$X = \begin{bmatrix} \lambda^{m-1}X_1 \\ \vdots \\ \lambda X_1 \\ X_1 \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ (\lambda A_m + A_{m-1})^H Y_1 \\ \vdots \\ (\lambda^{m-1}A_m + \lambda^{m-2}A_{m-1} + \cdots + A_1)^H Y_1 \end{bmatrix}, \tag{28}$$

where X_1 and Y_1 are matrices of right and left eigenvectors of P. For an infinite eigenvalue only the first blocks of X and Y are nonzero and equal to X_1 and Y_1, respectively.

The following lemma provides an explicit formula for the Hölder condition number and also shows $\alpha > 0$ (under the mentioned conditions on the weights), which—strictly speaking—is needed to justify Definition 4.8.

LEMMA 4.9. [illegible] λ [illegible]

$$\kappa(P, \lambda) = (n_1, (w_m|\lambda|^m + w_{m-1}|\lambda|^{m-1} + \cdots + w_0)\|X_1Y_1^H\|_2)$$

[illegible] $\|\cdot\|$ [illegible] X_1 [illegible] Y_1 [illegible] P [illegible] X [illegible] Y [illegible] (A, B) [illegible] (28) [illegible] $\kappa(P, \lambda) = (n_1, w_m\|X_1Y_1^H\|_2)$

[illegible]. The structure of the matrices E, F, X, and Y shown in (26) and (28) implies

$$Y^H(E - \lambda F)X = -Y_1^H(\lambda^m E_m + \lambda^{m-1}E_{m-1} + \cdots + E_0)X_1.$$

As in the proof of Lemma 4.4, this shows

$$\rho(Y^H(E - \lambda F)X) \le (w_m|\lambda|^m + w_{m-1}|\lambda|^{m-1} + \cdots + w_0)\|X_1Y_1^H\|_2. \tag{29}$$

Let u, v with $\|u\|_2 = \|v\|_2 = 1$ be the left/right singular vectors belonging to the largest singular value of $X_1Y_1^H$. Then equality in (29) is attained for the perturbation coefficients

$$E_0 = w_0 vu^H, \; E_1 = w_1 \frac{\bar{\lambda}}{|\lambda|} vu^H, \; \dots, \; E_m = w_m \frac{\bar{\lambda}^m}{|\lambda|^m} vu^H,$$

with $E_1 = \cdots = E_m = 0$ for $\lambda = 0$. This proves the result for finite λ. For an infinite eigenvalue, the result follows analogously after observing $Y^H F X = Y_1^H E_m X_1$. □

It should be emphasized that X_1 and Y_1 cannot be chosen arbitrarily in Lemma 4.9; the result depends on the normalization of the matrices X and Y imposed by (21). To illustrate the effect of this normalization, let λ be a [illegible] finite eigenvalue of

P and suppose that $\widetilde{X}_1$ and $\widetilde{Y}_1$ contain [illegible] bases of right and left eigenvectors belonging to λ. If we let $\widetilde{X}$ and $\widetilde{Y}$ denote the corresponding bases for eigenvectors of (A, B), then (28) implies

$$\widetilde{Y}^H B \widetilde{X} = \widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1.$$

Since λ is semisimple and finite, the matrix $\widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1$ is invertible and

$$X_1 = \widetilde{X}_1 (\widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1)^{-1}, \quad Y_1 = \widetilde{Y}_1$$

satisfy $Y_1^H P'(\lambda) X_1 = I$, which amounts to the condition imposed by (21) for a semisimple eigenvalue. By Lemma 4.9,

$$\kappa(P, \lambda) = \Big(1, (w_m|\lambda|^m + w_{m-1}|\lambda|^{m-1} + \cdots + w_0) \big\| \widetilde{X}_1 (\widetilde{Y}_1^H P'(\lambda) \widetilde{X}_1)^{-1} \widetilde{Y}_1^H \big\|_2 \Big).$$

For $r_1 = 1$, this formula coincides with a result by Tisseur [39, Theorem 5] on the condition number for a simple eigenvalue of a matrix polynomial.

Finally, let us emphasize again that the companion form linearization serves a purely theoretical purpose here. If one admits general, unstructured perturbations to the linearization, then the corresponding condition numbers do depend on the linearization; see the discussion in [10, 11]. In particular, [11] shows how to minimize the unstructured condition number for a simple eigenvalue of the linearization. This is useful when applying an unstructured method, such as the QZ algorithm, to compute the eigenvalue via the linearization. The extension of these results to multiple eigenvalues would require comparing the result of Lemma 4.9 with the unstructured Hölder condition numbers of a linearization. Also, it could be of interest to study the effect on the Hölder condition numbers if further structure is imposed on the coefficients of the matrix polynomial and this structure is preserved by the linearization [21]. Some results in this direction concerning structured pseudospectra can be found in [8, 41].

5. Conclusions.

A definition of structured Hölder condition number for multiple eigenvalues, both of matrices and of regular matrix pencils, has been introduced with the purpose of comparing structured and unstructured condition numbers for several classes of structured matrices and pencils. Moreover, eigenvalues of matrix polynomials can be treated within this framework via linearization through companion form. Like previous Hölder condition numbers in the literature, the structured condition number $\kappa(A, \lambda; \mathbb{S}) = (n_{\mathbb{S}}, \alpha_{\mathbb{S}})$ has two entries, the first one related to the leading exponent, the second one to the leading coefficient in the asymptotic expansions of perturbed eigenvalues. Although the present paper focuses on the case when the first entry $n_{\mathbb{S}}$ coincides with the one in the unstructured condition number, some examples are given when this does not happen (see, e.g., (6) and (7)).

According to the results in this paper, the behavior of multiple eigenvalues under structured perturbations does not differ much from the one for simple eigenvalues described in [3, 16, 32], in the sense that the influence of structure on the condition number is usually mild, except in a few, quite specific situations. All these situations seem to be related to a combination of symmetry with skew-symmetry, either for matrices which are skew-symmetric with respect to a symmetric bilinear form (Theorem 3.8(ii) and (iv)) or for symmetric/skew-symmetric pencils (Theorem 4.5(iv)). Palindromic pencils (Theorem 4.7) represent another instance of the interplay between symmetry and skew-symmetry; see (24). Understanding why this happens is one of the open questions raised by such results. Also, there are a few cases where all we can say is that $n_{\mathbb{S}} = n_1$, with no further information to compare $\alpha_{\mathbb{S}}$ and α. Such cases remain as objects of future study.

Another open problem is a more complete picture of what happens in those cases where $n_{\mathbb{S}} < n_1$, i.e., whenever structured perturbations induce a behavior qualitatively different from the one induced by unstructured ones.

Appendix. Structured canonical forms. This section collects known results on canonical forms for structured matrices and matrix pencils used in this paper. The forms are constructed as direct sums of the following $m \times m$ matrices:

[illegible] [illegible]

$$J_m(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}, \qquad F_m = \begin{bmatrix} & & & 1 \\ & & \cdot^{\cdot^{\cdot}} & \\ & \cdot^{\cdot^{\cdot}} & & \\ 1 & & & \end{bmatrix},$$

[illegible] m [illegible] [illegible] m [illegible]

$$\Sigma_m = \begin{bmatrix} I_{m/2} & \\ & -I_{m/2} \end{bmatrix}, \qquad \Sigma_m = \begin{bmatrix} 0 & I_{(m-1)/2} & \\ 0 & & -I_{(m-1)/2} \\ 0 & 0 & 0 \end{bmatrix}.$$

Proofs of the following theorems can be found in Thompson's overview paper [38] and in the more recent and more general work [25].

THEOREM A.1 (complex symmetric matrix pencils). [illegible] $A - \lambda B$ [illegible] A [illegible] B [illegible] S [illegible]

$$S^T(A - \lambda B)S = (A_1 - \lambda B_1) \oplus \cdots \oplus (A_p - \lambda B_p),$$

[illegible]

(i) $A_j - \lambda B_j = F_{n_j} - \lambda F_{n_j} J_{n_j}(0)$ [illegible]

(ii) $A_j - \lambda B_j = F_{n_j} J_{n_j}(\lambda_j) - \lambda F_{n_j}$ [illegible] λ_j

COROLLARY A.2.

(i) [illegible] λ_j [illegible] A [illegible] $A^T M = MA$ [illegible] M [illegible] X [illegible] Y [illegible] (11)–(12) [illegible] $Y = M\overline{X}$

(ii) [illegible] λ_j [illegible] $A - \lambda B$ [illegible] A [illegible] B [illegible] X [illegible] Y [illegible] (21) [illegible] $Y = \overline{X}$

[illegible]. The first part is proven by applying Theorem A.1 to the pencil $MA - \lambda M$. This yields a matrix S with $S^{-1} = (F_{n_1} \oplus \cdots \oplus F_{n_p})S^T M$ such that $S^{-1}AS$ is in Jordan canonical form. The result follows by inspection of (11)–(12). The second part follows directly from combining Theorem A.1 with (21). □

THEOREM A.3 (complex skew-symmetric/symmetric matrix pencils). [illegible] $A - \lambda B$ [illegible] A [illegible] B [illegible] S [illegible]

$$S^T(A - \lambda B)S = (A_1 - \lambda B_1) \oplus \cdots \oplus (A_p - \lambda B_p),$$

[illegible]

(i) $A_j - \lambda B_j = F_{n_j}\Sigma_{n_j} - \lambda F_{n_j} J_{n_j}(0)$ [illegible] n_j [illegible]

(ii) $A_j - \lambda B_j = [\begin{smallmatrix} 0 & F_{n_j}-\lambda F_{n_j}J_{n_j}(0) \\ -F_{n_j}-\lambda F_{n_j}J_{n_j}(0) & 0 \end{smallmatrix}]$, [illegible] n_j [illegible].

(iii) $A_j - \lambda B_j = F_{n_j}\Sigma_{n_j} - \lambda F_{n_j}$, [illegible] n_j [illegible].

(iv) $A_j - \lambda B_j = [\begin{smallmatrix} 0 & F_{n_j}J_{n_j}(\lambda_j)-\lambda F_{n_j} \\ -F_{n_j}J_{n_j}(\lambda_j)-\lambda F_{n_j} & 0 \end{smallmatrix}]$, [illegible] $\pm\lambda_j$, [illegible] λ_j [illegible] n_j [illegible].

Corollary A.4.

(i) [illegible] λ_j [illegible] A [illegible] $A^TM = -MA$, [illegible] M [illegible] X [illegible] Y [illegible] (11)–(12)

(a) [illegible] $\lambda_j = 0$ [illegible] n_j [illegible] $Y = M\overline{X}$.

(b) [illegible] $\lambda_j = 0$ [illegible] n_j [illegible] $Y = M\overline{X}[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}]$.

(c) [illegible] $\lambda_j \neq 0$ [illegible] $\widetilde{X}$ [illegible] $\widetilde{Y}$ [illegible] $-\lambda_j$ [illegible] $\widetilde{X} = -Y, \widetilde{Y} = X$.

(ii) [illegible] λ_j [illegible] $A - \lambda B$ [illegible] A [illegible] B [illegible] X [illegible] Y [illegible] (21)

(a) [illegible] $\lambda_j = \infty$ [illegible] n_j [illegible] $Y = \overline{X}$.

(b) [illegible] $\lambda_j = \infty$ [illegible] n_j [illegible] $Y = \overline{X}[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}]$.

(c) [illegible] $\lambda_j = 0$ [illegible] n_j [illegible] $Y = \overline{X}[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}]$.

(d) [illegible] $\lambda_j = 0$ [illegible] n_j [illegible] $Y = \overline{X}$.

(e) [illegible] $\lambda_j \neq 0$ [illegible] $\widetilde{X}$ [illegible] $\widetilde{Y}$ [illegible] $-\lambda_j$ [illegible] $\widetilde{X} = -Y, \widetilde{Y} = X$.

[illegible]. The first part is proven by applying Theorem A.3 to $A - \lambda M$. If $\lambda_j = 0$ and n_j is odd, then Theorem A.3(iii) yields the relation $Q = M\overline{P}(F_{n_j} \oplus \cdots \oplus F_{n_j})$ for the matrices P and Q defined in (8). An inspection of (11)–(12) verifies (a). If $\lambda_j = 0$ and n_j is even, then Theorem A.3(iv) yields

$$Q = M\overline{P}\left(\begin{bmatrix} 0 & F_{n_j} \\ -F_{n_j} & 0 \end{bmatrix} \oplus \cdots \oplus \begin{bmatrix} 0 & F_{n_j} \\ -F_{n_j} & 0 \end{bmatrix}\right)$$

and thus $Y = M\overline{X}([\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}] \otimes \cdots \otimes [\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}])$. A perfect shuffle of the columns of X and Y yields (b). A similar argument leads to (c).

The second part is proven by applying Theorem A.3 to $B - \lambda A$. □

Theorem A.5 (complex skew-symmetric matrix pencils). [illegible] $A - \lambda B$ [illegible] A [illegible] B [illegible] S [illegible]

$$S^T(A - \lambda B)S = (A_1 - \lambda B_1) \oplus \cdots \oplus (A_p - \lambda B_p),$$

[illegible]

(i) $A_j - \lambda B_j = [\begin{smallmatrix} 0 & F_{n_j}-\lambda F_{n_j}J_{n_j}(0) \\ -F_{n_j}+\lambda F_{n_j}J_{n_j}(0) & 0 \end{smallmatrix}]$, [illegible].

(ii) $A_j - \lambda B_j = [\begin{smallmatrix} 0 & F_{n_j}J_{n_j}(\lambda_j)-\lambda F_{n_j} \\ -F_{n_j}J_{n_j}(\lambda_j)+\lambda F_{n_j} & 0 \end{smallmatrix}]$, [illegible].

Corollary A.6.

(i) [illegible] λ_j [illegible] A [illegible] $A^TM = MA$, [illegible] M [illegible] X [illegible] Y [illegible] (11)–(12) [illegible] $Y = -M\overline{X}[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}]$

(ii) [illegible] λ_j [illegible] $A - \lambda B$ [illegible] A [illegible] B [illegible] X [illegible] Y [illegible] (21) [illegible] $Y = -\overline{X}[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}]$.

Proof. Using Theorem A.5, the proof is analogous to the proof of Corollary A.4(i)(b). □

Acknowledgments. The second author wishes to thank the hospitality of both the Institut für Mathematik at TU Berlin and the Department of Computing Science at Umeå University. Part of the research leading to this paper was conducted while she was visiting both institutions. Thanks also go to Volker Mehrmann for many helpful discussions.

REFERENCES

[1] S. Bora and V. Mehrmann, *Linear perturbation theory for structured matrix pencils arising in control theory*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 148–169.

[2] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.

[3] R. Byers and D. Kressner, *On the condition of a complex eigenvalue under real perturbations*, BIT, 44 (2004), pp. 209–215.

[4] F. Chaitin-Chatelin, A. Harrabi, and A. Ilahi, *About Hölder condition numbers and the stratification diagram for defective eigenvalues*, Math. Comput. Simulation, 54 (2000), pp. 397–402.

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

[6] F. de Terán, F. Dopico, and J. Moro, *First order spectral perturbation theory of square singular matrix pencils*, Linear Algebra Appl., 429 (2008), pp. 548–576.

[7] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[8] S. Graillat, *A note on structured pseudospectra*, J. Comput. Appl. Math., 191 (2006), pp. 68–76.

[9] 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.

[10] N. J. Higham, R.-C. Li, and F. Tisseur, *Backward error of polynomial eigenproblems solved by linearization*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1218–1241.

[11] N. J. Higham, D. S. Mackey, and F. Tisseur, *The conditioning of linearizations of matrix polynomials*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1005–1028.

[12] A. Hilliges, C. Mehl, and V. Mehrmann, *On the solution of palindromic eigenvalue problems*, in Proceedings of ECCOMAS, Jyväskylä, Finland, 2004.

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

[14] R. A. Horn and V. V. Sergeichuk, *Canonical forms for complex matrix congruence and* **congruence*, Linear Algebra Appl., 416 (2006), pp. 1010–1032.

[15] M. Karow, *Structured Pseudospectra and the Condition of a Nonderogatory Eigenvalue*, Technical report 407, DFG Research Center, MATHEON Mathematics for Key Technologies in Berlin, TU Berlin, Berlin, Germany, 2007.

[16] M. Karow, D. Kressner, and F. Tisseur, *Structured eigenvalue condition numbers*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1052–1068.

[17] H. Langer and B. Najman, *Remarks on the perturbation of analytic matrix functions.* II, Integral Equations Operator Theory, 12 (1989), pp. 392–407.

[18] H. Langer and B. Najman, *Remarks on the perturbation of analytic matrix functions.* III, Integral Equations Operator Theory, 15 (1992), pp. 796–806.

[19] H. Langer and B. Najman, *Leading coefficients of the eigenvalues of perturbed analytic matrix functions*, Integral Equations Operator Theory, 16 (1993), pp. 600–604.

[20] V. B. Lidskiĭ, *On the theory of perturbations of nonselfadjoint operators*, Ž. Vyčisl. Mat. i Mat. Fiz., 6 (1966), pp. 52–60.

[21] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Structured polynomial eigenvalue problems: Good vibrations from good linearizations*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1029–1051.

[22] D. S. Mackey, N. Mackey, and F. Tisseur, $\mathbb{G}$*-reflectors: Analogues of Householder transformations in scalar product spaces*, Linear Algebra Appl., 385 (2004), pp. 187–213.

[23] D. S. Mackey, N. Mackey, and F. Tisseur, *Structured mapping problems for matrices associated with scalar products. Part* I*: Lie and Jordan algebras*, SIAM J. Matrix Anal. Appl., 29 (2008), pp. 1389–1410.

[24] R. Mathias, *The singular values of the Hadamard product of a positive semidefinite and a skew-symmetric matrix*, Linear and Multilinear Algebra, 31 (1992), pp. 57–70.

[25] C. Mehl, *On classification of normal matrices in indefinite inner product spaces*, Electron. J. Linear Algebra, 15 (2006), pp. 50–83.

[26] V. Mehrmann and H. Xu, *Perturbation of purely imaginary eigenvalues of Hamiltonian matrices under structured perturbations*, Electron. J. Linear Algebra, 17 (2008), pp. 234–257.

[27] J. Moro, J. V. Burke, and M. L. Overton, *On the Lidskii–Vishik–Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 793–817.

[28] S. Noschese and L. Pasquini, *Eigenvalue condition numbers: Zero-structured versus traditional*, J. Comput. Appl. Math., 185 (2006), pp. 174–189.

[29] M. Peláez and J. Moro, *Structured condition numbers of multiple eigenvalues*, Proceedings in Applied Mathematics and Mechanics, 6 (2006), pp. 67–70.

[30] L. Rodman, *Bounded and stably bounded palindromic difference equations of first order*, Electron. J. Linear Algebra, 15 (2006), pp. 22–49.

[31] S. M. Rump, *Structured perturbations part* I*: Normwise distances*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 1–30.

[32] S. M. Rump, *Eigenvalues, pseudospectrum and structured perturbations*, Linear Algebra Appl., 413 (2006), pp. 567–593.

[33] C. Schröder, *A Canonical Form for Palindromic Pencils and Palindromic Factorizations*, Technical report 316, DFG Research Center, MATHEON Mathematics for Key Technologies in Berlin, TU Berlin, Berlin, Germany, 2006.

[34] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[35] J.-G. Sun, *A note on local behavior of multiple eigenvalues*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 533–541.

[36] J.-G. Sun, *Multiple eigenvalue sensitivity analysis*, Linear Algebra Appl., 137/138 (1990), pp. 183–211.

[37] J.-G. Sun, *Stability and Accuracy: Perturbation Analysis of Algebraic Eigenproblems*, Technical report UMINF 98-07, Department of Computing Science, University of Umeå, Umeå, Sweden, 1998. Revised 2002.

[38] R. C. Thompson, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147 (1991), pp. 323–371.

[39] F. Tisseur, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl., 309 (2000), pp. 339–361.

[40] F. Tisseur, *A chart of backward errors for singly and doubly structured eigenvalue problems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 877–897.

[41] F. Tisseur and N. J. Higham, *Structured pseudospectra for polynomial eigenvalue problems, with applications*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 187–208.

[42] M. I. Višik and L. A. Ljusternik, *Solution of some perturbation problems in the case of matrices and self-adjoint or non-selfadjoint differential equations.* I, Russian Math. Surveys, 15 (1960), pp. 1–73.

[43] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

[44] D. Xie, X. Hu, and L. Zhang, *The solvability conditions for inverse eigenproblem of symmetric and anti-persymmetric matrices and its approximation*, Numer. Linear Algebra Appl., 10 (2003), pp. 223–234.

[45] H. Xie and H. Dai, *On the sensitivity of multiple eigenvalues of nonsymmetric matrix pencils*, Linear Algebra Appl., 374 (2003), pp. 143–158.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 203–226

STRATIFICATION OF CONTROLLABILITY AND OBSERVABILITY PAIRS—THEORY AND USE IN APPLICATIONS[*]

ERIK ELMROTH[†], STEFAN JOHANSSON[†], AND BO KÅGSTRÖM[†]

Abstract. Cover relations for orbits and bundles of controllability and observability pairs associated with linear time-invariant systems are derived. The cover relations are combinatorial rules acting on integer sequences, each representing a subset of the Jordan and singular Kronecker structures of the corresponding system pencil. By representing these integer sequences as coin piles, the derived stratification rules are expressed as minimal coin moves between and within these piles, which satisfy and preserve certain monotonicity properties. The stratification theory is illustrated with two examples from systems and control applications, a mechanical system consisting of a thin uniform platform supported at both ends by springs, and a linearized Boeing 747 model. For both examples, nearby uncontrollable systems are identified as subsets of the complete closure hierarchy for the associated system pencils.

Key words. stratification, matrix pairs, controllability, observability, robustness, Kronecker structures, orbit, bundle, closure hierarchy, cover relations, StratiGraph

AMS subject classifications. 15A21, 15A22, 65F15, 93B05, 93B07

DOI. 10.1137/080717547

1. Introduction.

Computing the canonical structure of a linear time-invariant (LTI) system, $\dot{x}(t) = Ax(t) + Bu(t)$ with states $x(t)$ and inputs $u(t)$, is an ill-posed problem; i.e., small changes in the input data matrices A and B may drastically change the computed canonical structure of the associated system pencil $\begin{bmatrix} A - \lambda I & B \end{bmatrix}$ (e.g., see [13]). Besides knowing the canonical structure, it is equally important to be able to identify nearby canonical structures in order to explain the behavior and possibly determining the robustness of a state-space system under small perturbations. For example, a state-space system which is found to be controllable may be very close to an uncontrollable one; and can, therefore, by only a small change in some data, e.g., due to round-off or measurement errors, become uncontrollable. If the LTI system considered and all nearby systems in a given neighborhood are controllable, the system is called *robustly controllable* (e.g., see [46]).

The qualitative information about nearby linear systems is revealed by the theory of stratification for the corresponding system pencil. A *stratification* shows which canonical structures are near to each other (in the sense of small perturbations) and their relation to other structures; i.e., the theory reveals the closure hierarchy of orbits and bundles of canonical structures. A cover relation guarantees that two canonical structures are nearest neighbors in the closure hierarchy.

For square matrices, Arnold [1] examined nearby structures by small perturbations using versal deformations. For matrix pencils, Elmroth and Kågström [23] first investigated the set of 2-by-3 matrix pencils and later extended the theory, in collaboration with Edelman, to general matrices and matrix pencils [17, 18]. In line of this work, the theory has further been developed in [21], and for matrix pairs

[*] Received by the editors March 3, 2008; accepted for publication (in revised form) by P. Benner October 3, 2008; published electronically March 11, 2009. A preliminary version of this paper appeared as **Report UMINF 08-03**. Financial support was provided by the *Swedish Foundation for Strategic Research* under the frame program grant A3 02:128.

http://www.siam.org/journals/simax/31-2/71754.html

[†] Department of Computing Science, Umeå University, SE-901 87, Sweden (elmroth@cs.umu.se, stefanj@cs.umu.se, bokg@cs.umu.se).

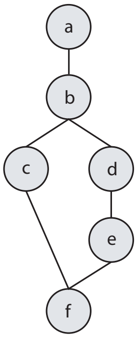


FIG. 1.1. *A graph presenting a hypothetical closure hierarchy, where letters (a–f) represent some canonical structures, the nodes represent orbits of these structures, and the edges represent covering relations.*

in [20, 42]. Several other people have worked on the theory of stratifications and similar topics, and we refer to [2, 27, 31, 35, 49] and references therein. Furthermore, the related topic distance to uncontrollability has recently been studied in, e.g., [6, 22, 30, 33, 34, 46].

In this paper, we derive the *cover relations* for independent controllability and observability pairs associated with LTI systems. These relations are combinatorial rules acting on integer sequences, each representing a subset of the Jordan and singular Kronecker structures (canonical form) of the corresponding system pencil. By following [17, 18], and representing these integer sequences as coin piles, the derived stratification rules are expressed as simple coin moves between and within these piles. Besides, only coin moves that satisfy and preserve certain monotonicity properties of the integer sequences are valid moves.

Before we go into further details, we outline the contents of the rest of the paper. In section 2, some linear systems background, including matrix pencil representations, are presented. In addition, a subsection introduces minimum coin moves for piles of coins representing integer partitions that frequently appear in the covering rules. Section 3 gives a concise presentation of the Kronecker canonical form (KCF) of a general matrix pencil and its invariants, as well as the Brunovsky canonical form for various system pencils. In section 4, system pencils for matrix pairs are considered. Concepts introduced include orbits and bundles for controllability and observability pairs, matrix representations for associated tangent spaces, and their codimensions expressed in terms of the KCF invariants. Equipped with all these concepts and notation, section 5 is devoted to the stratification theory, focusing on the derivation of cover relations for matrix pair orbits and bundles. In section 6, we illustrate the stratification theory by considering two examples from systems and control applications, a mechanical system consisting of a thin uniform platform supported at both ends by springs [44], and a linearized Boeing 747 model [51]. For both examples, we identify nearby uncontrollable systems as subsets of the complete closure hierarchy for the associated system pencils.

Following [18, 23], we present stratifications as graphs where each node represents an orbit or a bundle of a canonical structure and an edge represents a covering relation. A graph is organized with the most generic structure(s) at the top and other structures further down, ordered by increasing degeneracy (increasing codimension). Figure 1.1 illustrates how to interpret such a graph, assuming that each node represents the orbit of some canonical structure.

The topmost node shows the structure denoted a as the most generic structure. The edge to the node b illustrates that a covers b; i.e., the orbit of b is in the closure of that of a and there are no other structures between them in the closure hierarchy. Notably, all structures in the closure of b are also in the closure of a, although there are no covering relations between a and these structures since b appears between them in the hierarchy. Continuing downwards, b covers both c and d and there is no covering relation between c and d. Further down, the orbit of e is in the closure of that of d but not in the closure of c's orbit. The most degenerate structure is f, which is covered by both c and e, actually showing that f's orbit is in the intersection of the orbits of c and e. In this example, f is the most degenerate structure, whose orbit is in the closure of all other orbits.

In section 6, we make use of this type of graphs to illustrate closure hierarchies. The graphs presented are generated with StratiGraph [21, 38, 40, 41], which is a software tool for determining and presenting closure hierarchies based on the theory in [17, 18, 42]. The current version of StratiGraph (v. 2.2) has support for stratification of matrices, matrix pencils, and controllability and observability pairs. The theory of the latter is presented and illustrated in this paper.

2. Background and notation. A *linear time-invariant, finite dimensional system* (LTI system) is in continuous time represented as a state-space model by a system of the differential equations

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.1}$$

where $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, and $D \in \mathbb{C}^{p\times m}$. Such a state-space system is in short form represented by the *quadruple of matrices* (A, B, C, D).

System (2.1) is said to be *controllable* if there exists an input signal $u(t)$, $t_0 \leq t \leq t_{\mathrm{f}}$, that takes every state variable from an initial state $x(t_0)$ to a desired final state $x(t_{\mathrm{f}})$ in finite time. Otherwise it is said to be *uncontrollable.* The dual concept of controllability is observability. System (2.1) is said to be *observable* if it is possible to find the initial state $x(t_0)$ from the input signal $u(t)$ and the output signal $y(t)$ measured over a finite interval $t_0 \leq t \leq t_{\mathrm{f}}$. Otherwise it is said to be *unobservable.*

The controllability and observability of a system depend only on the matrix pairs (A, B) and (A, C), respectively, associated with the particular systems

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad \text{and} \qquad \begin{aligned} \dot{x}(t) &= Ax(t), \\ y(t) &= Cx(t), \end{aligned}$$

of (2.1). The matrix pairs (A, B) and (A, C) are referred to as the *controllability* and *observability pairs*, respectively.

2.1. The pencil representation. The set of matrices of the form $G - \lambda H$ with $\lambda \in \mathbb{C}$ corresponds to a general *matrix pencil*, where the two complex matrices G and H are of size $m_{\mathrm{p}} \times n_{\mathrm{p}}$. Notice that all matrix pencils where $m_{\mathrm{p}} \neq n_{\mathrm{p}}$ are singular, which is the case in most control applications.

A state-space system (2.1) can also be represented and analyzed in terms of a matrix pencil, which in this special form is called a *system pencil*, $\mathbf{S}(\lambda)$. Contrary to a general matrix pencil, a system pencil emphasizes the structure of the system. The associated system pencil for the state-space system (2.1) is

$$\mathbf{S}(\lambda) = G - \lambda H = \begin{bmatrix} A & B \\ C & D \end{bmatrix} - \lambda \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.2}$$

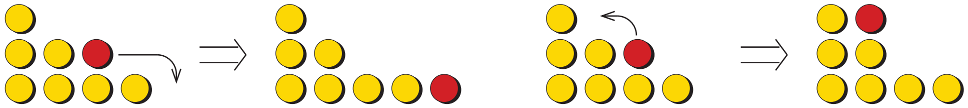


FIG. 2.1. *Minimum rightward and leftward coin moves illustrate that* $\kappa = (3, 2, 2, 1)$ *covers* $\nu = (3, 2, 1, 1, 1)$ *and* $\kappa = (3, 2, 2, 1)$ *is covered by* $\tau = (3, 3, 1, 1)$.

where G and H are of size $(n + p) \times (n + m)$ and, consequently, $m_{\mathrm{p}} = n + p$ and $n_{\mathrm{p}} = n+m$. The corresponding system pencils for the controllability and observability pairs are

$$\mathbf{S}_{\mathrm{C}}(\lambda) = \begin{bmatrix} A & B \end{bmatrix} - \lambda \begin{bmatrix} I_n & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{S}_{\mathrm{O}}(\lambda) = \begin{bmatrix} A \\ C \end{bmatrix} - \lambda \begin{bmatrix} I_n \\ 0 \end{bmatrix}.$$

In the rest of the paper, we are mainly only considering the controllability and observability pairs and their associated system pencils.

2.2. Integer partitions and coins. We give a brief introduction to integer partitions and minimum coin moves, which are used to represent the invariants of the matrix and system pencils and to define the stratification rules.

An *integer partition* $\kappa = (\kappa_1, \kappa_2, \dots)$ of an integer K is a monotonically decreasing sequence of integers ($\kappa_1 \geq \kappa_2 \geq \cdots \geq 0$) where $\kappa_1 + \kappa_2 + \cdots = K$. We denote the sum $\kappa_1 + \kappa_2 + \cdots$ as $\sum \kappa$. The *union* $\tau = (\tau_1, \tau_2, \dots)$ of two integer partitions κ and ν is defined as $\tau = \kappa \cup \nu$ where $\tau_1 \geq \tau_2 \geq \cdots$. The *difference* τ of two integer partitions κ and ν is defined as $\tau = \kappa \setminus \nu$, where τ includes the elements from κ except elements existing in both κ and ν, which are removed. Furthermore, the *conjugate partition* of κ is defined as $\nu = \mathrm{conj}(\kappa)$, where ν_i is equal to the number of integers in κ that is equal or greater than i, for $i = 1, 2, \dots$.

If ν is an integer partition, not necessarily of the same integer K as κ, and $\kappa_1 + \cdots + \kappa_i \geq \nu_1 + \cdots + \nu_i$ for $i = 1, 2, \dots$, then $\kappa \geq \nu$. When $\kappa \geq \nu$ and $\kappa \neq \nu$ then $\kappa > \nu$. If κ, ν and τ are integer partitions of the same integer K and there does not exist any τ such that $\kappa > \tau > \nu$ where $\kappa > \nu$, then κ *covers* ν. It follows that κ covers ν if and only if $\kappa > \nu$ and $\mathrm{conj}(\kappa) < \mathrm{conj}(\nu)$. A weaker definition of cover is *adjacent* [11, 35], where κ and ν can be partitions of different integers. We say that $\kappa > \nu$ are adjacent partitions if either κ covers ν or if $\kappa = \nu \cup (1)$.

An integer partition $\kappa = (\kappa_1, \dots, \kappa_n)$ can also be represented by n piles of coins, where the first pile has κ_1 coins, the second κ_2 coins and so on. An integer partition κ *covers* ν if ν can be obtained from κ by moving one coin *one* column rightward or *one* row downward, and keep κ monotonically decreasing. Or, equivalently, an integer partition κ *is covered by* τ if τ can be obtained from κ by moving one coin *one* column leftward or *one* row upward, and keep κ monotonically decreasing. These two types of coin moves are defined in [18] and called *minimum rightward* and *minimum leftward coin moves*, respectively (see Figure 2.1).

3. Canonical forms and invariants. In the following, we introduce the Kronecker canonical form (KCF) of a general matrix pencil and its invariants in terms of integer sequences, as well as the Brunovsky canonical form for various system pencils.

3.1. Kronecker canonical form. Any general $m_{\mathrm{p}} \times n_{\mathrm{p}}$ matrix pencil $G - \lambda H$ can be transformed into *Kronecker canonical form* (KCF) in terms of an equivalence transformation with two nonsingular matrices U and V [26]:

$$\begin{aligned} & U(G-\lambda H)V^{-1} \\ & \quad = \operatorname{diag}\left(L_{\epsilon_1},\dots,L_{\epsilon_{r_0}},J(\mu_1),\dots,J(\mu_q),N_{s_1},\dots,N_{s_{g_\infty}},L^T_{\eta_1},\dots,L^T_{\eta_{l_0}}\right), \end{aligned} \tag{3.1}$$

where $J(\mu_i)=\operatorname{diag}(J_{h_1}(\mu_i),\dots,J_{h_{g_i}}(\mu_i))$, $i=1,\dots,q$. The blocks $J_{h_k}(\mu_i)$ are $h_k\times h_k$ *Jordan blocks* associated with each distinct finite eigenvalue μ_i and the blocks N_{s_k} are $s_k\times s_k$ Jordan blocks for matrix pencils associated with the infinite eigenvalue. These two types of blocks constitute the *regular part* of a matrix pencil and are defined by

$$J_{h_k}(\mu_i)=\begin{bmatrix} \mu_i-\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \mu_i-\lambda & 1 \\ & & & \mu_i-\lambda \end{bmatrix}, \quad\text{and}\quad N_{s_k}=\begin{bmatrix} 1 & -\lambda & & \\ & \ddots & \ddots & \\ & & 1 & -\lambda \\ & & & 1 \end{bmatrix}.$$

If $m_{\mathrm{p}}\neq n_{\mathrm{p}}$ or $\det(G-\lambda H)\equiv 0$ for all $\lambda\in\mathbb{C}$, then $r_0\geq 1$ and/or $l_0\geq 1$ and the matrix pencil also includes a *singular part* which consists of the r_0 *right singular blocks* L_{ϵ_k} of size $\epsilon_k\times(\epsilon_k+1)$ and the l_0 *left singular blocks* $L^T_{\eta_k}$ of size $(\eta_k+1)\times\eta_k$:

$$L_{\epsilon_k}=\begin{bmatrix} -\lambda & 1 & & \\ & \ddots & \ddots & \\ & & -\lambda & 1 \end{bmatrix}, \quad\text{and}\quad L^T_{\eta_k}=\begin{bmatrix} -\lambda & & \\ 1 & \ddots & \\ & \ddots & -\lambda \\ & & 1 \end{bmatrix}.$$

L_0 and L_0^T blocks are of size 0×1 and 1×0, respectively, and each of them contributes with a column or row of zeros.

In general, a block diagonal matrix $A=\operatorname{diag}(A_1,A_2,\dots,A_b)$ with b blocks can also be represented as a direct sum

$$A\equiv A_1\oplus A_2\oplus\cdots\oplus A_b\equiv\bigoplus_{k=1}^{b}A_k.$$

Using this notation, the KCF (3.1) can compactly be rewritten as

$$U(G-\lambda H)V^{-1}\equiv\mathbb{L}\oplus\mathbb{L}^T\oplus\mathbb{J}(\mu_1)\oplus\cdots\oplus\mathbb{J}(\mu_q)\oplus\mathbb{N},$$

where

$$\mathbb{L}=\bigoplus_{k=1}^{r_0}L_{\epsilon_k},\quad \mathbb{L}^T=\bigoplus_{k=1}^{l_0}L^T_{\eta_k},\quad \mathbb{J}(\mu_i)=\bigoplus_{k=1}^{g_i}J_{h_k}(\mu_i),\quad\text{and}\quad \mathbb{N}=\bigoplus_{k=1}^{g_\infty}N_{s_k}.$$

Without loss of generality, we order the blocks of the KCF in the direct sum notation so that the singular blocks ($\mathbb{L}$ and $\mathbb{L}^T$) appear first.

3.2. Invariants of matrix pencils. The matrix pencil characteristics can equivalently be expressed in terms of column/row minimal indices and finite/infinite elementary divisors. Two matrix pencils are strictly equivalent if and only if they have the same minimal indices and elementary divisors or, equivalently, if they have the same KCF, i.e., the same L, L^T, J, and N blocks.

The four invariants are defined as follows [26]:

(i) The *column* (*right*) *minimal indices* are $\epsilon=(\epsilon_1,\dots,\epsilon_{r_0})$, where $\epsilon_1\geq\epsilon_2\geq\cdots\geq\epsilon_{r_1}>\epsilon_{r_1+1}=\cdots=\epsilon_{r_0}=0$ define the sizes of the L_{ϵ_k} blocks, $\epsilon_k\times(\epsilon_k+1)$.

From the conjugate partition $(r_1, \dots, r_{\epsilon_1}, 0, \dots)$ of ϵ we define the integer partition $\mathcal{R}(G - \lambda H) = (r_0) \cup (r_1, \dots, r_{\epsilon_1})$.

(ii) The *row* (*left*) *minimal indices* are $\eta = (\eta_1, \dots, \eta_{l_0})$, where $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_{l_1} > \eta_{l_1+1} = \cdots = \eta_{l_0} = 0$ define the sizes of the $L^T_{\eta_k}$ blocks, $(\eta_k + 1) \times \eta_k$. From the conjugate partition $(l_1, \dots, l_{\eta_1}, 0, \dots)$ of η we define the integer partition $\mathcal{L}(G - \lambda H) = (l_0) \cup (l_1, \dots, l_{\eta_1})$.

(iii) The *finite elementary divisors* are of the form $(\lambda - \mu_i)^{h_1^{(i)}}, \dots, (\lambda - \mu_i)^{h_{g_i}^{(i)}}$, with $h_1^{(i)} \geq \cdots \geq h_{g_i}^{(i)} \geq 1$ for each of the q distinct finite eigenvalue μ_i, $i = 1, \dots, q$. Here, g_i is the geometric multiplicity of μ_i and the sum of all $h_k^{(i)}$ for $k = 1, \dots, g_i$ is the algebraic multiplicity of μ_i. For each distinct eigenvalue μ_i, we introduce the integer partition $h_{\mu_i} = (h_1^{(i)}, \dots, h_{g_i}^{(i)})$, which is known as the *Segre characteristics*. These characteristics correspond to the sizes $h_k^{(i)} \times h_k^{(i)}$ of the $J_{h_k}(\mu_i)$ blocks (the largest first). The conjugate partition $\mathcal{J}_{\mu_i}(G - \lambda H) = (j_1, j_2, \dots)$ of h_{μ_i} is the *Weyr characteristics* of μ_i.

(iv) The *infinite elementary divisors* are of the form $\rho^{s_1}, \rho^{s_2}, \dots, \rho^{s_{g_\infty}}$, with $s_1 \geq \cdots \geq s_{g_\infty} \geq 1$, where g_∞ is the geometric multiplicity of the infinite eigenvalue and the sum of all s_k for $k = 1, \dots, g_\infty$ is the algebraic multiplicity. Similarly to case (iii), the integer partition $s = (s_1, \dots, s_{g_\infty})$ is the *Segre characteristics* for the infinite eigenvalue, which correspond to the sizes $s_k \times s_k$ of the N_{s_k} blocks. The conjugate partition $\mathcal{N}(G - \lambda H) = (n_1, n_2, \dots)$ of s is the *Weyr characteristics* of the infinite eigenvalue.

When it is clear from context, we use the abbreviated notation $\mathcal{R}$, $\mathcal{L}$, $\mathcal{J}$, and $\mathcal{N}$, for the above defined integer partitions corresponding to the right and left singular structures, and the Jordan structures of the finite and infinite eigenvalues, respectively. In the following, these integer partitions are referred to as *structure integer partitions*.

The system pencils $\mathbf{S}(\lambda)$, $\mathbf{S}_\mathrm{C}(\lambda)$, and $\mathbf{S}_\mathrm{O}(\lambda)$ can also be expressed in terms of the above invariants and their associated structure integer partitions. However, in general their corresponding invariants are different. For example, the system pencil $\mathbf{S}_\mathrm{C}(\lambda)$ of a completely controllable system associated with the pair (A, B) can only have L blocks in its KCF while $\mathbf{S}(\lambda)$ (2.2) may have both types of singular invariants (blocks) as well as eigenvalues in its KCF.

3.3. Brunovsky canonical form. When considering canonical forms of the system pencils $\mathbf{S}_\mathrm{C}(\lambda)$ and $\mathbf{S}_\mathrm{O}(\lambda)$ associated with pairs of matrices, we are (mainly) interested in canonical forms obtained from structure-preserving equivalence transformations. One such example is the Brunovsky canonical form. This canonical form explicitly reveals the system characteristics from the system pencils. This is in contrast to the KCF, which destroys the special block structure of $\mathbf{S}_\mathrm{C}(\lambda)$ and $\mathbf{S}_\mathrm{O}(\lambda)$, respectively, and only implicitly gives the system characteristics. Canonical and condensed forms for generalized matrix pairs appearing in descriptor systems [5, 43] are out of the scope of this paper.

Given a controllability pair (A, B) there exists a *feedback equivalent* (also known as Γ-equivalent or block similar) matrix pair (A_B, B_B) in *Brunovsky canonical form* (BCF) [4, 28, 31], such that

$$P \begin{bmatrix} A - \lambda I_n & B \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ R & Q^{-1} \end{bmatrix} = \left[\begin{array}{cc:c} A_\epsilon & 0 & B_\epsilon \\ 0 & A_\mu & 0 \end{array} \right], \tag{3.2}$$

where $A_\epsilon = \operatorname{diag}(J_{\epsilon_1}(0), \dots, J_{\epsilon_{r_1}}(0))$, $A_\mu = \operatorname{diag}(J(\mu_1), \dots, J(\mu_q))$, and $B_\epsilon = \operatorname{diag}(e_{\epsilon_1}, \dots, e_{\epsilon_{r_0}})$. The transformation matrices $P \in \mathbb{C}^{n \times n}$ and $Q \in \mathbb{C}^{m \times m}$ are nonsingular and $R \in \mathbb{C}^{m \times n}$. Each block $J(\mu_i)$ in A_μ is block diagonal with the Jordan blocks for the specified finite eigenvalue μ_i. $J_{\epsilon_i}(0)$ is a nilpotent matrix in its reduced Jordan form and $e_i = [0, \dots, 0, 1]^T \in \mathbb{C}^{i \times 1}$. Moreover, the matrix pair (A_ϵ, B_ϵ) is controllable and corresponds to the L blocks in the KCF of $\mathbf{S}_\mathrm{C}(\lambda)$. If $\mathrm{rank}(\mathbf{S}_\mathrm{C}(\lambda)) < n$ for some $\lambda \in \mathbb{C}$, then (A, B) is uncontrollable and there exists a regular pencil A_μ whose eigenvalues correspond to the *uncontrollable eigenvalues* (*modes*).

The dual form of BCF for the observability pair (A, C) is

$$\begin{bmatrix} P & S \\ 0 & T \end{bmatrix} \begin{bmatrix} A - \lambda I_n \\ C \end{bmatrix} P^{-1} = \begin{bmatrix} A_B - \lambda I_n \\ C_B \end{bmatrix} = \left[\begin{array}{cc} A_\eta & 0 \\ 0 & A_\mu \\ \hdashline C_\eta & 0 \end{array}\right], \tag{3.3}$$

where $A_\eta = \mathrm{diag}(J_{\eta_1}(0), \dots, J_{\eta_{l_1}}(0))$, $A_\mu = \mathrm{diag}(J(\mu_1), \dots, J(\mu_q))$, and $C_\eta = \mathrm{diag}(e^T_{\eta_1}, \dots, e^T_{\eta_{l_0}})$. The transformation matrices $P \in \mathbb{C}^{n \times n}$ and $T \in \mathbb{C}^{p \times p}$ are nonsingular and $S \in \mathbb{C}^{n \times p}$. The matrix pair (A_η, C_η) is observable and corresponds to the L^T blocks. If $\mathrm{rank}(\mathbf{S}_\mathrm{O}(\lambda)) < n$ for some $\lambda \in \mathbb{C}$, then (A, C) is unobservable and there exists a regular pencil A_μ whose eigenvalues correspond to the *unobservable eigenvalues* (*modes*).

Some of the system characteristics that the BCF directly reveals are as follows: (A, B) has exactly m L blocks, one for each column in B_ϵ, and $m - \mathrm{rank}(B_B)$ L_0 blocks. Likewise, (A, C) has exactly p L^T blocks, one for each row in C_η, and $p - \mathrm{rank}(C_B)$ L_0^T blocks. Since $\epsilon_{r_1+1} = \cdots = \epsilon_{r_0} = 0$, the column vectors $e_{\epsilon_{r_1+1}}, \dots, e_{\epsilon_{r_0}}$ are 0×1 and correspond to the L_0 blocks; $\mathrm{rank}(B) = m - \#(L_0 \text{ blocks})$. For each L_0 block one input signal $u_k(t)$ can be removed without losing controllability of (A_ϵ, B_ϵ). Likewise, the row vectors $e^T_{\eta_{l_1+1}}, \dots, e^T_{\eta_{l_0}}$ are 1×0 and correspond to the L_0^T blocks, where for each L_0^T block one output signal $y_k(t)$ can be removed without losing observability of (A_η, C_η).

4. The system pencil space. An $n \times (n + m)$ controllability pair (A, B) has $n^2 + nm$ free elements and, therefore, belongs to an $(n^2 + nm)$-dimensional (*system pencil*) *space*, one dimension for each parameter. A controllability pair (A, B) can be seen as a point in the $(n^2 + nm)$-dimensional space, and the union of equivalent matrix pairs as a manifold in this space [17, 18]. Similarly, the $(n + p) \times n$ observability pair (A, C) is a point in an $(n^2 + np)$-dimensional system pencil space. We say that the matrix pair "lives" in the space spanned by the manifold, and the dimension of the manifold is given from the number of parameters of the matrix pair, where each fixed parameter gives one less degree of freedom. The dimension of the complementary space to the manifold is called the *codimension*.

The *orbit* of a matrix pair, $\mathcal{O}(A, B)$ or $\mathcal{O}(A, C)$, is a manifold of all equivalent matrix pairs, i.e., manifolds in the $(n^2 + nm)$-dimensional and $(n^2 + np)$-dimensional spaces, respectively. In the following, when something holds for both (A, B) and (A, C) we denote the matrix pairs with $(*)$, e.g., $\mathcal{O}(*)$. Throughout this paper, we consider only orbits under feedback equivalence [4, 31], which for the controllability pairs is defined as

$$\mathcal{O}(A, B) = \left\{ P \begin{bmatrix} A - \lambda I & B \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ R & Q^{-1} \end{bmatrix} : \det(P) \cdot \det(Q) \neq 0 \right\},$$

and for observability pairs as

$$\mathcal{O}(A, C) = \left\{ \begin{bmatrix} P & S \\ 0 & T \end{bmatrix} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} P^{-1} : \det(P) \cdot \det(T) \neq 0 \right\}.$$

In other words, all matrix pairs in the same orbit have the same canonical form, with the eigenvalues and the sizes of the Jordan blocks fixed. A *bundle* defines the union of all orbits with the same canonical form but with the eigenvalues unspecified, $\bigcup_{\mu_i} \mathcal{O}(*)$ [1]. We denote the bundle of a matrix pair by $\mathcal{B}(*)$.

The dimension of the space $\mathcal{O}(A,B)$ is equal to the dimension of the *tangent space* to $\mathcal{O}(A,B)$ at (A,B), denoted by $\tan(A,B)$. Similar definitions hold for the matrix pair (A,C). The tangent spaces $\tan(A,B)$ and $\tan(A,C)$ can be represented in matrix form as

$$\begin{bmatrix} T_A & T_B \end{bmatrix} = X \begin{bmatrix} A & B \end{bmatrix} + \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} -X & 0 \\ V & W \end{bmatrix},$$

and

$$\begin{bmatrix} T_A \\ T_C \end{bmatrix} = \begin{bmatrix} X & Y \\ 0 & Z \end{bmatrix} \begin{bmatrix} A \\ C \end{bmatrix} + \begin{bmatrix} A \\ C \end{bmatrix} \begin{bmatrix} -X \end{bmatrix},$$

respectively, where X, Y, Z, V, and W are matrices of conforming sizes [7].

Using the technique in [17], the tangent vectors $\begin{bmatrix} T_A & T_B \end{bmatrix}$ can be expressed in terms of the vec-operator and Kronecker products (see also [7]):

$$\begin{bmatrix} \mathrm{vec}(T_A) \\ \mathrm{vec}(T_B) \end{bmatrix} = T_{(A,B)} \begin{bmatrix} \mathrm{vec}(X) \\ \mathrm{vec}(V) \\ \mathrm{vec}(W) \end{bmatrix},$$

where $\tan(A,B)$ is the range of the $(n^2+nm) \times (n^2+nm+m^2)$ matrix

$$T_{(A,B)} = \begin{bmatrix} A^T \otimes I_n - I_n \otimes A & I_n \otimes B & 0 \\ B^T \otimes I_n & 0 & I_m \otimes B \end{bmatrix}. \tag{4.1}$$

Similarly, $\tan(A,C)$ is the range of the $(n^2+np) \times (n^2+np+p^2)$ matrix

$$T_{(A,C)} = \begin{bmatrix} A^T \otimes I_n - I_n \otimes A & C^T \otimes I_n & 0 \\ -I_n \otimes C & 0 & C^T \otimes I_p \end{bmatrix}, \text{ where} \tag{4.2}$$

$$\begin{bmatrix} \mathrm{vec}(T_A) \\ \mathrm{vec}(T_C) \end{bmatrix} = T_{(A,C)} \begin{bmatrix} \mathrm{vec}(X) \\ \mathrm{vec}(Y) \\ \mathrm{vec}(Z) \end{bmatrix}.$$

The orthogonal complement of the tangent space is the *normal space*, $\mathrm{nor}(*)$. The dimension of the normal space is called the *codimension* of $\mathcal{O}(*)$ [12, 52], denoted by $\mathrm{cod}(*)$. Together, the tangent and the normal spaces span the complete (n^2+nm)-dimensional space for (A,B) and the complete (n^2+np)-dimensional space for (A,C).

Knowing the canonical structure, the explicit expression for the codimension of the controllability pair (A,B) is derived in [24]; see also [25]. By rewriting the result, it is obvious that the computation of the codimension of (A,B) can be done using parts of the expression for matrix pencils [12]. The codimension of the observability pair (A,C) is easily derived by its duality to (A,B). In summary, the codimension of the orbit of a controllability pair (A,B), with the column minimal indices $\epsilon_1, \dots, \epsilon_{r_0}$ and the finite elementary divisors $h_1^{(i)}, \dots, h_{g_i}^{(i)}$ for each distinct eigenvalue μ_i, is

$$\mathrm{cod}(A,B) = c_{\mathrm{Right}} + c_{\mathrm{Jor}} + c_{\mathrm{Jor,Right}}, \tag{4.3}$$

where

$$c_{\text{Right}} = \sum_{\epsilon_k > \epsilon_l} (\epsilon_k - \epsilon_l - 1), \; c_{\text{Jor}} = \sum_{i=1}^{q} \sum_{k=1}^{g_i} (2k-1) h_k^{(i)}, \; \text{and } c_{\text{Jor,Right}} = r_0 \sum_{i=1}^{q} \sum_{k=1}^{g_i} h_k^{(i)}.$$

The codimension of the orbit of a observability pair (A, C), with the row minimal indices $\eta_1, \ldots, \eta_{l_0}$ and the finite elementary divisors $h_1^{(i)}, \ldots, h_{g_i}^{(i)}$ for each distinct eigenvalue μ_i, is

$$\text{cod}(A, C) = c_{\text{Left}} + c_{\text{Jor}} + c_{\text{Jor,Left}}, \tag{4.4}$$

where

$$c_{\text{Left}} = \sum_{\eta_k > \eta_l} (\eta_k - \eta_l - 1), \; c_{\text{Jor}} = \sum_{i=1}^{q} \sum_{k=1}^{g_i} (2k-1) h_k^{(i)}, \; \text{and } c_{\text{Jor,Left}} = l_0 \sum_{i=1}^{q} \sum_{k=1}^{g_i} h_k^{(i)}.$$

The value of the eigenvalues make no contribution to the codimension in the bundle case. Therefore, knowing the codimension of an orbit, the codimension of the corresponding bundle is one less for each distinct eigenvalue: $\text{cod}(\mathcal{B}(*)) = \text{cod}(\mathcal{O}(*)) -$ (number of distinct eigenvalues). For example, if we are interested in a matrix pair (A, B) with k unspecified eigenvalues and the rest with known specified values, the codimension of $\mathcal{B}(A, B)$ is $\text{cod}(\mathcal{O}(A, B)) - k$.

5. Stratification of orbits and bundles.

In this section, we present the stratification of orbits and bundles of matrix pairs (A, B) and (A, C). The most and least generic cases are considered in section 5.1, and in section 5.2 the coin rules representing the closure and cover relations are derived.

A stratification is a closure hierarchy of orbits (or bundles). Following [18, 23], we represent the stratification by a connected graph where the nodes correspond to orbits (or bundles) of canonical structures and the edges to their covering relations; see Figures 1.1 and 6.2. The graph is organized from top to bottom with nodes in increasing order of codimension.

Given a node representing an orbit (or bundle) of a canonical structure, the *closure* of that orbit (or bundle) includes the orbit (or bundle) itself and all orbits (or bundles) represented by the nodes which can be reached by a *downward path.* A downward path is defined as a path for which all edges start in a node and end in another node below in the graph. An *upward path* is a path in the opposite direction. In the following, when it is clear from context we use the shorter term structure when we refer to a canonical structure.

Given a matrix pair and its corresponding node in the graph, it is always possible to make the pair more generic by a small perturbation, i.e., change the pair to one corresponding to a node along an upward path from the node. It is normally not possible to make a corresponding downward move by a small perturbation, i.e., a structure is not, in general, near any of the more degenerate structures below in the graph. However, the cases when a structure below in the hierarchy actually is nearby are often of particular interest, as it shows that a more degenerate structure can be found by a small perturbation.

5.1. Most and least generic cases.

Almost all matrix pairs of the same size and type (controllability or observability pairs) have the same canonical structure.

This canonical structure corresponds to the most generic case and has the lowest codimension in the closure hierarchy. The opposite case is the least generic case, or equivalently, the most degenerate case with the highest codimension. In the closure hierarchy graph, the most generic case is represented by the topmost node and the most degenerate case by the bottom node. The canonical structures in between correspond to degenerate (or nongeneric) cases, which from a computational point of view can be a real challenge [14, 15].

The most generic structure of the controllability pair (A, B) has $\mathcal{R} = (r_0, \ldots, r_\alpha, r_{\alpha+1})$ where $r_0 = \cdots = r_\alpha = m$, $r_{\alpha+1} = n \mod m$, and $\alpha = \lfloor n/m \rfloor$ [29, 53]. For the observability pair (A, C) the most generic structure has $\mathcal{L} = (l_0, \ldots, l_\alpha, l_{\alpha+1})$ where $l_0 = \cdots = l_\alpha = p$, $l_{\alpha+1} = n \mod p$, and $\alpha = \lfloor n/p \rfloor$. The most degenerate controllability pair has m L_0 blocks and n Jordan blocks of size 1×1 corresponding to an eigenvalue of multiplicity n. Similarly, the most degenerate observability pair has p L_0^T blocks and n 1×1 Jordan blocks. In other words, the most generic cases of the matrix pairs correspond to completely controllable and observable systems, while the most degenerate cases correspond to systems with n uncontrollable and n unobservable multiple modes, respectively.

We remark that the above formulae to compute the most generic structure only hold if there are no restrictions on the matrix pair. Otherwise, for example, when the matrix pair has a special structure or fixed rank, the restrictions must be considered when determining the most and least generic cases. There can even exist several most generic structures, but only one with codimension 0 (if it exists). This has recently been studied for general matrix pencils in, e.g., [9, 10, 37].

5.2. Closure and cover relations. To determine the closure hierarchy for $n \times (n + m)$ controllability pairs we stratify the $(n^2 + nm)$-dimensional system pencil space into feedback equivalent orbits (or bundles). Similarly, the closure hierarchy for $(n + p) \times n$ observability pairs is determined by the stratification of feedback equivalent orbits (or bundles) in the $(n^2 + np)$-dimensional system pencil space. The stratification of orbits or bundles is given from the closure relations and further the cover relations between these manifolds; see Arnold [1] and [17, 18]. An orbit *covers* another orbit if its closure includes the closure of the other orbit and there is no orbit in between in the closure hierarchy; i.e., they are nearest neighbors in the hierarchy. The closure and cover relations for bundles are defined analogously.

Before we give the closure and cover relations for matrix pairs, we review some results for matrices and general matrix pencils.

From the closure condition for nilpotent matrices derived in [1, 18] and the definition of covering partitions, the cover relations for orbits of nilpotent matrices are obtained [18]. The orbit of a matrix is the manifold of all similar matrices: $\mathcal{O}(A) = \{PAP^{-1} : \det(A) \neq 0\}$. If the matrix A has well-clustered eigenvalues but is not nilpotent, we order the Jordan blocks such that $A = \operatorname{diag}(A_1, \ldots, A_q)$, where A_i contains all Jordan blocks associated with the eigenvalue μ_i. Then for each matrix A_i, we consider $\widetilde{A}_i = A_i - \mu_i I$ which is nilpotent, and the closure and cover relations for nilpotent matrices are applicable. It follows that the number of eigenvalues and the total size of all blocks associated with the same eigenvalue are the same for all orbits in the closure hierarchy. This is in contrast to the bundle case where eigenvalues can coalesce or split apart.

THEOREM 5.1 ([1, 18]). *$\mathcal{O}(A_1)$ covers $\mathcal{O}(A_2)$* if and only if *some $\mathcal{J}_{\mu_i}(A_2)$ can be obtained from $\mathcal{J}_{\mu_i}(A_1)$ by a minimum leftward coin move, and $\mathcal{J}_{\mu_j}(A_2) = \mathcal{J}_{\mu_j}(A_1)$ for all $\mu_j \neq \mu_i$.*

In the case of not well-clustered eigenvalues, we have to consider the bundle case as defined by Arnold [1]. Even if testing for closure relations between nilpotent matrices is trivial, deciding if one bundle is in the closure of another bundle is an NP-complete problem [18, 32]. The solution to the closure decision problem for matrix bundles is given in [16, 18, 45], and the cover relations expressed in terms of coin moves in [18].

The necessary conditions for an orbit or a bundle of two matrix pencils to be closest neighbors in a closure hierarchy were derived in [3, 8, 50], where the orbit is the manifold of strictly equivalent matrix pencils: $\mathcal{O}(G-\lambda H) = \{U(G-\lambda H)V^{-1} : \det(U)\cdot\det(V) \neq 0\}$. These conditions were later complemented with the corresponding sufficient conditions in [18]. Notice that in the following theorem, for the structure integer partition $\mathcal{J}_{\mu_i}$ the eigenvalue μ_i belongs to the extended complex plane $\overline{\mathbb{C}}$, i.e., $\mu_i \in \mathbb{C} \cup \{\infty\}$. Furthermore, the restrictions on r_0 and l_0 in rules 1 and 2 correspond to the fact that the number of L_k and L_k^T blocks cannot change.

THEOREM 5.2 ([18]). *Given the structure integer partitions $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{J}_{\mu_i}$ of $G-\lambda H$, where $\mu_i \in \overline{\mathbb{C}}$, one of the following* if-and-only-if *rules finds $\widetilde{G}-\lambda\widetilde{H}$ such that $\mathcal{O}(G-\lambda H)$ covers $\mathcal{O}(\widetilde{G}-\lambda\widetilde{H})$:*

(1) *Minimum* rightward *coin move in $\mathcal{R}$ (or $\mathcal{L}$).*

(2) *If the rightmost column in $\mathcal{R}$ (or $\mathcal{L}$) is one single coin, move that coin to a new rightmost column of some $\mathcal{J}_{\mu_i}$ (which may be empty initially).*

(3) *Minimum* leftward *coin move in any $\mathcal{J}_{\mu_i}$.*

(4) *Let k denote the total number of coins in all of the longest (= lowest) rows from all of the $\mathcal{J}_{\mu_i}$. Remove these k coins, add one more coin to the set, and distribute $k+1$ coins to r_p, $p = 0,\ldots,t$ and l_q, $q = 0,\ldots,k-t-1$ such that at least all nonzero columns of $\mathcal{R}$ and $\mathcal{L}$ are given coins.*

Rules 1 *and* 2 *are not allowed to make coin moves that affect r_0 (or l_0).*

Necessary and sufficient conditions for closure relations between orbits of matrix pairs (A,B) have been studied in [31], and later in [35, 36]. These are a subset of those for general matrix pencils. Here we give our reformulation and slight modification of the theorem originally presented in [36, Theorem 4.6] for orbits and the corresponding theorem for bundles, where $\overline{\mathcal{O}}$ denotes the orbit closure and $\overline{\mathcal{B}}$ is the bundle closure.

THEOREM 5.3 ([36, 42]). $\overline{\mathcal{O}}(A,B) \supseteq \overline{\mathcal{O}}(\widetilde{A},\widetilde{B})$ if and only if *the following conditions hold:*

(1) $\mathcal{R}(A,B) \geq \mathcal{R}(\widetilde{A},\widetilde{B})$.

(2) $\mathcal{J}_{\mu_i}(A,B) \leq \mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B})$*, for all* $\mu_i \in \mathbb{C}$, $i = 1,\ldots,q$.

THEOREM 5.4. *If $\mathcal{B}(A,B)$ has at least as many distinct eigenvalues as $\mathcal{B}(\widetilde{A},\widetilde{B})$, then $\overline{\mathcal{B}}(A,B) \supseteq \overline{\mathcal{B}}(\widetilde{A},\widetilde{B})$* if and only if *the following conditions hold:*

(1) $\mathcal{R}(A,B) \geq \mathcal{R}(\widetilde{A},\widetilde{B})$.

(2) *It is possible to coalesce eigenvalues and apply the dominance ordering coin moves to $\mathcal{J}_{\mu_i}(A,B)$, for any μ_i, to reach $(\widetilde{A},\widetilde{B})$.*

Proof. The theorem follows directly from Theorem 5.3 and the closure condition for matrix bundles presented in [18]. □

The conditions for closure relations between two observability matrix pairs (A,C) are, from the duality with (A,B), equal to those for (A,B) except that $\mathcal{R}$ is replaced by $\mathcal{L}$.

In [35], also the necessary conditions for cover relations of matrix pencils with no row minimal indices have been derived. A matrix pencil $G-\lambda H$ with no row minimal indices differs from a controllability pair (A,B) in that it can have infinite elementary divisors, which is not the case for standard matrix pairs. The cover relations [35, Proposition 5.2] are summarized in Proposition 5.5 with some minor reformulations, where the invariants of $G - \lambda H$ and $\widetilde{G} - \lambda \widetilde{H}$ are

$$\epsilon = (\epsilon_1, \ldots, \epsilon_{r_0}),\ h_{\mu_i} = \left(h_1^{(i)}, \ldots, h_{g_i}^{(i)}\right),\ s = (s_1, \ldots, s_{g_\infty}), \quad \text{and}$$
$$\widetilde{\epsilon} = (\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_{\widetilde{r}_0}),\ \widetilde{h}_{\mu_j} = \left(\widetilde{h}_1^{(j)}, \ldots, \widetilde{h}_{\widetilde{g}_j}^{(j)}\right),\ \widetilde{s} = (\widetilde{s}_1, \ldots, \widetilde{s}_{\widetilde{g}_\infty}),$$

respectively. Remark, the integer partitions associated with the same invariants of $G - \lambda H$ and $\widetilde{G} - \lambda \widetilde{H}$, e.g., ϵ and $\widetilde{\epsilon}$, can be of different length.

Proposition 5.5 ([35]). *Let $G - \lambda H$ and $\widetilde{G} - \lambda \widetilde{H}$ be two $n \times (n+m)$ matrix pencils with no row minimal indices. If $\mathcal{O}(G - \lambda H)$ covers $\mathcal{O}(\widetilde{G} - \lambda \widetilde{H})$, then one of the following conditions holds:*

(1) $\operatorname{conj}(\epsilon) > \operatorname{conj}(\widetilde{\epsilon})$ *are adjacent,* $h_{\mu_i} = \widetilde{h}_{\mu_i}$ *for all eigenvalues* μ_i*, and* $s = \widetilde{s}$.

(2) $\sum_{i=1}^m \epsilon_i > \sum_{i=1}^m \widetilde{\epsilon}_i$, $\operatorname{conj}(\epsilon) > \operatorname{conj}(\widetilde{\epsilon})$ *are adjacent,* $\widetilde{h}_1^{(i)} = h_1^{(i)} + 1$ *for some eigenvalue* μ_i *(where* μ_i *can be a new eigenvalue), and* $s = \widetilde{s}$.

(3) $\sum_{i=1}^m \epsilon_i > \sum_{i=1}^m \widetilde{\epsilon}_i$, $\operatorname{conj}(\epsilon) > \operatorname{conj}(\widetilde{\epsilon})$ *are adjacent,* $h_{\mu_i} = \widetilde{h}_{\mu_i}$ *for all eigenvalues* μ_i*, and* $\widetilde{s}_1 = s_1 + 1$ *(where* s *and* $\widetilde{s}$ *can be empty partitions).*

(4) $\epsilon = \widetilde{\epsilon}$*,* $h_{\mu_i} > \widetilde{h}_{\mu_i}$ *for all eigenvalues* μ_i*, and* $s = \widetilde{s}$.

(5) $\epsilon = \widetilde{\epsilon}$*,* $h_{\mu_i} = \widetilde{h}_{\mu_i}$ *for all eigenvalues* μ_i*, and* $s > \widetilde{s}$.

From Theorem 5.3, Proposition 5.5, and the cover conditions for matrix pencils in Theorem 5.2, it is possible to derive both necessary and sufficient conditions for a covering relation between two controllability pairs (A, B). The result is given in Theorem 5.6, where $r_0(A, B)$ denotes the number of column minimal indices for (A, B). The proof is organized as follows. We modify Proposition 5.5 so that it fulfills the restrictions given by the structure of the controllability pair and then, where required, strengthen each condition so that they become not only necessary but also sufficient.

Theorem 5.6. *$\mathcal{O}(A, B)$ covers $\mathcal{O}(\widetilde{A}, \widetilde{B})$* if and only if *one of the following conditions holds:*

(1) $\mathcal{R}(A, B)$ *covers* $\mathcal{R}(\widetilde{A}, \widetilde{B})$ *where* $r_0(A, B) = r_0(\widetilde{A}, \widetilde{B})$*, and* $\mathcal{J}_{\mu_i}(A, B) = \mathcal{J}_{\mu_i}(\widetilde{A}, \widetilde{B})$ *for all eigenvalues* μ_i.

(2) *If* $r_{\epsilon_1} = 1$ *and* $\epsilon_1 \geq 1$ *for* $\mathcal{R}(A, B)$*, then* $\mathcal{R}(\widetilde{A}, \widetilde{B}) = \mathcal{R}(A, B) \setminus (r_{\epsilon_1})$*,* $\mathcal{J}_{\mu_i}(\widetilde{A}, \widetilde{B}) = \mathcal{J}_{\mu_i}(A, B) \cup (1)$ *for some eigenvalue* μ_i *(where* $\mathcal{J}_{\mu_i}(A, B)$ *can be an empty partition), and* $\mathcal{J}_{\mu_j}(A, B) = \mathcal{J}_{\mu_j}(\widetilde{A}, \widetilde{B})$ *for all* $\mu_j \neq \mu_i$.

(3) $\mathcal{R}(A, B) = \mathcal{R}(\widetilde{A}, \widetilde{B})$*,* $\mathcal{J}_{\mu_i}(A, B)$ *covers* $\mathcal{J}_{\mu_i}(\widetilde{A}, \widetilde{B})$ *for one eigenvalue* μ_i*, and* $\mathcal{J}_{\mu_j}(A, B) = \mathcal{J}_{\mu_j}(\widetilde{A}, \widetilde{B})$ *for all* $\mu_j \neq \mu_i$.

Proof. Let 5.5(n) denote condition n of Proposition 5.5, and similarly, 5.6(m) denotes condition m of Theorem 5.6.

A matrix pencil $G - \lambda H$ with no row minimal indices can have infinite elementary divisors which a controllability pair (A, B) cannot have. This restriction is introduced by only considering finite elementary divisors, which obviously exclude 5.5(3) and 5.5(5) (where $G - \lambda H$ and/or $\widetilde{G} - \lambda \widetilde{H}$ have infinite elementary divisors). The remaining three conditions are now considered, and we begin each proof by rewriting the conditions in the structure integer notation: $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{J}$.

First we consider 5.5(1) which can be rewritten as:

$$\mathcal{R}(A, B) > \mathcal{R}(\widetilde{A}, \widetilde{B}) \text{ are adjacent and } \mathcal{J}_{\mu_i}(A, B) = \mathcal{J}_{\mu_i}(\widetilde{A}, \widetilde{B}).$$

Since the two matrix pairs have the same Jordan structure, the size of the right singular parts of (A, B) and $(\widetilde{A}, \widetilde{B})$ must be equal, i.e., $\sum \mathcal{R}(A, B) = \sum \mathcal{R}(\widetilde{A}, \widetilde{B})$. Consequently, $\mathcal{R}(A, B) > \mathcal{R}(\widetilde{A}, \widetilde{B})$ *are adjacent* is strengthened to $\mathcal{R}(A, B)$ covers $\mathcal{R}(\widetilde{A}, \widetilde{B})$. This is also remarked in [35, proof of Theorem 5.1]. A consequence of the change of representation from column minimal indices to $\mathcal{R}$ is that we in 5.6(1) have to introduce the restriction that r_0 may not be affected. Otherwise, the number of column minimal indices may change. The new condition is given in 5.6(1).

Now consider 5.5(2) which can be rewritten as:

$\sum \mathcal{R}(A,B) > \sum \mathcal{R}(\widetilde{A},\widetilde{B})$, $\mathcal{R}(A,B) > \mathcal{R}(\widetilde{A},\widetilde{B})$ are adjacent, and $\mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B}) = \mathcal{J}_{\mu_i}(A,B) \cup (1)$ for some μ_i (where μ_i can be a new eigenvalue).

If $\sum \mathcal{R}(A,B) > \sum \mathcal{R}(\widetilde{A},\widetilde{B})$, then $\mathcal{R}(A,B) > \mathcal{R}(\widetilde{A},\widetilde{B})$ *are adjacent* if and only if $\mathcal{R}(\widetilde{A},\widetilde{B})$ can be derived from $\mathcal{R}(A,B)$ in the following way. If $r_{\epsilon_1} = 1$ and $\epsilon_1 \geq 1$ for $\mathcal{R}(A,B)$, then $\mathcal{R}(\widetilde{A},\widetilde{B}) = \mathcal{R}(A,B) \setminus (r_{\epsilon_1})$ [11]. Furthermore, the regular part is expanded by increasing the largest block for some eigenvalue by one, or by creating a 1 × 1 block for a new eigenvalue. It follows that condition 5.5(2) corresponds to rule (2) for orbits of matrix pencils, which already fulfills both the necessary and sufficient conditions, and we have 5.6(2).

Finally, 5.5(4) can be rewritten as:

$\mathcal{R}(A,B) = \mathcal{R}(\widetilde{A},\widetilde{B})$ and $\mathcal{J}_{\mu_i}(A,B) < \mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B})$ for all μ_i.

This condition considers the case when the two matrix pairs have equal right singular parts, as opposed to 5.5(1) where the regular parts are the same. The conditions $\mathcal{R}(A,B) = \mathcal{R}(\widetilde{A},\widetilde{B})$ and $\mathcal{J}_{\mu_i}(A,B) < \mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B})$ do not guarantee that (A,B) covers $(\widetilde{A},\widetilde{B})$. To guarantee that (A,B) covers $(\widetilde{A},\widetilde{B})$ the corresponding integer partitions $\mathcal{J}_{\mu_i}(A,B)$ and $\mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B})$ must also cover each other, which corresponds to the matrix case (Theorem 5.1). The new condition is given in 5.6(3). ∎

Theorem 5.7. *$\mathcal{B}(A,B)$ covers $\mathcal{B}(\widetilde{A},\widetilde{B})$* if and only if *one of the following conditions holds:*

(1) *$\mathcal{R}(A,B)$ covers $\mathcal{R}(\widetilde{A},\widetilde{B})$ where $r_0(A,B) = r_0(\widetilde{A},\widetilde{B})$, and $\mathcal{J}_{\mu_i}(A,B) = \mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B})$ for all eigenvalues μ_i.*

(2) *If $r_{\epsilon_1} = 1$ and $\epsilon_1 \geq 1$ for $\mathcal{R}(A,B)$, then $\mathcal{R}(\widetilde{A},\widetilde{B}) = \mathcal{R}(A,B) \setminus (r_{\epsilon_1})$, $\mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B}) = (1)$ for a new eigenvalue μ_i, and $\mathcal{J}_{\mu_j}(A,B) = \mathcal{J}_{\mu_j}(\widetilde{A},\widetilde{B})$ for all $\mu_j \neq \mu_i$.*

(3) *$\mathcal{R}(A,B) = \mathcal{R}(\widetilde{A},\widetilde{B})$, $\mathcal{J}_{\mu_i}(A,B)$ covers $\mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B})$ for one eigenvalue μ_i, and $\mathcal{J}_{\mu_j}(A,B) = \mathcal{J}_{\mu_j}(\widetilde{A},\widetilde{B})$ for all $\mu_j \neq \mu_i$.*

(4) *$\mathcal{R}(A,B) = \mathcal{R}(\widetilde{A},\widetilde{B})$, $\mathcal{J}_{\mu_i}(\widetilde{A},\widetilde{B}) = \mathcal{J}_{\mu_i}(A,B) \cup \mathcal{J}_{\mu_j}(A,B)$ for one pair of eigenvalues μ_i and μ_j, $\mu_i \neq \mu_j$, and $\mathcal{J}_{\mu_k}(A,B) = \mathcal{J}_{\mu_k}(\widetilde{A},\widetilde{B})$ for all $\mu_k \neq \mu_i, \mu_j$.*

Proof. The proof of the bundle case follows directly from Theorem 5.6 and the covering rules for bundles of matrix pencils given in [18]. ∎

Notably, Theorem 5.7 has four rules in contrary to Theorem 5.6 which has three rules. The additional rule (4) follows from the fact that eigenvalues can coalesce in the bundle case.

From the dual relation between the controllability pair (A,B) and the observability pair (A,C), it follows that replacing partition $\mathcal{R}$ by $\mathcal{L}$ in Theorems 5.6 and 5.7 give the cover conditions for the observability pair (A,C). We remark that the theorems are valid only for independent matrix pairs (A,B) and (A,C), respectively. They cannot be applied straightforwardly to the related matrix triple (A,B,C) or matrix quadruple (A,B,C,D). The covering relations for orbits and bundles of the controllability and observability pairs in terms of coin rules are given in Corollaries 5.8 and 5.9. The reformulations are done using the definition of integer partitions in section 2.2.

Table 5.1

Given the structure integer partitions $\mathcal{R}$, $\mathcal{L}$, and $\mathcal{J}_{\mu_i}$ of a matrix pair, one of the following if-and-only-if *rules finds $(\widetilde{A},\widetilde{B})$ or $(\widetilde{A},\widetilde{C})$ fulfilling orbit or bundle covering relations with (A,B) or (A,C), respectively.*

A. $\mathcal{O}(A,B)$ covers $\mathcal{O}(\widetilde{A},\widetilde{B})$: (1) Minimum rightward coin move in $\mathcal{R}$. (2) If the rightmost column in $\mathcal{R}$ is one single coin, move that coin to a new rightmost column of some $\mathcal{J}_{\mu_i}$ (which may be empty initially). (3) Minimum leftward coin move in any $\mathcal{J}_{\mu_i}$. Rules 1 and 2 are not allowed to do coin moves that affect r_0. **B.** $\mathcal{B}(A,B)$ covers $\mathcal{B}(\widetilde{A},\widetilde{B})$ (1) Same as rule **A**(1). (2) Same as rule **A**(2), except it is only allowed to start a new set corresponding to a new eigenvalue (i.e., no appending to nonempty sets). (3) Same as rule **A**(3). (4) Let any pair of eigenvalues coalesce, i.e., take the union of their sets of coins.	**E.** $\mathcal{O}(A,C)$ covers $\mathcal{O}(\widetilde{A},\widetilde{C})$: (1) Minimum rightward coin move in $\mathcal{L}$. (2) If the rightmost column in $\mathcal{L}$ is one single coin, move that coin to a new rightmost column of some $\mathcal{J}_{\mu_i}$ (which may be empty initially). (3) Minimum leftward coin move in any $\mathcal{J}_{\mu_i}$. Rules 1 and 2 are not allowed to do coin moves that affect l_0. **F.** $\mathcal{B}(A,C)$ covers $\mathcal{B}(\widetilde{A},\widetilde{C})$: (1) Same as rule **E**(1). (2) Same as rule **E**(2), except it is only allowed to start a new set corresponding to a new eigenvalue (i.e., no appending to nonempty sets). (3) Same as rule **E**(3). (4) Let any pair of eigenvalues coalesce, i.e., take the union of their sets of coins.
C. $\mathcal{O}(A,B)$ is covered by $\mathcal{O}(\widetilde{A},\widetilde{B})$ (1) Minimum leftward coin move in $\mathcal{R}$, without affecting r_0. (2) If the rightmost column in some $\mathcal{J}_{\mu_i}$ consists of one coin only, move that coin to a new rightmost column in $\mathcal{R}$. (3) Minimum rightward coin move in any $\mathcal{J}_{\mu_i}$. **D.** $\mathcal{B}(A,B)$ is covered by $\mathcal{B}(\widetilde{A},\widetilde{B})$ (1) Same as rule **C**(1). (2) Same as rule **C**(2), except that $\mathcal{J}_{\mu_i}$ must consist of one coin only. (3) Same as rule **C**(3). (4) For any $\mathcal{J}_{\mu_i}$, divide the set of coins into two new sets so that their union is $\mathcal{J}_{\mu_i}$.	**G.** $\mathcal{O}(A,C)$ is covered by $\mathcal{O}(\widetilde{A},\widetilde{C})$: (1) Minimum leftward coin move in $\mathcal{L}$, without affecting l_0. (2) If the rightmost column in some $\mathcal{J}_{\mu_i}$ consists of one coin only, move that coin to a new rightmost column in $\mathcal{L}$. (3) Minimum rightward coin move in any $\mathcal{J}_{\mu_i}$. **H.** $\mathcal{B}(A,C)$ is covered by $\mathcal{B}(\widetilde{A},\widetilde{C})$: (1) Same as rule **G**(1). (2) Same as rule **G**(2), except that $\mathcal{J}_{\mu_i}$ must consist of one coin only. (3) Same as rule **G**(3). (4) For any $\mathcal{J}_{\mu_i}$, divide the set of coins into two new sets so that their union is $\mathcal{J}_{\mu_i}$.

Corollary 5.8. *Given the structure integer partitions $\mathcal{R}$ and $\mathcal{J}_{\mu_i}$ of (A,B), one of the* if-and-only-if *rules of* **A**–**D** *in Table* 5.1 *finds $(\widetilde{A},\widetilde{B})$ fulfilling orbit or bundle covering relations with (A,B).*

Corollary 5.9. *Given the structure integer partitions $\mathcal{L}$ and $\mathcal{J}_{\mu_i}$ of (A,C), one of the* if-and-only-if *rules of* **E**–**H** *in Table* 5.1 *finds $(\widetilde{A},\widetilde{C})$ fulfilling orbit or bundle covering relations with (A,C).*

The major difference between the rules for matrix pencils and matrix pairs is that rule (4) in Theorem 5.2 does not apply to matrix pairs, since there is only one type of singular blocks (L_i or L_j^T) in each matrix pair type. Moreover, rules (1) and (2) of **A**–**D** in Table 5.1 apply only to the structure integer partition $\mathcal{R}$ and rules (1) and (2) of **E**–**H** in Table 5.1 apply only to $\mathcal{L}$.

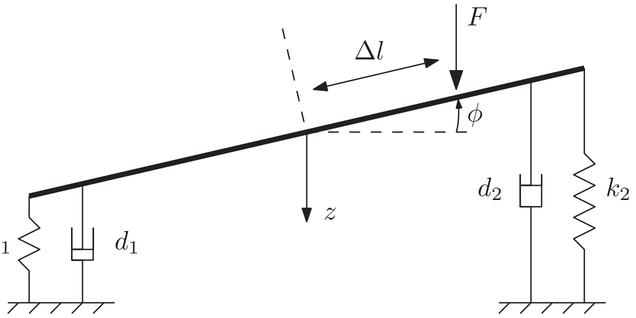


FIG. 6.1. *Mechanical system consisting of a uniform platform controlled by a vertical force* [44].

6. Illustrating the stratification. To illustrate the concept of stratification we consider two examples from systems and control applications. We use the software tool StratiGraph [38, 41] for computing and visualizing the closure hierarchy graphs for the different matrix pairs in the examples. The numerical results regarding Kronecker structure information and upper/lower bounds are computed using the prototype of the matrix canonical structure (MCS) toolbox for MATLAB [39, 22].

6.1. Mechanical system. The first example is a mechanical system studied by Mailybaev [44]; see Figure 6.1. It consists of a thin uniform platform supported at both ends by springs, where the platform has mass m and length $2l$, and the springs have elasticity coefficients k_1, k_2 and viscous damping coefficients d_1, d_2. The position of the platform is determined by the vertical coordinate z of its center and the angle ϕ between the platform and the horizontal axis.

At distance Δl, $-1 \le \Delta \le 1$, from the center of the platform a force F is applied, which is the control parameter of the system. The equilibrium of the system when $F = 0$ is assumed to be $z = 0$ and $\phi = 0$. For a zero force F and a nonzero z and/or ϕ, the system oscillates with a decaying amplitude until it reaches equilibrium asymptotically. If the system is controllable, there exists a control action such that the system can be put into equilibrium in finite time. Otherwise, if it is uncontrollable or close to an uncontrollable system this task becomes difficult or even impossible.

By linearizing the equations of motion of the system near the equilibrium the system can be expressed by the state-space model $\dot{x} = Ax(\tau) + Bu(\tau)$, where the derivative is taken with respect to time $\tau = t/\omega$ and ω is a time scale coefficient. The resulting state-space model is

$$\begin{bmatrix} \omega\dot{z}/l \\ \omega\dot{\phi} \\ \omega^2\ddot{z}/l \\ \omega^2\ddot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -c_1 & -c_2 & -f_1 & -f_2 \\ -3c_2 & -3c_1 & -3f_2 & -3f_1 \end{bmatrix} \begin{bmatrix} z/l \\ \phi \\ \omega\dot{z}/l \\ \omega\dot{\phi} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ -3\Delta \end{bmatrix} \frac{\omega^2}{ml} F, \tag{6.1}$$

where

$$c_1 = \frac{(k_1 + k_2)\omega^2}{m}, \ c_2 = \frac{(k_1 - k_2)\omega^2}{m}, \ f_1 = \frac{(d_1 + d_2)\omega}{m}, \text{ and } f_2 = \frac{(d_1 - d_2)\omega}{m}.$$

Let us consider a controllability pair of (6.1), denoted (A_0, B_0), with the parameters $d_1 = 4$, $d_2 = 4$, $k_1 = 6$, $k_2 = 6$, $m = 3$, $l = 1$, $\omega = 0.01$, and $\Delta = 0$. The KCF of

(A_0, B_0) is $L_2 \oplus J_1(\alpha) \oplus J_1(\beta)$ with the corresponding Brunovsky canonical form

$$\begin{bmatrix} A_B & B_B \end{bmatrix} - \lambda \begin{bmatrix} I_4 & 0 \end{bmatrix} = \left[\begin{array}{cccc|c} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & \alpha & 0 & 0 \\ 0 & 0 & 0 & \beta & 0 \end{array}\right] - \lambda \left[\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right],$$

where $\alpha = -0.02$ and $\beta = -0.06$. From the BCF of (A_0, B_0) we can directly see that the system is uncontrollable with the uncontrollable modes α and β; $\operatorname{rank}\big(\begin{bmatrix} A_B & B_B \end{bmatrix} - \lambda \begin{bmatrix} I_4 & 0 \end{bmatrix}\big) = 3$ for $\lambda \in \{\alpha, \beta\}$. The two uncontrollable modes correspond to that the angle ϕ and its velocity $\dot{\phi}$ cannot be controlled by the force F.

In [44], Mailybaev developed a quantitative perturbation method for local analysis of the uncontrollability set for a linear dynamical system depending on parameters. A *uncontrollability set* is defined as the set of values of a parameter vector p for which (A, B) depending on p is uncontrollable. In [44], an uncontrollable set for (A_0, B_0) is computed by letting the parameters c_1 and f_1 be fixed and varying the parameter vector $p = (c_2, f_2, \Delta)$ in the range of $-c_1 < c_2 < c_1$ and $-f_1 < f_2 < f_1$. It is also shown how the modes of (A_0, B_0) are changing over this set.

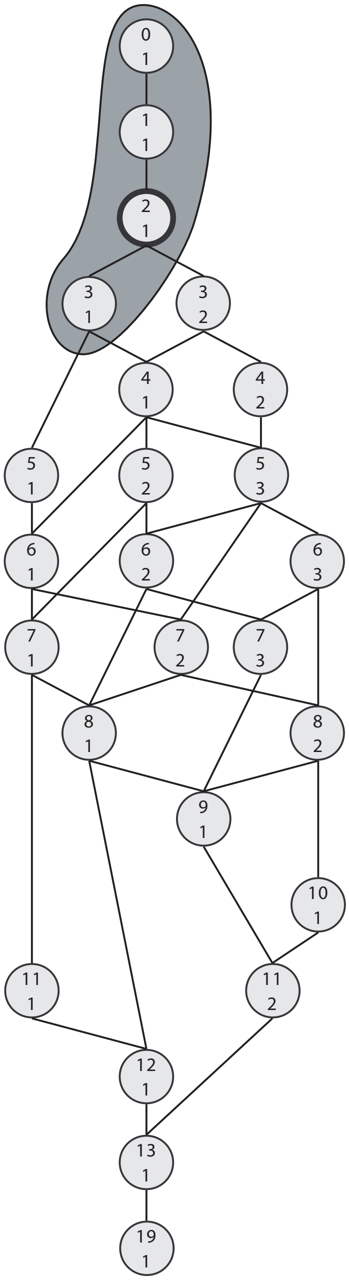
With the stratification theory, the quantitative results presented in [22, 44] and additional results like distance to uncontrollability [34, 46] are complemented with new qualitative information. In the following, we step-by-step illustrate the procedure to obtain the bundle stratification of the controllability pencil $\mathbf{S}_{\mathrm{C}}(\lambda)$ of size 4×5, which (A_0, B_0) is part of. Note that we can change only the values of the parameters c_1, c_2, f_1, f_2, and Δ in the state-space model (6.1); The first two rows of the matrix A and the first three of B are fixed. As we will see, due to the special structure of A and B not all bundles or parts of these exist for (A, B), which would exist for a controllability pair with unrestricted matrices A and B. We only show in details how to get the subgraph representing the stratification of possible structures. The complete bundle stratification of (A, B) is displayed in Figure 6.2, where the nodes corresponding to the bundles of possible structures are highlighted by the grey area. Let $c\!:\!k$ denote node $\textcircled{\scriptsize\substack{c\\k}}$ in Figure 6.2, where c is the codimension of the corresponding bundle and k is an order number that identifies individual nodes with the same codimension.

The first step is to compute the codimension of (A_0, B_0) using (4.3): $\operatorname{cod}(\mathcal{O}(A_0, B_0)) = 0 + (1 + 1) + 1(1 + 1) = 4$. To get the codimension of the bundle the number of distinct eigenvalues are subtracted: $\operatorname{cod}(\mathcal{B}(A_0, B_0)) = 4 - 2 = 2$. In Figure 6.2, $\mathcal{B}(A_0, B_0)$ corresponds to node $2\!:\!1$. To find covered or covering bundle(s) we use the set of rules $\mathbf{B}$ and $\mathbf{D}$, respectively, in Table 5.1. To apply these rules we express the KCF of (A_0, B_0) in terms of its structure integer partitions: $\mathcal{R} = (1, 1, 1)$, $\mathcal{J}_\alpha = (1)$, and $\mathcal{J}_\beta = (1)$. We are now ready to determine which bundle(s) that covers $\mathcal{B}(A_0, B_0)$.

Rule $\mathbf{D}(1)$ is not applicable because it would affect r_0 (the first column of $\mathcal{R}$). Rule $\mathbf{D}(2)$ can be applied to either $\mathcal{J}_\alpha$ or $\mathcal{J}_\beta$; we choose the former:

$\mathcal{R}$: ○○○ , $\mathcal{J}_\alpha$: ● , $\mathcal{J}_\beta$: ○ $\Rightarrow$ $\mathcal{R}$: ○○○● , $\mathcal{J}_\beta$: ○ ,

which gives the structure $L_3 \oplus J_1(\beta)$. The rules $\mathbf{D}(3)$ and $\mathbf{D}(4)$ are not applicable because $\mathcal{J}_\alpha$ and $\mathcal{J}_\beta$ only have one coin each. So the only bundle covering $\mathcal{B}(A_0, B_0)$ is the bundle with KCF $L_3 \oplus J_1(\beta)$, which has codimension 1 and is represented by node $1\!:\!1$ in Figure 6.2. Furthermore, this system is uncontrollable with one uncontrollable



Codimension 0
1 L_4
Codimension 1
1 $L_3 \oplus J_1(\mu_1)$
Codimension 2
1 $L_2 \oplus J_1(\mu_1) \oplus J_1(\mu_2)$
Codimension 3
1 $L_2 \oplus J_2(\mu_1)$
2 $L_1 \oplus J_1(\mu_1) \oplus J_1(\mu_2) \oplus J_1(\mu_3)$
Codimension 4
1 $L_1 \oplus J_2(\mu_1) \oplus J_1(\mu_2)$
2 $L_0 \oplus J_1(\mu_1) \oplus J_1(\mu_2) \oplus J_1(\mu_3) \oplus J_1(\mu_4)$
Codimension 5
1 $L_2 \oplus 2J_1(\mu_1)$
2 $L_1 \oplus J_3(\mu_1)$
3 $L_0 \oplus J_2(\mu_1) \oplus J_1(\mu_2) \oplus J_1(\mu_3)$
Codimension 6
1 $L_1 \oplus 2J_1(\mu_1) \oplus J_1(\mu_2)$
2 $L_0 \oplus J_3(\mu_1) \oplus J_1(\mu_2)$
3 $L_0 \oplus J_2(\mu_1) \oplus J_2(\mu_2)$
Codimension 7
1 $L_1 \oplus J_2(\mu_1) \oplus J_1(\mu_1)$
2 $L_0 \oplus 2J_1(\mu_1) \oplus J_1(\mu_2) \oplus J_1(\mu_3)$
3 $L_0 \oplus J_4(\mu_1)$
Codimension 8
1 $L_0 \oplus J_2(\mu_1) \oplus J_1(\mu_1) \oplus J_1(\mu_2)$
2 $L_0 \oplus 2J_1(\mu_1) \oplus J_2(\mu_2)$
Codimension 9
1 $L_0 \oplus J_3(\mu_1) \oplus J_1(\mu_1)$
Codimension 10
1 $L_0 \oplus 2J_1(\mu_1) \oplus 2J_1(\mu_2)$
Codimension 11
1 $L_1 \oplus 3J_1(\mu_1)$
2 $L_0 \oplus 2J_2(\mu_1)$
Codimension 12
1 $L_0 \oplus 3J_1(\mu_1) \oplus J_1(\mu_2)$
Codimension 13
1 $L_0 \oplus J_2(\mu_1) \oplus 2J_1(\mu_1)$
Codimension 19
1 $L_0 \oplus 4J_1(\mu_1)$

FIG. 6.2. *The graph shows the complete bundle stratification of a* 4×5 *controllability pencil* $\mathbf{S}_{\mathrm{C}}(\lambda)$, *where the grey area marks the possible structures for the mechanical system* (6.1). *The upper number in each node is the codimension of the corresponding bundle. The lower number is an order number that identifies individual nodes with the same codimension. The table to the right of the graph displays the corresponding KCF structures associated with the nodes in the graph.*

mode $\beta = -0.06$, which also can be seen from its BCF:

$$\begin{bmatrix} A_B & B_B \end{bmatrix} - \lambda \begin{bmatrix} I_4 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & : 0 \\ 0 & 0 & 1 & 0 & : 0 \\ 0 & 0 & 0 & 0 & : 1 \\ 0 & 0 & 0 & -0.06 & : 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 & 0 : 0 \\ 0 & 1 & 0 & 0 : 0 \\ 0 & 0 & 1 & 0 : 0 \\ 0 & 0 & 0 & 1 : 0 \end{bmatrix}.$$

For the system (6.1), we can at least find two cases which belong to this bundle. The first one[1] occurs when the elasticity coefficients k_1 and k_2 are zero. This case is not of practical interest, since it corresponds to a system with no springs. The second case occurs when element $A(4,2) = 1.2e{-}3$ becomes zero and element $A(4,3)$ is perturbed with $\epsilon \geq 1e{-}12$. The KCF of this system is $L_3 \oplus J_1(0)$.

We continue by repeating the procedure for $L_3 \oplus J_1(\beta)$. As for the previous structure, the only rule applicable is $\mathbf{D}(2)$. So, we take the single coin in $\mathcal{J}_\beta$ and move that to a new right-most column of $\mathcal{R}$:

$\mathcal{R}$: ○○○○ , $\mathcal{J}_\beta$: ● ⇒ $\mathcal{R}$: ○○○○● ,

which gives the KCF L_4 with BCF:

$$\begin{bmatrix} A_B & B_B \end{bmatrix} - \lambda \begin{bmatrix} I_4 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0:0 \\ 0 & 0 & 1 & 0:0 \\ 0 & 0 & 0 & 1:0 \\ 0 & 0 & 0 & 0:1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 & 0:0 \\ 0 & 1 & 0 & 0:0 \\ 0 & 0 & 1 & 0:0 \\ 0 & 0 & 0 & 1:0 \end{bmatrix}.$$

This is the most generic case represented by the topmost node 0:1 in Figure 6.2 and has codimension 0. As we can see from its BCF, it is controllable; $\operatorname{rank}([A_B \quad B_B] - \lambda[I_4 \quad 0]) = 4$ for all $\lambda \in \mathbb{C}$. In other words, there exists a control parameter F such that any state of z and ϕ can be reached in finite time.

After having reached the most generic case and the top of the closure-hierarchy graph, we continue by determining the bundle(s) covered by $\mathcal{B}(A_0, B_0)$ using the set of rules $\mathbf{B}$ in Table 5.1. But first, we remark that the mechanical system represented by the state-space system (6.1) must have an L block of at least size 2; i.e., it has at most two uncontrollable modes. This can be seen by studying the system with all parameters set to zero:

$$\begin{bmatrix} \omega\dot{z}/l \\ \omega\dot{\phi} \\ \omega^2\ddot{z}/l \\ \omega^2\ddot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z/l \\ \phi \\ \omega\dot{z}/l \\ \omega\dot{\phi} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \frac{\omega^2}{ml} F,$$

which has the KCF $L_2 \oplus J_2(0)$. The bundle of this canonical structure has codimension 3 and is represented by node 3:1 in Figure 6.2. Indeed, it is the most degenerate structure possible for the state-space system (6.1). As we can see from the graph in Figure 6.2, $\mathcal{B}(L_2 \oplus J_2(0))$ is covered by $\mathcal{B}(A_0, B_0)$. This closure relation is obtained by applying rule $\mathbf{B}(4)$ to (A_0, B_0):

$\mathcal{R}$: ○○○ , $\mathcal{J}_\alpha$: ○ $\bigcup$ $\mathcal{J}_\beta$: ● ⇒ $\mathcal{R}$: ○○○ , $\mathcal{J}_\alpha$: ○● .

We can also reach this bundle by changing the value of m in (A_0, B_0). Let $(\widetilde{A}_0, \widetilde{B}_0)$ have the same parameters as (A_0, B_0) but with m unfixed. With $m = 4$, $(\widetilde{A}_0, \widetilde{B}_0)$ has KCF $L_2 \oplus J_2(-0.2)$ and by a small perturbation on m we again reach the bundle of (A_0, B_0), $\mathcal{B}(L_2 \oplus J_1(\mu_1) \oplus J_1(\mu_2))$. Actually, for $m < 4$ $(\widetilde{A}_0, \widetilde{B}_0)$ has KCF $L_2 \oplus J_1(\mu_1) \oplus J_1(\mu_2)$ with two real eigenvalues, and for $m > 4$ the system has instead one complex conjugate pair of eigenvalues.

The only other rule that can be applied to (A_0, B_0) is rule $\mathbf{B}(2)$, producing the structure $L_1 \oplus J_1(\mu_1) \oplus J_1(\mu_2) \oplus J_1(\mu_3)$. However, this structure has three uncontrol-

[1] The parameters in A are $c_1 = c_2 = 0$, and one of f_1 and f_2 is nonzero while the other one is equal to zero (Δ is arbitrary).

lable modes which is not possible for the mechanical system considered. So, the closure hierarchy for the state-space system (6.1) corresponds to the highlighted subgraph of the complete bundle stratification of 4 × 5 controllability pencil in Figure 6.2.

Notice, there also exists a structure $L_2 \oplus 2J_1(\mu)$ (node 5 : 1) in the closure hierarchy which has an L block of size at least two, and, therefore, also should be possible. However, since the codimension of $\mathcal{B}(L_2 \oplus 2J_1(\mu))$ is less than the most degenerate case $L_2 \oplus J_2(0)$, this case cannot appear for this example.

6.2. Boeing 747. As the second example, we study the orbit closure hierarchy of a linearized nominal longitudinal model of a Boeing 747 considered in [51]. In our model we have joined nine inputs into five, which results in a model with 5 states, 6 outputs, and 5 inputs:

$$x = \begin{bmatrix} \delta q \\ \delta V_{TAS} \\ \delta\alpha \\ \delta\theta \\ \delta h_e \end{bmatrix} \begin{pmatrix} \text{pitch rate } (rad/s) \\ \text{true airspeed } (m/s) \\ \text{angle of attack } (rad) \\ \text{pitch angle } (rad) \\ \text{altitude } (m) \end{pmatrix}, \quad y = \begin{bmatrix} \delta\alpha \\ \delta\dot{V}_{TAS} \\ \delta\theta \\ \delta q \\ \delta V_z \\ \delta h_e \end{bmatrix} \begin{pmatrix} \text{angle of attack } (rad) \\ \text{acceleration } (m/s^2) \\ \text{pitch angle } (rad) \\ \text{pitch rate } (rad/s) \\ \text{vertical velocity } (m/s) \\ \text{altitude } (m) \end{pmatrix},$$

$$u = \begin{bmatrix} \delta_{ei} \\ \delta_{eo} \\ \delta_{ih} \\ \delta EPR_{1,4} \\ \delta EPR_{2,3} \end{bmatrix} \begin{pmatrix} \text{total inner elevator } (rad) \\ \text{total outer elevator } (rad) \\ \text{stabilizer trim angle } (rad) \\ \text{total thrust engine \#1 and \#4 } (rad) \\ \text{total thrust engine \#2 and \#3 } (rad) \end{pmatrix},$$

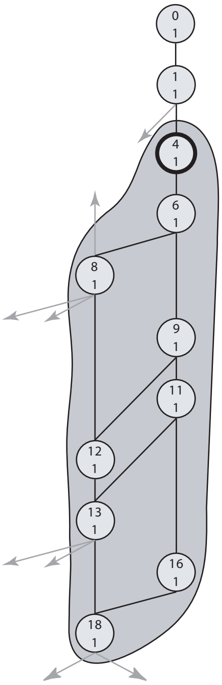
and the state-space matrices:

$$A = \begin{bmatrix} -0.4861 & 0.000317 & -0.5588 & 0 & -2.04 \cdot 10^{-6} \\ 0 & -0.0199 & 3.0796 & -9.8048 & 8.98 \cdot 10^{-5} \\ 1.0053 & -0.0021 & -0.5211 & 0 & 9.30 \cdot 10^{-6} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -92.6 & 92.6 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} -0.291 & -0.2988 & -1.286 & 0.0026 & 0.007 \\ 0 & 0 & -0.3122 & 0.3998 & 0.3998 \\ -0.0142 & -0.0148 & -0.0676 & -0.0008 & -0.0008 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & -0.0199 & 3.0796 & -9.8048 & 8.98 \cdot 10^{-5} \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -92.6 & 92.6 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.3122 & 0.3988 & 0.3988 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$



Codimension 0
1 $5L_1$
Codimension 1
1 $L_2 \oplus 3L_1 \oplus L_0$
Codimension 4
1 $2L_2 \oplus L_1 \oplus 2L_0$
Codimension 6
1 $L_3 \oplus 2L_1 \oplus 2L_0$
Codimension 8
1 $L_2 \oplus 2L_1 \oplus 2L_0 \oplus J_1(\mu)$
Codimension 9
1 $L_3 \oplus L_2 \oplus 3L_0$
Codimension 11
1 $L_4 \oplus L_1 \oplus 3L_0$
Codimension 12
1 $2L_2 \oplus 3L_0 \oplus J_1(\mu)$
Codimension 13
1 $L_3 \oplus L_1 \oplus 3L_0 \oplus J_1(\mu)$
Codimension 16
1 $L_5 \oplus 4L_0$
Codimension 18
1 $L_4 \oplus 4L_0 \oplus J_1(\mu)$

FIG. 6.3. *Subgraph of the complete orbit stratification of a controllability pencil of size* 5×10*, where the grey area marks the possible structures for the Boeing* 747 *model. The node with codimension* 4 *represents the orbit to a system corresponding to a Boeing* 747 *under flight. The four nodes in the left-most branch of the graph represent the orbits of uncontrollable systems with one uncontrollable mode.*

These state-space matrices correspond to a Boeing 747 under straight-and-level flight at altitude 600 m with speed 92.6 m/s, flap setting at 20°, and landing gears up. The aircraft has mass = 317,000 kg and the center of gravity coordinates are $X_{cg} = 25\%$, $Y_{cg} = 0$, and $Z_{cg} = 0$.

The corresponding controllability pencil of the state-space system is of size 5×10 and the observability pencil of size 11×5. First, let us consider the controllability pencil. Using StratiGraph the complete stratification of the orbit to a 5×10 controllability pencil can be computed, which has 74 nodes and 133 edges. In our case, we are only interested to know the closest uncontrollable systems which can be reached by a perturbation of the system matrices. Instead of generating the complete stratification, we derive only the controllable and the nearest uncontrollable systems, starting with the controllability pencil given by the state-space matrices A and B above.

As in the previous example, we begin by determining the KCF of the controllability pair (A, B) which is $2L_2 \oplus L_1 \oplus 2L_0$ with codimension 4. From the KCF (and BCF) we can see that the system is controllable with only three of the five input signals.[2]

Using the set of rules **A** and **C** in Table 5.1, the closure hierarchy around (A, B) can be determined. The resulting stratification graph is shown in Figure 6.3, where node $4\!:\!1$ corresponds to the orbit which (A, B) belongs to. We now take the structural

[2] The other two inputs (corresponding to the L_0 blocks) can be removed without loss of controllability. However, for safety reasons it is customary to have redundancy in the actuation system and the corresponding control surface in critical systems.

Table 6.1

Lower and upper bounds from the controllability pair (A, B) of a Boeing 747 *under flight with KCF* $2L_2 \oplus L_1 \oplus 2L_0$ *to the less generic orbits shown in Figure* 6.3.

Imposed structure from $2L_2 \oplus L_1 \oplus 2L_0$	cod	Lower bound	Upper bound
$L_3 \oplus 2L_1 \oplus 2L_0$	6	$1.29e-4$	$4.02e-2$
$L_2 \oplus 2L_1 \oplus 2L_0 \oplus J_1(\mu)$	8	$4.33e-4$	1.0
$L_3 \oplus L_2 \oplus 3L_0$	9	$5.97e-4$	$1.59e-3$
$L_4 \oplus L_1 \oplus 3L_0$	11	$8.47e-4$	$1.59e-3$
$2L_2 \oplus 3L_0 \oplus J_1(\mu)$	12	$1.09e-3$	$2.48e-1$
$L_3 \oplus L_1 \oplus 3L_0 \oplus J_1(\mu)$	13	$1.33e-3$	$1.79e-1$
$L_5 \oplus 4L_0$	16	$1.78e-2$	$5.56e-1$
$L_4 \oplus 4L_0 \oplus J_1(\mu)$	18	$7.57e-2$	$5.56e-1$

restrictions of A and B into consideration. By keeping all zeros and ones constant and choosing all free elements in A and B nonzero, it follows that the most generic orbit must have at least $2L_0$ blocks: The number of L_0 blocks is $m - \text{rank}(B) = 5 - 3 = 2$. This excludes $\mathcal{O}(5L_1)$ and $\mathcal{O}(L_2 \oplus 3L_1 \oplus L_0)$ from possible orbits, and the most generic orbit is indeed the one (A, B) belongs to. The most degenerate orbit has KCF $5L_1 \oplus J_2(\mu_1) \oplus 3J_1(\mu_2)$, which is obtained by considering the system with all parameters set to zero. This orbit is, however, more degenerate than those of interest.

Using the stratification graph together with bounds on the distance to uncontrollability we can validate the robustness of the system. For a controllable pair (A, B), the distance to uncontrollability [48] is defined as

$$\tau(A, B) = \min \left\{ \| \begin{bmatrix} \Delta A & \Delta B \end{bmatrix} \| : (A + \Delta A, B + \Delta B) \text{ is uncontrollable} \right\},$$

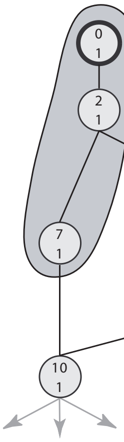
where $\| \cdot \|$ denotes the 2-norm or Frobenius norm. Equivalently,

$$\tau(A, B) = \inf_{\lambda \in \mathbb{C}} \sigma_{\min} \left(\begin{bmatrix} A - \lambda I & B \end{bmatrix} \right),$$

where $\sigma_{\min}(X)$ denotes the smallest singular value of $X \in \mathbb{C}^{n \times (n+m)}$ [19]. Using the Matlab implementation [47] of the methods presented in [34, 46], the distance to uncontrollability can be computed where $\tau(A, B)$ is bounded within an interval $(l, u]$ with any desired accuracy $tol \geq u - l$. For the above system, the computed distance to uncontrollability is within $(3.0323e{-}2,\ 3.0332e{-}2]$, where $tol = 10^{-5}$.

Furthermore, using the technique presented in [22], the upper and lower bounds to all less generic controllability pairs shown in Figure 6.3 can be computed; see Table 6.1. The upper bounds are based on staircase regularizing perturbations, and the lower bounds are of Eckart–Young type and are derived from the matrix representations $T_{(A,B)}$ (4.1) and $T_{(A,C)}$ (4.2) of $\tan(A, B)$ and $\tan(A, C)$, respectively. For the upper bounds, the implemented algorithm uses a naive approach to find a nearby matrix pair and the computed upper bounds are sometimes too conservative. However, we can observe that the above computed distance to uncontrollability is within the bounds of the uncontrollable systems with codimensions 8, 12, and 13.

Briefly, we also consider the 11×5 observability pencil $\mathbf{S}_{\text{O}}(\lambda)$ given by the above state-space matrices. This matrix pair has the KCF $5L_1^T \oplus L_0^T$ with codimension 0, i.e., it is completely observable. Considering the structural restrictions of (A, C), the most degenerate orbit possible has the KCF $4L_1^T \oplus 2L_0^T \oplus J_1(\mu)$ with codimension 7. This can be seen by studying the matrix C with all parameters set to zero; at most two L_0^T blocks can exist, $p - \text{rank}(C) = 5 - 3 = 2$. Using the set of rules $\mathbf{E}$ (and $\mathbf{G}$) in Table 5.1, the closure hierarchy shown in Figure 6.4 is derived.



Codimension 0
1 $5L_1^T \oplus L_0^T$
Codimension 2
1 $L_2^T \oplus 3L_1^T \oplus 2L_0^T$
Codimension 6
1 $2L_2^T \oplus L_1^T \oplus 3L_0^T$
Codimension 7
1 $4L_1^T \oplus 2L_0^T \oplus J_1(\mu)$
Codimension 8
1 $L_3^T \oplus 2L_1^T \oplus 3L_0^T$
Codimension 10
1 $L_2^T \oplus 2L_1^T \oplus 3L_0^T \oplus J_1(\mu)$

FIG. 6.4. *Subgraph of the complete orbit stratification of an observability pencil of size* 11 × 5*, where the grey area marks the possible structures for the Boeing* 747 *model. The node with codimension* 0 *represents the orbit to a system corresponding to a Boeing* 747 *under flight. The two nodes* 7 : 1 *and* 10 : 1 *represent the orbits of unobservable systems with one unobservable mode.*

7. Conclusions. We have derived the closure and cover conditions for orbits and bundles of matrix pairs, where the cover conditions are new results. In line with previous work on matrices and matrix pencils [17, 18], we have derived the stratification rules for matrix pairs, both for controllability pairs (A, B) and observability pairs (A, C), in terms of coin moves.

The results are illustrated with two examples taken from real applications in systems and control. We show how the rules are used and how they provide qualitative information of a system, which together with distance information are useful for validating an LTI state-space system.

Acknowledgments. A special thanks to Pedher Johansson for implementing our stratification rules for matrix pairs in StratiGraph. The authors are also grateful to Andras Varga for providing the data for the Boeing 747 model and for constructive comments around this example. Finally, we appreciate the constructive comments from the referees.

REFERENCES

[1] V. I. ARNOLD, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.

[2] J. M. BERG AND H. G. KWATNY, *Unfolding the zero structure of a linear control system*, Linear Algebra Appl., 258 (1997), pp. 19–39.

[3] K. BONGARTZ, *On degenerations and extensions of finite dimensional modules*, Adv. Math., 121 (1996), pp. 245–287.

[4] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–188.

[5] A. BUNSE-GERSTNER, R. BYERS, V. MEHRMANN, AND N. K. NICHOLS, *Feedback design for regularizing descriptor systems*, Linear Algebra Appl., 299 (1999), pp. 119–151.

[6] J. V. BURKE, A. S. LEWIS, AND M. L. OVERTON, *Pseudospectral components and the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 350–361.

[7] J. CLOTET, M. I. GARCÍA-PLANAS, AND M. D. MAGRET, *Estimating distances from quadruples satisfying stability properties to quadruples not satisfying them*, Linear Algebra Appl., 332–334 (2001), pp. 541–567.

[8] I. DE HOYOS, *Points of continuity of the Kronecker canonical form*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 278–300.

[9] F. De Terán and F. Dopico, *Low rank perturbation of Kronecker structures without full rank*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 496–529.

[10] F. De Terán and F. Dopico, *A note on generic Kronecker orbits of matrix pencils with fixed rank*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 491–496.

[11] C. DeConcini, D. Eisenbud, and C. Procesi, *Young diagrams and determinantal varieties*, Invent. Math., 56 (1980), pp. 129–165.

[12] J. Demmel and A. Edelman, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Linear Algebra Appl., 230 (1995), pp. 61–87.

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

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

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

[16] H. Den Boer and P. A. Thijsse, *Semi-stability of sums of partial multiplicities under additive perturbation*, Integral Equations Operator Theory, 3 (1980), pp. 23–42.

[17] 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.

[18] A. Edelman, E. Elmroth, and B. Kågström, *A geometric approach to perturbation theory of matrices and matrix pencils. Part* II: *A stratification-enhanced staircase algorithm*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 667–669.

[19] R. Eising, *Between controllable and uncontrollable*, Systems Control Lett., 4 (1984), pp. 263–264.

[20] E. Elmroth, P. Johansson, S. Johansson, and B. Kågström, *Orbit and bundle stratification of controllability and observability matrix pairs in StratiGraph*, in Proceedings of the Sixteenth International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), B. D. M. et.al., ed., Leuven, Belgium, July 2004. On CD.

[21] E. Elmroth, P. Johansson, and B. Kågström, *Computation and presentation of graph displaying closure hierarchies of Jordan and Kronecker structures*, Numer. Linear Algebra Appl., 8 (2001), pp. 381–399.

[22] E. Elmroth, P. Johansson, and B. Kågström, *Bounds for the distance between nearby Jordan and Kronecker structures in a closure hierarchy*, J. Math. Sci., 114 (2003), pp. 1765–1779.

[23] E. Elmroth and B. Kågström, *The set of* 2*-by-*3 *matrix pencils — Kronecker structures and their transitions under perturbations*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1–34.

[24] J. Ferrer, M. I. García, and F. Puerta, *Brunowsky local form of a holomorphic family of pairs of matrices*, Linear Algebra Appl., 253 (1997), pp. 175–198.

[25] J. Ferrer, M. I. García, and F. Puerta, *Regularity of the Brunovsky-Kronecker stratification*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 724–742.

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

[27] M. I. García-Planas and M. D. Magret, *Stratification of linear systems. Bifurcation diagrams for families of linear systems*, Linear Algebra Appl., 297 (1999), pp. 23–56.

[28] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, Wiley, New York, 1986.

[29] J.-M. Gracia and I. De Hoyos, *Puntos de continuidad de formas canónicas de matrices*, in the Homage Book of Prof. Luis de Albuquerque of Coimbra, Coimbra, 1987.

[30] J.-M. Gracia and I. De Hoyos, *Nearest pair with more nonconstant invariant factors and pseudospectrum*, Linear Algebra Appl., 298 (1999), pp. 143–158.

[31] J.-M. Gracia, I. De Hoyos, and I. Zaballa, *Perturbation of linear control systems*, Linear Algebra Appl., 121 (1989), pp. 353–383.

[32] M. Gu, *Finding well-conditioned similarities to block-diagonalize non-symmetric matrices is NP-hard*, J. Complexity, 11 (1995), pp. 377–391.

[33] M. Gu, *New methods for estimating the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 989–1003.

[34] M. Gu, E. Mengi, M. L. Overton, J. Xia, and J. Zhu, *Fast methods for estimating the distance to uncontrollability*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 477–502.

[35] D. Hinrichsen and J. O'Halloran, *Orbit closures of singular matrix pencils*, J. Pure Appl. Algebra, 81 (1992), pp. 117–137.

[36] D. Hinrichsen and J. O'Halloran, *A pencil approach to high gain feedback and generalized state space systems*, Kybernetika, 31 (1995), pp. 109–139.

[37] S. Iwata and R. Shimizu, *Combinatorial analysis of singular matrix pencils*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 245–259.

[38] P. Johansson, *StratiGraph User's Guide*, Technical report UMINF 03.21, Department of Computing Science, Umeå University, Sweden, 2003.

[39] P. Johansson, *Matrix Canonical Structure Toolbox*, Technical report UMINF 06.15, Department of Computing Science, Umeå University, Sweden, 2006.

[40] P. Johansson, *StratiGraph Developer's Guide*, Technical report UMINF 06.14, Department of Computing Science, Umeå University, Sweden, 2006.

[41] P. Johansson, *StratiGraph homepage.* Department of Computing Science, Umeå University, Sweden, Feb. 2008. http://www.cs.umu.se/research/nla/singular_pairs/stratigraph/.

[42] S. Johansson, *Canonical forms and stratification of orbits and bundles of system pencils*, Technical report UMINF 05.16, Department of Computing Science, Umeå University, Sweden, 2005.

[43] J. J. Loiseau, K. Özçaldiran, M. Malabre, and N. Karcanias, *Feedback canonical forms of singular systems*, Kybernetika, 27 (1991), pp. 289–305.

[44] A. Mailybaev, *Uncontrollability for linear autonomous multi-input dynamical systems depending on parameters*, SIAM J. Control Optim., 42 (2003), pp. 1431–1450.

[45] A. S. Markus and E. É. Parilis, *The change of the Jordan structure of a matrix under small perturbations*, Linear Algebra Appl., 54 (1983), pp. 139–152.

[46] E. Mengi, *Measures for robust stability and controllability*, Ph.D. Thesis, Courant Institute of Mathematical Sciences, New York, NY, 2006.

[47] E. Mengi, *Software for robust stability and controllability measures.* New York University, Computer Science Department, NY, Feb. 2008. http://www.cs.nyu.edu/~mengi/robuststability.html.

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

[49] D. D. Pervouchine, *Hierarchy of closures of matrix pencils*, J. Lie Theory, 14 (2004), pp. 443–479.

[50] A. Pokrzywa, *On perturbations and the equivalence orbit of a matrix pencil*, Linear Algebra Appl., 82 (1986), pp. 99–121.

[51] A. Varga, *On designing least order residual generators for fault detection and isolation*, in Proceedings of the 16th International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, 2007, pp. 323–330.

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

[53] J. Willems, *Topological classification and structural stability of linear systems*, J. Differential Equations, 35 (1980), pp. 306–318.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 2, pp. 227–247

CONVERGENCE ANALYSIS OF THE DOUBLING ALGORITHM FOR SEVERAL NONLINEAR MATRIX EQUATIONS IN THE CRITICAL CASE[*]

CHUN-YUEH CHIANG[†], ERIC KING-WAH CHU[‡], CHUN-HUA GUO[§], TSUNG-MING HUANG[¶], WEN-WEI LIN[||], AND SHU-FANG XU[**]

Abstract. In this paper, we review two types of doubling algorithm and some techniques for analyzing them. We then use the techniques to study the doubling algorithm for three different nonlinear matrix equations in the critical case. We show that the convergence of the doubling algorithm is at least linear with rate 1/2. As compared to earlier work on this topic, the results we present here are more general, and the analysis here is much simpler.

Key words. nonlinear matrix equation, minimal nonnegative solution, maximal positive definite solution, critical case, doubling algorithm, cyclic reduction, convergence rate

AMS subject classifications. 15A24, 15A48, 65F30, 65H10

DOI. 10.1137/080717304

1. Introduction.

The doubling algorithm has been studied for various nonlinear matrix equations in [1, 6, 7, 19, 21, 24, 27, 28, 34]. Its convergence behavior in the critical case, however, has not been fully investigated. The doubling algorithm is said to be structure-preserving (and denoted by SDA) because it preserves certain block structures for matrix pairs (or pencils) related to matrix equations.

In section 2, we review two types of doubling algorithm and some techniques for analyzing them. The presentation here is more general than in [34] and [24], to allow direct application to various matrix equations. In sections 3–5, the techniques reviewed in section 2 are used to study the convergence behavior of the doubling algorithm for three different nonlinear matrix equations in the critical case. As compared to previous papers, the results here are obtained with only basic assumptions. In particular, the results we obtain about a quadratic matrix equation arising from quasi-birth-death processes are more general than previous results, and the analysis here is much simpler. A connection between the doubling algorithm and the cyclic reduction algorithm is also pointed out for that quadratic matrix equation. Some concluding remarks are made in section 6.

[*] Received by the editors March 3, 2008; accepted for publication (in revised form) by D. A. Bini November 3, 2008; published electronically March 11, 2009.
http://www.siam.org/journals/simax/31-2/71730.html

[†] Department of Mathematics, National Tsinghua University, Hsinchu 300, Taiwan (d907201@oz.nthu.edu.tw).

[‡] School of Mathematical Sciences, Monash University, Building 28, VIC 3800, Australia (eric.chu@sci.monash.edu.au).

[§] Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (chguo@math.uregina.ca). The work of this author was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

[¶] Department of Mathematics, National Taiwan Normal University, Taipei 11677, Taiwan (min@math.ntnu.edu.tw). The work of this author was partially supported by the National Science Council and the National Center for Theoretical Sciences in Taiwan.

[||] Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan (wwlin@math.nctu.edu.tw). The work of this author was partially supported by the National Science Council and the National Center for Theoretical Sciences in Taiwan.

[**] LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China (xsf@pku.edu.cn).

2. The doubling algorithm.

The first three subsections are based on [34], [24], and [27], but the presentation here is more general. The last subsection is directly from [27].

2.1. SDA-1.

For a given matrix pair

$$L_0 = \begin{bmatrix} I & -G_0 \\ 0 & F_0 \end{bmatrix}, \quad M_0 = \begin{bmatrix} E_0 & 0 \\ -H_0 & I \end{bmatrix}, \tag{2.1}$$

where E_0, F_0, G_0, H_0 are $n \times n$, $m \times m$, $n \times m$, $m \times n$, respectively, we are going to define

$$L_k = \begin{bmatrix} I & -G_k \\ 0 & F_k \end{bmatrix}, \quad M_k = \begin{bmatrix} E_k & 0 \\ -H_k & I \end{bmatrix} \tag{2.2}$$

for all $k \geq 0$. Assume that L_k and M_k have been defined and $I - G_kH_k$ (and, thus, $I - H_kG_k$) is nonsingular for $k \geq 0$. Then we can define the matrices

$$\widetilde{L}_k = \begin{bmatrix} I & -E_k(I - G_kH_k)^{-1}G_k \\ 0 & F_k(I - H_kG_k)^{-1} \end{bmatrix}, \quad \widetilde{M}_k = \begin{bmatrix} E_k(I - G_kH_k)^{-1} & 0 \\ -F_k(I - H_kG_k)^{-1}H_k & I \end{bmatrix}.$$

It is easily verified that $\widetilde{L}_kM_k = \widetilde{M}_kL_k$. We then define

$$L_{k+1} = \widetilde{L}_kL_k = \begin{bmatrix} I & -\big(G_k + E_k(I - G_kH_k)^{-1}G_kF_k\big) \\ 0 & F_k(I - H_kG_k)^{-1}F_k \end{bmatrix},$$
$$M_{k+1} = \widetilde{M}_kM_k = \begin{bmatrix} E_k(I - G_kH_k)^{-1}E_k & 0 \\ -\big(H_k + F_k(I - H_kG_k)^{-1}H_kE_k\big) & I \end{bmatrix}.$$

Therefore, the sequence $\{L_k, M_k\}$ can be defined by the following doubling algorithm if no breakdown occurs.

ALGORITHM 2.1. (SDA-1) *Given* E_0, F_0, G_0, H_0.
For $k = 0, 1, \ldots$ *compute*

$$E_{k+1} = E_k(I - G_kH_k)^{-1}E_k, \tag{2.3}$$
$$F_{k+1} = F_k(I - H_kG_k)^{-1}F_k, \tag{2.4}$$
$$G_{k+1} = G_k + E_k(I - G_kH_k)^{-1}G_kF_k, \tag{2.5}$$
$$H_{k+1} = H_k + F_k(I - H_kG_k)^{-1}H_kE_k. \tag{2.6}$$

The algorithm requires about $\frac{14}{3}m^3 + 6m^2n + 6mn^2 + \frac{14}{3}n^3$ flops each iteration. Note that the flop count is $\frac{64}{3}n^3$ when $m = n$.

2.2. SDA-2.

For a given matrix pair

$$L_0 = \begin{bmatrix} -P_0 & I \\ T_0 & 0 \end{bmatrix}, \quad M_0 = \begin{bmatrix} V_0 & 0 \\ Q_0 & -I \end{bmatrix},$$

where all matrix blocks are $n \times n$, we are going to define

$$L_k = \begin{bmatrix} -P_k & I \\ T_k & 0 \end{bmatrix}, \quad M_k = \begin{bmatrix} V_k & 0 \\ Q_k & -I \end{bmatrix} \tag{2.7}$$

for all $k \geq 0$. Assume that L_k and M_k have been defined and $Q_k - P_k$ is nonsingular for $k \geq 0$. Then we can define the matrices

$$\widetilde{L}_k = \begin{bmatrix} I & -V_k(Q_k - P_k)^{-1} \\ 0 & T_k(Q_k - P_k)^{-1} \end{bmatrix}, \quad \widetilde{M}_k = \begin{bmatrix} V_k(Q_k - P_k)^{-1} & 0 \\ -T_k(Q_k - P_k)^{-1} & I \end{bmatrix}.$$

It is easily verified that $\widetilde{L}_k M_k = \widetilde{M}_k L_k$. We then define

$$L_{k+1} = \widetilde{L}_k L_k = \begin{bmatrix} -\left(P_k + V_k(Q_k - P_k)^{-1}T_k\right) & I \\ T_k(Q_k - P_k)^{-1}T_k & 0 \end{bmatrix},$$
$$M_{k+1} = \widetilde{M}_k M_k = \begin{bmatrix} V_k(Q_k - P_k)^{-1}V_k & 0 \\ Q_k - T_k(Q_k - P_k)^{-1}V_k & -I \end{bmatrix}.$$

Therefore, the sequence $\{L_k, M_k\}$ can be defined by the following doubling algorithm if no breakdown occurs.

ALGORITHM 2.2. (SDA-2) *Given* V_0, T_0, Q_0, P_0.
For $k = 0, 1, \ldots,$ *compute*

$$\begin{aligned} V_{k+1} &= V_k(Q_k - P_k)^{-1}V_k, \\ T_{k+1} &= T_k(Q_k - P_k)^{-1}T_k, \\ Q_{k+1} &= Q_k - T_k(Q_k - P_k)^{-1}V_k, \\ P_{k+1} &= P_k + V_k(Q_k - P_k)^{-1}T_k. \end{aligned}$$

This algorithm requires about $\frac{38}{3}n^3$ flops each iteration.

2.3. Relation between $\boldsymbol{L_k}$ and $\boldsymbol{M_k}$. Suppose we have

$$L_0 U = M_0 U E, \tag{2.8}$$

where the matrix pair (L_0, M_0) is the initialization for either SDA-1 or SDA-2, E is a square matrix, and U is any matrix of suitable dimension.

Premultiplying (2.8) with $\widetilde{L}_0$ and using $\widetilde{L}_0 M_0 = \widetilde{M}_0 L_0$, we get $L_1 U = M_1 U E^2$. In general, we have for each $k \geq 0$

$$L_k U = M_k U E^{2^k}. \tag{2.9}$$

Suppose that there are nonsingular matrices V and Z such that

$$V L_0 Z = J_L, \quad V M_0 Z = J_M, \tag{2.10}$$

and $J_L J_M = J_M J_L$. Then it follows that

$$M_0 Z J_L = V^{-1} J_M J_L = V^{-1} J_L J_M = L_0 Z J_M,$$

and

$$M_1 Z J_L^2 = \widetilde{M}_0 M_0 Z J_L^2 = \widetilde{M}_0 L_0 Z J_M J_L = \widetilde{L}_0 M_0 Z J_L J_M = \widetilde{L}_0 L_0 Z J_M^2 = L_1 Z J_M^2.$$

In general, we have for each $k \geq 0$

$$M_k Z J_L^{2^k} = L_k Z J_M^{2^k}. \tag{2.11}$$

2.4. Result on special Jordan blocks. Let $J_{\omega,p}$ be the $p \times p$ Jordan block with a unimodular eigenvalue $\omega = e^{i\theta}$:

$$J_{\omega,p} \equiv \begin{bmatrix} \omega & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & \omega \end{bmatrix}. \tag{2.12}$$

When $p = 2m$, let $\Gamma_{k,m}$ be determined through the partition

$$J_{\omega,2m}^{2^k} = \begin{bmatrix} J_{\omega,m}^{2^k} & \Gamma_{k,m} \\ 0 & J_{\omega,m}^{2^k} \end{bmatrix}. \tag{2.13}$$

The following useful Lemma is proved in [27].

Lemma 2.1. *The matrix $\Gamma_{k,m}$ is invertible and satisfies*

$$\left\|\Gamma_{k,m}^{-1} J_{\omega,m}^{2^k}\right\| = O(2^{-k}), \quad \left\|J_{\omega,m}^{2^k}\Gamma_{k,m}^{-1} J_{\omega,m}^{2^k}\right\| = O\left(2^{-k}\right) \quad \textit{as } k \to \infty. \tag{2.14}$$

In the next three sections, we will apply the techniques reviewed in this section to three different nonlinear matrix equations. Although the general approach will be the same, we will need to fully exploit the special properties of each equation. Among other things, the following two issues deserve special attention: (1) Given a nonlinear matrix equation, how should we rewrite it in its equivalent form (2.8)? If possible, we should try to get a form (2.8) that would lead to SDA-2 rather than SDA-1, since SDA-2 is less expensive. (2) How should we choose the matrices J_L and J_M in (2.10)? The matrices must satisfy $J_L J_M = J_M J_L$, and the resulting equation (2.11) and an equation from a similar procedure should be easy to handle together. We will keep these issues in mind when we carry out the convergence analysis for the three equations.

3. A special nonlinear matrix equation.

In this section we consider the nonlinear matrix equation (NME)

$$X + A^T X^{-1} A = Q, \tag{3.1}$$

where $A, Q \in \mathbb{R}^{n\times n}$ with Q being symmetric positive definite. Various aspects of the NME, like solvability, numerical solution, perturbation, and applications, can be found in [8, 9, 13, 17, 22, 35, 38, 39, 40, 41] and the references therein.

For symmetric matrices X and Y, we write $X \geq Y$ $(X > Y)$ if $X - Y$ is positive semidefinite (definite). We use this definition of ordering only in this section, and will use the elementwise order in sections 4 and 5. We assume that (3.1) has a symmetric positive definite solution. Then [9] it has a maximal symmetric positive definite solution X_+ ($X_+ \geq X$ for any symmetric positive definite solution X of (3.1)), and $\rho(X_+^{-1}A) \leq 1$, where $\rho(\cdot)$ is the spectral radius.

Let

$$L_0 = \begin{bmatrix} 0 & I \\ A^T & 0 \end{bmatrix}, \quad M_0 = \begin{bmatrix} A & 0 \\ Q & -I \end{bmatrix}. \tag{3.2}$$

It is easy to verify that the pencil $M_0 - \lambda L_0$ (also denoted by (M_0, L_0)) is symplectic, i.e.,

$$M_0 J M_0^T = L_0 J L_0^T \;\text{ for }\; J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

Using Algorithm 2.2 with $V_0 = A, T_0 = A^T, Q_0 = Q, P_0 = 0$, we have $T_k = V_k^T, Q_k^T = Q_k, P_k^T = P_k$. So Algorithm 2.2 is simplified to the following, where we have used A_k for V_k.

Algorithm 3.1. *Let* $A_0 = A, Q_0 = Q, P_0 = 0.$
For $k = 0, 1, \ldots,$ *compute*

$$\begin{aligned} A_{k+1} &= A_k(Q_k - P_k)^{-1}A_k, \\ Q_{k+1} &= Q_k - A_k^T(Q_k - P_k)^{-1}A_k, \\ P_{k+1} &= P_k + A_k(Q_k - P_k)^{-1}A_k^T. \end{aligned}$$

The matrices L_k, M_k in (2.7) are now given by

$$L_k = \begin{bmatrix} -P_k & I \\ A_k^T & 0 \end{bmatrix}, \quad M_k = \begin{bmatrix} A_k & 0 \\ Q_k & -I \end{bmatrix}. \tag{3.3}$$

It is noted in [34] that the cyclic reduction algorithm in [35] is recovered from Algorithm 3.1 when $Q_k - P_k$ and Q_k are replaced by Q_k and X_k, respectively, where the latter Q_k and X_k are the notations used in [35, Algorithm 3.1]. So we know from [35] that $Q_k - P_k > 0$ in Algorithm 3.1. Thus, the algorithm is well defined and $0 \le P_k < Q_k \le Q$. This fact is also proved in [34] without using the results in [35].

It is easy to verify that

$$M_0 \begin{bmatrix} I \\ X_+ \end{bmatrix} = L_0 \begin{bmatrix} I \\ X_+ \end{bmatrix} X_+^{-1}A. \tag{3.4}$$

We are interested in the case with $\rho(X_+^{-1}A) = 1$. It follows from [13, Theorem 2.4] that the eigenvalues of $X_+^{-1}A$ have the following characterization.

Theorem 3.1. *For* (3.1), *the eigenvalues of the matrix* $X_+^{-1}A$ *are precisely the eigenvalues of the matrix pencil* $M_0 - \lambda L_0$ *inside or on the unit circle, with half of the* (*necessarily even*) *partial multiplicities for each unimodular eigenvalue of the pencil.*

In view of the connection between Algorithm 3.1 and the cyclic reduction algorithm in [35], we know from [13] that the sequence Q_k in Algorithm 3.1 converges to X_+ at least linearly with rate 1/2, as long as all eigenvalues of $X_+^{-1}A$ on the unit circle are semisimple. With the tools in section 2, we are going to prove more convergence results for Algorithm 3.1, without any assumption on the unimodular eigenvalues of $X_+^{-1}A$.

Suppose there are r Jordan blocks associated with unimodular eigenvalues of (M_0, L_0). Then they have the form

$$J_{\omega_j,2m_j} = \begin{bmatrix} J_{\omega_j,m_j} & \Gamma_{0,m_j} \\ 0 & J_{\omega_j,m_j} \end{bmatrix}, \quad \Gamma_{0,m_j} \equiv e_{m_j}e_1^T, \tag{3.5}$$

where $\omega_j = e^{i\theta_j}$ for $j = 1, \ldots, r$.

By the results on Kronecker canonical form for a symplectic pencil (see [11] and [33]), there exist nonsingular matrices V and Z such that

$$VL_0Z = \begin{bmatrix} I_n & 0_n \\ 0_n & J_s^H \oplus I_m \end{bmatrix} \equiv J_L, \tag{3.6}$$

$$VM_0Z = \begin{bmatrix} J_s \oplus J_1 & 0_l \oplus \Gamma_0 \\ 0_n & I_l \oplus J_1 \end{bmatrix} \equiv J_M, \tag{3.7}$$

where $J_s \in \mathbb{C}^{l \times l}$ consists of stable Jordan blocks (so $\rho(J_s) < 1$), $J_1 = J_{\omega_1,m_1} \oplus \cdots \oplus J_{\omega_r,m_r}$, $\Gamma_0 \equiv \Gamma_{0,m_1} \oplus \cdots \oplus \Gamma_{0,m_r}$, $m = m_1 + \cdots + m_r$, $l = n - m$, $\oplus$ denotes the direct sum of matrices and $(\cdot)^H$ the conjugate transpose. Moreover, the nonsingular matrix Z can be taken to be of the form $Z = Z_a Z_b$ with Z_a symplectic and $Z_b = I_n \oplus Z_c$. It follows that span$\{Z(:, 1 : n)\}$ forms the unique weakly stable Lagrangian deflating subspace of (M_0, L_0) corresponding to $J_s \oplus J_1$.

Let Γ_{k,m_j} be given by (2.13) with $\omega = \omega_j$ and $m = m_j$. Since $J_L J_M = J_M J_L$, we have by (2.11)

$$M_k Z \begin{bmatrix} I & 0 \\ 0 & \left(J_s^H\right)^{2^k} \oplus I \end{bmatrix} = L_k Z \begin{bmatrix} J_s^{2^k} \oplus J_1^{2^k} & 0 \oplus \Gamma_k \\ 0 & I \oplus J_1^{2^k} \end{bmatrix}, \tag{3.8}$$

where $\Gamma_k = \Gamma_{k,m_1} \oplus \cdots \oplus \Gamma_{k,m_r}$.

Similarly, there exist nonsingular matrices T and W such that

$$TM_0W = J_L, \quad TL_0W = J_M, \tag{3.9}$$

and

$$L_k W \begin{bmatrix} I & 0 \\ 0 & \left(J_s^H\right)^{2^k} \oplus I \end{bmatrix} = M_k W \begin{bmatrix} J_s^{2^k} \oplus J_1^{2^k} & 0 \oplus \Gamma_k \\ 0 & I \oplus J_1^{2^k} \end{bmatrix}. \tag{3.10}$$

By Lemma 2.1 we have

$$\left\|\Gamma_k^{-1} J_1^{2^k}\right\| = O\left(2^{-k}\right), \quad \left\|J_1^{2^k} \Gamma_k^{-1} J_1^{2^k}\right\| = O\left(2^{-k}\right) \quad \text{as } k \to \infty. \tag{3.11}$$

We now prove some convergence results for Algorithm 3.1. Partition Z and W as

$$Z = \begin{bmatrix} Z_1 & Z_3 \\ Z_2 & Z_4 \end{bmatrix}, \quad W = \begin{bmatrix} W_1 & W_3 \\ W_2 & W_4 \end{bmatrix}, \tag{3.12}$$

where $Z_i, W_i \in \mathbb{C}^{n\times n}$ $(i = 1, \ldots, 4)$.

THEOREM 3.2. *When $\rho(X_+^{-1}A) = 1$, the sequences $\{A_k, Q_k, P_k\}$ generated by Algorithm* 3.1 *satisfy*

(a) $\|A_k\| = O(2^{-k})$;

(b) $\|Q_k - X_+\| = O(2^{-k})$ *and* $X_+ = Z_2 Z_1^{-1}$;

(c) $\|P_k - X_-\| = O(2^{-k})$ *for* $X_- = W_2 W_1^{-1}$ *if* W_1 *is invertible; if* A *is also invertible, then* X_- *is a solution of* (3.1) *and the eigenvalues of* $X_-^{-1}A$ *are the reciprocals of the eigenvalues of* $X_+^{-1}A$;

(d) $Q_k - P_k$ *converges to a singular matrix as* $k \to \infty$.

Proof. (a) Substituting L_k and M_k of (3.3) and Z of (3.12) into (3.8), we obtain

$$A_k Z_1 = (-P_k Z_1 + Z_2)\left(J_s^{2^k} \oplus J_1^{2^k}\right), \tag{3.13}$$

$$A_k Z_3 \left(\left(J_s^H\right)^{2^k} \oplus I\right) = (-P_k Z_1 + Z_2)(0 \oplus \Gamma_k) + (-P_k Z_3 + Z_4)\left(I \oplus J_1^{2^k}\right), \tag{3.14}$$

$$Q_k Z_1 - Z_2 = A_k^T Z_1 \left(J_s^{2^k} \oplus J_1^{2^k}\right), \tag{3.15}$$

$$(Q_k Z_3 - Z_4)\left(\left(J_s^H\right)^{2^k} \oplus I\right) = A_k^T Z_1 (0 \oplus \Gamma_k) + A_k^T Z_3 \left(I \oplus J_1^{2^k}\right). \tag{3.16}$$

From (3.6) and (3.7) we have

$$M_0 \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = L_0 \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} (J_s \oplus J_1).$$

By Theorem 3.1, $X_+^{-1}A$ is similar to $J_s \oplus J_1$. Then from (3.4) and the uniqueness of weakly stable Lagrangian deflating subspaces of (M_0, L_0) corresponding to $J_s \oplus J_1$, we have

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} I \\ X_+ \end{bmatrix} R$$

for a nonsingular matrix R. It follows that Z_1^{-1} exists and $X_+ = Z_2 Z_1^{-1}$.

Postmultiplying (3.14) by $(0 \oplus \Gamma_k^{-1} J_1^{2^k}) Z_1^{-1}$ and using (3.13), we have

$$\begin{aligned} &A_k \left[I - Z_3 \left(0 \oplus \Gamma_k^{-1} J_1^{2^k} \right) Z_1^{-1} \right] \\ &= (-P_k Z_1 + Z_2) \left(J_s^{2^k} \oplus 0 \right) Z_1^{-1} - (-P_k Z_3 + Z_4) \left(0 \oplus J_1^{2^k} \Gamma_k^{-1} J_1^{2^k} \right) Z_1^{-1}. \end{aligned}$$

It follows from (3.11) and the boundedness of $\{P_k\}$ that

$$\|A_k\| = O\left(2^{-k}\right). \tag{3.17}$$

(b) Postmultiplying (3.16) by $(0 \oplus \Gamma_k^{-1} J_1^{2^k}) Z_1^{-1}$ and using (3.15), we get

$$\begin{aligned} &Q_k \left[I - Z_3 \left(0 \oplus \Gamma_k^{-1} J_1^{2^k} \right) Z_1^{-1} \right] - X_+ \\ &= \left[A_k^T Z_1 \left(J_s^{2^k} \oplus 0 \right) - A_k^T Z_3 \left(0 \oplus J_1^{2^k} \Gamma_k^{-1} J_1^{2^k} \right) - Z_4 \left(0 \oplus \Gamma_k^{-1} J_1^{2^k} \right) \right] Z_1^{-1}. \end{aligned} \tag{3.18}$$

By (3.11) and (3.17), we have

$$\|Q_k - X_+\| = O\left(2^{-k}\right).$$

(c) Substituting L_k and M_k of (3.3) and W of (3.12) into (3.10), we have

$$W_2 - P_k W_1 = A_k W_1 \left(J_s^{2^k} \oplus J_1^{2^k} \right), \tag{3.19}$$

$$(W_4 - P_k W_3) \left(\left(J_s^H \right)^{2^k} \oplus I \right) = A_k W_1 (0 \oplus \Gamma_k) + A_k W_3 \left(I \oplus J_1^{2^k} \right). \tag{3.20}$$

Let $X_- = W_2 W_1^{-1}$. As before, postmultiplying (3.20) by $(0 \oplus \Gamma_k^{-1} J_1^{2^k}) W_1^{-1}$ and using (3.19), we get

$$\begin{aligned} &X_- - P_k \left[I - W_3 \left(0 \oplus \Gamma_k^{-1} J_1^{2^k} \right) W_1^{-1} \right] \\ &= \left[W_4 \left(0 \oplus \Gamma_k^{-1} J_1^{2^k} \right) + A_k W_1 \left(J_s^{2^k} \oplus 0 \right) - A_k W_3 \left(0 \oplus J_1^{2^k} \Gamma_k^{-1} J_1^{2^k} \right) \right] W_1^{-1}. \end{aligned} \tag{3.21}$$

By (3.11) and the result of (a), we have

$$\|X_- - P_k\| = O\left(2^{-k}\right).$$

From (3.9) we get

$$\begin{bmatrix} 0 & I \\ A^T & 0 \end{bmatrix} \begin{bmatrix} I \\ X_- \end{bmatrix} = \begin{bmatrix} A & 0 \\ Q & -I \end{bmatrix} \begin{bmatrix} I \\ X_- \end{bmatrix} R_-,$$

where $R_- = W_1 (J_s \oplus J_1) W_1^{-1}$. It follows that

$$X_- = A R_-, \quad A^T = (Q - X_-) R_-.$$

When A is invertible, the matrices $X_+^{-1}A, R_-, X_-$ are all invertible and we obtain

$$X_- + A^T X_-^{-1} A = Q.$$

Moreover, the eigenvalues of $X_-^{-1}A$ are the reciprocals of the eigenvalues of R_- (and, thus, $X_+^{-1}A$).

(d) From (3.13) and (3.15), we get

$$\begin{aligned} -P_k Z_1 \left(J_s^{2^k} \oplus J_1^{2^k}\right) &= A_k Z_1 - Z_2 \left(J_s^{2^k} \oplus J_1^{2^k}\right), \\ Q_k Z_1 \left(J_s^{2^k} \oplus J_1^{2^k}\right) &= Z_2 \left(J_s^{2^k} \oplus J_1^{2^k}\right) + A_k^T Z_1 \left(J_s^{2\cdot 2^k} \oplus J_1^{2\cdot 2^k}\right). \end{aligned}$$

This implies that

$$(Q_k - P_k) Z_1 \begin{bmatrix} 0 \\ I_m \end{bmatrix} = A_k Z_1 \begin{bmatrix} 0 \\ J_1^{-2^k} \end{bmatrix} + A_k^T Z_1 \begin{bmatrix} 0 \\ J_1^{2^k} \end{bmatrix}. \tag{3.22}$$

Since $0 \le P_k \le P_{k+1} \le Q$, the sequence P_k converges even if W_1 is singular. Let $\lim(Q_k - P_k) = R_*$. It follows from (3.22) and the result of (a) that

$$R_* Z_1 \begin{bmatrix} 0 \\ I_m \end{bmatrix} = 0.$$

Thus, R_* is singular. □

The most important conclusion in Theorem 3.2 is that the sequence Q_k from the doubling algorithm converges to X_+ at least linearly with rate 1/2, regardless of the values of m_j $(j = 1, 2, \ldots, r)$. This is in sharp contrast with the behavior of Newton's method. The NME (3.1) is a special case of the discrete algebraic Riccati equation studied in [12]. It is conjectured in [12] that the convergence of Newton's method is linear with rate $1/\sqrt[q]{2}$, where $q = \max_{1 \le j \le r} m_j$. This conjecture is confirmed in numerical experiments on (3.1) with A being a $q \times q$ Jordan block with eigenvalue 1 and $Q = I + A^T A$, for small values of q. We know from [13] that $X_+ = I$ in all those examples. Newton's method is given in [13, Algorithm 3.3].

4. A quadratic matrix equation from quasi-birth-death problems. A discrete-time quasi-birth-death (QBD) process is a Markov chain with state space $\{(i,j) \,|\, i \ge 0, 1 \le j \le n\}$, and with a transition probability matrix of the form

$$P = \begin{bmatrix} B_0 & B_1 & 0 & 0 & \cdots \\ A_0 & A_1 & A_2 & 0 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where B_0, B_1, A_0, A_1, and A_2 are $n \times n$ nonnegative matrices such that P is stochastic. In particular, $(A_0 + A_1 + A_2)e = e$, where $e = (1, 1, \ldots, 1)^T$.

We make the standard assumption that the matrix P and the matrix $A = A_0 + A_1 + A_2$ are both irreducible. Thus, $A_0 \ne 0$ and $A_2 \ne 0$. Moreover, there exists a unique positive vector α with $\alpha^T e = 1$ and $\alpha^T A = \alpha^T$. The QBD is positive recurrent if $\alpha^T A_0 e > \alpha^T A_2 e$, transient 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

$$G = A_0 + A_1G + A_2G^2 \tag{4.1}$$

plays an important role in the study of the QBD process (see [32]). We will also need the dual equation

$$F = A_2 + A_1F + A_0F^2, \tag{4.2}$$

and we let F be its minimal nonnegative solution. It is well known (see [32], 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 transient, then F is stochastic and G is substochastic with $\rho(G) < 1$; if the QBD is null recurrent, then G and F are both stochastic.

The Latouche–Ramaswami (LR) algorithm [31] and the cyclic reduction (CR) algorithm [5] are both efficient iterative methods for finding the minimal solution G. The convergence of these two algorithms is quadratic for positive recurrent and transient QBDs. A convergence analysis has been performed in [15] for the LR algorithm in the null recurrent case under two additional assumptions. The first assumption is that $\lambda = 1$ is a simple eigenvalue of G and F and there are no other eigenvalues of G or F on the unit circle; the second assumption is made under the first assumption and is more technical. The convergence rate for the LR algorithm is the same in view of the relationship between CR and LR, given in [3].

We can also use the doubling algorithm (SDA-1 or SDA-2) to find the minimal solution G. We will choose to use SDA-2 since it is less expensive. Moreover, there is a close connection between the CR algorithm and SDA-2. In this section we determine the convergence rate of SDA-2 in the null recurrent case, without the two additional assumptions in [15]. The convergence rate for the CR (or LR) algorithm in the null recurrent case is the same in view of their connections to SDA-2. As compared to [15], the result here is more general and the analysis here is much simpler.

We mention that a doubling algorithm is also derived in [26] for finding the minimal nonnegative solution of a polynomial equation that is more general than (4.1). The algorithm there is different from SDA-2 when applied to (4.1).

The CR algorithm for (4.1), or for $-A_0 + (I - A_1)G - A_2G^2 = 0$, is the following:

Algorithm 4.1. *Set* $T_0 = A_0,\quad U_0 = I - A_1,\quad V_0 = A_2,\quad S_0 = I - A_1.$ *For* $k = 0, 1, \ldots,$ *compute*

$$\begin{aligned}
T_{k+1} &= T_kU_k^{-1}T_k,\\
U_{k+1} &= U_k - T_kU_k^{-1}V_k - V_kU_k^{-1}T_k,\\
V_{k+1} &= V_kU_k^{-1}V_k,\\
S_{k+1} &= S_k - V_kU_k^{-1}T_k.
\end{aligned}$$

The above CR algorithm is as presented in [3], but with one minor change: if we follow [3] exactly, T_k and V_k here would have to be replaced by $-T_k$ and $-V_k$ for $k \geq 0$.

The following result is known from the discussions in [4] and [32].

Theorem 4.1. *The sequences* $\{T_k\}, \{U_k\}, \{V_k\}, \{S_k\}$ *in Algorithm* 4.1 *are well defined. For each* $k \geq 0$, T_k *and* V_k *are nonnegative, and* U_k*and* S_k *are nonsingular* M*-matrices. When the QBD is positive recurrent or transient, the sequence* $\{S_k\}$ *converges quadratically to a nonsingular* M*-matrix* S_* *and* $S_*^{-1}A_0 = G$.

We note that Algorithm 4.1 may break down if we do not assume the irreducibility of the transition matrix P. As an example, we consider

$$A_0 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad A_1 = 0, \quad A_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

It is easy to see that P is not irreducible, although $A_0 + A_1 + A_2$ is. For this example, $U_1 = 0$ in Algorithm 4.1, so the algorithm breaks down. The LR algorithm also breaks down for this example.

To use the doubling algorithm to find G, we may rewrite (4.1) as

$$\begin{bmatrix} 0 & I \\ A_0 & A_1 - I \end{bmatrix} \begin{bmatrix} I \\ G \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & -A_2 \end{bmatrix} \begin{bmatrix} I \\ G \end{bmatrix} G.$$

Multiplying the second block row by $-(I - A_1)^{-1}$ and eliminating the I in the (1, 2) block of the leftmost matrix, we get

$$\begin{bmatrix} (I - A_1)^{-1}A_0 & 0 \\ -(I - A_1)^{-1}A_0 & I \end{bmatrix} \begin{bmatrix} I \\ G \end{bmatrix} = \begin{bmatrix} I & -(I - A_1)^{-1}A_2 \\ 0 & (I - A_1)^{-1}A_2 \end{bmatrix} \begin{bmatrix} I \\ G \end{bmatrix} G.$$

We can then use SDA-1 to find the matrix G. However, the less expensive SDA-2 can also be used if we rewrite (4.1) as

$$L_0 \begin{bmatrix} I \\ A_2G \end{bmatrix} = M_0 \begin{bmatrix} I \\ A_2G \end{bmatrix} G, \tag{4.3}$$

where

$$L_0 = \begin{bmatrix} 0 & I \\ A_0 & 0 \end{bmatrix}, \quad M_0 = \begin{bmatrix} A_2 & 0 \\ I - A_1 & -I \end{bmatrix}.$$

It is easily seen that $L_0 - \lambda M_0$ is a linearization of $-A_0 + \lambda(I - A_1) - \lambda^2 A_2$.

If we use SDA-1, the matrix G can be approximated directly by a sequence generated by SDA-1. One may have some concern about the SDA-2 approach: how can one get G if A_2G is obtained and A_2 is singular? This concern will turn out to be unnecessary.

In this section SDA-2 is Algorithm 2.2 with the initialization

$$T_0 = A_0, \quad Q_0 = I - A_1, \quad P_0 = 0, \quad V_0 = A_2. \tag{4.4}$$

The algorithm generates the sequence $\{L_k, M_k\}$ (see (2.7)) if no breakdown occurs.

It is readily seen that Algorithm 4.1 is recovered from SDA-2 by letting $U_k = Q_k - P_k$ and $S_k = S_0 - P_k$. By Theorem 4.1, $Q_k - P_k = U_k$ are nonsingular M-matrices for all $k \geq 0$. So SDA-2 is also well defined.

In view of (2.9) we have for each $k \geq 0$

$$L_k \begin{bmatrix} I \\ A_2G \end{bmatrix} = M_k \begin{bmatrix} I \\ A_2G \end{bmatrix} G^{2^k}.$$

So

$$-P_k + A_2G = V_kG^{2^k}, \quad T_k = Q_kG^{2^k} - A_2G^{2^k+1}. \tag{4.5}$$

Similarly we have

$$\widehat{L}_0 \begin{bmatrix} I \\ A_0F \end{bmatrix} = \widehat{M}_0 \begin{bmatrix} I \\ A_0F \end{bmatrix} F,$$

where

$$\widehat{L}_0 = \begin{bmatrix} 0 & I \\ V_0 & 0 \end{bmatrix}, \quad \widehat{M}_0 = \begin{bmatrix} T_0 & 0 \\ Q_0 & -I \end{bmatrix}.$$

It is easily seen that $\widehat{M}_0 - \lambda\widehat{L}_0$ is also a linearization of $-A_0 + \lambda(I - A_1) - \lambda^2 A_2$.

For each $k \geq 0$ we now have

$$\widehat{L}_k \begin{bmatrix} I \\ A_0F \end{bmatrix} = \widehat{M}_k \begin{bmatrix} I \\ A_0F \end{bmatrix} F^{2^k},$$

where

$$\widehat{L}_k = \begin{bmatrix} -\widehat{P}_k & I \\ V_k & 0 \end{bmatrix}, \quad \widehat{M}_k = \begin{bmatrix} T_k & 0 \\ \widehat{Q}_k & -I \end{bmatrix}$$

with

$$\widehat{P}_k = I - A_1 - Q_k, \quad \widehat{Q}_k = I - A_1 - P_k. \tag{4.6}$$

So

$$-\widehat{P}_k + A_0F = T_kF^{2^k}, \quad V_k = \widehat{Q}_kF^{2^k} - A_0F^{2^k+1}. \tag{4.7}$$

We mentioned before that the S_k in Algorithm 4.1 satisfies $S_k = S_0 - P_k = I - A_1 - P_k$. So we have $\widehat{Q}_k = S_k$.

When the QBD is positive recurrent or transient, we know by Theorem 4.1 that $\widehat{Q}_k$ converges quadratically to a nonsingular M-matrix $\widehat{Q}_*$ and $\widehat{Q}_*^{-1}A_0 = G$. Here we give a quick proof using the doubling algorithm. By the first equation in (4.5) and the second equation in (4.6), we have

$$\widehat{Q}_k - I + A_1 + A_2G = V_kG^{2^k}.$$

Eliminating V_k using the second equation in (4.7) gives

$$\widehat{Q}_k\left(I - F^{2^k}G^{2^k}\right) = I - A_1 - A_2G - A_0F^{2^k+1}G^{2^k}.$$

It follows that

$$\limsup_{k\to\infty} \sqrt[2^k]{\left\|\widehat{Q}_k - (I - A_1 - A_2G)\right\|} \leq \rho(F)\rho(G) < 1.$$

Since $\widehat{Q}_* = I - A_1 - A_2G$ is a nonsingular M-matrix and $A_0 = \widehat{Q}_*G$, we have $G = \widehat{Q}_*^{-1}A_0$. Similarly, Q_k converges quadratically to the nonsingular M-matrix $Q_* = I - A_1 - A_0F$ and $F = Q_*^{-1}A_2$.

Our main purpose of this section, however, is to determine the convergence rate of SDA-2 for the null recurrent case.

We start with a review of an important result about the spectral properties of the quadratic pencil $-A_0 + \lambda(I - A_1) - \lambda^2A_2$ and of the matrices G and F when the QBD is null recurrent. See Proposition 14 and Theorem 4 of [10] and Theorem 4.10 of [4].

Theorem 4.2. *Let the QBD be null recurrent. Then*

(a) *For some integer $r \geq 1$ the quadratic pencil $-A_0 + \lambda(I - A_1) - \lambda^2 A_2$ has $n - r$ eigenvalues inside the unit circle, $n - r$ eigenvalues outside the unit circle (which include eigenvalues at infinity), and $2r$ eigenvalues on the unit circle, which are the rth roots of unity, each with multiplicity two.*

(b) *The partial multiplicity of each eigenvalue on the unit circle is exactly two.*

(c) *The eigenvalues of G are the $n - r$ eigenvalues of the pencil inside the unit circle plus the r simple eigenvalues at the rth roots of unity, the eigenvalues of F are the reciprocals of the $n - r$ eigenvalues of the pencil outside the unit circle plus the r simple eigenvalues at the rth roots of unity.*

Using the Kronecker form for matrix pairs, we have nonsingular matrices V and Z such that

$$VM_0Z = \begin{bmatrix} I_n & 0 \\ 0 & J_2 \oplus I_r \end{bmatrix} \equiv J_M, \tag{4.8}$$

$$VL_0Z = \begin{bmatrix} J_1 \oplus D_r & 0 \oplus I_r \\ 0 & I_{n-r} \oplus D_r \end{bmatrix} \equiv J_L, \tag{4.9}$$

where J_1 and J_2 are $(n-r) \times (n-r)$ matrices consisting of the Jordan blocks with diagonal elements inside the unit circle, and D_r is a $r \times r$ diagonal matrix with the rth roots of unity on the diagonal.

Similarly, we have nonsingular matrices T and W such that

$$T\widehat{L}_0W = \begin{bmatrix} I_n & 0 \\ 0 & J_2 \oplus I_r \end{bmatrix} = J_M, \tag{4.10}$$

$$T\widehat{M}_0W = \begin{bmatrix} J_1 \oplus D_r & 0 \\ 0 \oplus I_r & I_{n-r} \oplus D_r \end{bmatrix} \equiv \widehat{J}_L. \tag{4.11}$$

We have for each $k \geq 0$

$$M_kZJ_L^{2^k} = L_kZJ_M^{2^k}, \quad \widehat{L}_kW\widehat{J}_L^{2^k} = \widehat{M}_kWJ_M^{2^k}. \tag{4.12}$$

Let Z and W be partitioned as in (3.12). From (4.8) and (4.9) we have

$$L_0 \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = M_0 \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} (J_1 \oplus D_r).$$

Comparing this with (4.3) and using Theorem 4.2, we know that Z_1 is nonsingular and $Z_2Z_1^{-1} = A_2G$. Similarly, W_3 is nonsingular and $W_4W_3^{-1} = A_0F$.

Using block matrix multiplication for (4.12), we have

$$V_kZ_1\left(J_1^{2^k} \oplus D_r^{2^k}\right) = -P_kZ_1 + Z_2, \tag{4.13}$$

$$(Q_kZ_1 - Z_2)\left(J_1^{2^k} \oplus D_r^{2^k}\right) = T_kZ_1, \tag{4.14}$$

$$V_kZ_1\left(0 \oplus 2^kD_r^{2^k-1}\right) + V_kZ_3\left(I \oplus D_r^{2^k}\right) = (-P_kZ_3 + Z_4)\left(J_2^{2^k} \oplus I\right), \tag{4.15}$$

$$(Q_kZ_1 - Z_2)\left(0 \oplus 2^kD_r^{2^k-1}\right) + (Q_kZ_3 - Z_4)\left(I \oplus D_r^{2^k}\right) = T_kZ_3\left(J_2^{2^k} \oplus I\right), \tag{4.16}$$

$$\left(-\widehat{P}_kW_1 + W_2\right)\left(J_1^{2^k} \oplus D_r^{2^k}\right) + \left(-\widehat{P}_kW_3 + W_4\right)\left(0 \oplus 2^kD_r^{2^k-1}\right) = T_kW_1, \tag{4.17}$$

$$V_kW_1\left(J_1^{2^k} \oplus D_r^{2^k}\right) + V_kW_3\left(0 \oplus 2^kD_r^{2^k-1}\right) = \widehat{Q}_kW_1 - W_2, \tag{4.18}$$

$$\left(-\widehat{P}_k W_3 + W_4)(I \oplus D_r^{2^k}\right) = T_k W_3 \left(J_2^{2^k} \oplus I\right), \tag{4.19}$$

$$V_k W_3 \left(I \oplus D_r^{2^k}\right) = \left(\widehat{Q}_k W_3 - W_4\right) \left(J_2^{2^k} \oplus I\right). \tag{4.20}$$

Postmultiplying (4.16) by $0 \oplus 2^{-k} D_r$ and subtracting the result from (4.14), we get

$$T_k \left(Z_1 - Z_3 \left(0 \oplus 2^{-k} D_r\right)\right) = \left(Q_k Z_1 - Z_2\right) \left(J_1^{2^k} \oplus 0\right) - \left(Q_k Z_3 - Z_4\right) \left(0 \oplus 2^{-k} D_r^{2^k+1}\right). \tag{4.21}$$

By (4.19) we have

$$-\widehat{P}_k = -W_4 W_3^{-1} + T_k W_3 \left(J_2^{2^k} \oplus D_r^{-2^k}\right) W_3^{-1}. \tag{4.22}$$

Thus, in view of (4.6),

$$Q_k = I - A_1 - W_4 W_3^{-1} + T_k W_3 \left(J_2^{2^k} \oplus D_r^{-2^k}\right) W_3^{-1}. \tag{4.23}$$

Inserting (4.23) into (4.21) and letting $Q_* = I - A_1 - W_4 W_3^{-1}$, we get

$$\begin{aligned} T_k\Big[Z_1 - Z_3 \left(0 \oplus 2^{-k} D_r\right) - W_3\left(J_2^{2^k} \oplus D_r^{-2^k}\right) W_3^{-1}\left(Z_1 \left(J_1^{2^k} \oplus 0\right) - Z_3\left(0 \oplus 2^{-k} D_r^{2^k+1}\right)\right)\Big] \\ = \left(Q_* Z_1 - Z_2\right) \left(J_1^{2^k} \oplus 0\right) - \left(Q_* Z_3 - Z_4\right) \left(0 \oplus 2^{-k} D_r^{2^k+1}\right), \end{aligned}$$

from which it follows that

$$\|T_k\| = O\left(2^{-k}\right).$$

It then follows from (4.23) that

$$\left\|Q_k - \left(I - A_1 - W_4 W_3^{-1}\right)\right\| = O\left(2^{-k}\right).$$

Postmultiplying (4.15) by $0 \oplus 2^{-k} D_r$ and subtracting the result from (4.13), we get

$$-P_k Z_1 + Z_2 - \left(-P_k Z_3 + Z_4\right) \left(0 \oplus 2^{-k} D_r\right) = V_k \left(Z_1 \left(J_1^{2^k} \oplus 0\right) - Z_3 \left(0 \oplus 2^{-k} D_r^{2^k+1}\right)\right). \tag{4.24}$$

By (4.20),

$$V_k = \left(\widehat{Q}_k W_3 - W_4\right) \left(J_2^{2^k} \oplus D_r^{-2^k}\right) W_3^{-1}. \tag{4.25}$$

Inserting (4.25) into (4.24) and using $\widehat{Q}_k = I - A_1 - P_k$, we get

$$-P_k Z_1 + Z_2 - \left(-P_k Z_3 + Z_4\right) \left(0 \oplus 2^{-k} D_r\right) = \left((I - A_1 - P_k) W_3 - W_4\right) C_k$$

for some C_k with $\|C_k\| = O(2^{-k})$. Thus,

$$P_k \left(Z_1 - Z_3 \left(0 \oplus 2^{-k} D_r\right) - W_3 C_k\right) = Z_2 - Z_4 \left(0 \oplus 2^{-k} D_r\right) - \left((I - A_1) W_3 - W_4\right) C_k.$$

It follows that

$$\left\|P_k - Z_2 Z_1^{-1}\right\| = O\left(2^{-k}\right).$$

Postmultiplying (4.18) by $0 \oplus 2^{-k} D_r^{1-2^k}$, we get

$$V_k W_1 \left(0 \oplus 2^{-k} D_r\right) + V_k W_3 (0 \oplus I) = \left(\widehat{Q}_k W_1 - W_2\right)\left(0 \oplus 2^{-k} D_r^{1-2^k}\right). \tag{4.26}$$

Postmultiplying (4.20) by $I \oplus 0$, we get

$$V_k W_3 (I \oplus 0) = \left(\widehat{Q}_k W_3 - W_4\right)\left(J_2^{2^k} \oplus 0\right). \tag{4.27}$$

Adding (4.26) and (4.27) gives

$$\begin{aligned} &V_k \left(W_3 + W_1 \left(0 \oplus 2^{-k} D_r\right)\right) \\ &\qquad = \left(\widehat{Q}_k W_1 - W_2\right)\left(0 \oplus 2^{-k} D_r^{1-2^k}\right) + \left(\widehat{Q}_k W_3 - W_4\right)\left(J_2^{2^k} \oplus 0\right). \end{aligned}$$

It follows that

$$\|V_k\| = O\left(2^{-k}\right),$$

since W_3 is nonsingular and $\{\widehat{Q}_k\}$ has been shown to be bounded.

In summary, we have proved the following result.

THEOREM 4.3. *Let the QBD be null recurrent. Then for SDA-2sf we have*

$$\begin{aligned} &\|V_k\| = O\left(2^{-k}\right), \quad \|T_k\| = O\left(2^{-k}\right), \\ &\|Q_k - (I - A_1 - A_0 F)\| = O\left(2^{-k}\right), \quad \|P_k - A_2 G\| = O\left(2^{-k}\right). \end{aligned}$$

COROLLARY 4.4. *Let* $\lim Q_k = Q_*$ *and* $\lim P_k = P_*$. *Then* Q_* *is nonsingular and* $Q_*^{-1} A_2 = F$, $I - A_1 - P_*$ *is nonsingular and* $(I - A_1 - P_*)^{-1} A_0 = G$. *The matrix* $Q_* - P_*$ *is a singular* M*-matrix.*

Proof. By Theorem 4.3, $Q_* = I - A_1 - A_0 F$ and $I - A_1 - P_* = I - A_1 - A_2 G$. These two matrices are known to be nonsingular [32]. Since $Q_* F = (I - A_1 - A_0 F)F = A_2$, $Q_*^{-1} A_2 = F$. Since $(I - A_1 - P_*)G = (I - A_1 - A_2 G)G = A_0$, $(I - A_1 - P_*)^{-1} A_0 = G$. $Q_* - P_*$ is a singular M-matrix since

$$(Q_* - P_*)e = (I - A_1 - A_0 F - A_2 G)e = e - (A_1 + A_0 + A_2)e = 0.$$

This completes the proof. □

When the QBD is null recurrent, the interpretation of the CR algorithm as a doubling algorithm has allowed us to show that the minimal solutions G and F can be found by the CR algorithm (or the closely related LR algorithm) simultaneously and with at least linear convergence with rate 1/2. It is important to note that we no longer need the assumption that the matrices G and F have no eigenvalues on the unit circle other than the simple eigenvalue 1. With that assumption, one would use the shift technique as studied in [25], [16], and [4], and apply the CR algorithm or the LR algorithm to the shifted equation. When G and F have more than one eigenvalue on the unit circle, the shift technique is not helpful and the CR algorithm or the LR algorithm will be applied directly to the equation (4.1).

5. A nonsymmetric algebraic Riccati equation.

In this section we consider the nonsymmetric algebraic Riccati equation (NARE)

$$XCX - XD - AX + B = 0, \tag{5.1}$$

where A, B, C, D are real matrices of sizes $m \times m, m \times n, n \times m, n \times n$, respectively, and the matrix

$$K = \begin{bmatrix} D & -C \\ -B & A \end{bmatrix} \tag{5.2}$$

is a nonsingular M-matrix or an irreducible singular M-matrix. The NARE arises in the study of Wiener–Hopf factorization of Markov chains [37], and it includes the NARE arising from transport theory [29, 30]. We will also need the dual equation of (5.1)

$$YBY - YA - DY + C = 0, \tag{5.3}$$

which is in the same form of (5.1).

We will use the elementwise order for matrices: for any matrices $A = [a_{ij}], B = [b_{ij}] \in \mathbb{R}^{m\times n}$, we write $A \geq B(A > B)$ if $a_{ij} \geq b_{ij}(a_{ij} > b_{ij})$ for all i, j.

A basic result about (5.1) and (5.3) is the following [14].

Theorem 5.1. *If the matrix K in* (5.2) *is a nonsingular M-matrix or an irreducible singular M-matrix, then the NARE* (5.1) *and the NARE* (5.3) *have minimal nonnegative solutions X and Y, respectively. Moreover, $D - CX$ and $A - BY$ are M-matrices.*

The minimal nonnegative solution of the NARE is the solution of practical interest. There have been a number of methods for finding this solution. The methods and their analyses can be found in [2, 14, 18, 20, 21, 23, 24, 36]. Among the iterative methods, the doubling algorithm proposed in [24] stands out for its overall efficiency. The algorithm is analyzed in [24] for the case when K is a nonsingular M-matrix, and is analyzed in [21] for the case when K is an irreducible singular M-matrix. When K is an irreducible singular M-matrix, we let $[v_1^T, v_2^T]^T > 0$ and $[u_1^T, u_2^T]^T > 0$ be the right and the left null vectors of K in (5.2), respectively. If $u_1^T v_1 \neq u_2^T v_2$, then the convergence of the doubling algorithm is still quadratic; if $u_1^T v_1 = u_2^T v_2$, then the convergence is observed to be linear with rate 1/2 (see [21]). The latter case will be referred to as the critical case for the NARE. For this critical case, the convergence of Newton's method has been shown to at least linear with rate 1/2 [14, 20, 23]. We will reach the same conclusion for the doubling algorithm.

We start with a brief review of the doubling algorithm in [24]. Let

$$H = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix}, \tag{5.4}$$

and

$$R = D - CX, \quad S = A - BY, \tag{5.5}$$

where X and Y are given in Theorem 5.1. Then the NAREs (5.1) and (5.3) can be rewritten as

$$H \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix} R \tag{5.6}$$

and

$$H \begin{bmatrix} Y \\ I_m \end{bmatrix} = \begin{bmatrix} Y \\ I_m \end{bmatrix} (-S). \tag{5.7}$$

Applying the Cayley transform to (5.6) with a scalar $\gamma > 0$ we have

$$(H - \gamma I) \begin{bmatrix} I_n \\ X \end{bmatrix} = (H + \gamma I) \begin{bmatrix} I_n \\ X \end{bmatrix} R_\gamma,$$

where $R_\gamma = (R + \gamma I_n)^{-1}(R - \gamma I_n)$. Premultiplying the above equation by a proper nonsingular matrix gives

$$M_0 \begin{bmatrix} I_n \\ X \end{bmatrix} = L_0 \begin{bmatrix} I_n \\ X \end{bmatrix} R_\gamma. \tag{5.8}$$

Here L_0 and M_0 are given by (2.1) with

$$\begin{aligned} E_0 &= I_n - 2\gamma V_\gamma^{-1}, & F_0 &= I_m - 2\gamma W_\gamma^{-1}, \\ G_0 &= 2\gamma D_\gamma^{-1} C W_\gamma^{-1}, & H_0 &= 2\gamma W_\gamma^{-1} B D_\gamma^{-1}, \end{aligned} \tag{5.9}$$

where

$$\begin{aligned} A_\gamma &= A + \gamma I_m, & D_\gamma &= D + \gamma I_n, \\ W_\gamma &= A_\gamma - B D_\gamma^{-1} C, & V_\gamma &= D_\gamma - C A_\gamma^{-1} B. \end{aligned} \tag{5.10}$$

Similarly,

$$M_0 \begin{bmatrix} Y \\ I_m \end{bmatrix} S_\gamma = L_0 \begin{bmatrix} Y \\ I_m \end{bmatrix}, \tag{5.11}$$

where $S_\gamma = (S + \gamma I_m)^{-1}(S - \gamma I_m)$.

In this section SDA-1 denotes Algorithm 2.1 with E_0, F_0, G_0, H_0 given by (5.9).

The following result from [21] improves the original results given in [24].

Theorem 5.2. *Let the matrix K in* (5.2) *be a nonsingular M-matrix or an irreducible singular M-matrix, and $X, Y \geq 0$ be the minimal nonnegative solutions of the NAREs* (5.1) *and* (5.3), *respectively. If γ satisfies*

$$\gamma \geq \gamma_0 \equiv \max \left\{ \max_{1 \leq i \leq m} a_{ii}, \max_{1 \leq i \leq n} d_{ii} \right\}, \tag{5.12}$$

where a_{ii} and d_{ii} are the diagonal entries of A and D, respectively, then the sequence $\{E_k, F_k, H_k, G_k\}$ in SDA-1 is well defined. Moreover, we have

(a) *$E_0, F_0 < 0$ and $E_k, F_k > 0$ for $k \geq 1$;*

(b) *For $k \geq 0$, $0 \leq H_k < H_{k+1} < X$, $0 \leq G_k < G_{k+1} < Y$;*

(c) *For $k \geq 0$, $I_m - H_k G_k$ and $I_n - G_k H_k$ are nonsingular M-matrices.*

From now on we assume that K in (5.2) is an irreducible singular M-matrix, and consider the critical case of the NARE (5.1). We always assume that γ satisfies (5.12).

The Kronecker form for the pencil (M_0, L_0) can be determined with the help of the following result [14], where $\mathbb{C}_-$ and $\mathbb{C}_+$ denote the open left and the open right half planes, respectively.

Theorem 5.3. *For the critical case of the NARE* (5.1), *the matrix H has $n-1$ eigenvalues in $\mathbb{C}_+$, $m-1$ eigenvalues in $\mathbb{C}_-$, and two zero eigenvalues with a quadratic divisor. Moreover, R and S in* (5.5) *are irreducible singular M-matrices* (*so each of them has a simple eigenvalue* 0 *and the remaining eigenvalues are in $\mathbb{C}_+$*).

In view of Theorem 5.3, the properties of the Cayley transform, and the process leading to (5.8) and (5.11), we know that there are nonsingular matrices V and Z such that

$$VL_0Z = \begin{bmatrix} I_n & 0_{n,m} \\ 0_{m,n} & J_{2,s}\oplus[1] \end{bmatrix} \equiv J_L, \tag{5.13}$$

$$VM_0Z = \begin{bmatrix} J_1 & \Gamma \\ 0_{m,n} & I_{m-1}\oplus[-1] \end{bmatrix} \equiv J_M, \tag{5.14}$$

in which

$$J_1 = J_{1,s}\oplus[-1] \overset{s}{\sim} R_\gamma, \quad J_2 \equiv J_{2,s}\oplus[-1] \overset{s}{\sim} S_\gamma, \quad \Gamma = 0_{n-1,m-1}\oplus[1] \equiv e_n e_m^T, \tag{5.15}$$

where $\rho(J_{1,s}) < 1$, $\rho(J_{2,s}) < 1$, and "$\overset{s}{\sim}$" denotes the similarity transformation. Since $J_L J_M = J_M J_L$, for the matrices L_k and M_k given by (2.2) we have by (2.11)

$$M_k Z J_L^{2^k} = L_k Z J_M^{2^k}. \tag{5.16}$$

On the other hand, there are nonsingular matrices T and W such that

$$TL_0W = \begin{bmatrix} J_2 & \widehat{\Gamma} \\ 0_{n,m} & I_{n-1}\oplus[-1] \end{bmatrix} \equiv \widehat{J}_L, \tag{5.17}$$

$$TM_0W = \begin{bmatrix} I_m & 0_{m,n} \\ 0_{n,m} & J_{1,s}\oplus[1] \end{bmatrix} \equiv \widehat{J}_M, \tag{5.18}$$

where $\widehat{\Gamma} = e_m e_n^T$. We now have

$$L_k W \widehat{J}_M^{2^k} = M_k W \widehat{J}_L^{2^k}. \tag{5.19}$$

The following result determines the convergence rate of SDA-1 in the critical case.

THEOREM 5.4. *Let $X, Y \geq 0$ be the minimal nonnegative solutions of the NAREs* (5.1) *and* (5.3)*, respectively, and let $\{E_k, F_k, G_k, H_k\}$ be generated by SDA-*1*. Then for the critical case*

$$\|E_k\| = O\left(2^{-k}\right), \quad \|F_k\| = O\left(2^{-k}\right), \quad \|H_k - X\| = O\left(2^{-k}\right), \quad \|G_k - Y\| = O\left(2^{-k}\right).$$

Proof. Partition the matrices Z and W as

$$Z = \begin{bmatrix} Z_1 & Z_3 \\ Z_2 & Z_4 \end{bmatrix}, \quad W = \begin{bmatrix} W_1 & W_3 \\ W_2 & W_4 \end{bmatrix}, \tag{5.20}$$

where $Z_1, W_3 \in \mathbb{R}^{n\times n}$, and $Z_4, W_2 \in \mathbb{R}^{m\times m}$. Then from (5.13) and (5.14), and from (5.17) and (5.18), we have

$$M_0 \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = L_0 \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} J_1, \quad M_0 \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} J_2 = L_0 \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}. \tag{5.21}$$

Comparing (5.21) with (5.8) and (5.11), and using (5.15), we know that Z_1 and W_2 are invertible and $X = Z_2 Z_1^{-1}$, $Y = W_1 W_2^{-1}$.

Note that for $k \geq 1$ we have

$$J_L^{2^k} = \begin{bmatrix} I_n & 0 \\ 0 & J_2^{2^k} \end{bmatrix}, \; J_M^{2^k} = \begin{bmatrix} J_1^{2^k} & \Gamma_k \\ 0 & I_m \end{bmatrix}, \; \widehat{J}_M^{2^k} = \begin{bmatrix} I_m & 0 \\ 0 & J_1^{2^k} \end{bmatrix}, \; \widehat{J}_L^{2^k} = \begin{bmatrix} J_2^{2^k} & \widehat{\Gamma}_k \\ 0 & I_n \end{bmatrix},$$

where $\Gamma_k = -2^k\Gamma = -2^k e_n e_m^T$, $\widehat{\Gamma}_k = -2^k\widehat{\Gamma} = -2^k e_m e_n^T$. It follows from (5.16) and (5.19) that for $k \geq 1$

$$E_k Z_1 = (Z_1 - G_k Z_2) J_1^{2^k}, \tag{5.22}$$

$$E_k Z_3 J_2^{2^k} = (Z_1 - G_k Z_2)\Gamma_k + (Z_3 - G_k Z_4), \tag{5.23}$$

$$-H_k Z_1 + Z_2 = F_k Z_2 J_1^{2^k}, \tag{5.24}$$

$$(-H_k Z_3 + Z_4) J_2^{2^k} = F_k Z_2 \Gamma_k + F_k Z_4, \tag{5.25}$$

$$W_1 - G_k W_2 = E_k W_1 J_2^{2^k}, \tag{5.26}$$

$$(W_3 - G_k W_4) J_1^{2^k} = E_k W_1 \widehat{\Gamma}_k + E_k W_3, \tag{5.27}$$

$$F_k W_2 = (W_2 - H_k W_1) J_2^{2^k}, \tag{5.28}$$

$$F_k W_4 J_1^{2^k} = (W_2 - H_k W_1)\widehat{\Gamma}_k + (W_4 - H_k W_3). \tag{5.29}$$

Postmultiplying (5.29) by $\widehat{\Gamma}_k^\dagger = -2^{-k}\Gamma$, the Moore–Penrose pseudo inverse of $\widehat{\Gamma}_k$, subtracting the result from (5.28), and noting that $\widehat{\Gamma}_k \widehat{\Gamma}_k^\dagger = 0_{m-1} \oplus [1]$, we get

$$F_k\left(W_2 + 2^{-k} W_4 J_1^{2^k}\Gamma\right) = (W_2 - H_k W_1)\left(J_{2,s}^{2^k} \oplus [0]\right) + 2^{-k}(W_4 - H_k W_3)\Gamma. \tag{5.30}$$

Since W_2 is invertible and $\{H_k\}$ is bounded by Theorem 5.2(b), it follows from (5.30) that $\|F_k\| = O(2^{-k})$. It then follows from (5.24) that $\|H_k - X\| = O(2^{-k})$.

Similarly, postmultiplying (5.23) by $\Gamma_k^\dagger = -2^{-k}\widehat{\Gamma}$, subtracting the result from (5.22), and noting that $\Gamma_k \Gamma_k^\dagger = 0_{n-1} \oplus [1]$, we get

$$E_k\left(Z_1 + 2^{-k} Z_3 J_2^{2^k}\widehat{\Gamma}\right) = (Z_1 - G_k Z_2)\left(J_{1,s}^{2^k} \oplus [0]\right) + 2^{-k}(Z_3 - G_k Z_4)\widehat{\Gamma}. \tag{5.31}$$

Since Z_1 is invertible and $\{G_k\}$ is bounded by Theorem 5.2(b), it follows from (5.31) that $\|E_k\| = O(2^{-k})$. It then follows from (5.26) that $\|G_k - Y\| = O(2^{-k})$. □

We note that $\lim(I - G_k H_k) = I - YX$ and $\lim(I - H_k G_k) = I - XY$ are both singular M-matrices (see [21]).

The critical case we have considered is a singular case, and the singularity can be removed by applying a proper shift technique. Indeed, a shift technique has been introduced in [21] and SDA-1 applied to the shifted NARE has quadratic convergence if no breakdown happens. However, whether breakdown is possible remains an open problem in general, although some partial results have been obtained in [21].

Since K is an irreducible singular M-matrix, we may assume without loss of generality that $Ke = 0$. In this case, one can transform the NARE to a quadratic matrix equation of the type in section 4, but with $(m + n) \times (m + n)$ matrices in the equation (see [36]). One can then apply CR and LR to the transformed equation (see [2, 18]). A specific shift technique (following [25]) is introduced in [18] to the transformed equation, and quadratic convergence is recovered for the LR algorithm (thus, also for the CR algorithm) if no breakdown happens. It has been shown in [20] that the LR algorithm is indeed well-defined when the shift technique is used. However, when $m = n$, the computational work required in each iteration is nearly twice that for SDA-1, due to the dimension expansion from n to $2n$. If we use the shift technique in [18] with the CR approach in [2], then no breakdown happens and the complexity is down to $34n^3$ flops each iteration when $m = n$.

Although it is preferable to use a shift technique for the critical case of the NARE (with an irreducible singular M-matrix K), our convergence results in Theorem 5.4

still provide some insights about the convergence behavior of SDA-1 for nearby NAREs with a nonsingular M-matrix K (where the shift technique is no longer appplicable). The exact solution of a singular NARE is quite sensitive to the input data in the NARE (see [20]). For the singular NARE and nearby NAREs, it would be reasonable to stop the iteration when $\|H_k - H_{k-1}\| < \epsilon^{1/2}$, where ϵ is the machine epsilon, and take H_k as an approximation to the exact solution X. Further iterations for SDA-1 may not be able to improve the accuracy significantly in view of the perturbation behavior of X and the fact that $I - G_kH_k$ and $I - H_kG_k$ are nearly singular for large k. So we are mainly interested in the behavior of SDA-1 for iterations up to the point where $\|H_k - H_{k-1}\| < \epsilon^{1/2}$ (assuming this is achievable). And up to that point, the behavior of SDA-1 for those nearby NAREs would be very much similar to that of SDA-1 for the singular NARE. We use one example to illustrate this point.

Example 5.1. Let T be a 16×16 doubly stochastic matrix given by $T = \frac{1}{2056}\texttt{magic}(16)$, where `magic` is the Matlab function that generates magic squares. Let $K = I - T$, and let the 8×8 matrices A, B, C, D be determined through (5.2). The matrix K is an irreducible singular M-matrix, and we have the critical case for the NARE (5.1). We take γ to be the largest diagonal entry of K (which is the last diagonal entry of K) and apply SDA-1. We find that $\|H_k - H_{k-1}\| < 10^{-7}$ is satisfied for $k = 24$. The convergence rate of $H_k - X$ is determined through that of F_k (see the proof of Theorem 5.4). We find that the values of $\sqrt[k]{\|F_k\|_\infty}$ are between 0.4924 and 0.5001 for $k = 4 : 24$.

We then increase the (1,1) entry of K by 10^{-12}. So K is now a nonsingular M-matrix. The matrix D is changed accordingly. The change in K does not change the largest diagonal entry of K. So we apply SDA-1 to the new NARE with the same γ. We find that $\|H_k - H_{k-1}\| < 10^{-7}$ is satisfied for $k = 23$, and that the values of $\sqrt[k]{\|F_k\|_\infty}$ are between 0.4924 and 0.5000 for $k = 4 : 21$ (the values are 0.4855 and 0.4570 for $k = 22$ and $k = 23$, respectively). Thus, the (nonterminal and more important) convergence behavior of SDA-1 for this nearby NARE is largely dictated by our theoretical results in Theorem 5.4.

6. Conclusion.

We have determined the convergence rate of the doubling algorithm in the critical (or singular) case for three different nonlinear matrix equations. It is possible to apply the techniques we reviewed in section 2 to other nonlinear matrix equations. Through this study, we have also gained more insights for the convergence behavior for the doubling algorithm for nearly singular cases.

REFERENCES

[1] B. D. O. Anderson, *Second-order convergent algorithms for the steady-state Riccati equation*, Internat. J. Control, 28 (1978), pp. 295–306.

[2] D. A. Bini, B. Iannazzo, G. Latouche, and B. Meini, *On the solution of algebraic Riccati equations arising in fluid queues*, Linear Algebra Appl., 413 (2006), pp. 474–494.

[3] D. A. Bini, G. Latouche, and B. Meini, *Solving matrix polynomial equations arising in queueing problems*, Linear Algebra Appl., 340 (2002), pp. 225–244.

[4] D. A. Bini, G. Latouche, and B. Meini, *Numerical Methods for Structured Markov Chains*, Oxford University Press, Oxford, 2005.

[5] 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.

[6] E. K.-W. Chu, H.-Y. Fan, and W.-W. Lin, *A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations*, Linear Algebra Appl., 396 (2005), pp. 55–80.

[7] E. K.-W. Chu, H.-Y. Fan, W.-W. Lin, and C. S. Wang, *Structure-preserving algorithms for periodic descrete-time algebraic Riccati equations*, Internat. J. Control, 77 (2004), pp. 767–788.

[8] J. C. Engwerda, *On the existence of a positive definite solution of the matrix equation $X + A^T X^{-1} A = I$*, Linear Algebra Appl., 194 (1993), pp. 91–108.

[9] J. C. Engwerda, A. C. M. Ran, and A. L. Rijkeboer, *Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X + A^* X^{-1} A = Q$*, Linear Algebra Appl., 186 (1993), pp. 255–275.

[10] H. R. Gail, S. L. Hantler, and B. A. Taylor, *Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains*, Adv. Appl. Probab., 28 (1996), pp. 114–165.

[11] F. R. Gantmacher, *The Theory of Matrices, Vol.* II, Chelsea Publishing Company, New York, 1987.

[12] C.-H. Guo, *Newton's method for discrete algebraic Riccati equations when the closed-loop matrix has eigenvalues on the unit circle*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 279–294.

[13] C.-H. Guo, *Convergence rate of an iterative method for a nonlinear matrix equation*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 295–302.

[14] C.-H. Guo, *Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 225–242.

[15] 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.

[16] C.-H. Guo, *Comments on a shifted cyclic reduction algorithm for quasi-birth-death problems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 1161–1166.

[17] C.-H. Guo, *Numerical solution of a quadratic eigenvalue problem*, Linear Algebra Appl., 385 (2004), pp. 391–406.

[18] C.-H. Guo, *Efficient methods for solving a nonsymmetric algebraic Riccati equation arising in stochastic fluid models*, J. Comput. Appl. Math., 192 (2006), pp. 353–373.

[19] C.-H. Guo, *A new class of nonsymmetric algebraic Riccati equations*, Linear Algebra Appl., 426 (2007), pp. 636–649.

[20] C.-H. Guo and N. J. Higham, *Iterative solution of a nonsymmetric algebraic Riccati equation*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 396–412.

[21] C.-H. Guo, B. Iannazzo, and B. Meini, *On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1083–1100.

[22] C.-H. Guo and P. Lancaster, *Iterative solution of two matrix equations*, Math. Comp., 68 (1999), pp. 1589–1603.

[23] C.-H. Guo and A. J. Laub, *On the iterative solution of a class of nonsymmetric algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 376–391.

[24] X.-X. Guo, W.-W. Lin, and S.-F. Xu, *A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation*, Numer. Math., 103 (2006), pp. 393–412.

[25] 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.

[26] C. He, B. Meini, N. H. Rhee, and K. Sohraby, *A quadratically convergent Bernoulli-like algorithm for solving matrix polynomial equations in Markov chains*, Electron. Trans. Numer. Anal., 17 (2004), pp. 151–167.

[27] T.-M. Huang and W.-W. Lin, *Structured doubling algorithms for weakly stabilizing Hermitian solutions of algebraic Riccati equations*, Linear Algebra Appl., 430 (2009), pp. 1452–1478.

[28] T.-M. Hwang, E. K.-W. Chu, and W.-W. Lin, *A generalized structure-preserving doubling algorithm for generalized discrete-time algebraic Riccati equations*, Internat. J. Control, 78 (2005), pp. 1063–1075.

[29] J. Juang, *Existence of algebraic matrix Riccati equations arising in transport theory*, Linear Algebra Appl., 230 (1995), pp. 89–100.

[30] J. Juang and W.-W. Lin, *Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 228–243.

[31] G. Latouche and V. Ramaswami, *A logarithmic reduction algorithm for quasi-death-birth processes*, J. Appl. Probab., 30 (1993), pp. 650–674.

[32] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, SIAM, Philadelphia, 1999.

[33] W.-W. Lin, V. Mehrmann, and H. Xu, *Canonical forms for Hamiltonian and symplectic matrices and pencils*, Linear Algbera Appl., 302/303 (1999), pp. 469–533.

[34] W.-W. Lin and S.-F. Xu, *Convergence analysis of structure-preserving doubling algorithms for Riccati-type matrix equations*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 26–39.

[35] B. Meini, *Efficient computation of the extreme solutions of $X + A^* X^{-1} A = Q$ and $X - A^* X^{-1} A = Q$*, Math. Comp., 71 (2002), pp. 1189–1204.

[36] V. Ramaswami, *Matrix analytic methods for stochastic fluid flows*, in Proceedings of the 16th International Teletraffic Congress, Elsevier Science B. V., Edinburgh, 1999, pp. 1019–1030.

[37] L. C. G. Rogers, *Fluid models in queueing theory and Wiener–Hopf factorization of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 390–413.

[38] J.-G. Sun and S.-F. Xu, *Perturbation analysis of the maximal solution of the matrix equation $X + A^T X^{-1} A = P$, II*, Linear Algebra Appl., 362 (2003), pp. 211–228.

[39] S.-F. Xu, *Numerical methods for the maximal solution of the matrix equation $X + A^T X^{-1} A = I$*, Acta Sci. Natur. Univ. Pekinensis, 36 (2000), pp. 29–38.

[40] X. Zhan, *Computing the extremal positive definite soluions of a matrix equation*, SIAM J. Sci. Comput., 17 (1996), pp. 1167–1174.

[41] X. Zhan and J. Xie, *On the matrix equation $X + A^T X^{-1} A = I$*, Linear Algebra Appl., 247 (1996), pp. 337–345.

SIAM J. Matrix Anal. Appl.
Vol. 31, No. 2, pp. 248–271

A NEWTON–GRASSMANN METHOD FOR COMPUTING THE BEST MULTILINEAR RANK-(r_1, r_2, r_3) APPROXIMATION OF A TENSOR[*]

LARS ELDÉN[†] AND BERKANT SAVAS[†]

Abstract. We derive a Newton method for computing the best rank-(r_1, r_2, r_3) approximation of a given $J \times K \times L$ tensor $\mathcal{A}$. The problem is formulated as an approximation problem on a product of Grassmann manifolds. Incorporating the manifold structure into Newton's method ensures that all iterates generated by the algorithm are points on the Grassmann manifolds. We also introduce a consistent notation for matricizing a tensor, for contracted tensor products and some tensor-algebraic manipulations, which simplify the derivation of the Newton equations and enable straightforward algorithmic implementation. Experiments show a quadratic convergence rate for the Newton–Grassmann algorithm.

Key words. tensor, multilinear, rank, approximation, Grassmann manifold, Newton

AMS subject classifications. 65F99, 65K10, 15A69, 14M15

DOI. 10.1137/070688316

1. Introduction.

The problem of approximating a tensor $\mathcal{A} \in \mathbb{R}^{J\times K\times L}$ by another tensor $\mathcal{B}$ of equal dimensions but of lower rank

$$\min_{\mathcal{B}} \|\mathcal{A} - \mathcal{B}\|$$

occurs, e.g., in signal processing [6, 4] and pattern classification [27]. Throughout the paper, we will use the Frobenius norm (we will state the precise meaning of this and other concepts in section 2). There is no unique definition of the rank of a tensor (as opposed to the case of matrices); see, e.g., [7]. Here we will deal with the concept of *multilinear rank*, defined by Hitchcock [12] (see also [6, 7]) and assume that $\mathrm{rank}(\mathcal{B}) = (r_1, r_2, r_3)$, which means that the tensor $\mathcal{B}$ can be written as a product of a *core* tensor $\mathcal{S}$ and three matrices,

$$\mathcal{B} = (X, Y, Z) \cdot \mathcal{S}, \qquad b_{ijk} = \sum_{\lambda,\mu,\nu} x_{i\lambda} y_{j\mu} z_{k\nu} s_{\lambda\mu\nu}, \tag{1.1}$$

with matrices of full column rank, $X \in \mathbb{R}^{J\times r_1}$, $Y \in \mathbb{R}^{K\times r_2}$, and $Z \in \mathbb{R}^{L\times r_3}$. The tensor $\mathcal{S}$ has dimensions $r_1 \times r_2 \times r_3$. It is no restriction to assume that X, Y, and Z have orthonormal columns. Thus, we want to solve the problem

$$\min_{\mathcal{S},X,Y,Z} \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{S}\| \quad \text{subject to} \quad X^{\mathsf{T}}X = I,\ Y^{\mathsf{T}}Y = I,\ Z^{\mathsf{T}}Z = I, \tag{1.2}$$

which is the problem of computing a *Tucker decomposition* [24, 25] approximating the tensor. The approximation problem is illustrated in Figure 1.1.

Unlike the matrix case, there is no known closed-form solution of the approximation problem (1.2). It can be shown that the minimization problem is well defined [7, Corollary 4.5].

[*] Received by the editors April 16, 2007; accepted for publication (in revised form) by H. Park October 16, 2008; published electronically March 18, 2009.
http://www.siam.org/journals/simax/31-2/68831.html

[†] Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden (laeld@math.liu.se, besav@math.liu.se).

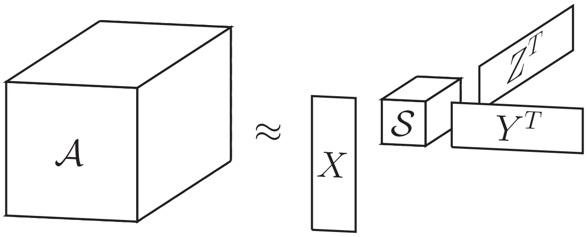


FIG. 1.1. *The approximation of a tensor $\mathcal{A}$ by another tensor $\mathcal{B} = (X, Y, Z) \cdot \mathcal{S}$ of lower multilinear rank.*

Let $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ be an *order-N* tensor (or N-tensor for short). We will restrict ourselves to considering the approximation problem (1.2) for a 3-tensor $\mathcal{A}$ in this paper. The main contribution is the derivation of a Newton method for the solution of (1.2). The constraints on the unknown matrices X, Y, and Z are taken into account by formulating the problem as an optimization problem on a product of three Grassmann manifolds. To be able to differentiate the objective function and derive the Newton equations without extensive index manipulation (as is sometimes used in tensor algebra), we develop an algebraic framework based on tensor contractions. Within this framework, it is also straightforward to generalize the derivations to tensors of order four and higher, and we sketch this in section 4.6.

In view of the lack of a standard terminology and notation in the field of tensor computations, we define the concepts used in this paper in section 2. There we also propose a "canonical" tensor matricization, contracted tensor products, and a few tensor-algebraic identities. The optimization problem on the product of three Grassmann manifolds is formulated in section 3, and the Newton–Grassmann (NG) method is derived in section 4.2. In section 5, the numerical implementation of the method is briefly described, and some numerical experiments are reported.

2. Tensor concepts and identities. For simplicity of notation and presentation, we will mostly, in this and the following sections, present the basic concepts using examples in terms of 3-tensors or 5-tensors. Some more general definitions are given in [5, 2, 3, 7]. We will use Roman letters written with a calligraphic font to denote tensors, capital Roman letters to denote matrices (2-tensors), and lowercase Roman letters to denote vectors. However, we will also use Roman letters in the middle of the alphabet, $J, K, L, \ldots$, and $j, k, l, \ldots$, to denote tensor dimensions and subscripts.

Let $\mathcal{A}$ denote a tensor in $\mathbb{R}^{J \times K \times L}$. The three "dimensions" of the tensor are referred to as *modes*. In the approximation problem (1.2), we will not consider the tensor as a multilinear operator,[1] and therefore, there is no need to make a distinction between *contravariant* and *covariant* tensor modes [16][2] in the tensor notation. We will use both standard subscripts and "MATLAB-like" notation: a particular tensor element will be denoted in two equivalent ways:

$$\mathcal{A}(j, k, l) = a_{jkl}.$$

We will refer to subtensors in the following way. A subtensor obtained by fixing one

[1] However, when we derive the Newton equations for solving the minimization problem, then we will deal with a Hessian, which, of course, is a linear operator constructed in terms of tensors.

[2] All our citations of [16] will refer to Chapter 2 of the book.

of the indices is called a *slice*, e.g.,

$$\mathcal{A}(j,:,:).$$

A *fiber* is a subtensor, where all indices but one are fixed:

$$\mathcal{A}(j,:,l).$$

When in the following, we use tensors, matrices, and vectors in operations, it is assumed that the dimensions of the respective quantities are conforming in the sense that all the operations are well defined.

2.1. Tensor-matrix multiplication. We define mode-p *multiplication of a tensor by a matrix* as follows. For concreteness, we first let $p = 1$. The mode-1 product of a tensor $\mathcal{A} \in \mathbb{R}^{J\times K\times L}$ by a matrix $W \in \mathbb{R}^{M\times J}$ is defined by

$$\mathbb{R}^{M\times K\times L} \ni \mathcal{B} = (W)_1 \cdot \mathcal{A}, \qquad b_{mkl} = \sum_{j=1}^{J} w_{mj} a_{jkl}. \tag{2.1}$$

This means that all column vectors (mode-1 fibers) in the 3-tensor are multiplied by the matrix W. Similarly, mode-2 multiplication by a matrix X means that all row vectors (mode-2 fibers) are multiplied by the matrix X. Mode-3 multiplication is analogous.

It is easy to see that for integers $p \neq q$, mode-p and mode-q multiplication commute

$$(W)_p \cdot \Big((X)_q \cdot \mathcal{A}\Big) = (X)_q \cdot \Big((W)_p \cdot \mathcal{A}\Big).$$

Therefore, it makes sense to define

$$(W,X)_{p,q} \cdot \mathcal{A} = (W)_p \cdot \Big((X)_q \cdot \mathcal{A}\Big).$$

Obviously, the following identity holds:

$$(W_1)_p \cdot \Big((W_2)_p \cdot \mathcal{A}\Big) = (W_1 W_2)_p \cdot \mathcal{A}, \tag{2.2}$$

where the matrix and tensor dimensions are assumed to be conforming and product W_1W_2 is standard matrix multiplication.

In the case when tensor-matrix multiplication is performed in all modes in the same formula, we omit the subscripts and write

$$(X,Y,Z) \cdot \mathcal{A}, \tag{2.3}$$

where the mode of each multiplication is understood from the order in which the matrices are given. Thus, we have the identity

$$(Y)_2 \cdot \mathcal{A} = (I,Y,I) \cdot \mathcal{A}.$$

The notation (2.3) was suggested by Lim [7]. An alternative notation was earlier given in [5]. Our $(W)_p \cdot \mathcal{A}$ is the same as $\mathcal{A} \times_p W$ in that system, and the identity (2.2) reads $(\mathcal{A} \times_p W_2) \times_p W_1 = \mathcal{A} \times_p (W_1 W_2)$.

One can also write the standard matrix multiplication of three matrices in the form

$$XFY^{\mathsf{T}} = (X,Y) \cdot F, \tag{2.4}$$

where, at the same time, F is considered as a matrix and a 2-tensor.

It is convenient to introduce a separate notation for multiplication by a transposed matrix $V \in \mathbb{R}^{J\times M}$:

$$\mathbb{R}^{M\times K\times L} \ni \mathcal{C} = \left(V^{\mathsf{T}}\right)_1 \cdot \mathcal{A} = \mathcal{A} \cdot (V)_1, \qquad c_{mkl} = \sum_{j=1}^{J} a_{jkl} v_{jm}. \tag{2.5}$$

2.2. A "canonical" tensor matricization. In the following sections, we will occasionally rearrange the elements of a tensor so that they form a matrix. We will refer to this as *matricizing* [2] the tensor.[3] In particular, when the Newton equations are to be solved numerically, they must be arranged as standard "matrix-vector" linear equations. Sometimes the matricization is performed *along one specific mode* [5, 15, 26]. Given an N-tensor $\mathcal{A} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, its matricization along the nth mode is a matrix of dimensions $I_n \times I_1 \cdots I_{n-1} I_{n+1} \cdots I_N$. Here we will introduce a more general tensor matricization which is intuitively and directly related to the matrix-tensor multiplication. In this matricization, we will map some modes of a tensor to the rows of the matrix and the rest to the columns. A similar, but not identical, generalized matricization is given in Bader and Kolda [2, 3]. The difference between the two definitions is explained later in this section.

Let $\mathsf{r} = [r_1, \ldots, r_L]$ be the modes of $\mathcal{A}$ mapped to the rows and $\mathsf{c} = [c_1, \ldots, c_M]$ be the modes of $\mathcal{A}$ mapped to the columns. The matricization is denoted

$$A^{(\mathsf{r};\mathsf{c})} \in \mathbb{R}^{J\times K}, \qquad \text{where} \qquad J = \prod_{i=1}^{L} I_{r_i} \qquad \text{and} \qquad K = \prod_{i=1}^{M} I_{c_i}. \tag{2.6}$$

Of course, many different one-to-one functions can map the tensor $\mathcal{A}$ onto a matrix with dimensions as specified in (2.6). The different maps differ in the ordering of the row and column indices of specific tensor elements.

We consider it useful, for analysis and consistency with tensor-matrix products, if the matricization operation has the following properties, which are best illustrated with a few examples. Let $\mathcal{A}$ be a 5-tensor and consider the product $\mathcal{B} = \mathcal{A} \cdot (V, W, X, Y, Z)$, where V, W, X, Y, Z are matrices of appropriate dimensions multiplied with $\mathcal{A}$ along its different modes. We want to point out that the box brackets used to specify r and c will be omitted when writing out a specific matricization, like in the following examples.

$$\begin{aligned}
B^{(2;1,3\ldots5)} \equiv B^{(2)} &= W^{\mathsf{T}} A^{(2)} (V \otimes X \otimes Y \otimes Z), && \mathsf{r} = [2], && \mathsf{c} = [1,3,4,5],\\
B^{(3,2;1,4,5)} \equiv B^{(3,2)} &= (X \otimes W)^{\mathsf{T}} A^{(3,2)} (V \otimes Y \otimes Z), && \mathsf{r} = [3,2], && \mathsf{c} = [1,4,5],\\
B^{(2,4,1;5,3)} &= (W \otimes Y \otimes V)^{\mathsf{T}} A^{(2,4,1;5,3)} (Z \otimes X), && \mathsf{r} = [2,4,1], && \mathsf{c} = [5,3],\\
B^{(1,2,4;5,3)} \equiv B^{(;5,3)} &= (V \otimes W \otimes Y)^{\mathsf{T}} A^{(;5,3)} (Z \otimes X), && \mathsf{r} = [1,2,4], && \mathsf{c} = [5,3].
\end{aligned}$$

Above $\otimes$ denotes the Kronecker product of matrices. Observe that the ordering of the matrices in the Kronecker products is specified by matricization indices r and c. Specifying only the row (column) modes assumes the column (row) modes to be in increasing order. In the above examples, we have used multiplication (2.3). For variant (2.5), the transpose will be introduced on the other side. For instance, with $\mathcal{C} = (V, W, X, Y, Z) \cdot \mathcal{A}$, we have

$$\begin{aligned}
C^{(2)} &= W A^{(2)} (V \otimes X \otimes Y \otimes Z)^{\mathsf{T}}, && \mathsf{r} = [2], && \mathsf{c} = [1,3,4,5],\\
C^{(3,2)} &= (X \otimes W) A^{(3,2)} (V \otimes Y \otimes Z)^{\mathsf{T}}, && \mathsf{r} = [3,2], && \mathsf{c} = [1,4,5].
\end{aligned}$$

[3] Alternative terms are *unfolding* [5] or *flattening* [26].

For a given N-tensor $\mathcal{A}$, the matricization to $A^{(\mathsf{r};\mathsf{c})}$ has the desired properties if element $\mathcal{A}(i_1, \ldots, i_N)$ is mapped to $A^{(\mathsf{r};\mathsf{c})}(j,k)$, where

$$j = 1 + \sum_{l=1}^{L} \left[\left(i_{r_{L-l+1}} - 1\right) \prod_{l'=1}^{l-1} I_{r_{L-l'+1}} \right], \tag{2.7}$$

$$k = 1 + \sum_{m=1}^{M} \left[\left(i_{c_{M-m+1}} - 1\right) \prod_{m'=1}^{m-1} I_{c_{M-m'+1}} \right]. \tag{2.8}$$

The matricization mapping presented in Bader and Kolda [2, 3] is different from ours in that it reverses the ordering of the matrices in both sides of matricized forms[4] of the tensor-matrix products. As already stated, tensor matricization can be represented in many different ways, and from a theoretical point of view, this is not an issue as long as the matricization mapping is applied consistently on both sides of an equation. But we want to emphasize that from a practical, implementational, and analytical point of view, it is important that the matricization is as simple and straightforward as possible, specifically when the matricization is applied on matrix-tensor products.

Applying the matricizing on the matrix products $B = (X, Y) \cdot A = XAY^{\mathsf{T}}$, we obtain

$$B^{(1)} = XA^{(1)}Y^{\mathsf{T}}, \qquad B^{(2)} = YA^{(2)}X^{\mathsf{T}}.$$

Of course, this is trivial since for matrices $A^{(1)} \equiv A$ and $A^{(2)} \equiv A^{\mathsf{T}}$.

Observe that this framework enables vectorization as well. Then, one of r or c has to be the empty set denoted $\varnothing$, and the other contains all modes. Consider first matrix case $B = (X, Y) \cdot A = XAY^{\mathsf{T}}$. Vectorizing B with $\mathsf{r} = [1, 2]$ and $\mathsf{c} = \emptyset$, we obtain

$$B^{(1,2;\varnothing)} = (X \otimes Y)A^{(1,2;\varnothing)},$$

where $A^{(1,2;\varnothing)}$ and $B^{(1,2;\varnothing)}$ are the row-wise vectorizations of A and B, giving a column vector. Changing the row modes to $\mathsf{r} = [2, 1]$, we obtain the more familiar

$$B^{(2,1;\varnothing)} = \operatorname{vec}(B) = \operatorname{vec}\left(XAY^{\mathsf{T}}\right) = (Y \otimes X) \operatorname{vec}(A) = (Y \otimes X)A^{(2,1;\varnothing)},$$

where, by convention, $\operatorname{vec}(\cdot)$ denotes the columnwise vectorization. Further, with a 3-tensor $\mathcal{B} = \mathcal{A} \cdot (X, Y, Z)$, we have

$$B^{(2,1,3;\varnothing)} = (Y \otimes X \otimes Z)^{\mathsf{T}} A^{(2,1,3;\varnothing)} \quad \text{and} \quad B^{(\varnothing;2,1,3)} = A^{(\varnothing;2,1,3)}(Y \otimes X \otimes Z),$$

where, in the first case, the vectorization gives a column vector, and, in the second case, the vectorization gives a row vector.

Finally, for later reference, we specify two special cases with tensor-matrix product along one mode only. Let $\mathcal{A}$ be a general N-tensor. Then

$$\mathcal{B} = \mathcal{A} \cdot (X)_p, \qquad \Leftrightarrow \qquad B^{(p)} = X^{\mathsf{T}} A^{(p)}, \tag{2.9}$$

$$\mathcal{C} = (X)_p \cdot \mathcal{A}, \qquad \Leftrightarrow \qquad C^{(p)} = XA^{(p)}. \tag{2.10}$$

The notation in this paper emphasizes the connection between multilinear tensor-matrix products and their matricized form. Other notations are found in [17, 5, 15, 11].

[4] For example, in the Bader–Kolda mapping, the matricization of $\mathcal{B}$ would be $B^{(2,4,1;5,3)} = (V \otimes Y \otimes W)^{\mathsf{T}} A^{(2,4,1;5,3)}(X \otimes Z)$.

2.3. Inner product, tensor product, and contracted product. Given two tensors $\mathcal{A}$ and $\mathcal{B}$ of the same dimensions, we define the *inner product*

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{j,k,l} a_{jkl} b_{jkl}. \tag{2.11}$$

This is, of course, the standard Euclidean inner product when we identify the space of tensors $\mathbb{R}^{J \times K \times L}$ with the vector space $\mathbb{R}^{JKL}$. The corresponding *tensor norm* is

$$\|\mathcal{A}\| = \langle \mathcal{A}, \mathcal{A} \rangle^{1/2}. \tag{2.12}$$

This *Frobenius norm* will be used throughout the paper. As in the matrix case, the norm is invariant under orthogonal transformations, i.e.,

$$\|\mathcal{A}\| = \|(U, V, W) \cdot \mathcal{A}\| = \|\mathcal{A} \cdot (U, V, W)\,\|$$

for orthogonal matrices U, V, and W. This follows immediately from the fact that mode-p multiplication by an orthogonal matrix does not change the Euclidean length of the mode-p fibers.

The *tensor product* or *outer product* of two tensors $\mathcal{A} \in \mathbb{R}^{J \times K \times L}$ and $\mathcal{B} \in \mathbb{R}^{M \times N}$, say, is a tensor of higher dimensionality, here a 5-tensor

$$\mathbb{R}^{J \times K \times L \times M \times N} \ni \mathcal{C} = \mathcal{A} \circ \mathcal{B}, \qquad c_{jklmn} = a_{jkl} b_{mn}.$$

The inner product (2.11) can be considered as a special case of the *contracted product of two tensors* (cf. [16, Chapter 2]), which is a tensor (outer) product followed by a contraction along specified modes. Thus, if $\mathcal{A}$ and $\mathcal{B}$ are 3-tensors, we define, using essentially the notation of [2],

$$\begin{aligned}
\mathcal{C} &= \langle \mathcal{A}, \mathcal{B} \rangle_1 \,, & c_{jklm} &= \sum_{\lambda} a_{\lambda jk} b_{\lambda lm} & &\text{(4-tensor)}\,,\\
D &= \langle \mathcal{A}, \mathcal{B} \rangle_{1:2} \,, & d_{jk} &= \sum_{\lambda,\mu} a_{\lambda\mu j} b_{\lambda\mu k} & &\text{(2-tensor)},\\
e = \langle \mathcal{A}, \mathcal{B} \rangle &= \langle \mathcal{A}, \mathcal{B} \rangle_{1:3} \,, & e &= \sum_{\lambda,\mu,\nu} a_{\lambda\mu\nu} b_{\lambda\mu\nu} & &\text{(scalar)}.
\end{aligned}$$

It is required that contracted dimensions are equal in the two tensors. We will refer to the first two as *partial contractions.*

Observe that we let the ordering of the modes in contracted tensor products be implicitly given in the summation. Thus, given $\mathcal{A} \in \mathbb{R}^{J \times K \times L}$ and $\mathcal{B} \in \mathbb{R}^{J \times M \times N}$, then

$$\mathcal{C} = \langle \mathcal{A}, \mathcal{B} \rangle_1 \in \mathbb{R}^{K \times L \times M \times N}.$$

In general, the modes of the product are those of the noncontracted modes of the first argument, followed by those of the noncontracted modes of the second argument, in their respective orders.

We will also use negative subscripts when the contraction is made in all but a few modes. For 3-tensors, we have

$$\langle \mathcal{A}, \mathcal{B} \rangle_{2:3} \equiv \langle \mathcal{A}, \mathcal{B} \rangle_{-1} \,, \qquad \langle \mathcal{A}, \mathcal{B} \rangle_{2} \equiv \langle \mathcal{A}, \mathcal{B} \rangle_{-(1,3)} \,.$$

The contracted product can be defined also for tensors of different numbers of modes, and contractions can be made along any two conforming modes. For example, with a 4-tensor $\mathcal{F}$ and matrices (2-tensors) F and G, we could have

$$\langle \mathcal{A}, F \rangle_{3,4;1,2} = G, \qquad \sum_{\mu,\nu} a_{jk\mu\nu} f_{\mu\nu} = g_{jk}, \tag{2.13}$$

where subscripts $3,4$ and $1,2$ indicate the contracting modes of the two arguments. Obviously, (2.13) defines a linear system of equations.

In the following sections, we will need a few lemmas. The first result relates contraction to matricization.

Lemma 2.1. *Let $\mathcal{A}$ and $\mathcal{B}$ be N-tensors of matching dimensions in all but (possibly) the ith mode and $A^{(i)}$ and $B^{(i)}$ the corresponding ith mode matricizations. Then*

$$\langle \mathcal{A}, \mathcal{B} \rangle_{-i} = A^{(i)} \left(B^{(i)}\right)^{\mathsf{T}}. \tag{2.14}$$

If all dimensions match, then

$$\langle \mathcal{A}, \mathcal{B} \rangle = \operatorname{tr}\left(\langle \mathcal{A}, \mathcal{B} \rangle_{-i}\right) = \operatorname{tr}\left(A^{(i)} \left(B^{(i)}\right)^{\mathsf{T}}\right). \tag{2.15}$$

Proof. For simplicity, we give the proof for only 3-tensors and partial contraction in all but the first mode. The general case is completely analogous. Let $\mathcal{A} \in \mathbb{R}^{J\times L\times M}$ and $\mathcal{B} \in \mathbb{R}^{K\times L\times M}$. Then

$$\langle \mathcal{A}, \mathcal{B} \rangle_{-1}(j,k) = \sum_{l,m} a_{jlm} b_{klm}. \tag{2.16}$$

With $C = A^{(1)}(B^{(1)})^{\mathsf{T}}$, we get

$$C(j,k) = \sum_{\lambda} a^{(1)}_{j\lambda} b^{(1)}_{k\lambda}, \tag{2.17}$$

where $A^{(1)}(j,\lambda) = a^{(1)}_{j\lambda}$ and $B^{(1)}(k,\lambda) = b^{(1)}_{k\lambda}$. By (2.8), element $\mathcal{A}(j,l,m)$ is mapped to $A^{(1)}(j,\lambda)$, where $\lambda = m + (l-1)M$ and similarly for elements of $\mathcal{B}$. The equality of (2.14) follows by observing that the λ-summation for the right-hand side actually consists of a summation over m and l.

The identity (2.15) follows from (2.16) by inspection. □

The partial contracted products of two matrices A and B are

$$\langle A, B \rangle_{-2} = \langle A, B \rangle_{1} = A^{\mathsf{T}} B, \qquad \langle A, B \rangle_{-1} = \langle A, B \rangle_{2} = A B^{\mathsf{T}}, \tag{2.18}$$

which shows that partial contraction is related to matrix transposition and multiplication. In the next lemma, we show that partial contractions play the role of taking the adjoint with respect to the inner product (2.11).

Lemma 2.2. *Let the N-tensors $\mathcal{B}$ and $\mathcal{C}$ and the matrix Q be of conforming dimensions. Then*

$$\left\langle \mathcal{B} \cdot (Q)_i, \mathcal{C} \right\rangle = \left\langle Q, \langle \mathcal{B}, \mathcal{C} \rangle_{-i} \right\rangle, \tag{2.19}$$

$$\left\langle \mathcal{B} \cdot (Q)_i, \mathcal{C} \right\rangle_{-i} = Q^{\mathsf{T}} \langle \mathcal{B}, \mathcal{C} \rangle_{-i} = \left\langle Q, \langle \mathcal{B}, \mathcal{C} \rangle_{-i} \right\rangle_1, \tag{2.20}$$

$$\left\langle \mathcal{B}, \mathcal{C} \cdot (Q^{\mathsf{T}})_i \right\rangle_{-i} = \langle \mathcal{B}, \mathcal{C} \rangle_{-i} Q^{\mathsf{T}} = \left\langle \langle \mathcal{B}, \mathcal{C} \rangle_{-i}, Q \right\rangle_2. \tag{2.21}$$

Proof. Equation (2.19) follows from

$$\langle Q, \langle \mathcal{B}, \mathcal{C} \rangle_{-i} \rangle = \left\langle Q, B^{(i)} \left(C^{(i)} \right)^{\mathsf{T}} \right\rangle = \operatorname{tr} \left(Q^{\mathsf{T}} B^{(i)} \left(C^{(i)} \right)^{\mathsf{T}} \right)$$
$$= \operatorname{tr} \left(\left(\mathcal{B} \cdot (Q)_i \right)^{(i)} \left(C^{(i)} \right)^{\mathsf{T}} \right) = \langle \mathcal{B} \cdot (Q)_i , \mathcal{C} \rangle ,$$

where we have used (2.15) and (2.9). The second and third identities follow directly by matricizing the expressions along the ith mode and using (2.14). □

The following lemma can be motivated as follows: Obviously, from the definition of a contracted product, the mapping

$$Q \longrightarrow \left\langle \mathcal{B} \cdot (Q)_j , \mathcal{C} \right\rangle_{-i}$$

is linear from matrices to matrices. In order to solve a linear system involving such a mapping, we need to write it in the form (2.13).

Lemma 2.3. *Let the N-tensors $\mathcal{B}$ and $\mathcal{C}$ and the matrix Q be of conforming dimensions. If $j \neq i$, then*

$$\left\langle \mathcal{B} \cdot (Q)_j , \mathcal{C} \right\rangle_{-i} = \begin{cases} \left\langle \langle \mathcal{B}, \mathcal{C} \rangle_{-(i,j)} , Q \right\rangle_{1,3;1,2} & \text{if } j < i, \\ \left\langle \langle \mathcal{B}, \mathcal{C} \rangle_{-(i,j)} , Q \right\rangle_{2,4;1,2} & \text{if } j > i, \end{cases} \tag{2.22}$$

$$\left\langle \mathcal{B}, \mathcal{C} \cdot (Q)_j \right\rangle_{-i} = \begin{cases} \left\langle \langle \mathcal{B}, \mathcal{C} \rangle_{-(i,j)} , Q \right\rangle_{3,1;1,2} & \text{if } j < i, \\ \left\langle \langle \mathcal{B}, \mathcal{C} \rangle_{-(i,j)} , Q \right\rangle_{4,2;1,2} & \text{if } j > i. \end{cases} \tag{2.23}$$

The proof is given in Appendix A.

2.4. Multilinear rank and higher order SVD. The *multilinear* rank of a 3-tensor is a triplet (r_1, r_2, r_3) such that

$$r_i = \dim \left(R \left(A^{(i)} \right) \right) = \operatorname{rank} \left(A^{(i)} \right), \qquad i = 1, 2, 3,$$

where $R(A) = \{ y \,|\, y = Ax \}$ is the *range space* of the matrix A and rank(A) is the matrix rank. Multilinear rank [12, 6] is discussed in [7], as well as other rank concepts. In this paper, we will deal only with multilinear rank, and we use the notation rank-(r_1, r_2, r_3) and rank$(\mathcal{A}) = (r_1, r_2, r_3)$.

For matrices, the rank is obtained via the *singular value decomposition (SVD)*; see, e.g., [10, Chapter 2]. One generalization of the SVD to tensors, the *higher order SVD*, was given in [5]. We here present the HOSVD for the case when $\mathcal{A}$ is a 3-tensor. The general case is an obvious generalization.

Theorem 2.4 (HOSVD). *Any 3-tensor $\mathcal{A} \in \mathbb{R}^{J \times K \times L}$ can be factorized*

$$\mathcal{A} = (U, V, W) \cdot \mathcal{S}, \tag{2.24}$$

where $U \in \mathbb{R}^{J \times J}$, $V \in \mathbb{R}^{K \times K}$, and $W \in \mathbb{R}^{L \times L}$ are orthogonal matrices and $\mathcal{S} \in \mathbb{R}^{J \times K \times L}$ is all-orthogonal. *The matrices $\langle \mathcal{S}, \mathcal{S} \rangle_{-i}$, $i = 1, 2, 3$ are diagonal, and*

$$\| \mathcal{S}(1, :, :) \| \geq \| \mathcal{S}(2, :, :) \| \geq \cdots \geq 0, \tag{2.25}$$
$$\| \mathcal{S}(:, 1, :) \| \geq \| \mathcal{S}(:, 2, :) \| \geq \cdots \geq 0, \tag{2.26}$$
$$\| \mathcal{S}(:, :, 1) \| \geq \| \mathcal{S}(:, :, 2) \| \geq \cdots \geq 0 \tag{2.27}$$

are the 1-mode, 2-mode, and 3-mode singular values, also denoted $\sigma_i^{(1)}$, $\sigma_i^{(2)}$, $\sigma_i^{(3)}$.

Partitioning the orthogonal matrices in terms of columns $U = (u_1, \ldots, u_J)$, $V = (v_1, \ldots, v_K)$, $W = (w_1, \ldots, w_L)$, the HOSVD equation can be written

$$\mathcal{A} = \sum_{j,k,l} s_{jkl}\, u_j \circ v_k \circ w_l,$$

where $\circ$ denotes the tensor (outer) product: for vectors x, y, and z, we have

$$(x \circ y \circ z)_{\lambda\mu\nu} = x_\lambda y_\mu z_\nu.$$

Assume that the 1-, 2-, and 3-mode singular values of $\mathcal{A}$ satisfy

$$\begin{aligned} \sigma^{(1)}_{r_1} &> 0, & \sigma^{(1)}_{r_1+1} &= 0, \\ \sigma^{(2)}_{r_2} &> 0, & \sigma^{(2)}_{r_2+1} &= 0, \\ \sigma^{(3)}_{r_3} &> 0, & \sigma^{(3)}_{r_3+1} &= 0 \end{aligned}$$

for some constants r_1, r_2, and r_3. It is easy to show that, in this case, the multilinear rank of $\mathcal{A}$ is (r_1, r_2, r_3).

3. Best rank-(r_1, r_2, r_3) approximation. Assume that we want to approximate, using the norm (2.12), the tensor $\mathcal{A}$ by another tensor $\mathcal{B}$ of rank (r_1, r_2, r_3). Thus, we want to solve

$$\min_{\operatorname{rank}(\mathcal{B})=(r_1,r_2,r_3)} \|\mathcal{A} - \mathcal{B}\|. \tag{3.1}$$

This problem is treated in [6]. In the matrix case, the solution of the corresponding problem is given by the truncated SVD (the Eckart–Young property; a simple proof is given in [9, Theorem 6.7]). In view of the fact that the HOSVD "orders the mass" of the tensor in a similar way as the SVD (see (2.25)–(2.27)), one might think that a truncated HOSVD would give the solution of (3.1). However, this is not the case [6].

Some theoretical questions concerning the best rank-(r_1, r_2, r_3) approximation problem are studied in [7]. In particular, the following result is proved (Corollary 4.5).

Proposition 3.1. *Let $(r_1, r_2, \ldots, r_k)$ be an arbitray k-tuple satisfying $r_i \leq s_i$, $i = 1, 2, \ldots, k$, where $(s_1, s_2, \ldots, s_k)$ is the multilinear rank of a given k-tensor $\mathcal{A}$. Then $\mathcal{A}$ has a best approximation $\mathcal{B}$, in the norm* (2.12), *with*

$$\operatorname{rank}(\mathcal{B}) \leq (r_1, r_2, \ldots, r_k).$$

The rank constraint in (3.1) implies (see [7, 6] and section 2.4) that $\mathcal{B}$ can be written

$$\mathcal{B} = (X, Y, Z) \cdot \overline{\mathcal{B}}, \qquad \overline{\mathcal{B}} \in \mathbb{R}^{r_1 \times r_2 \times r_3},$$

where $X \in \mathbb{R}^{J \times r_1}$, $Y \in \mathbb{R}^{K \times r_2}$, and $Z \in \mathbb{R}^{L \times r_3}$, with

$$X^{\mathsf{T}} X = I, \qquad Y^{\mathsf{T}} Y = I, \qquad Z^{\mathsf{T}} Z = I. \tag{3.2}$$

The identity matrices in (3.2) have dimensions r_1, r_2, and r_3, respectively.

Define three orthogonal matrices

$$\begin{aligned} \widehat{X} &= \begin{pmatrix} X & X_\perp \end{pmatrix}, & X_\perp &\in \mathbb{R}^{J \times (J - r_1)}, \\ \widehat{Y} &= \begin{pmatrix} Y & Y_\perp \end{pmatrix}, & Y_\perp &\in \mathbb{R}^{K \times (K - r_2)}, \\ \widehat{Z} &= \begin{pmatrix} Z & Z_\perp \end{pmatrix}, & Z_\perp &\in \mathbb{R}^{L \times (L - r_3)}. \end{aligned}$$

Further, define three sets of indices

$$\begin{aligned} S &= \{(j,k,l) \,|\, 1 \le j \le J, 1 \le k \le K, 1 \le l \le L\}, \\ S_r &= \{(j,k,l) \,|\, 1 \le j \le r_1, 1 \le k \le r_2, 1 \le l \le r_3\}, \\ S' &= S \backslash S_r. \end{aligned}$$

Define $\widehat{\mathcal{A}} = (\widehat{X}^{\mathsf{T}}, \widehat{Y}^{\mathsf{T}}, \widehat{Z}^{\mathsf{T}}) \cdot \mathcal{A}$ and $\widehat{\mathcal{B}} = (\widehat{X}^{\mathsf{T}}, \widehat{Y}^{\mathsf{T}}, \widehat{Z}^{\mathsf{T}}) \cdot \mathcal{B}$. Then the residual in the approximation problem becomes

$$\begin{aligned} \|\mathcal{A} - \mathcal{B}\|^2 = \left\|\widehat{\mathcal{A}} - \widehat{\mathcal{B}}\right\|^2 &= \sum_{(j,k,l) \in S_r} \left(\hat{a}_{jkl} - \hat{b}_{jkl}\right)^2 + \sum_{(j,k,l) \in S'} \left(\hat{a}_{jkl} - \hat{b}_{jkl}\right)^2 \\ &= \sum_{(j,k,l) \in S_r} \left(\hat{a}_{jkl} - \hat{b}_{jkl}\right)^2 + \sum_{(j,k,l) \in S'} \hat{a}_{jkl}^2 \, , \end{aligned}$$

due to the rank constraint. The first term can be made equal to zero by choosing $\hat{a}_{jkl} = \hat{b}_{jkl}$, and the residual becomes

$$\left\|\widehat{\mathcal{A}} - \widehat{\mathcal{B}}\right\|^2 = \sum_{(j,k,l) \in S'} \hat{a}_{jkl}^2.$$

We now see that the problem of solving (3.1), i.e., making the residual as small as possible, is equivalent to determining X, Y, and Z so that

$$\left\|(X^{\mathsf{T}}, Y^{\mathsf{T}}, Z^{\mathsf{T}}) \cdot \mathcal{A}\right\| = \|\mathcal{A} \cdot (X, Y, Z)\|$$

is maximized. We thus define the objective function to be maximized:

$$\Phi(X,Y,Z) = \frac{1}{2} \|\mathcal{A} \cdot (X,Y,Z)\|^2 = \frac{1}{2} \sum_{j,k,l} \left(\sum_{\lambda,\mu,\nu} a_{\lambda\mu\nu} x_{\lambda j} y_{\mu k} z_{\nu l} \right)^2, \tag{3.3}$$

where $x_{\lambda j}$, $y_{\mu k}$, and $z_{\nu l}$ are elements of X, Y, and Z, respectively.

4. Solving the maximization problem by Newton's method. It follows from the invariance of the norm under orthogonal transformations that

$$\Phi(X,Y,Z) = \Phi(XU, YV, ZW) \tag{4.1}$$

for orthogonal matrices $U \in \mathbb{R}^{r_1 \times r_1}$, $V \in \mathbb{R}^{r_2 \times r_2}$, and $W \in \mathbb{R}^{r_3 \times r_3}$. This means that the problem of maximizing Φ under the orthogonality constraint (3.2) is not yet well defined: the problem is overparameterized, and any straightforward constrained optimization method would have difficulties. It follows that we should maximize the function Φ not just over matrices with orthonormal columns but over equivalence classes of such matrices, for instance,

$$[X] = \{XU \mid U \text{ orthogonal}\}. \tag{4.2}$$

This means that we should maximize over the *Grassmann manifold* [8] or more precisely, over a product of Grassmann manifolds.

4.1. Newton's method on the Grassmann manifold. The Grassmann manifold can be considered as a set of equivalence classes of matrices (4.2) with orthonormal columns that span the same subspace. Here we give a very brief description of Newton's method for maximizing a function $G(X)$ defined on the Grassmann manifold and then we state Newton's method on the product manifold. Our presentation is based on that in [8], where detailed definitions and derivations are given. For a comprehensive treatment of optimization on matrix manifolds, see [1].

Assume that $X \in \mathbb{R}^{J \times r_1}$ is a point on the Grassmann manifold $\mathrm{Gr}(J, r_1)$. This is clearly abuse of notation, because, strictly speaking, we should say that the matrix X with orthonormal columns is a representative of the equivalence class of matrices that represents a point on the manifold. In order not to burden the presentation with too many abstract details, we allow ourselves this laxness.

The manifold is a "curved object" (i.e., not a vector space), and therefore, we perform our computations on a "linear approximation" at X of the manifold, the *tangent space* $\mathbb{T}_X$, which is an affine vector space with elements in $\mathbb{R}^{J \times r_1}$. It can be shown [8] that any *tangent*[5] $\Delta \in \mathbb{T}_X$ satisfies

$$X^{\mathsf{T}} \Delta = 0.$$

The projection on the tangent space is

$$\Pi_X = I - XX^{\mathsf{T}}. \tag{4.3}$$

The canonical way of defining an inner product on the tangent space is

$$\langle \Delta_1, \Delta_2 \rangle = \mathrm{tr}\left(\Delta_1^{\mathsf{T}} \Delta_2\right), \tag{4.4}$$

where Δ_1 and Δ_2 are tangents at the same point X.

Recall that we want to maximize $G(X)$ on $\mathrm{Gr}(J, r_1)$. In Newton's method on the Grassmann manifold, we make a local quadratic approximation of a function defined on the manifold. The natural approach is to consider the function along geodesic curves. Let Δ be a tangent at X, and let $X_\Delta(t)$ be a parameterization of a geodesic curve in the direction Δ. With the thin SVD, $\Delta = U \Sigma V^{\mathsf{T}}$, where $U \in \mathbb{R}^{J \times r_1}$ and $\Sigma \in \mathbb{R}^{r_1 \times r_1}$, the geodesic is given by [8]

$$X_\Delta(t) = XV \cos(t\Sigma) V^{\mathsf{T}} + U \sin(t\Sigma) V^{\mathsf{T}}. \tag{4.5}$$

Naturally, $dX(t)/dt|_{t=0} = \Delta$.

The objective of the quadratic approximation is to determine a tangent Δ at X that maximizes

$$\begin{aligned} G(X_\Delta(1)) &\approx G(X) + \left.\frac{dG}{dt}\right|_{t=0} + \frac{1}{2} \left.\frac{d^2G}{dt^2}\right|_{t=0} \\ &= G(X) + \langle \Delta, \nabla G \rangle + \frac{1}{2} \langle \Delta, H(\Delta) \rangle, \end{aligned} \tag{4.6}$$

where $\langle \cdot, \cdot \rangle$ is the inner product (4.4). ∇G is the gradient on the tangent space

$$\nabla G = \Pi_X G_x, \qquad (G_x)_{jk} = \frac{\partial G}{\partial x_{jk}}, \tag{4.7}$$

and the Hessian $H(\Delta)$ is a linear operator on the tangent space $H(\cdot) : \mathbb{T}_X \to \mathbb{T}_X$.

[5] We could say *tangent vectors*, since they are elements of an affine vector space or *tangent matrices*, since, in our case, they have the form of a matrix. To avoid confusion, we will simply call them *tangents*.

Maximizing (4.6) with respect to Δ leads us to the Newton equation $H(\Delta) = -\nabla G$. It is shown in [8] that the Newton equation for determining $\Delta \in \mathbb{T}_X$ is a Sylvester-like equation, which in our notation becomes

$$\Pi_X \left\langle \mathcal{G}_{xx}, \Delta \right\rangle_{1:2} - \Delta \left\langle X, G_x \right\rangle_1 = -\nabla G, \qquad (\mathcal{G}_{xx})_{jklm} = \frac{\partial^2 G}{\partial x_{jk}\, \partial x_{lm}}. \tag{4.8}$$

Here, the contracted product of the 4-tensor $\mathcal{G}_{xx}$ and the matrix Δ defines a linear operator. $\langle \mathcal{G}_{xx}, \Delta \rangle_{1:2}$ is a matrix, which can be multiplied by Π_X to project it to the tangent space $\mathbb{T}_X$.

In order to solve the Newton equation (4.8) numerically, there are essentially three approaches.

Solve the problem in the ambient Euclidean space. Using the coordinates given by X itself, we could disregard that the problem is defined on the Grassmann manifold and solve the Newton equation in the ambient Euclidean space $\mathbb{R}^{Jr_1}$. Since X is constrained, i.e., $X^{\mathsf{T}} X = I$, the overparameterized coordinate representation will cause NG equation (4.8) to be singular. A pseudoinverse solution combined with a projection might be used to keep the iterates on the manifold.

Solve the problem on the tangent space. The NG equation (4.8) is nonsingular in the neighborhood of a local maximum when considered *on the tangent space* $\mathbb{T}_X$. Using a coordinate representation on the tangent space, one can obtain a smaller problem with a full rank Hessian operator. We will do this in the case of a product manifold in section 4.3.

Solve the problem by introducing Lagrange multipliers. The third approach, which is more efficient for large problems with $J \gg r_1$, is to effectively introduce Lagrange multipliers for the constraint and simultaneously solve for those and Δ; see, e.g., [19, Algorithm 2].

4.2. Newton's method on the product manifold. Our constrained optimization problem is

$$\max_{(X,Y,Z) \in \mathrm{Gr}^3} \Phi(X, Y, Z), \qquad \mathrm{Gr}^3 = \mathrm{Gr}(J, r_1) \times \mathrm{Gr}(K, r_2) \times \mathrm{Gr}(L, r_3), \tag{4.9}$$

where the objective function is defined in (3.3). The tangent space at (X, Y, Z) is $\mathbb{T}^3 = \mathbb{T}_X \times \mathbb{T}_Y \times \mathbb{T}_Z$, and the inner product is the sum of the inner products on the respective manifolds. The dimensions of the tangent spaces are usually different, and therefore, we write $\Delta = (\Delta_x, \Delta_y, \Delta_z)$ as a triplet rather than as blocks of a matrix. This is the case both for the gradient and the Hessian. We will now derive the Newton equation on the product manifold corresponding to (4.8). First, we will differentiate Φ in direction Δ and then we will identify the terms in the expansion corresponding to (4.6).

A geodesic curve in direction $(\Delta_x, \Delta_y, \Delta_z)$ is given by $(X(t), Y(t), Z(t))$, where the components are defined according to (4.5). From the definition of a tangent (see also, (4.5)), we have

$$\left. \frac{dx_{\nu\mu}}{dt} \right|_{t=0} = (\Delta_x)_{\nu\mu}$$

and correspondingly in the other two directions. We therefore get

$$\left. \left(\frac{dX(t)}{dt}, \frac{dY(t)}{dt}, \frac{dZ(t)}{dt} \right) \right|_{t=0} = (\Delta_x , \Delta_y , \Delta_z),$$

and since

$$\mathcal{A}\cdot(X,Y,Z)\,(j,k,l)=\sum_{\lambda,\mu,\nu} a_{\lambda\mu\nu}x_{\lambda j}y_{\mu k}z_{\nu l},$$

every $x_{\lambda j}$, etc., will be replaced by $(\Delta_x)_{\lambda j}$, etc., in the differentiation of $\mathcal{A}\cdot(X,Y,Z)$:

$$\left.\frac{d\,(\mathcal{A}\cdot(X,Y,Z))}{dt}\right|_{t=0}=\mathcal{A}\cdot(\Delta_x,Y,Z)+\mathcal{A}\cdot(X,\Delta_y,Z)+\mathcal{A}\cdot(X,Y,\Delta_z)\,.$$

4.2.1. Grassmann gradient. The first derivative of Φ becomes

$$\begin{aligned}\left.\frac{d\Phi}{dt}\right|_{t=0} &= \frac{1}{2}\frac{d}{dt}\left.\langle\mathcal{A}\cdot(X,Y,Z)\,,\mathcal{A}\cdot(X,Y,Z)\rangle\right|_{t=0}\\ &= \langle\mathcal{A}\cdot(\Delta_x,Y,Z)\,,\mathcal{A}\cdot(X,Y,Z)\rangle && (4.10)\\ &\quad+\ \langle\mathcal{A}\cdot(X,\Delta_y,Z)\,,\mathcal{A}\cdot(X,Y,Z)\rangle && (4.11)\\ &\quad+\ \langle\mathcal{A}\cdot(X,Y,\Delta_z)\,,\mathcal{A}\cdot(X,Y,Z)\rangle. && (4.12)\end{aligned}$$

First, we will identify the gradient $\nabla\Phi$, and to do this, we need to rewrite (4.10)–(4.12) in the form of the first derivative term in (4.6).

It is convenient to define the tensor $\mathcal{F}=\mathcal{A}\cdot(X,Y,Z)$, since it will be used in many expressions. From (2.19), we see that

$$\langle\mathcal{A}\cdot(\Delta_x,Y,Z)\,,\mathcal{F}\rangle=\left\langle\Delta_x,\langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{F}\rangle_{-1}\right\rangle=:\langle\Delta_x,\Phi_x\rangle \tag{4.13}$$

and correspondingly, for the other terms in (4.11) and (4.12). The x-part of the Grassmann gradient (see (4.7)) then becomes

$$\begin{aligned}\Pi_X\Phi_x &= \Pi_X\,\langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{F}\rangle_{-1}\\ &= \langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{A}\cdot(X,Y,Z)\rangle_{-1}-XX^{\mathsf{T}}\,\langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{F}\rangle_{-1}\\ &= \langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{A}\cdot(I,Y,Z)\rangle_{-1}X-X\,\langle\mathcal{F},\mathcal{F}\rangle_{-1}\,, && (4.14)\end{aligned}$$

where we have used Lemma 2.1, (2.20), and (2.21). The factors in (4.14) have an interpretation in terms of subtensors: $\mathcal{F}$ is a tensor in $\mathbb{R}^{r_1\times r_2\times r_3}$, and the contracted product

$$\langle\mathcal{F},\mathcal{F}\rangle_{-1}=\langle\mathcal{F},\mathcal{F}\rangle_{2:3}=\langle\mathcal{A}\cdot(X,Y,Z)\,,\mathcal{A}\cdot(X,Y,Z)\rangle_{2:3}$$

is a symmetric matrix in $\mathbb{R}^{r_1\times r_1}$, whose (j,k) element is the inner product between $\mathcal{F}(j,:,:)$ and $\mathcal{F}(k,:,:)$, i.e., first mode jth and kth slices of $\mathcal{F}$. Multiplying from the left by X results in a $J\times r_1$ matrix. Similarly, $\langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{A}\cdot(I,Y,Z)\rangle_{-1}$ is a symmetric $J\times J$ matrix, where the elements are inner products between the slices of $\mathcal{A}\cdot(I,Y,Z)$.

Using analogous reformulations for (4.11) and (4.12), the complete Grassmann gradient becomes $\nabla\Phi=(\Pi_X\Phi_x,\Pi_Y\Phi_y,\Pi_Z\Phi_z)$, where

$$\Pi_X\Phi_x=\langle\mathcal{A}\cdot(I,Y,Z)\,,\mathcal{A}\cdot(I,Y,Z)\rangle_{-1}X-X\,\langle\mathcal{F},\mathcal{F}\rangle_{-1}\,, \tag{4.15}$$

$$\Pi_Y\Phi_y=\langle\mathcal{A}\cdot(X,I,Z)\,,\mathcal{A}\cdot(X,I,Z)\rangle_{-2}Y-Y\,\langle\mathcal{F},\mathcal{F}\rangle_{-2}\,, \tag{4.16}$$

$$\Pi_Z\Phi_z=\langle\mathcal{A}\cdot(X,Y,I)\,,\mathcal{A}\cdot(X,Y,I)\rangle_{-3}Z-Z\,\langle\mathcal{F},\mathcal{F}\rangle_{-3}\,. \tag{4.17}$$

4.2.2. Grassmann Hessian. Computing the second derivative of Φ, using the same technique as for the gradient, we obtain

$$\begin{aligned}\left.\frac{d^2\Phi}{dt^2}\right|_{t=0} &= \langle \mathcal{A}\cdot(\Delta_x, Y, Z), \mathcal{A}\cdot(\Delta_x, Y, Z)\rangle + \langle \mathcal{A}\cdot(\Delta_x, \Delta_y, Z), \mathcal{A}\cdot(X, Y, Z)\rangle \\ &\quad + \langle \mathcal{A}\cdot(\Delta_x, Y, Z), \mathcal{A}\cdot(X, \Delta_y, Z)\rangle + \langle \mathcal{A}\cdot(\Delta_x, Y, \Delta_z), \mathcal{A}\cdot(X, Y, Z)\rangle \\ &\quad + \langle \mathcal{A}\cdot(\Delta_x, Y, Z), \mathcal{A}\cdot(X, Y, \Delta_z)\rangle + \cdots, \end{aligned} \tag{4.18}$$

where, for simplicity of the present discussion, we have omitted 10 analogous terms. The first term, which gives the "xx" derivative, can be dealt with using Lemma 2.2. We get

$$\begin{aligned}\langle \mathcal{A}\cdot(\Delta_x, Y, Z), \mathcal{A}\cdot(\Delta_x, Y, Z)\rangle &= \left\langle \Delta_x, \langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(\Delta_x, Y, Z)\rangle_{-1}\right\rangle \\ &= \left\langle \Delta_x, \langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(I, Y, Z)\rangle_{-1}\Delta_x\right\rangle.\end{aligned}$$

From (4.8) and (4.13), we now see that the "xx" part of the Grassmann Hessian is a Sylvester operator

$$\begin{aligned}\mathcal{H}_{xx}(\Delta_x) &= \Pi_X \langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(I, Y, Z)\rangle_{-1}\Delta_x - \Delta_x X^{\mathsf{T}}\Phi_x \\ &= \Pi_X \langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(I, Y, Z)\rangle_{-1}\Delta_x - \Delta_x\langle \mathcal{F}, \mathcal{F}\rangle_{-1},\end{aligned} \tag{4.19}$$

where Φ_x is defined in (4.13) and we have used Lemma 2.2.

For the second term in (4.18), we get, using Lemmas 2.2 and 2.3,

$$\begin{aligned}\left\langle \mathcal{A}\cdot(\Delta_x, \Delta_y, Z), \mathcal{A}\cdot(X, Y, Z)\right\rangle &= \left\langle \Delta_x, \left\langle \mathcal{A}\cdot(I, \Delta_y, Z), \mathcal{A}\cdot(X, Y, Z)\right\rangle_{-1}\right\rangle \\ &= \left\langle \Delta_x, \left\langle \mathcal{F}^1_{xy}, \Delta_y\right\rangle_{2,4;1,2}\right\rangle,\end{aligned} \tag{4.20}$$

where $\mathcal{F}^1_{xy}$ is the 4-tensor

$$\begin{aligned}\mathbb{R}^{J\times K\times r_1\times r_2} \ni \mathcal{F}^1_{xy} &= \langle \mathcal{A}\cdot(I, I, Z), \mathcal{A}\cdot(X, Y, Z)\rangle_{-(1,2)} \\ &= \langle \mathcal{A}\cdot(I, I, Z), \mathcal{A}\cdot(X, Y, Z)\rangle_{3}.\end{aligned}$$

Obviously, $\left\langle \mathcal{F}^1_{xy}, \cdot\right\rangle_{2,4;1,2}$ defines a linear operator that maps matrices on matrices.

The third term in (4.18) becomes, again, using Lemmas 2.2 and 2.3,

$$\begin{aligned}\left\langle \mathcal{A}\cdot(\Delta_x, Y, Z), \mathcal{A}\cdot(X, \Delta_y, Z)\right\rangle &= \left\langle \Delta_x, \left\langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(X, \Delta_y, Z)\right\rangle_{-1}\right\rangle \\ &= \left\langle \Delta_x, \left\langle \mathcal{F}^2_{xy}, \Delta_y\right\rangle_{4,2;1,2}\right\rangle,\end{aligned} \tag{4.21}$$

where $\mathcal{F}^2_{xy}$ is a 4-tensor

$$\begin{aligned}\mathbb{R}^{J\times r_2\times r_1\times K} \ni \mathcal{F}^2_{xy} &= \langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(X, I, Z)\rangle_{-(1,2)} \\ &= \langle \mathcal{A}\cdot(I, Y, Z), \mathcal{A}\cdot(X, I, Z)\rangle_{3}.\end{aligned}$$

We now have

$$\mathcal{F}_{xy}(\Delta_y) = \left\langle \mathcal{F}^1_{xy}, \Delta_y\right\rangle_{2,4;1,2} + \left\langle \mathcal{F}^2_{xy}, \Delta_y\right\rangle_{4,2;1,2}. \tag{4.22}$$

The fourth and fifth terms in (4.18) can be dealt with similarly and give the $\mathcal{F}_{xz}$ operator.

In order for the second derivative operators to be in the tangent space, we must multiply also $\mathcal{F}_{xy}$, and $\mathcal{F}_{xz}$ by Π_X. If we rewrite all the terms in the second derivative (4.18) in an analogous way, we get a Hessian operator

$$\mathcal{H}(\Delta) = (\Phi_{x*}(\Delta), \Phi_{y*}(\Delta), \Phi_{z*}(\Delta)) \,:\, \mathbb{T}^3 \mapsto \mathbb{T}^3,$$

where

$$\begin{aligned} \Phi_{x*}(\Delta) &= \mathcal{H}_{xx}(\Delta_x) + \mathcal{H}_{xy}(\Delta_y) + \mathcal{H}_{xz}(\Delta_z), & \Phi_{x*}(\cdot) &: \mathbb{T}^3 \to \mathbb{T}_X, \\ \Phi_{y*}(\Delta) &= \mathcal{H}_{yx}(\Delta_x) + \mathcal{H}_{yy}(\Delta_y) + \mathcal{H}_{yz}(\Delta_z), & \Phi_{y*}(\cdot) &: \mathbb{T}^3 \to \mathbb{T}_Y, \\ \Phi_{z*}(\Delta) &= \mathcal{H}_{zx}(\Delta_x) + \mathcal{H}_{zy}(\Delta_y) + \mathcal{H}_{zz}(\Delta_z), & \Phi_{z*}(\cdot) &: \mathbb{T}^3 \to \mathbb{T}_Z, \end{aligned} \tag{4.23}$$

and each "$\mathcal{H}_{**}$" is a linear operator specified below. The diagonal operators[6] are (recall that $\mathcal{F} = \mathcal{A} \cdot (X, Y, Z)$)

$$\begin{aligned} \mathcal{H}_{xx}(\Delta_x) &= \Pi_X \langle \mathcal{B}_x, \mathcal{B}_x \rangle_{-1} \Delta_x - \Delta_x \langle \mathcal{F}, \mathcal{F} \rangle_{-1}, & \mathcal{B}_x &= \mathcal{A} \cdot (I, Y, Z), \\ \mathcal{H}_{yy}(\Delta_y) &= \Pi_Y \langle \mathcal{B}_y, \mathcal{B}_y \rangle_{-2} \Delta_y - \Delta_y \langle \mathcal{F}, \mathcal{F} \rangle_{-2}, & \mathcal{B}_y &= \mathcal{A} \cdot (X, I, Z), \\ \mathcal{H}_{zz}(\Delta_z) &= \Pi_Z \langle \mathcal{B}_z, \mathcal{B}_z \rangle_{-3} \Delta_z - \Delta_z \langle \mathcal{F}, \mathcal{F} \rangle_{-3}, & \mathcal{B}_z &= \mathcal{A} \cdot (X, Y, I). \end{aligned} \tag{4.24}$$

Since the Hessian operator is self-adjoint,[7] we give only the blocks of the "upper triangular part"

$$\begin{aligned} \mathcal{H}_{xy}(\Delta_y) &= \Pi_X \left(\left\langle \langle \mathcal{C}_{xy}, \mathcal{F} \rangle_{-(1,2)}, \Delta_y \right\rangle_{2,4;1,2} + \left\langle \langle \mathcal{B}_x, \mathcal{B}_y \rangle_{-(1,2)}, \Delta_y \right\rangle_{4,2;1,2} \right), \\ \mathcal{H}_{xz}(\Delta_z) &= \Pi_X \left(\left\langle \langle \mathcal{C}_{xz}, \mathcal{F} \rangle_{-(1,3)}, \Delta_z \right\rangle_{2,4;1,2} + \left\langle \langle \mathcal{B}_x, \mathcal{B}_z \rangle_{-(1,3)}, \Delta_z \right\rangle_{4,2;1,2} \right), \\ \mathcal{H}_{yz}(\Delta_z) &= \Pi_Y \left(\left\langle \langle \mathcal{C}_{yz}, \mathcal{F} \rangle_{-(2,3)}, \Delta_z \right\rangle_{2,4;1,2} + \left\langle \langle \mathcal{B}_y, \mathcal{B}_z \rangle_{-(2,3)}, \Delta_z \right\rangle_{4,2;1,2} \right), \end{aligned}$$

where we have also introduced $\mathcal{C}_{xy} = \mathcal{A} \cdot (I, I, Z)$, $\mathcal{C}_{xz} = \mathcal{A} \cdot (I, Y, I)$, and $\mathcal{C}_{yz} = \mathcal{A} \cdot (X, I, I)$. Observe that diagonal operators are Sylvester operators, and the off-diagonal operators have the form of 4-tensors mapping elements of one tangent space to another tangent space, for instance, $\mathcal{H}_{xz}(\cdot) : \mathbb{T}_Z \to \mathbb{T}_X$.

4.3. Coordinate representation for the gradient and the Hessian operator on the tangent space. Hessian (4.23) is still given in terms of the ambient Euclidean coordinate system. In order to obtain a linear system of equations with the correct dimension that is nonsingular in a neighborhood of a maximum, we introduce local coordinate expressions for the unknowns on the tangent space. We first see that the projections onto the tangent spaces can be represented as

$$\Pi_X = X_\perp X_\perp^\mathsf{T}, \qquad \Pi_Y = Y_\perp Y_\perp^\mathsf{T}, \qquad \Pi_Z = Z_\perp Z_\perp^\mathsf{T},$$

[6] Even if the Hessian is not a block matrix, we will refer to the operators $\mathcal{H}_{xx}$, $\mathcal{H}_{yy}$, etc., as diagonal operators and $\mathcal{H}_{xy}$, $\mathcal{H}_{xz}$, etc., as off-diagonal operators.

[7] The operator is still somewhat abstract in the sense that we have not specified any coordinate representation on the tangent space $\mathbb{T}^3$. However, considered as an operator on $\mathbb{T}^3$, it can be seen that the operator is self-adjoint.

where $X_\perp$, $Y_\perp$, $Z_\perp$, are defined as in section 3. In order to get a coordinate representation for the unknown tangents, we write them as [8, section 2.5]

$$\begin{aligned} &\Delta_x = X_\perp D_x, && D_x \in \mathbb{R}^{(J-r_1)\times r_1}, \\ &\Delta_y = Y_\perp D_y, && D_y \in \mathbb{R}^{(K-r_2)\times r_2}, \\ &\Delta_z = Z_\perp D_z, && D_x \in \mathbb{R}^{(L-r_3)\times r_3} \end{aligned} \tag{4.25}$$

(note that the coordinate matrices D_* are *not assumed to be diagonal*, even if the notation might be interpreted in that direction). With these coordinate expressions, we can repeat the derivation from after (4.18) and write the Hessian as a linear operator acting on $D = (D_x, D_y, D_z)$. We get

$$\widehat{\mathcal{H}}(D) = \left(\widehat{\Phi}_{x*}(D), \widehat{\Phi}_{y*}(D), \widehat{\Phi}_{z*}(D)\right),$$

where

$$\begin{aligned} \widehat{\Phi}_{x*}(D) &= X_\perp^{\mathsf{T}}\Phi_{x*}(\Delta) = \widehat{\mathcal{H}}_{xx}(D_x) + \widehat{\mathcal{H}}_{xy}(D_y) + \widehat{\mathcal{H}}_{xz}(D_z), \\ \widehat{\Phi}_{y*}(D) &= Y_\perp^{\mathsf{T}}\Phi_{y*}(\Delta) = \widehat{\mathcal{H}}_{yx}(D_x) + \widehat{\mathcal{H}}_{yy}(D_y) + \widehat{\mathcal{H}}_{yz}(D_z), \\ \widehat{\Phi}_{z*}(D) &= Z_\perp^{\mathsf{T}}\Phi_{z*}(\Delta) = \widehat{\mathcal{H}}_{zx}(D_x) + \widehat{\mathcal{H}}_{zy}(D_y) + \widehat{\mathcal{H}}_{zz}(D_z), \end{aligned} \tag{4.26}$$

and the "$\widehat{\mathcal{H}}_{**}$" terms are, as before, linear operators mapping elements from one tangent space to another (possibly the same) tangent space. The diagonal operators are

$$\begin{aligned} &\widehat{\mathcal{H}}_{xx}(D_x) = \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_x \right\rangle_{-1} D_x - D_x \langle \mathcal{F}, \mathcal{F}\rangle_{-1}, && \widehat{\mathcal{B}}_x = \mathcal{A} \cdot (X_\perp, Y, Z), \\ &\widehat{\mathcal{H}}_{yy}(D_y) = \left\langle \widehat{\mathcal{B}}_y, \widehat{\mathcal{B}}_y \right\rangle_{-2} D_y - D_y \langle \mathcal{F}, \mathcal{F}\rangle_{-2}, && \widehat{\mathcal{B}}_y = \mathcal{A} \cdot (X, Y_\perp, Z), \\ &\widehat{\mathcal{H}}_{zz}(D_z) = \left\langle \widehat{\mathcal{B}}_z, \widehat{\mathcal{B}}_z \right\rangle_{-3} D_z - D_z \langle \mathcal{F}, \mathcal{F}\rangle_{-3}, && \widehat{\mathcal{B}}_z = \mathcal{A} \cdot (X, Y, Z_\perp). \end{aligned} \tag{4.27}$$

The Hessian operator $\widehat{\mathcal{H}}$ is self-adjoint with respect to the inner product

$$\left\langle D, \widehat{\mathcal{H}}(E) \right\rangle_{\mathbb{T}^3} = \left\langle \widehat{\mathcal{H}}(D), E \right\rangle_{\mathbb{T}^3},$$

where

$$\langle D, E\rangle_{\mathbb{T}^3} = \langle D_x, E_x\rangle + \langle D_y, E_y\rangle + \langle D_z, E_z\rangle$$

and $D = (D_x, D_y, D_z)$ and $E = (E_x, E_y, E_z)$ are the coordinates for two tangents. Therefore, we give only the blocks of the "upper triangular part"

$$\begin{aligned} \widehat{\mathcal{H}}_{xy}(D_y) &= \left(\left\langle \left\langle \widehat{\mathcal{C}}_{xy}, \mathcal{F} \right\rangle_{-(1,2)}, D_y \right\rangle_{2,4;1,2} + \left\langle \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_y \right\rangle_{-(1,2)}, D_y \right\rangle_{4,2;1,2} \right), \\ \widehat{\mathcal{H}}_{xz}(D_z) &= \left(\left\langle \left\langle \widehat{\mathcal{C}}_{xz}, \mathcal{F} \right\rangle_{-(1,3)}, D_z \right\rangle_{2,4;1,2} + \left\langle \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_z \right\rangle_{-(1,3)}, D_z \right\rangle_{4,2;1,2} \right), \\ \widehat{\mathcal{H}}_{yz}(D_z) &= \left(\left\langle \left\langle \widehat{\mathcal{C}}_{yz}, \mathcal{F} \right\rangle_{-(2,3)}, D_z \right\rangle_{2,4;1,2} + \left\langle \left\langle \widehat{\mathcal{B}}_y, \widehat{\mathcal{B}}_z \right\rangle_{-(2,3)}, D_z \right\rangle_{4,2;1,2} \right), \end{aligned} \tag{4.28}$$

where $\widehat{\mathcal{C}}_{xy} = \mathcal{A} \cdot (X_\perp, Y_\perp, Z)$, $\widehat{\mathcal{C}}_{xz} = \mathcal{A} \cdot (X_\perp, Y, Z_\perp)$, and $\widehat{\mathcal{C}}_{yz} = \mathcal{A} \cdot (X, Y_\perp, Z_\perp)$.

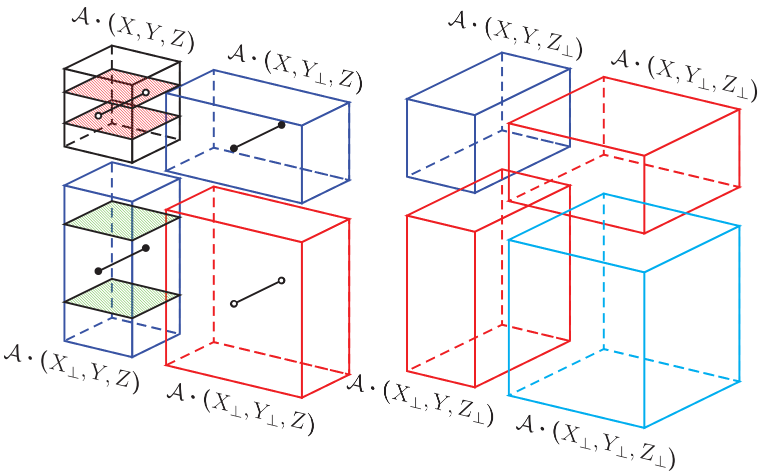

$\mathcal{A}\cdot(X,Y,Z)$
$\mathcal{A}\cdot(X,Y_\perp,Z)$
$\mathcal{A}\cdot(X,Y,Z_\perp)$
$\mathcal{A}\cdot(X,Y_\perp,Z_\perp)$
$\mathcal{A}\cdot(X_\perp,Y,Z)$
$\mathcal{A}\cdot(X_\perp,Y_\perp,Z)$
$\mathcal{A}\cdot(X_\perp,Y,Z_\perp)$
$\mathcal{A}\cdot(X_\perp,Y_\perp,Z_\perp)$

FIG. 4.1. *Illustration of the partial contractions in the Hessian. For better visibility, we have slided part of the tensor $\widehat{\mathcal{A}}$ to the right.*

In the coordinate representation (4.25), the Grassmann gradient (4.15)–(4.17) is given by

$$\nabla\widehat{\Phi} = \left(X_\perp^{\mathsf T}\Phi_x, Y_\perp^{\mathsf T}\Phi_y, Z_\perp^{\mathsf T}\Phi_z\right) = \left(\left\langle \widehat{\mathcal{B}}_x, \mathcal{F}\right\rangle_{-1}, \left\langle \widehat{\mathcal{B}}_y, \mathcal{F}\right\rangle_{-2}, \left\langle \widehat{\mathcal{B}}_z, \mathcal{F}\right\rangle_{-3}\right). \tag{4.29}$$

It is known [6] that, in general, the objective function (4.9) is not concave. In fact, it is easy to construct nonconcave examples using the coordinate representation of the Hessian.

PROPOSITION 4.1. *The maximization problem* (4.9) *can have local maxima.*

Proof. Consider the $2 \times 2 \times 2$ tensor

$$\mathcal{A}(:,:,1) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathcal{A}(:,:,2) = \begin{pmatrix} 0 & 0 \\ 0 & 8 \end{pmatrix},$$

and let $x = y = z = e_1$. The gradient is equal to zero, and the Hessian is $-I \in \mathbb{R}^{3\times 3}$ so that the point (x, y, z) is a local maximum. Clearly, it is not a global maximum. □

4.4. Interpretation of operators in the Hessian. The Hessian operator $\widehat{\mathcal{H}}$ consists of partial contractions involving the tensors

$$\begin{array}{llll} \mathcal{A}\cdot(X,Y,Z)\,, & \mathcal{A}\cdot(X_\perp,Y,Z)\,, & \mathcal{A}\cdot(X,Y_\perp,Z)\,, & \mathcal{A}\cdot(X,Y,Z_\perp)\,, \\ & \mathcal{A}\cdot(X_\perp,Y_\perp,Z)\,, & \mathcal{A}\cdot(X_\perp,Y,Z_\perp)\,, & \mathcal{A}\cdot(X,Y_\perp,Z_\perp)\,. \end{array}$$

These are blocks of the tensor $\widehat{\mathcal{A}} = \mathcal{A}\cdot((X\ X_\perp),(Y\ Y_\perp),(Z\ Z_\perp))$. The only block in $\widehat{\mathcal{A}}$ that does not occur in $\widehat{\mathcal{H}}$ is $\mathcal{A}\cdot(X_\perp,Y_\perp,Z_\perp)$. $\widehat{\mathcal{A}}$ is illustrated in Figure 4.1.

The partial contractions $\langle\cdot,\cdot\rangle_{-p}$ are matrices, whose elements are inner products between the slices in a subtensor. In Figure 4.1, we illustrate the inner products in $\widehat{\mathcal{H}}_{xx}$. In the off-diagonal operators, the inner products are between fibers in subtensors. For instance, in $\widehat{\mathcal{H}}_{xy}$, the inner products in $\langle\widehat{\mathcal{C}}_{xy},\mathcal{F}\rangle_{-(1,2)}$ are between fibers, illustrated with the ∘—∘ symbol, from $\mathcal{A}\cdot(X_\perp,Y_\perp,Z)$ and $\mathcal{A}\cdot(X,Y,Z)$. Similarly, the elements of $\langle\widehat{\mathcal{B}}_x,\widehat{\mathcal{B}}_y\rangle_{-(1,2)}$ are inner products between the •—• fibers from $\mathcal{A}\cdot(X_\perp,Y,Z)$ and $\mathcal{A}\cdot(X,Y_\perp,Z)$.

4.5. Matricizing the Hessian operator. It is now straightforward to matricize the operators in the Hessian and vectorize D_x, D_y, and D_z to obtain a standard matrix-vector linear system.

The "xx" block in (4.27) has the form

$$\widehat{\mathcal{H}}_{xx}(D_x) = \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_x \right\rangle_{-1} D_x - D_x \left\langle \mathcal{F}, \mathcal{F} \right\rangle_{-1}. \tag{4.30}$$

Observing that the contracted tensors are matrices and with straightforward vectorization of matrix products [13, Chapter 4.3], we get

$$\operatorname{vec}\left(\widehat{\mathcal{H}}_{xx}(D_x)\right) = \left(I \otimes \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_x \right\rangle_{-1} + \left\langle \mathcal{F}, \mathcal{F} \right\rangle_{-1} \otimes I\right) d_x \equiv \widehat{H}_{xx} d_x,$$

where $d_x = \operatorname{vec}(D_x)$. The other diagonal blocks are treated analogously.

The off-diagonal blocks in (4.28) consist of two 4-tensors acting on matrices. The "xy" block is given by

$$\widehat{\mathcal{H}}_{xy}(D_y) = \left(\left\langle \widehat{\mathcal{H}}^1_{xy}, D_y \right\rangle_{2,4;1,2} + \left\langle \widehat{\mathcal{H}}^2_{xy}, D_y \right\rangle_{4,2;1,2} \right),$$

where $\widehat{\mathcal{H}}^1_{xy} = \langle \widehat{\mathcal{C}}_{xy}, \mathcal{F} \rangle_{-(1,2)}$ and $\widehat{\mathcal{H}}^2_{xy} = \langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_y \rangle_{-(1,2)}$. In $\widehat{\mathcal{H}}^1_{xy}$, we map the first and third modes to the rows and second and fourth modes to the columns of the matrix. In $\widehat{\mathcal{H}}^2_{xy}$, the ordering of the row modes is the same, but the column modes are four and two. The vectorized form of the operation $\widehat{\mathcal{H}}_{xy}(D_y)$ is

$$\operatorname{vec}\left(\widehat{\mathcal{H}}_{xy}(D_y)\right) = \left(\widehat{H}^{1\,(1,3;2,4)}_{xy} + \widehat{H}^{2\,(1,3;4,2)}_{xy}\right) d_y \equiv \widehat{H}_{xy} d_y,$$

where $d_y = \operatorname{vec}(D_y)$.

After matricizing all blocks of $\widehat{\mathcal{H}}$ and vectorizing the gradients, we obtain the matrix form for the Newton equation

$$\widehat{H} d = \begin{pmatrix} \widehat{H}_{xx} & \widehat{H}_{xy} & \widehat{H}_{xz} \\ \widehat{H}_{yx} & \widehat{H}_{yy} & \widehat{H}_{yz} \\ \widehat{H}_{zx} & \widehat{H}_{zy} & \widehat{H}_{zz} \end{pmatrix} \begin{pmatrix} d_x \\ d_y \\ d_z \end{pmatrix} = - \begin{pmatrix} g_x \\ g_y \\ g_z \end{pmatrix} = -g, \tag{4.31}$$

where $g_x = \operatorname{vec}(\langle \widehat{\mathcal{B}}_x, \mathcal{F} \rangle_{-1})$, $g_y = \operatorname{vec}(\langle \widehat{\mathcal{B}}_y, \mathcal{F} \rangle_{-2})$, and $g_z = \operatorname{vec}(\langle \widehat{\mathcal{B}}_z, \mathcal{F} \rangle_{-3})$ are the vectorized gradients from (4.29).

4.6. Generalizing to higher order tensors. Note that the representations for the Grassmann gradient and Hessian in section 4.2 can easily be generalized to the case of 4-tensors and higher. Assume that the objective function $\Phi(X, Y, Z, W) = \frac{1}{2}\|\mathcal{A}\cdot(X, Y, Z, W)\|_F$ is to be maximized over a product of four Grassmann manifolds. Then the diagonal operators in the Hessian (4.24) have to be modified by introducing an extra matrix W, i.e., we put $\mathcal{B}_x = \mathcal{A}\cdot(I, Y, Z, W)$, etc., and then we add a fourth diagonal block

$$\mathcal{H}_{ww}(\Delta_w) = \Pi_W \left\langle \mathcal{B}_w, \mathcal{B}_w \right\rangle_{-4} \Delta_w - \Delta_w \left\langle \mathcal{F}, \mathcal{F} \right\rangle_{-4},$$

where now $\mathcal{F} = \mathcal{A}\cdot(X, Y, Z, W)$ and $\mathcal{B}_w = \mathcal{A}\cdot(X, Y, Z, I)$. The off-diagonal operators are modified analogously. For instance,

$$\mathcal{H}_{xw}(\Delta_w) = \Pi_X \left(\left\langle \langle \mathcal{C}_{xw}, \mathcal{F} \rangle_{-(1,4)}, \Delta_w \right\rangle_{2,4;1,2} + \left\langle \langle \mathcal{B}_x, \mathcal{B}_w \rangle_{-(1,4)}, \Delta_w \right\rangle_{4,2;1,2} \right),$$

where $\mathcal{B}_x$ and $\mathcal{B}_w$ are as above and $\mathcal{C}_{xw} = \mathcal{A}\cdot(I, Y, Z, I)$. The modification needed for the gradient is one additional term: $\Pi_W \Phi_w$, which is a simple modification of (4.13).

5. Implementation and experimental results.

Given the analysis from the previous section together with the TensorToolbox [2], the algorithmic implementation in MATLAB is straightforward. A pseudocode is given in Algorithm 1. See also Appendix B for details on computational complexity.

Algorithm 1 NG algorithm.

```
Given tensor $\mathcal{A}$ and starting points $(X_0, Y_0, Z_0) \in \mathrm{Gr}^3$
repeat
    compute Grassmann gradient $\nabla\widehat{\Phi}$ given in (4.29)
    compute Grassmann Hessian $\widehat{\mathcal{H}}$ from (4.26)
    matricize $\widehat{\mathcal{H}}$ and vectorize $\nabla\widehat{\Phi}$ to form NG equations (4.31)
    solve $D = (D_x, D_y, D_z)$ from the Newton equation on the tangent spaces
    take a geodesic step along the direction given by $D$ to obtain new iterates (X,Y,Z)
until $\|\nabla\widehat{\Phi}\|/\Phi <$ TOL.
```

In this section, we report the results of some preliminary numerical experiments where we compare the NG algorithm with higher order orthogonal iteration (HOOI) [6]. Each HOOI iteration consists of three steps, where, in each step, two of the unknown matrices are considered as fixed and the third is updated. But, first, we would like to make a few remarks on the convergence of HOOI, which is related to the alternating least square (ALS) method [18].

The sequence of iterates generated by HOOI, and other ALS methods, converge asymptotically linearly, proved by Ruhe and Wedin [22], to a point. But this point need not be a local stationary point of the considered objective function [20, pp. 53–54], [21]. Thus, as an algorithm, HOOI is not guaranteed to converge to a local minimizer. This scenario was, in fact, occasionally encountered during our tests. Further, the two algorithms NG and HOOI need not converge to the same local minimizer (under the assumption that HOOI does converge to a local minimizer). This was also encountered during our tests.

5.1. Test 1—signal and noise tensors.

Our first experiment was tailored to simulate a “signal tensor” with low rank and added normally distributed noise. We used two $20 \times 20 \times 20$ tensors $\mathcal{A}_1 = \mathcal{B}_1 + \rho\mathcal{E}_1$ and $\mathcal{A}_2 = \mathcal{B}_2 + \rho\mathcal{E}_2$, where we chose $\mathcal{B}_1$ and $\mathcal{B}_2$ as random tensors with ranks $(10, 10, 10)$ and $(15, 15, 15)$, respectively. Thus, $\mathcal{B}_1$ was constructed from a $10 \times 10 \times 10$ tensor with normally distributed $(N(0, 1))$ elements; that tensor was then projected up to dimension $20 \times 20 \times 20$ by multiplying it in each mode by a 20×10 matrix with orthonormal columns. The elements of the noise tensors $\mathcal{E}_1$ and $\mathcal{E}_2$ were chosen normally distributed $(N(0, 1))$, and the level of noise was controlled by ρ, which was taken equal to 0.1. In both cases, we computed a rank-$(5, 5, 5)$ approximation. As initial iterates we chose random matrices with orthonormal columns and performed 10 HOOI iterations before the NG method was started. The plots marked with NG-1 and NG-2 in Figure 5.1 show the convergence history of both methods. Observe that the initial points are the same, therefore, the two plots are on top of each other for the first 10 iterations.

5.2. Test 2—random tensors.

We approximated two random $20 \times 20 \times 20$ tensors (the elements were in $N(0, 1)$) by a rank-$(5, 5, 5)$ tensor. Both algorithms were initialized by HOSVD, and we performed 20 HOOI iterations before NG was employed. The plots NG-3 and NG-4 in Figure 5.1 show the convergence history. The quadratic convergence of the NG algorithm is clearly visible.

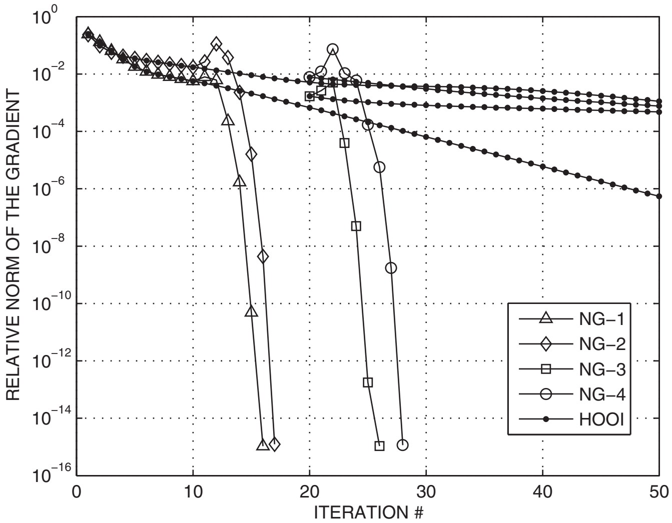


FIG. 5.1. *Convergence history—iterations versus the relative gradient norm* $\|\nabla\widehat{\Phi}\|/\Phi$. *All tests are rank-*$(5,5,5)$ *approximations of* $20 \times 20 \times 20$ *tensors. NG-*1 : *signal tensor with rank-*$(10,10,10)$ *and noise. NG-2: signal tensor with rank-*$(15,15,15)$ *and noise. NG-*3 *and NG-*4*: random tensors. The* 20 *initial HOOI iterations for these two tests are omitted for clarity.*

5.3. Test 3—noncubic tensors. We also did a third set of tests, now with tensors of dimension $25 \times 30 \times 20$, which were approximated with tensors of rank-$(7,3,5)$. The first two tests were set up as in section 5.1, i.e., one case with a signal tensor of rank-$(10,10,10)$ and another case with a signal tensor of rank-$(15,15,15)$, same noise level $\rho = 0.1$. The NG method was initialized with random matrices and 10 HOOI iterations. Then two sets were set up as in section 5.2, i.e., with random tensors, in which the NG algorithm was initiated with HOSVD and 20 HOOI iterations. The corresponding plots were very similar to those given in Figure 5.1.

5.4. Comments on more tests and relative fit. In our experience, the HOOI method may have an acceptable convergence rate for low-rank signal tensors with noise of small magnitude. In general, the closer the rank of the approximating tensor to the correct rank of the signal tensor, the faster the convergence.

We performed tests where the ranks of the "signal tensor" and the approximating tensor coincided. Then, for values of ρ less than or approximately equal to 0.1, the convergence of HOOI was very rapid. In these tests, there was a clear gap in the higher order singular values (see Theorem 2.4), indicating a relation of the convergence rate in HOOI to the gap at the cut-off level.

On the other hand, approximating a full rank tensor with HOOI can have very slow convergence. For example, the HOOI run, with the same initial iterates as NG-3 shown in Figure 5.1, requires more than 800 iterations to achieve the same accuracy as NG does in 6 iterations. In some cases, HOOI also requires a large number of iterations before the convergence is stabilized to a constant linear rate.

The relative fit $\|\mathcal{A} - \mathcal{B}\|/\|\mathcal{A}\|$, where $\mathcal{B}$ is the approximating tensor of lower rank, was in the order 0.3–0.8 for the performed tests. The fit depends on the rank and the dimensions of the original tensor, the rank of the approximating tensor, and the noise level.

Considering the subspace angles between the final iterate and all previous iterates, both algorithms generate plots very similar to those given in Figure 5.1.

5.5. Computational complexity. Naturally, the price to be paid for the fast convergence of the NG method is a higher computational cost per iteration. Assume, for simplicity, that we have an $n \times n \times n$ tensor which is approximated by an $r \times r \times r$ tensor. Each iteration in HOOI involves six tensor by matrix products and three maximization problems, e.g., $\mathcal{A} \cdot (I, Y, Z)$ and

$$\max_{X^{\mathsf{T}}X=I} \left\| X^{\mathsf{T}} A^{(1)}(Y \otimes Z) \right\|.$$

The solution is the dominant r-dimensional left singular subspace of the matrix $A^{(1)}(Y \otimes Z)$, which we assume is computed with SVD [10, section 5.4.5]. Then, the approximate amount of flops (floating point additions and multiplications) per iteration is $6n^3r$ for the tensor-matrix product and $18nr^4 + 33r^6$ for the dominant subspace (based on the table in [10, section 5.4.5]; note that faster SVD algorithms are available and will be implemented in the next version of LAPACK), which gives

$$\text{flops(HOOI)} \approx 6n^3r + 18nr^4 + 33r^6.$$

Each iteration in the NG algorithm is dominated by the solution of the Newton's equations (4.31), which amounts to

$$\text{flops(NG)} \approx 9(n-r)^3r^3. \tag{5.1}$$

This is for computing the Cholesky factorization of the $3(n-r)r \times 3(n-r)r$ Hessian matrix. The computation of the Hessian itself is of lower complexity:

$$\text{flops(Hessian)} \approx 18n^3r + 12(n-r)^2r^3. \tag{5.2}$$

Details on how this expression was derived are given in Appendix B. Observe that, in local coordinates, no tangent vectors are transported, only the iterates (X_k, Y_k, Z_k) are moved along geodesics. The price for this is the computation of basis matrices $X_\perp$, $Y_\perp$, and $Z_\perp$ for the three tangent space at every point. The amount of computations due to the geodesic movements on the Grassmann manifolds is negligible.

5.6. Optimization issues. In our formulation of problem (3.1), the original constraints due to the manifold structure have been incorporated in the computation of the Grassmann gradient and the Grassmann Hessian of the objective function. Together with the parametrization of section 4.3, the problem reduces to an unconstrained optimization problem. When solving this, one must deal with standard optimization issues, such as obtaining good starting points, indefiniteness of the Hessian, line search, etc.; see, e.g., [20] for details.

As is always the case with Newton's method, the choice of a good starting point is important. One obvious alternative is to start with X_0, Y_0, and Z_0 given by the first r_1, r_2, and r_3 columns of the HOSVD matrices U, V, and W, respectively. But this choice is often not good enough. In our experiments with random tensors, the Hessian was almost always indefinite for points given by the HOSVD. When we, in addition, performed initial HOOI iterations, then, within a reasonable amount of steps, we got to the proximity of the local minimum where we could employ the Newton algorithm. Alternatively, one could incorporate a trust region method together with the NG method [14] or perform the initial steps with a conjugate gradient algorithm on the product of Grassmann manifolds.

6. Conclusion and future work. In this paper, we have formulated the tensor approximation problem defined on the product of Grassmann manifolds and derived Newton's method for this problem. We have showed quadratic convergence of the algorithm in the proximity of a local minimizer.

The general tensor matricization introduced in section 2.2, the contracted tensor products, and the tensor algebraic identities from section 2.3 have been used both for the analysis of the differentiated expressions of the objective function and for the algorithmic implementation. The generalization from 3-tensors to higher order tensors is straightforward with the presented tensor algebraic analysis.

Our present and future work include further analysis of the theoretical aspects of the best approximation problem. For computational and memory efficiency, the implementation details for the NG algorithms need to be investigated. An alternative approach for this and similar problems, which we are presently pursuing, is to develop quasi-Newton methods on (products of) Grassmann manifolds [23].

Appendix A. Proof of Lemma 2.3. To prove the identities, we will use the definition for a contracted tensor product to verify that elements of the resulting matrices in both sides are the same. Let $j < i$ and assume we have the following dimensions:

$$\begin{aligned}&\mathcal{B} \in \mathbb{R}^{K_1 \times \cdots \times K_j \times \cdots \times K_i \times \cdots \times K_N},\\&Q \in \mathbb{R}^{K_j \times L_j}.\end{aligned}$$

The dimensions of the modes of tensor $\mathcal{C}$ are assumed to be the same as those in $\mathcal{B}$, except modes j and i, which are taken to be L_j and L_i. We will show that

$$\left\langle \mathcal{B} \cdot (Q)_j \,, \mathcal{C} \right\rangle_{-i} = \left\langle \langle \mathcal{B}, \mathcal{C} \rangle_{-(i,j)} \,, Q \right\rangle_{1,3;1,2}. \tag{A.1}$$

Then, for the first argument on the left-hand side, we have

$$\mathcal{B} \cdot (Q)_j =: \mathcal{D} \in \mathbb{R}^{K_1 \times \cdots \times L_j \times \cdots \times K_i \times \cdots \times K_N},$$

where the elements are given by

$$d_{k_1 \cdots l_j \cdots k_i \cdots k_N} = \sum_{k_j} a_{k_1 \cdots k_j \cdots k_i \cdots k_N} q_{k_j l_j}.$$

The expression on the left-hand side of (A.1) becomes

$$\langle \mathcal{D}, \mathcal{C} \rangle_{-i} =: \mathcal{E} \in \mathbb{R}^{K_i \times L_i},$$

where the entries are

$$\begin{aligned}e_{k_i l_i} &= \sum_{\substack{k_1,\ldots,k_{j-1},l_j\\ k_{j+1},\ldots,k_{i-1}\\ k_{i+1},\ldots,k_N}} d_{k_1 \cdots l_j \cdots k_i \cdots k_N} c_{k_1 \cdots l_j \cdots l_i \cdots k_N}\\&= \sum_{\substack{k_1,\ldots,k_j,l_j\\ k_{j+1},\ldots,k_{i-1}\\ k_{i+1},\ldots,k_N}} a_{k_1 \cdots k_j \cdots k_i \cdots k_N} q_{k_j l_j} c_{k_1 \cdots l_j \cdots l_i \cdots k_N}\\&= \sum_{k_j,l_j} q_{k_j l_j} \sum_{\substack{k_1,\ldots,k_{j-1}\\ k_{j+1},\ldots,k_{i-1}\\ k_{i+1},\ldots,k_N}} a_{k_1 \cdots k_j \cdots k_i \cdots k_N} c_{k_1 \cdots l_j \cdots l_i \cdots k_N},\end{aligned}$$

which shows that (A.1) holds. The other cases are analogous.

Appendix B. Computation of the gradient and the Hessian. The computation of the gradient $\nabla\widehat{\Phi}$ from (4.29) and the Hessian $\widehat{\mathcal{H}}$ from (4.26) involve the following terms: $\mathcal{F} = \mathcal{A} \cdot (X, Y, Z)$ and

$$\begin{aligned} \widehat{\mathcal{B}}_x &= \mathcal{A} \cdot (X_\perp, Y, Z)\,, & \widehat{\mathcal{B}}_y &= \mathcal{A} \cdot (X, Y_\perp, Z)\,, & \widehat{\mathcal{B}}_z &= \mathcal{A} \cdot (X, Y, Z_\perp)\,, \\ \widehat{\mathcal{C}}_{xy} &= \mathcal{A} \cdot (X_\perp, Y_\perp, Z)\,, & \widehat{\mathcal{C}}_{xz} &= \mathcal{A} \cdot (X_\perp, Y, Z_\perp)\,, & \widehat{\mathcal{C}}_{yz} &= \mathcal{A} \cdot (X, Y_\perp, Z_\perp)\,. \end{aligned}$$

For efficiency, the computation of these terms are arranged into a series[8] of tensor times matrix multiplies. These computations amount to $18n^3r$ flops, where we approximate $(n - r) \approx n$ and omit terms of lower complexity. This gives the first term in (5.2).

Given the $\widehat{\mathcal{B}}_*$ and $\widehat{\mathcal{C}}_{**}$ tensors, there follow contracted tensor products in two modes between 3-tensors yielding matrices

$$\begin{array}{lll} \langle \mathcal{F}, \mathcal{F} \rangle_{-1}\,, & \langle \mathcal{F}, \mathcal{F} \rangle_{-2}\,, & \langle \mathcal{F}, \mathcal{F} \rangle_{-3}\,, \\ \left\langle \widehat{\mathcal{B}}_x, \mathcal{F} \right\rangle_{-1}, & \left\langle \widehat{\mathcal{B}}_y, \mathcal{F} \right\rangle_{-2}, & \left\langle \widehat{\mathcal{B}}_z, \mathcal{F} \right\rangle_{-3}, \\ \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_x \right\rangle_{-1}, & \left\langle \widehat{\mathcal{B}}_y, \widehat{\mathcal{B}}_y \right\rangle_{-2}, & \left\langle \widehat{\mathcal{B}}_z, \widehat{\mathcal{B}}_z \right\rangle_{-3}, \end{array}$$

and contracted tensor products in one mode between 3-tensors yielding 4-tensors

$$\begin{array}{lll} \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_y \right\rangle_{-(1,2)}, & \left\langle \widehat{\mathcal{B}}_x, \widehat{\mathcal{B}}_z \right\rangle_{-(1,3)}, & \left\langle \widehat{\mathcal{B}}_y, \widehat{\mathcal{B}}_z \right\rangle_{-(2,3)}, \\ \left\langle \widehat{\mathcal{C}}_{xy}, \mathcal{F} \right\rangle_{-(1,2)}, & \left\langle \widehat{\mathcal{C}}_{xz}, \mathcal{F} \right\rangle_{-(1,3)}, & \left\langle \widehat{\mathcal{C}}_{yz}, \mathcal{F} \right\rangle_{-(2,3)}. \end{array}$$

The contribution to the second term in (5.2) comes from the computation of the 4-tensors. The computations of the matrices, which are involved in the gradient, are negligible.

Acknowledgments. We are grateful to the reviewers and to Lek-Heng Lim for their constructive criticism, which has helped us to improve the presentation.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2007.

[2] B. Bader and T. Kolda, *Algorithm* 862: *MATLAB tensor classes for fast algorithm prototyping,* ACM Trans. Math. Software, 32 (2006), pp. 635–653.

[3] B. W. Bader and T. G. Kolda, *Efficient MATLAB computations with sparse and factored tensors,* SIAM J. Sci. Comput., 30 (2007), pp. 205–231.

[4] L. De Lathauwer, L. Hoegaerts, and J. Vandewalle, *A Grassmann-Rayleigh quotient iteration for dimensionality reduction in ICA,* in Proceedings of the 5th International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2004), Granada, Spain, 2004, pp. 335–342.

[5] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition,* SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.

[6] 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 Anal. Appl., 21 (2000), pp. 1324–1342.

[7] V. de Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem,* SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127.

[8] A total of 14 tensor times matrix multiplies.

[8] A. Edelman, T. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints,* SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[9] L. Eldén, *Matrix Methods in Data Mining and Pattern Recognition,* SIAM, Philadelphia, 2007.

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

[11] R. A. Harshman, *An index formalism that generalizes the capabilities of matrix notation and algebra to n-way arrays,* J. Chemometrics, 15 (2001), pp. 689–714.

[12] F. L. Hitchcock, *Multiple invariants and generalized rank of a p-way matrix or tensor,* J. Math. Phys. Camb., 7 (1927), pp. 39–70.

[13] R. J. Horn and C. R. Johnson, *Topics in Matrix Analysis,* Cambridge University Press, Cambridge, 1991.

[14] M. Ishteva, L. De Lathauwer, P.-A. Absil, and S. Van Huffel, *Dimensionality reduction for higher-order tensors: Algorithms and applications,* Int. J. Pure Appl. Math., 42 (2008), pp. 337–343.

[15] H. A. L. Kiers, *Towards a standardized notation and terminology in multiway analysis,* J. Chemometrics, 14 (2000), pp. 106–125.

[16] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry,* Interscience, New York, 1963.

[17] T. G. Kolda, *Multilinear Operators for Higher-order Decompositions,* Technical report SAND2006-2081, Sandia National Laboratories, Albuquerque, NM, 2006.

[18] P. M. Kroonenberg, *Principal component analysis of three-mode data by means of alternating least squares algorithms,* Psychometrika, 45 (1980), pp. 69–97.

[19] E. Lundström and L. Eldén, *Adaptive eigenvalue computations using Newton's method on the Grassmann manifold,* SIAM J. Matrix Anal. Appl., 23 (2002), pp. 819–839.

[20] J. Nocedal and S. J. Wright, *Numerical Optimization,* Springer Ser. Oper. Res., Springer, New York, 1999.

[21] M. J. D. Powell, *On search directions for minimization algorithms,* Math. Program., 4 (1973), pp. 193–201.

[22] A. Ruhe and P.-Å Wedin, *Algorithms for separable nonlinear least squares problems,* SIAM Rev., 22 (1980), pp. 318–337.

[23] B. Savas and L.-H. Lim, *Best Multilinear Rank Approximation of Tensors with Quasi-Newton Methods on Grassmannians,* Technical report, Department of Mathematics, Linköping University, Linköping, Sweden 2008.

[24] L. R. Tucker, *The extension of factor analysis to three-dimensional matrices,* in Contributions to Mathematical Psychology, H. Gulliksen and N. Frederiksen, eds., Holt, Rinehart and Winston, New York, 1964, pp. 109–127.

[25] L. R. Tucker, *Some mathematical notes on three-mode factor analysis,* Psychometrika, 31 (1966), pp. 279–311.

[26] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear analysis of image ensembles: Tensorfaces,* in Proceedings of the 7th European Conference on Computer Vision (ECCV'02), Lect. Notes Comput. Sci. 2350, Springer, New York, 2002, pp. 447–460.

[27] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear subspace analysis of image ensembles,* in IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03), Madison WI, 2003, pp. 93–99.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 272–288

INEXACT KLEINMAN–NEWTON METHOD FOR RICCATI EQUATIONS*

F. FEITZINGER[†], T. HYLLA[†], AND E. W. SACHS[‡]

Abstract. In this paper we consider the numerical solution of the algebraic Riccati equation using Newton's method. We propose an inexact variant which allows one control the number of the inner iterates used in an iterative solver for each Newton step. Conditions are given under which the monotonicity and global convergence result of Kleinman also hold for the inexact Newton iterates. Numerical results illustrate the efficiency of this method.

Key words. Riccati, Kleinman–Newton, inexact Newton

AMS subject classifications. 49M15, 49N35, 93B52

DOI. 10.1137/070700978

1. Introduction.

The numerical solution of Riccati equations for large scale feedback control systems is still a formidable task. In order to reduce computing time in the context of Kleinman–Newton methods, it is mandatory that one uses iterative solvers for the solution of the linear systems occurring at each iteration.

In such an approach, it is important to control the accuracy of the solution of the linear systems at each Newton step in order to gain efficiency, but not to lose the overall fast convergence properties of Newton's method. This can be achieved in the framework of inexact Newton's methods.

In his classical paper, Kleinman [13] applied Newton's method to the algebraic Riccati equation, a quadratic equation for matrices of the type:

$$A^T X + XA - XBB^T X + C^T C = 0,$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{l\times n}$.

At each Newton step, a Lyapunov equation

$$X_{k+1}\left(A - BB^T X_k\right) + \left(A - BB^T X_k\right)^T X_{k+1} = -X_k BB^T X_k - C^T C$$

needs to be solved to obtain the next iterate X_{k+1}.

In the literature, several variants of this method have been proposed. The approach taken by Banks and Ito [1] suggests to apply Chandrasekhar's method for the initial iterates and then use the computed iterate as a starting matrix for the Kleinman–Newton method. Rosen and Wang [25] apply a multilevel approach to the solution of large scale Riccati equations. Navasca and Morris [19], [20] use the Kleinman–Newton method with a modified ADI method for the Lyapunov solvers to find the feedback gain matrix directly for a discretized version of a parabolic optimal

*Received by the editors August 23, 2007; accepted for publication (in revised form) by J. H. Brandts November 25, 2008; published electronically March 18, 2009.
http://www.siam.org/journals/simax/31-2/70097.html

[†]Fachbereich IV, Abteilung Mathematik, Universität Trier, 54286 Trier, Germany (f.feitzinger@gmx.de, hylla@uni-trier.de).

[‡]Fachbereich IV, Abteilung Mathematik, Universität Trier, 54286 Trier, Germany (sachs@uni-trier.de) and Interdisciplinary Center for Applied Mathematics, Virginia Tech, Blacksburg, VA 24061 (sachs@icam.vt.edu).

control problem. They also consider a version where the gain matrix is computed directly.

Benner and his co-authors use variants of the Kleinman–Newton algorithm to solve Riccati equations in their papers [2], [4], [5]. They utilize ADI methods for solving the Lyapunov equations at each step and address issues like parameter selection for ADI and parallelization among others. Incorporating line searches in a Newton procedure can often lead to a reduction in the number of iterations. This has been the focus of the research in Benner and Byers [3] and Guo and Laub [11]. The case, when the Jacobian of the nonlinear map describing the Riccati equation is singular at the solution, has been considered by Guo and Lancaster in [10].

In a recent paper Burns, Sachs, and Zietsman [6] give conditions under which the Kleinman–Newton method is mesh independent, i.e., the number of iterates remains virtually constant when the discretization of the underlying optimal control problem is refined.

Large scale Lyapunov equations usually require the use of iterative solvers like Smith's method or versions of the ADI method; for systems like this, the inexact Newton's method gives a rigorous guideline for the termination of the inner iteration for the Lyapunov equation while retaining the fast local rate of convergence. Another major effect in saving computing time is the possibility to terminate the inner iteration early when the iterates X_k are still far away from the solution of the Riccati equation. For a discussion on these methods see, for example, Kelley [12].

Whereas these aspects are typical for inexact Newton methods, the application to a Riccati equation bears some special features. Kleinman observed that the convergence is more global than usual; i.e., the starting matrix X_0 does not need to lie in a neighborhood of the solution X_∞. The proof is based on monotonicity properties of the iterates X_k as pointed out below. Obviously, this monotonicity is lost, when the iterates X_k are computed inexactly and, as a consequence, the global convergence feature of Kleinman–Newton does no longer hold. In this paper we also address the question, under which conditions the monotone convergence behavior and, hence, the larger convergence radius is maintained for the inexact Kleinman–Newton method.

The paper is organized as follows: In the next section we state the well-known convergence results for the exact Kleinman–Newton method in the case of Riccati equations and for the inexact Newton method in the general case. In the following section we formulate the inexact Newton method applied to the Riccati equation and give the convergence statement. Section 4 contains convergence results of the inexact Kleinman–Newton method including monotonicity statements for the iterates. This is achieved under certain assumptions on the residuals of the inexact Lyapunov solver. A condition on the size of the residual guarantees that the inexact Newton iterates are well defined. A stronger condition on the residuals yields the quadratic rate of convergence. The assumption on the starting data is the same as for the exact Kleinman–Newton method.

In section 5 we consider several iterative solvers for the Lyapunov equations like Smith's method and variants of the ADI method. We show how the previous conditions for the monotone convergence relate to the iterative solvers. This is followed by a section on numerical results for a discretized two-dimensional parabolic control problem. The convergence is illustrated and the savings in computing time is documented. The last section deals with another variant of the Kleinman–Newton method. We show that the inexact version of this method is unstable. The residuals accumulate as the iteration progresses, and, hence, this version should not be used in an inexact framework.

2. Inexact Newton method.

The algebraic Riccati equation presented in the introduction can be written as a nonlinear system of equations.

The goal is to find a symmetric matrix $X \in \mathbb{R}^{n\times n}$ with $\mathcal{F}(X) = 0$, where the map $\mathcal{F} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is defined by

$$\mathcal{F}(X) = A^T X + XA - XBB^T X + C^T C. \tag{2.1}$$

If one applies Newton's method to this system, one has to compute the derivative at X, symmetric, given by

$$\begin{aligned}\mathcal{F}'(X)(Y) &= A^T Y + YA - YBB^T X - XBB^T Y \\ &= \left(A - BB^T X\right)^T Y + Y\left(A - BB^T X\right) \qquad \forall\ Y \in \mathbb{R}^{n\times n}.\end{aligned} \tag{2.2}$$

In Newton's method, the next iterate is obtained by solving the Newton system

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) = -\mathcal{F}(X_k) \quad \text{or} \quad \mathcal{F}'(X_k)X_{k+1} = \mathcal{F}'(X_k)X_k - \mathcal{F}(X_k). \tag{2.3}$$

For the Riccati equation the computation of a Newton step requires the solution of a Lyapunov equation. Corresponding to the second part of (2.3) we obtain

$$X_{k+1}\left(A - BB^T X_k\right) + \left(A - BB^T X_k\right)^T X_{k+1} = -X_k BB^T X_k - C^T C, \tag{2.4}$$

which is a Lyapunov equation for X_{k+1}. This method is well understood and analyzed. It does not only exhibit locally a quadratic rate of convergence, but has also a monotone convergence property which is not so common for Newton's method and which is due to the quadratic form of $\mathcal{F}$ and the monotonicity of $\mathcal{F}'$. For this to hold, we introduce and impose the following definition and assumption:

DEFINITION 2.1. *Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{l\times n}$. A pair (A, BB^T) is called stabilizable if there is a feedback matrix $K \in \mathbb{R}^{n\times n}$ such that $A - BB^T K$ is stable, which means that $A - BB^T K$ has only eigenvalues in the open left halfplane. $(C^T C, A)$ is called detectable if and only if $(A^T, C^T C)$ is stabilizable.*

In the following sections we make the assumption:

Assumption 2.2. (A, BB^T) is stabilizable and $(C^T C, A)$ is detectable.

Note that by [14, Lemma 4.5.4] the first assumption implies the existence of a matrix X_0 such that $A - BB^T X_0$ is stable.

As a common abbreviation we set

$$A_k := \left(A - BB^T X_k\right), \quad k \in \mathbb{N}_0$$

and $A \le B$ means that the matrix $A - B$ is negative semidefinite. Then the next theorem is well known; see, e.g., Kleinman [13], Mehrmann [18], or Lancaster and Rodman [14].

THEOREM 2.3. *Let $X_0 \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite such that $A - BB^T X_0$ is stable and let Assumption 2.2 hold. Then the Newton iterates X_k defined by*

$$X_{k+1}A_k + A_k^T X_{k+1} = -X_k BB^T X_k - C^T C$$

converge to some X_∞ such that $A - BB^T X_\infty$ is stable, and it solves the Riccati equation $\mathcal{F}(X_\infty) = 0$. Furthermore, the iterates have a monotone convergence behavior

$$0 \le X_\infty \le \cdots \le X_{k+1} \le X_k \le \cdots \le X_1$$

and quadratic convergence.

In the past decade, a variant of Newton's method has become quite popular in several areas of applications, the so-called inexact Newton's method. In this variant, it is no longer necessary to solve the Newton equation exactly for the Newton step, but it is possible to allow for errors in the residual. In particular, this is useful if iterative solvers are used for the solution of the linear Newton equation. We cite a theorem in Kelley [12, p. 99].

THEOREM 2.4. *Let $F : \mathbb{R}^N \to \mathbb{R}^N$ have a Lipschitz-continuous derivative in a neighborhood of some $x_\infty \in \mathbb{R}^N$ with $F(x_\infty) = 0$ and $F'(x_\infty)$ invertible. Then there exist $\delta > 0$ and $\bar{\eta}$ such that for all $x_0 \in \mathcal{B}(x_\infty, \delta)$ the inexact Newton iterates*

$$x_{k+1} = x_k + s_k,$$

where s_k satisfies

$$\|F'(x_k)s_k + F(x_k)\| \le \eta_k \|F(x_k)\|, \qquad \eta_k \in [0, \bar{\eta}]$$

converge to x_∞. Furthermore, we have the following rate estimates:

The rate of convergence is at least linear. If, in addition, $\eta_k \to 0$, then we obtain a superlinear rate and if $\eta_k \le K_\eta \|F(x_k)\|$ for some $K_\eta > 0$, then we have a quadratic rate of convergence.

Our goal in this paper is to analyze how we can apply the last theorem to the Riccati equation and extend the convergence Theorem 2.3 to the inexact Kleinman–Newton method. This seems to be promising, especially for this application, since the resulting linear Newton equations are Lyapunov equations which are usually solved iteratively by Smith's method or versions of the ADI method.

3. Inexact Kleinman–Newton method.

Here we introduce for Riccati equations the inexact Kleinman–Newton method in the context presented in the previous chapter. Formally, the new iterate is determined by solving

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) + \mathcal{F}(X_k) = R_k \tag{3.1}$$

for X_{k+1}. This can be written more explicitly as a solution of X_{k+1}

$$X_{k+1}A_k + A_k^T X_{k+1} = -X_k BB^T X_k - C^T C + R_k. \tag{3.2}$$

Before we come to the convergence properties, we recall an existence and uniqueness theorem for Lyapunov equations, which need to be solved at each step of the algorithm.

THEOREM 3.1. *If $A \in \mathbb{R}^{n\times n}$ is stable, then for each $Z \in \mathbb{R}^{n\times n}$ the Lyapunov equation*

$$A^T Y + YA - Z = 0$$

is uniquely solvable and its solution is given by

$$Y = -\int_0^\infty e^{A^T t} Z e^{At} dt.$$

For a proof see [14, Theorem 8.5.1].

Before we state the convergence theorem, we summarize the algorithm proposed.

Inexact Kleinman–Newton algorithm.

```
Step 0:  Choose $X_0$ and set $k = 0$.
Step 1:  Determine a solution $X_{k+1}$,
         which solves the Lyapunov equation up to a residual $R_k$
         $X_{k+1}A_k + A_k^T X_{k+1} = -X_k BB^T X_k - C^T C + R_k$.
Step 2:  Set $k = k + 1$ and return to Step 1.
```

We can formulate the local convergence properties of this method by applying the standard theorem from the previous section.

THEOREM 3.2. *Let $X_\infty \in \mathbb{R}^{n\times n}$ be a symmetric solution of* (2.1) *such that $A - BB^T X_\infty$ is stable. Then there exist $\delta > 0$ and $\bar{\eta} > 0$ such that for all starting values $X_0 \in \mathbb{R}^{n\times n}$ with $\|X_0 - X_\infty\| \leq \delta$ the iterates X_k of the inexact Kleinman–Newton algorithm converge to X_∞, if the residuals R_k satisfy*

$$\|R_k\| \leq \eta_k \|\mathcal{F}(X_k)\| = \eta_k \left\|A^T X_k + X_k A - X_k BB^T X_k + C^T C\right\|. \tag{3.3}$$

The rate of convergence is linear if $\eta_k \in (0, \bar{\eta}]$, it is superlinear if $\eta_k \to 0$, and quadratic if $\eta_k \leq K_\eta \|\mathcal{F}(X_k)\|$ for some $K_\eta > 0$.

Proof. We apply Theorem 2.4 to the equation $\mathcal{F}(X) = A^T X + XA - XBB^T X + C^T C = 0$. This map is differentiable and has a Lipschitz continuous derivative. Since $A - BB^T X_\infty$ is assumed to be stable, $\mathcal{F}'(X_\infty)Y = 0$ implies $Y = 0$ by Theorem 3.1, and, hence, $\mathcal{F}'(X_\infty)$ is an invertible linear map. Since all assumptions in Theorem 2.4 hold, the conclusions can be applied and yield the statements in the theorem. □

4. Monotone convergence properties. An interesting fact about the Kleinman–Newton method is that the iterates exhibit monotonicity and a global convergence property, once the initial iterate X_0 is symmetric, positive semidefinite, and A_0 is stable. These properties are not common for Newton methods and depend on applications of the concavity and monotonicity results; see also Damm and Hinrichsen [7] or Ortega and Rheinboldt [21]. For the inexact version, these identities are perturbed and those results are much harder to obtain. In order to retain these properties, we have to impose certain conditions on the residuals.

Let us summarize at first a few monotonicity properties for the Lyapunov operators.

THEOREM 4.1. *The map $\mathcal{F}$ is concave in the following sense:*

$$\mathcal{F}'(X)(Y - X) \geq \mathcal{F}(Y) - \mathcal{F}(X) \quad \textit{for all symmetric} \quad X, Y \in \mathbb{R}^{n\times n}. \tag{4.1}$$

Proof. The proof follows easily from an identity due to the quadratic nature of the Riccati equation:

$$\mathcal{F}(Y) = \mathcal{F}(X) + \mathcal{F}'(X)(Y - X) + \frac{1}{2}\mathcal{F}''(X)(Y - X, Y - X), \tag{4.2}$$

where the quadratic term

$$\frac{1}{2}\mathcal{F}''(Z)(W, W) = -WBB^T W \tag{4.3}$$

is independent of Z. □

THEOREM 4.2. *Let $A - BB^T X$ be stable. Then*

$$Z = \mathcal{F}'(X)(Y) \Longleftrightarrow Y = -\int_0^\infty e^{\left(A-BB^TX\right)^T t} Z e^{\left(A-BB^TX\right)t} dt \tag{4.4}$$

and, hence, $\mathcal{F}'(X)(Y) \geq 0$ implies $Y \leq 0$.

Proof. We have

$$Z = \mathcal{F}'(X)(Y) = \left(A - BB^T X\right)^T Y + Y\left(A - BB^T X\right).$$

Since $(A - BB^T X)$ is stable, Theorem 3.1 yields the result. □

The next theorem shows that we can weaken the condition on the starting point and that the inexact Kleinman–Newton iteration is still well defined.

THEOREM 4.3. *Let X_k be symmetric and positive semidefinite such that $A - BB^T X_k$ is stable and*

$$R_k \le C^T C \tag{4.5}$$

holds. Then

(i) *the iterate X_{k+1} of the inexact Kleinman–Newton method is well defined, symmetric and positive semidefinite,*

(ii) *and the matrix $A - BB^T X_{k+1}$ is stable.*

Proof. The inexact Newton step (3.1) is given by the solution of a Lyapunov equation

$$X_{k+1} A_k + A_k^T X_{k+1} = -X_k BB^T X_k - C^T C + R_k.$$

Since A_k is stable, the unique solution X_{k+1} exists and is symmetric by Theorem 3.1. Furthermore, requirement (4.5) leads to

$$X_{k+1} A_k + A_k^T X_{k+1} \le 0$$

and Theorem 4.2 implies $X_{k+1} \ge 0$. Equation (3.2) is equivalent to

$$\begin{aligned} X_{k+1} A_{k+1} + A_{k+1}^T X_{k+1} = &- C^T C - X_{k+1} BB^T X_{k+1} \\ &- (X_{k+1} - X_k) BB^T (X_{k+1} - X_k) + R_k =: W. \end{aligned} \tag{4.6}$$

We define W as the right side of (4.6).

Let us assume $A_{k+1} x = \lambda x$ for λ with $Re(\lambda) \ge 0$ and $x \ne 0$. Then (4.6) implies

$$(\bar{\lambda} + \lambda) \bar{x}^T X_{k+1} x = \bar{x}^T A_{k+1}^T X_{k+1} x + \bar{x}^T X_{k+1} A_{k+1} x = \bar{x}^T W x.$$

On the one hand, the definition of W combined with requirement (4.5) leads to $W \le 0$. On the other hand, $X_{k+1} \ge 0$ implies $\bar{x}^T W x = 0$. Using the definition of W and a similar argument as before again, we obtain

$$\bar{x}^T (X_{k+1} - X_k) BB^T (X_{k+1} - X_k) x = 0. \tag{4.7}$$

Since $BB^T \ge 0$ we have $B^T (X_{k+1} - X_k) x = 0$, and, hence, $BB^T X_{k+1} x = BB^T X_k x$, so that by definition of A_k, A_{k+1}

$$A_{k+1} x = A_k x = \lambda x,$$

contradicting the stability of A_k. Hence, A_{k+1} is also stable. □

The requirements on the residuals can be weakened, e.g.,

$$R_k \le C^T C + X_j BB^T X_j \qquad j = k, k+1 \tag{4.8}$$

will also provide the previous proof.

In the following theorem we show under which requirements the monotonicity of the iterates X_k can be preserved also for the inexact Kleinman–Newton method.

THEOREM 4.4. *Let Assumption* 2.2 *be satisfied and let* X_0*, symmetric and positive semidefinite, be such that* A_0 *is stable. Assume that* (4.5) *and*

$$0 \leq R_k \leq (X_{k+1} - X_k)BB^T(X_{k+1} - X_k) \tag{4.9}$$

hold for all $k \in \mathbb{N}$*. Then the iterates* (3.2) *satisfy*

(i) $\lim_{k\to\infty} X_k = X_\infty$ *and* $0 \leq X_\infty \leq \cdots \leq X_{k+1} \leq X_k \leq \cdots \leq X_1$*,*

(ii) $(A - BB^T X_\infty)$ *is stable and* X_∞ *is the maximal solution of* $\mathcal{F}(X_\infty) = 0$*,*

(iii) $\|X_{k+1} - X_\infty\| \leq c\|X_k - X_\infty\|^2, k \in \mathbb{N}$.

Proof. i) Using the definition of an inexact Newton step and (4.2)

$$\begin{aligned} R_k &= \mathcal{F}'(X_k)(X_{k+1} - X_k) + \mathcal{F}(X_k) \\ &= \mathcal{F}(X_{k+1}) + (X_{k+1} - X_k)BB^T(X_{k+1} - X_k). \end{aligned}$$

This can be inserted into the next Newton step

$$\begin{aligned} \mathcal{F}'(X_{k+1})(X_{k+2} - X_{k+1}) &= -\mathcal{F}(X_{k+1}) + R_{k+1} \\ = R_{k+1} - R_k + (X_{k+1} - X_k)BB^T&(X_{k+1} - X_k) \geq R_{k+1} \geq 0 \end{aligned}$$

by assumption (4.9). Then from Theorem 4.2 we can infer

$$X_{k+2} - X_{k+1} \leq 0, \quad k = 0, 1, 2, \ldots$$

Therefore, $(X_k)_{k\in\mathbb{N}}$ is a monotone sequence of symmetric and positive semidefinite matrices and $X_k \geq 0$ due to Theorem 4.3. Hence, it is convergent to some symmetric and positive semidefinite limit matrix

$$\lim_{k\to\infty} X_k = X_\infty.$$

ii) Passing to the limit in (3.1) and (4.9) we deduce that X_∞ satisfies the Riccati equation, $X_\infty \leq X_k$ and $\mathcal{F}(X_\infty) = 0$.

We show that X_∞ is the maximal symmetric solution of the Riccati equation (2.1), which means $X_\infty \geq X$ for every symmetric solution X of (2.1). For this to hold we assume that X is a symmetric solution of the Riccati equation. Then Theorem 4.1 and (4.2) imply

$$\begin{aligned} \mathcal{F}'(X_k)(X - X_k) \geq -\,\mathcal{F}(X_k) &= -\mathcal{F}(X_{k-1}) - \mathcal{F}'(X_{k-1})(X_k - X_{k-1}) \\ &\quad - \frac{1}{2}\mathcal{F}''(X_{k-1})(X_k - X_{k-1}, X_k - X_{k-1}) \geq -R_{k-1}. \end{aligned}$$

Therefore, there exists $Q_k \geq 0$ with

$$\mathcal{F}'(X_k)(X - X_k) = Q_k - R_{k-1},$$

and since A_k is stable Theorem 3.1 implies

$$X - X_k = -\int_0^\infty e^{A_k^T t}(Q_k - R_{k-1})e^{A_k t}dt \leq \int_0^\infty e^{A_k^T t}R_{k-1}e^{A_k t}dt. \tag{4.10}$$

Passing to the limits leads to the desired result

$$X - X_\infty \leq 0$$

and X_∞ is the maximal solution. We can deduce from [14, Theorem 9.1.2] that the matrix $A - BB^T X_\infty$ is stable.

iii) To prove the quadratic rate of convergence we use the inexact Newton step

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) + \mathcal{F}(X_k) - R_k = 0$$

and rewrite it using (4.2)

$$\begin{aligned} \mathcal{F}'(X_\infty)(X_{k+1} - X_\infty) &= \mathcal{F}'(X_\infty)(X_{k+1} - X_\infty) - \mathcal{F}(X_{k+1}) + \mathcal{F}(X_\infty) \\ &\quad - (\ \mathcal{F}'(X_k)(X_{k+1} - X_k) - \mathcal{F}(X_{k+1}) + \mathcal{F}(X_k)) + R_k \\ &= (X_{k+1} - X_\infty)BB^T(X_{k+1} - X_\infty) \\ &\quad - (X_{k+1} - X_k)BB^T(X_{k+1} - X_k) + R_k. \end{aligned}$$

Since $A_\infty := (A - BB^T X_\infty)$ is stable, Theorem 4.2 shows

$$\begin{aligned} 0 \leq X_{k+1} - X_\infty &= \int_0^\infty e^{A_\infty^T t}\big\{ - (X_{k+1} - X_\infty)BB^T(X_{k+1} - X_\infty) \\ &\qquad + (X_{k+1} - X_k)BB^T(X_{k+1} - X_k) - R_k\big\} e^{A_\infty t}\, dt \\ &\leq \int_0^\infty e^{A_\infty^T t}((X_{k+1} - X_k)BB^T(X_{k+1} - X_k))e^{A_\infty t}\, dt. \end{aligned} \tag{4.11}$$

Note, that for all symmetric $A, B \in \mathbb{R}^{n\times n}$, $A \leq B$ implies $\|A\|_2 \leq \|B\|_2$, due to

$$\lambda_{\max}(A) = \max_{\|x\|_2=1} \frac{\bar{x}^T A x}{\bar{x}^T x} = \frac{\bar{x}_*^T A x_*}{\bar{x}_*^T x_*} \leq \frac{\bar{x}_*^T B x_*}{\bar{x}_*^T x_*} \leq \max_{\|x\|_2=1} \frac{\bar{x}^T B x}{\bar{x}^T x} = \lambda_{\max}(B).$$

Taking norms in (4.11) we obtain due to the stability of A_∞

$$\begin{aligned} 4\|X_{k+1} - X_\infty\|_2 &\leq \|X_{k+1} - X_k\|_2^2 \|BB^T\|_2 \int_0^\infty \left\|e^{A_\infty t}\right\|_2 \left\|e^{A_\infty^T t}\right\|_2\, dt \\ &\leq c\|X_{k+1} - X_k\|_2^2, \end{aligned} \tag{4.12}$$

and using the monotonicity of the iterates

$$0 \leq X_k - X_{k+1} \leq X_k - X_\infty \quad \Rightarrow \quad \|X_k - X_{k+1}\|_2 \leq \|X_k - X_\infty\|_2, \tag{4.13}$$

and, therefore,

$$\|X_{k+1} - X_\infty\|_2 \leq c\|X_k - X_\infty\|_2^2,$$

which implies quadratic convergence in any matrix norm. □

We impose several requirements on the residuals in Theorem 4.3 and Theorem 4.4. Some of them restrict the size of R_k in dependence on the step, see (4.9) and (4.5); others assume the positive definiteness. The first assumption on the size of the residuals depends on the quantity X_{k+1}, which has to be computed by the iterative procedure. However, the inequalities involved can be tested as the iteration for X_{k+1} progresses. The latter assumption is a condition, which the iterative Lyapunov solver has to satisfy.

5. Methods for solving the Lyapunov equation. There is a sizeable amount of literature on how to solve Lyapunov equations with direct solvers and iterative methods. In the inexact context we do not address direct Lyapunov solvers as presented in Laub [15], Roberts [24], or Grasedyck [8], but only iterative solvers, like Smith's [27], cyclic Smith(l) [23], or ADI methods [17]. In particular, we analyze these solvers with respect to the additonal requirements for maintaining the monotonicity as stated in the previous section.

Smith's and the ADI method are iterative solvers, which can be used to solve the Lyapunov equation at each Newton step. The inexact Newton method developed previously allows for early termination of these iterations, because the convergence criterion is not so stringent far away from the solution.

We review some basic properties of these methods.

Recall that at Newton iteration step k the following Lyapunov equation needs to be solved:

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) + \mathcal{F}(X_k) = 0,$$

or as in (3.2) we solve for $X = X_{k+1}$

$$XA_k + A_k^T X + S_k = 0 \tag{5.1}$$

with a stable matrix A_k

$$A_k = A - BB^T X_k \quad \text{and} \quad S_k = X_k BB^T X_k + C^T C.$$

This equation is equivalent to a Stein's equation.

LEMMA 5.1. *Given any $\mu \in \mathbb{R}^-$, then a solution X of the Lyapunov equation* (5.1) *is also a solution of Stein's equation and vice versa. Stein's equation is*

$$X = A_{k,\mu}^T X A_{k,\mu} + S_{k,\mu} \tag{5.2}$$

with

$$A_{k,\mu} = (A_k - \mu I)(A_k + \mu I)^{-1}, \quad S_{k,\mu} = -2\mu(A_k + \mu I)^{-T} S_k (A_k + \mu I)^{-1}.$$

Note that (5.1) is equivalent to

$$(A_k + \mu I)^T X (A_k + \mu I) - (A_k - \mu I)^T X (A_k - \mu I) = -2\mu S_k,$$

and from this (5.2) follows, since $A_k + \mu I$ is invertible for $\mu < 0$ due to the stability of A_k.

Smith's method—here we consider a simple version with one shift—is a fixed point iteration for (5.2) for given starting value $Z_k^{(0)}$

$$Z_k^{(l+1)} = A_{k,\mu}^T Z_k^{(l)} A_{k,\mu} + S_{k,\mu}, \quad l = 0, 1, \ldots \quad \text{and } \mu < 0 \text{ fixed.}$$

ADI method is a fixed point iteration for (5.2) for given starting value $Z_k^{(0)}$

$$Z_k^{(l+1)} = A_{k,\mu_l}^T Z_k^{(l)} A_{k,\mu_l} + S_{k,\mu_l}, \quad l = 0, 1, \ldots \quad \text{and } \mu_l < 0 \text{ varies.} \tag{5.3}$$

In practice, cyclic versions of both methods, where a given set of shift parameter $\mu_0, \ldots, \mu_s$ is used in a cyclic manner, became quite popular; see, e.g., [23] and [9].

Since Smith's method and the cyclic versions are special cases of the ADI method, we consider the ADI method in the following statements.

LEMMA 5.2. *Let Z_k be the solution of the Lyapunov equation* (5.1) *and let $Z_k^{(l)}$ be an iterate of the ADI method. Then*

$$Z_k^{(l+1)} - Z_k = A_{k,\mu_l}^T \dots A_{k,\mu_0}^T \left(Z_k^{(0)} - Z_k \right) A_{k,\mu_0} \dots A_{k,\mu_l}. \tag{5.4}$$

Proof. Recall that by Lemma 5.1 Z_k satisfies a Stein's equation for any $\mu \in \mathbb{R}^-$; hence, for all μ_l in the ADI method

$$Z_k = A_{k,\mu_l}^T Z_k A_{k,\mu_l} + S_{k,\mu_l} \quad l = 0, 1, \dots$$

Therefore, we have for any l

$$Z_k^{(l+1)} - Z_k = A_{k,\mu_l}^T Z_k^{(l)} A_{k,\mu_l} + S_{k,\mu_l} - (A_{k,\mu_l}^T Z_k A_{k,\mu_l} + S_{k,\mu_l}) = A_{k,\mu_l}^T \left(Z_k^{(l)} - Z_k \right) A_{k,\mu_l}.$$

If we apply this identity to $Z_k^{(l)} - Z_k$ consecutively, then we obtain the statement of the lemma. □

To estimate the residual of the Lyapunov equation using some iterate from the ADI method, we prove the following lemma.

LEMMA 5.3. *Let $Z_k^{(l)}$ be an iterate of the ADI method, and then for the residuals of the Lyapunov equation we obtain*

$$\begin{aligned} R_k^{(l)} &:= Z_k^{(l)} A_k + A_k^T Z_k^{(l)} + S_k \\ &= A_{k,\mu_{l-1}}^T \dots A_{k,\mu_0}^T \left(Z_k^{(0)} A_k + A_k^T Z_k^{(0)} + S_k \right) A_{k,\mu_0} \dots A_{k,\mu_{l-1}}. \end{aligned} \tag{5.5}$$

If, in particular, the initial residual $R_k^{(0)}$ is positive semidefinite, then all residuals $R_k^{(l)}$ are also positive semidefinite.

Proof. Note that

$$\begin{aligned} Z_k^{(l)} A_k + A_k^T Z_k^{(l)} + S_k &= Z_k^{(l)} A_k + A_k^T Z_k^{(l)} - Z_k A_k - A_k^T Z_k \\ &= \left(Z_k^{(l)} - Z_k \right) A_k + A_k^T \left(Z_k^{(l)} - Z_k \right). \end{aligned}$$

Next we insert (5.4) to obtain

$$\begin{aligned} Z_k^{(l)} A_k + A_k^T Z_k^{(l)} + S_k = {} & A_{k,\mu_{l-1}}^T \dots A_{k,\mu_0}^T \left(Z_k^{(0)} - Z_k \right) A_{k,\mu_0} \dots A_{k,\mu_{l-1}} A_k \\ & + A_k^T A_{k,\mu_{l-1}}^T \dots A_{k,\mu_0}^T \left(Z_k^{(0)} - Z_k \right) A_{k,\mu_0} \dots A_{k,\mu_{l-1}}. \end{aligned} \tag{5.6}$$

Since A_k and $A_{k,\mu}$ commute for any μ, we have

$$\begin{aligned} & Z_k^{(l)} A_k + A_k^T Z_k^{(l)} + S_k \\ & \quad = A_{k,\mu_{l-1}}^T \dots A_{k,\mu_0}^T \left(\left(Z_k^{(0)} - Z_k \right) A_k + A_k^T \left(Z_k^{(0)} - Z_k \right) \right) A_{k,\mu_0} \dots A_{k,\mu_{l-1}} \end{aligned}$$

from which (5.5) follows. From this equation we obtain the result that if the initial residual is positive semidefinite, then this also holds for all residuals in the Lyapunov equation using any ADI iterate. □

In particular, with the zero starting matrix we get the following.

LEMMA 5.4. *Let $Z_k^{(0)} = 0$. Then the residuals of* (5.1) *for the ADI iterates satisfy*

$$R_k^{(l)} \geq 0.$$

Proof. The residuals of (5.1) for the iterates $Z_k^{(l)}$ of the ADI method are given by Lemma 5.3:

$$R_k^{(l)} = Z_k^{(l)} A_k + A_k^T Z_k^{(l)} + S_k = A_{k,\mu_{l-1}}^T \dots A_{k,\mu_0}^T S_k A_{k,\mu_0} \dots A_{k,\mu_{l-1}} \geq 0$$

since $S_k = X_k B B^T X_k + C^T C \geq 0$. □

LEMMA 5.5. *Let us consider a cyclic ADI method with a finite set of shift parameter $\mu_0, \dots, \mu_s \in \mathbb{R}^-$. If $C^T C$ is positive definite and $Z_k^{(0)} = 0$, there is l_k such that for all $l \geq l_k$*

$$0 \leq R_k^{(l)} \leq C^T C$$

holds.

Proof. $R_k^{(l)} \geq 0$ is proved in the previous Lemma. Furthermore, if $C^T C > 0$, there exists $\zeta > 0$ such that for all $x \in \mathbb{C}^n$

$$\bar{x}^T C^T C x \geq \zeta \|x\|_2^2.$$

We have $\rho(A_{k,\mu}) = \max_{\lambda \in \sigma(A_k)} |\frac{\lambda - \mu}{\lambda + \mu}| < 1$ for every $\mu \in \mathbb{R}^-$. Due to the special structure of the matrices $A_{k,\mu}, \mu \in \mathbb{R}^-$, it follows that

$$\rho(A_{k,\mu_0} \dots A_{k,\mu_s}) = \max_{\lambda \in \sigma(A_k)} \left| \prod_{i=1}^{s} \frac{\lambda - \mu_i}{\lambda + \mu_i} \right| \leq \prod_{i=1}^{s} \max_{\lambda \in \sigma(A_k)} \left| \frac{\lambda - \mu_i}{\lambda + \mu_i} \right| < 1.$$

Therefore, a consistent matrix norm $\| \cdot \|_*$ exists with $\|A_{k,\mu_0} \dots A_{k,\mu_s}\|_* < 1$.

For l large enough (depending on k) we obtain with $m := l \mod (s+1)$

$$\begin{aligned} \left\| R_k^{(l)} \right\|_2 = \Bigg\| & A_{k,\mu_m}^T \dots A_{k,\mu_0}^T \underbrace{A_{k,\mu_s}^T \dots A_{k,\mu_0}^T \dots A_{k,\mu_s}^T \dots A_{k,\mu_0}^T}_{\lfloor \frac{l}{s+1} \rfloor \text{ times}} S_k \\ & \underbrace{A_{k,\mu_0} \dots A_{k,\mu_s} \dots A_{k,\mu_0} \dots A_{k,\mu_s}}_{\lfloor \frac{l}{s+1} \rfloor \text{ times}} A_{k,\mu_0} \dots A_{k,\mu_m} \Bigg\|_2 \\ & \leq c \|A_{k,\mu_0} \dots A_{k,\mu_s}\|_2^{2\lfloor \frac{l}{s+1} \rfloor} \|S_k\|_2 \\ & \leq c_* \|A_{k,\mu_0} \dots A_{k,\mu_s}\|_*^{2\lfloor \frac{l}{s+1} \rfloor} \leq \zeta. \end{aligned}$$

Hence, for all $x \in \mathbb{C}^n$

$$\bar{x}^T R_k^{(l)} x \leq \|x\|_2^2 \left\| R_k^{(l)} \right\|_2 \leq \zeta \|x\|_2^2 \leq \bar{x}^T C^T C x,$$

which is to be shown. □

According to (4.8) it might be possible to introduce a weaker requirement compared to the positive definiteness of the matrix $C^T C$ to achieve the same results.

6. Numerical results. In this section we analyze the efficiency of the inexact Kleinman–Newton versions, developed in Theorem 3.2, compared to the standard Kleinman–Newton method. Note that we concentrate on the local convergence properties and not on the monotonicity of the iterates. The efficiency of the inexact versions cannot be tested without special consideration of the applied Lyapunov solver.

Many iterative solvers for Lyapunov equations are presented in the literature, e.g., Smith's method [27], ADI method [17], and low-rank ADI methods [22], [16]. Other iterative methods can be found in [9], [26], or [23].

In order to indicate the benefits of the inexact Kleinman–Newton method, we implement Smith's method, the ADI method, and a low-rank ADI version to solve the Newton steps.

We consider an example arising from optimal control problems. The example has been considered by Morris and Navasca [19] and is described as a two-dimensional optimal control problem with parabolic partial differential equations including convection:

$$\min_u J(u) = \frac{1}{2}\int_0^\infty \left(\| \; Cz(t) \; \|_2^2 + \| \; u(t) \; \|_2^2\right) dt$$

s. t.

$$\begin{aligned} \frac{\partial z}{\partial t} = \frac{\partial z}{\partial x^2} + \frac{\partial z}{\partial y^2} + 20\frac{\partial z}{\partial y} + 100z &= f(x,y)u(t) \qquad && (x,y) \in \Omega \\ z(x,y,t) &= 0 && (x,y) \in \partial\Omega \qquad \forall t \end{aligned}$$

with $\Omega = (0,1) \times (0,1)$ and

$$f(x,y) := \begin{cases} 100 & 0.1 < x < 0.3 \quad \& \quad 0.4 < y < 0.6, \\ 0 & \text{else.} \end{cases}$$

The discretization is carried out on a 23 × 23 grid and central differences are used for the approximation. We choose $C = (0.1, \ldots, 0.1)$, $X_0 = 0$. The optimal matrix X_∞ has been computed beforehand with a higher accuracy. We compare the number of the Newton steps (outer), the number of inner iterations (inner) which are needed to solve each Newton step, and the cumulative number of inner iterations (cumul). According to Theorem 3.2 we test an inexact Kleinman–Newton version with a superlinear rate of convergence. In Tables 6.1 and 6.2 we use Smith's method

TABLE 6.1
Smith's method: Exact Kleinman–Newton method.

outer	inner	cumul	$\| \mathcal{F}(X_k) \|$	$\| X_k - X_\infty \|$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|}$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|^2}$
1	97	97	$7.639\text{e}+005$	$3.495\text{e}+001$	$1.056\text{e}+003$	$3.484\text{e}+004$
2	211	308	$1.911\text{e}+005$	$2.333\text{e}+001$	$6.677\text{e}-001$	$1.911\text{e}-002$
3	144	452	$4.794\text{e}+004$	$1.755\text{e}+001$	$7.521\text{e}-001$	$3.224\text{e}-002$
4	83	535	$1.213\text{e}+004$	$1.453\text{e}+001$	$8.279\text{e}-001$	$4.719\text{e}-002$
5	66	601	$3.172\text{e}+003$	$1.245\text{e}+001$	$8.571\text{e}-001$	$5.901\text{e}-002$
6	49	650	$8.973\text{e}+002$	$9.594\text{e}+000$	$7.704\text{e}-001$	$6.187\text{e}-002$
7	66	716	$2.357\text{e}+002$	$4.481\text{e}+000$	$4.671\text{e}-001$	$4.869\text{e}-002$
8	58	774	$1.801\text{e}+001$	$5.031\text{e}-001$	$1.123\text{e}-001$	$2.505\text{e}-002$
9	43	817	$8.544\text{e}-002$	$4.162\text{e}-003$	$8.272\text{e}-003$	$1.645\text{e}-002$
10	33	850	$8.230\text{e}-004$	$1.263\text{e}-005$	$3.036\text{e}-003$	$7.457\text{e}-001$
11	21	871	$3.281\text{e}-008$	$6.488\text{e}-010$	$5.135\text{e}-005$	$4.149\text{e}+000$

Table 6.2
Smith's method: Inexact K-N method with superlinear convergence $\eta_k = 1/(k^3+1)$.

outer	inner	cumul	$\| \mathcal{F}(X_k) \|$	$\| X_k - X_\infty \|$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|}$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|^2}$
1	32	32	7.628e + 005	3.490e + 001	1.054e + 003	3.479e + 004
2	15	47	2.294e + 005	2.442e + 001	6.997e − 001	2.006e − 002
3	15	62	6.088e + 004	1.827e + 001	7.481e − 001	3.065e − 002
4	14	76	1.581e + 004	1.496e + 001	8.191e − 001	4.486e − 002
5	13	89	4.141e + 003	1.284e + 001	8.582e − 001	5.738e − 002
6	7	96	1.162e + 003	1.053e + 001	8.203e − 001	6.389e − 002
7	12	108	3.171e + 002	5.560e + 000	5.279e − 001	5.012e − 002
8	19	127	3.797e + 001	9.554e − 001	1.718e − 001	3.091e − 002
9	17	144	2.741e − 001	1.483e − 002	1.552e − 002	1.625e − 002
10	18	162	3.396e − 003	5.113e − 005	3.448e − 003	2.343e − 001
11	15	177	1.400e − 006	2.267e − 008	4.433e − 004	8.886e + 000
12	14	191	3.477e − 010	3.333e − 012	1.471e − 004	6.539e + 003

Table 6.3
ADI method: Exact Kleinman–Newton method.

outer	inner	cumul	$\| \mathcal{F}(X_k) \|$	$\| X_k - X_\infty \|$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|}$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|^2}$
1	24	24	7.639e + 005	3.495e + 001	1.056e + 003	3.484e + 004
2	20	44	1.911e + 005	2.333e + 001	6.677e − 001	1.911e − 002
3	22	66	4.794e + 004	1.755e + 001	7.521e − 001	3.224e − 002
4	22	88	1.213e + 004	1.453e + 001	8.279e − 001	4.719e − 002
5	21	109	3.172e + 003	1.245e + 001	8.571e − 001	5.901e − 002
6	20	129	8.973e + 002	9.594e + 000	7.704e − 001	6.187e − 002
7	22	151	2.357e + 002	4.481e + 000	4.671e − 001	4.869e − 002
8	27	178	1.801e + 001	5.031e − 001	1.123e − 001	2.505e − 002
9	34	212	8.544e − 002	4.162e − 003	8.272e − 003	1.645e − 002
10	32	244	8.230e − 004	1.263e − 005	3.036e − 003	7.457e − 001
11	24	268	3.273e − 008	6.273e − 010	4.965e − 005	4.012e + 000

Table 6.4
ADI method: Inexact K-N method with superlinear convergence $\eta_k = 1/(k^3+1)$.

outer	inner	cumul	$\| \mathcal{F}(X_k) \|$	$\| X_k - X_\infty \|$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|}$	$\frac{\|X_k - X_\infty\|}{\|X_{k-1} - X_\infty\|^2}$
1	13	13	7.639e + 005	3.495e + 001	1.056e + 003	3.484e + 004
2	4	17	1.911e + 005	2.334e + 001	6.677e − 001	1.911e − 002
3	5	22	4.794e + 004	1.755e + 001	7.520e − 001	3.224e − 002
4	5	27	1.213e + 004	1.453e + 001	8.279e − 001	4.719e − 002
5	6	33	3.175e + 003	1.249e + 001	8.597e − 001	5.918e − 002
6	3	36	9.542e + 002	1.109e + 001	8.882e − 001	7.112e − 002
7	8	44	2.073e + 002	3.978e + 000	3.586e − 001	3.232e − 002
8	5	49	1.829e + 001	5.119e − 001	1.287e − 001	3.235e − 002
9	11	60	1.037e − 001	5.069e − 003	9.903e − 003	1.935e − 002
10	16	76	9.042e − 004	1.502e − 005	2.962e − 003	5.966e − 001
11	16	92	4.119e − 007	3.114e − 009	2.073e − 004	1.416e + 001
12	15	107	1.905e − 010	1.087e − 012	3.491e − 004	1.165e + 005

for the Lyapunov solver, whereas in Tables 6.3 and 6.4 we apply the ADI method and in Tables 6.5 and 6.6 its low-rank version. The shift parameters are determined with a heuristic introduced by Penzl [23]. All computations were done within MATLAB.

In Table 6.7 we present a comparison of CPU times in order to test the efficiency of the inexact versions. We include alternative stopping criteria for the inner iteration which result according to Theorem 3.2 in a linear and a superlinear rate of convergence. We observed for our examples a rather small linear convergence rate factor and a

Table 6.5
Low-rank ADI method: Exact Kleinman–Newton method.

outer	inner	cumul	$\Vert \mathcal{F}(X_k) \Vert$	$\Vert X_k - X_\infty \Vert$	$\frac{\Vert X_k - X_\infty \Vert}{\Vert X_{k-1} - X_\infty \Vert}$	$\frac{\Vert X_k - X_\infty \Vert}{\Vert X_{k-1} - X_\infty \Vert^2}$
1	24	24	7.639e + 005	3.495e + 001	1.056e + 003	3.484e + 004
2	41	65	1.911e + 005	2.333e + 001	6.677e − 001	1.911e − 002
3	25	90	4.794e + 004	1.755e + 001	7.521e − 001	3.224e − 002
4	23	113	1.213e + 004	1.453e + 001	8.279e − 001	4.719e − 002
5	24	137	3.172e + 003	1.245e + 001	8.571e − 001	5.901e − 002
6	24	161	8.973e + 002	9.594e + 000	7.704e − 001	6.187e − 002
7	25	186	2.357e + 002	4.481e + 000	4.671e − 001	4.869e − 002
8	26	212	1.801e + 001	5.031e − 001	1.123e − 001	2.505e − 002
9	33	245	8.544e − 002	4.162e − 003	8.272e − 003	1.645e − 002
10	41	286	8.230e − 004	1.263e − 005	3.036e − 003	7.457e − 001
11	42	328	3.209e − 008	5.735e − 010	4.539e − 005	3.667e + 000

Table 6.6
Low-rank ADI method: Inexact K-N method with superlinear convergence $\eta_k = 1/(k^3 + 1)$.

outer	inner	cumul	$\Vert \mathcal{F}(X_k) \Vert$	$\Vert X_k - X_\infty \Vert$	$\frac{\Vert X_k - X_\infty \Vert}{\Vert X_{k-1} - X_\infty \Vert}$	$\frac{\Vert X_k - X_\infty \Vert}{\Vert X_{k-1} - X_\infty \Vert^2}$
1	13	13	7.639e + 005	3.495e + 001	1.056e + 003	3.484e + 004
2	4	17	1.908e + 005	1.853e + 001	5.302e − 001	1.518e − 002
3	3	20	4.760e + 004	6.054e + 000	3.267e − 001	1.763e − 002
4	6	26	1.208e + 004	1.374e + 001	2.270e + 000	3.750e − 001
5	8	34	3.185e + 003	1.254e + 001	9.122e − 001	6.638e − 002
6	7	41	8.998e + 002	9.653e + 000	7.700e − 001	6.143e − 002
7	7	48	2.363e + 002	4.494e + 000	4.655e − 001	4.823e − 002
8	6	54	1.788e + 001	4.977e − 001	1.107e − 001	2.464e − 002
9	9	63	8.554e − 002	3.983e − 003	8.002e − 003	1.608e − 002
10	21	84	8.871e − 004	1.245e − 005	3.126e − 003	8.029e − 001
11	31	115	5.657e − 007	2.088e − 009	1.677e − 004	1.368e + 001
12	47	162	1.449e − 010	9.694e − 013	4.643e − 004	2.740e + 005

Table 6.7
Comparison of computing time.

Lyapunov solver	Convergence rate					
	Exact K-N		Linear		Superlinear	
	$\Vert \mathcal{F}(X_\infty) \Vert$	Time	$\Vert \mathcal{F}(X_\infty) \Vert$	Time	$\Vert \mathcal{F}(X_\infty) \Vert$	Time
Smith	3.281e − 008	289	6.115e − 008	68	3.477e − 010	76
ADI	3.273e − 008	204	1.588e − 008	65	1.905e − 010	81
Low-rank ADI	3.209e − 008	145	1.447e − 008	81	1.449e − 010	68

rather late onset of the superlinear convergence behavior, which results in a rather good performance of the linearly convergent version compared to the superlinearly convergent version. The time needed to compute the shift parameter is not included in Table 6.7. This is an additional advantage of the inexact versions because they need fewer inner iterations and, therefore, a smaller number of shift parameters.

7. Robustness. Let us note that there is another implementation of Newton's method for the Riccati equation presented in the literature, e.g., [1], [19]. Here the Newton step is computed by a Lyapunov equation for the increment $X_{k+1} - X_k$ in the following way:

$$\begin{aligned}(X_{k+1} - X_k)\left(A - BB^T X_k\right) + \left(A - BB^T X_k\right)^T (X_{k+1} - X_k) \\ = (X_k - X_{k-1}) BB^T (X_k - X_{k-1}),\end{aligned} \tag{7.1}$$

in contrast to (2.4)

$$X_{k+1}\left(A - BB^TX_k\right) + \left(A - BB^TX_k\right)^T X_{k+1} = -X_kBB^TX_k - C^TC. \tag{7.2}$$

Note that the inhomogeneous terms in the Lyapunov equations for both variants of Newton's method differ quite substantially. The authors of [1] pointed out that (7.1) exhibits some advantages compared to the standard implementation (2.4), e.g., if BB^T has low rank.

Equation (7.1) is the matrix notation of

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) = -\mathcal{F}(X_k) \tag{7.3}$$

with the following modification due to (4.2):

$$\begin{aligned} -\mathcal{F}(X_k) &= -\mathcal{F}(X_{k-1}) - \mathcal{F}'(X_{k-1})(X_k - X_{k-1}) \\ &\quad -\frac{1}{2}\mathcal{F}''(X_{k-1})(X_k - X_{k-1}, X_k - X_{k-1}) \\ &= -\frac{1}{2}\mathcal{F}''(X_{k-1})(X_k - X_{k-1}, X_k - X_{k-1}). \end{aligned} \tag{7.4}$$

Both methods are identical for the exact Newton step:

LEMMA 7.1. *If a sequence X_k satisfies* (7.2), *then it also fulfills* (7.1). *If, conversely, a sequence X_k satisfies* (7.1), *then it also fulfills* (7.2), *provided the starting points X_0, X_1 satisfy* (7.2) *for $k = 0$.*

Proof. The first conclusion was shown above. For the reverse to hold, we use (7.4) and obtain

$$\begin{aligned} \mathcal{F}'(X_k)(X_{k+1} - X_k) &= -\tfrac{1}{2}\mathcal{F}''(X_{k-1})(X_k - X_{k-1}, X_k - X_{k-1}) \\ &= -\mathcal{F}(X_k) + \mathcal{F}(X_{k-1}) + \mathcal{F}'(X_{k-1})(X_k - X_{k-1}) \end{aligned}$$

and, hence,

$$\mathcal{F}(X_k) + \mathcal{F}'(X_k)(X_{k+1} - X_k) = \mathcal{F}(X_{k-1}) + \mathcal{F}'(X_{k-1})(X_k - X_{k-1})$$

for all $k \geq 0$. Since it is assumed that for the starting iterates

$$\mathcal{F}(X_0) + \mathcal{F}'(X_0)(X_1 - X_0) = 0, \tag{7.5}$$

the X_k also satisfy (2.4). □

Although both methods are identical in the exact case, an inexact version of the Kleinman–Newton method based on implementation (7.1) is unstable. A reformulation of an inexact Kleinman–Newton method using (7.1) leads to

$$\begin{aligned} \mathcal{F}'(X_k)(X_{k+1} - X_k) &= -\frac{1}{2}\mathcal{F}''(X_{k-1})(X_k - X_{k-1}, X_k - X_{k-1}) + \tilde{R}_k \\ &= \mathcal{F}(X_{k-1}) + \mathcal{F}'(X_{k-1})(X_k - X_{k-1}) - \mathcal{F}(X_k) + \tilde{R}_k \end{aligned}$$

or, equivalently,

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) + \mathcal{F}(X_k) = \mathcal{F}'(X_{k-1})(X_k - X_{k-1}) + \mathcal{F}(X_{k-1}) + \tilde{R}_k.$$

Using this recursively shows that the residuals accumulate during the course of the iteration

$$\mathcal{F}'(X_k)(X_{k+1} - X_k) + \mathcal{F}(X_k) = \sum_{i=1}^{k} \tilde{R}_i.$$

This means that one has to limit $R_k = \sum_{i=1}^{k} \tilde{R}_i$ (if X_1 is computed by an exact Newton step) according to the convergence Theorem 2.4 which seems to be a rather strong assumption because the residuals are cumulative.

8. Conclusion. In this paper we propose a modification of the classical Kleinman–Newton method for the numerical solution of Riccati equations. The iterative Lyapunov equation solvers for the Newton steps are terminated early to save computing time. Based on the theory of inexact Newton methods, we give termination criteria which warrant the fast local rates. In addition, we derive conditions, which guarantee the more global convergence statement for the Kleinman–Newton method. We show how these requirements can be addressed, for example, for Smith's method or the ADI method. The numerical example for a parabolic control problem illustrates the potential for substantial savings in the number of iterations and computing time.

Acknowledgments. The authors thank John Burns, Virginia Tech, for several discussions about the topics of this paper and two anonymous referees for their valuable comments. Furthermore, financial support by the Forschungsfond 2007 Universität Trier is acknowledged.

REFERENCES

[1] H. Banks and K. Ito, *A numerical algorithm for optimal feedback gains in high dimensional linear quadratic regulator problems*, SIAM J. Control Optim., 29 (1991), pp. 499–515.

[2] P. Benner, *Solving large-scale control problems*, IEEE Control Systems Magazine, 24 (2004), pp. 44–59.

[3] P. Benner and R. Byers, *An exact line search method for solving generalized continuous-time algebraic Riccati equations*, IEEE Trans. Automatic Control, 43 (1998), pp. 101–107.

[4] P. Benner, H. Mena, and J. Saak, *On the parameter selection problem in the Newton-ADI iteration for large-scale Riccati equations*, Electronic Trans. Numer. Anal., 29 (2008), pp. 136–149.

[5] P. Benner, E. Quintana-Orti, and G. Quintana-Orti, *Solving linear-quadratic optimal control problems on parallel computers*, Optimization Methods and Software, 23 (2008), pp. 879–909.

[6] J. Burns, E. Sachs, and L. Zietsman, *Mesh independence of Kleinman-Newton iterations for Riccati equations in Hilbert spaces*, SIAM J. Control Optim., 47 (2008), pp. 2663–2692.

[7] T. Damm and D. Hinrichsen, *Newton's method for a rational matrix equation occuring in stochastic control*, Linear Algebra Appl., 332–334 (2001), pp. 81–109.

[8] L. Grasedyck, W. Hackbusch, and B. N. Khoromskij, *Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices*, Computing, 70 (2003), pp. 121–165.

[9] S. Gugercin, D. C. Sorensen, and A. C. Antoulas, *A modified low-rank Smith method for large-scale Lyapunov equations*, Numerical Algorithms, 32 (2003), pp. 27–55.

[10] C.-H. Guo and P. Lancaster, *Analysis and modification of Newton's method for algebraic Riccati equations*, Mathematics of Computation, 67 (1998), pp. 1089–1105.

[11] C.-H. Guo and A. Laub, *On a Newton-like method for solving algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 694–698.

[12] C. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[13] D. Kleinman, *On an iterative technique for Riccati equation computations*, IEEE Trans. Automatic Control, 13 (1968), pp. 114–115.

[14] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.

[15] A. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automatic Control, 24 (1979), pp. 913–921.

[16] J.-R. Li and J. White, *Low rank solution of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 260–280.

[17] A. Lu and E. Wachspress, *Solution of Lyapunov equations by alternating direction implicit iteration*, Comp. Math. Appl., 21 (1991), pp. 43–58.

[18] V. L. Mehrmann, *The Autonomous Linear Quadratic Control Problem*, Springer, Berlin-Heidelberg, 1991.

[19] K. Morris and C. Navasca, *Solution of algebraic Riccati equations arising in control of partial differential equations*, in Control and Boundary Analysis, J. P. Zolesio and J. Cagnol, eds., Lecture Notes in Pure Appl. Math. 240, CRC Press, 2005, pp. 257–280.

[20] K. Morris and C. Navasca, *Iterative solution of algebraic Riccati equations for damped systems*, 2006 45th IEEE Conference on Decision and Control (2006), pp. 2436–2440.

[21] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[22] T. Penzl, *Numerische Lösung grosser Lyapunov-Gleichungen*, Ph.D. Thesis, Fakultät für Mathematik, TU Chemnitz, Berlin, 1998.

[23] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (1999), pp. 1401–1418.

[24] J. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Int. J. Control, 32 (1980), pp. 677–687.

[25] I. Rosen and C. Wang, *A multilevel technique for the approximate solution of operator Lyapunov and algebraic Riccati equations*, SIAM J. Numer. Anal., 32 (1995), pp. 514–541.

[26] V. Simoncini, *A new iterative method for solving large-scale Lyapunov matrix equations*, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288.

[27] R. Smith, *Matrix equation $XA + BX = C$*, SIAM J. Appl. Math., 16 (1968), pp. 198–201.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 289–315

HADAMARD FUNCTIONS OF INVERSE M-MATRICES[*]

CLAUDE DELLACHERIE[†], SERVET MARTINEZ[‡], AND JAIME SAN MARTIN[‡]

Abstract. We prove that the class of generalized ultrametric matrices (GUM) is the largest class of bipotential matrices stable under Hadamard increasing functions. We also show that any power $\alpha \geq 1$, in the sense of Hadamard functions, of an inverse M-matrix is also inverse M-matrix. This was conjectured for $\alpha = 2$ by Neumann in [*Linear Algebra Appl.*, 285 (1998), pp. 277–290], and solved for integer $\alpha \geq 1$ by Chen in [*Linear Algebra Appl.*, 381 (2004), pp. 53–60]. We study the class of filtered matrices, which include naturally the GUM matrices, and present some sufficient conditions for a filtered matrix to be a bipotential.

Key words. M-matrices, Hadamard functions, ultrametric matrices, potential matrices

AMS subject classifications. 15A48, 15A51, 60J45

DOI. 10.1137/060651082

1. Introduction and basic notations.

In this article we study stability properties under Hadamard functions for the class of inverse M-matrices and the class of filtered matrices, which includes GUM (generalized ultrametric matrices).

A nonnegative matrix U is said to be a *potential* if it is nonsingular and its inverse satisfies

$$\forall i \neq j,\ U^{-1}_{ij} \leq 0, \qquad \forall i,\ U^{-1}_{ii} > 0,$$

$$\forall i,\ \sum_j U^{-1}_{ij} \geq 0,$$

that is, if U^{-1} is an M-matrix which is row diagonally dominant. We denote this class of matrices by $\mathcal{P}$. In addition we say that U is a *bipotential* if U^{-1} is also column diagonally dominant. This class of matrices is denoted by $bi\mathcal{P}$. We note that $\mathcal{P}, bi\mathcal{P}$ are contained in $\mathcal{M}^{-1}$, the class of inverses of M-matrices.

The class of potential matrices play an important role in probability theory. They represent the potential (from which we have taken the name) of a transient continuous time Markov chain $(X_t)_{t\geq 0}$, with generator $-U^{-1}$. That is,

$$U_{ij} = \int_0^\infty (e^{-U^{-1}t})_{ij}\, dt = \int_0^\infty \mathbb{P}_i\{X_t = j\} dt$$

is the mean expected time expended at site j when the chain starts at site i. Clearly U is a bipotential if both U and U' are potentials.

To get a discrete time interpretation take $K_0 = \max_i\{U^{-1}_{ii}\}$. For any $K \geq K_0$ the matrix $P_K = \mathbb{I} - \frac{1}{K}U^{-1}$ is nonnegative, substochastic, and verifies

$$U^{-1} = k(\mathbb{I} - P_K).$$

[*] Received by the editors January 30, 2006; accepted for publication (in revised form) by R. Nabben November 27, 2008; published electronically March 18, 2009. This research was supported by Nucleus Millennium Information and Randomness P04-069-F.

http://www.siam.org/journals/simax/31-2/65108.html

[†] Laboratoire Raphaël Salem, UMR 6085, Université de Rouen, Site Colbert, 76821 Mont Saint Aignan Cedex, France (Claude.Dellacherie@univ-rouen.fr).

[‡] Center for Mathematical Modelling and Department of Mathematical Engineering (CMM-DIM), Universidad de Chile, UMI-CNRS 2807, Casilla 170-3 Correo 3 Santiago, Chile (smartine@dim.uchile.cl, jsanmart@dim.uchile.cl).

If we can take $K = 1$, then $U^{-1} = \mathbb{I} - P$ (with $P = P_1$) and U is the mean expected number of visits of a Markov chain $(Y_n)_{n\in\mathbb{N}}$ whose transition probability is given by P. Indeed,

$$U_{ij} = \sum_{n\geq 0} P_{ij}^n = \sum_{n\geq 0} \mathbb{P}_i\{Y_n = j\}.$$

We notice that if U is a potential, then for all i, j we have $U_{ii} \geq U_{ji}$. The probabilistic proof of this fact is based on the so-called strong Markov property which allows us to conclude

$$U_{ji} = f_{ji} U_{ii},$$

where $f_{ji} \leq 1$ is the probability that the Markov process (X_t), starting from j, ever reaches the state i. If U is a bipotential, then $U_{ii} \geq \max\{U_{ij}, U_{ji}\}$.

For any nonnegative matrix U we define the quantity

$$\tau(U) = \inf\{t \geq 0 :\ \mathbb{I} + tU \notin bi\mathcal{P}\},$$

which is invariant under permutations; that is, $\tau(U) = \tau(\Pi U \Pi')$. We point out that if U is a positive matrix, then $\tau(U) > 0$. We shall study some properties of this function τ. In particular we are interested on matrices for which $\tau(U) = \infty$, generalizing the class $bi\mathcal{P}$ as the next result shows.

Proposition 1.1. *Assume that U is a nonnegative matrix, which is nonsingular and $\tau(U) = \infty$; then $U \in bi\mathcal{P}$.*

Proof. It is direct from the observation that

$$t(\mathbb{I} + tU)^{-1} \underset{t\to\infty}{\rightarrow} U^{-1}. \qquad \square$$

Remark 1.1. Later on, we shall prove that the converse is also true: if U is in class $bi\mathcal{P}$, then $\tau(U) = \infty$.

The following notion will play an important role in this article.

Definition 1.2. *Given a matrix B, a vector μ is said to be a right equilibrium potential if*

$$B\mu = \mathbf{1},$$

where $\mathbf{1}$ is the constant vector of ones. Similarly it is defined the notion of a left equilibrium potential, which is the right equilibrium potential for B'. When B is nonsingular the unique right and left equilibrium potentials are, respectively, denoted by μ_B and ν_B.

We denote by $\bar{\mu} = \mathbf{1}'\mu$ the total mass of μ. In the nonsingular case, it is not difficult to see that $\bar{\nu} = \bar{\mu}$.

Notice that for a matrix $U \in bi\mathcal{P}$ the right and left equilibrium potentials are nonnegative. This is exactly the fact that the inverse of a bipotential matrix is row and column diagonally dominant.

Definition 1.3. *The constant block form (CBF) matrices are defined recursively in the following way: given two CBF matrices A, B of sizes p and $n - p$, respectively, and numbers α, β, we produce the new CBF matrix by*

$$U = \begin{pmatrix} A & \alpha \mathbf{1}_p \mathbf{1}'_{n-p} \\ \beta \mathbf{1}_{n-p} \mathbf{1}'_p & B \end{pmatrix}, \tag{1.1}$$

where the vector $\mathbf{1}_p$ *is the vector of ones, of size* p. *We also say that* U *is in* increasing *CBF if* $\min\{A, B\} \geq \min\{\alpha, \beta\}$.

The Definitions 1.4 and 1.6 below were introduced in [12] and [15], generalizing Definition 1.5 of ultrametric matrices introduced in [11] (see also [14]).

Definition 1.4. *A nonnegative CBF matrix* U *is in nested block form (NBF) if in* (1.1) A *and* B *are NBF matrices and*

- $0 \leq \alpha \leq \beta$;
- $\min\{A_{ij}, A_{ji}\} \geq \alpha$ *and* $\min\{B_{kl}, B_{lk}\} \geq \alpha$;
- $\max\{A_{ij}, A_{ji}\} \geq \beta$ *and* $\max\{B_{kl}, B_{lk}\} \geq \beta$.

Definition 1.5. *A nonnegative symmetric matrix* U *is said to be an* ultrametric matrix *if*

(1) *for all* $i, j, U_{ii} \geq U_{ij}$,

(2) *for all* i, j, k, *the inequality* $U_{ij} \geq \min\{U_{ik}, U_{kj}\}$ *is satisfied.*

The matrix U *is* strictly *ultrametric if in* (1) *the inequality is strict.*

Remark 1.2. The name ultrametric comes from ultrametric distances. One may think as $U_{ij} = \frac{1}{\delta_{ij}}$ (for $i \neq j$), where δ is an ultrametric distance.

A possible generalization of this concept to the nonsymmetric case is the following.

Definition 1.6. *A nonnegative matrix* U *of size* n *is said to be a GUM if, for all* $i, j, U_{ii} \geq \max\{U_{ij}, U_{ji}\}$ *and, when* $n > 2$, *every three distinct elements* i, j, k *have a* preferred element. *Assume that this element is* i *which means*

- $U_{ij} = U_{ik}$;
- $U_{ji} = U_{ki}$;
- $\min\{U_{jk}, U_{kj}\} \geq \min\{U_{ji}, U_{ij}\}$;
- $\max\{U_{jk}, U_{kj}\} \geq \max\{U_{ji}, U_{ij}\}$.

By definition the transpose of a GUM is also a GUM. We note that an ultrametric matrix is a symmetric GUM. The study of the incidence graph for the inverse of an ultrametric matrix was done in [6] and for a GUM in [7] (this is the one step graph induced by a Markov chain associated with the matrix).

In the next result we summarize the main results obtained in [12] and [15] concerning GUM.

Theorem 1.7. *Let* U *be a nonnegative matrix.*

- U *is a GUM if and only if it is a permutation similar to a NBF.*
- *If* U *is a GUM, then it is nonsingular if and only if it does not contain a row of zeros and no two rows are the same.*
- *If* U *is a nonsingular GUM, then* $U \in bi\mathcal{P}$.

It is clear that if U is a GUM, then $\mathbb{I} + tU$ is a nonsingular GUM. In particular, $\tau(U) = \infty$.

We introduce a main object of this article.

Definition 1.8. *Given a function* f *and a matrix* U, *the matrix* $f(U)$ *is defined as* $f(U)_{ij} = f(U_{ij})$. *We shall say that* $f(U)$ *is a Hadamard function of* U.

Given two matrices A, B *of the same size, we denote by* $A \odot B$ *the Hadamard product of them, where* $(A \odot B)_{ij} = A_{ij} B_{ij}$.

Given a vector a, we denote by D_a the diagonal matrix whose diagonal is a. We have $D_a D_b = D_a \odot D_b = D_{a \odot b}$.

The class of CBF matrices (and its permutations) is closed under Hadamard functions. Similarly, the class of increasing CBF (and its permutations) is closed under increasing Hadamard functions.

On the other hand, the class of NBF, and therefore also the class of GUM, is stable under Hadamard nonnegative increasing functions. We summarize this result in the following proposition.

PROPOSITION 1.9. *Assume that U is a GUM and $f : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function. Then $f(U)$ is a GUM. In particular, $\tau(f(U)) = \infty$, and if $f(U)$ is nonsingular, then $f(U) \in bi\mathcal{P}$. A sufficient condition for $f(U)$ to be nonsingular is that U is nonsingular and f is strictly increasing.*

Proof. It is clear that $f(U)$ is a GUM, and therefore $\tau(f(U)) = \infty$. Then, from Proposition 1.1 we have that $f(U) \in bi\mathcal{P}$ as long as $f(U)$ is nonsingular. If U is nonsingular, then it does not contain a row (or column) of zeros, and there are not two equal rows (or columns). This condition is stable under strictly increasing nonnegative functions, so the result follows. □

One of our main results is a sort of reciprocal of the previous one. We shall prove that if $\tau(f(U)) = \infty$ for all increasing nonnegative functions f, then U must be a GUM (see Theorem 2.4).

Let us introduce the following index.

DEFINITION 1.10. *We say that a nonnegative matrix U is in class $\mathcal{T}$ if*

$$\tau(U) = \inf\{t > 0 : \ (\mathbb{I} + tU)^{-1}\mathbf{1} \ngeqq 0 \ \textit{or} \ \mathbf{1}'(\mathbb{I} + tU)^{-1} \ngeqq 0\},$$

and $\mathbb{I} + \tau(U)\,U$ is nonsingular whenever $\tau(U) < \infty$.

We shall prove that every nonnegative matrix U that is a permutation of an increasing CBF is in class $\mathcal{T}$.

We remark here that our purpose is to study Hadamard functions of matrices and not spectral functions of matrices, which are quite different concepts. For spectral functions of matrices there are deep and beautiful results for the same classes of matrices we consider here. See, for example, the work of Bouleau [3] for filtered operators. For M matrices, see the works of Varga [17], Micchelli and Willoughby [13], Ando [1], Fiedler and Schneider [9], and the recent work of Bapat, Catral, and Neummann [2] for M-matrices and inverse M-matrices.

2. Main results.

THEOREM 2.1. *Assume $U \in \mathcal{P}$ and that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a nonnegative strictly increasing convex function. Then $f(U)$ is nonsingular and $det(f(U)) > 0$. Also $f(U)$ has a right nonnegative equilibrium potential. Moreover, if $f(0) = 0$, we have that $M = U^{-1}f(U)$ is an M-matrix. If $U \in bi\mathcal{P}$, then $f(U)$ also has a left nonnegative equilibrium potential.*

Note that $H = f(U)^{-1}$ is not necessarily a Z-matrix; that is, for some $i \neq j$ it can happen that $H_{ij} > 0$, as the following example will show. Therefore the existence of a nonnegative right equilibrium potential, which is

$$\forall i, \ \ H_{ii} + \sum_{j \neq i} H_{ij} \geq 0,$$

does not necessarily imply that the inverse is row diagonally dominant, that is,

$$\forall i, \ \ H_{ii} \geq \sum_{j \neq i} |H_{ij}|.$$

Example 2.1. Consider the matrix

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

Then $U = (\mathbb{I} - P)^{-1} \in bi\mathcal{P}$. Consider the nonnegative strictly convex function $f(x) = x^2 - \cos(x) + 1$. A numerical computation gives

$$(f(U))^{-1} \approx \begin{pmatrix} 0.3590 & -0.0975 & 0.0027 \\ -0.0975 & 0.2372 & -0.0975 \\ 0.0027 & -0.0975 & 0.3590 \end{pmatrix},$$

which is not a Z-matrix.

We denote by $U^{(\alpha)}$ the Hadamard transformation of U under $f(x) = x^\alpha$. In particular, $U^{(2)} = U \odot U$. It was conjectured by Neumann in [16] that $U^{(2)}$ is an inverse M-matrix if U is so. This was solved by Chen in the beautiful article [4] for any positive integer power of U. Our next result is a generalization of Chen's result. His proof depends on the following interesting result: U is an inverse M-matrix if and only if its adjoint is a Z-matrix, and each proper principal submatrix is an inverse M-matrix. Our technique is entirely different and is based strongly on the idea of an equilibrium potential.

This result has the following probabilistic interpretation. If U is the potential of a transient continuous time Markov process, then $U^{(\alpha)}$ is also the potential of a transient continuous time Markov process. In Theorem 2.3 we show that the same is true for a potential of a Markov chain. An interesting open question is what is the relation between the Markov chain associated with U and that associated with $U^{(\alpha)}$.

THEOREM 2.2. *Assume $U \in \mathcal{M}^{-1}$ and $\alpha \geq 1$. Then $U^{(\alpha)} \in \mathcal{M}^{-1}$. If $U^{-1} \in \mathcal{P}$, then $(U^{(\alpha)})^{-1} \in \mathcal{P}$. If $U \in bi\mathcal{P}$, then $U^{(\alpha)} \in bi\mathcal{P}$.*

THEOREM 2.3. *Assume that $U^{-1} = \mathbb{I} - P$, where P is a sub-Markov kernel, that is, $P \geq 0$, $P\mathbf{1} \leq \mathbf{1}$. Then for all $\alpha \geq 1$ there is a sub-Markov kernel $Q(\alpha)$ such that $(U^{(\alpha)})^{-1} = \mathbb{I} - Q(\alpha)$. Moreover, if $P'\mathbf{1} \leq \mathbf{1}$, then $Q(\alpha)'\mathbf{1} \leq \mathbf{1}$.*

The next result establishes that the class of GUM is the largest class of potentials stable under increasing Hadamard functions.

THEOREM 2.4. *Let U be a nonnegative matrix such that $\tau(f(U)) = \infty$ for all increasing nonnegative functions f. Then, U must be a GUM.*

Example 2.2. Given $a, b, c, d \in \mathbb{R}_+$, consider the nonsingular matrix

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ a & b & 1 & 0 \\ c & d & 0 & 1 \end{pmatrix}.$$

For all increasing nonnegative functions f and all $t > 0$, $(\mathbb{I}+tf(U))^{-1}$ is an M-matrix, while U is not a GUM. Moreover, U is not a permutation of an increasing CBF. This shows that the last theorem does not hold if, in the definition of τ, we replace the class $bi\mathcal{P}$ by the class $\mathcal{M}^{-1}$.

THEOREM 2.5. *Let $U \in bi\mathcal{P}$ and $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a strictly increasing convex function. $f(U)$ is in $bi\mathcal{P}$ if and only if $f(U)$ belongs to the class $\mathcal{T}$.*

THEOREM 2.6. *If U is a nonnegative increasing CBF matrix, then U is in the class $\mathcal{T}$.*

As a corollary of the two previous theorems we obtain the following important result.

THEOREM 2.7. *Assume that $U \in bi\mathcal{P}$ is an increasing CBF matrix and that $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a nonnegative strictly increasing convex function. Then $f(U) \in bi\mathcal{P}$.*

3. Proofs of Theorems 2.1, 2.2, 2.3, and 2.5. Let us start with a useful lemma.

LEMMA 3.1. *Assume $U \in \mathcal{M}^{-1}$. Then for all $t \geq 0$, $(\mathbb{I} + tU) \in \mathcal{M}^{-1}$. Moreover, if $U \in \mathcal{P}$, then $(\mathbb{I} + tU) \in \mathcal{P}$ and its right equilibrium potential is strictly positive. In particular if $U \in bi\mathcal{P}$, then so is $\mathbb{I} + tU$, and its equilibrium potentials are strictly positive. Similarly, let $0 \leq s < t$ and assume $\mathbb{I} + tU \in bi\mathcal{P}$; then $\mathbb{I} + sU \in bi\mathcal{P}$, and its equilibrium potentials are strictly positive.*

Proof. For some $K > 0$ large enough, $U^{-1} = K(\mathbb{I} - N)$, where $N \geq 0$ (and $N\mathbf{1} \leq \mathbf{1}$ for the row diagonally dominant case). In what follows we can assume that $K = 1$. (It is enough to consider the matrix KU instead of U.)

From the equality $(\mathbb{I} - N)(\mathbb{I} + N + N^2 + \cdots + N^p) = \mathbb{I} - N^{p+1}$ we get that

$$\mathbb{I} + N + N^2 + \cdots + N^p = U(\mathbb{I} - N^{p+1}) \leq U.$$

We deduce that the series $\sum_{l=1}^{\infty} N^l$ is convergent and its limit is U.

Consider now the matrix

$$N_t = t\left(\left(\mathbb{I} - \frac{1}{1+t}N\right)^{-1} - \mathbb{I}\right) = t\sum_{l=1}^{\infty}\left(\frac{1}{1+t}\right)^l N^l.$$

We have $N_t \geq 0$. In the case $N\mathbf{1} \leq \mathbf{1}$, since N is a nonnegative matrix we deduce that $N^l\mathbf{1} \leq \mathbf{1}$. This allows us to prove

$$N_t\mathbf{1} = t\sum_{l=1}^{\infty}\left(\frac{1}{1+t}\right)^l N^l\mathbf{1} \leq t\sum_{l=1}^{\infty}\left(\frac{1}{1+t}\right)^l \mathbf{1} = \mathbf{1}.$$

Therefore the matrix $\mathbb{I} - N_t$ is a Z-matrix (which is row diagonally dominant when U^{-1} is). On the other hand, we have

$$\mathbb{I} + tU = \mathbb{I} + t(\mathbb{I} - N)^{-1} = (t\mathbb{I} + \mathbb{I} - N)(\mathbb{I} - N)^{-1} = (1+t)\left(\mathbb{I} - \frac{1}{1+t}N\right)(\mathbb{I} - N)^{-1},$$

and we deduce that $\mathbb{I} + tU$ is nonsingular and its inverse is

$$\begin{aligned}(\mathbb{I} + tU)^{-1} &= \frac{1}{1+t}(\mathbb{I} - N)\left(\mathbb{I} - \frac{1}{1+t}N\right)^{-1} \\ &= \frac{1}{1+t}\left(\left(\mathbb{I} - \frac{1}{1+t}N\right)^{-1} - N\left(\mathbb{I} - \frac{1}{1+t}N\right)^{-1}\right) \\ &= \frac{1}{1+t}\left(\sum_{l=0}^{\infty}(1+t)^{-l}N^l - \sum_{l=0}^{\infty}(1+t)^{-l}N^{l+1}\right) \\ &= \frac{1}{1+t}(\mathbb{I} - N_t).\end{aligned}$$

This shows that the inverse of $\mathbb{I} - N_t$ is nonnegative, and therefore this matrix is an M-matrix. We conclude $\mathbb{I} + tU \in \mathcal{M}^{-1}$.

The only thing left to prove is that $N_t\mathbf{1} < \mathbf{1}$ in the row diagonally dominant case, that is, when $N\mathbf{1} \leq \mathbf{1}$. Notice that from the convergence of the series $\sum_{l \geq 0} N^l$ we deduce that $N^l \to 0$ as $l \to \infty$. Then for large l, say $l > l_0$, we have $N^l\mathbf{1} \leq \frac{1}{2}\mathbf{1}$. Thus

$$N_t\mathbf{1} = t\sum_{l=1}^{\infty}\left(\frac{1}{1+t}\right)^l N^l\mathbf{1} \leq t\left(\sum_{l=1}^{l_0}\left(\frac{1}{1+t}\right)^l + \frac{1}{2}\sum_{l=l_0+1}^{\infty}\left(\frac{1}{1+t}\right)^l\right)\mathbf{1} < \mathbf{1}.$$

For a general $K > 0$ we have the equality $(\mathbb{I}+tU)^{-1} = \frac{K}{t+K}(\mathbb{I}-\frac{t}{K}\sum_{l=1}^{\infty}(\frac{K}{t+K})^l N^l)$, where $N = \mathbb{I} - \frac{1}{K}U^{-1}$.

Finally, assume that $\mathbb{I} + tU \in bi\mathcal{P}$. Hence $\mathbb{I} + \beta(\mathbb{I} + tU) \in bi\mathcal{P}$ for all $\beta \geq 0$. This implies that

$$\mathbb{I} + \frac{\beta}{1+\beta} tU \in bi\mathcal{P}.$$

Now it is enough to take $\beta \geq 0$ such that $s = \frac{\beta}{1+\beta}t$. □

This lemma has two immediate consequences.

Corollary 3.2. *If $U \in bi\mathcal{P}$, then $\tau(U) = \infty$.*

Corollary 3.3. *Let U be a nonnegative matrix; then*

$$\tau(U) = \sup\{t \geq 0 : \ \mathbb{I} + tU \in bi\mathcal{P}\}.$$

Proof. It is clear that $\tau(U) \leq \sup\{t \geq 0 : \ \mathbb{I} + tU \in bi\mathcal{P}\}$. On the other hand, if $\mathbb{I} + tU \in bi\mathcal{P}$, we get $\mathbb{I} + sU \in bi\mathcal{P}$ for all $0 \leq s \leq t$. This fact and the definition of $\tau(U)$ imply the result. □

Proof of Theorem 2.1. We first assume that $f(0) = 0$. We have that $U^{-1} = K(\mathbb{I} - P)$ for some $K > 0$ and P a substochastic matrix. Without loss of generality we can assume $K = 1$, because it is enough to consider KU instead of U and $\tilde{f}(x) = f(x/K)$ instead of f.

Consider $M = (U^{-1}f(U))$. For $i \neq j$ let us compute

$$M_{ij} = (U^{-1}f(U))_{ij} = (1 - p_{ii})f(U_{ij}) - \sum_{k \neq i} p_{ik} f(U_{kj}).$$

Since $1 - p_{ii} - \sum_{k \neq i} p_{ik} \geq 0$ (which is equivalent to $\sum_k p_{ik} \leq 1$) and f is convex, we obtain

$$\left(1 - \sum_k p_{ik}\right) f(0) + \sum_k p_{ik} f(U_{kj}) \geq f\left(\sum_k p_{ik} U_{kj}\right) = f(U_{ij}).$$

The last equality follows from the fact that $U^{-1} = \mathbb{I} - P$. This shows that $M_{ij} \leq 0$. Consider now a positive vector x such that $y' = x'U^{-1} > 0$ (for its existence, see [10, Theorem 2.5.3]). Then

$$x'M = x'U^{-1}f(U) = y'f(U) > 0,$$

which implies, by the same cited theorem in [10], that M is an M-matrix. In particular, M is nonsingular and $\det(M) > 0$. So $f(U)$ is nonsingular and $\det(f(U)) > 0$. Now consider ρ the right equilibrium potential of $f(U)$. We have

$$M\rho = U^{-1}f(U)\rho = U^{-1}\mathbf{1} = \mu_U \geq 0,$$

then $\rho = M^{-1}\mu_U \geq 0$, because M^{-1} is a nonnegative matrix. This means that $f(U)$ possesses a nonnegative right equilibrium potential. Since $f(U)$ is nonsingular, we also have a left equilibrium potential, but we do not know whether it is nonnegative. Then the first part is proven under the extra hypothesis that $f(0) = 0$.

Assume now $a = f(0) > 0$, and consider $g(x) = f(x) - a$, which is a strictly increasing convex function. Obviously $f(U) = g(U) + a\mathbf{1}\mathbf{1}'$, so

$$\mu_{f(U)} = \frac{1}{1 + a\bar{\mu}_{g(U)}} \mu_{g(U)} \geq 0, \qquad \nu_{f(U)} = \frac{1}{1 + a\bar{\mu}_{g(U)}} \nu_{g(U)},$$

where $\bar{\mu}_{g(U))} = \mathbf{1}'\mu_{g(U))} > 0$. We have used the fact that $\bar{\mu}_{g(U))} = \bar{\nu}_{g(U))}$. Thus $f(U)$ has a nonnegative right equilibrium potential and a left equilibrium potential. We need to prove that $f(U)$ is nonsingular and $\det(f(U)) > 0$. This follows immediately from the equality

$$f(U) = g(U)(\mathbb{I} + a\mu_{g(U)}\mathbf{1}'),$$

which implies

$$f(U)^{-1} = g(U)^{-1} - \frac{a}{1 + a\bar{\mu}_{g(U)}}\mu_{g(U)}(\nu_{g(U)});$$
$$\det(f(U)) = \det(g(U))(1 + a\bar{\mu}_{g(U)}).$$

Then the first part of the result is proven.

In the bipotential case use U' instead of U to obtain the existence of a nonnegative left equilibrium potential for $f(U)$. □

Proof of Theorem 2.5. Using the same ideas as above, we can assume that $f(0) = 0$. Also we have that $U^{-1}(\mathbb{I} + tf(U)) = M_t$ is an M-matrix for all $t \geq 0$. Therefore $\mathbb{I}+tf(U)$ is nonsingular for all t, and we denote by μ_t and ν_t the equilibrium potentials for $\mathbb{I} + tf(U)$.

Assume first that $f(U)$ is in class $\mathcal{T}$ (see Definition 1.10), which means that

$$\tau(f(U)) = \min\{t > 0 : \ \mu_t \ngeq 0 \text{ or } \nu_t \ngeq 0\}.$$

We prove that for all $t \geq 0$, μ_t, ν_t are nonnegative. Since

$$M_t\mu_t = U^{-1}\mathbf{1} = \mu_U,$$

we obtain that $\mu_t = M_t^{-1}\mu_U \geq 0$, because M_t^{-1} is a nonnegative matrix. Thus, $\tau(f(U)) = \infty$, and since $f(U)$ is nonsingular we get from Proposition 1.1 that $f(U) \in bi\mathcal{P}$. Conversely if $f(U) \in bi\mathcal{P}$, then $\tau(f(U)) = \infty$, and the result follows. □

For the rest of the section n denotes the size of U.

Lemma 3.4. *Assume that $U \in \mathcal{P}$. Then any principal square submatrix A of U is also in class $\mathcal{P}$. The same is true if we replace $\mathcal{P}$ by $bi\mathcal{P}$.*

Proof. By induction and a suitable permutation the restriction of U to $\{1, \dots, n-1\} \times \{1, \dots, n-1\}$ is enough to prove the result for A. Assume that

$$U = \begin{pmatrix} A & b \\ c' & d \end{pmatrix} \quad \text{and} \quad U^{-1} = \begin{pmatrix} \Lambda & -\zeta \\ -\varrho' & \theta \end{pmatrix}.$$

Since $A^{-1} = \Lambda - \frac{1}{\theta}\zeta\varrho'$ we get that the off-diagonal elements of A^{-1} are nonpositive. It is quite easy to see that the result will follow as soon as $A^{-1}\mathbf{1} \geq 0$.

Since $U \in \mathcal{P}$ we have that $\Lambda\mathbf{1} - \zeta \geq 0$ and $\theta \geq \varrho'\mathbf{1}$. Therefore,

$$A^{-1}\mathbf{1} = \Lambda\mathbf{1} - \frac{1}{\theta}\zeta\varrho'\mathbf{1} = \Lambda\mathbf{1} - \frac{\varrho'\mathbf{1}}{\theta}\,\zeta \geq \Lambda\mathbf{1} - \zeta \geq 0. \qquad □$$

Recall that for a vector a, D_a is the associated diagonal matrix.

Lemma 3.5. *Assume $U \in bi\mathcal{P}$ and $\alpha \geq 1$. If*

$$U = \begin{pmatrix} A & b \\ c' & d \end{pmatrix},$$

then there exists a nonnegative vector η *such that*

$$A^{(\alpha)}\eta = b^{(\alpha)}.$$

Proof. We first perturb the matrix U to have a positive matrix. Consider $\epsilon > 0$ and the positive matrix $U_\epsilon = U + \epsilon\mathbf{1}\mathbf{1}'$. It is direct to prove that

$$U_\epsilon^{-1} = U^{-1} - \frac{\epsilon}{1+\epsilon\bar{\mu}_U}\mu_U(\nu_U)',$$

where $\bar{\mu}_U = \mathbf{1}'\mu_U$ is the total mass of μ_U. Then $U_\epsilon \in bi\mathcal{P}$, and its equilibrium potentials are given by

$$\mu_{U_\epsilon} = \frac{1}{1+\epsilon\bar{\mu}_U}\mu_U, \qquad \nu_{U_\epsilon} = \frac{1}{1+\epsilon\bar{\nu}_U}\nu_U.$$

We decompose the inverse of U_ϵ as

$$U_\epsilon^{-1} = \begin{pmatrix} \Lambda_\epsilon & \zeta_\epsilon \\ \varrho_\epsilon' & \theta_\epsilon \end{pmatrix},$$

and we notice that $A_\epsilon\zeta_\epsilon + \theta_\epsilon b_\epsilon = 0$, which implies that

$$b_\epsilon = A_\epsilon\lambda_\epsilon,$$

with $\lambda_\epsilon = -\frac{1}{\theta_\epsilon}\zeta_\epsilon \geq 0$. Also we mention here that λ_ϵ is a subprobability vector, that is, $\mathbf{1}'\lambda_\epsilon \leq 1$. This follows from the fact that U_ϵ^{-1} is column diagonally dominant.

Take now the matrix $V_\epsilon = D_{b_\epsilon}^{-1}A_\epsilon$. It is direct to check that $V_\epsilon \in \mathcal{M}^{-1}$ and that its equilibrium potentials are

$$\mu_{V_\epsilon} = \lambda_\epsilon, \qquad \nu_{V_\epsilon} = D_{b_\epsilon}\nu_{A_\epsilon}.$$

Thus $V_\epsilon \in bi\mathcal{P}$, and we can apply Theorem 2.1 to get that the matrix $V_\epsilon^{(\alpha)}$ possesses a right equilibrium potential $\eta_\epsilon \geq 0$; that is, for all i,

$$\sum_j (V_\epsilon^{(\alpha)})_{ij}(\eta_\epsilon)_j = 1,$$

which is equivalent to

$$\sum_j \frac{(A_\epsilon)_{ij}^\alpha}{(b_\epsilon)_i^\alpha}(\eta_\epsilon)_j = 1.$$

Hence

$$A_\epsilon^{(\alpha)}\eta_\epsilon = b_\epsilon^{(\alpha)}.$$

Recall that the matrix $A^{(\alpha)}$ is nonsingular. Since obviously $A_\epsilon^{(\alpha)} \to A^{(\alpha)}$ as $\epsilon \to 0$, we get

$$\eta_\epsilon \to \eta = (A^{(\alpha)})^{-1}b^{(\alpha)},$$

and the result follows. □

Proof of Theorem 2.2. Consider first the case where $U \in bi\mathcal{P}$. We already know that $U^{(\alpha)}$ is nonsingular and that it has left and right nonnegative equilibrium potentials. Therefore, in order to prove $U^{(\alpha)} \in bi\mathcal{P}$, it is enough to prove that $(U^{(\alpha)})^{-1}$ is a Z-matrix; that is, we need to show $((U^{(\alpha)})^{-1})_{ij} \le 0$ for $i \neq j$. An argument based on permutations shows that it is enough to prove the claim for $i = 1, j = n$.

Decompose $U^{(\alpha)}$ and its inverse as follows:

$$U^{(\alpha)} = \begin{pmatrix} A^{(\alpha)} & b^{(\alpha)} \\ (c^{(\alpha)})' & d^{\alpha} \end{pmatrix} \quad \text{and} \quad (U^{(\alpha)})^{-1} = \begin{pmatrix} \Omega & -\beta \\ -\alpha' & \delta \end{pmatrix}.$$

We will show $\beta \ge 0$. We notice that $\delta = \frac{\det(A^{(\alpha)})}{\det(U^{(\alpha)})} > 0$ and $-A^{(\alpha)}\beta + \delta b^{(\alpha)} = 0$, and we deduce

$$b^{(\alpha)} = A^{(\alpha)} \left(\frac{\beta}{\delta}\right).$$

Therefore, $\frac{\beta}{\delta} = \eta \ge 0$, where η is the vector given in Lemma 3.5. Thus $\beta \ge 0$, and the result is proven for the case $U \in bi\mathcal{P}$.

Now, consider $U = M^{-1}$ the inverse of the M-matrix M. Using Theorem 2.5.3 in [10], we get the existence of two positive diagonal matrices D, E such that DME is a strictly row and column diagonally dominant M-matrix. Thus $V = E^{-1}UD^{-1}$ is in $bi\mathcal{P}$, from which it follows that $V^{(\alpha)} \in bi\mathcal{P}$. Hence, $U^{(\alpha)} = E^{(\alpha)}V^{(\alpha)}D^{(\alpha)}$ is the inverse of an M-matrix. The rest of the result is proven in a similar way. □

Proof of Theorem 2.3. By hypothesis we have $U = \mathbb{I} - P$, where $P \ge 0$ and $P\mathbf{1} \le \mathbf{1}$. We notice that U is diagonally dominant on each column, which means that for all i, j

$$U_{ii} \ge U_{ji}.$$

Also we notice that $U = \mathbb{I} + PU$ and therefore $U_{ii} \ge 1$.

According to Theorem 2.2 we know that $H = (U^{(\alpha)})^{-1}$ is a row diagonally dominant M-matrix. The only thing left to prove is that the diagonal elements of H are dominated by one: $H_{ii} \le 1$ for all i. We will prove it for $i = n$.

Consider the following decompositions:

$$U = \begin{pmatrix} A & b \\ c' & d \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} \Lambda & -\omega \\ -\eta' & \gamma \end{pmatrix}, \quad (U^{(\alpha)})^{-1} = \begin{pmatrix} \Omega & -\beta \\ -\alpha' & \delta \end{pmatrix},$$

$$U^{-1}U^{(\alpha)} = \begin{pmatrix} \Xi & -\zeta \\ -\chi' & \rho \end{pmatrix}.$$

A direct computation gives that

$$\gamma = \rho\delta + \chi'\beta \ge \rho\delta.$$

We need to show that $\delta \le 1$. By hypothesis, $\gamma \le 1$; then it is enough to prove that $\rho \ge 1$. On the one hand, we have

$$\rho = (1 - p_{nn})U_{nn}^{\alpha} - \sum_{j \neq n} p_{nj} U_{jn}^{\alpha} = U_{nn}^{\alpha} - \sum_{j} p_{nj} U_{jn}^{\alpha} = U_{nn}^{\alpha} - \sum_{j} p_{nj} U_{jn} U_{jn}^{\alpha - 1}.$$

On the other hand, we also have $U_{jn}^{\alpha-1} \le U_{nn}^{\alpha-1}$ and $\sum_j p_{nj} U_{jn} = U_{nn} - 1$. Hence we deduce

$$\rho \ge U_{nn}^{\alpha-1} \ge 1.$$

This finishes the first part of the theorem . The rest of the result is proven by using U' instead of U. $\square$

4. Proof of Theorem 2.4. Notice that U is a GUM if and only if $n \le 2$ or every principal submatrix of size 3 is a GUM.

Since by hypothesis the matrix $\mathbb{I} + tU$ is a bipotential, it is diagonally dominant,

$$1 + tU_{ii} \ge tU_{ij},$$

and by taking $t \to \infty$, we find $U_{ii} \ge U_{ij}$. This proves the result when $n \le 2$. So, in what follows we assume $n \ge 3$.

Consider A any principal submatrix of U, of size 3×3. Since $\mathbb{I}+tf(A)$ is a principal submatrix of $\mathbb{I} + tf(U)$, we deduce that $\mathbb{I} + tf(A) \in bi\mathcal{P}$ (as long as $\mathbb{I} + tf(U) \in bi\mathcal{P}$). If the result holds for the 3×3 matrices, we deduce that A is a GUM, implying that U is also a GUM.

Thus, in the rest of the proof we can assume that U is a 3×3 matrix that verifies the hypothesis of the theorem. After a suitable permutation we can further assume that

$$U = \begin{pmatrix} a & b_1 & b_2 \\ c_1 & d & \alpha \\ c_2 & \beta & e \end{pmatrix},$$

where $\alpha = \min\{U_{ij} :\ i \ne j\} = \min\{U\}$ and $\beta = \min\{U_{ji} :\ U_{ij} = \alpha,\ i \ne j\}$.

Since U is diagonally dominant we have $\min\{a, d, e\} \ge \alpha$. Take f increasing such that $f(\alpha) = 0$ and $f(x) > 0$ for $x > \alpha$. Then,

$$\mathbb{I} + f(U) = \begin{pmatrix} 1 + f(a) & f(b_1) & f(b_2) \\ f(c_1) & 1 + f(d) & 0 \\ f(c_2) & f(\beta) & 1 + f(e) \end{pmatrix}$$

is a $bi\mathcal{P}$-matrix whose inverse we denote by

$$\begin{pmatrix} \delta & -\rho_1 & -\rho_2 \\ -\theta_1 & \gamma_1 & -\gamma_2 \\ -\theta_2 & -\gamma_3 & \gamma_4 \end{pmatrix}.$$

In particular we obtain

$$\begin{pmatrix} 1 + f(d) & 0 \\ f(\beta) & 1 + f(e) \end{pmatrix}^{-1} = \begin{pmatrix} \gamma_1 & -\gamma_2 \\ -\gamma_3 & \gamma_4 \end{pmatrix} - \frac{1}{\delta} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}',$$

and we deduce that

$$0 = \gamma_2 = \theta_1 \rho_2. \tag{4.1}$$

- *Case* $\rho_2 = 0$. We get $f(b_2) = 0$, which implies further

$$b_2 = \alpha \quad \text{and} \quad c_2 \geq \beta. \tag{4.2}$$

The last conclusion follows from the definition of β. Therefore,

$$U = \begin{pmatrix} a & b_1 & \alpha \\ c_1 & d & \alpha \\ c_2 & \beta & e \end{pmatrix}. \tag{4.3}$$

We must prove that U is GUM.
Consider another increasing function g such that $g(\beta) = 0$ and $g(x) > 0$ for $x > \beta$. Then,

$$\mathbb{I} + g(U) = \begin{pmatrix} 1+g(a) & g(b_1) & 0 \\ g(c_1) & 1+g(d) & 0 \\ g(c_2) & 0 & 1+g(e) \end{pmatrix}.$$

Its inverse is of the form

$$\begin{pmatrix} \tilde{\delta} & -\tilde{\rho}_1 & 0 \\ -\tilde{\theta}_1 & \tilde{\gamma}_1 & 0 \\ -\tilde{\theta}_2 & -\tilde{\gamma}_3 & \tilde{\gamma}_4 \end{pmatrix}.$$

As before, we deduce that $0 = \tilde{\gamma}_3 = \tilde{\theta}_2 \tilde{\rho}_1$.

  - *Subcase* $\tilde{\theta}_2 = 0$. We have $g(c_2) = 0$, which implies $c_2 = \beta$. In this situation we have

$$U = \begin{pmatrix} a & b_1 & \alpha \\ c_1 & d & \alpha \\ \beta & \beta & e \end{pmatrix}.$$

By permuting rows and columns $1, 2$, if necessary, we can assume that $b_1 \leq c_1$. Consider the situation where $c_1 < \beta$; of course, implicitly we should have $\alpha < \beta$. Under a suitable increasing transformation h, we get

$$\mathbb{I} + h(U) = \begin{pmatrix} 1+h(a) & 0 & 0 \\ 0 & 1+h(d) & 0 \\ h(\beta) & h(\beta) & 1+h(e) \end{pmatrix}$$

and its inverse

$$\begin{pmatrix} \frac{1}{1+h(a)} & 0 & 0 \\ 0 & \frac{1}{1+h(d)} & 0 \\ -\frac{h(\beta)}{(1+h(a))(1+h(e))} & -\frac{h(\beta)}{(1+h(d))(1+h(e))} & \frac{1}{1+h(e)} \end{pmatrix}.$$

The sum of the third row is then

$$\frac{1}{1+h(e)} \left(1 - h(\beta) \left(\frac{1}{1+h(a)} + \frac{1}{1+h(d)} \right) \right),$$

and this quantity can be made negative by choosing an appropriate function h. The idea is to make $h(\beta) \to \infty$ and

$$\frac{h(\beta)}{\max\{h(a), h(d)\}} \to 1.$$

Therefore, $c_1 \geq \beta$ and U is a GUM.

- *Subcase* $\tilde{\rho}_1 = 0$. We have $g(b_1) = 0$ and then $b_1 \le \beta$. Take again an increasing function ℓ such that

$$\mathbb{I} + \ell(U) = \begin{pmatrix} 1+\ell(a) & 0 & 0 \\ \ell(c_1) & 1+\ell(d) & 0 \\ \ell(c_2) & 0 & 1+\ell(e) \end{pmatrix}$$

and its inverse

$$\begin{pmatrix} \frac{1}{1+\ell(a)} & 0 & 0 \\ -\frac{\ell(c_1)}{(1+\ell(a))(1+\ell(d))} & \frac{1}{1+\ell(d)} & 0 \\ -\frac{\ell(c_2)}{(1+\ell(a))(1+\ell(e))} & 0 & \frac{1}{1+\ell(e)} \end{pmatrix}.$$

The sum of the first column is

$$\frac{1}{1+\ell(a)}\left(1 - \frac{\ell(c_1)}{(1+\ell(d))} - \frac{\ell(c_2)}{(1+\ell(e))}\right),$$

which can be made negative by repeating a similar argument as before if both $c_1 > \beta$ and $c_2 > \beta$.

Therefore if we assume that $c_1 > \beta$, we necessarily have $c_2 \le \beta$. On the other hand, from (4.2) we know $c_2 \ge \beta$, proving that $c_2 = \beta$. The conclusion is $\alpha \le b_1 \le \beta < c_1$ and

$$U = \begin{pmatrix} a & b_1 & \alpha \\ c_1 & d & \alpha \\ \beta & \beta & e \end{pmatrix},$$

which is a GUM.

Therefore we can continue under the hypothesis $c_1 \le \beta \le c_2$.

  * *Subsubcase* $b_1 < \beta$. Again we must have $\alpha < \beta$. Under this condition we have that $c_2 > \alpha$. Using an increasing function ω, we get

$$\mathbb{I} + \omega(U) = \begin{pmatrix} 1+\omega(a) & 0 & 0 \\ \omega(c_1) & 1+\omega(d) & 0 \\ \omega(c_2) & \omega(\beta) & 1+\omega(e) \end{pmatrix},$$

and its inverse is

$$\begin{pmatrix} \frac{1}{1+\omega(a)} & 0 & 0 \\ -\frac{\omega(c_1)}{(1+\omega(a))(1+\omega(d))} & \frac{1}{1+\omega(d)} & 0 \\ -\frac{\omega(c_2)(1+\omega(d))-\omega(\beta)\omega(c_1)}{(1+\omega(a))(1+\omega(d))(1+\omega(e))} & -\frac{\omega(\beta)}{(1+\omega(d))(1+\omega(e))} & \frac{1}{1+\omega(e)} \end{pmatrix}.$$

The sum of the third row is

$$\frac{1}{(1+\omega(e))}\left(1 - \frac{\omega(c_2)}{1+\omega(a)} + \frac{\omega(\beta)\omega(c_1)}{(1+\omega(a))(1+\omega(d))} - \frac{\omega(\beta)}{1+\omega(d)}\right). \tag{4.4}$$

If $c_1 < \beta$, we can assume $\omega(c_1) = 0$. With this choice we can make the sum in (4.4) negative by a suitable selection of ω as we did before. Thus we must have $c_1 = \beta$, in which case the sum under study is proportional to

$$1 - \frac{\omega(c_2)}{1+\omega(a)} + \frac{\omega(\beta)^2}{(1+\omega(a))(1+\omega(d))} - \frac{\omega(\beta)}{1+\omega(d)}. \tag{4.5}$$

If $c_2 = \beta$, then

$$U = \begin{pmatrix} a & b_1 & \alpha \\ \beta & d & \alpha \\ \beta & \beta & e \end{pmatrix}$$

is a GUM. So, we must analyze the case where $c_2 > \beta$ in (4.5). We will arrive at a contradiction by taking an asymptotic as before. Consider a fixed number $\lambda \in (0,1)$. Choose a family of functions $(\omega_r)_{r\in\mathbb{N}}$ such that, as $r \to \infty$,

$$\omega_r(\beta) \to \infty, \quad \frac{\omega_r(\beta)}{\omega_r(c_2)} \to \lambda, \quad \frac{\omega_r(c_2)}{\omega_r(a)} \to 1, \quad \frac{\omega_r(d)}{\omega_r(a)} \to \phi,$$

where $\phi = 1$ if $d > \beta$, and $\phi = \lambda$ if $d = \beta$. The asymptotic of (4.5) is then

$$1 - 1 + \frac{\lambda^2}{\phi} - \frac{\lambda}{\phi}.$$

This quantity is strictly negative for the two possible values of ϕ, which is a contradiction, and therefore $c_2 = \beta$.

To finish with the Subcase $\tilde{\rho}_1 = 0$, which will in turn finish with Case $\rho_2 = 0$, we consider a further subcase.

∗ *Subsubcase* $b_1 = \beta$. We recall that we are under the restrictions $c_1 \le \beta \le c_2$ and

$$U = \begin{pmatrix} a & \beta & \alpha \\ c_1 & d & \alpha \\ c_2 & \beta & e \end{pmatrix}.$$

Notice that if $c_2 = \beta$, then U is GUM. So, we may assume in this part that $c_2 > \beta$. If $c_1 = \alpha$, we can permute 1 and 2 to get

$$\Pi U \Pi' = \begin{pmatrix} d & \alpha & \alpha \\ \beta & a & \alpha \\ \beta & c_2 & e \end{pmatrix},$$

which is also in NBF, and U is a GUM. Thus we can assume $c_1 > \alpha$, and again we have $\alpha < \beta$.

Take an increasing function m such that

$$\mathbb{I} + m(U) = \begin{pmatrix} 1+m(a) & m(\beta) & 0 \\ m(c_1) & 1+m(d) & 0 \\ m(c_2) & m(\beta) & 1+m(e) \end{pmatrix}.$$

We take the asymptotic under the following restrictions:

$$\frac{m(\beta)}{m(a)} \to \lambda \in (0,1), \ \frac{m(c_1)}{m(a)} \to \lambda, \ \frac{m(e)}{m(a)} \to 1, \ \frac{m(c_2)}{m(a)} \to 1, \ \frac{m(d)}{m(a)} \to \phi,$$

where $\phi = 1$ if $d > \beta$, and $\phi = \lambda$ if $d = \beta$. The limiting matrix for $\frac{1}{m(a)}(\mathbb{I} + m(U))$ is

$$V = \begin{pmatrix} 1 & \lambda & 0 \\ \lambda & \phi & 0 \\ 1 & \lambda & 1 \end{pmatrix},$$

whose determinant is $\Delta = \phi - \lambda^2 > 0$. Therefore it is nonsingular, and as the limit of matrices in $bi\mathcal{P}$, V itself must belong to $bi\mathcal{P}$. On the other hand, the inverse of V is given by

$$V^{-1} = \frac{1}{\Delta} \begin{pmatrix} \phi & -\lambda & 0 \\ -\lambda & 1 & 0 \\ -(\phi - \lambda^2) & 0 & \phi - \lambda^2 \end{pmatrix},$$

and the sum of the first column is

$$\frac{\lambda^2 - \lambda}{\Delta} < 0,$$

which is a contradiction.

This finishes with the subcase $\rho_2 = 0$, and we return to (4.1) to consider now the following case.

- *Case* $\theta_1 = 0$. Under this condition we get $c_1 = \alpha$ and

$$U = \begin{pmatrix} a & b_1 & b_2 \\ \alpha & d & \alpha \\ c_2 & \beta & e \end{pmatrix}.$$

Consider the transpose of U and permute on it 2 and 3, to obtain the matrix

$$\tilde{U} = \begin{pmatrix} a & c_2 & \alpha \\ b_2 & e & \alpha \\ b_1 & \beta & d \end{pmatrix},$$

where now $b_1 \geq \beta$. Clearly the matrix $\tilde{U}$ verifies the hypothesis of the theorem and has the shape of (4.3); that is, we are in the "case $\rho_2 = 0$," which, we already know, implies that $\tilde{U}$ is a GUM. Then U itself is a GUM, and the theorem is proven. □

5. Filtered matrices and sufficient conditions for classes $bi\mathcal{P}$ and $\mathcal{T}$.

The class of filtered matrices, which turn out to be a generalization of GUM, gives a good framework to study a potential theory of matrices. They were introduced as operators in [8] to generalize the class of self-adjoint operators whose spectral decomposition is written in terms of conditional expectations (see, for instance, [3], [5], and [11]).

The basic tool to construct these matrices is partitions of $\mathcal{J}_n = \{1, \ldots, n\}$. The components of a partition $\mathcal{R}$ are called atoms, and we denote by $\overset{\mathcal{R}}{\sim}$ the equivalence relation induced by $\mathcal{R}$. Then i, j are in the same atom of $\mathcal{R}$ if and only if $i \overset{\mathcal{R}}{\sim} j$.

A partition $\mathcal{R}$ is coarser than or equal to a partition $\mathcal{Q}$ if the atoms of $\mathcal{Q}$ are contained in the atoms of $\mathcal{R}$. This (partial) order relation is denoted by $\mathcal{R} \preccurlyeq \mathcal{Q}$. It is also said that $\mathcal{Q}$ is finer than $\mathcal{R}$. For example, in $\mathcal{J}_4$ we have $\mathcal{R} = \{\{1, 2\}, \{3, 4\}\} \preccurlyeq$

$\mathcal{Q} = \{\{1\},\{2\},\{3,4\}\}$. The coarsest partition is the trivial one $\mathcal{N} = \{\mathcal{J}_n\}$, and the finest one is the discrete partition $\mathcal{F} = \{\{1\},\{2\},\ldots,\{n\}\}$.

Definition 5.1. *A* filtration $\mathbb{F} = \{\mathcal{R}_0 \prec \mathcal{R}_1 \prec \cdots \prec \mathcal{R}_k\}$ *is a strictly increasing sequence of comparable partitions.* $\mathbb{F}$ *is said to be* dyadic *if each nontrivial atom of* $\mathcal{R}_s$ *is divided into two atoms of* $\mathcal{R}_{s+1}$.

A filtration in the wide sense is an increasing sequence of comparable partitions $\mathbb{G} = \{\mathcal{R}_0 \preccurlyeq \mathcal{R}_1 \preccurlyeq \cdots \preccurlyeq \mathcal{R}_k\}$.

The difference between a filtration and a filtration in the wide sense is that in the latter case repetition of partitions is allowed.

Each partition $\mathcal{R}$ induces an incidence matrix $F =: F(\mathcal{R})$ given by

$$F_{ij} = \begin{cases} 1 & \text{if } i \overset{\mathcal{R}}{\sim} j, \\ 0 & \text{otherwise .} \end{cases}$$

A vector $v \in \mathbb{R}^n$ is said to be $\mathcal{R}$-measurable if v is constant on the atoms of $\mathcal{R}$, that is,

$$i \overset{\mathcal{R}}{\sim} j \Rightarrow v_i = v_j.$$

This can be expressed in terms of standard matrix operations as

$$F(\mathcal{R})v = D_{w_\mathcal{R}} v,$$

where $w_\mathcal{R} = F(\mathcal{R})\mathbf{1}$ is the vector constant on each atom, and this constant is the size of the respective atom (recall that D_z is the diagonal matrix associated with the vector z). The set of $\mathcal{R}$-measurable vectors is a linear subspace of $\mathbb{R}^n$. Notice that if the partition is $\mathcal{F}$, then the associated incidence matrix is the identity and the subspace of measurable vectors is just $\mathbb{R}^n$. On the other hand, if the partition is the trivial one $\mathcal{N}$, then the incidence matrix is $\mathbf{1}\mathbf{1}'$ and the measurable vectors in this case are the constant ones.

Definition 5.2. *A matrix* U *is said to be* filtered *if there exists a filtration in the wide sense* $\mathbb{G} = \{\mathcal{Q}_0 \preccurlyeq \mathcal{Q}_1 \preccurlyeq \cdots \preccurlyeq \mathcal{Q}_l\}$, *vectors* $\mathfrak{a}_0,\ldots,\mathfrak{a}_l$, $\mathfrak{b}_0,\ldots,\mathfrak{b}_l$ *with the restriction that* $\mathfrak{a}_s, \mathfrak{b}_s$ *are* $\mathcal{Q}_{s+1}$*-measurable (we take* $\mathcal{Q}_{l+1} = \mathcal{F}$ *the discrete partition), and*

$$U = \sum_{s=0}^{\ell} D_{\mathfrak{a}_s} F(\mathcal{Q}_s) D_{\mathfrak{b}_s}. \tag{5.1}$$

There is no loss of generality if we assume that $\mathcal{Q}_0 = \mathcal{N}$ and $\mathcal{Q}_\ell = \mathcal{F}$, that is, $F(\mathcal{Q}_0) = \mathbf{1}\mathbf{1}'$ and $F(\mathcal{Q}_\ell) = \mathbb{I}$. Let us see that (5.1) can be simply written in terms of a filtration. Indeed, notice that if $\mathfrak{a}_s$ and $\mathfrak{b}_s$ are $\mathcal{Q}_s$-measurable, then

$$D_{\mathfrak{a}_s} F(\mathcal{Q}_s) D_{\mathfrak{b}_s} = D_{\mathfrak{a}_s} D_{\mathfrak{b}_s} F(\mathcal{Q}_s) = D_{\mathfrak{a}_s \odot \mathfrak{b}_s} F(\mathcal{Q}_s),$$

where the vector $\mathfrak{a}_s \odot \mathfrak{b}_s$ is the Hadamard product of $\mathfrak{a}_s$ and $\mathfrak{b}_s$, which is also $\mathcal{Q}_s$-measurable. Hence a sum of terms of the form

$$D_{\mathfrak{a}_\mathfrak{s}} F(\mathcal{Q}_s) D_{\mathfrak{b}_\mathfrak{s}} + D_{\mathfrak{a}_{s+1}} F(\mathcal{Q}_{s+1}) D_{\mathfrak{b}_{s+1}} + \cdots + D_{\mathfrak{a}_{s+r}} F(\mathcal{Q}_{s+r}) D_{\mathfrak{b}_{s+r}},$$

with $\mathcal{R} = \mathcal{Q}_s = \cdots = \mathcal{Q}_{s+r}$, can be reduced to the sum of two terms as

$$D_C F(\mathcal{R}) + D_{\mathfrak{a}_{s+r}} F(\mathcal{R}) D_{\mathfrak{b}_{s+r}},$$

where $C = \sum_{h=0}^{r-1} \mathfrak{a}_{s+h} \odot \mathfrak{b}_{s+h}$ is $\mathcal{R}$-measurable. In this way the representation (5.1) can be written as

$$U = \sum_{s=0}^{k} D_{C_s} F(\mathcal{R}_s) + D_{\mathfrak{m}_s} F(\mathcal{R}_s) D_{\mathfrak{n}_s}, \tag{5.2}$$

where $\mathbb{F} = \{\mathcal{R}_0 \prec \mathcal{R}_1 \prec \cdots \prec \mathcal{R}_k\}$ is a filtration, $\mathcal{N} = \mathcal{R}_0$, $\mathcal{F} = \mathcal{R}_k$, C_s is $\mathcal{R}_s$-measurable, $\mathfrak{m}_s, \mathfrak{n}_s$ are $\mathcal{R}_{s+1}$-measurable, and $\mathfrak{m}_k = 0$ (again we assume that $\mathcal{R}_{k+1} = \mathcal{F}$). We shall always consider this reduced representation of (5.1), and we shall say that U is *filtered* with respect to the filtration $\mathbb{F}$.

If all $\mathfrak{m}_s, \mathfrak{n}_s$ are $\mathcal{R}_s$-measurable, then (5.1) reduces to the form

$$U = \sum_{s=0}^{k} D_{C_s + \mathfrak{m}_s \odot \mathfrak{n}_s} F(\mathcal{R}_s), \tag{5.3}$$

and U is a symmetric matrix.

We are mainly interested in a decomposition like (5.2) with the vectors $\mathfrak{m}_s, \mathfrak{n}_s$ having the following special structure:

$$\mathfrak{m}_s = \Gamma_s \odot p_s, \qquad \mathfrak{n}_s = q_s, \tag{5.4}$$

where Γ_s is $\mathcal{R}_s$-measurable and $\{p_s, q_s\}$ is an $\mathcal{R}_{s+1}$-measurable partition; that is, $\{p_s, q_s\}$ are $\mathcal{R}_{s+1}$-measurable $\{0,1\}$-valued vectors with disjoint support $p_s \odot q_s = 0$ and $p_s + q_s = 1$. If this is the case, U is said to be a special filtered matrix (SFM),

$$U = \sum_{s=0}^{k} D_{C_s}\ F(\mathcal{R}_s) + D_{\Gamma_s} D_{p_s}\ F(\mathcal{R}_s)\ D_{q_s}. \tag{5.5}$$

Notice that $\Gamma_k = 0$.

It is not difficult to see that every CBF matrix is filtered. This is done by induction. Assume that

$$U = \begin{pmatrix} A & \alpha \mathbf{1}_p \mathbf{1}'_{n-p} \\ \beta \mathbf{1}_{n-p} \mathbf{1}'_p & B \end{pmatrix}.$$

Define $\mathcal{R}_0 = \mathcal{N}$ and $\mathcal{R}_1 = \{\{1, \dots, p\}, \{p+1, \dots, n\}\}$. Take

$$C_0 = \alpha \mathbf{1}_n, \quad \Gamma_0 = (\beta - \alpha) \mathbf{1}_n, \quad p_0 = (\mathbf{0}_p, \mathbf{1}_{n-p})', \quad q_0 = (\mathbf{1}_p, \mathbf{0}_{n-p})';$$

then we obtain

$$D_{C_0} F(\mathcal{R}_0) + D_{\Gamma_0} D_{p_0} F(\mathcal{R}_0) D_{q_0} = \begin{pmatrix} \alpha \mathbf{1}_p \mathbf{1}'_p & \alpha \mathbf{1}_p \mathbf{1}'_{n-p} \\ \beta \mathbf{1}_{n-p} \mathbf{1}'_p & \alpha \mathbf{1}_{n-p} \mathbf{1}'_{n-p} \end{pmatrix}.$$

The key step is that $A - \alpha, B - \alpha$ are also in CBF. We have that C_0, Γ_0 are $\mathcal{R}_0$-measurable and p_0, q_0 is an $\mathcal{R}_1$-measurable partition. We also notice that if $0 \le \alpha \le \beta$, then $C_0 \ge 0, \Gamma_0 \ge 0$.

The induction also shows that U can be decomposed as in (5.5), where $\mathbb{F} = \{\mathcal{R}_0 \prec \cdots \prec \mathcal{R}_k\}$ is a dyadic filtration; C_s, Γ_s are $\mathcal{R}_s$-measurable; and $\{p_s, q_s\}$ is a $\mathcal{R}_{s+1}$-measurable partition.

We now summarize the representation form for the class of CBF, NBF, and GUM matrices.

Proposition 5.3. *V is a permutation of a CBF matrix if and only if there exists a dyadic filtration $\mathbb{F} = \{\mathcal{R}_0 \prec \cdots \prec \mathcal{R}_k\}$; a sequence of vectors $C_0, \dots, C_k$, $\Gamma_0, \dots, \Gamma_k$ verifying C_s, Γ_s are $\mathcal{R}_s$-measurable, and a sequence $\{p_s, q_s\}$ of $\mathcal{R}_{s+1}$-measurable partitions such that*

$$V = \sum_{s=0}^{k} D_{C_s} F(\mathcal{R}_s) + D_{\Gamma_s} D_{p_s} F(\mathcal{R}_s) D_{q_s}.$$

That is V is an SFM.

Also V is a permutation of an increasing CBF matrix if and only if there is a decomposition where Γ_0, C_s, Γ_s, $s = 1, \dots, k$, are nonnegative. Furthermore, V is a nonnegative matrix if and only if C_0 is nonnegative.

Moreover, V is a GUM if and only if C_s, Γ_s, $s = 0, \dots, k$, are nonnegative and for $s = 0, \dots, k-1$ it holds that

$$\Gamma_s \le C_{s+1} + \Gamma_{s+1}. \tag{5.6}$$

Finally, V is an ultrametric matrix if and only if there is a decomposition with $\Gamma_s = 0$ for all s.

Remark 5.1. We can assume without loss of generality that each p_s, q_s is obtained as follows. The nontrivial atoms $\mathcal{A}_1, \dots, \mathcal{A}_r$ of $\mathcal{R}_s$ are divided into the new atoms

$$\mathcal{A}_{1,1}, \mathcal{A}_{1,2}, \dots, \mathcal{A}_{r,1}, \mathcal{A}_{r,2}$$

of $\mathcal{R}_{s+1}$. Consider $\mathcal{B}_1, \dots, \mathcal{B}_r$ the set of trivial atoms in $\mathcal{R}_s$ (that is, the atoms which are singletons). Let q_s be the indicator of $\mathcal{A}_{1,1} \cup \cdots \cup \mathcal{A}_{r,1}$, p_s be the indicator of $\mathcal{A}_{1,2} \cup \cdots \cup \mathcal{A}_{r,2} \cup \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_r$, and $\Gamma_s = 0$ on the $\mathcal{R}_s$-measurable set $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_r$. We point out that the partition $\mathcal{R}_{s+1}$ is obtained from $\mathcal{R}_s$ refined by p_s. The following consistency relation,

$$D_{p_s} F(\mathcal{R}_s) p_s = D_{p_s} F(\mathcal{R}_{s+1}) \mathbf{1}, \tag{5.7}$$

will be used further in order to give sufficient treatable conditions for an SFM to be a bipotential.

Example 5.1. Consider the CBF matrix

$$U = \begin{pmatrix} a & \alpha_2 & \alpha_1 & \alpha_1 \\ \beta_2 & b & \alpha_1 & \alpha_1 \\ \beta_1 & \beta_1 & c & \hat{\alpha}_2 \\ \beta_1 & \beta_1 & \hat{\beta}_2 & d \end{pmatrix}.$$

U is an NBF matrix if the constraints $\alpha_1 \le \beta_1$, $\alpha_1 \le \min\{\alpha_2, \hat{\alpha}_2\}$, $\beta_1 \le \min\{\beta_2, \hat{\beta}_2\}$, $\alpha_2 \le \beta_2$, $\hat{\alpha}_2 \le \hat{\beta}_2$ are verified and finally the diagonal elements dominate on each row and column, that is, $\beta_2 \le \min\{a, b\}$, $\hat{\beta}_2 \le \min\{c, d\}$.

U is filtered with respect to the dyadic filtration $\mathcal{R}_0 = \{1,2,3,4\} \prec \mathcal{R}_1 = \{\{1,2\},\{3,4\}\} \prec \mathcal{R}_2 = \{\{1\},\{2\},\{3\},\{4\}\}$ and can be written as

$$U = D_{C_0} F(\mathcal{R}_0) + D_{\Gamma_0} D_{p_0} F(\mathcal{R}_0) D_{q_0} + D_{C_1} F(\mathcal{R}_1) + D_{\Gamma_1} D_{p_1} F(\mathcal{R}_1) D_{q_1} + D_{C_2} F(\mathcal{R}_2), \tag{5.8}$$

where

$$C_0 = \begin{pmatrix} \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \alpha_1 \end{pmatrix}, \quad \Gamma_0 = \begin{pmatrix} \beta_1 - \alpha_1 \\ \beta_1 - \alpha_1 \\ \beta_1 - \alpha_1 \\ \beta_1 - \alpha_1 \end{pmatrix}, \quad p_0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \quad q_0 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

$$C_1 = \begin{pmatrix} \alpha_2 - \alpha_1 \\ \alpha_2 - \alpha_1 \\ \hat{\alpha}_2 - \alpha_1 \\ \hat{\alpha}_2 - \alpha_1 \end{pmatrix}, \quad \Gamma_1 = \begin{pmatrix} \beta_2 - \alpha_2 \\ \beta_2 - \alpha_2 \\ \hat{\beta}_2 - \hat{\alpha}_2 \\ \hat{\beta}_2 - \hat{\alpha}_2 \end{pmatrix}, \quad p_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad q_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix},$$

and

$$C_2 = \begin{pmatrix} a - \alpha_2 \\ b - \alpha_2 \\ c - \hat{\alpha}_2 \\ d - \hat{\alpha}_2 \end{pmatrix}.$$

The decomposition in (5.8) is then

$$U = \begin{pmatrix} \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & \alpha_1 & \alpha_1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \beta_1 - \alpha_1 & \beta_1 - \alpha_1 & 0 & 0 \\ \beta_1 - \alpha_1 & \beta_1 - \alpha_1 & 0 & 0 \end{pmatrix}$$
$$+ \begin{pmatrix} \alpha_2 - \alpha_1 & \alpha_2 - \alpha_1 & 0 & 0 \\ \alpha_2 - \alpha_1 & \alpha_2 - \alpha_1 & 0 & 0 \\ 0 & 0 & \hat{\alpha}_2 - \alpha_1 & \hat{\alpha}_2 - \alpha_1 \\ 0 & 0 & \hat{\alpha}_2 - \alpha_1 & \hat{\alpha}_2 - \alpha_1 \end{pmatrix}$$
$$+ \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta_2 - \alpha_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\beta}_2 - \hat{\alpha}_2 & 0 \end{pmatrix} + \begin{pmatrix} a - \alpha_2 & 0 & 0 & 0 \\ 0 & b - \alpha_2 & 0 & 0 \\ 0 & 0 & c - \hat{\alpha}_2 & 0 \\ 0 & 0 & d - \hat{\alpha}_2 & 0 \end{pmatrix}.$$

The constraints are translated into the positivity of the vectors C and Γ and the ones induced by (5.6). We point out that we can also choose, for example, $\Gamma_1 = (0, \beta_2 - \alpha_2, 0, \hat{\beta}_2 - \hat{\alpha}_2)'$, but in this case Γ_1 is not $\mathcal{R}_1$-measurable. As we will see in subsection (5.1), this measurability condition will play an important role.

Example 5.2. Consider the nonnegative CBF matrix

$$U = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}.$$

This matrix is an SFM and can be decomposed as in (5.5). Nevertheless, none of these decompositions can have all its terms nonnegative. In particular, no permutation of U is an increasing CBF matrix.

Remark 5.2. Notice that the class of CBF matrices is stable under Hadamard functions. Nevertheless there are examples of filtered matrices for which $f(U)$ is not filtered. Consider the matrix

$$U = D_\alpha F_1 + D_{\mathfrak{a}} F_1 D_{\mathfrak{b}} + D_\beta F_2,$$

where $F_1 = F(\mathcal{N}) = \mathbf{1}\mathbf{1}'$ and $F_2 = \mathbb{I}$. The vector α is constant, and we confound it with the constant $\alpha \in \mathbb{R}$. The vectors $\mathfrak{a}$, $\mathfrak{b}, \beta$ are all $\mathcal{F}$-measurable. Then U is filtered and, moreover,

$$U = \alpha + \mathfrak{a}\mathfrak{b}' + D_\beta. \tag{5.9}$$

Take $\alpha = \beta = 0$, $\mathfrak{a} = (2,3,5,7)'$, and $\mathfrak{b} = (11,13,17,19)'$. Then all the entries of U are different. As f runs over all possible functions, $f(U)$ runs over all 4×4 matrices. This implies that some of them can not be written as in (5.9), because in this representation we have at most 13 free variables. Still is possible that each $f(U)$ is decomposable as in (5.1), using maybe a different filtration. A more detailed analysis shows that this is not the case. For example, if we choose the filtration $\mathcal{N} \prec \{\{1,2\},\{3,4\}\} \prec \mathcal{F}$, then every matrix V filtered with respect to this filtration verifies that $V_{13} = V_{23} = V_{14} = V_{24}$.

Matrices of the type $F(\mathcal{R})$ are related to conditional expectations (in probability theory). Indeed, let $\mathcal{R} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_r\}$ and $n_\ell = \#(\mathcal{A}_\ell)$ be the size of each atom. It is direct that $w = w_{\mathcal{R}} = F(\mathcal{R})\mathbf{1}$ is an $\mathcal{R}$-measurable vector that verifies $w_i = n_\ell$ for $i \in \mathcal{A}_\ell$. Then

$$\mathbb{E}_{\mathcal{R}} = D_w^{-1}\, F(\mathcal{R}) = F(\mathcal{R})\, D_w^{-1}$$

is the matrix of conditional expectation with respect to the σ-algebra generated by $\mathcal{R}$. This matrix $\mathbb{E} = \mathbb{E}_{\mathcal{R}}$ satisfies

$$\begin{aligned}&\mathbb{E}\mathbb{E} = \mathbb{E},\ \mathbb{E}' = \mathbb{E},\ \mathbb{E}\mathbf{1} = \mathbf{1};\\ &\forall v,\ \mathbb{E}v \text{ is } \mathcal{R}\text{-measurable};\\ &\text{ if } v \text{ is } \mathcal{R}\text{-measurable, then } \mathbb{E}v = v.\end{aligned}$$

Therefore, $\mathbb{E}$ is the orthogonal projection over the subspace of all $\mathcal{R}$-measurable vectors. In the case of the trivial partition $\mathcal{N}$, one gets $\mathbb{E}_{\mathcal{N}} = \frac{1}{n}\mathbf{1}\mathbf{1}'$ as the mean operator.

Remark 5.3. The L^2 space associated with $\{1,\dots,n\}$ endowed with the counting measure is identified with $\mathbb{R}^n$ with the standard Euclidean scalar product. In this way each vector of $\mathbb{R}^n$ can be seen as a function in L^2, and $\mathbb{E}$ is an orthogonal projection. The product $D_v\mathbb{E}$ (as matrices) is the product of the operators D_v and $\mathbb{E}$, where D_v is the multiplication by the function v. Notice that $\mathbb{E}D_v$ and $\mathbb{E}(v)$ are quite different. The former is an operator (a matrix), and the latter is a function (vector). They are related by $\mathbb{E}(v) = \mathbb{E}D_v(\mathbf{1})$, where $\mathbf{1}$ is the constant function.

Let $\mathcal{R}$, $\mathcal{Q}$ be two partitions; then $\mathcal{R} \preccurlyeq \mathcal{Q}$ is equivalent to $\mathbb{E}_{\mathcal{R}}\mathbb{E}_{\mathcal{Q}} = \mathbb{E}_{\mathcal{Q}}\mathbb{E}_{\mathcal{R}} = \mathbb{E}_{\mathcal{R}}$. This commutation relation can be written as a commutation relation for $F(\mathcal{R})$ and $F(\mathcal{Q})$. In fact,

$$\begin{aligned}F(\mathcal{R})F(\mathcal{Q}) &= \mathbb{E}_{\mathcal{R}}D_{w_{\mathcal{R}}}\mathbb{E}_{\mathcal{Q}}D_{w_{\mathcal{Q}}} = \mathbb{E}_{\mathcal{R}}\mathbb{E}_{\mathcal{Q}}D_{w_{\mathcal{R}}}D_{w_{\mathcal{Q}}}\\ &= \mathbb{E}_{\mathcal{R}}D_{w_{\mathcal{R}}}D_{w_{\mathcal{Q}}} = F(\mathcal{R})D_{w_{\mathcal{Q}}},\end{aligned}$$

$$F(\mathcal{Q})F(\mathcal{R}) = (F(\mathcal{R})F(\mathcal{Q}))' = D_{w_{\mathcal{Q}}}F(\mathcal{R}).$$

5.1. An algorithm for filtered matrices: Conditions to be in *biP*. In this section we introduce a backward algorithm that gives a sufficient condition for a filtered matrix to be in class *bi*$\mathcal{P}$. For that purpose assume that U has a representation as in (5.1):

$$U = \sum_{s=0}^{\ell} D_{\mathfrak{a}_s} F(\mathcal{Q}_s) D_{\mathfrak{b}_s},$$

where we assume further that $\mathfrak{a}_s, \mathfrak{b}_s$ are all nonnegative. In particular, U is a nonnegative matrix.

We introduce the conditional expectations $\mathbb{E}_s = \mathbb{E}_{\mathcal{Q}_s} = D^{-1}_{F(\mathcal{Q}_s)\mathbf{1}}\, F(\mathcal{Q}_s)$ and the normalized factors $a_s = \mathfrak{a}_s \odot F(\mathcal{Q}_s)\mathbf{1},\ b_s = \mathfrak{b}_s$. Then U can be written as

$$U = \sum_{s=0}^{\ell} D_{a_s}\mathbb{E}_s D_{b_s} = \sum_{s=0}^{\ell} a_s\mathbb{E}_s b_s, \tag{5.10}$$

where we have identified vectors (functions) and the operator of multiplication they induce. We shall use this notation throughout this section. Finally, we recall that $\mathbb{E}_\ell = \mathbb{I}$.

We can now use the algorithm developed in [8] to study the inverse of $\mathbb{I}+U$. In what follows, we take the convention $0\cdot\infty = 0/0 = 0$. This algorithm is defined by the backward recursion starting with the values $\lambda_\ell = \mu_\ell = \kappa_\ell = 1,\ \sigma_\ell = (1+a_\ell b_\ell)^{-1}$ and for $s = \ell-1,\ldots,0$,

$$\begin{aligned}
\lambda_s &= \lambda_{s+1}[1-\sigma_{s+1}a_{s+1}\mathbb{E}_{s+1}(\kappa_{s+1}b_{s+1})],\\
\mu_s &= \mu_{s+1}[1-\sigma_{s+1}b_{s+1}\mathbb{E}_{s+1}(\kappa_{s+1}a_{s+1})],\\
\kappa_s &= \mathbb{E}_{s+1}(\lambda_s) = \mathbb{E}_{s+1}(\mu_s),\\
\sigma_s &= (1+\mathbb{E}_s(\kappa_s a_s b_s))^{-1}.
\end{aligned} \tag{5.11}$$

We get the recursion

$$\kappa_{s-1} = \mathbb{E}_s(\kappa_s) - \frac{\mathbb{E}_s(\kappa_s a_s)\mathbb{E}_s(\kappa_s b_s)}{1+\mathbb{E}_s(\kappa_s a_s b_s)}. \tag{5.12}$$

The algorithm continues until some λ or μ is negative; otherwise we arrive at $s=0$. If this is the case, then $\mathbb{I}+U$ is nonsingular and its inverse is of the form $\mathbb{I}-N$, where

$$N = \sum_{s=0}^{\ell} \sigma_s\lambda_s a_s\mathbb{E}_s b_s\mu_s.$$

We also have that

$$\lambda_{-1} = (\mathbb{I}-N)\mathbf{1} \quad\text{and}\quad \mu_{-1} = (\mathbb{I}-N)'\mathbf{1},$$

where λ_{-1},μ_{-1} are obtained from the first two formulae in (5.11) for $s=-1$. Therefore, if they are also nonnegative, the matrix $\mathbb{I}+U$ is a *bi*$\mathcal{P}$-matrix.

In this way we have that a sufficient condition for $\mathbb{I}+U$ to be a *bi*$\mathcal{P}$-matrix is that the algorithm works for $s=\ell,\ldots,0$ and that all the λ,μ are nonnegative, including λ_{-1},μ_{-1}. In this situation we have that λ (and μ) is a decreasing nonnegative sequence of vectors. Sufficient treatable conditions on the coefficients of the expansion (5.10) involve the recurrence (5.12). Starting from $\kappa_\ell = 1$, we assume that this recurrence has a solution such that $\kappa_s \in [0,1]$ for all $s=\ell,\ldots,-1$. We shall study closely this recursion for the class of SFM, and we shall obtain sufficient conditions to have $\mathbb{I}+U$ in *bi*$\mathcal{P}$.

Before studying this problem, we further discuss the algorithm. We have the following relations:

$$\left(\mathbb{I}+\sum_{k=s}^{\ell} a_k\mathbb{E}_k b_k\right)^{-1} = \mathbb{I} - \sum_{k=s}^{\ell}\sigma_k\lambda_k a_k\mathbb{E}_k b_k\mu_k = \mathbb{I}-N_s,$$

$$\lambda_{s-1} = (\mathbb{I}-N_s)\mathbf{1}, \qquad \mu_{s-1} = (\mathbb{I}-N_s)'\mathbf{1}.$$

That is, our condition is to impose that all the matrices

$$\mathbb{I} + a_\ell \mathbb{E}_\ell b_\ell, \dots, \mathbb{I} + \sum_{k=s}^{\ell} a_k \mathbb{E}_k b_k, \dots, \mathbb{I} + \sum_{k=0}^{\ell} a_k \mathbb{E}_k b_k = \mathbb{I} + U$$

are in class $bi\mathcal{P}$.

We now assume that U is an SFM with a decomposition like

$$U = \sum_{s=0}^{k} D_{C_s}\ F(\mathcal{R}_s) + D_{\Gamma_s} D_{p_s}\ F(\mathcal{R}_s)\ D_{q_s},$$

where $\mathbb{F} = \mathcal{R}_0 \prec \cdots \prec \mathcal{R}_k$ is a filtration; C_s, Γ_s are nonnegative $\mathcal{R}_s$-measurable; and $\{p_s, q_s\}$ is a $\mathcal{R}_{s+1}$-measurable partition. Again we set $\mathbb{E}_s = D^{-1}_{F(\mathcal{R}_s)\mathbf{1}}\ F(\mathcal{R}_s)$ and the normalized $\mathcal{R}_s$-measurable factors

$$c_s = C_s \odot F(\mathcal{R}_s)\mathbf{1}, \qquad \gamma_s = \Gamma_s \odot F(\mathcal{R}_s)\mathbf{1}.$$

Since diagonal matrices commute, we get that U has a representation of the form

$$U = \sum_{s=0}^{k} c_s \mathbb{E}_s + \gamma_s p_s \mathbb{E}_s q_s,$$

with $\gamma_k = 0$. In the previous algorithm we can make two steps at each time and consider κ_s in place of κ_{2s}, λ_s instead of λ_{2s+1}, l_s instead of λ_{2s}. We also introduce $d_s = 1/\kappa_s$ to simplify certain formulae (this vector can take the value ∞). We get, starting from $\kappa_k = l_k = 1, \sigma_k = (1 + c_k)^{-1}$, that for $s = k-1, \dots, 0$

$$\begin{aligned} \lambda_s &= \sigma_{s+1} l_{s+1}, \\ l_s &= \lambda_s [1 - \gamma_s p_s \mathbb{E}_s (q_s/(c_{s+1} + d_{s+1}))], \\ \kappa_s &= \mathbb{E}_s(l_s), \\ \sigma_s &= 1/(1 + \kappa_s c_s) = d_s/(c_s + d_s). \end{aligned}$$

Similar recursions hold for μ, m, which are the analogues of λ, l. Relation (5.12) takes the form

$$\frac{1}{d_s} = \mathbb{E}_s\left(\frac{1}{c_{s+1} + d_{s+1}}\right) - \gamma_s\, \mathbb{E}_s\left(\frac{p_s}{c_{s+1} + d_{s+1}}\right) \mathbb{E}_s\left(\frac{q_s}{c_{s+1} + d_{s+1}}\right). \tag{5.13}$$

The inverse of $\mathbb{I} + U$ is $\mathbb{I} - N$, where

$$N = \sum_{s=0}^{k} c_s \sigma_s l_s \mathbb{E}_s m_s + \sum_{s=0}^{k-1} \gamma_s \lambda_s p_s \mathbb{E}_s q_s \mu_s = \sum_{s=0}^{k} c_s \sigma_s l_s \mathbb{E}_s m_s + \gamma_s \lambda_s p_s \mathbb{E}_s q_s \mu_s. \tag{5.14}$$

Again $\lambda_{-1} = (\mathbb{I} - N)\mathbf{1} = \sigma_0 l_0$, and similarly $\mu_{-1} = \sigma_0 m_0$.

Let us introduce the following function:

$$\rho_s = \mathbb{E}_s(p_s) p_s + \mathbb{E}_s(q_s) q_s.$$

THEOREM 5.4. *Assume that the backward recursion* (5.13) *has a nonnegative solution starting with* $d_k = 1$. *Assume, moreover, that this solution verifies for* $s = k-1, \dots, 0$

$$\rho_s \gamma_s \le c_{s+1} + d_{s+1}. \tag{5.15}$$

Then $\lambda_s, l_s, \mu_s, m_s, \sigma_s$, *for* $s = k, \dots, 0$, *as well as* λ_{-1}, μ_{-1} *are well defined and nonnegative. Therefore,* $\mathbb{I} + U \in bi\mathcal{P}$, *and its inverse is* $\mathbb{I} - N$, *where* N *is given by* (5.14).

The proof of this result is based on the following lemma.

LEMMA 5.5. *Let* x, y *be nonnegative vectors, and* $\mathbb{E}$ *be a conditional expectation. If* $x\mathbb{E}(y) \le 1$, *then* $\mathbb{E}(xy) \le 1$.

Proof. We first assume that y is strictly positive. Since $x \le 1/\mathbb{E}(y)$ and $\mathbb{E}$ is an increasing operator, we have

$$\mathbb{E}(xy) \le \mathbb{E}\left(\frac{1}{\mathbb{E}(y)} y\right) = \frac{\mathbb{E}(y)}{\mathbb{E}(y)} = 1.$$

For the general case consider $(y + \epsilon \mathbf{1})/(1 + \epsilon |x|_\infty)$ instead of y and pass to the limit $\epsilon \to 0$. □

Proof of Theorem 5.4. We notice that condition (5.15) implies that

$$\frac{q_s \gamma_s}{c_{s+1} + d_{s+1}} \mathbb{E}_s(q_s) \le 1.$$

Since γ_s is $\mathbb{E}_s$-measurable and $q_s = q_s^2$, we obtain

$$\gamma_s \mathbb{E}_s\left(\frac{q_s}{c_{s+1} + d_{s+1}}\right) = \mathbb{E}_s\left(\frac{\gamma_s q_s^2}{c_{s+1} + d_{s+1}}\right).$$

This last quantity is bounded by one by Lemma 5.5. Similarly we have

$$\gamma_s \mathbb{E}_s\left(\frac{p_s}{c_{s+1} + d_{s+1}}\right) \le 1,$$

which implies that the algorithm is not stopped, and all the coefficients are nonnegative including λ_{-1}, μ_{-1}. □

COROLLARY 5.6. *Assume that for* $s = k-1, \dots, 0$ *we have*

$$\rho_s \gamma_s \le c_{s+1} + \gamma_{s+1}. \tag{5.16}$$

Then the recursion (5.13) *has a nonnegative solution that verifies* (5.15). *In particular,* $\mathbb{I} + tU$ *is in class* $bi\mathcal{P}$ *for all* $t \ge 0$, *and* U *is in* $bi\mathcal{P}$ *if it is nonsingular.*

Proof. Let us consider first the case $t = 1$. We prove by induction that $\gamma_s \le d_s$. For $s = k$ we have $0 = \gamma_k \le d_k = 1$. We point out that if we multiply in (5.13) by γ_s, we get

$$\frac{\gamma_s}{d_s} = \mathbb{E}_s\left(\frac{\gamma_s}{c_{s+1} + d_{s+1}}\right) - \mathbb{E}_s\left(\frac{\gamma_s p_s}{c_{s+1} + d_{s+1}}\right) \mathbb{E}_s\left(\frac{\gamma_s q_s}{c_{s+1} + d_{s+1}}\right),$$

which is of the form $x + y - xy$, where $x = \mathbb{E}_s\left(\frac{\gamma_s p_s}{c_{s+1}+d_{s+1}}\right)$. The inequality (5.16), the induction hypothesis $\gamma_{s+1} \le d_{s+1}$, and Lemma 5.5 imply $0 \le x \le 1,\ 0 \le y \le 1$. In particular,

$$0 \le \frac{\gamma_s}{d_s} \le 1,$$

and the induction is completed. Theorem 5.4 shows that $\mathbb{I} + U$ is in class $bi\mathcal{P}$. We notice that tU also verifies condition (5.16) because this condition is homogeneous, and the result follows. □

Remark 5.4. We notice that condition (5.16) can be expressed in terms of the original coefficients C, Γ in the dyadic case. In fact (see (5.7)),

$$p_s \mathbb{E}_s(p_s) = D_{p_s} D^{-1}_{F(\mathcal{R}_s)\mathbf{1}}\, F(\mathcal{R}_s) p_s = D_{p_s} D^{-1}_{F(\mathcal{R}_s)\mathbf{1}} F(\mathcal{R}_{s+1})\mathbf{1},$$

which implies that

$$\rho_s = (1/F(\mathcal{R}_s)\mathbf{1}) \odot (F(\mathcal{R}_{s+1})\mathbf{1}).$$

Then, inequality (5.16) is

$$\Gamma_s \le C_{s+1} + \Gamma_{s+1},$$

which is the condition for having a GUM (see (5.6)) . We mention here that condition (5.16) is more general than having a GUM, as the following example shows.

Remark 5.5. Consider the matrix U_β,

$$U_\beta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \beta & \beta & 1 & 0 \\ \beta & \beta & 0 & 1 \end{pmatrix} = D_{\Gamma_0} D_{p_0} F(\mathcal{R}_0) D_{q_0} + \mathbb{I},$$

where $\mathcal{R}_0 = \mathcal{N}$, $\Gamma_0 = \beta(1,1,1,1)' \le C_1 = (1,1,1,1)'$. We compute $c_0 = 0$, $\gamma_0 = 4\beta$, $c_1 = C_1$, $\gamma_1 = 0$ and also $\rho_0 = 1/2$.

It is direct to check that $U_\beta^{-1} = U_{-\beta}$. Then for all $\beta \ge 0$ the matrix $U_\beta \in \mathcal{M}^{-1}$. Also $U_\beta \in bi\mathcal{P}$ if and only if $0 \le \beta \le 1/2$. When $\beta \ge 0$ the condition (5.6), $\Gamma_0 \le C_1 + \Gamma_1$, is equivalent to $\beta \le 1$. Then, this condition does not ensure that $U \in bi\mathcal{P}$ (this happens because the filtration is not dyadic). Nevertheless, the analogous condition in terms of the normalized factors (5.16),

$$\rho_0 \gamma_0 \le c_1 + \gamma_1,$$

is equivalent to $\beta \le 1/2$, which is the correct condition.

Corollary 5.7. *Assume that*

$$\rho_s \gamma_s \le \sum_{r=s+1}^{k} c_r \tag{5.17}$$

hold for $s = k-1, \ldots, 0$. *Then the recursion* (5.13) *has a nonnegative solution that verifies* (5.15). *In particular,* $\mathbb{I} + tU$ *is in class* $bi\mathcal{P}$ *for all* $t \ge 0$, *and* U *is in* $bi\mathcal{P}$ *if it is nonsingular.*

Proof. Consider the set of inequalities

$$\rho_s \gamma_s \vee \xi_s \le c_{s+1} + \xi_{s+1},$$

for $s = k-1, \ldots, 0$. A nonnegative solution is given by

$$\xi_s = \sup\left\{0,\ \gamma_0\rho_0 - \sum_{r=1}^{s} c_r, \ldots, \gamma_k\rho_k - \sum_{r=k+1}^{s} c_r, \ldots, \gamma_{s-1}\rho_{s-1} - c_s\right\}.$$

The hypothesis of the corollary is that $\xi_k = 0$. We also notice that ξ_s is $\mathcal{R}_s$-measurable.

We show, using a backward recursion, that $\xi_s \le d_s$. Indeed, by construction, $1/\xi_s = \mathbb{E}_s(1/\xi_s) \ge (c_{s+1} + \xi_{s+1})^{-1}$ while $1/d_s \le \mathbb{E}_s((c_{s+1} + d_{s+1})^{-1})$. Then the inequality $\rho_s\gamma_s \le c_{s+1} + \xi_{s+1}$ implies $\rho_s\gamma_s \le c_{s+1} + d_{s+1}$, so the result holds (see Theorem 5.4). □

5.2. Conditions for class $\mathcal{T}$ and proof of Theorem 2.6.

THEOREM 5.8. *Assume that* U *has a decomposition*

$$U = \sum_{s=0}^{\ell} a_s \, \mathbb{E}_s \, b_s,$$

where a_s, b_s *are nonnegative* $\mathbb{E}_{s+1}$*-measurable. Then* U *belongs to the class* $\mathcal{T}$ *and, moreover,*

$$\tau(U) = \inf\{t > 0 : \ (\mathbb{I} + tU)^{-1}\mathbf{1} \not\geq 0 \ \text{or} \ \mathbf{1}'(\mathbb{I} + tU)^{-1} \not\geq 0\}.$$

In particular, if $\tau(U) < \infty$*, then* $\mathbb{I} + \tau(U)\, U \in bi\mathcal{P}$*.*

Remark 5.6. In the case $\tau(U) < \infty$ we have that $\mathbb{I}+t\,U$ is nonsingular for $t > \tau(U)$ sufficiently close to $\tau(U)$. This follows from the fact that the set of nonsingular matrices is open.

Theorem 5.8 states that every filtered matrix with a nonnegative decomposition is in class $\mathcal{T}$, which proves Theorem 2.6.

Proof of Theorem 5.8. A warning about the use of vectors and functions. Here we consider vectors or functions on $\{1, \dots, n\}$ indiscriminately. Thus for two vectors a, b the product ab makes sense as the product of two functions, which corresponds to the Hadamard product of the vectors. Also an expression as $(1 + ab)^{-1}$ is the vector whose components are the reciprocals of the components of $1 + ab$. We also recall that $(a)_i$ is the ith component of a.

Now, for $p = 0, \dots, \ell$ consider the matrices

$$U(p) = \sum_{s=p}^{\ell} a_s \, \mathbb{E}_s \, b_s.$$

We notice that $U(0) = U$. We shall prove that $\tau_p = \tau(U(p))$ is increasing in p and $\tau_\ell = \infty$.

We rewrite the algorithm for $\mathbb{I} + tU$. This takes the form $\lambda_\ell(t) = \mu_\ell(t) = \kappa_\ell(t) = 1$, $\sigma_\ell(t) = (1 + t\, a_\ell b_\ell)^{-1}$, and for $p = \ell - 1, \dots, 0$

$$\begin{aligned}
&\lambda_p(t) = \lambda_{p+1}(t)[1 - \sigma_{p+1}(t)\, t\, a_{p+1}\mathbb{E}_{p+1}(\kappa_{p+1}(t) b_{p+1})],\\
&\mu_p(t) = \mu_{p+1}(t)[1 - \sigma_{p+1}(t)\, t\, b_{p+1}\mathbb{E}_{p+1}(\kappa_{p+1}(t) a_{p+1})],\\
&\kappa_p(t) = \mathbb{E}_{p+1}(\lambda_p(t)) = \mathbb{E}_{p+1}(\mu_p(t)),\\
&\sigma_p(t) = (1 + \mathbb{E}_p(\kappa_p(t) t a_p b_p))^{-1}.
\end{aligned} \tag{5.18}$$

Also $\lambda_{-1}(t)$, $\mu_{-1}(t)$ are defined similarly. If $\lambda_s(t)$, $\mu_s(t)$, $\sigma_s(t)$, $s = \ell, \dots, p$, are well defined, then

$$(\mathbb{I} + tU(p))^{-1} = \mathbb{I} - N(p, t),$$

where

$$N(p, t) = \sum_{s=p}^{\ell} \sigma_s(t) \lambda_s(t)\, t\, a_s \mathbb{E}_s b_s \mu_s(t). \tag{5.19}$$

If $\lambda_s(t),\ \mu_s(t),\ \sigma_s(t),\ s=\ell,\dots,p$, are nonnegative, then $N(p,t)\geq 0$ and $(\mathbb{I}+tU(p))\in\mathcal{M}^{-1}$. Moreover, $\lambda_{p-1}(t)$ and $\mu_{p-1}(t)$ are the right and left equilibrium potentials of $(\mathbb{I}+tU(p))$,

$$(\mathbb{I}+tU(p))\lambda_{p-1}(t)=\mathbf{1}\quad\text{ and }\quad \mu'_{p-1}(t)(\mathbb{I}+tU(p))=\mathbf{1}'.$$

So, if they are nonnegative, we have $\mathbb{I}+tU(p)\in bi\mathcal{P}$. In particular, for $p=\ell$ we get

$$(\mathbb{I}+ta_\ell\,\mathbb{E}_\ell\,b_\ell)^{-1}=(\mathbb{I}+tU(\ell))^{-1}=\mathbb{I}-t(1+t\,a_\ell b_\ell)^{-1}a_\ell\,\mathbb{E}_\ell\,b_\ell.$$

Since $\mathbb{E}_\ell=\mathbb{I}$ we obtain that $\lambda_{\ell-1}=\mu_{\ell-1}=(1+t\,a_\ell b_\ell)^{-1}$. This means that $\mathbb{I}+tU(\ell)\in bi\mathcal{P}$ for all $t\geq 0$. Therefore $\tau_\ell=\infty$, and the result is true for $U(\ell)$. In particular, $\tau_{\ell-1}\leq\tau_\ell$. We shall prove by induction that

- $\tau_{p+1}\leq\cdots\leq\tau_\ell$

and for $q=p+1,\dots,\ell$

- $\tau_q=\inf\{t>0:\lambda_{q-1}(t)\ngeq 0\text{ or }\mu_{q-1}(t)\ngeq 0\}=\inf\{t>0:\lambda_{q-1}(t)\not>0\text{ or }\mu_{q-1}(t)\not>0\}$;
- $\lambda_s(t),\mu_s(t)$, for $s=\ell,\dots,q-1$, are strictly positive for $t\in[0,\tau_q)$;
- if $\tau_q<\infty$, we have $\mathbb{I}+\tau_qU(q)\in bi\mathcal{P}$.

The case $\tau_{p+1}=\infty$ is simple. Indeed, fix $t\geq 0$. From Lemma 3.1, $\mathbb{I}+tU(p+1)\in bi\mathcal{P}$ and its equilibrium potential are strictly positive; that is, $\lambda_p(t)>0,\ \mu_p(t)>0$. Thus, $\mathbb{I}+tU(p)$ is nonsingular; its inverse is $\mathbb{I}-N(p,t)$, where $N(p,t)\geq 0$ is given by (5.19). Hence, $\mathbb{I}+tU(p)\in\mathcal{M}^{-1}$. We conclude that

$$\tau_p=\inf\{t>0:\mathbb{I}+tU(p)\notin bi\mathcal{P}\}=\inf\{t>0:\lambda_{p-1}(t)\ngeq 0\text{ or }\mu_{p-1}(t)\ngeq 0\}.$$

If $\tau_p=\infty$, Lemma 3.1 gives

$$\lambda_{p-1}(t)>0,\qquad \mu_{p-1}(t)>0,$$

and the induction step holds in this case.

Now if $\tau_p<\infty$, by continuity we have $\mathbb{I}+\tau_pU(p)\in bi\mathcal{P}$. We shall prove later on that $\lambda_{p-1}(t),\mu_{p-1}(t)$ are strictly positive in $[0,\tau_p)$.

We now analyze the case $\tau_{p+1}<\infty$. We first notice that in the algorithm the only possible problem could arise with the definition of $\sigma_p(t)$. Since $\sigma_p(\tau_{p+1})>0$, the algorithm is well defined, by continuity, for steps $\ell,\dots,p$ on an interval $[0,\tau_{p+1}+\epsilon]$ for $\epsilon>0$ small enough. This proves that the matrix $\mathbb{I}+tU(p)$ is nonsingular in that interval, and that $\lambda_{p-1},\ \mu_{p-1}$ exist in the same interval.

Now, for a sequence $t_n\downarrow\tau_{p+1}$, either $\lambda_p(t_n)$ or $\mu_p(t_n)$ has a negative component. Since there are a finite number of components, we can assume without loss of generality that for a fixed component i we have $(\lambda_p(t_n))_i<0$. Then, by continuity we get that $(\lambda_p(\tau_{p+1}))_i=0$, which implies (by the algorithm) that $(\lambda_{p-1}(\tau_{p+1}))_i=0$.

Assume now that for some $t>\tau_{p+1}$ the matrix $\mathbb{I}+tU(p)\in bi\mathcal{P}$. By Lemma 3.1 we will have that $\mathbb{I}+\tau_{p+1}U(p)\in bi\mathcal{P}$, but its equilibrium potential will satisfy $\lambda_{p-1}(\tau_{p+1})>0$, which is a contradiction. Therefore we conclude that $\tau_p\leq\tau_{p+1}$.

The conclusion of this discussion is that the matrix $\mathbb{I}+tU(p)$, for $t\in[0,\tau_{p+1}]$, is nonsingular and its inverse is $\mathbb{I}-N(p,t)$, with $N(p,t)\geq 0$. That is, $\mathbb{I}+tU(p)\in\mathcal{M}^{-1}$ and therefore

$$\tau_p=\inf\{t>0:\mathbb{I}+tU(p)\notin bi\mathcal{P}\}=\inf\{t>0:\lambda_{p-1}(t)\ngeq 0\text{ or }\mu_{p-1}(t)\ngeq 0\},$$

and by continuity $\mathbb{I}+\tau_pU(p)\in bi\mathcal{P}$.

To finish the proof we need to show that τ_p coincides with

$$S = \inf\{t > 0 : \lambda_{p-1}(t) \not> 0 \text{ or } \mu_{p-1}(t) \not> 0\}.$$

It is clear that $S \leq \tau_p$. If $S < \tau_p$, then, due to Lemma 3.1, we have that both $\lambda_{p-1}(S) > 0$ and $\mu_{p-1}(S) > 0$, which is a contradiction, and then $S = \tau_p$. This shows that $\lambda_{p-1}(t)$, $\mu_{p-1}(t)$ are strictly positive for $t \in [0, \tau_p)$, and the induction is proven. □

Remark 5.7. It is possible to prove that $\kappa_p(\tau_p) > 0$ when $\tau_p < \infty$, but this is not central to our discussion.

REFERENCES

[1] T. Ando, *Inequalities for M-matrices*, Linear and Multilinear Algebra, 8 (1980), pp. 291–316.

[2] R. B. Bapat, M. Catral, and M. Neumnann, *On functions that preserve M-matrices and inverse M-matrices*, Linear and Multilinear Algebra, 53 (2005), pp. 193–201.

[3] N. Bouleau, *Autour de la variance comme forme de Dirichlet*, in Séminaire de Théorie du Potentiel 8, Lecture Notes in Math. 1235, Springer, New York, 1989, pp. 39–53.

[4] S. Chen, *A property concerning the Hadamard powers of inverse M-matrices*, Linear Algebra Appl., 381 (2004), pp. 53–60.

[5] P. Dartnell, S. Martínez, and J. San Martín, *Opérateurs filtrés et chaînes de tribus invariantes sur un espace probabilisé dénombrable*, in Séminaire de Probabilités XXII, Lecture Notes in Math. 1321, Springer-Verlag, New York, Berlin, 1988, pp. 197–213.

[6] C. Dellacherie, S. Martínez, and J. San Martín, *Ultrametric matrices and induced Markov chains*, Adv. Appl. Math., 17 (1996), pp. 169–183.

[7] 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.

[8] C. Dellacherie, S. Martínez, J. San Martín, and D. Taïbi, *Noyaux potentiels associés à une filtration*, Ann. Inst. H. Poincaré Probab. Statist., 34 (1998), pp. 707–725.

[9] M. Fiedler and H. Schneider, *Analytic functions of M-matrices and generalizations*, Linear and Multilinear Algebra, 13 (1983), pp. 185–201.

[10] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, London, 1991.

[11] S. Martínez, G. Michon, and J. San Martín, *Inverses of strictly ultrametric matrices are of Stieltjes types*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 98–106.

[12] 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.

[13] C. A. Micchelli and R. A. Willoughby, *On functions which preserve Stieltjes matrices*, Linear Algebra Appl., 23 (1979), pp. 141–156.

[14] R. Nabben and R. S. Varga, *A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieljes matrix*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 107–113.

[15] R. Nabben and R. S. Varga, *Generalized ultrametric matrices—A class of inverse M-matrices*, Linear Algebra Appl., 220 (1995), pp. 365–390.

[16] M. Neummann, *A conjecture concerning the Hadamard product of inverses of M-matrices*, Linear Algebra Appl., 285 (1998), pp. 277–290.

[17] R. S. Varga, *Nonnegatively posed problems and completely monotonic functions*, Linear Algebra Appl., 1 (1968), pp. 329–347.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 316–330

FAST SIMULTANEOUS ORTHOGONAL REDUCTION TO TRIANGULAR MATRICES[*]

I. V. OSELEDETS[†], D. V. SAVOSTYANOV[†], AND E. E. TYRTYSHNIKOV[†]

Abstract. A new algorithm is presented for simultaneous reduction of a given finite sequence of square matrices to upper triangular matrices by means of orthogonal transformations. The reduction is performed through a series of deflation steps, where each step contains a *simultaneous eigenvalue problem* being a direct generalization of the generalized eigenvalue problem. To solve the latter, a fast variant of the Gauss–Newton algorithm is proposed with some results on its local convergence properties (quadratic for the exact and linear for the approximate reduction) and numerical examples are provided.

Key words. simultaneous reduction of matrices, fast algorithms, convergence estimates

AMS subject classifications. 65F15, 15A69

DOI. 10.1137/060650738

1. Introduction.

Numerical algorithms for simultaneous reduction of several matrices to a specific structured form (like triangular or diagonal) by means of one and the same transformation applied to all given matrices is one of the most challenging problems in matrix analysis [1, 14, 9, 8]. The transformations of interest can be similarity, congruence, or equivalence with possible additional constraints. The problem we are going to tackle is the following problem of *approximate simultaneous reduction to triangular matrices.*

Problem. Given $n \times n$ real matrices $A_1, \dots, A_r$, find orthogonal $n \times n$ matrices Q and Z such that matrices

$$B_k = QA_kZ$$

are as upper triangular as possible.

This problem arises, for example, during the computation of the canonical decomposition in tensor algebra; see [9] and references therein. This problem is a key ingredient in many applications like signal processing [5, 12]. The simultaneous reduction of several matrices to upper triangular form is to some extent related to computing the product Schur form of a set of matrices, considered in [6].

In the case of two matrices ($r = 2$) such a transformation is well-studied, can be constructed explicitly, and often is referred to as the *generalized Schur decomposition.* It justifies the name of *simultaneous generalized Schur decomposition* (SGSD) used for those decompositions in the case of $r > 2$ [9]. It is clear that arbitrary matrices A_k may not admit such a reduction with good accuracy, so we impose on A_k the following *existence assumption*: Real matrices A_k are such that orthogonal matrices

[*] Received by the editors January 24, 2006; accepted for publication (in revised form) by B. T. Kågström December 4, 2008; published electronically March 25, 2009. This research was supported by the Russian Fund of Basic Research (grants 05-01-00721, 06-01-08052, and 08-01-00115) and a Priority Research grant OMN-3 of the Department of Mathematical Sciences of the Russian Academy of Sciences.

http://www.siam.org/journals/simax/31-2/65073.html

[†] Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina Street, 8, 119991 Moscow, Russia (ivan@bach.inm.ras.ru, draug@bach.inm.ras.ru, tee@bach.inm.ras.ru). The research of the first and second authors was supported by the Human Capital Foundation grant.

Q and Z exist complying with the equations

$$QA_kZ = T_k + E_k,$$

where T_k are upper triangular and the residue matrices E_k satisfy

$$\left(\sum_{k=1}^{r} ||E_k||^2\right)^{1/2} = \varepsilon,$$

where ε is considered to be "small." When $\varepsilon = 0$, we will say that matrices $A_1, \ldots, A_r$ possess an *exact SGSD*, otherwise an *approximate SGSD* or *ε-SGSD*.

Numerical algorithms for simultaneous reduction of matrices are often obtained as generalizations of the corresponding algorithms for one matrix or a pair of matrices. We should mention three basic approaches for solving problems of simultaneous reduction of matrices. The first is a very general approach of Chu [3]; an efficient numerical implementation requires the integration of a possibly very stiff ODE, and that is a very hard problem. The second approach is a Jacobi-type algorithm, where each step consists of two (or one) Jacobi rotations Q_i and Z_i that are sought to minimize the merit function of the form

$$\sum_{k=1}^{r} ||Q_i(QA_kZ)Z_i||_{LF}^2,$$

where $|| \; ||_{LF}$ is the sum of squares of all elements in the strictly lower triangular part of the matrix and Q and Z are orthogonal matrices that are already obtained at the previous steps. This algorithm was proposed in [9]; it was shown that at each step it requires finding the roots of a polynomial of order 8. This algorithm is a generalization of the well-known algorithm of Cardoso and Souloumiac [2] for simultaneous diagonalization of a sequence of matrices using Jacobi rotations.

The third family of methods (besides Jacobi and continuous approaches) are *alternating least squares* methods. The problem of nonorthogonal joint diagonalization of matrices is very close to our problem (indeed it is the main application of the SGSD [9]). For nonorthogonal joint diagonalization many methods are available, for example, a very popular AC-DC (alternating columns–diagonal centers) method of Yeredor [12]. Also we can mention [13] where a noniterative method for the simultaneous diagonalization is presented.

The alternating-type methods are simple to implement, and each iteration is quite fast; however, they suffer from several drawbacks. One of them is that we can converge into a local minima of the objective function; another is that the number of iterations can be large.

In the case of $r = 2$ (a pair of matrices), a well-developed tool for the computation of the generalized Schur decomposition is the QZ algorithm [10, 7]. Its generalization to the case of several matrices was developed in [5] where it was called an extended QZ algorithm. However, the QZ algorithm is efficient only when the shifts are used. It is not clear how the technique of shifts can be generalized to the case $r > 2$.

In this paper we propose a new algorithm to compute the SGSD. We reach the purpose through a sequence of *deflation steps*. At each step we solve an optimization problem in $2n+r$ variables which is a direct generalization of the generalized eigenvalue problem (we call it the *simultaneous eigenvalue problem*). The main ingredient of the algorithm is our fast version for the Gauss–Newton method applied to the latter problem. We suggest an inexpensive update technique for the matrices at successive iterative steps.

We prove the local quadratic convergence in the case of exact SGSD for the given $A_1, \dots, A_r$ and linear convergence in the case of ε-SGSD, the convergence factor being proportional to ε. Numerical experiments confirm the effectiveness of our approach. For example, the computation of the SGSD for 128 matrices of size 128×128[1] in our examples takes less than a minute. The word "fast" in the title should be understood in the sense that our method is faster than a straightforward implementation of the Gauss–Newton approach due to some clever update tricks. We also give the asymptotic complexity of our method with respect to n and r. However, we did not perform any comparisons with other methods. Such results will be reported elsewhere.

2. Simultaneous eigenvalue problem and its solution.

2.1. Transformation of the initial problem. Given real matrices $A_1, \dots, A_r$, we want to find orthogonal matrices Q and Z making the matrices $QA_1Z, \dots, QA_rZ$ as upper triangular as possible. In order to do this, first consider a related problem of finding orthogonal matrices Q and Z such that the matrices QA_iZ are reduced to the following block triangular form:

$$QA_kZ \approx \begin{pmatrix} \lambda_k & v_k^\top \\ 0 & B_k \end{pmatrix}, \tag{2.1}$$

where B_k is an $(n-1) \times (n-1)$ matrix, λ_k is a scalar, and v_k is a vector of size $n-1$. As soon as (2.1) is found, we accomplish a *deflation step* reducing the problem to the same but smaller problem for the matrices B_k. Proceeding in the same way with B_k we finally reduce A_k to the upper triangular form. The question is, Can these matrices, in fact, be reduced to triangular form? At present we do not have any theoretical estimates on the error of the triangularization of B_k by means of orthogonal transformations. One can expect the following situation to happen: The matrices A_k are reducible to triangular form with the accuracy ε, B_k are reducible to triangular form with accuracy $c_1\varepsilon$ for some $c_1 > 1$, and so on. If ε is large (say, 10^{-3}), it may happen that after n steps of the algorithm no appropriate accuracy is obtained. A separate theoretical study of this issue is needed, and the results will be reported elsewhere. However, at least in our numerical experiments we do not observe such growth; i.e., if the initial set of matrices A_k can be reduced to the triangular form with accuracy ε, the approximation obtained by successive deflation (i.e., n steps are made) has accuracy $c\varepsilon$ with some small constant c, see the numerical examples section for details. Thus, our main goal is to find Q and Z that satisfy (2.1).

The approximate equations (2.1) are equivalent to

$$QA_kZe_1 \approx \lambda_k e_1,$$

where e_1 is the first unit vector. Since Q is orthogonal, we can multiply these equations from the left by $Q^\top$ without changing the residue:

$$A_kZe_1 \approx \lambda_k Q^\top e_1.$$

Now, introducing two vectors $x = Ze_1$ and $y = Q^\top e_1$, we end up with the following overdetermined system of equations to be solved in the least squares sense:

$$A_kx = \lambda_k y. \tag{2.2}$$

[1] $r \approx n$ is often the case in tensor applications of SGSD.

(It is easy to see that unknowns x, λ, and y are defined up to the multiplication by a constant; thus, a normalizing condition is needed, i.e., $||x|| = 1$.)

If $r = 2$, then this is a well-known generalized eigenvalue problem for a pair of matrices A_1, A_2. Note that since we work in real arithmetics, then our *existence assumption* of the reduction matrices Q and Z means that the matrix pencil $A_1 - \lambda A_2$ has no complex eigenvalues. However, it is worth pointing out that the method in this paper can be straightforwardly extended to the complex case if all computations are performed in complex arithmetic. The algorithm can also be modified to the case when matrices A_k are real and complex conjugate pairs may appear, but it is not trivial.

So we may refer to the problem (2.2) as the *simultaneous generalized eigenvalue problem*, or simply the *simultaneous eigenvalue problem* (SEP). Since x, y, and λ_k are defined up to the multiplication by a constant, we use the following normalizing condition:

$$||x|| = 1.$$

(The norm is a Euclidean norm of a vector; for matrices we use the Frobenius norm.)

The SEP (2.2) is the key component of our algorithm. How it can be solved will be described in the next section. If we consider this solver as a "black box," then the algorithm for calculation of the SGSD reads as follows.

ALGORITHM 2.1. Given r real matrices $A_1, \ldots, A_r$ of size $n \times n$, find orthogonal matrices Q and Z such that the matrices QA_kZ are as upper triangular as possible:

1. Set
$$m = n, \quad B_i = A_i, \quad i = 1, \ldots, r, \quad Q = Z = I.$$
2. If $m = 1$, then stop.
3. Solve the SEP
$$B_k x = \lambda_k y, \quad k = 1, \ldots, r.$$
4. Find $m \times m$ Householder matrices Q_m, Z_m such that
$$x = \beta_1 Q_m^\top e_1, \quad y = \beta_2 Z_m e_1.$$
5. Calculate C_k as $(m-1) \times (m-1)$ submatrices of $\widehat{B}_k$ defined as follows:
$$\widehat{B}_k = Q_m B_k Z_m = \begin{pmatrix} \alpha_k & v_k^\top \\ \varepsilon_k & C_k \end{pmatrix}.$$
6. Set
$$Q \leftarrow \begin{pmatrix} I_{(n-m)\times(n-m)} & 0 \\ 0 & Q_m \end{pmatrix} Q, \quad Z \leftarrow Z \begin{pmatrix} I_{(n-m)\times(n-m)} & 0 \\ 0 & Z_m \end{pmatrix}.$$
7. Set $m = m - 1$, $B_k = C_k$, and proceed to step 2.

2.2. Gauss–Newton algorithm for the simultaneous eigenvalue problem. In this section we present an algorithm for solving the SEP (2.2). To begin with, let us write (2.2) elementwise as follows:

$$\sum_{j=1}^{n} (A_k)_{ij} x_j = \lambda_k y_i, \quad i = 1, \ldots, n. \tag{2.3}$$

Now introduce $r \times n$ matrices A'_j with the entries

$$(A'_j)_{ki} = (A_k)_{ij}, \quad k = 1, \ldots, r, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n,$$

and a column vector $\lambda = [\lambda_1, \ldots, \lambda_r]^\top$. Then (2.3) becomes

$$\sum_{j=1}^{n} x_j A'_j = \lambda y^\top. \tag{2.4}$$

The set of equations (2.4) can be considered as an overdetermined system of nonlinear equations. To solve it, we derive a variant of the Gauss–Newton method and propose a fast scheme to implement it. Then we obtain some convergence estimates. It is worthy to note that in the case of two matrices ($r = 2$), the Gauss–Newton method is equivalent to the simple Newton method for the computation of the generalized eigenvectors of a matrix pencil.

The idea behind the Gauss–Newton method is to linearize the system (as in the standard Newton method) producing an overdetermined linear system and then solve it in the least squares sense.

From the linearization of (2.4) at some point (x, λ, y), we obtain the following overdetermined system:

$$\sum_{j=1}^{n} \widehat{x}_j A'_j = \lambda y^\top + \triangle\lambda y^\top + \lambda \triangle y^\top, \quad \widehat{x} = x + \triangle x, \quad ||\widehat{x}|| = 1. \tag{2.5}$$

At each iterative step, the system (2.5) has to be solved in the least squares sense. To cope with this problem, let us observe, first of all, that the unknowns $\triangle y$ and $\triangle \lambda_k$ can be easily excluded in the following way. To this end, find an $n \times n$ Householder matrix H such that

$$Hy = he_1$$

and an $r \times r$ Householder matrix C such that

$$C\lambda = ce_1.$$

(In the above equations, e_1 denotes the first column in the identity matrices of different sizes and h and c are scalars.) Premultiplying (2.5) by C and postmultiplying it by $H^\top$, we obtain

$$\sum_{j=1}^{n} \widehat{x}_j \widehat{A}'_j = che_1 e_1^\top + h\triangle\widehat{\lambda} e_1^\top + ce_1 \triangle\widehat{y}^\top, \tag{2.6}$$

where

$$\widehat{A}'_j = CA'_j H^\top, \quad \triangle\widehat{y} = H\triangle y, \quad \triangle\widehat{\lambda} = C\triangle\lambda.$$

The equivalence of (2.6) and (2.5) follows from the orthogonality of H and C. In particular, the coefficients $\widehat{x}_j$ are the same in both problems.

Now we can split the problem (2.6) into two independent problems. Indeed,

$$\begin{aligned}
&\left|\left|\sum_{j=1}^{n} \widehat{x}_j \widehat{A}'_j - che_1 e_1^\top + ce_1 \triangle \widehat{y}^\top + h \triangle \widehat{\lambda} e_1^\top \right|\right|^2 \\
&\quad = \left|\left|\sum_{j=1}^{n} \widehat{x}_j B'_j\right|\right|^2 + \sum_{k=2}^{r} \left(\sum_{j=1}^{n} \widehat{x}_j (\widehat{A}'_j)_{k1} - h\triangle\widehat{\lambda}_k\right)^2 \\
&\qquad + \sum_{i=2}^{r} \left(\sum_{j=1}^{n} \widehat{x}_j (\widehat{A}'_j)_{1i} - c\triangle\widehat{y}_i\right)^2 + \left(\sum_{j=1}^{n} \widehat{x}_j (\widehat{A}'_j)_{11} - ch - h\triangle\widehat{\lambda}_1 - c\triangle\widehat{y}_1\right)^2,
\end{aligned}$$

where the matrices B'_j are obtained from $\widehat{A}'_j$ by replacing the elements in the first row and column by zeroes. Therefore, if $c \neq 0$ and $h \neq 0$, then we can exclude $\triangle\widehat{\lambda}$ and $\triangle\widehat{y}$ (one can verify the equations below by direct algebraic manipulations) as follows:

$$\triangle\widehat{\lambda}_k = \frac{1}{h} \sum_{j=1}^{n} \widehat{x}_j (\widehat{A}'_j)_{k1}, \qquad \triangle\widehat{y}_i = \frac{1}{c} \sum_{j=1}^{n} \widehat{x}_j (\widehat{A}'_j)_{1i}, i = 2, \dots, n,$$

and

$$h\triangle\widehat{\lambda}_1 + ch + c\triangle\widehat{y}_1 = \sum_{j=1}^{n} \widehat{x}_j (\widehat{A}'_j)_{11}.$$

After eliminating $\triangle\widehat{\lambda}$ and $\triangle\widehat{y}$, we have the minimization problem only in terms of $\widehat{x}$:

$$\left|\left|\sum_{j=1}^{n} \widehat{x}_j B'_j\right|\right|^2 \to \min, \quad ||\widehat{x}|| = 1. \tag{2.7}$$

Once $\widehat{x}$ is found, $\triangle\widehat{y}$ and $\triangle\widehat{\lambda}$ can be determined from the equations

$$\left(\sum_{j=1}^{n} \widehat{x}_j \widehat{A}'_j\right)_{k1} = h\triangle\widehat{\lambda}_k, \quad k = 2, \dots, r, \qquad \left(\sum_{j=1}^{n} \widehat{x}_j \widehat{A}'_j\right)_{1i} = c\triangle\widehat{y}_i, \quad i = 2, \dots, n.$$

For the two unknowns $\triangle\widehat{y}_1$ and $\triangle\widehat{\lambda}_1$, we have only one equation, so one of these unknowns can be chosen arbitrarily.

What happens if c or h is zero? Consider, for example, the case $c = 0$. That means that $\lambda = 0$, and that, in turn, means that there exists a vector x such that

$$\sum_{j=1}^{n} x_j A'_j \approx 0,$$

so the matrices A'_j are approximately linearly dependent. To tackle this problem, we should first find an orthogonal basis in the set of matrices A'_j and only after that solve our problem.

Equation (2.5) is the *first order optimality condition* for our minimization problem. Since (2.6) and (2.7) follow from (2.5) if $||\lambda|| \neq 0$ and $||y|| \neq 0$, then we have the following.

Theorem 2.1. *If* (x^*, λ^*, y^*) *with* $||x^*|| = 1$ *minimizes*

$$\left|\left|\sum_{j=1}^{n} x_j A'_j - \lambda y^\top\right|\right|$$

and λ^*, y^* *are both nonzero vectors, then* x^* *solves* (2.7).

2.2.1. Practical way for calculation of $\widehat{x}$. However, the approach described above requires an explicit computation of the Householder matrices H and C and evaluation of $\widehat{A}_j$. It will be shown later that $\widehat{x}$ can be computed without any reference to the Householder matrices. Therefore, we need a way to find the new λ and y directly from the known $\widehat{x}$. Having obtained the new $\widehat{x}$, we propose to evaluate y and λ by the power method as follows:

$$\widetilde{\lambda} = B'y, \quad \widetilde{y} = B'^\top \lambda, \tag{2.8}$$

where

$$B' = \sum_{j=1}^{n} \widehat{x}_j A'_j.$$

The explanation is simple. If we know λ and y with accuracy δ (that is, $||\lambda - \lambda^*|| \leq \delta, \quad ||y - y^*|| \leq \delta$, where λ^*, y^* are the solution vectors), then, as it will be proved in section 3, the computed approximation x to the vector x^* satisfies

$$||x - x^*|| = \mathcal{O}(\delta^2 + \delta\varepsilon),$$

where ε is the smallest attainable residue in the SEP. Now we have to find new λ and y as left and right singular vectors of the matrix

$$\widetilde{B} = \sum_{j=1}^{n} x_j A'_j$$

corresponding to the maximal singular value. This can be done by a power method. The convergence speed of the power method is determined by the ratio of the two largest singular values. Since x is close to the solution, then the second largest singular value of $\widetilde{B}$ is small ($\sigma_2 \sim \varepsilon$) and the λ and y converge linearly with convergence speed proportional to ε.

To determine $\widehat{x}_j$, observe that the problem (2.7) is, in fact, a problem of finding the minimal singular value of a matrix

$$\widehat{B} = [\mathrm{vec}(B'_1), \ldots, \mathrm{vec}(B'_n)],$$

where the operator vec transforms a matrix into a vector taking the elements column-by-column. Therefore, $\widehat{x}$ is an eigenvector (normalized to have a unit norm) corresponding to the minimal eigenvalue of the $n \times n$ matrix $\Gamma = \widehat{B}^\top \widehat{B}$:

$$\Gamma\widehat{x} = \gamma_{\min}\widehat{x}.$$

This matrix Γ plays the key role in the solution process. Its elements are given by

$$\Gamma_{sl} = (B'_s, B'_l)$$

where $(\cdot, \cdot)$ is the Frobenius (Euclidean) scalar product of matrices.[2] To calculate the new vector $\widehat{x}$, we need to find the minimal eigenvalue and the corresponding eigenvector of the matrix Γ.

The solution of the problem (2.5) consists of two parts:

1. calculation of the matrix Γ;
2. finding the minimal eigenvalue and the corresponding eigenvector of the matrix Γ.

Since only one eigenvector for Γ is to be found, we propose to use the shifted inverse iteration using $\widehat{x}$ from the previous iteration as an initial guess. The complexity is then $\mathcal{O}(n^3)$.

Let us estimate the number of arithmetic operations required for step 1. The straightforward implementation of this step includes $\mathcal{O}(n^2 r + nr^2)$ operations for the calculation of B'_j and $\mathcal{O}(n^2 rn)$ operations for the calculation of $\Gamma = \widehat{B}^\top \widehat{B}$. The total cost of step 1 is

$$\mathcal{O}(n^3 r + n^2 r + nr^2).$$

However, Γ can be computed more efficiently without the explicit computation of the Householder matrices, which is described below.

2.2.2. Calculation of the matrix Γ. In this section we suggest an efficient method to acquire the entries of Γ. Recall that

$$\Gamma_{sl} = (B'_s, B'_l), \quad i, j = 1, \dots, n.$$

Thus, we need the scalar products (B'_s, B'_l). From the definition of B'_j it follows that B'_j and $\widehat{A}'_j$ are connected in the following way:

$$B'_j = \widehat{A}'_j - \widehat{A}'_j e_1 e_1^\top - e_1 e_1^\top \widehat{A}'_j + (\widehat{A}'_j)_{11} e_1 e_1^\top.$$

Hence, the required scalar products are expressed as

$$(B'_s, B'_l) = (\widehat{A}'_s, \widehat{A}'_l) - (\widehat{A}'_s e_1, \widehat{A}'_l e_1) - (\widehat{A}'^\top_s e_1, \widehat{A}'^\top_l e_1) + (\widehat{A}'_s)_{11} (\widehat{A}'_l)_{11}.$$

Taking into account that $\widehat{A}'_j = C A'_j H^\top$, we have

$$\Gamma_{sl} = (B'_s, B'_l) = (A'_s, A'_l) - \frac{(A'_s y, A'_l y)}{||y||^2} - \frac{((A'_s)^\top \lambda, (A'_l)^\top \lambda)}{||\lambda||^2} + \frac{(A'_s y, \lambda)(A'_l y, \lambda)}{||y||^2 ||\lambda||^2}. \tag{2.9}$$

A fast computation of Γ can be based on (2.9). Note also that

$$(A'_s, A'_l) = \sum_{ki} (A'_s)_{ki} (A'_l)_{ki} = \sum_{ki} (A_k)_{is} (A_k)_{il} = \left(\sum_{k=1}^{n} A_k^\top A_k \right)_{sl};$$

therefore, the first summand (A'_s, A'_l) can be computed once and for all y and λ in $\mathcal{O}(n^3 r)$ operations. The cost of computing vectors $A_s y$ and $(A')_s^\top \lambda$ for all $s = 1, \dots, n$ is $\mathcal{O}(n^2 r + r^2 n)$. The cost of computing scalar products $(A'_s y, A'_l y)$ and $((A')_s^\top \lambda, (A')_l^\top \lambda)$ is of the same order. The total complexity of computing Γ is

$$\mathcal{O}(n^3 r)$$

[2] For two matrices X and Y of dimensions $n \times m$ the Frobenius scalar product is $(X, Y)_F = \sum_{i=1}^n \sum_{j=1}^m x_{ij} y_{ij}$.

operations on the zero (initialization) step for the given matrices $A_1, \dots, A_k$ plus

$$\mathcal{O}(n^2 r + nr^2)$$

operations for each y and λ arising during iterations.

Now we are ready to describe the algorithm for solving the SEP.

Algorithm 2.2. Given a sequence of $n \times n$ matrices $A_1, \dots, A_r$ and an initial approximation to the solution of the SEP (2.2) x^0, y^0, λ^0, proceed as follows:

1. Set $k = 0$, and calculate the initial Gram matrix

$$\Gamma_0 = \sum_{i=1}^{r} A_i^\top A_i.$$

2. If x^k, y^k, and λ^k have converged, then stop, else continue.
3. Calculate the vectors $A'_s y^k$ and $A'_s \lambda^k$ for all $s = 1, \dots, n$.
4. Calculate the matrix Γ using (2.9).
5. Set x^{k+1} to the eigenvector corresponding to the minimal eigenvalue of Γ.
6. Calculate y^{k+1} and λ^{k+1} from (2.8).
7. Increase k by 1, and proceed to the step 2.

The selection of the stopping criterion is a very delicate issue and may be problem-dependent. In our experiments we used a criterion based on the functional value: If the functional being minimized changes less than a prescribed tolerance, then we stop.

It is important to note that the matrix Γ can be updated fast during the work of Algorithm 2.2. Indeed, the "hardest" work on each step of Algorithm 2.2 is the calculation of

$$\Gamma_0 = \sum_{k=1}^{r} B_k^\top B_k.$$

After the QZ equivalence transformation of each B_k, Γ_0 becomes

$$\sum_{k=1}^{r} \widehat{B}_k^\top \widehat{B}_k = \sum_{k=1}^{r} (Q_m B_k Z_m)^\top Q_m B_k Z_m = Z_m^\top \Gamma_0 Z_m.$$

We need to calculate

$$\widehat{\Gamma}_0 = \sum_{k=1}^{r} C_k^\top C_k,$$

where C_k is an $(n-1) \times (n-1)$ leading principal submatrix of $\widehat{B}_k$ starting from position $(2, 2)$. It is straightforward to see that

$$(C_k^\top C_k)_{ij} = (\widehat{B}_k^\top \widehat{B}_k)_{(i+1)(j+1)} - (\widehat{B}_k)_{i1} (\widehat{B}_k)_{j1}, \quad i = 1, \dots, n-1, \ j = 1, \dots, n-1.$$

Consequently,

$$(\widehat{\Gamma}_0)_{ij} = (Z_m^\top \Gamma_0 Z_m)_{(i+1)(j+1)} - \sum_{k=1}^{r} (\widehat{B}_k)_{i1} (\widehat{B}_k)_{j1}, \quad i = 1, \dots, n-1, \ j = 1, \dots, n-1.$$

The complexity of this update is $\mathcal{O}(n^2 r)$.

It remains to analyze the convergence properties of the algorithm.

3. Convergence. Assume that x^*, y^*, and λ^* solve the nonlinear minimization problem

$$\sum_{k=1}^{r} ||A_k x - \lambda_k y||^2 \to \min, \qquad ||x|| = 1, \tag{3.1}$$

and

$$A_k x^* = \lambda_k^* y^* + \varepsilon_k. \tag{3.2}$$

Suppose we have constructed approximations

$$y = y^* + \delta y,$$
$$\lambda = \lambda^* + \delta\lambda$$

and then computed the new approximation x to x^* using the Algorithm 2.1. What can we say about $||x - x^*||$?

Recall that (3.2) can be written in terms of matrices A'_j; so (2.6) takes on the form

$$\sum_{j=1}^{n} x_j^* A'_j = \lambda^* (y^*)^\top + \varepsilon, \tag{3.3}$$

where the residue ε is inserted. Obviously, it can be assumed that $||y^*|| = ||\lambda^*||$. Also we will need the normalized vectors

$$\widetilde{y} = \frac{y}{||y||}, \quad \widetilde{\lambda} = \frac{\lambda}{||\lambda||}, \quad \widetilde{y}^* = \frac{y^*}{||y^*||}, \quad \widetilde{\lambda}^* = \frac{\lambda^*}{||\lambda^*||}.$$

The vector x is an eigenvector of the matrix Γ defined by (2.9); this is a direct corollary from Theorem 2.1. Denote by Γ^* the matrix for y^* and λ^*. Then the following lemma holds true.

Lemma 3.1. *The following inequality holds:*

$$\begin{aligned} &|(\Gamma x^* - \Gamma^* x^*)_s| \\ &\quad \le ||A'_s||(||y^*||\ ||\lambda^*||(4\ ||\delta\widetilde{y}||^2 + ||\delta\widetilde{y}||\ ||\delta\widetilde{\lambda}|| + 4\ ||\widetilde{\lambda}||^2) + ||\varepsilon||(||\delta\widetilde{y}|| + ||\delta\widetilde{\lambda}||)) \\ &\qquad + \mathcal{O}(\delta^3 + ||\varepsilon||\delta^2), \end{aligned}$$

where $\delta = \max(||\delta y||, ||\delta\lambda||)$.

Proof. Using the definition of the matrix Γ and the equality (3.3) we have

$$\begin{aligned} ((\Gamma - \Gamma_0)x^*)_s = &- \left(A'_s \widetilde{y}, \sum_{l=1}^{n} x_l^* A'_l \widetilde{y} \right) - \left((A'_s)^\top \widetilde{\lambda}, \sum_{l=1}^{n} x_l^* (A'_l)^\top \widetilde{\lambda} \right) \\ &+ (a_s \widetilde{y}, \widetilde{\lambda}) \left(\sum_{l=1}^{n} x_l^* A'_l \widetilde{y}, \widetilde{\lambda} \right) \\ = &-(A'_s \widetilde{y}, \lambda^*)(y^*, \widetilde{y}) - (A'_s \widetilde{y}, \varepsilon \widetilde{y}) - ((A'_s)^\top \widetilde{\lambda}, y^*)(\lambda^*, \widetilde{\lambda}) \\ &- ((A'_s)^\top \widetilde{\lambda}, \varepsilon^\top \widetilde{\lambda}) + (A'_s \widetilde{y}, \widetilde{\lambda})((y^*, \widetilde{y})(\lambda^*, \widetilde{\lambda}) + (\varepsilon \widetilde{y}, \widetilde{\lambda})). \end{aligned}$$

Now set

$$\widetilde{y} = \widetilde{y}^* + \delta\widetilde{y}, \quad \widetilde{\lambda} = \widetilde{\lambda}^* + \delta\widetilde{\lambda}.$$

Since $||\widetilde{y}|| = ||\widetilde{y}^*|| = 1$ and

$$(\delta\widetilde{y}, y^*) = -2\ ||y^*||\ ||\delta\widetilde{y}||^2,$$

we obtain the following first order terms (denoted by Φ_1):

$$\begin{aligned}\Phi_1 = &-(A'_s\delta\widetilde{y}, \varepsilon\widetilde{y}^*) - (A'_s\widetilde{y}^*, \varepsilon\delta\widetilde{y}) - ((A'_s)^\top\widetilde{\lambda}^*, \varepsilon^\top\delta\widetilde{\lambda}) - ((A'_s)^\top\delta\widetilde{\lambda}, \varepsilon^\top\widetilde{\lambda}^*)\\ &+ (A'_s\widetilde{y}^*, \delta\widetilde{\lambda})(\varepsilon\widetilde{y}^*, \widetilde{\lambda}^*) + (A'_s\delta\widetilde{y}, \widetilde{\lambda}^*)(\varepsilon\widetilde{y}^*, \widetilde{\lambda}^*)\\ &+ (A'_s\widetilde{y}^*, \widetilde{\lambda}^*)((\varepsilon\delta\widetilde{y}, \widetilde{\lambda}^*) + \ (\varepsilon\widetilde{y}^*, \delta\widetilde{\lambda})).\end{aligned}$$

Φ_1 is estimated from above by

$$4\ ||A'_s||\ ||\varepsilon||\ (||\delta\widetilde{y}|| + ||\delta\widetilde{\lambda}||).$$

The second order terms are estimated in the same way (we omit terms of order $\mathcal{O}(\delta^2||\varepsilon||)$). We give here only the final result:

$$|\Phi_2| \le ||y^*||\ ||\lambda^*||\ ||A'_s||\ ||\delta\widetilde{y}||\ ||\delta\widetilde{\lambda}|| + 4\ ||A'_s||\ ||y^*||\ ||\lambda^*||\ (||\delta\widetilde{y}||^2 + ||\delta\widetilde{\lambda}||^2).$$

To finish the proof, it is left to note that $|((\Gamma - \Gamma^*)x^*)_s| \le (|\Phi_1| + |\Phi_2|) + \mathcal{O}(\delta^3 + \delta^2\varepsilon)$. □

Now we can estimate the error in the approximate x.

THEOREM 3.2. *If x is computed from y and λ using the Algorithm* 2.1*, x^*, y^*, and λ^* are the solution of the minimization problem* (3.1)*, and $y^* \ne 0$ and $\lambda^* \ne 0$, then*

$$||x - x^*|| \le \frac{1}{\gamma_{n-1} - \gamma_n}\sqrt{\sum_{s=1}^{n} ||A'_s||^2\ (C_1\delta^2 + C_2||\varepsilon||\ \delta)} + \mathcal{O}(\delta^3 + ||\varepsilon||\delta^2), \tag{3.4}$$

where

$$\delta = \max(||\delta y||, ||\delta\lambda||),$$
$$C_1 = 36, \quad C_2 = 8\left(\frac{1}{||y^*||} + \frac{1}{||\lambda^*||}\right),$$

and γ_n, γ_{n-1} are the two smallest eigenvalues of the matrix Γ^*.

Proof. If we consider matrix Γ as a perturbation of Γ^*, then x is a perturbation of the eigenvector x^*. The Davis–Kahan $\sin\theta$ theorem [4, 11] states that the angle θ between the original and the perturbed eigenvectors is bounded by

$$\sin\theta \le \frac{||r||}{\mathrm{g}},$$

where g is an absolute spectral gap (matrix Γ^* is a symmetric positive definite matrix; in that case the spectral gap for the smallest eigenvalue is $\gamma_{n-1} - \gamma_n$; see, for example, [11]) and r is the residual:

$$r = (\Gamma - \lambda^* I)x^* = (\Gamma - \Gamma^*)x^*.$$

Therefore,

$$||x - x^*|| \leq \frac{1}{\gamma_{n-1} - \gamma_n} ||r|| + \mathcal{O}(||r||^2).$$

Now, applying the inequality of Lemma 3.1 and taking into account that

$$||\delta\widetilde{y}|| \leq 2\frac{||\delta y||}{||y^*||} + \mathcal{O}(\delta^2), \quad ||\delta\widetilde{\lambda}|| \leq 2\frac{||\delta\lambda||}{||\lambda^*||} + \mathcal{O}(\delta^2),$$

we arrive at (3.4). □

The estimate (3.4) fully describes the local convergence of our method. If $||\varepsilon|| = 0$ (that is, matrices A_k can be exactly reduced to the triangular form), then the convergence is quadratic. In the case of nonzero but sufficiently small $||\varepsilon||$, the convergence is linear, but the convergence speed is proportional to $||\varepsilon||$.

Remark 1. An important requirement for the matrix Γ^* is that the minimal eigenvalue is simple and sufficiently well-separated. What happens if the minimal eigenvalue is not well separated from others is still an open question, both from theoretical and algorithmical points of view. A good solution seems to use a Levenberg–Marquardt modification of the Gauss–Newton method. It would be also interesting to analyze the relationship between the conditioning of the initial problem and the eigenvalues of the matrix Γ^*.

Remark 2. The numerical experiments show that when $||\varepsilon||$ is small enough, the algorithm converges globally (that means that it converges from *any* initial approximation (x_0, λ_0, y_0) to the solution of the minimization problem). But at present we have no rigorous formulations of the conditions required (and/or sufficient) for the algorithm to be globally convergent.

Remark 3. The use of Gramians is known to be numerically unstable. Why do we use Gramians in that case? The answer is that we sacrifice accuracy for speed. The direct solution of the minimization problem (3.1) (a main step of the algorithm) by means of, for example, SVD requires n^3r operations, and since a suitable updating technique during deflation is not currently known, it leads to n^4r complexity (compared with the n^3r complexity of our method). The price is that we may have only (in the ill-conditioned problem) $\mathcal{O}(\sqrt{\eta})$ accuracy, where η is the relative machine precision. In our case $\eta = 10^{-16}$; thus, eight correct digits are available. This is more than enough in most of the applications (especially when the *approximation error* is large, say, $\varepsilon \sim 10^{-3} - 10^{-5}$).

Remark 4. As was already mentioned, the requirement $\lambda^* \neq 0$ is not restrictive, because that means that the matrices A'_k are approximately linearly dependent and we can replace their linear combination $\sum_{j=1}^n x_j A'_j$ by a linear combination of a smaller number of matrices that form a basis in a linear subspace spanned by A'_k. Accurate analysis and the implementation (which vectors can be considered linearly dependent, the threshold selection, etc.) are the subject for future research.

4. Numerical experiments.

In this section we present some numerical experiments confirming the efficiency of our method. It was implemented in Fortran. The numerical experiments were performed on a Pentium 4 machine, 3.2 Ghz with g77 3.4.3 compiler optimized with -O3 option, with Lapack 2.0 and Atlas 3.4.1 libraries for matrix operations.

The first series of examples is created in the following way. We generate two random $n \times n$ matrices X and Y and r diagonal matrices Λ_k, $k = 1, \ldots, r$, of the

Table 4.1
Timings (in seconds) for the computation of SGSD of 10 *n-by-n matrices.*

n	Time
16	0.01
32	0.11
64	1.6
128	14.77
256	210.61

Table 4.2
Timings (in seconds) for the computation of SGSD of n n-by-n matrices.

n	Time
16	0.02
32	0.14
64	3.41
128	54.96
256	810.77

same size, and a set of matrices

$$A_k = X\Lambda_k Y, \quad k = 1, \dots, r.$$

The elements of X and Y and Λ_k are uniformly distributed on the interval $[-1, 1]$. As it was shown in [9], these sequences of matrices have an exact *SGSD*, because we can find orthogonal Q and Z such that

$$X = QR_1, \quad Y = R_2 Z,$$

with R_1 and R_2 being upper triangular. We also corrupt these matrices with multiplicative noise, setting

$$(\widehat{A}_k)_{ij} = (A_k)_{ij}(1 + \sigma\phi),$$

where ϕ are taken from the uniform distribution on $[-1, 1]$ and σ is a "noise level." We are interested in the following quantities:

- the convergence speed, its dependence on n,r, and σ;
- the stability. The dependence of the *residue* of the SGSD on σ. The residue is defined as

$$\left(\sum_{k=1}^{r} ||A_k - QT_k Z||_F^2 \right)^{1/2}.$$

We have observed that the speed of the algorithm does not depend pronouncedly on σ. The independence from σ means that the computational cost of the iterations (proportional to the number of iterations) is small relative to the cost of the update procedures done in each deflation step.

We perform two experiments. First we fix r and σ setting them to 10 and 10^{-6}, respectively, and change n. The timings (in seconds) are given in Table 4.1.

In the second experiment we set $r = n$. Corresponding timings are given in Table 4.2.

To check stability we take fixed $r = n = 64$ and vary the noise level. For each noise level 10 test sequences of matrices are generated, and the mean, maximal, and minimal values of the residue are reported in Table 4.3.

Table 4.3
Residues for different noise levels.

σ	Mean residue	Min residue	Max residue
10^{-16}	$2 \cdot 10^{-15}$	$9 \cdot 10^{-16}$	$5 \cdot 10^{-15}$
10^{-15}	$7 \cdot 10^{-15}$	$1 \cdot 10^{-15}$	$2 \cdot 10^{-14}$
10^{-14}	$3 \cdot 10^{-14}$	$4 \cdot 10^{-15}$	$8 \cdot 10^{-14}$
10^{-13}	$5 \cdot 10^{-14}$	$4 \cdot 10^{-14}$	$8 \cdot 10^{-14}$
10^{-12}	$2 \cdot 10^{-12}$	$4 \cdot 10^{-13}$	$4 \cdot 10^{-12}$
10^{-11}	$6 \cdot 10^{-11}$	$4 \cdot 10^{-12}$	$1 \cdot 10^{-11}$
10^{-10}	$4 \cdot 10^{-10}$	$4 \cdot 10^{-11}$	$1 \cdot 10^{-9}$
10^{-9}	$2 \cdot 10^{-8}$	$4 \cdot 10^{-10}$	$5 \cdot 10^{-8}$
10^{-8}	$4 \cdot 10^{-8}$	$4 \cdot 10^{-9}$	$1 \cdot 10^{-7}$
10^{-7}	$2 \cdot 10^{-7}$	$4 \cdot 10^{-8}$	$7 \cdot 10^{-7}$
10^{-6}	$6 \cdot 10^{-7}$	$4 \cdot 10^{-7}$	$1 \cdot 10^{-6}$
10^{-5}	$2 \cdot 10^{-5}$	$4 \cdot 10^{-6}$	$5 \cdot 10^{-5}$
10^{-4}	$4 \cdot 10^{-4}$	$4 \cdot 10^{-5}$	$1 \cdot 10^{-3}$
10^{-3}	$2 \cdot 10^{-3}$	$4 \cdot 10^{-4}$	$6 \cdot 10^{-3}$

5. Conclusion.

In this paper a problem of the calculation of the SGSD is considered. It is shown that this problem can be reduced to a series of smaller optimization problems which are a direct generalization of the generalized eigenvalue problem; that is why we called it the *SEP*. We have proposed the fast Gauss–Newton algorithm for the solution of the SEP and shown that the computations can be performed efficiently using careful update techniques. A local quasi-quadratic convergence result is obtained. If the number of matrices r is of order n (what frequently happens, if we use the SGSD for the computation of the canonical decomposition), then the complexity of the algorithm is $\mathcal{O}(n^4)$ arithmetic operations. The efficiency and robustness of the algorithm was demonstrated by some numerical examples. These examples, however, are rather artificial. The comparison with other methods and applications to real life problems will be reported elsewhere.

REFERENCES

[1] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *Numerical methods for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 927–949.

[2] J.-F. Cardoso and A. Souloumiac, *Jacobi angles for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 161–164.

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

[4] C. Davis and W. Kahan, *Some new bounds on perturbation of subspaces*, Bull. Amer. Math. Soc., 75 (1969), pp. 863–868.

[5] A.-J. Van der Veen and A. Paulraj, *An analytical constant modulus algorithm*, IEEE Trans. Signal Process., 44 (1996), pp. 1136–1155.

[6] R. Granat, B. Kågström, and D. Kressner, *Computing periodic deflating subspaces associated with a specified set of eigenvalues*, BIT, 47 (2007), pp. 763–791.

[7] B. Kågström and D. Kressner, *Multishift variants of the QZ algorithm with aggressive early deflation*, SIAM J. Matrix Anal. Appl., 29 (2006), pp. 199–227.

[8] L. D. Lathauwer and J. Castaing, *Blind identification of underdetermined mixtures by simultaneous matrix diagonalization*, IEEE Trans. Signal Process., 55 (2008), pp. 1096–1105.

[9] L. D. Lathauwer, B. De Moor, and J. Vanderwalle, *Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 295–327.

[10] C. B. Moler and G. W. Stewart, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[11] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Comput. Sci. Sci. Comput., Academic Press, Boston, MA, 1990.

[12] A. Yeredor, *Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation*, IEEE Trans. Signal Process., 50 (2002), pp. 1545–1553.

[13] A. Yeredor, *On using exact joint diagonalization for noniterative approximate joint diagonalization*, Signal Process. Lett., 12 (2005), pp. 645–648.

[14] A. Ziehe, M. Kawanabe, S. Hamerling, and K.-R. Müller, *A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation*, J. Mach. Learn. Res., 5 (2004), pp. 801–818.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 331–359

STRUCTURED BACKWARD ERRORS AND PSEUDOSPECTRA OF STRUCTURED MATRIX PENCILS*

BIBHAS ADHIKARI[†] AND RAFIKUL ALAM[‡]

Abstract. Structured backward perturbation analysis plays an important role in the accuracy assessment of computed eigenelements of structured eigenvalue problems. We undertake a detailed structured backward perturbation analysis of approximate eigenelements of linearly structured matrix pencils. The structures we consider include, for example, symmetric, skew-symmetric, Hermitian, skew-Hermitian, even, odd, palindromic, and Hamiltonian matrix pencils. We also analyze structured backward errors of approximate eigenvalues and structured pseudospectra of structured matrix pencils.

Key words. structured backward error, pseudospectrum, structured matrix pencils

AMS subject classifications. 65F15, 15A57, 15A18, 65F35

DOI. 10.1137/070696866

1. Introduction.

Backward perturbation analysis and condition numbers play an important role in the accuracy assessment of computed solutions of eigenvalue problems. Backward perturbation analysis determines the smallest perturbation for which a computed solution is an exact solution of the perturbed problem. On the other hand, condition numbers measure the sensitivity of solutions to small perturbations in the data of the problem. Thus, backward errors when combined with condition numbers provide approximate upper bounds on the errors in the computed solutions.

Structured eigenvalue problems occur in many applications (see, for example, [16, 21, 25] and the references therein). With a view to preserving structures and their associated properties, structured preserving algorithms for structured eigenproblems have been proposed in the literature (see, for example, [4, 5, 7, 11, 20, 21] and the references therein). Consequently, there is a growing interest in the structured perturbation analysis of structured eigenproblems (see, for example, [10, 13, 12, 24, 22, 6] for sensitivity analysis of structured eigenproblems).

The main purpose of this paper is to undertake a detailed structured backward perturbation analysis of approximate eigenelements of linearly structured matrix pencils. Needless to mention that structured backward errors when combined with structured condition numbers provide approximate upper bounds on the errors in the computed eigenelements. Hence, structured backward perturbation analysis plays an important role in the accuracy assessment of approximate eigenelements of structured pencils. Further, it also plays an important role in the selection of an optimum structured linearization of a structured matrix polynomial [1]. This assumes significance due to the fact that linearization is a standard approach to solving a polynomial eigenvalue problem (see, for example, [15] and the references therein).

We consider regular matrix pencils of the form $\mathrm{L}(\lambda) = A + \lambda B$, where A and B are square matrices of size n. We assume L to be linearly structured, that is, L to be an element of a real or a complex linear subspace $\mathbb{S}$ of the space of pencils. More

*Received by the editors July 11, 2007, accepted for publication (in revised form) by D. Boley December 10, 2008; published electronically March 25, 2009.
http://www.siam.org/journals/simax/31-2/69686.html

[†]Department of Mathematics, IIT Guwahati, Guwahati-781039, India (bibhas@iitg.ernet.in).

[‡]Department of Mathematics, IIT Guwahati, Guwahati-781039, India (rafik@iitg.ernet.in, rafikul@yahoo.com).

specifically, we consider ten special classes of linearly structured pencils, namely, T-symmetric, T-skew-symmetric, T-odd, T-even, T-palindromic, H-Hermitian, H-skew-Hermitian, H-even and H-odd, and H-palindromic. These structures, defined in the next section, are prototypes of structured pencils which occur in many applications (see, [16, 21] and the references therein). We also consider $\mathbb{S}$ to be the space of pencils whose coefficient matrices are elements of Jordan and/or Lie algebras associated with the scalar product $(x, y) \mapsto y^T M x$ or $(x, y) \mapsto y^H M x$, where M is unitary and $M^T = \pm M$ or $M^H = \pm M$. For example, when $M := \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$, the Lie and Jordan algebras associated with the scalar product $(x, y) \mapsto y^H M x$ consist of Hamiltonian and skew-Hamiltonian matrices, respectively. The structures so considered encompass a wide variety of structured pencils and, in particular, include pencils whose coefficient matrices are Hamiltonian and skew-Hamiltonian. We show, however, that analyzing these wide classes of structured pencils ultimately boils down to analyzing one of the ten special classes of structured pencils considered above. Consequently, in this paper, we consider these ten special classes of structured pencils and investigate structured backward perturbation analysis of approximate eigenelements.

So, let $\mathbb{S}$ be the space of pencils having one of the ten structures. Let $\mathrm{L} \in \mathbb{S}$ and $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ with $x^H x = 1$. Then we define the structured backward error $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ of (λ, x) by

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) := \inf\{|||\Delta\mathrm{L}||| : \Delta\mathrm{L} \in \mathbb{S} \text{ and } \mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0\}.$$

Here the pencil norm $|||\mathrm{L}|||$ is given by $|||\mathrm{L}||| := \sqrt{\|A\|^2 + \|B\|^2}$, where $\mathrm{L}(z) = A + zB$ and $\|\cdot\|$ is either the spectral norm or the Frobenius norm on $\mathbb{C}^{n\times n}$. The main contributions of this paper are as follows.

Given $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ with $x^H x = 1$ and $\mathrm{L} \in \mathbb{S}$, we show that there is a pencil $\mathrm{K} \in \mathbb{S}$ such that $\mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0$. Consequently, $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) < \infty$. We determine $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and construct a pencil $\Delta\mathrm{L} \in \mathbb{S}$ such that $|||\Delta\mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$. Moreover, we show that $\Delta\mathrm{L}$ is unique for the Frobenius norm on $\mathbb{C}^{n\times n}$, but there are infinitely many such $\Delta\mathrm{L}$ for the spectral norm on $\mathbb{C}^{n\times n}$. Further, for the spectral norm, we show how to construct all such $\Delta\mathrm{L}$. In either case, we show that if $\mathrm{K} \in \mathbb{S}$ is such that $\mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0$, then $\mathrm{K} = \Delta\mathrm{L} + (I - xx^H)^* \,\mathrm{N}(I - xx^H)$ for some $\mathrm{N} \in \mathbb{S}$, where $(I - xx^H)^*$ denotes the transpose or the conjugate transpose of $(I - xx^H)$ depending upon the structure defined by $\mathbb{S}$. Furthermore, we show that the unstructured backward error $\eta(\lambda, x, \mathrm{L})$ of (λ, x) is a lower bound of $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and is attained by $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ for certain $\lambda \in \mathbb{C}$. However, $\eta(\lambda, x, \mathrm{L}) \neq \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ for most $\lambda \in \mathbb{C}$.

Next, we consider structured pseudospectra of structured matrix pencils. It is a well-known fact that pseudospectra of matrices and matrix pencils are powerful tools for sensitivity and perturbation analysis (see, [26] and the references therein). We consider structured and unstructured ϵ-pseudospectra

$$\Lambda^{\mathbb{S}}_{\epsilon}(\mathrm{L}) := \left\{\lambda \in \mathbb{C} : \eta^{\mathbb{S}}(\lambda, \mathrm{L}) \leq \epsilon\right\} \text{ and } \Lambda_{\epsilon}(\mathrm{L}) := \{\lambda \in \mathbb{C} : \eta(\lambda, \mathrm{L}) \leq \epsilon\}$$

of L, where $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) := \min_{x^H x = 1} \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and $\eta(\lambda, \mathrm{L}) := \min_{x^H x=1} \eta(\lambda, x, \mathrm{L})$, respectively, are structured and unstructured backward errors of an approximate eigenvalue λ. When L is T-symmetric or T-skew-symmetric pencils, we show that $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$ for the spectral norm and $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \sqrt{2}\,\eta(\lambda, \mathrm{L})$ for the Frobenius norm. Consequently, for these structures, we show that $\Lambda^{\mathbb{S}}_{\epsilon}(\mathrm{L}) = \Lambda_{\epsilon}(\mathrm{L})$ for the spectral norm and $\Lambda^{\mathbb{S}}_{\epsilon}(\mathrm{L}) = \Lambda_{\epsilon/\sqrt{2}}(\mathrm{L})$ for the Frobenius norm. For the rest of the structures, we show that there is a set $\Omega \subset \mathbb{C}$ such that $\Lambda^{\mathbb{S}}_\epsilon(\mathrm{L}) \cap \Omega = \Lambda_\epsilon(\mathrm{L}) \cap \Omega$. For example, $\Omega = \mathbb{R}$ when L is H-Hermitian or H-skew-Hermitian and $\Omega = i\mathbb{R}$ when L is H-even or H-odd. Often the spectrum of L is symmetric with respect to Ω. When Ω does not contain an eigenvalue of L, it is of practical importance to determine the smallest perturbation $\Delta\mathrm{L} \in \mathbb{S}$ of L such that $\mathrm{L} + \Delta\mathrm{L}$ has an eigenvalue in Ω. We show how to construct such a $\Delta\mathrm{L}$. Indeed, we show that the equality $\Lambda^{\mathbb{S}}_\epsilon(\mathrm{L}) \cap \Omega = \Lambda_\epsilon(\mathrm{L}) \cap \Omega$ plays a crucial role in the construction of such a $\Delta\mathrm{L}$.

The paper is organized as follows. In section 2, we define the ten special classes of structured pencils mentioned above. We also discuss some basic facts about spectral symmetry of structured pencils and, given $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and a structured pencil L, we show that there exists a structured pencil K such that $\mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0$. In section 3, we undertake a detailed structured backward perturbation analysis when $\mathbb{C}^{n\times n}$ is equipped with the Frobenius norm. For each of the ten structures, we derive $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and a unique $\Delta\mathrm{L} \in \mathbb{S}$ such that $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$. In section 4, we undertake a detailed structured backward perturbation analysis for each of the ten classes of structured pencils when $\mathbb{C}^{n\times n}$ is equipped with the spectral norm. We show that the choice of a norm on $\mathbb{C}^{n\times n}$ plays a crucial role in the structured backward perturbation analysis. Finally, in section 5, we analyze structured pseudospectra of structured pencils.

Notation. We consider 2-norm on $\mathbb{C}^n$ defined by $\|x\|_2 := (x^H x)^{1/2}$, where x^H is the conjugate transpose of x. We denote the set of n-by-n matrices with real or complex entries by $\mathbb{C}^{n\times n}$. For $A \in \mathbb{C}^{n\times n}$, we denote the transpose of A by A^T and the conjugate transpose of A by A^H. We consider spectral norm and the Frobenius norm on $\mathbb{C}^{n\times n}$. For $A \in \mathbb{C}^{n\times n}$, the spectral norm of A is given by $\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2$ and the Frobenius norm of A is given by $\|A\|_F := (\mathrm{trace}(A^H A))^{1/2}$. We denote the smallest singular value of $A \in \mathbb{C}^{n\times n}$ by $\sigma_{\min}(A)$. The Moore–Penrose inverse of A is denoted by $A^\dagger$. As usual, the conjugate of a complex number z is denoted by $\overline{z}$. For a matrix A, $\overline{A}$ denotes the matrix whose entries are conjugate of that of A. The spectrum of $A \in \mathbb{C}^{n\times n}$ is denoted by $\Lambda(A)$.

2. Structured matrix pencils.

We consider n-by-n matrix pencils of the form $\mathrm{L}(\lambda) := A + \lambda B$, where $A, B \in \mathbb{C}^{n\times n}$, and $\lambda \in \mathbb{C}$. Thus, the set of n-by-n matrix pencils consists of affine transformations from $\mathbb{C}$ to $\mathbb{C}^{n\times n}$ which we denote by $\mathbb{A}^{n\times n}$. Hence, $\mathbb{A}^{n\times n}$ is a vector space which we endow with an appropriate norm $|\!|\!|\cdot|\!|\!|$ as follows. Let $\mathrm{L} \in \mathbb{A}^{n\times n}$ be given by $\mathrm{L}(\lambda) = A + \lambda B$. Then we define the pencil norm $|\!|\!|\mathrm{L}|\!|\!|$ by

$$|\!|\!|\mathrm{L}|\!|\!| := \left(\|A\|^2 + \|B\|^2\right)^{1/2}, \tag{2.1}$$

where $\|\cdot\|$ is either the spectral norm or the Frobenius norm on $\mathbb{C}^{n\times n}$. We refer to [3] for various other norms on $\mathbb{A}^{n\times n}$. It is evident that $\|\mathrm{L}(\lambda)\| \leq |\!|\!|\mathrm{L}|\!|\!| \; \|(1, \lambda)\|_2$.

The spectrum $\Lambda(\mathrm{L})$ of a regular pencil $\mathrm{L} \in \mathbb{A}^{n\times n}$ is given by

$$\Lambda(\mathrm{L}) := \{\lambda \in \mathbb{C} : \mathrm{rank}(\mathrm{L}(\lambda)) < n\}.$$

To be precise, $\Lambda(\mathrm{L})$ consists of finite eigenvalues of L. When B is singular, the pencil L has an infinite eigenvalue. In this paper, we consider only finite eigenvalues of matrix pencils. By convention, if $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$, then x is assumed to be nonzero, that is, $x \neq 0$. Treating (λ, x) as an approximate eigenpair of L, we define the backward error of (λ, x) by

$$\eta(\lambda, x, \mathrm{L}) := \inf\{|\!|\!|\Delta\mathrm{L}|\!|\!| : \Delta\mathrm{L} \in \mathbb{A}^{n\times n} \text{ and } \mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0\}.$$

We follow the convention that if L is given by $\mathrm{L}(\lambda) = A + \lambda B$, then the pencil $\Delta\mathrm{L}$ to be of the form $\Delta\mathrm{L}(\lambda) = \Delta A + \lambda \Delta B$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$. Then setting $r := -\mathrm{L}(\lambda)x$, we have

$$\eta(\lambda, x, \mathrm{L}) = \frac{\|r\|_2}{\|x\|_2 \|(1, \lambda)\|_2}.$$

Indeed, defining $\Delta A := \frac{r x^H}{x^H x (1+|\lambda|^2)}$ and $\Delta B := \frac{\overline{\lambda} r x^H}{x^H x (1+|\lambda|^2)}$, and considering the pencil $\Delta\mathrm{L}(z) = \Delta A + z \Delta B$, we have $|||\Delta\mathrm{L}||| = \|r\|_2 / \|x\|_2 \|(1, \lambda)\|_2$ and $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$.

Next, let $\mathbb{S}$ be a (real or complex) linear subspace of $\mathbb{A}^{n\times n}$. Pencils in $\mathbb{S}$ will be referred to as structured pencils. Let $\mathrm{L} \in \mathbb{S}$. Then treating $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^{n\times n}$ as an approximate eigenpair of L, we define the structured backward error of (λ, x) by

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) := \inf\{|||\Delta\mathrm{L}||| : \Delta\mathrm{L} \in \mathbb{S} \text{ and } \mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0\}.$$

Obviously, we have $\eta(\lambda, x, \mathrm{L}) \le \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$. Let L be given by $\mathrm{L}(z) = A + zB$. Then the ten special structures of L we consider in this paper are as follows.

- ***T*-symmetric:** $\mathrm{L}(\lambda)^T = \mathrm{L}(\lambda)$ for all $\lambda \in \mathbb{C}$, that is, $A^T = A$ and $B^T = B$.
- ***T*-skew-symmetric:** $\mathrm{L}(\lambda)^T = -\mathrm{L}(\lambda)$ for all $\lambda \in \mathbb{C}$, that is, $A^T = -A$ and $B^T = -B$.
- ***T*-even:** $\mathrm{L}(\lambda)^T = \mathrm{L}(-\lambda)$ for all $\lambda \in \mathbb{C}$, that is, $A^T = A$ and $B^T = -B$.
- ***T*-odd:** $\mathrm{L}(\lambda)^T = -\mathrm{L}(-\lambda)$ for all $\lambda \in \mathbb{C}$, that is, $A^T = -A$ and $B^T = B$.
- ***T*-palindromic:** $\mathrm{L}(\lambda)^T = \lambda \mathrm{L}(1/\lambda)$ for all $\lambda \neq 0$, that is, $B = A^T$.
- ***H*-Hermitian:** $\mathrm{L}(\lambda)^H = \mathrm{L}(\overline{\lambda})$ for all $\lambda \in \mathbb{C}$, that is, $A^H = A$ and $B^H = B$.
- ***H*-skew-Hermitian:** $\mathrm{L}(\lambda)^H = -\mathrm{L}(\overline{\lambda})$ for all $\lambda \in \mathbb{C}$, that is, $A^H = -A$ and $B^H = -B$.
- ***H*-even:** $\mathrm{L}(\lambda)^H = \mathrm{L}(-\overline{\lambda})$ for all $\lambda \in \mathbb{C}$, that is, $A^H = A$ and $B^H = -B$.
- ***H*-odd:** $\mathrm{L}(\lambda)^H = -\mathrm{L}(-\overline{\lambda})$ for all $\lambda \in \mathbb{C}$, that is, $A^H = -A$ and $B^H = B$.
- ***H*-palindromic:** $\mathrm{L}(\lambda)^H = \overline{\lambda}\mathrm{L}(1/\overline{\lambda})$ for all $\lambda \neq 0$, that is, $B = A^H$.

Let L be a regular pencil. We say that (λ, x, y) is an eigentriple of L if λ is an eigenvalue of L and x and y, respectively, are right and left eigenvectors of L corresponding to λ; that is, $\mathrm{L}(\lambda)x = 0$ and $y^H \mathrm{L}(\lambda) = 0$. An eigentriple (λ, x, y) is said to be normalized if $y^H y = x^H x = 1$. We consider only normalized eigentriples. Now, for ready reference, we collect some basic facts about eigenpairs of structured pencils in the following theorem.

THEOREM 2.1. *Let* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zB$. *Let* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ *be an eigenpair of* L. *Then we have the following.*

$\mathbb{S}$	***eigenvalue pairing***	***eigentriple***	$x^T A x$	$x^T B x$
T-symmetric	λ	$(\lambda, x, \overline{x})$	*in* $\mathbb{C}$	*in* $\mathbb{C}$
T-skew-symmetric	λ	$(\lambda, x, \overline{x})$	0	0
T-even	$(\lambda, -\lambda)$	$(\lambda, x, \overline{y}), (-\lambda, y, \overline{x})$	0	0
T-odd	$(\lambda, -\lambda)$	$(\lambda, x, \overline{y}), (-\lambda, y, \overline{x})$	0	0 *if* $\lambda \neq 0$
T-palindromic	$(\lambda, 1/\lambda)$	$(\lambda, x, \overline{y}), (1/\lambda, y, \overline{x})$	0 *if* $\lambda \neq -1$	0 *if* $\lambda \neq -1$
	eigenvalue pairing	***eigentriple***	$x^H A x$	$x^H B x$
H-Hermitian / *H-skew-Hermitian*	$(\lambda, \overline{\lambda})$	(λ, x, y) $(\overline{\lambda}, y, x)$	0 *if* $\mathsf{im}\,\lambda \neq 0$	0 *if* $\mathsf{im}\,\lambda \neq 0$
H-even/ *H-odd*	$(\lambda, -\overline{\lambda})$	(λ, x, y) $(-\overline{\lambda}, y, x)$	0 *if* $\mathsf{re}\,\lambda \neq 0$	0 *if* $\mathsf{re}\,\lambda \neq 0$
H-palindromic	$(\lambda, 1/\overline{\lambda})$	$(\lambda, x, y), (1/\overline{\lambda}, y, x)$	0 *if* $\lvert\lambda\rvert \neq 1$	0 *if* $\lvert\lambda\rvert \neq 1$

Proof. Note that when L is T-symmetric or T-skew-symmetric, we have $\mathrm{L}(\lambda)x = 0$ and $\overline{x}^H\mathrm{L}(\lambda) = 0$. Hence, $(\lambda, x, \overline{x})$ is an eigentriple of L. In particular, if L is T-skew-symmetric, then both A and B are skew-symmetric and, hence, $x^TAx = 0 = x^TBx$.

When L is T-even or T-odd, we have $\mathrm{L}(\lambda)^T = \mathrm{L}(-\lambda)$ or $\mathrm{L}(\lambda)^T = -\mathrm{L}(-\lambda)$. Hence, if $\mathrm{L}(\lambda)x = 0$ and $\mathrm{L}(-\lambda)y = 0$, then $\overline{x}^H\mathrm{L}(-\lambda) = 0$ and $\overline{y}^H\mathrm{L}(\lambda) = 0$. This shows $(\lambda, -\lambda)$ pairing of eigenvalues and that $(\lambda, x, \overline{y})$ and $(-\lambda, y, \overline{x})$ are eigentriples. When L is T-even, B is skew-symmetric and, hence, $x^TBx = 0$. Consequently, $x^T\mathrm{L}(\lambda)x = 0 \Rightarrow x^TAx = 0$. Similarly, when L is T-odd, A is skew-symmetric and, hence, $x^TAx = 0$. Consequently, $x^T\mathrm{L}(\lambda)x = 0 \Rightarrow x^TBx = 0$ whenever $\lambda \neq 0$. The proof is similar for H-Hermitian, H-skew-Hermitian, H-odd, and H-even pencils.

Now let L be T-palindromic given by $\mathrm{L}(z) = A + zA^T$. Suppose that $\lambda \neq 0$. Then $\mathrm{L}(\lambda)x = 0 \Rightarrow \overline{x}^H\mathrm{L}(1/\lambda) = 0$ which shows $(\lambda, 1/\lambda)$ pairing of eigenvalues. It also follows that $(\lambda, \overline{y}, x)$ is an eigentriple of L if and only if $(1/\lambda, \overline{x}, y)$ is an eigentriple of L. Note that $x^T\mathrm{L}(\lambda)x = 0 \Rightarrow x^TAx + \lambda x^TAx = 0$. Thus, if $\lambda \neq -1$, then $x^TAx = 0$.

Similarly, when L is H-palindromic it follows that (λ, y, x) is an eigentriple of L if and only if $(1/\overline{\lambda}, x, y)$ is an eigentriple of L. Now $x^H\mathrm{L}(\lambda)x \Rightarrow x^HAx + \lambda\, \overline{x^HAx} = 0 \Rightarrow |x^HAx| = |\lambda|\ |x^HAx| \Rightarrow (1 - |\lambda|)\ |x^HAx| = 0$. Hence, for $|\lambda| \neq 1$, we have $x^H Ax = 0$. □

Next, we show that if $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $\mathrm{L} \in \mathbb{S}$, then there exists $\Delta\mathrm{L} \in \mathbb{S}$ such that (λ, x) is an eigenpair of $\mathrm{L} + \Delta\mathrm{L}$, that is, $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$. Consequently, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) < \infty$.

THEOREM 2.2. *Let* $\mathbb{S} \in \{$*T-symmetric, T-skew-symmetric, T-odd, T-even, H-Hermitian, H-skew-Hermitian, H-odd, H-even*$\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zB$. *Let* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ *be such that* $x^Hx = 1$. *Set* $r := -\mathrm{L}(\lambda)x$ *and define*

$$\Delta A := \begin{cases} -\overline{x}x^TAxx^H + \frac{1}{1+|\lambda|^2}\left[\overline{x}r^T + rx^H - 2\left(x^Tr\right)\overline{x}x^H\right], & \textit{if } A = A^T, \\ -\frac{1}{1+|\lambda|^2}\left[\overline{x}r^T - rx^H\right], & \textit{if } A = -A^T, \end{cases}$$

$$\Delta B := \begin{cases} -\overline{x}x^TBxx^H + \frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T + rx^H - 2\left(x^Tr\right)\overline{x}x^H\right], & \textit{if } B = B^T, \\ -\frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T - rx^H\right], & \textit{if } B = -B^T, \end{cases}$$

and

$$\Delta A := \begin{cases} -xx^HAxx^H + \frac{1}{1+|\lambda|^2}\left[xr^H\left(I - xx^H\right) + \left(I - xx^H\right)rx^H\right], & \textit{if } \ A = A^H, \\ -xx^HAxx^H - \frac{1}{1+|\lambda|^2}\left[xr^H\left(I - xx^H\right) - \left(I - xx^H\right)rx^H\right], & \textit{if } \ A = -A^H. \end{cases}$$

$$\Delta B := \begin{cases} -xx^HBxx^H + \frac{1}{1+|\lambda|^2}\left[\lambda xr^H\left(I - xx^H\right) + \overline{\lambda}\left(I - xx^H\right)rx^H\right], & \textit{if } \ B = B^H \\ -xx^HBxx^H - \frac{1}{1+|\lambda|^2}\left[\lambda xr^H\left(I - xx^H\right) - \overline{\lambda}\left(I - xx^H\right)rx^H\right], & \textit{if } \ B = -B^H. \end{cases}$$

Consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L} \in \mathbb{S}$ *and* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$.

Proof. The proof is computational and is easy to check. □

For palindromic pencils, we have the following result.

THEOREM 2.3. *Let* $\mathbb{S} \in \{$*T-palindromic, H-palindromic*$\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zA^*$, *where* $A^* = A^T$ *or* $A^* = A^H$. *Let* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ *be such that*

$x^Hx = 1$. *Set* $r := -\mathrm{L}(\lambda)x$ *and define*

$$\Delta A := \begin{cases} -\overline{x}x^TAxx^H + \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}\overline{x}r^T\left(I - xx^H\right) + \left(I - \overline{x}x^T\right)rx^H\right], & \textit{if } B = A^T, \\ -xx^HAxx^H + \frac{1}{1+|\lambda|^2}\left[\lambda x r^H\left(I - xx^H\right) + \left(I - xx^H\right)rx^H\right], & \textit{if } B = A^H. \end{cases}$$

Consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z(\Delta A)^*$. *Then* $\Delta\mathrm{L} \in \mathbb{S}$ *and* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$.

Proof. The proof is computational and is easy to check. $\square$

In section 3, we consider general classes of linearly structured pencils whose coefficient matrices are elements of certain Jordan and/or Lie algebras and show that for these pencils structured backward perturbation analysis ultimately reduces to that of one of the ten classes of structured pencils discussed above.

3. Frobenius norm and structured backward errors. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$. Unless stated otherwise, we always assume that $x^Hx = 1$. Let $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) = A + zB$. In this section, we determine the structured backward error $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ when $\mathbb{C}^{n\times n}$ is equipped with the Frobenius norm. Recall that the pencil norm defined in (2.1) is then given by $|||\mathrm{L}||| := \sqrt{\|A\|_F^2 + \|B\|_F^2} = \|[A\ B]\|_F$. Also recall that the unstructured backward error $\eta(\lambda, x, \mathrm{L})$ for the spectral norm as well as for the Frobenius norm on $\mathbb{C}^{n\times n}$ is given by $\eta(\lambda, x, \mathrm{L}) = \|\mathrm{L}(\lambda)x\|_2/\|(1, \lambda)\|_2$.

THEOREM 3.1. *Let* $\mathbb{S}$ *be the space of* T*-symmetric pencils and let* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zB$. *Then for* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$, *setting* $r := -\mathrm{L}(\lambda)x$, *we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \frac{\sqrt{2\|r\|_2^2 - |x^Tr|^2}}{\|(1, \lambda)\|_2} \leq \sqrt{2}\,\eta(\lambda, x, \mathrm{L}).$$

Define $\Delta A := \frac{1}{1+|\lambda|^2}[\overline{x}r^T + rx^H - (r^Tx)\overline{x}x^H]$ *and* $\Delta B := \frac{\overline{\lambda}}{1+|\lambda|^2}[\overline{x}r^T + rx^H - (r^Tx)$ $\overline{x}x^H]$ *and consider the pencil* $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L}$ *is* T*-symmetric,* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ *and* $|||\Delta\mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. By Theorem 2.2 there is a $\Delta\mathrm{L} \in \mathbb{S}$ such that $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$. Let $\Delta\mathrm{L}$ be given by $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. Then we have $(\Delta A + \lambda\Delta B)x = r$. Choose $Q_1 \in \mathbb{C}^{n\times(n-1)}$ such that $Q := [x,\ Q_1]$ is unitary. Then

$$\widetilde{\Delta A} := Q^T\Delta AQ = \begin{pmatrix} a_{11} & a_1^T \\ a_1 & A_1 \end{pmatrix},$$
$$\widetilde{\Delta B} := Q^T\Delta BQ = \begin{pmatrix} b_{11} & b_1^T \\ b_1 & B_1 \end{pmatrix},$$
$$Q^Tr = \begin{pmatrix} x^Tr \\ Q_1^Tr \end{pmatrix},$$

where $A_1 = A_1^T$ and $B_1 = B_1^T$ are of size $n-1$. Since $\overline{Q}Q^T = I$, we have

$$\left(\overline{Q}\widetilde{\Delta A}Q^H + \lambda\overline{Q}\widetilde{\Delta B}Q^H\right)x = r \Rightarrow \left(\widetilde{\Delta A}Q^H + \lambda\widetilde{\Delta B}Q^H\right)x = Q^Tr = \begin{pmatrix} x^Tr \\ Q_1^Tr \end{pmatrix}.$$

As $Q^Hx = e_1$, the first column of the identity matrix, we have

$$\left(\widetilde{\Delta A} + \lambda\widetilde{\Delta B}\right)Q^Hx = \begin{pmatrix} x^Tr \\ Q_1^Tr \end{pmatrix} \Rightarrow \begin{pmatrix} a_{11} + \lambda b_{11} \\ a_1 + \lambda b_1 \end{pmatrix} = \begin{pmatrix} x^Tr \\ Q_1^Tr \end{pmatrix}.$$

This gives $a_{11} + \lambda b_{11} = x^Tr$ and $a_1 + \lambda b_1 = Q_1^Tr$ whose minimum norm solutions are

$$(a_1\ \ b_1) = Q_1^Tr\ \begin{pmatrix} 1 \\ \lambda \end{pmatrix}^{\dagger} \Rightarrow a_1 = \frac{1}{1+|\lambda|^2}Q_1^Tr,\ \ b_1 = \frac{\overline{\lambda}}{1+|\lambda|^2}Q_1^Tr$$

and $(a_{11} \;\; b_{11}) = x^T r \begin{pmatrix} 1 \\ \lambda \end{pmatrix}^{\dagger} \Rightarrow a_{11} = \frac{1}{1+|\lambda|^2} x^T r, \;\; b_{11} = \frac{\overline{\lambda}}{1+|\lambda|^2} x^T r$. Hence, we have

$$\widetilde{\Delta A} = \begin{pmatrix} \frac{1}{1+|\lambda|^2} x^T r & \frac{1}{1+|\lambda|^2} \left(Q_1^T r\right)^T \\ \frac{1}{1+|\lambda|^2} Q_1^T r & A_1 \end{pmatrix}, \quad \widetilde{\Delta B} = \begin{pmatrix} \frac{\overline{\lambda}}{1+|\lambda|^2} x^T r & \frac{\overline{\lambda}}{1+|\lambda|^2} \left(Q_1^T r\right)^T \\ \frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^T r & B_1 \end{pmatrix}.$$

This shows that the Frobenius norms of $\widetilde{\Delta A}$ and $\widetilde{\Delta B}$ are minimized when $A_1 = 0$ and $B_1 = 0$. Hence, $\|\Delta A\|_F^2 = \|\widetilde{\Delta A}\|_F^2 = |a_{11}|^2 + 2\,\|a_1\|_2^2$ and $\|\Delta B\|_F^2 = \|\widetilde{\Delta B}\|_F^2 = |b_{11}|^2 + 2\,\|b_1\|_2^2$. Note that $QQ^H = I \Rightarrow Q_1 Q_1^H = I - xx^H \Rightarrow \overline{Q}_1 Q_1^T = I - \overline{x}x^T$. Consequently, we have

$$|\!|\!|\Delta \mathrm{L}|\!|\!| = \left(\|\Delta A\|_F^2 + \|\Delta B\|_F^2\right)^{1/2} = \frac{\sqrt{|x^T r|^2 + 2\,\|(I - \overline{x}x^T)r\|_2^2}}{\|(1,\lambda)\|_2} = \frac{\sqrt{2\|r\|_2^2 - |x^T r|^2}}{\|(1,\lambda)\|_2}.$$

Next, we have

$$\begin{aligned} \Delta A = \overline{Q}\widetilde{\Delta A}Q^H &= \frac{1}{1+|\lambda|^2}\overline{x}x^T r x^H + \frac{1}{1+|\lambda|^2}\left[\overline{x}r^T Q_1 Q_1^H + \overline{Q}_1 Q_1^T r x^H\right] + \overline{Q}_1 A_1 Q_1^H \\ &= \frac{1}{1+|\lambda|^2}\left[\overline{x}r^T + r x^H - \left(r^T x\right)\overline{x}x^H\right] + \overline{Q}_1 A_1 Q_1^H, \\ \Delta B = \overline{Q}\widetilde{\Delta B}Q^H &= \frac{\overline{\lambda}}{1+|\lambda|^2}\overline{x}x^T r x^H + \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}\overline{x}r^T Q_1 Q_1^H + \overline{\lambda Q}_1 Q_1^T r x^H\right] + \overline{Q}_1 B_1 Q_1^H \\ &= \frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T + r x^H - \left(r^T x\right)\overline{x}x^H\right] + \overline{Q}_1 B_1 Q_1^H \end{aligned}$$

from which we obtain the desired pencil by setting $A_1 = 0$ and $B_1 = 0$. This completes the proof. □

Observe that if Y is symmetric and $Yx = 0$, then $Y = (I - xx^H)^T Z (I - xx^H)$ for some symmetric matrix Z. Consequently, we have $\overline{Q_1} A_1 Q_1^H = (I - xx^H)^T Z_1 (I - xx^H)$ and $\overline{Q_1} B_1 Q_1^H = (I - xx^H)^T Z_2 (I - xx^H)$ for some symmetric matrices Z_1 and Z_2. Hence, from the proof of Theorem 3.1 we have following.

COROLLARY 3.2. *Let* L *be a* T*-symmetric pencil and* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$. *Set* $r := -\mathrm{L}(\lambda)x$. *Let* K *be a* T*-symmetric pencil. Then* $\mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0$ *if and only if* $\mathrm{K}(z) = \Delta\mathrm{L}(z) + (I - xx^H)^T \mathrm{N}(z)(I - xx^H)$ *for some* T*-symmetric pencil* N, *where* ΔL *is the* T*-symmetric pencil given in Theorem* 3.1.

Next, we consider T-skew-symmetric pencils.

THEOREM 3.3. *Let* $\mathbb{S}$ *be the space of* T*-skew-symmetric pencils and let* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zB$. *Let* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ *and* $r := -\mathrm{L}(\lambda)x$. *Then* $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{2}\,\|r\|_2/\|(1,\lambda)\|_2 = \sqrt{2}\,\eta(\lambda, x, \mathrm{L})$. *Further, for the* T*-skew-symmetric pencil* ΔL *given in Theorem* 2.2, *we have* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ *and* $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. As A and B are skew-symmetric, from the proof of Theorem 3.1, we have

$$\begin{aligned} \widetilde{\Delta A} &= Q^T \Delta A Q = \begin{pmatrix} 0 & a_1^T \\ -a_1 & A_1 \end{pmatrix}, \\ \widetilde{\Delta B} &= Q^T \Delta B Q = \begin{pmatrix} 0 & b_1^T \\ -b_1 & B_1 \end{pmatrix}, \\ Q^T r &= \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix}, \end{aligned}$$

where A_1 and B_1 are skew-symmetric matrices of size $n-1$. Consequently, as before, we have $(\widetilde{\Delta A}+\lambda\widetilde{\Delta B})Q^H x = \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix}$ which gives $\begin{pmatrix} 0 \\ -a_1-\lambda b_1 \end{pmatrix} = \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix}$. Note that $x^T r = 0$ and the smallest norm solution of $-a_1 - \lambda b_1 = Q_1^T r$ is given by

$$(a_1 \;\; b_1) = Q_1^T r \begin{pmatrix} -1 \\ -\lambda \end{pmatrix}^{\dagger} \Rightarrow a_1 = -\frac{1}{1+|\lambda|^2} Q_1^T r, \quad b_1 = -\frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^T r.$$

Hence, we have

$$\Delta A = \overline{Q}\begin{pmatrix} 0 & -\frac{1}{1+|\lambda|^2}\left(Q_1^T r\right)^T \\ \frac{1}{1+|\lambda|^2} Q_1^T r & A_1 \end{pmatrix} Q^H,$$
$$\Delta B = \overline{Q}\begin{pmatrix} 0 & -\frac{\overline{\lambda}}{1+|\lambda|^2}\left(Q_1^T r\right)^T \\ \frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^T r & B_1 \end{pmatrix} Q^H.$$

Setting $A_1 = 0$ and $B_1 = 0$ we obtain $\Delta\mathrm{L}$ such that $|||\Delta\mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{2}\|r\|_2/\|(1,\lambda)\|_2$.

Since $\overline{Q}_1 Q_1^T = I - \overline{x}x^T$, we have

$$\Delta A = -\frac{1}{1+|\lambda|^2}\left[\overline{x}r^T - r x^H\right] + \overline{Q}_1 A_1 Q_1^H \text{ and } \Delta B = -\frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T - r x^H\right] + \overline{Q}_1 B_1 Q_1^H.$$

Setting $A_1 = B_1 = 0$ we obtain the T-skew-symmetric pencil $\Delta\mathrm{L}$ given in Theorem 2.2. □

Using the fact that if Y is skew-symmetric and $Yx = 0$ then $Y = (I - xx^H)^T Z (I - xx^H)$ for some skew-symmetric matrix Z, we obtain an analogue of Corollary 3.2 for T-skew-symmetric pencils.

Next, we derive structured backward errors for T-even and T-odd pencils.

Theorem 3.4. *Let $\mathbb{S} \in \{T\text{-even}, T\text{-odd}\}$ and $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) = A + zB$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{|x^T A x|^2 + \frac{2\|r\|_2^2 - 2|x^T r|^2}{1+|\lambda|^2}} = \frac{\sqrt{2\|r\|_2^2 + (|\lambda|^2 - 1)|x^T r|^2}}{\|(1,\lambda)\|_2}$$

when L *is T-even and*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{|x^T B x|^2 + \frac{2\|r\|_2^2 - 2|x^T r|^2}{1+|\lambda|^2}} = \dfrac{\sqrt{2\|r\|_2^2 + (|\lambda|^{-2} - 1)|x^T r|^2}}{\|(1,\lambda)\|_2}, & \text{if } \lambda \neq 0. \\ \sqrt{2}\,\eta(\lambda, x, \mathrm{L}), & \text{if } \lambda = 0, \end{cases}$$

when L *is T-odd. The pencil $\Delta\mathrm{L} \in \mathbb{S}$ given in Theorem 2.2 satisfies $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ and $|||\Delta\mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.*

Proof. First, assume that L is T-even. Then noting that $A = A^T$ and $B = -B^T$, the proof follows from similar arguments as those employed for T-symmetric and T-skew-symmetric pencils. Indeed, considering a unitary matrix $Q := [x,\ Q_1]$, we have

$$\widetilde{\Delta A} := Q^T \Delta A Q = \begin{pmatrix} a_{11} & a_1^T \\ a_1 & A_1 \end{pmatrix},$$
$$\widetilde{\Delta B} := Q^T \Delta B Q = \begin{pmatrix} 0 & b_1^T \\ -b_1 & B_1 \end{pmatrix},$$
$$Q^T r = \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix},$$

where $A_1 = A_1^T$ and $B_1 = -B_1^T$ are of size $n-1$. Consequently, we have $(\widetilde{\Delta A} + \lambda \widetilde{\Delta B})Q^H x = \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix} \Rightarrow \begin{pmatrix} a_{11} \\ a_1 - \lambda b_1 \end{pmatrix} = \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix}$. This gives $a_{11} = -x^T A x$. The smallest norm solution of $a_1 + \lambda b_1 = Q_1^T r$ is given by

$$(a_1 \;\; b_1) = Q_1^T r \begin{pmatrix} 1 \\ -\lambda \end{pmatrix}^{\dagger} \Rightarrow a_1 = \frac{1}{1+|\lambda|^2} Q_1^T r, \;\; b_1 = -\frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^T r.$$

Consequently, we have

$$\Delta A = \overline{Q} \begin{pmatrix} -x^T A x & \left(\frac{1}{1+|\lambda|^2} Q_1^T r\right)^T \\ \frac{1}{1+|\lambda|^2} Q_1^T r & A_1 \end{pmatrix} Q^H,$$
$$\Delta B = \overline{Q} \begin{pmatrix} 0 & \left(-\frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^T r\right)^T \\ \frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^T r & B_1 \end{pmatrix} Q^H.$$

Setting $A_1 = B_1 = 0$ and using the fact that $\overline{Q}_1 Q_1^T = I - \overline{x} x^T$, we obtain the pencil $\Delta \mathrm{L}$ such that

$$|||\Delta \mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{|x^T A x|^2 + \frac{2\|r\|_2^2 - 2|x^T r|^2}{1+|\lambda|^2}}.$$

Now simplifying expressions for ΔA and ΔB, we obtain

$$\Delta A = -\overline{x} x^T A x x^H + \frac{1}{1+|\lambda|^2} \left[\overline{x} r^T + r x^H - 2\left(x^T r\right) \overline{x} x^H\right] + \overline{Q}_1 A_1 Q_1^H,$$
$$\Delta B = -\frac{\overline{\lambda}}{1+|\lambda|^2} \left[\overline{x} r^T - r x^H\right] + \overline{Q}_1 B_1 Q_1^H.$$

Setting $A_1 = B_1 = 0$ we obtain the T-even pencil $\Delta \mathrm{L}$ given in Theorem 2.2.

When L is T-odd, the results follow by interchanging the role of A and B. □

It follows from Theorem 3.4 that for a T-even pencil, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \sqrt{2}\, \eta(\lambda, x, \mathrm{L})$ when $|\lambda| \leq 1$ and $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \|(1, \lambda)\|_2\, \eta(\lambda, x, \mathrm{L})$ when $|\lambda| > 1$. Similarly, for a T-odd pencil, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \sqrt{2}\, \eta(\lambda, x, \mathrm{L})$ when $|\lambda| \geq 1$ and $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \|(1, \lambda^{-1})\|_2\, \eta(\lambda, x, \mathrm{L})$ when $\lambda \neq 0$ and $|\lambda| < 1$.

We mention that an analogue of Corollary 3.2 holds for T-even and T-odd pencils as well. Now, we consider a T-palindromic pencil $\mathrm{L}(z) = A + zA^T$.

Theorem 3.5. *Let $\mathbb{S}$ be the space of T-palindromic pencils and $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) = A + zA^T$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{2}\sqrt{|x^T A x|^2 + \frac{\|r\|_2^2 - |x^T r|^2}{1+|\lambda|^2}} = \sqrt{2}\, \frac{\sqrt{\|r\|_2^2 - 2\mathsf{re}\lambda\, |x^T A x|^2}}{\|(1,\, \lambda)\|_2}, & \text{if } \lambda \neq -1, \\ \sqrt{2}\; \eta(\lambda, x, \mathrm{L}), & \text{if } \lambda = -1. \end{cases}$$

In particular, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{2}\; \eta(\lambda, x, \mathrm{L}),$ if $\lambda \in i\mathbb{R}$.

Now define

$$\Delta A = \begin{cases} \frac{1}{1+|\lambda|^2} \left[\overline{\lambda} \overline{x} r^T \left(I - x x^H\right) + \left(I - \overline{x} x^T\right) r x^H\right], & \text{if } \lambda = -1, \\ -\overline{x} x^T A x x^H + \frac{1}{1+|\lambda|^2} \left[\overline{\lambda} \overline{x} r^T \left(I - x x^H\right) + \left(I - \overline{x} x^T\right) r x^H\right], & \text{if } \lambda \neq -1, \end{cases}$$

and consider the pencil $\Delta \mathrm{L}(z) = \Delta A + z(\Delta A)^T$. *Then* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ *and* $|||\Delta \mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. By Theorem 2.3, there exists a T-palindromic pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta A^T$ such that $(\mathrm{L}(\lambda) + \Delta\mathrm{L}(\lambda))x = 0$. Let $Q_1 \in \mathbb{C}^{n\times(n-1)}$ be such that $Q := [x \quad Q_1]$ is unitary. Then

$$\widetilde{\Delta A} := Q^T \Delta A Q = \begin{pmatrix} a_{11} & a_1^T \\ b_1 & A_1 \end{pmatrix}, \quad Q^T r = \begin{pmatrix} x^T r \\ Q_1^T r \end{pmatrix}.$$

Now, if $\lambda \neq -1$ then by Theorem 2.1, we have $x^T(\Delta A + A)x = 0 \Rightarrow x^T \Delta A x = -x^T A x$. Hence, we have $a_{11} = -x^T A x$. When $\lambda = -1$, we have $\lambda a_{11}^T + a_{11} = x^T r = 0$ for any a_{11}. Since the aim is to minimize the Frobenius norm of ΔA, we set $a_{11} = 0$.

Next, the minimum norm solution of $a_1 \lambda + b_1 = Q_1^T r$ is given by

$$\begin{pmatrix} a_1 & b_1 \end{pmatrix} = Q_1^T r \begin{pmatrix} \lambda \\ 1 \end{pmatrix}^{\dagger} \Rightarrow a_1 = \frac{\overline{\lambda} Q_1^T r}{1 + |\lambda|^2}, \quad b_1 = \frac{Q_1^T r}{1+|\lambda|^2}.$$

Therefore, when $\lambda = -1$, we have

$$\Delta A = \overline{Q} \begin{pmatrix} 0 & \left(\frac{\overline{\lambda} Q_1^T r}{1+|\lambda|^2}\right)^T \\ \frac{Q_1^T r}{1+|\lambda|^2} & A_1 \end{pmatrix} Q^H.$$

Setting $A_1 = 0$, we obtain $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = |||\Delta\mathrm{L}||| = \sqrt{2}\|r\|_2/\sqrt{1+|\lambda|^2} = \sqrt{2}\,\eta(\lambda, x, \mathrm{L})$. Since $Q_1 Q_1^H = I - xx^H \Rightarrow \overline{Q}_1 Q_1^T = I - \overline{x}x^T$, simplifying the expression for ΔA, we obtain

$$\Delta A = \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}\overline{x}r^T\left(I - xx^H\right) + \left(I - \overline{x}x^T\right)rx^H\right] + \overline{Q}_1 A_1 Q_1^H.$$

When $\lambda \neq -1$, we have

$$\Delta A = \overline{Q} \begin{pmatrix} -x^T A x & \left(\frac{\overline{\lambda} Q_1^T r}{1+|\lambda|^2}\right)^T \\ \frac{Q_1^T r}{1+|\lambda|^2} & A_1 \end{pmatrix} Q^H.$$

Setting $A_1 = 0$, we obtain

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = |||\Delta\mathrm{L}||| = \sqrt{2|x^T A x|^2 + \frac{2\|[I - \overline{x}x^T]r\|_2^2}{1+|\lambda|^2}} = \sqrt{2}\sqrt{|x^T A x|^2 + \frac{\|r\|_2^2 - |x^T r|^2}{1+|\lambda|^2}}$$

from which the result follows. Since $|x^T r|^2 = |x^T A x|^2(1+|\lambda|^2)$ when $\lambda \in i\mathbb{R}$, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{2}\|r\|_2/\|(1,\lambda)\|_2$, for $\lambda \in i\mathbb{R}$. Again, simplifying the expression for ΔA, we obtain $\Delta A = -\overline{x}x^T A x x^H + \frac{1}{1+|\lambda|^2}[\overline{\lambda}\overline{x}r^T(I - xx^H) + (I - \overline{x}x^T)rx^H] + \overline{Q}_1 A_1 Q_1^H$. This completes the proof. □

Observe that from Theorem 3.5 we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \sqrt{2}\,\eta(\lambda, x, \mathrm{L})$ when $\mathsf{re}\lambda > 0$ and $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \|(1, \sqrt{|\mathsf{re}\lambda|}/|1+\lambda|)\|_2\,\eta(\lambda, x, \mathrm{L})$ when $\lambda \neq -1$ and $\mathsf{re}\lambda < 0$.

Note that if $Y \in \mathbb{C}^{n\times n}$ is such that $Yx = 0$ and $Y^T x = 0$, then $Y = (I - xx^H)^T Z (I - xx^H)$ for some matrix Z. Hence, from the proof of Theorem 3.5, we obtain an analogue of Corollary 3.2 for T-palindromic pencil. Indeed, if K is a T-palindromic pencil such that $\mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0$, then $\mathrm{K}(z) = \Delta\mathrm{L}(z) + (I - xx^H)^T \mathrm{N}(z)(I - xx^H)$ for some T-palindromic pencil N, where $\Delta\mathrm{L}$ is given in Theorem 3.5.

Now we turn to H-Hermitian, H-skew-Hermitian, H-even, H-odd, and H-palindromic pencils.

THEOREM 3.6. *Let* $\mathbb{S} \in \{$*H-Hermitian, H-skew-Hermitian*$\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zB$. *For* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$, *set* $r := -\mathrm{L}(\lambda)x$. *Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \dfrac{\sqrt{2\|r\|_2^2 - |x^H r|^2}}{\|(1, \lambda)\|_2} \leq \sqrt{2}\,\eta(\lambda, x, \mathrm{L}) & \text{if } \lambda \in \mathbb{R}, \\ \sqrt{|x^H A x|^2 + |x^H B x|^2 + \dfrac{2\|r\|_2^2 - 2|x^H r|^2}{1 + |\lambda|^2}}, & \text{if } \lambda \in \mathbb{C} \setminus \mathbb{R}. \end{cases}$$

In particular, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \|r\|_2 = \sqrt{2}\ \eta(\lambda, x, \mathrm{L})$, *if* $\lambda = \pm i$.

When $\lambda \in \mathbb{R}$, *define*

$$\Delta A := \begin{cases} \frac{1}{1+\lambda^2}\left[x r^H + r x^H - \left(r^H x\right) x x^H\right], & \text{if } A = A^H \\ \frac{1}{1+\lambda^2}\left[r x^H - x r^H + \left(r^H x\right) x x^H\right], & \text{if } A = -A^H \end{cases}$$

$$\Delta B := \begin{cases} \frac{\lambda}{1+\lambda^2}\left[x r^H + r x^H - \left(r^H x\right) x x^H\right], & \text{if } B = B^H \\ \frac{\lambda}{1+\lambda^2}\left[r x^H - x r^H + \left(r^H x\right) x x^H\right], & \text{if } B = -B^H \end{cases}$$

and consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L} \in \mathbb{S}$, $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$, *and* $\|\Delta\mathrm{L}\| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

When $\lambda \in \mathbb{C} \setminus \mathbb{R}$, *the H-Hermitian/H-skew-Hermitian pencil* $\Delta\mathrm{L}$ *given in Theorem* 2.2 *satisfies* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ *and* $\|\Delta\mathrm{L}\| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. Suppose that $\mathrm{L}(z) = A + zB$ is H-Hermitian so that $A = A^H$ and $B = B^H$. By Theorem 2.2 there exists H-Hermitian pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$ such that $(\Delta A + \lambda\Delta B)x = r$. Again, choosing a unitary matrix $Q := [x,\ Q_1]$, we have

$$\begin{aligned} \widetilde{\Delta A} &:= Q^H \Delta A Q = \begin{pmatrix} a_{11} & a_1^H \\ a_1 & A_1 \end{pmatrix}, \\ \widetilde{\Delta B} &:= Q^H \Delta B Q = \begin{pmatrix} b_{11} & b_1^H \\ b_1 & B_1 \end{pmatrix}, \\ Q^H r &= \begin{pmatrix} x^H r \\ Q_1^H r \end{pmatrix}, \end{aligned}$$

where $A_1 = A_1^H$ and $B_1 = B_1^H$ are of size $n - 1$. This gives

$$(\widetilde{\Delta A} + \lambda\widetilde{\Delta B})Q^H x = \begin{pmatrix} x^H r \\ Q_1^H r \end{pmatrix} \Rightarrow \begin{pmatrix} a_{11} + \lambda b_{11} \\ a_1 + \lambda b_1 \end{pmatrix} = \begin{pmatrix} x^H r \\ Q_1^H r \end{pmatrix}.$$

The minimum norm solution of $a_1 + \lambda b_1 = Q_1^H r$ is given by

$$(a_1 \ \ b_1) = Q_1^H r \begin{pmatrix} 1 \\ \lambda \end{pmatrix}^{\dagger} \Rightarrow a_1 = \frac{1}{1 + |\lambda|^2} Q_1^H r, \quad b_1 = \frac{\overline{\lambda}}{1 + |\lambda|^2} Q_1^H r.$$

For the equation $a_{11} + \lambda b_{11} = x^H r$, two cases arise.

Case-I: When $\lambda \in \mathbb{R}$, the minimum norm solution is given by

$$(a_{11} \ \ b_{11}) = x^H r \begin{pmatrix} 1 \\ \lambda \end{pmatrix}^{\dagger} \Rightarrow a_{11} = \frac{1}{1 + \lambda^2} x^H r \in \mathbb{R}, \quad b_{11} = \frac{\lambda}{1 + \lambda^2} x^H r \in \mathbb{R}.$$

Hence, we have

$$\Delta A = Q \left(\begin{array}{cc} \frac{1}{1+\lambda^2} x^H r & \frac{1}{1+\lambda^2} \left(Q_1^H r\right)^H \\ \frac{1}{1+\lambda^2} Q_1^H r & A_1 \end{array} \right) Q^H,$$

$$\Delta B = Q \left(\begin{array}{cc} \frac{\lambda}{1+\lambda^2} x^H r & \left(\frac{\lambda}{1+\lambda^2} Q_1^H r\right)^H \\ \frac{\lambda}{1+\lambda^2} Q_1^H r & B_1 \end{array} \right) Q^H.$$

Setting $A_1 = B_1 = 0$ and using the fact that $Q_1 Q_1^H = I - xx^H$, we have

$$\eta^{\mathbb{S}}(x, \lambda, \mathrm{L}) = |||\Delta \mathrm{L}||| = \frac{\sqrt{2\|r\|_2^2 - |x^H r|^2}}{\|(1, \lambda)\|_2}.$$

Now simplifying the expressions for ΔA and ΔB, we have

$$\begin{aligned} \Delta A &= \frac{1}{1+\lambda^2} xx^H r x^H + \frac{1}{1+\lambda^2} \left[x r^H Q_1 Q_1^H + Q_1 Q_1^H r x^H \right] + Q_1 A_1 Q_1^H \\ &= \frac{1}{1+\lambda^2} \left[x r^H + r x^H - \left(r^H x\right) x x^H \right] + Q_1 A_1 Q_1^H, \\ \Delta B &= \frac{\lambda}{1+\lambda^2} xx^H r x^H + \frac{\lambda}{1+\lambda^2} \left[x r^H Q_1 Q_1^H + Q_1 Q_1^H r x^H \right] + Q_1 B_1 Q_1^H \\ &= \frac{\lambda}{1+\lambda^2} \left[x r^H + r x^H - \left(r^H x\right) x x^H \right] + Q_1 B_1 Q_1^H. \end{aligned}$$

Hence, the results follow.

Case-II: Suppose that $\lambda \in \mathbb{C} \setminus \mathbb{R}$. Then by Theorem 2.1, we have $x^H(A + \Delta A)x = 0$ and $x^H(B + \Delta B)x = 0$. Hence, we have $a_{11} = -x^H A x$ and $b_{11} = -x^H B x$. Consequently,

$$\Delta A = Q \left(\begin{array}{cc} -x^H A x & \left(\frac{1}{1+|\lambda|^2} Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2} Q_1^H r & A_1 \end{array} \right) Q^H,$$

$$\Delta B = Q \left(\begin{array}{cc} -x^H B x & \left(\frac{\lambda}{1+|\lambda|^2} Q_1^H r\right)^H \\ \frac{\lambda}{1+|\lambda|^2} Q_1^H r & B_1 \end{array} \right) Q^H.$$

Setting $A_1 = B_1 = 0$, we obtain

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = |||\Delta \mathrm{L}||| = \sqrt{|x^H A x|^2 + |x^H B x|^2 + \frac{2\|(I - xx^H) r\|_2^2}{1 + |\lambda|^2}}.$$

Hence, the result follows.

Now simplifying the expressions for ΔA and ΔB, we have

$$\Delta A = -xx^H A x x^H + \frac{1}{1+|\lambda|^2} \left[x r^H \left(I - xx^H\right) + \left(I - xx^H\right) r x^H \right] + Q_1 A_1 Q_1^H,$$

$$\Delta B = -xx^H B x x^H + \frac{1}{1+|\lambda|^2} \left[\lambda x r^H \left(I - xx^H\right) + \overline{\lambda} \left(I - xx^H\right) r x^H \right] + Q_1 B_1 Q_1^H.$$

Setting $A_1 = B_1 = 0$, we obtain the H-Hermitian pencil $\Delta \mathrm{L}$ given in Theorem 2.2.

The proof is similar for the case when L is H-skew-Hermitian. □

Needless to mention that an analogue of Corollary 3.2 holds for H-Hermitian/H-skew-Hermitian pencils.

THEOREM 3.7. *Let* $\mathbb{S} \in \{H\text{-even, } H\text{-odd}\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) = A + zB$. *For* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$, *set* $r := -\mathrm{L}(\lambda)x$. *Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \dfrac{\sqrt{2\|r\|_2^2 - |x^H r|^2}}{\|(1,\,\lambda)\|_2} \leq \sqrt{2}\,\eta(\lambda, x, \mathrm{L}) & \text{if } \lambda \in i\mathbb{R}, \\ \sqrt{|x^H A x|^2 + |x^H B x|^2 + \dfrac{2\|r\|_2^2 - 2|x^H r|^2}{1 + |\lambda|^2}}, & \text{if } \lambda \in \mathbb{C} \setminus i\mathbb{R}. \end{cases}$$

In particular, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \|r\|_2 = \sqrt{2}\ \eta(\lambda, x, \mathrm{L})$, *if* $\lambda = \pm 1$.

When $\lambda \in i\mathbb{R}$, *define*

$$\Delta A := \begin{cases} \frac{1}{1+|\lambda|^2}\left[x r^H + r x^H - \left(r^H x\right) x x^H\right], & \text{if } A = A^H \\ \frac{1}{1+|\lambda|^2}\left[r x^H - x r^H + \left(r^H x\right) x x^H\right], & \text{if } A = -A^H \end{cases}$$

$$\Delta B := \begin{cases} \frac{-\lambda}{1+|\lambda|^2}\left[r x^H - x r^H + \left(r^H x\right) x x^H\right], & \text{if } B = B^H \\ \frac{-\lambda}{1+|\lambda|^2}\left[r x^H + x r^H - \left(r^H x\right) x x^H\right], & \text{if } B = -B^H \end{cases}$$

and consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L} \in \mathbb{S}$, $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$, *and* $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

When $\lambda \in \mathbb{C} \setminus i\mathbb{R}$, *the H-even/H-odd pencil* $\Delta\mathrm{L}$ *given in Theorem* 2.2 *satisfies* $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ *and* $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. First, suppose that $\mathrm{L}(z) = A + zB$ is H even. Then $A = A^H$ and $B = -B^H$. By Theorem 2.2 there exists H-even pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$ such that $\Delta\mathrm{L}(\lambda)x = r$. Now choosing a unitary matrix $Q := [x, Q_1]$ and noting that $\Delta A = \Delta A^H,\ \Delta B = -\Delta B^H$, we have

$$\Delta A := Q \begin{pmatrix} a_{11} & a_1^H \\ a_1 & A_1 \end{pmatrix} Q^H \text{ and } \Delta B = Q \begin{pmatrix} b_{11} & b_1^H \\ -b_1 & B_1 \end{pmatrix} Q^H,$$

where $A_1 = A_1^H$ and $B_1 = -B_1^H$ are matrices of size $n-1$. Then $\Delta\mathrm{L}(\lambda)x = r$ gives $\begin{pmatrix} a_{11} + \lambda b_{11} \\ a_1 - \lambda b_1 \end{pmatrix} = \begin{pmatrix} x^H r \\ Q_1^H r \end{pmatrix}$. The minimum norm solution of $a_1 - \lambda b_1 = Q_1^H r$ is given by

$$(a_1 \;\; b_1) = Q_1^H r \begin{pmatrix} 1 \\ -\lambda \end{pmatrix}^{\dagger} \Rightarrow a_1 = \frac{1}{1+|\lambda|^2} Q_1^H r, \;\; b_1 = -\frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^H r.$$

For the solution of $a_{11} + \lambda b_{11} = x^H r$ two cases arise. When $\lambda \in i\mathbb{R}$, the minimum norm solution is given by

$$(a_{11} \;\; b_{11}) = x^H r \begin{pmatrix} 1 \\ \lambda \end{pmatrix}^{\dagger} \Rightarrow a_{11} = \frac{1}{1+|\lambda|^2} x^H r \in \mathbb{R}, \;\; b_{11} = \frac{\overline{\lambda}}{1+|\lambda|^2} x^H r \in i\mathbb{R}.$$

When $\lambda \in \mathbb{C} \setminus i\mathbb{R}$, by Theorem 2.1, $x^H(A + \Delta A)x = 0 = x^H(B + \Delta B)x \Rightarrow a_{11} =$

$-x^HAx$ and $b_{11} = -x^HBx$. Consequently, we have

$$\Delta A = Q \left(\begin{array}{cc} \frac{1}{1+|\lambda|^2} x^H r & \left(\frac{1}{1+|\lambda|^2} Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2} Q_1^H r & A_1 \end{array} \right) Q^H,$$

$$\Delta B = Q \left(\begin{array}{cc} \frac{\overline{\lambda}}{1+|\lambda|^2} x^H r & \left(-\frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^H r\right)^H \\ \frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^H r & B_1 \end{array} \right) Q^H$$

when $\lambda \in i\mathbb{R}$ and

$$\Delta A = Q \left(\begin{array}{cc} -x^H Ax & \left(\frac{1}{1+|\lambda|^2} Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2} Q_1^H r & A_1 \end{array} \right) Q^H,$$

$$\Delta B = Q \left(\begin{array}{cc} -x^H Bx & \left(-\frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^H r\right)^H \\ \frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^H r & B_1 \end{array} \right) Q^H$$

when $\lambda \in \mathbb{C} \setminus i\mathbb{R}$. Hence, the desired results follow. Finally, reversing the role of A and B we obtain the results for the case when $\mathrm{L}(z) = A + zB$ is H-odd. □

We have the following result for H-palindromic pencils.

THEOREM 3.8. *Let $\mathbb{S}$ be the space of H-palindromic pencils and $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) = A + zA^H$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{2}\,\sqrt{|x^HAx|^2 + \dfrac{\|r\|_2^2 - |x^Hr|^2}{1+|\lambda|^2}} & \text{if } |\lambda| \neq 1, \\ \sqrt{\|r\|_2^2 - \frac{1}{2}|x^Hr|^2}, & \text{if } |\lambda| = 1. \end{cases}$$

Now define

$$\Delta A := \begin{cases} \frac{1}{1+|\lambda|^2}\left[rx^H + \lambda x r^H \left(I - xx^H\right)\right], & \text{if } |\lambda| = 1, \\ -xx^HAxx^H + \frac{1}{1+|\lambda|^2}\left[\lambda x r^H\left(I - xx^H\right) + \left(I - xx^H\right) r x^H\right], & \text{if } |\lambda| \neq 1, \end{cases}$$

and consider $\Delta\mathrm{L}(z) := \Delta A + z(\Delta A)^H$. Then $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ and $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. Let $Q := [x, \;\; Q_1]$ be unitary. Then $\widetilde{\Delta A} := Q^H\Delta A Q = \left(\begin{smallmatrix} a_{11} & a_1^H \\ b_1 & A_1 \end{smallmatrix}\right)$ and $Q^Hr = \left(\begin{smallmatrix} x^Hr \\ Q_1^Hr \end{smallmatrix}\right)$. Hence, $\Delta\mathrm{L}(\lambda)x = r$ gives $\left(\begin{smallmatrix} \lambda a_{11}^H + a_{11} \\ \lambda a_1 + b_1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} x^Hr \\ Q_1^Hr \end{smallmatrix}\right)$. If $|\lambda| \neq 1$, then by Theorem 2.1, we have $x^H(\Delta A + A)x = 0 \Rightarrow x^H\Delta Ax = -x^HAx$. Hence, we have $a_{11} = -x^HAx$. On the other hand, when $|\lambda| = 1$, the minimum norm solution is given by

$$(\overline{a}_{11} \;\; a_{11}) = x^Hr \left(\begin{array}{c} \lambda \\ 1 \end{array} \right)^{\dagger} = \left(\frac{\overline{\lambda}x^Hr}{1+|\lambda|^2} \;\; \frac{x^Hr}{1+|\lambda|^2} \right).$$

Note that when $|\lambda| = 1$ we have $\overline{x^Hr} = \overline{\lambda}x^Hr$. Next, the minimum solution of $a_1\lambda + b_1 = Q_1^Hr$ is given by $(a_1, \;\; b_1) = Q_1^Hr\left(\begin{smallmatrix} \lambda \\ 1 \end{smallmatrix}\right)^{\dagger} = \left(\begin{smallmatrix} \frac{\overline{\lambda}Q_1^Hr}{1+|\lambda|^2} & \frac{Q_1^Hr}{1+|\lambda|^2} \end{smallmatrix}\right)$. Consequently, when

$|\lambda| \neq 1$, we have

$$\Delta A = Q \begin{pmatrix} -x^H A x & \left(\frac{\overline{\lambda} Q_1^H r}{1+|\lambda|^2}\right)^H \\ \frac{Q_1^H r}{1+|\lambda|^2} & A_1 \end{pmatrix} Q^H.$$

Setting $A_1 = 0$, we obtain

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = |||\Delta \mathrm{L}||| = \sqrt{2|x^H A x|^2 + \frac{2\|[I - xx^H]r\|_2^2}{1+|\lambda|^2}} = \sqrt{2}\,\sqrt{|x^H A x|^2 + \frac{\|r\|_2^2 - |r^H x|^2}{1+|\lambda|^2}}.$$

Using the fact that $Q_1 Q_1^H = I - xx^H$, we have

$$\Delta A = -xx^H A x x^H + \frac{1}{1+|\lambda|^2}\left[\lambda x r^H \left(I - xx^H\right) + \left(I - xx^H\right) r x^H\right] + Q_1 A_1 Q_1^H.$$

Setting $A_1 = 0$, the result follows.

For the case when $|\lambda| = 1$, we have

$$\Delta A = Q \begin{pmatrix} \frac{x^H r}{1+|\lambda|^2} & \left(\frac{\overline{\lambda} Q_1^H r}{1+|\lambda|^2}\right)^H \\ \frac{Q_1^H r}{1+|\lambda|^2} & A_1 \end{pmatrix} Q^H.$$

Again, setting $A_1 = 0$ we obtain

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = |||\Delta \mathrm{L}||| = \sqrt{\|r\|_2^2 - \frac{1}{2}|x^H r|^2}.$$

Since $Q_1 Q_1^H = (I - xx^H)$, simplifying the expression for ΔA, we obtain

$$\Delta A := \frac{1}{1+|\lambda|^2}\left[r x^H + \lambda x r^H \left(I - xx^H\right)\right] + Q_1 A_1 Q_1^H.$$

Hence, the proof. □

Remark 3.9. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ with $x^H x = 1$ and $\mathbb{S} \in \{T$-symmetric, T-skew-symmetric, T-odd, T-even, T-palindromic, H-Hermitian, H-skew-Hermitian, H-odd, H-even, H-palindromic$\}$. For $\mathrm{L} \in \mathbb{S}$, consider the set

$$\mathbb{S}(\lambda, x, \mathrm{L}) := \{\mathrm{K} \in \mathbb{S} : \mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0\}.$$

Then $\mathbb{S}(\lambda, x, \mathrm{L}) \neq \emptyset$ and there exists a unique $\Delta \mathrm{L} \in \mathbb{S}(\lambda, x, \mathrm{L})$ such that

$$\min\{|||\mathrm{K}||| : \mathrm{K} \in \mathbb{S}(\lambda, x, \mathrm{L})\} = |||\Delta \mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L}).$$

Further, each pencil in $\mathbb{S}(\lambda, x, \mathrm{L})$ is of the form $\Delta \mathrm{L} + (I - xx^H)^* \, \mathrm{Z}(I - xx^H)$ for some $\mathrm{Z} \in \mathbb{S}$, where $*$ is either the transpose or the conjugate transpose depending upon the structure defined by $\mathbb{S}$. In other words, we have $\mathbb{S}(\lambda, x, \mathrm{L}) = \Delta \mathrm{L} + (I - xx^H)^* \, \mathbb{S}\,(I - xx^H)$.

We mention that the results obtained above are easily extended to the case of pencils having more general structures. Indeed, let M be a unitary matrix such that $M^T = M$ or $M^T = -M$. Consider the Jordan algebra $\mathbb{J} := \{A \in \mathbb{C}^{n \times n} : M^{-1} A^T M = A\}$ and the Lie algebra $\mathbb{L} := \{A \in \mathbb{C}^{n \times n} : M^{-1} A^T M = -A\}$ associated with the scalar product $(x, y) \mapsto y^T M x$. Consider a pencil $\mathrm{L}(z) = A + zB$, where A and B

are in $\mathbb{J}$ and/or in $\mathbb{L}$. Then the pencil $M\mathrm{L}$ given by $M\mathrm{L}(z) = MA + zMB$ is either T-symmetric, T-skew-symmetric, T-even, or T-odd. Hence, replacing A, B, and r by MA, MB, and Mr, respectively, in the above results, we obtain corresponding results for the pencil L.

Similarly, when M is unitary and $M = M^H$ or $M = -M^H$, we consider the Jordan algebra $\mathbb{J} := \{A \in \mathbb{C}^{n\times n} : M^{-1}A^HM = A\}$ and the Lie algebra $\mathbb{L} := \{A \in \mathbb{C}^{n\times n} : M^{-1}A^HM = -A\}$ associated with the scalar product $(x, y) \mapsto y^HMx$. Now, let $\mathrm{L}(z) = A + zB$ be a pencil where A and B are in $\mathbb{J}$ and/or in $\mathbb{L}$. Then the pencil $M\mathrm{L}(z) = MA + zMB$ is either H-Hermitian, H-skew-Hermitian, H-even, or H-odd. Hence, replacing A, B, and r by MA, MB, and Mr, respectively, in the above results, we obtain corresponding results for the pencil L. In particular, when $M := J$, where $J := \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \in \mathbb{C}^{2n\times 2n}$, the Jordan algebra $\mathbb{J}$ consists of skew-Hamiltonian matrices and the Lie algebra $\mathbb{L}$ consists of Hamiltonian matrices. So, for example, considering the pencil $\mathrm{L}(z) := A+zB$, where A is Hamiltonian and B is skew-Hamiltonian, we see that the pencil $J\mathrm{L}(z) = JA + zJB$ is H-even. Hence, extending the results obtained for H-even pencil to the case of L, we have the following.

THEOREM 3.10. *Let $\mathbb{S}$ be the space of pencils of the form $\mathrm{L}(z) = A + zB$, where A is Hamiltonian and B is skew-Hamiltonian. For $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$, set $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \dfrac{\sqrt{2\|r\|_2^2 - |x^HJr|^2}}{\|(1,\ \lambda)\|_2} \leq \sqrt{2}\,\eta(\lambda, x, \mathrm{L}) & \text{if } \lambda \in i\mathbb{R}, \\ \sqrt{|x^HJAx|^2 + |x^HJBx|^2 + \dfrac{2\|r\|_2^2 - 2|x^HJr|^2}{1 + |\lambda|^2}}, & \text{if } \lambda \in \mathbb{C} \setminus i\mathbb{R}. \end{cases}$$

We mention that Remark 3.9 remains valid for structured pencils in $\mathbb{S}$ whose coefficient matrices are elements of Jordan and/or Lie algebras associated with a scalar product considered above. In such a case the $*$ in $(I - xx^H)^*$ is the adjoint induced by the scalar product that defines the Jordan and Lie algebras.

4. Spectral norm and structured backward errors. Considering Frobenius norm on $\mathbb{C}^{n\times n}$, in the previous section, we have obtained structured backward error of an approximate eigenpair. In this section, we derive structured backward errors when $\mathbb{C}^{n\times n}$ is equipped with the spectral norm. Recall that the norm of a pencil $\mathrm{L}(z) = A + zB$ as defined in (2.1) is then given by $|||\mathrm{L}||| := (\|A\|_2^2 + \|B\|_2^2)^{1/2}$. Derivations of structured backward errors of approximate eigenpairs turn out to be much more difficult when $\mathbb{C}^{n\times n}$ is equipped with the spectral norm than in the case when $\mathbb{C}^{n\times n}$ is equipped with the Frobenius norm. We mention that for certain structures (e.g., T-symmetric and T-skew-symmetric) it is indeed possible to use structured mapping theorems given in [18, 2] to derive structured backward errors of approximate eigenpairs. However, for most structures (e.g., even, odd, palindromic, Hermitian, skew-Hermitian), the structured mapping theorems are not of much help for deriving structured backward errors. We overcome this difficulty by employing Davis–Kahan–Weinberger solutions of norm preserving dilation problem for Hilbert space operators.

The Davis–Kahan–Weinberger (DKW, in short) solutions of norm-preserving dilations of matrices can be stated as follows (for a more general version of the DKW theorem, see [9]).

THEOREM 4.1 (Davis–Kahan–Weinberger, [9]). *Let A, B, C satisfy $\left\|\begin{pmatrix} A \\ B \end{pmatrix}\right\|_2 = \mu$ and $\left\|\begin{pmatrix} A & C \end{pmatrix}\right\|_2 = \mu$. Then there exists D such that $\left\|\begin{pmatrix} A & C \\ B & D \end{pmatrix}\right\|_2 = \mu$. Indeed, those D*

which have this property are exactly those of the form

$$D = -KA^HL + \mu\left(I - KK^H\right)^{1/2} Z \left(I - L^HL\right)^{1/2},$$

where $K^H := (\mu^2\mathrm{I} - A^HA)^{-1/2}B^H,\ L := (\mu^2\mathrm{I} - AA^H)^{-1/2}C$, *and* Z *is an arbitrary contraction, that is,* $\|Z\|_2 \leq 1$.

We now use the DKW theorem with $Z = 0$ and derive structured backward error of an approximate eigenpair. Recall that for $(\lambda, x) \in \mathbb{C}\times\mathbb{C}^n$, our standing assumption is that $x^Hx = 1$.

THEOREM 4.2. *Let* $\mathbb{S} \in \{T$*-symmetric,* T*-skew-symmetric*$\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) := A + zB$. *Let* $(\lambda, x) \in \mathbb{C}\times\mathbb{C}^n$ *and* $r := -\mathrm{L}(\lambda)x$. *Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \frac{\|r\|_2}{\|(1,\lambda)\|_2} = \eta(\lambda, x, \mathrm{L}).$$

Now define

$$\Delta A := \begin{cases} \frac{1}{1+|\lambda|^2}\left[\overline{x}r^T + rx^H - \left(r^Tx\right)\overline{x}x^H - \frac{\overline{x^Tr}\ \left(I-\overline{x}x^T\right)rr^T\left(I-xx^H\right)}{\|r\|_2^2-|x^Tr|^2}\right], & \textit{if } A = A^T, \\ -\frac{1}{1+|\lambda|^2}\left[\overline{x}r^T - rx^H\right], & \textit{if } A = -A^T. \end{cases}$$

$$\Delta B := \begin{cases} \frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T + rx^H - \left(r^Tx\right)\overline{x}x^H - \frac{\overline{x^Tr}\ \left(I-\overline{x}x^T\right)rr^T\left(I-xx^H\right)}{\|r\|_2^2-|x^Tr|^2}\right], & \textit{if } B = B^T, \\ -\frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T - rx^H\right], & \textit{if } B = -B^T, \end{cases}$$

and consider the pencil $\Delta\mathrm{L}(z) := \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L} \in \mathbb{S},\ \mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$, *and* $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. Suppose that L is T-symmetric. Then from the proof of Theorem 3.1, we have

$$\Delta A = \overline{Q}\begin{pmatrix} \frac{x^Tr}{1+|\lambda|^2} & \frac{1}{1+|\lambda|^2}\left(Q_1^Tr\right)^T \\ \frac{1}{1+|\lambda|^2}\left(Q_1^Tr\right) & A_1 \end{pmatrix} Q^H,$$
$$\Delta B = \overline{Q}\begin{pmatrix} \frac{\overline{\lambda}\ x^Tr}{1+|\lambda|^2} & \frac{\overline{\lambda}}{1+|\lambda|^2}\left(Q_1^Tr\right)^T \\ \frac{\overline{\lambda}}{1+|\lambda|^2}\left(Q_1^Tr\right) & B_1 \end{pmatrix} Q^H,$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. Now, for $\mu_{\Delta A} := \frac{\|r\|_2}{1+|\lambda|^2}$ and $\mu_{\Delta B} := \frac{|\lambda|\ \|r\|_2}{1+|\lambda|^2}$, by the DKW Theorem 4.1, we have $A_1 = -\frac{\overline{x^Tr}\,(Q_1^Tr)(Q_1^Tr)^T}{(1+|\lambda|^2)\,(\|r\|_2^2-|x^Tr|^2)}$ and $B_1 = -\frac{\overline{\lambda}\,\overline{x^Tr}\,(Q_1^Tr)(Q_1^Tr)^T}{(1+|\lambda|^2)\,(\|r\|_2^2-|x^Tr|^2)}$. This gives $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = (\|\Delta A\|_2^2 + \|\Delta B\|_2^2)^{1/2} = \frac{\|r\|_2}{\|(1,\lambda)\|_2}$. Simplifying expressions for ΔA and ΔB, we obtain the desired results.

When L is T-skew-symmetric, from the proof of Theorem 3.3, we have

$$\Delta A = \overline{Q}\begin{pmatrix} 0 & -\frac{\left(Q_1^Tr\right)^T}{1+|\lambda|^2} \\ \frac{1}{1+|\lambda|^2}Q_1^Tr & A_1 \end{pmatrix} Q^H,\ \Delta B = \overline{Q}\begin{pmatrix} 0 & -\frac{\overline{\lambda}\left(Q_1^Tr\right)^T}{1+|\lambda|^2} \\ \frac{\overline{\lambda}}{1+|\lambda|^2}Q_1^Tr & B_1 \end{pmatrix} Q^H,$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. Now, for $\mu_{\Delta A} := \frac{\|r\|_2}{1+|\lambda|^2}$ and $\mu_{\Delta B} := \frac{|\lambda|\ \|r\|_2}{1+|\lambda|^2}$, by the DKW Theorem 4.1, we obtain $A_1 = 0 = B_1$. Consequently, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) =$

$(\|\Delta A\|_2^2 + \|\Delta B\|_2^2)^{1/2} = \|r\|_2/\|(1,\lambda)\|_2$. Simplifying the expressions for ΔA and ΔB, we obtain the desired results. □

Remark 4.3. If $|x^T r| = \|r\|_2$, then $\|Q_1^T r\|_2 = 0$. In such a case, considering $A_1 = 0 = B_1$ we obtain the desired results.

Next, we consider T-even and T-odd pencils. Recall that for $z \in \mathbb{C}$, $\mathrm{sign}(z) := \overline{z}/|z|$ when $z \neq 0$ and $\mathrm{sign}(z) := 1$ when $z = 0$.

THEOREM 4.4. *Let* $\mathbb{S} \in \{T\text{-}even,\ T\text{-}odd\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) := A + zB$. *Let* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ *and* $r := -\mathrm{L}(\lambda)x$. *Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{|x^T Ax|^2 + \dfrac{\|r\|_2^2 - |x^T r|^2}{1+|\lambda|^2}} = \dfrac{\sqrt{\|r\|_2^2 + |\lambda|^2 |x^T r|^2}}{\|(1,\,\lambda)\|_2}, & \textit{if } \mathrm{L} \textit{ is } T\textit{-even},\\ \sqrt{|x^T Bx|^2 + \frac{\|r\|_2^2 - |x^T r|^2}{1+|\lambda|^2}} = \dfrac{\sqrt{\|r\|_2^2 + |\lambda|^{-2}\,|x^T r|^2}}{\|(1,\,\lambda)\|_2}, & \textit{if } \mathrm{L} \textit{ is } T\textit{-odd}, \lambda \neq 0,\\ \eta(\lambda, x, \mathrm{L}), & \textit{if } \mathrm{L} \textit{ is } T\textit{-odd}, \lambda = 0. \end{cases}$$

Now, define

$$\Delta A := \begin{cases} -\,\overline{x}x^T Axx^H + \frac{1}{1+|\lambda|^2}\left[\overline{x}r^T + rx^H - 2\left(x^T r\right)\overline{x}x^H\right] \\ +\, \frac{\overline{x^T Ax}\ \left(I - \overline{x}x^T\right) rr^T \left(I - xx^H\right)}{\|r\|_2^2 - |x^T r|^2}, \qquad \textit{if } A = A^T, \\ \frac{1}{1+|\lambda|^2}\left[rx^H - \overline{x}r^T\right],\ \textit{if } A = -A^T. \end{cases}$$

$$\Delta B := \begin{cases} -\,\overline{x}x^T Bxx^H + \frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T + rx^H - 2\left(x^T r\right)\overline{x}x^H\right] \\ -\, \frac{\mathrm{sign}(\lambda)^2\, \overline{x^T Bx}\ \left(I - \overline{x}x^T\right) rr^T \left(I - xx^H\right)}{\|r\|_2^2 - |x^T r|^2}, \qquad \textit{if } B = B^T, \\ -\, \frac{\overline{\lambda}}{1+|\lambda|^2}\left[\overline{x}r^T - rx^H\right],\ \textit{if } B = -B^T, \end{cases}$$

and consider the pencil $\Delta\mathrm{L}(z) := \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L} \in \mathbb{S},\ \mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ *and* $\|\Delta\mathrm{L}\| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. Suppose that L is T-even. Then from the proof of Theorem 3.4, we have

$$\Delta A = \overline{Q}\begin{pmatrix} -x^T Ax & \frac{(Q_1^T r)^T}{1+|\lambda|^2} \\ \frac{Q_1^T r}{1+|\lambda|^2} & A_1 \end{pmatrix} Q^H,\ \Delta B = \overline{Q}\begin{pmatrix} 0 & -\frac{\overline{\lambda}}{1+|\lambda|^2}\left(Q_1^T r\right)^T \\ \frac{\overline{\lambda}}{1+|\lambda|^2}\left(Q_1^T r\right) & B_1 \end{pmatrix} Q^H,$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. Now, for

$$\mu_{\Delta A} := \sqrt{|x^T Ax|^2 + \frac{\|r\|_2^2 - |x^T r|^2}{(1+|\lambda|^2)^2}} \text{ and } \mu_{\Delta B} := \sqrt{\frac{|\lambda|^2(\|r\|_2^2 - |x^T r|^2)}{(1+|\lambda|^2)^2}}$$

by the DKW Theorem 4.1, we have $A_1 = \frac{\overline{x^T Ax}}{\|r\|_2^2 - |x^T r|^2}(Q_1^T r)(Q_1^T r)^T$ and $B_1 = 0$. This gives

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{|x^T Ax|^2 + \frac{\|r\|_2^2 - |x^T r|^2}{1+|\lambda|^2}} = \frac{\sqrt{\|r\|_2^2 + |\lambda|^2|x^T r|^2}}{\|(1,\,\lambda)\|_2}.$$

Simplifying the expressions for ΔA and ΔB, we obtain the desired results. When L is T-odd, the results follow by interchanging the role of A and B. □

It follows that for a T-even pencil we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \|(1, \lambda)\|_2 \, \eta(\lambda, x, \mathrm{L})$ whereas for a T-odd pencil we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) \leq \|(1, \lambda^{-1})\|_2 \, \eta(\lambda, x, \mathrm{L})$ when $\lambda \neq 0$.

Theorem 4.5. *Let $\mathbb{S} \in \{H$-Hermitian, H-skew-Hermitian$\}$ and $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) := A + zB$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \eta(\lambda, x, \mathrm{L}), \quad \textit{if} \quad \lambda \in \mathbb{R}, \\ \sqrt{|x^H A x|^2 + |x^H B x|^2 + \dfrac{\|r\|_2^2 - |x^H r|^2}{1 + |\lambda|^2}}, \quad \textit{if } \lambda \in \mathbb{C} \setminus \mathbb{R}. \end{cases}$$

When $\lambda \in \mathbb{R}$, define

$$\Delta A := \begin{cases} \frac{1}{1+\lambda^2} \left[x r^H + r x^H - \left(r^H x\right) x x^H - \frac{x^H r \, \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2} \right], \ \textit{if } A = A^H, \\ \frac{1}{1+\lambda^2} [r x^H - x r^H + (r^H x) x x^H + \frac{r^H x (I - x x^H) r r^H (I - x x^H)}{\|r\|_2^2 - |x^H r|^2}], \quad \textit{if } A = -A^H. \end{cases}$$

$$\Delta B := \begin{cases} \frac{\lambda}{1+\lambda^2} \left[x r^H + r x^H - \left(r^H x\right) x x^H - \frac{x^H r \, \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2} \right], \quad \textit{if } B = B^H, \\ \frac{\lambda}{1+\lambda^2} \left[r x^H - x r^H + \left(r^H x\right) x x^H + \frac{r^H x \, \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2} \right], \ \textit{if } B = -B^H. \end{cases}$$

When $\lambda \in \mathbb{C} \setminus \mathbb{R}$, define

$$\Delta A := \begin{cases} - x x^H A x x^H + \frac{1}{1+|\lambda|^2} \left[x r^H \left(I - x x^H\right) + \left(I - x x^H\right) r x^H \right] \\ + \frac{x^H A x \, \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2}, \qquad \textit{if } A = A^H, \\ - x x^H A x x^H + \frac{1}{1+|\lambda|^2} \left[\left(I - x x^H\right) r x^H - x r^H \left(I - x x^H\right) \right] \\ + \frac{x^H A x \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2}, \qquad \textit{if } A = -A^H. \end{cases}$$

$$\Delta B := \begin{cases} - x x^H B x x^H + \frac{1}{1+|\lambda|^2} \left[\lambda x r^H \left(I - x x^H\right) + \overline{\lambda} \left(I - x x^H\right) r x^H \right] \\ + \frac{x^H B x \, \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2}, \qquad \textit{if } B = B^H, \\ - x x^H B x x^H - \frac{\lambda}{1+|\lambda|^2} x r^H \left(I - x x^H\right) + \frac{\overline{\lambda}}{1+|\lambda|^2} \left(I - x x^H\right) r x^H \\ + \frac{x^H B x \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{\|r\|_2^2 - |x^H r|^2}, \qquad \textit{if } B = -B^H. \end{cases}$$

Consider $\Delta \mathrm{L}(z) := \Delta A + z \Delta B$. Then $\Delta \mathrm{L} \in \mathbb{S}$, $\mathrm{L}(\lambda)x + \Delta \mathrm{L}(\lambda) x = 0$, and $\|\Delta \mathrm{L}\| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. First, suppose that L is H-Hermitian. Assume that $\lambda \in \mathbb{R}$. Then $x^H r \in \mathbb{R}$. Now from the proof of Theorem 3.6, we have

$$\Delta A = Q \begin{pmatrix} \frac{1}{1+\lambda^2} x^H r & \frac{1}{1+\lambda^2} \left(Q_1^H r\right)^H \\ \frac{1}{1+\lambda^2} Q_1^H r & A_1 \end{pmatrix} Q^H$$

and

$$\Delta B = Q \begin{pmatrix} \frac{\lambda}{1+\lambda^2} x^H r & \frac{\lambda}{1+\lambda^2} \left(Q_1^H r\right)^H \\ \frac{\lambda}{1+\lambda^2} Q_1^H r & B_1 \end{pmatrix} Q^H$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. For $\mu_{\Delta A} := \frac{\|r\|_2}{1+\lambda^2}$ and $\mu_{\Delta B} := \frac{|\lambda|\,\|r\|_2}{1+\lambda^2}$ by the DKW Theorem 4.1, we have $A_1 = -\frac{x^H r\ (Q_1^H r)(Q_1^H r)^H}{(1+\lambda^2)\ (\|r\|_2^2 - |x^H r|^2)}$, $B_1 = -\frac{\lambda\ x^H r\ (Q_1^H r)(Q_1^H r)^H}{(1+\lambda^2)\ (\|r\|_2^2 - |x^H r|^2)}$. This gives $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = (\|\Delta A\|_2^2 + \|\Delta B\|_2^2)^{1/2} = \frac{\|r\|_2}{\|(1,\lambda)\|_2}$. Now simplifying the expressions for ΔA and ΔB, we obtain the desired results.

Next, suppose that $\lambda \in \mathbb{C}\backslash\mathbb{R}$. Then again from the proof of Theorem 3.6, we have

$$\Delta A = Q \begin{pmatrix} -x^H A x & \frac{1}{1+|\lambda|^2}\left(Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2} Q_1^H r & A_1 \end{pmatrix} Q^H,$$
$$\Delta B = Q \begin{pmatrix} -x^H B x & \frac{\lambda}{1+|\lambda|^2}\left(Q_1^H r\right)^H \\ \frac{\overline{\lambda}}{1+|\lambda|^2} Q_1^H r & B_1 \end{pmatrix} Q^H.$$

For, $\mu_{\Delta A} := \sqrt{|x^H A x|^2 + \frac{\|r\|_2^2 - |x^H r|^2}{(1+|\lambda|^2)^2}}$, $\mu_{\Delta B} := \sqrt{|x^H B x|^2 + \frac{|\lambda|^2(\|r\|_2^2 - |x^H r|^2)}{(1+|\lambda|^2)^2}}$, by the DKW Theorem 4.1, we have

$$A_1 = \frac{x^H A x}{\|r\|_2^2 - |x^H r|^2}\left(Q_1^H r\right)\left(Q_1^H r\right)^H \text{ and } B_1 = \frac{x^H B x}{\|r\|_2^2 - |x^T r|^2}\left(Q_1^H r\right)\left(Q_1^H r\right)^H.$$

Hence, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{|x^H A x|^2 + |x^H B x|^2 + \frac{\|r\|_2^2 - |x^H r|^2}{1+|\lambda|^2}}$. Now, simplifying the expressions for ΔA and ΔB, we obtain the desired results. The proof is similar for the case when L is H-skew-Hermitian. □

We mention that when $Q_1^H r = 0$, the desired results follow by considering $A_1 = 0 = B_1$.

Theorem 4.6. *Let* $\mathbb{S} \in \{H$*-even*, H*-odd*$\}$ *and* $\mathrm{L} \in \mathbb{S}$ *be given by* $\mathrm{L}(z) := A + zB$. *Let* $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ *and* $r := -\mathrm{L}(\lambda)x$. *Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \eta(\lambda, x, \mathrm{L}), & \textit{if } \lambda \in i\mathbb{R}, \\ \sqrt{|x^H A x|^2 + |x^H B x|^2 + \dfrac{\|r\|_2^2 - |x^H r|^2}{1+|\lambda|^2}}, & \textit{if } \lambda \in \mathbb{C}\setminus i\mathbb{R}. \end{cases}$$

When $\lambda \in i\mathbb{R}$, *define*

$$\Delta A := \begin{cases} \frac{1}{1+|\lambda|^2}\left[xr^H + rx^H - \left(r^H x\right)xx^H - \frac{x^H r\ \left(I - xx^H\right)rr^H\left(I - xx^H\right)}{\|r\|_2^2 - |x^H r|^2}\right], & \textit{if } A = A^H, \\ \frac{1}{1+|\lambda|^2}\left[rx^H - xr^H + \left(r^H x\right)xx^H + \frac{r^H x\left(I - xx^H\right)rr^H\left(I - xx^H\right)}{\|r\|_2^2 - |x^H r|^2}\right], & \textit{if } A = -A^H. \end{cases}$$

$$\Delta B := \begin{cases} \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}rx^H + \lambda xr^H - \lambda\left(r^H x\right)xx^H + \frac{\lambda x^H r\ \left(I - xx^H\right)rr^H\left(I - xx^H\right)}{\|r\|_2^2 - |x^H r|^2}\right], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{if } B = B^H, \\ \frac{1}{1+|\lambda|^2}\Big[\overline{\lambda}xx^H rx^H - \lambda xr^H\left(I - xx^H\right) + \overline{\lambda}\left(I - xx^H\right)rx^H \\ \qquad + \frac{\lambda r^H x\left(I - xx^H\right)rr^H\left(I - xx^H\right)}{\|r\|_2^2 - |x^H r|^2}\Big], \qquad \textit{if } B = -B^H. \end{cases}$$

When $\lambda \in \mathbb{C} \setminus i\mathbb{R}$, *define*

$$\Delta A := \begin{cases} -xx^H Axx^H + \frac{1}{1+|\lambda|^2}\left[xr^H\left(I-xx^H\right)+\left(I-xx^H\right)rx^H\right] \\ + \frac{x^H Ax\ \left(I-xx^H\right)rr^H\left(I-xx^H\right)}{\|r\|_2^2-|x^H r|^2}, & \textit{if } A = A^H, \\ -xx^H Axx^H + \frac{1}{1+|\lambda|^2}\left[\left(I-xx^H\right)rx^H - xr^H\left(I-xx^H\right)\right] \\ +\frac{x^H Ax\left(I-xx^H\right)rr^H\left(I-xx^H\right)}{\|r\|_2^2-|x^H r|^2}, & \textit{if } A = -A^H. \end{cases}$$

$$\Delta B := \begin{cases} -xx^H Bxx^H + \frac{1}{1+|\lambda|^2}\left[\lambda xr^H\left(I-xx^H\right)+\overline{\lambda}\left(I-xx^H\right)rx^H\right] \\ + \frac{x^H Bx\ \left(I-xx^H\right)rr^H\left(I-xx^H\right)}{\|r\|_2^2-|x^H r|^2}, & \textit{if } B = B^H, \\ -xx^H Bxx^H - \frac{\lambda xr^H\left(I-xx^H\right)}{1+|\lambda|^2} + \frac{\overline{\lambda}\left(I-xx^H\right)rx^H}{1+|\lambda|^2} + \frac{x^H Bx\left(I-xx^H\right)rr^H\left(I-xx^H\right)}{\|r\|_2^2-|x^H r|^2}, \\ & \textit{if } B = -B^H. \end{cases}$$

Consider $\Delta\mathrm{L}(z) := \Delta A + z\Delta B$. *Then* $\Delta\mathrm{L} \in \mathbb{S}$, $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$, *and* $\|\Delta\mathrm{L}\| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. First, suppose that L is H-even. Next, assume that $\lambda \in i\mathbb{R}$. Then it follows that $x^H r \in \mathbb{R}$. Now from the proof of Theorem 3.7, we have

$$\Delta A = Q\begin{pmatrix} \frac{1}{1+|\lambda|^2}x^H r & \frac{1}{1+|\lambda|^2}(Q_1^H r)^H \\ \frac{1}{1+|\lambda|^2}Q_1^H r & A_1 \end{pmatrix}Q^H,$$

$$\Delta B := Q\begin{pmatrix} \frac{\overline{\lambda}}{1+|\lambda|^2}x^H r & -\frac{\lambda}{1+|\lambda|^2}(Q_1^H r)^H \\ \frac{\overline{\lambda}}{1+|\lambda|^2}Q_1^H r & B_1 \end{pmatrix}Q^H$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. For $\mu_{\Delta A} := \frac{\|r\|_2}{1+|\lambda|^2}$, $\mu_{\Delta B} := \frac{|\lambda|\ \|r\|_2}{1+|\lambda|^2}$, by the DKW Theorem 4.1 we have

$$A_1 = -\frac{x^H r\ \left(Q_1^H r\right)\left(Q_1^H r\right)^H}{(1+|\lambda|^2)\ (\|r\|_2^2 - |x^H r|^2)} \text{ and } B_1 = \frac{\lambda\ x^H r\ \left(Q_1^H r\right)\left(Q_1^H r\right)^H}{(1+|\lambda|^2)\ (\|r\|_2^2 - |x^H r|^2)}.$$

This gives $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = (\|\Delta A\|_2^2 + \|\Delta B\|_2^2)^{1/2} = \frac{\|r\|_2}{\|(1,\lambda)\|_2}$. Simplifying expressions for ΔA and ΔB, we obtain the desired result.

Now suppose that $\lambda \in \mathbb{C} \setminus i\mathbb{R}$. The again from the proof of Theorem 3.7, we have

$$\Delta A = Q\begin{pmatrix} -x^H Ax & \frac{1}{1+|\lambda|^2}\left(Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2}Q_1^H r & A_1 \end{pmatrix}Q^H,$$

$$\Delta B := Q\begin{pmatrix} -x^H Bx & -\frac{\lambda}{1+|\lambda|^2}\left(Q_1^H r\right)^H \\ \frac{\overline{\lambda}}{1+|\lambda|^2}Q_1^H r & B_1 \end{pmatrix}Q^H.$$

For $\mu_{\Delta A} = \sqrt{|x^H Ax|^2 + \frac{\|r\|_2^2 - |x^H r|^2}{(1+|\lambda|^2)^2}}$ and $\mu_{\Delta B} = \sqrt{|x^H Bx|^2 + \frac{|\lambda|^2(\|r\|_2^2 - |x^H r|^2)}{(1+|\lambda|^2)^2}}$, by the DKW Theorem 4.1, we have

$$A_1 = \frac{x^H Ax}{\|r\|_2^2 - |x^H r|^2}\left(Q_1^H r\right)\left(Q_1^H r\right)^H \text{ and } B_1 = \frac{x^H Bx}{\|r\|_2^2 - |x^T r|^2}\left(Q_1^H r\right)\left(Q_1^H r\right)^H.$$

Consequently, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{|x^HAx|^2 + |x^HBx|^2 + \frac{\|r\|_2^2 - |x^Hr|^2}{1+|\lambda|^2}}$. Now, simplifying the expressions for ΔA and ΔB, we obtain the desired results.

When L is H-odd, the desired results follow by interchanging the role of A and B. □

As before, the above results are easily extended to the case of general structured pencils where the coefficient matrices are elements of Jordan and/or Lie algebras. In particular, for the pencil $\mathrm{L}(z) := A + zB$, where A is Hamiltonian and B is skew-Hamiltonian, we have the following result.

Theorem 4.7. *Let $\mathbb{S}$ be the space of pencils of the form $\mathrm{L}(z) = A + zB$, where A is Hamiltonian and B is skew-Hamiltonian. Let $\mathrm{L} \in \mathbb{S}$ and $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$. Set $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \eta(\lambda, x, \mathrm{L}), & \textit{if } \lambda \in i\mathbb{R} \\ \sqrt{|x^HJAx|^2 + |x^HJBx|^2 + \dfrac{\|r\|_2^2 - |x^HJr|^2}{1+|\lambda|^2}}, & \textit{if } \lambda \in \mathbb{C} \setminus i\mathbb{R}. \end{cases}$$

Now we consider palindromic pencils.

Theorem 4.8. *Let $\mathbb{S}$ be the space of T-palindromic pencils and $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) := A + zA^T$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{2}\,\sqrt{|x^TAx|^2 + \dfrac{|\lambda|^2\ (\|r\|_2^2 - |x^Tr|^2)}{(1+|\lambda|^2)^2}}, & \textit{if } |\lambda| > 1, \\ \sqrt{2}\,\sqrt{|x^TAx|^2 + \dfrac{\|r\|_2^2 - |x^Tr|^2}{(1+|\lambda|^2)^2}}, & \textit{if } |\lambda| \leq 1 \textit{ and } \lambda \neq \pm 1, \\ \eta(\lambda, x, \mathrm{L}), & \textit{if } \lambda = \pm 1. \end{cases}$$

Now define

$$\Delta A := \begin{cases} -\overline{x}x^TAxx^H + \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}\overline{x}r^T\left(I - xx^H\right) + \left(I - \overline{x}x^T\right)rx^H\right] \\ + \frac{\overline{\lambda}\ \overline{x^TAx}\ \left(I - \overline{x}x^T\right)rr^T\left(I - xx^H\right)}{|\lambda|^2\ (\|r\|_2^2 - |x^Tr|^2)}, & \textit{if } |\lambda| > 1, \\ -\overline{x}x^TAxx^H + \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}\overline{x}r^T\left(I - xx^H\right) + \left(I - \overline{x}x^T\right)rx^H\right] \\ + \frac{\overline{\lambda}\ \overline{x^TAx}\ \left(I - \overline{x}x^T\right)rr^T\left(I - xx^H\right)}{\|r\|_2^2 - |x^Tr|^2}, & \textit{if } |\lambda| \leq 1 \textit{ and } \lambda \neq -1, \\ \frac{1}{1+|\lambda|^2}\left[\overline{\lambda}\overline{x}r^T\left(\mathrm{I} - xx^H\right) + \left(\mathrm{I} - \overline{x}x^T\right)rx^H\right], & \textit{if } \lambda = -1. \end{cases}$$

Consider the pencil $\Delta\mathrm{L}(z) := \Delta A + z(\Delta A)^T$. Then $\Delta\mathrm{L} \in \mathbb{S}$, $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ and $|||\Delta\mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. Suppose that $\lambda \neq -1$. Then from the proof of Theorem 3.5, we have

$$\Delta A = \overline{Q}\begin{pmatrix} -x^TAx & \frac{\overline{\lambda}}{1+|\lambda|^2}\left(Q_1^Tr\right)^T \\ \frac{1}{1+|\lambda|^2}Q_1^Tr & A_1 \end{pmatrix}Q^H$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. Now for

$$\mu_{\Delta A} := \begin{cases} \sqrt{|x^TAx|^2 + \frac{|\lambda|^2(\|r\|_2^2 - |x^Tr|^2)}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| > 1, \\ \sqrt{|x^TAx|^2 + \frac{\|r\|_2^2 - |x^Tr|^2}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| \leq 1, \end{cases}$$

by the DKW Theorem 4.1, we have

$$A_1 = \begin{cases} \frac{\overline{\lambda}\ \overline{x^TAx}}{|\lambda|^2\ (\|r\|_2^2 - |x^Tr|^2)} Q_1^T r \left(Q_1^T r\right)^T, & \text{if } |\lambda| > 1, \\ \frac{\overline{\lambda}\ \overline{x^TAx}}{\|r\|_2^2 - |x^Tr|^2} Q_1^T r \left(Q_1^T r\right)^T, & \text{if } |\lambda| \le 1. \end{cases}$$

Consequently, we have

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{2}\ \sqrt{|x^TAx|^2 + \frac{|\lambda|^2\ (\|r\|_2^2 - |x^Tr|^2)}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| > 1, \\ \sqrt{2}\ \sqrt{|x^TAx|^2 + \frac{\|r\|_2^2 - |x^Tr|^2}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| \le 1. \end{cases}$$

Now simplifying the expression for ΔA, we obtain the desired results.

Next, suppose that $\lambda = -1$. Then again from the proof of Theorem 3.5, we have

$$\Delta A = \overline{Q} \begin{pmatrix} 0 & \frac{\overline{\lambda}}{1+|\lambda|^2} \left(Q_1^T r\right)^T \\ \frac{1}{1+|\lambda|^2} Q_1^T r & A_1 \end{pmatrix} Q^H. \text{ For } \mu_{\Delta A} := \frac{\|r\|_2}{1+|\lambda|^2},$$

by the DKW Theorem 4.1, we have $A_1 = 0$. Hence, $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \frac{1}{\sqrt{2}}\|r\|_2$. Simplifying the expression for ΔA, we obtain the desired result. □

For H-palindromic pencils we have the following.

THEOREM 4.9. *Let $\mathbb{S}$ be the space of H-palindromic pencils and $\mathrm{L} \in \mathbb{S}$ be given by $\mathrm{L}(z) := A + zA^H$. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ and $r := -\mathrm{L}(\lambda)x$. Then we have*

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{2}\ \sqrt{|x^HAx|^2 + \dfrac{|\lambda|^2\ (\|r\|_2^2 - |x^Hr|^2)}{(1+|\lambda|^2)^2}}, & \textit{if } |\lambda| > 1, \\ \sqrt{2}\ \sqrt{|x^HAx|^2 + \dfrac{\|r\|_2^2 - |x^Hr|^2}{(1+|\lambda|^2)^2}}, & \textit{if } |\lambda| < 1, \\ \eta(\lambda, x, \mathrm{L}), & \textit{if } |\lambda| = 1. \end{cases}$$

Now define

$$\Delta A := \begin{cases} -xx^HAxx^H + \frac{1}{1+|\lambda|^2}\left[\lambda x r^H \left(I - xx^H\right) + \left(I - xx^H\right) r x^H\right] \\ + \frac{\overline{x^HAx}\ \left(I - xx^H\right) r r^H \left(I - xx^H\right)}{\overline{\lambda}\,(\|r\|_2^2 - |x^Hr|^2)}, & \textit{if } |\lambda| > 1, \\ -xx^HAxx^H + \frac{1}{1+|\lambda|^2}\left[\lambda x r^H \left(I - xx^H\right) + \left(I - xx^H\right) r x^H\right] \\ + \frac{\lambda\ \overline{x^HAx}\ \left(I - xx^H\right) r r^H \left(I - xx^H\right)}{\|r\|_2^2 - |x^Hr|^2}, & \textit{if } |\lambda| < 1, \\ \frac{1}{1+|\lambda|^2}\left[r x^H + \lambda x r^H \left(\mathrm{I} - xx^H\right) - \frac{x^Hr\ \left(I - xx^H\right) r r^H \left(I - xx^H\right)}{(\|r\|_2^2 - |x^Hr|^2)}\right], & \textit{if } |\lambda| = 1, \end{cases}$$

and consider the pencil $\Delta\mathrm{L}(z) := \Delta A + z(\Delta A)^H$. Then $\Delta\mathrm{L} \in \mathbb{S}$, $\mathrm{L}(\lambda)x + \Delta\mathrm{L}(\lambda)x = 0$ and $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$.

Proof. First, suppose that $|\lambda| \neq 1$. Then from the proof of Theorem 3.8, we have

$$\Delta A = Q \begin{pmatrix} -x^HAx & \frac{\lambda}{1+|\lambda|^2} \left(Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2} Q_1^H r & A_1 \end{pmatrix} Q^H$$

such that $\Delta\mathrm{L}(\lambda)x + \mathrm{L}(\lambda)x = 0$. In this case, we have

$$\mu_{\Delta A} = \begin{cases} \sqrt{|x^HAx|^2 + \frac{|\lambda|^2(\|r\|_2^2 - |x^Hr|^2)}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| > 1, \\ \sqrt{|x^HAx|^2 + \frac{\|r\|_2^2 - |x^Hr|^2}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| < 1. \end{cases}$$

Hence, by the DKW Theorem 4.1, we have

$$A_1 = \begin{cases} \frac{\lambda \ \overline{x^H A x}}{|\lambda|^2 \ (\|r\|_2^2 - |x^H r|^2)} Q_1^H r (Q_1^H r)^H, & \text{if } |\lambda| > 1, \\ \frac{\lambda \ \overline{x^H A x}}{\|r\|_2^2 - |x^H r|^2} Q_1^H r (Q_1^H r)^H, & \text{if } |\lambda| < 1. \end{cases}$$

This gives

$$\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \begin{cases} \sqrt{2} \ \sqrt{|x^H A x|^2 + \frac{|\lambda|^2 \ (\|r\|_2^2 - |x^H r|^2)}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| > 1, \\ \sqrt{2} \ \sqrt{|x^H A x|^2 + \frac{\|r\|_2^2 - |x^H r|^2}{(1+|\lambda|^2)^2}}, & \text{if } |\lambda| < 1. \end{cases}$$

Simplifying the expression for ΔA, we obtain the desired result.

When $|\lambda| = 1$, again from the proof of Theorem 3.8, we have

$$\Delta A = Q \begin{pmatrix} \frac{x^H r}{1+|\lambda|^2} & \frac{\lambda}{1+|\lambda|^2} \left(Q_1^H r\right)^H \\ \frac{1}{1+|\lambda|^2} Q_1^H r & A_1 \end{pmatrix} Q^H.$$

Now, we have $\mu_{\Delta A} = \frac{\|r\|_2}{1+|\lambda|^2}$. Hence, by the DKW Theorem 4.1, we have

$$A_1 = -\frac{x^H r \ \left(I - x x^H\right) r r^H \left(I - x x^H\right)}{(1+|\lambda|^2)(\|r\|_2^2 - |x^H r|^2)}.$$

Consequently, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \frac{\|r\|_2}{\sqrt{2}}$. Simplifying the expression for ΔA, we obtain the desired result. □

Remark 4.10. Let $(\lambda, x) \in \mathbb{C} \times \mathbb{C}^n$ with $x^H x = 1$ and $\mathbb{S} \in \{T$-symmetric, T-skew-symmetric, T-odd, T-even, T-palindromic, H-Hermitian, H-skew-Hermitian, H-odd, H-even, H-palindromic$\}$. For $\mathrm{L} \in \mathbb{S}$, consider the set

$$\mathbb{S}(\lambda, x, \mathrm{L}) := \{\mathrm{K} \in \mathbb{S} : \mathrm{L}(\lambda)x + \mathrm{K}(\lambda)x = 0\}.$$

Then $\mathbb{S}(\lambda, x, \mathrm{L}) \neq \emptyset$ and $\min\{|\!|\!|\mathrm{K}|\!|\!| : \mathrm{K} \in \mathbb{S}(\lambda, x, \mathrm{L})\} = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$. Further,

$$\mathbb{S}_{\mathrm{opt}}(\lambda, x, \mathrm{L}) := \{\Delta\mathrm{L} \in \mathbb{S}(x, \lambda, \mathrm{L}) : |\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})\}$$

is an infinite set and is characterized by the DKW Theorem 4.1 by taking into account the nonzero contractions. Let $\Delta\mathrm{L} \in \mathbb{S}_{\mathrm{opt}}(\lambda, x, \mathrm{L})$. Then each pencil in $\mathbb{S}(\lambda, x, \mathrm{L})$ is of the form $\Delta\mathrm{L} + (I - x x^H)^* \, \mathrm{Z}(I - x x^H)$ for some $\mathrm{Z} \in \mathbb{S}$, where $*$ is either the transpose or the conjugate transpose depending upon the structure defined by $\mathbb{S}$. In other words, we have

$$\mathbb{S}(\lambda, x, \mathrm{L}) = \Delta\mathrm{L} + \left(I - x x^H\right)^* \ \mathbb{S} \ \left(I - x x^H\right).$$

Needless to mention that Remark 4.10 remains valid for structured pencils in $\mathbb{S}$ whose coefficient matrices are elements of Jordan and/or Lie algebras associated with a scalar product considered in the previous section. In such a case the $*$ in $(I - x x^H)^*$ is the adjoint induced by the scalar product that defines the Jordan and Lie algebras.

We now illustrate various structured and unstructured backward errors by numerical examples. We use MATLAB.7.0 for our computation. We generate A and B as follows:

```
>> randn('state',15), A = randn(50)+ i*randn(50); A = A ± A*;
 >> randn('state',25), B = randn(50)+i*randn(50); B = B ± B*;
```

For T-palindromic/H-palindromic pencils, we generate A and B by

```
>> randn('state',15), A = randn(50)+ i*randn(50); B = A*;
```

Here $A^* = A^T$ or $A^* = A^H$. Finally, we compute (λ, x) by

```
>> [V,D] = eig(A,B); λ = -D(2,2); x = V(:,2)/norm(V(:,2));
```

We denote by $\eta_F^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and $\eta_2^{\mathbb{S}}(\lambda, x, \mathrm{L})$ the backward error $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ when $\mathbb{C}^{n\times n}$ is equipped with the Frobenius norm and the spectral norm, respectively. Note that $\eta(\lambda, x, \mathrm{L})$ is the same for the spectral and the Frobenius norms. Then we have the following.

$\mathbb{S}$	$\eta(\lambda, x, \mathrm{L})$	$\eta_F^{\mathbb{S}}(\lambda, x, \mathrm{L})$	$\eta_2^{\mathbb{S}}(\lambda, x, \mathrm{L})$
T-symm	1.387705737323579e−014	1.959539856593202e−014	1.387705737323579e−014
T-skew-symm	1.796046101865378e−014	2.539992755905347e−014	1.796046101865378e−014
T-even	2.219610496439476e−014	3.211055813711074e−014	2.324926535413804e−014
T-odd	1.559070464273151e−014	2.204626223816091e−014	1.559075824083717e−014
T-palindromic	1.068704043320177e−014	1.512088705618463e−014	1.487010794022381e−014
H-Herm	2.076731533185186e−014	2.947235222707197e−014	2.106896507205170e−014
H-skew-Herm	1.714743310005108e−014	2.489567700503872e−014	1.811820338752170e−014
H-even	1.590165856939442e−014	2.299115486213681e−014	1.663718384482337e−014
H-odd	2.343032834027323e−014	3.472518481940936e−014	2.566511276851151e−014
H-palindromic	9.161344100487524e−015	1.298942035829892e−014	1.296310627878570e−014

Note that structured backward errors are bigger than or equal to unstructured backward errors but they are marginally so. On the other hand, structured condition numbers are less than or equal to unstructured condition numbers [13, 6]. Consequently, structured backward errors when combined with structured condition numbers provide almost the same approximate upper bounds on the errors in the computed eigenelements as do their unstructured counterparts. We mention that the MATLAB `eig` command does not ensure spectral symmetry in the computed eigenvalues.

5. Structured pseudospectra of structured pencils. Let $\mathrm{L} \in \mathbb{A}^{n\times n}$ be a regular pencil. For $\lambda \in \mathbb{C}$, the backward error of λ as an approximate eigenvalue of L is given by $\eta(\lambda, \mathrm{L}) := \min\{\eta(\lambda, x, \mathrm{L}) : x \in \mathbb{C}^n \text{ and } \|x\|_2 = 1\}$. Since $\eta(\lambda, x, \mathrm{L}) = \|\mathrm{L}(\lambda)x\|_2/\|(1, \lambda)\|_2$, it follows that for the spectral norm as well as for the Frobenius norm on $\mathbb{C}^{n\times n}$, we have $\eta(\lambda, \mathrm{L}) := \sigma_{\min}(\mathrm{L}(\lambda))/\|(1, \lambda)\|_2$. Similarly, we define structured backward error of an approximate eigenvalue λ of $\mathrm{L} \in \mathbb{S}$ by

$$\eta^{\mathbb{S}}(\lambda, \mathrm{L}) := \min\left\{\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) : x \in \mathbb{C}^n \text{ and } \|x\|_2 = 1\right\}.$$

Note that backward errors of approximate eigenvalues and pseudospectra of a pencil are closely related. For $\epsilon > 0$, the unstructured ϵ-pseudospectrum of L, denoted by $\Lambda_\epsilon(\mathrm{L})$, is given by [3]

$$\Lambda_\epsilon(\mathrm{L}) = \bigcup_{|||\Delta\mathrm{L}|||\leq\epsilon} \{\Lambda(\mathrm{L} + \Delta\mathrm{L}) : \Delta\mathrm{L} \in \mathbb{A}^{n\times n}\}.$$

Obviously, we have $\Lambda_\epsilon(\mathrm{L}) = \{z \in \mathbb{C} : \eta(z, \mathrm{L}) \leq \epsilon\}$, assuming, for simplicity, that $\infty \notin \Lambda_\epsilon(\mathrm{L})$. For the sake of simplicity, for rest of this section, we make an implicit assumption that $\infty \notin \Lambda_\epsilon(\mathrm{L})$. We observe the following.

- Since $\eta(\lambda, \mathrm{L})$ is the same for the spectral norm and the Frobenius norm on $\mathbb{C}^{n\times n}$, it follows that $\Lambda_\epsilon(\mathrm{L})$ is the same for the spectral and the Frobenius norms.

Similarly, when $\mathrm{L} \in \mathbb{S}$, we define the structured ϵ-pseudospectrum of L, denoted by $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L})$, by

$$\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) := \bigcup_{|\!|\!| \Delta\mathrm{L} |\!|\!| \leq \epsilon} \{\Lambda(\mathrm{L} + \Delta\mathrm{L}) : \Delta\mathrm{L} \in \mathbb{S}\}.$$

Then it follows that $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) = \{z \in \mathbb{C} : \eta^{\mathbb{S}}(\lambda, \mathrm{L}) \leq \epsilon\}$.

Theorem 5.1. *Let $\mathbb{S} \in \{T$-symmetric, T-skew-symmetric$\}$ and $\mathrm{L} \in \mathbb{S}$. Let $\lambda \in \mathbb{C}$. Then for the spectral norm on $\mathbb{C}^{n\times n}$, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$ and $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) = \Lambda_\epsilon(\mathrm{L})$. For the Frobenius norm on $\mathbb{C}^{n\times n}$, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \sqrt{2}\,\eta(\lambda, \mathrm{L})$ and $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) = \Lambda_{\epsilon/\sqrt{2}}(\mathrm{L})$ when L is T-skew-symmetric, and $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$ and $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) = \Lambda_\epsilon(\mathrm{L})$ when L is T-symmetric.*

Proof. For the spectral norm, by Theorem 4.2, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \eta(\lambda, x, \mathrm{L})$ for all x. Consequently, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$. Hence, the result follows.

For the Frobenius norm, the result follows from Theorem 3.3 when L is T-skew-symmetric. So, suppose that L is T-symmetric. Then $\mathrm{L}(\lambda) \in \mathbb{C}^{n\times n}$ is symmetric. Consider the Takagi factorization $\mathrm{L}(\lambda) = U\Sigma U^T$, where U is unitary and Σ is a diagonal matrix containing singular values of $\mathrm{L}(\lambda)$ (appear in descending order). Set $\sigma := \Sigma(n, n)$ and $u := U(:, n)$. Then we have $\mathrm{L}(\lambda)\overline{u} = \sigma u$. Now define $\Delta A := -\frac{\sigma\, uu^T}{1+|\lambda|^2}, \Delta B := -\frac{\overline{\lambda}\,\sigma\, uu^T}{1+|\lambda|^2}$, and consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. Then $\Delta\mathrm{L}$ is T-symmetric and $\mathrm{L}(\lambda)\overline{u} + \Delta\mathrm{L}(\lambda)\overline{u} = 0$. Notice that, for the spectral norm and the Frobenius norm on $\mathbb{C}^{n\times n}$, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) \leq |\!|\!|\Delta\mathrm{L}|\!|\!| = \sigma/\|(1, \lambda)\|_2 = \eta(\lambda, \mathrm{L})$ and, hence, $\Lambda_\epsilon(\mathrm{L}) = \Lambda_\epsilon^{\mathbb{S}}(\mathrm{L})$. This completes the proof. □

When L is T-symmetric, the above proof shows how to construct a T-symmetric pencil $\Delta\mathrm{L}$ such that $\lambda \in \Lambda(\mathrm{L} + \Delta\mathrm{L})$ and $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, \mathrm{L})$. When L is T-skew-symmetric, using Takagi factorization of the complex skew-symmetric matrix $\mathrm{L}(\lambda)$, one can construct a T-skew-symmetric pencil $\Delta\mathrm{L}$ such that $\lambda \in \Lambda(\mathrm{L} + \Delta\mathrm{L})$ and $|\!|\!|\Delta\mathrm{L}|\!|\!| = \eta^{\mathbb{S}}(\lambda, \mathrm{L})$. Indeed, consider the Takagi factorization $\mathrm{L}(\lambda) = U\mathrm{diag}(d_1, \ldots, d_m)$ U^T, where U is unitary, $d_j := \left(\begin{smallmatrix} 0 & s_j \\ -s_j & 0 \end{smallmatrix}\right)$, $s_j \in \mathbb{C}$ is nonzero, and $|s_j|$ are singular values of $\mathrm{L}(\lambda)$. Here the blocks d_j appear in descending order of magnitude of $|s_j|$. Note that $\mathrm{L}(\lambda)\overline{U} = U\mathrm{diag}(d_1, \ldots, d_m)$. Let $u := U(:, n-1 : n)$. Then $\mathrm{L}(\lambda)\overline{u} = ud_m = ud_m u^T\overline{u}$. Now define

$$\Delta A := -\frac{ud_m u^T}{1+|\lambda|^2}, \;\; \Delta B := -\frac{\overline{\lambda}\, ud_m u^T}{1+|\lambda|^2}$$

and consider $\Delta\mathrm{L}(z) := \Delta A + z\Delta B$. Then $\Delta\mathrm{L}$ is T-skew-symmetric and $\mathrm{L}(\lambda)\overline{u} + \Delta\mathrm{L}(\lambda)\overline{u} = 0$. For the spectral norm on $\mathbb{C}^{n\times n}$, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = |\!|\!|\Delta\mathrm{L}|\!|\!| = \sigma_{\min}(\mathrm{L}(\lambda))/\|(1, \lambda)\|_2 = \eta(\lambda, \mathrm{L})$ and for the Frobenius norm on $\mathbb{C}^{n\times n}$, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = |\!|\!|\Delta\mathrm{L}|\!|\!| = \sqrt{2}\,\sigma_{\min}(\mathrm{L}(\lambda))/\|(1, \lambda)\|_2 = \sqrt{2}\,\eta(\lambda, \mathrm{L})$.

We denote the unit circle in $\mathbb{C}$ by $\mathbb{T}$, that is, $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$. Then for T-even and T-odd pencils we have the following.

Theorem 5.2. *Let $\mathbb{S} \in \{T$-even, T-odd$\}$ and $\mathrm{L} \in \mathbb{S}$. Let $\lambda \in \mathbb{T}$. Then for the Frobenius norm on $\mathbb{C}^{n\times n}$, we have $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \sqrt{2}\,\eta(\lambda, \mathrm{L})$ and $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap \mathbb{T} = \Lambda_{\epsilon/\sqrt{2}}(\mathrm{L}) \cap \mathbb{T}$.*

Proof. Let $\lambda \in \mathbb{T}$. Then by Theorem 3.4, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \frac{\sqrt{2}\,\|\mathrm{L}(\lambda)x\|_2}{\|(1,\lambda)\|_2}$ for all x such that $\|x\|_2 = 1$. Hence, taking minimum over $\|x\|_2 = 1$, we obtain the desired results. □

THEOREM 5.3. *Let* $\mathbb{S} \in \{H\text{-Hermitian, } H\text{-skew-Hermitian}\}$ *and* $\mathrm{L} \in \mathbb{S}$. *Let* $\lambda \in \mathbb{R}$. *Then for the spectral and the Frobenius norms on* $\mathbb{C}^{n\times n}$, *we have* $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$ *and* $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap \mathbb{R} = \Lambda_\epsilon(\mathrm{L}) \cap \mathbb{R}$. *Also when* $\lambda = \pm i$, *for the Frobenius norm, we have* $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \sqrt{2}\,\eta(\lambda, \mathrm{L})$.

Proof. Note that $\mathrm{L}(\lambda)$ is either Hermitian or skew-Hermitian. Let (μ, u) be an eigenpair of the matrix $\mathrm{L}(\lambda)$ such that $|\mu| = \sigma_{\min}(\mathrm{L}(\lambda))$ and $u^H u = 1$. Then $\mathrm{L}(\lambda)u = \mu u$. Define $\Delta A := -\frac{\mu\, uu^H}{1+|\lambda|^2}$, $\Delta B := -\frac{\overline{\lambda}\mu\, uu^H}{1+|\lambda|^2}$, and consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z\Delta B$. Then $\Delta\mathrm{L} \in \mathbb{S}$ and $\lambda \in \Lambda(\mathrm{L} + \Delta\mathrm{L})$. Further, for the spectral and the Frobenius norms, we have $|||\Delta\mathrm{L}||| = \sigma_{\min}(\mathrm{L}(\lambda))/\|(1, \lambda)\|_2$. Hence, the result follows. Finally, when $\lambda = \pm i$, the result follows from Theorem 3.6. □

THEOREM 5.4. *Let* $\mathbb{S} \in \{H\text{-even, } H\text{-odd}\}$ *and* $\mathrm{L} \in \mathbb{S}$. *Let* $\lambda \in i\mathbb{R}$. *Then for the spectral and the Frobenius norms on* $\mathbb{C}^{n\times n}$, *we have* $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$ *and* $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap i\mathbb{R} = \Lambda_\epsilon(\mathrm{L}) \cap i\mathbb{R}$. *Also when* $\lambda = \pm 1$, *for the Frobenius norm, we have* $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \sqrt{2}\,\eta(\lambda, \mathrm{L})$.

Proof. Note for $\lambda \in i\mathbb{R}$, the matrix $\mathrm{L}(\lambda)$ is again either Hermitian or skew-Hermitian. Hence, the result follows from the proof of Theorem 5.3. When $\lambda = \pm 1$, the result follows from Theorem 3.7. □

We mention that the above results are easily extended to the case of general structured pencils where the coefficients matrices are elements of Jordan and/or Lie algebras.

Finally, for T-palindromic and H-palindromic pencils we have the following result.

THEOREM 5.5. *Let* $\mathbb{S}$ *be the space of* T*-palindromic pencils and* $\mathrm{L} \in \mathbb{S}$. *Let* $\lambda \in i\mathbb{R}$. *Then for the Frobenius norm on* $\mathbb{C}^{n\times n}$, $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \sqrt{2}\,\eta(\lambda, \mathrm{L})$ *and* $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap i\mathbb{R} = \Lambda_{\epsilon/\sqrt{2}}(\mathrm{L}) \cap i\mathbb{R}$.

Proof. Let $\lambda \in i\mathbb{R}$. Then by Theorem 3.5, we have $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \sqrt{2}\,\|\mathrm{L}(\lambda)x\|_2/\|(1, \lambda)\|_2$ for all x such that $\|x\|_2 = 1$. Hence, taking minimum over $\|x\|_2 = 1$, we obtain the desired results. □

THEOREM 5.6. *Let* $\mathbb{S}$ *be the space of* H*-palindromic matrix pencils and* $\mathrm{L} \in \mathbb{S}$. *Let* $\lambda \in \mathbb{T}$. *Then for the spectral and the Frobenius norms on* $\mathbb{C}^{n\times n}$, *we have* $\eta^{\mathbb{S}}(\lambda, \mathrm{L}) = \eta(\lambda, \mathrm{L})$ *and* $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap \mathbb{T} = \Lambda_\epsilon(\mathrm{L}) \cap \mathbb{T}$.

Proof. Let L be given by $\mathrm{L}(\lambda) = A + \lambda A^H$. For $\lambda \in \mathbb{T}$, we have $\mathrm{L}(\lambda)^H = \overline{\lambda}\mathrm{L}(\lambda)$. This shows that $\mathrm{L}(\lambda)$ is a normal matrix. Let (μ, u) be an eigenpair of $\overline{\lambda}\mathrm{L}(\lambda)$ such that $|\mu| = \sigma_{\min}(\overline{\lambda}\mathrm{L}(\lambda)) = \sigma_{\min}(\mathrm{L}(\lambda))$. Define $\Delta A := -\frac{1}{2}\lambda\mu\, uu^H$ and consider the pencil $\Delta\mathrm{L}(z) = \Delta A + z(\Delta A)^H$. Noting the fact that $\overline{\lambda}\mathrm{L}(\lambda)u = \mu u$ and $\overline{\mu}u = (\overline{\lambda}\mathrm{L}(\lambda))^H u = \mu u$, we have $\mathrm{L}(\lambda)u + \Delta\mathrm{L}(\lambda)u = \lambda\mu\, u - \lambda\mu u = 0$. Further, we have $|||\Delta\mathrm{L}||| = |\mu|/\sqrt{2} = \sigma_{\min}(\mathrm{L}(\lambda))/\|(1, \lambda)\|_2 = \eta(\lambda, \mathrm{L})$. Hence, the results follow. □

For structured pencils, we have seen that $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap \Omega = \Lambda_\epsilon(\mathrm{L}) \cap \Omega$ for appropriate $\Omega \subset \mathbb{C}$. We now show that this result plays an important role in solving certain distance problems associated with structured pencils. For illustration, we consider an H-even pencil $\mathrm{L}(z) = A + zB$. Then by Theorem 5.4, we have $\Omega = i\mathbb{R}$, that is, $\Lambda_\epsilon^{\mathbb{S}}(\mathrm{L}) \cap i\mathbb{R} = \Lambda_\epsilon(\mathrm{L}) \cap i\mathbb{R}$. The spectrum of L has Hamiltonian eigensymmetry, that is, the eigenvalues of L occur in $\lambda, -\overline{\lambda}$ pairs so that the eigenvalues are symmetric with respect to the imaginary axis $i\mathbb{R}$.

Question: Suppose that L is H-even and is of size $2n$. Suppose also that L has n eigenvalues in the open left half complex plane and n eigenvalues in the open right half complex plane. What is the smallest value of $|||\Delta\mathrm{L}|||$ such that $\Delta\mathrm{L}$ is H-even and $\mathrm{L} + \Delta\mathrm{L}$ has a purely imaginary eigenvalue?

Distance problems of this kind occur in many applications (see, for example, [8]). Let $\mathrm{d}(\mathrm{L})$ denote the smallest value of $|||\Delta\mathrm{L}|||$ such that $\mathrm{L} + \Delta\mathrm{L}$ has a purely imaginary eigenvalue. Then by Theorem 5.4, we have

$$\mathrm{d}(\mathrm{L}) = \inf_{t\in\mathbb{R}} \eta^{\mathbb{S}}(it, \mathrm{L}) = \min\left\{\epsilon : \Lambda^{\mathbb{S}}_{\epsilon}(\mathrm{L}) \cap i\mathbb{R} \neq \emptyset\right\} = \min\{\epsilon : \Lambda_{\epsilon}(\mathrm{L}) \cap i\mathbb{R} \neq \emptyset\} = \inf_{t\in\mathbb{R}} \eta(it, \mathrm{L}).$$

Hence, d(L) can be read off from the unstructured pseudospectra of L. Note that $\eta(z, \mathrm{L}) = \sigma_{\min}(A + zB)/\sqrt{1 + |z|^2}$. Thus, if the infimum of $\eta(z, \mathrm{L})$ is attained at $\mu \in i\mathbb{R}$, then as in the proof of Theorem 5.4 we can construct an H-even pencil $\Delta\mathrm{L}$ such that μ is an eigenvalue of $\mathrm{L} + \Delta\mathrm{L}$ and that $|||\Delta\mathrm{L}||| = \eta(\mu, \mathrm{L}) = \mathrm{d}(\mathrm{L})$.

6. Conclusions. We have analyzed structured backward perturbations of ten special classes of structured pencils. Given a structured pencil $\mathrm{L} \in \mathbb{S}$ and an approximate eigenpair (λ, x), we have determined the structured backward error $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and a structured pencil $\Delta\mathrm{L} \in \mathbb{S}$ such that $|||\Delta\mathrm{L}||| = \eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$. We have shown that such a $\Delta\mathrm{L}$ is unique when $\mathbb{C}^{n\times n}$ is equipped with the Frobenius norm. On the other hand, we have shown that there are infinitely many such $\Delta\mathrm{L}$ when $\mathbb{C}^{n\times n}$ is equipped with the spectral norm and that all such $\Delta\mathrm{L}$ are characterized by adopting Davis–Kahan–Weinberger solutions of norm-preserving dilation problem for structured matrices. More specifically, for the Frobenius norm on $\mathbb{C}^{n\times n}$, we have determined $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and a unique $\Delta\mathrm{L}$ for T-symmetric (Theorem 3.1), T-skew-symmetric (Theorem 3.3), T-even and T-odd (Theorem 3.4), T-palindromic (Theorem 3.5), H-Hermitian and H-skew-Hermitian (Theorem 3.6), H-even and H-odd (Theorem 3.7), and H-palindromic (Theorem 3.8) pencils. On the other hand, when $\mathbb{C}^{n\times n}$ is equipped with the spectral norm, we have shown that $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L}) = \eta(\lambda, x, \mathrm{L})$ for T-symmetric and T-skew-symmetric pencils (Theorem 4.2), and have determined $\eta^{\mathbb{S}}(\lambda, x, \mathrm{L})$ and a $\Delta\mathrm{L}$ for T-even and T-odd (Theorem 4.4), H-Hermitian and H-skew-Hermitian (Theorem 4.5), H-even and H-odd (Theorem 4.6), T-palindromic (Theorem 4.8), and H-palindromic (Theorem 4.9) pencils. We have shown that structured and unstructured pseudospectra are the same for T-symmetric and T-skew-symmetric pencils. For the rest of the structures, we have shown that $\Lambda^{\mathbb{S}}_{\epsilon}(\mathrm{L}) \cap \Omega = \Lambda_{\epsilon}(\mathrm{L}) \cap \Omega$ for some $\Omega \subset \mathbb{C}$. We have also shown that the equality $\Lambda^{\mathbb{S}}_{\epsilon}(\mathrm{L}) \cap \Omega = \Lambda_{\epsilon}(\mathrm{L}) \cap \Omega$ plays an important role in constructing solution of certain distance problems.

REFERENCES

[1] B. Adhikari and R. Alam, *Structured backward errors and pseudospectra of structured matrix polynomials,* preprint.

[2] B. Adhikari and R. Alam, *Structured mapping problems for linearly structured matrices,* preprint.

[3] S. Ahmad, R. Alam, and R. Byers, *On pseudospectra, critical points and multiple eigenvalues of matrix pencils,* SIAM J. Matrix Anal. Appl., to appear.

[4] 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.

[5] 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.

[6] S. Bora, *Structured eigenvalue conditioning and backward error of a class of polynomial eigenvalue problems,* Preprint 417-2007, Institute of Mathematics, Technische Universität Berlin, 2007.

[7] 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.

[8] R.W. Freund and F. Jarre, *An extension of the positive real lemma to descriptor systems,* Optim. Methods Software, 19 (2004), pp. 69–87.

[9] 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.

[10] 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.

[11] T.-M. Hwang, W.-W. Lin, and V. Mehrmann, *Numerical solution of quadratic eigenvalue problems with structure-preserving methods,* SIAM J. Sci. Comp., 24 (2003), pp. 1283–1302.

[12] M. Karow, D. Kressner, and F. Tisseur, *Structured eigenvalue condition numbers,* SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1052–1068.

[13] D. Kressner, M.J. Pelaez, and J. Moro, *Structured Hölder condition numbers for multiple eigenvalues,* Uminf report, Department of Computing Science, Umeä University, Sweden, October 2006.

[14] X.-G. Liu and Z.-X. Wang, *A note on the backward errors for Hermite eigenvalue problems,* Appl. Math. Comput., 165 (2005), pp. 405–417.

[15] D.S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Vector spaces of linearizations for matrix polynomials,* SIAM J. Matrix Anal. Appl., 28 (2006), pp. 971–1004.

[16] D.S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann, *Palindromic polynomial eigenvalue problems: Good vibrations from good linearizations,* SIAM J. Matrix Anal. Appl., 28 (2006), pp. 1029–1051.

[17] D.S. Mackey, N. Mackey, and F. Tisseur, *Structured factorizations in scalar product spaces,* SIAM J. Matrix Anal. Appl., 27 (2006), pp. 821–850.

[18] D.S. Mackey, N. Mackey, and F. Tisseur, *Structured mapping problems for matrices associated with scalar products, Part* I: *Lie and Jordan Algebras,* SIAM J. Matrix Anal. Appl., 29 (2008), pp. 1389–1410.

[19] V. Mehrmann and H. Xu, *Structured Jordan canonical forms for structured matrices that are Hermitian, skew Hermitian or unitary with respect to indefinite inner products,* Electron. J. Linear Algebra, 5 (1999), pp. 67–103.

[20] 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.

[21] V. Mehrmann and D. Watkins, *Polynomial eigenvalue problems with Hamiltonian structure,* Electron. Trans. Numer. Anal., 13 (2002), pp. 106–113.

[22] S.M. Rump, *Eigenvalues, pseudospectrum and structured perturbation,* Linear Algebra Appl., 413 (2006), pp. 567–593.

[23] F. Tisseur, *Stability of structured Hamiltonian eigensolvers,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 103–125.

[24] F. Tisseur, *A chart of backward errors and condition numbers for singly and doubly structured eigenvalue problems,* SIAM J. Matrix Anal. Appl., 24 (2003), pp. 877–897.

[25] F. Tisseur and N.J. Higham, *Structured pseudospectra for polynomial eigenvalue problems, with applications,* SIAM J. Matrix Anal. Appl., 23 (2001), pp. 187–208.

[26] L.N. Trefethen and M. Embree, *Spectra and Pseudospectra: The behaviour of nonnormal matrices and operators,* Princeton University Press, Princeton, NJ, 2005.